CLC **Cancer Research** Workbench

REFERENCE MANUAL

Manual for

CLC Cancer Research Workbench 1.5

Windows, Mac OS X and Linux

April 16, 2015

This software is for research purposes only.

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Part I

Introduction

Chapter 1

Introduction to *CLC Cancer Research Workbench*

Contents

Welcome to *CLC Cancer Research Workbench* — a software package supporting your daily bioinformatics work.

CLC Cancer Research Workbench is a virtual lab bench. It gives you access to atomic level insights in protein-ligand interaction, and allows new ideas for improved binders to be quickly tested and visualized.

CLC Cancer Research Workbench comes with drug design and sequence analysis tools that allow you to analyze and visualize protein targets and ligands binding to them. The interface is designed to communicate with all chemists, with no assumptions about their level of theoretical training.

We strongly encourage you to read this user manual in order to get the best possible basis for working with the software package.

We also recommend that you check out our cancer web page that can be found here: `http://clccancer.com/`

This software is for research purposes only.

1.1 Contact information

The *CLC Cancer Research Workbench* is developed by:

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
8000 Aarhus C
Denmark

http://www.clcbio.com

VAT no.: DK 28 30 50 87

Telephone: 45 70 22 32 44
Fax: +45 86 20 12 22

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you can contact us through the support team as described here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Getting_help.html.

1.2 Download and installation

The *CLC Cancer Research Workbench* is developed for Windows, Mac OS X and Linux. The software for either platform can be downloaded from http://www.clcbio.com/download.

1.2.1 Program download

The program is available for download on http://www.clcbio.com/download.

Before you download the program you are asked to fill in the **Download** dialog.

In the dialog you must choose:

- Which operating system you use
- Whether you would like to receive information about future releases

Depending on your operating system and your Internet browser, you are taken through some download options.

When the download of the installer (an application which facilitates the installation of the program) is complete, follow the platform specific instructions below to complete the installation procedure. [1]

1.2.2 Installation on Microsoft Windows

Starting the installation process is done in one of the following ways:

[1] You must be connected to the Internet throughout the installation process.

When you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
- Choose a name for the Start Menu folder used to launch *CLC Cancer Research Workbench* and click **Next**.
- Choose if *CLC Cancer Research Workbench* should be used to open CLC files and click **Next**.
- Choose where you would like to create shortcuts for launching *CLC Cancer Research Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Cancer Research Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Cancer Research Workbench*.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Cancer Research Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from the Start Menu or from one of the shortcuts you chose to create.

1.2.3 Installation on Mac OS X

Starting the installation process is done in the following way:

When you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.
Launch the installer by double-clicking on the "*CLC Cancer Research Workbench*" icon.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
- Choose if *CLC Cancer Research Workbench* should be used to open CLC files and click **Next**.

- Choose whether you would like to create desktop icon for launching *CLC Cancer Research Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Cancer Research Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Cancer Research Workbench*.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Cancer Research Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from your Applications folder, or from the desktop shortcut you chose to create. If you like, you can drag the application icon to the dock for easy access.

1.2.4 Installation on Linux with an installer

Navigate to the directory containing the installer and execute it. This can be done by running a command similar to:

```
# sh CLCCancerResearchWorkbench_1_JRE.sh
```

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
  For a system-wide installation you can choose for example /opt or /usr/local. If you do not have root privileges you can choose to install in your home directory.
- Choose where you would like to create symbolic links to the program
  DO NOT create symbolic links in the same location as the application.
  Symbolic links should be installed in a location which is included in your environment PATH. For a system-wide installation you can choose for example /usr/local/bin. If you do not have root privileges you can create a 'bin' directory in your home directory and install symbolic links there. You can also choose not to create symbolic links.
- Wait for the installation process to complete and click **Finish**.

If you choose to create symbolic links in a location which is included in your PATH, the program can be executed by running the command:

```
# clccancerresearchwb1
```

Otherwise you start the application by navigating to the location where you choose to install it and running the command:

```
# ./clccancerresearchwb1
```

1.2.5 Installation on Linux with an RPM-package

Navigate to the directory containing the rpm-package and install it using the rpm-tool by running a command similar to:

```
# rpm -ivh CLCCancerResearchWorkbench_1_JRE.rpm
```

Installation of RPM-packages usually requires root-privileges.

When the installation process is finished the program can be executed by running the command:

```
# clccancerresearchwb1
```

1.3 System requirements

- Windows Vista, Windows 7, Windows 8 or Windows Server 2008
- Mac OS X 10.7 or later.
- Linux: Red Hat 5.0 or later. SUSE 10.2 or later. Fedora 6 or later.
- 8 GB RAM required
- 16 GB RAM recommended
- 1024 x 768 display required
- 1600 x 1200 display recommended
- Intel or AMD CPU required
- Minimum 10 GB free disc space in the tmp directory
- Minimum 80 GB free disc space required in the CLC_References directory (if you are not connected to a server). If you have less free disc space available it is possible to change the reference data location. How to do this is described in section 10.1.4
- **Special requirements for read mapping**. The numbers below give minimum and recommended memory for systems running mapping and analysis tasks. The requirements suggested are based on the genome size. Systems with less memory than specified below will benefit from installing the legacy read mapper plugin (see http://www.clcbio.com/plugins). This is slower than the standard mapper but adjusts to the amount of memory available.
  - **Human (3.2 gigabases)** and **Mouse (2.7 gigabases)**
    - Minimum: 6 GB RAM
    - Recommended: 8 GB RAM

1.3.1 Limitations on maximum number of cores

Most modern CPUs implements hyper threading or a similar technology which makes each physical CPU core appear as two logical cores on a system. In this manual the therm "core" always refer to a logical core unless otherwise stated.

For static licenses, there is a limitation on the number of logical cores on the computer. If there are more than 64 logical cores, the *CLC Cancer Research Workbench* cannot be started. In this case, a network license is needed (read more at http://www.clcbio.com/desktop-applications/licensing/).

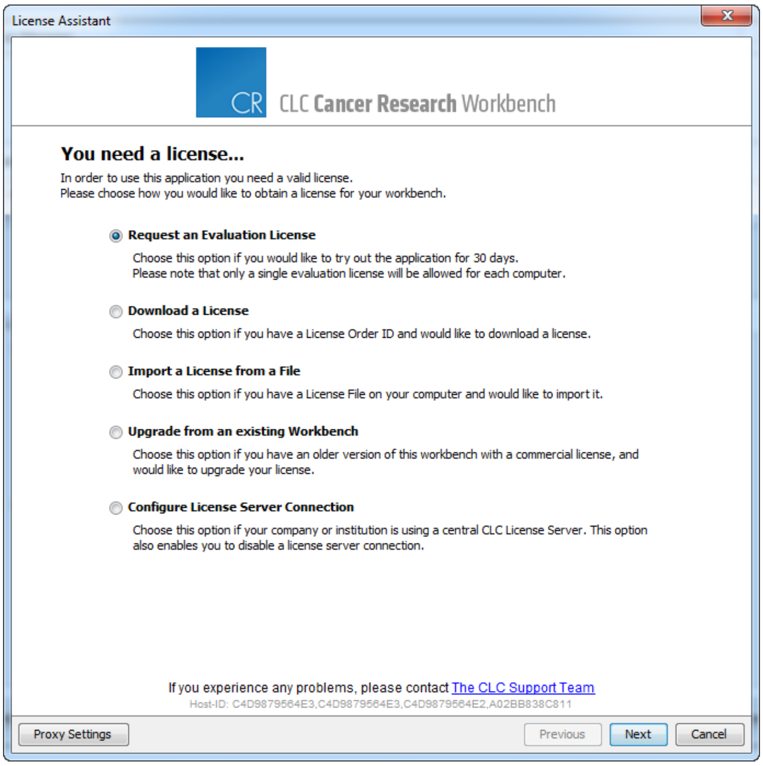
1.4 Licenses

When you have installed the *CLC Cancer Research Workbench*, and start it for the first time, you will meet the license assistant, shown in figure 1.1. The **License Manager** can also be accessed from the menu bar in the Workbench:

Help | License Manager

This can be useful if you wish to use a different license or want to view information about the license(s) the Workbench is currently using. The **License Manager** is described in detail in section 1.4.5 and can be seen in figure 1.23.

To install a license, you must be running the program in administrative mode [2].



You need a license...

In order to use this application you need a valid license.
Please choose how you would like to obtain a license for your workbench.

● **Request an Evaluation License**
Choose this option if you would like to try out the application for 30 days.
Please note that only a single evaluation license will be allowed for each computer.

○ **Download a License**
Choose this option if you have a License Order ID and would like to download a license.

○ **Import a License from a File**
Choose this option if you have a License File on your computer and would like to import it.

○ **Upgrade from an existing Workbench**
Choose this option if you have an older version of this workbench with a commercial license, and would like to upgrade your license.

○ **Configure License Server Connection**
Choose this option if your company or institution is using a central CLC License Server. This option also enables you to disable a license server connection.

If you experience any problems, please contact The CLC Support Team
Host-ID: C4D9879564E3,C4D9879564E3,C4D9879564E2,A02BB838C811

Figure 1.1: *The license assistant showing you the options for getting started.*

The following options are available. They are described in detail in the sections that follow.

[2] How to do this differs for different operating systems. To run the program in administrator mode on Windows Vista, or 7, right-click the program shortcut and choose "Run as Administrator.

- **Request an evaluation license.** Request a fully functional, time-limited license (see below).
- **Download a license.** Use the license order ID received when you purchase the software to download and install a license file.
- **Import a license from a file.** Import an existing license file, for example a file downloaded from the web-based licensing system.
- **Upgrade license.** If you have used a previous version of the *CLC Cancer Research Workbench*, and you are entitled to upgrade to a new major version of the *CLC Cancer Research Workbench 1.5*, select this option to upgrade your license file.
- **Configure license server connection.** If your organization has a CLC License Server, select this option to configure the connection to it.

Select the appropriate option and click on button labeled **Next**.

To use the Download option in the License Manager, your machine must be able to access the external network. If this is not the case, please see section 1.4.6.

If for some reason you don't have a license order ID or access to a license, you can click the **Limited Mode** button (see section 1.4.7).

1.4.1 Request an evaluation license

We offer a fully functional version of the *CLC Cancer Research Workbench* for evaluation purposes, free of charge.

Each user is entitled to 14 days demo of *CLC Cancer Research Workbench*.

If you are unable to complete your assessment in the available time, please send an email to sales@clcbio.com to request an additional evaluation period.

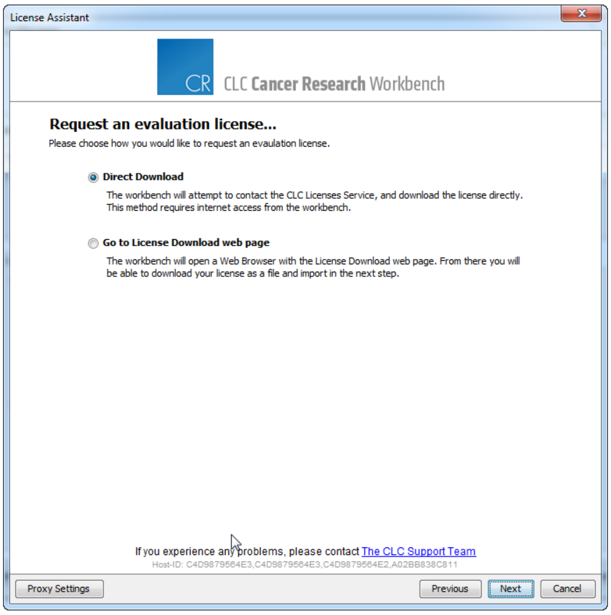
When you choose the option **Request an evaluation license**, you will see the dialog shown in figure 1.2.

In this dialog, there are two options:

- **Direct download.** Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page.** In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.



License Assistant

CLC Cancer Research Workbench

Request an evaluation license...

Please choose how you would like to request an evaulation license.

● **Direct Download**

The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.

○ **Go to License Download web page**

The workbench will open a Web Browser with the License Download web page. From there you will
be able to download your license as a file and import in the next step.

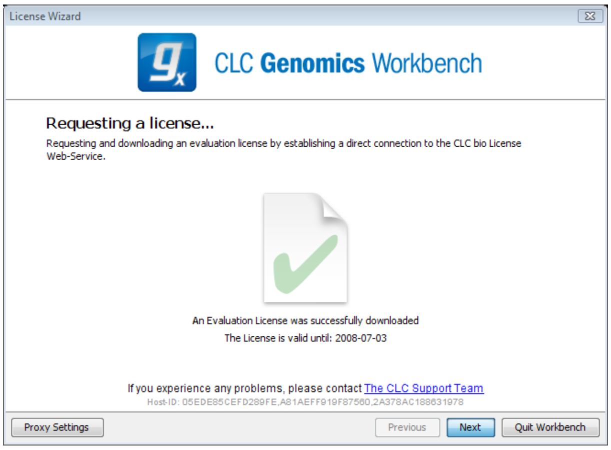
If you experience any problems, please contact The CLC Support Team

Host-ID: C4D9879564E3,C4D9879564E3,C4D9879564E2,A02BB838C811

Figure 1.2: *Choosing between direct download or download web page.*

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.3 appears.



License Wizard

CLC Genomics Workbench

Requesting a license...

Requesting and downloading an evaluation license by establishing a direct connection to the CLC bio License
Web-Service.

An Evaluation License was successfully downloaded

The License is valid until: 2008-07-03

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.3: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

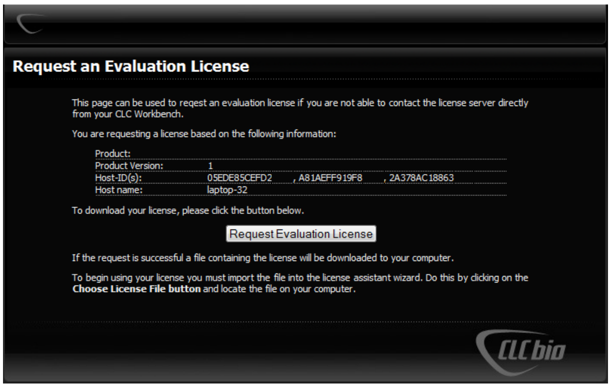
Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.4.

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.5.

Click the **Choose License File** button and browse to find the license file you saved. When you



Request an Evaluation License

This page can be used to reqest an evaluation license if you are not able to contact the license server directly from your CLC Workbench.

You are requesting a license based on the following information:

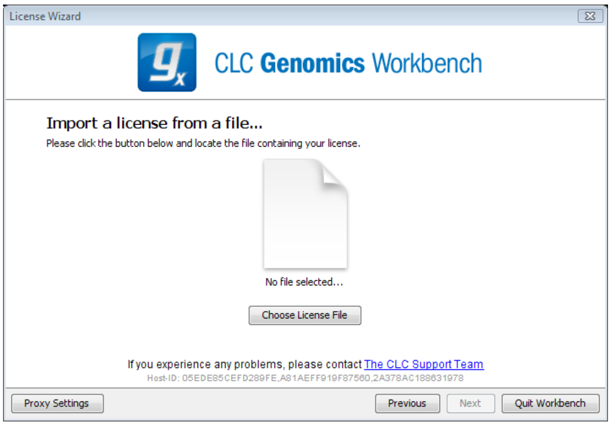
Product:	
Product Version:	1
Host-ID(s):	05EDE85CEFD2 , A81AEFF919F8 , 2A378AC18863
Host name:	laptop-32

To download your license, please click the button below.

If the request is successful a file containing the license will be downloaded to your computer.

To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the **Choose License File button** and locate the file on your computer.

Figure 1.4: *The license download web page.*



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

If you experience any problems, please contact The CLC Support Team

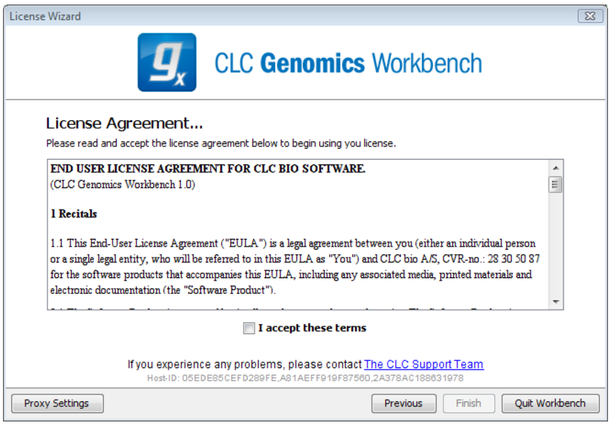
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.5: *Importing the license file downloaded from the web page.*

have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.6.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team

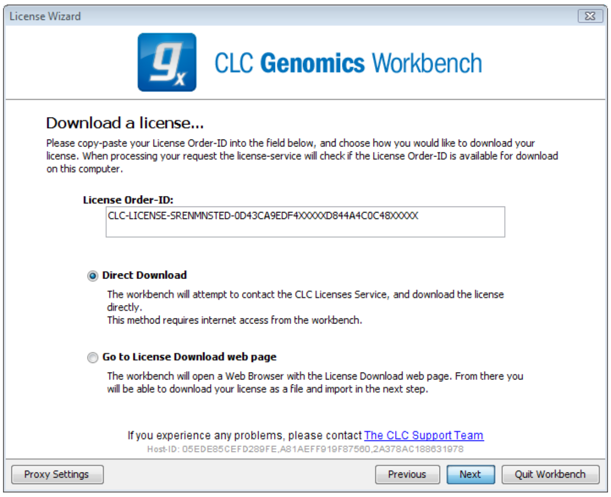
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.6: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.2 Download a license using a license order ID ID

Using a license order ID, you can download a license file via the Workbench or using an online form. When you have chosen this option and clicked **Next** button, you will see the dialog shown in 1.7. Enter your license order ID into the text field under the title License Order-ID. (The ID can be pasted into the box after copying it and then using menus or key combinations like Ctrl+V on some system or ⌘ + V on Mac).



Download a license...

Please copy-paste your License Order-ID into the field below, and choose how you would like to download your license. When processing your request the license-service will check if the License Order-ID is available for download on this computer.

License Order-ID: CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX

● **Direct Download**

The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.

○ **Go to License Download web page**

The workbench will open a Web Browser with the License Download web page. From there you will be able to download your license as a file and import in the next step.

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.7: *Enter a license order ID for the software.*

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

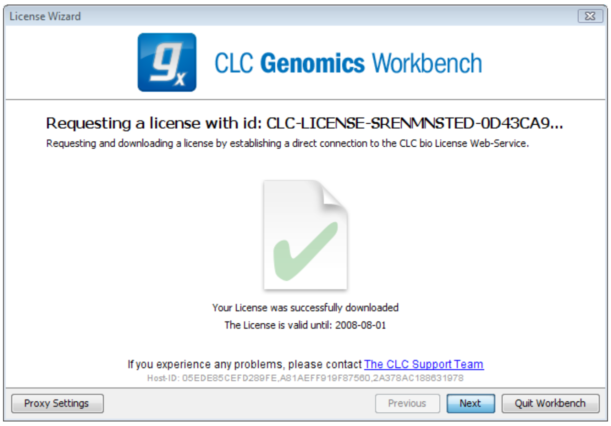
Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.8 appears.

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.9.



Requesting a license with id: CLC-LICENSE-SRENMNSTED-0D43CA9...

Requesting and downloading a license by establishing a direct connection to the CLC bio License Web-Service.

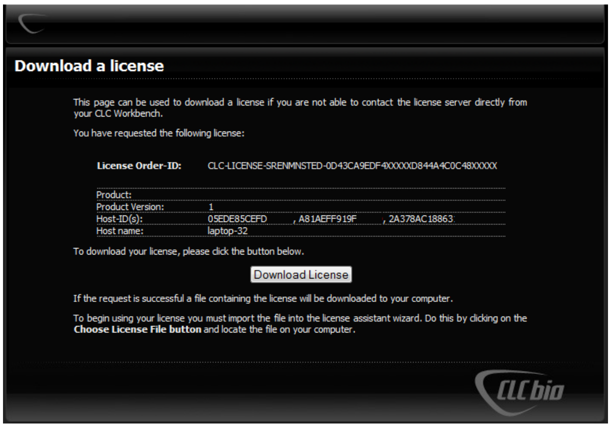
Your License was successfully downloaded

The License is valid until: 2008-08-01

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.8: *A license has been downloaded.*



Download a license

This page can be used to download a license if you are not able to contact the license server directly from your CLC Workbench.

You have requested the following license:

License Order-ID:	CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX
Product:	
Product Version:	1
Host-ID(s):	05EDE85CEFD , A81AEFF919F , 2A378AC18863
Host name:	laptop-32

To download your license, please click the button below.

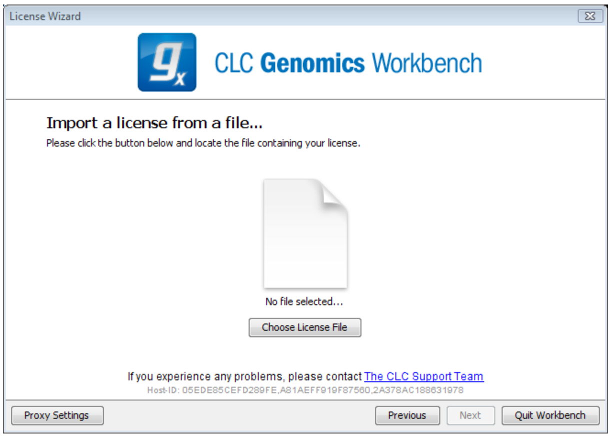
If the request is successful a file containing the license will be downloaded to your computer.

To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the **Choose License File button** and locate the file on your computer.

Figure 1.9: *The license download web page.*

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.10.



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

If you experience any problems, please contact The CLC Support Team

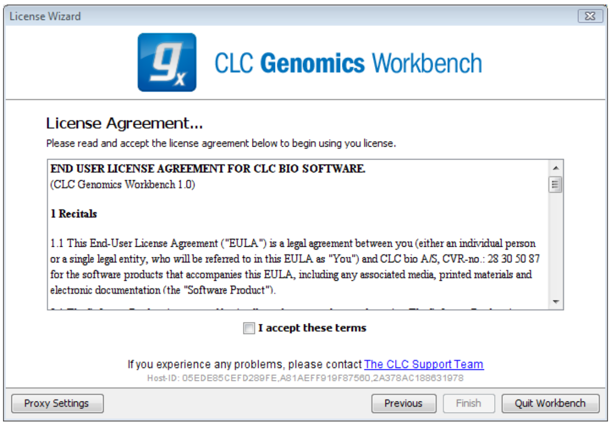
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.10: *Importing the license file downloaded from the web page.*

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.11.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

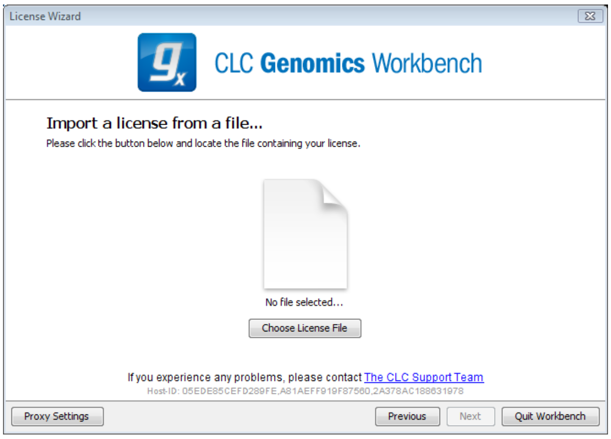
Figure 1.11: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.3 Import a license from a file

If you already have a license file associated with the host ID of your machine,it can be imported using this option.

When you have clicked on the **Next** button, you will see the dialog shown in 1.12.



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

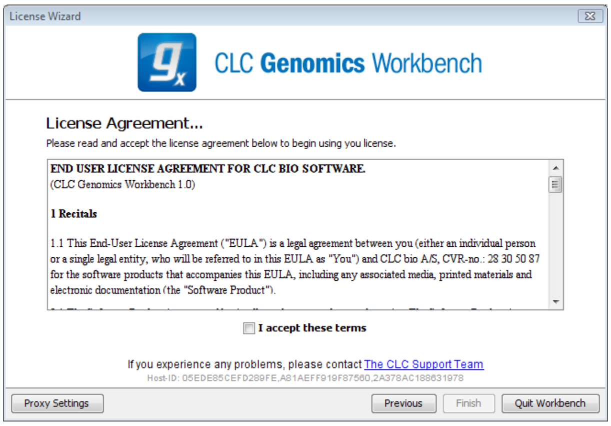
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.12: *Selecting a license file .*

Click the **Choose License File** button and browse to find the license file. When you have selected the file, click on the **Next** button.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.13.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.13: *Read the license agreement carefully.*

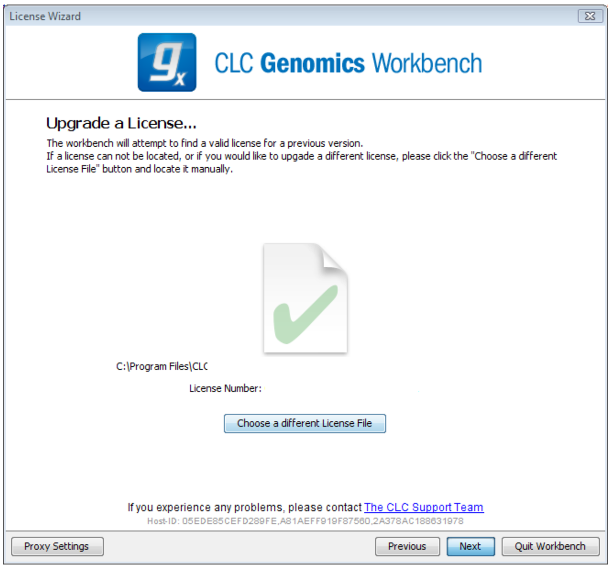
Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.4 Upgrade license

This option is used when you already have used a previous version of *CLC Cancer Research Workbench*, and you are entitled to upgrade to a new major version of the *CLC Cancer Research Workbench 1.5*. The Workbench will need direct access to the external network to use this option.

When you click on the **Next** button, the Workbench will search for a previous installation of *CLC Cancer Research Workbench*. It will then locate the old license.

If the Workbench finds an existing license file, the next dialog will look like figure 1.14.



Upgrade a License...

The workbench will attempt to find a valid license for a previous version.
If a license can not be located, or if you would like to upgade a different license, please click the "Choose a different License File" button and locate it manually.

C:\Program Files\CLC

License Number:

If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.14: *An license from an older installation is found.*

When you click on the **Next** button, the Workbench checks if you are entitled to upgrade your license. This is done by contacting CLC bio's servers.

If the Workbench cannot connect to the external network directly, please see the section on downloading a license for non-networked machines. You will need your license order ID for this.

Your license must be covered by our Maintenance, Upgrades and Support (MUS) program to be eligible to upgrade your license. If the license is covered for upgrades and there are any problems with this, please contact licenses@clcbio.com.

In this dialog, there are two options:

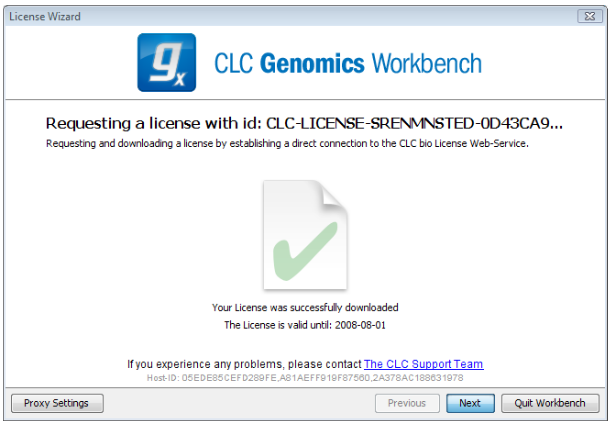
- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.15 appears.



Requesting a license with id: CLC-LICENSE-SRENMNSTED-0D43CA9...

Requesting and downloading a license by establishing a direct connection to the CLC bio License Web-Service.

Your License was successfully downloaded

The License is valid until: 2008-08-01

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.15: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

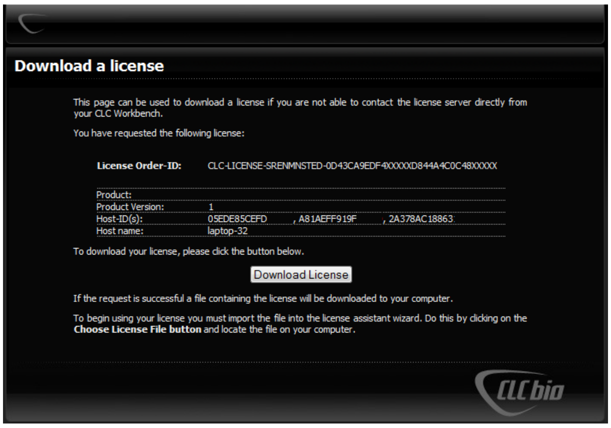
Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.16.

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.17.

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.



Download a license

This page can be used to download a license if you are not able to contact the license server directly from your CLC Workbench.

You have requested the following license:

License Order-ID: CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX

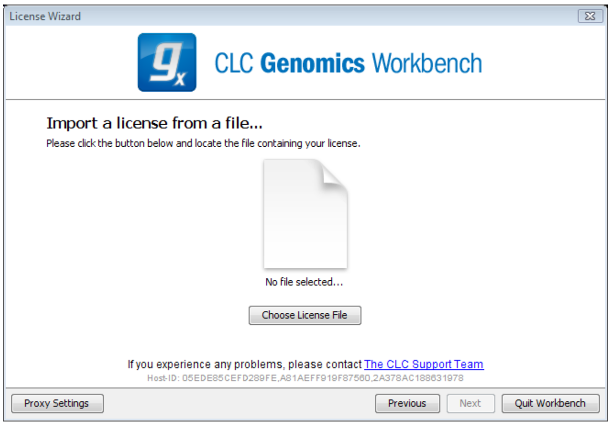
Product:	
Product Version:	1
Host-ID(s):	05EDE85CEFD , A81AEFF919F , 2A378AC18863
Host name:	laptop-32

To download your license, please click the button below.

If the request is successful a file containing the license will be downloaded to your computer.

To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the **Choose License File button** and locate the file on your computer.

Figure 1.16: *The license download web page.*



Import a license from a file...

Please click the button below and locate the file containing your license.

No file selected...

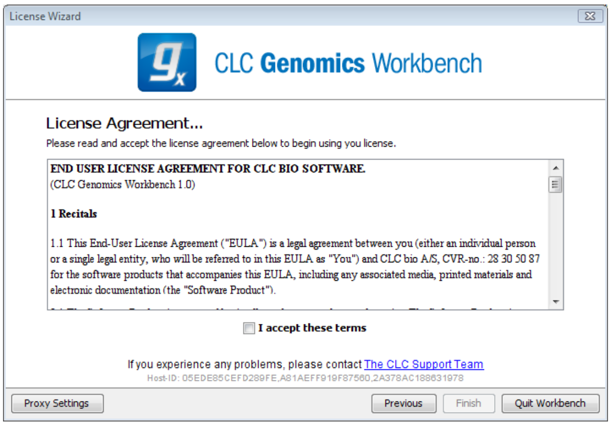
If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

Figure 1.17: *Importing the license file downloaded from the web page.*

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.18.



License Agreement...

Please read and accept the license agreement below to begin using you license.

END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)

1 Recitals

1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").

□ **I accept these terms**

If you experience any problems, please contact The CLC Support Team

Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978

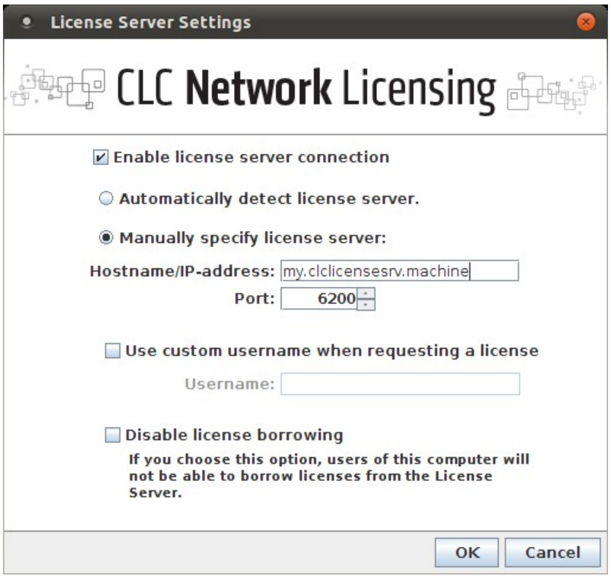
Figure 1.18: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.5 Configure license server connection

If your organization is running a CLC License Server, you can configure your Workbench to connect to it to get a license.

To do this, select this option and click on the **Next** button. A dialog like that shown in figure 1.19 then appears. Here, you configure how to connect to the CLC License Server.



☑ Enable license server connection

○ Automatically detect license server.

● Manually specify license server:

Hostname/IP-address: my.clclicensesrv.machine

Port: 6200

□ Use custom username when requesting a license

Username:

□ Disable license borrowing

If you choose this option, users of this computer will not be able to borrow licenses from the License Server.

Figure 1.19: *Connecting to a CLC License Server.*

- **Enable license server connection.** This box must be checked for the Workbench is to contact the CLC License Server to get a license for *CLC Cancer Research Workbench*.
- **Automatically detect license server**. By checking this option the Workbench will look for a CLC License Server accessible from the Workbench[3].
- **Manually specify license server**. If there are technical limitations such that the CLC License Server cannot be detected automatically, use this option to provides details of machine the CLC License Server software is on, and the port used by the software to receive requests. After selecting this option, please enter:
  - **Host name**. The address for the machine the CLC Licenser Server software is running on.
  - **Port**. The port used by the CLC License Server to receive requests.
- **Disable license borrowing on this computer**. If you do not want users of the computer to borrow a license from the set of licenses available, then (see section 1.4.5), select this option.

[3] Automatic server discovery sends UDP broadcasts from the Workbench on a fixed port, 6200. Available license servers respond to the broadcast. The Workbench then uses TCP communication for to get a license, assuming one is available. Automatic server discovery works only on local networks and will not work on WAN or VPN connections. Automatic server discovery is not guaranteed to work on all networks. If you are working on an enterprise network on where local firewalls or routers cut off UDP broadcast traffic, then you may need to configure the details of the CLC License server manually instead.

Borrowing a license

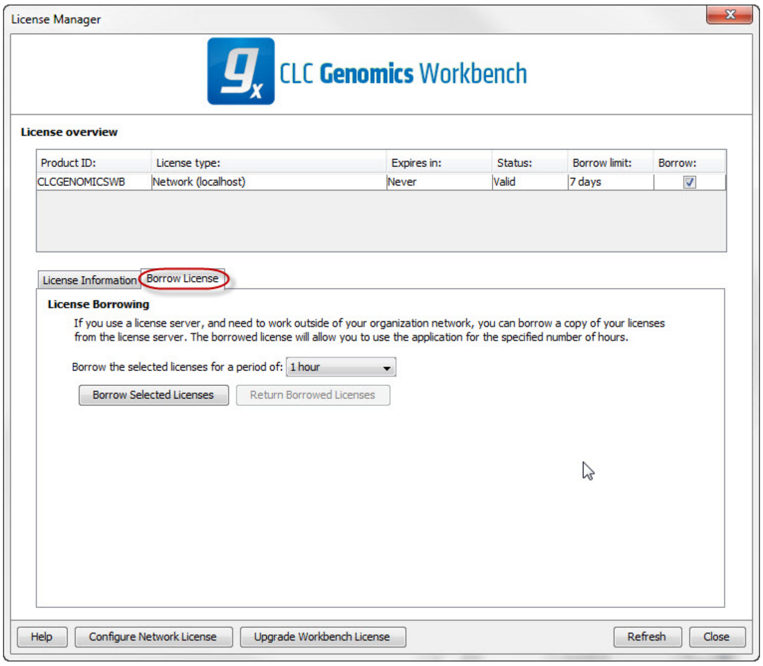
A network license can only be used when you are connected to the a license server. If you wish to use the *CLC Cancer Research Workbench* when you are not connected to the CLC License Server, you can *borrow* an available license for a period of time. During this time, there will be one less network license available on the for other users. The Workbench must have a connection to the CLC License Server at the point in time when you wish to borrow a license.

The procedure for borrowing a license is:

1. Go to the Workbench menu option:

   Help | **License Manager**

2. Click on the "Borrow License" tab to display the dialog shown in figure 1.20.



License overview

Product ID:	License type:	Expires in:	Status:	Borrow limit:	Borrow:
CLCGENOMICSWB	Network (localhost)	Never	Valid	7 days	☑

License Borrowing

If you use a license server, and need to work outside of your organization network, you can borrow a copy of your licenses from the license server. The borrowed license will allow you to use the application for the specified number of hours.

Borrow the selected licenses for a period of: 1 hour

Figure 1.20: *Borrow a license.*

3. Use the checkboxes at the right hand sideof the table in the License overview section of the window to select the license(s) that you wish to borrow.

4. Select the length of time you wish to borrow the license(s).

5. Click on the button labeled **Borrow Licenses**.

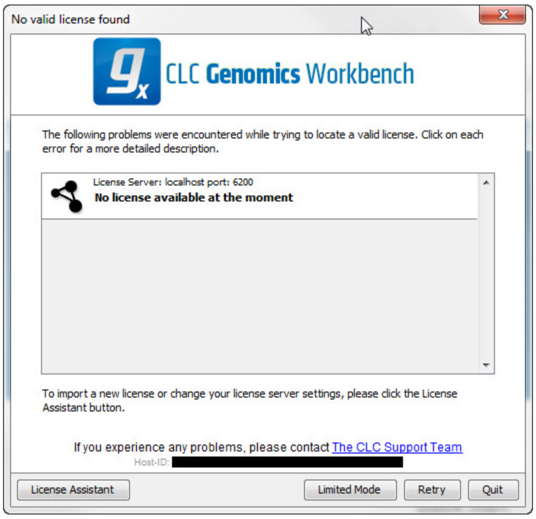
6. Close the License Manager when you are done.

You can now go offline and work with the *CLC Cancer Research Workbench*. When the time period you borrowed the license for has elapsed, the network license you borrowed is made available again for other users to access. To continue using the *CLC Cancer Research Workbench* with a license, you will need to connect the Workbench to the network again so it can contact the CLC Licene Server to obtain one.

Note! Your CLC License Server administrator can choose to disable to the option allowing the borrowing of licenses. If this has been done, you will not be able to borrow a network license using your Workbench.

Common issues when using a network license

No license available at the moment If all the network licenses or *CLC Cancer Research Workbench*are in use, you will see a dialog like that shown in figure 1.21 when you start up the Workbench.



No valid license found

The following problems were encountered while trying to locate a valid license. Click on each error for a more detailed description.

License Server: localhost port: 6200
No license available at the moment

To import a new license or change your license server settings, please click the License Assistant button.

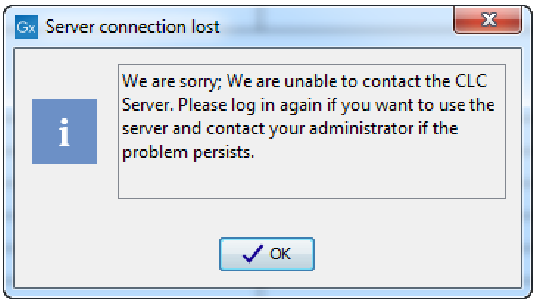
If you experience any problems, please contact The CLC Support Team
Host-ID: [redacted]

Figure 1.21: *This window appears when there are no available network licenses for the software you are running.*

This means others are using the network licenses. You will need to wait for them to return their licenses before you can continue to work with a fully functional copy of software. If this is a frequent issue, you may wish to discuss this with your CLC License Server administrator.

Clicking on the **Limited Mode** button in the dialog allows you to start the Workbench with a subset of features available. This includes the ability your access CLC data.

Lost connection to the CLC License Server If the Workbench connection to the CLC License Server is lost, you will see a dialog as shown in figure 1.22.



Server connection lost

We are sorry; We are unable to contact the CLC Server. Please log in again if you want to use the server and contact your administrator if the problem persists.

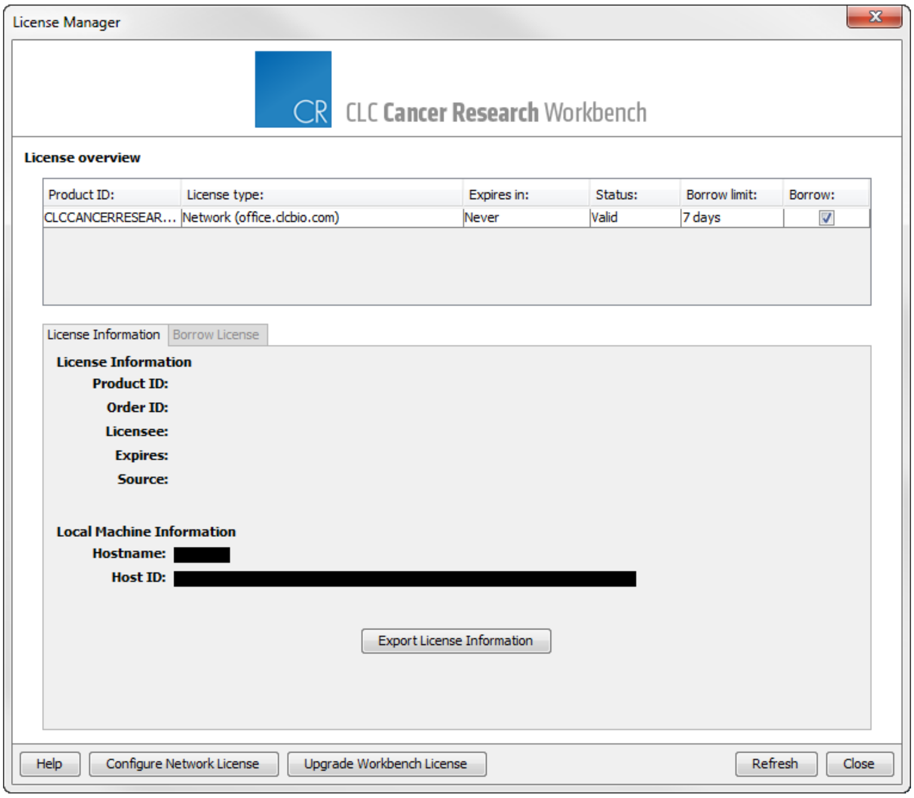
Figure 1.22: *This message appears if the Workbench is unable to establish a connection to a CLC License server.*

If you have chosen the option to **Automatically detect license server** and you have not succeeded in connecting to the License Server before, please check with your local IT support that automatic detection will be possible to do at your site. If it is not possible at your site, you will need to manually configure the CLC License Server settings using the License Manager, as described earlier in this section.

If you have successfully contacted the CLC License Server from your Workbench previously, please consider discussing this issue with your CLC License Server administrator or your local IT support, to make sure that the CLC License Server is running and that your Workbench can connect to it. There may be situations where you wish to use a different license or view information about the license(s) the Workbench is currently using. To do this, open the License Manager using the menu option:

Help | **License Manager** ( )

The license manager is shown in figure 1.23.


License Manager
CR CLC Cancer Research Workbench
License overview
License Information
Borrow License
Export License Information
Help
Configure Network License
Upgrade Workbench License
Refresh
Close

Product ID:	License type:	Expires in:	Status:	Borrow limit:	Borrow:
CLCCANCERRESEAR...	Network (office.clcbio.com)	Never	Valid	7 days	☑

License Information
Product ID:
Order ID:
Licensee:
Expires:
Source:

Local Machine Information
Hostname: [redacted]
Host ID: [redacted]

Figure 1.23: *The license manager.*

This dialog can be used to:

- See information about the license (e.g. what kind of license, when it expires)
- Configure how to connect to a license server (**Configure License Server** the button at the lower left corner). Clicking this button will display a dialog similar to figure 1.19.
- Upgrade from an evaluation license by clicking the **Upgrade license** button. This will display the dialog shown in figure 1.1.
- Export license information to a text file.

- Borrow a license

If you wish to switch away from using a network license, click on the button to **Configure License Server** and uncheck the box beside the text **Enable license server connection** in the dialog. When you restart the Workbench, you can set up the new license as described in section 1.4.

1.4.6 Download a static license on a non-networked machine

To download a static license for a machine that does not have direct access to the external network, you can follow the steps below:

- Install the *CLC Cancer Research Workbench* on the machine you wish to run the software on.
- Start up the software as an administrative user and find the host ID of the machine that you will run the CLC Workbench on. You can see the host ID the machine reported at the bottom of the License Manager window in grey text.
- Make a copy of this host ID such that you can use it on a machine that has internet access.
- Go to a computer with internet access, open a browser window and go to the relevant network license download web page:
- For Workbenches released from January 2013 and later, (e.g. the Genomics Workbench version 6.0 or higher, and the Main Workbench, version 6.8 or higher), please go to:

  `https://secure.clcbio.com/LmxWSv3/GetLicenseFile`

  For earlier Workbenches, including any DNA, Protein or RNA Workbench, please go to:

  `http://licensing.clcbio.com/LmxWSv1/GetLicenseFile`

  It is vital that you choose the license download page appropriate to the version of the software you plan to run.
- Paste in your license order ID and the host ID that you noted down in the relevant boxes on the webpage.
- Click 'download license' and save the resulting .lic file.
- Open the Workbench on your non-networked machine. In the Workbench license manager choose 'Import a license from a file'. In the resulting dialog click 'choose license file' to browse the location of the .lic file you have just downloaded.

  If the License Manager does not start up by default, you can start it up by going to the Help menu and choosing License Manager.
- Click on the **Next** button and go through the remaining steps of the license manager wizard.

1.4.7 Limited mode

We have created the limited mode to prevent a situation where you are unable to access your data because you do not have a license. When you run in limited mode, a lot of the tools in the Workbench are not available, but you still have access to your data (also when stored in a

CLC Bioinformatics Database) . To get out of the limited mode and run the Workbench normally, restart the Workbench. When you restart the Workbench will try to find a proper license and if it does, it will start up normally. If it can't find a license, you will again have the option of running in limited mode.

1.5 About CLC Workbenches

In November 2005 CLC bio released two Workbenches: *CLC Free Workbench* and *CLC Protein Workbench*. *CLC Protein Workbench* is developed from the free version, giving it the well-tested user friendliness and look & feel. However, the *CLC Protein Workbench* includes a range of more advanced analyses.

In March 2006, *CLC DNA Workbench* (formerly *CLC Gene Workbench*) and *CLC Main Workbench* were added to the product portfolio of CLC bio. Like *CLC Protein Workbench*, *CLC DNA Workbench* builds on *CLC Free Workbench*. It shares some of the advanced product features of *CLC Protein Workbench*, and it has additional advanced features. *CLC Main Workbench* holds all basic and advanced features of the *CLC Workbenches*.

In June 2007, *CLC RNA Workbench* was released as a sister product of *CLC Protein Workbench* and *CLC DNA Workbench*. *CLC Main Workbench* now also includes all the features of *CLC RNA Workbench*.

In March 2008, the *CLC Free Workbench* changed name to *CLC Sequence Viewer*.

In June 2008, the first version of the *CLC Genomics Workbench* was released due to an extraordinary demand for software capable of handling sequencing data from all new high-throughput sequencing platforms such as Roche-454, Illumina and SOLiD in addition to Sanger reads and hybrid data.

For an overview of which features all the applications include, see http://www.clcbio.com/features.

In December 2006, CLC bio released a **Software Developer Kit** which makes it possible for anybody with a knowledge of programming in Java to develop plugins. The plugins are fully integrated with the CLC Workbenches and the Viewer and provide an easy way to customize and extend their functionalities.

In April 2012, *CLC Protein Workbench*, *CLC DNA Workbench*and *CLC RNA Workbench* were discontinued. All customers with a valid license for any of these products were offered an upgrade to the *CLC Main Workbench*.

In February 2014, CLC bio expanded the product repertoire with the release of *CLC Drug Discovery Workbench*, a product that enables studies of protein-ligand interactions for drug discovery.

1.5.1 New program feature request

The CLC team is continuously improving the *CLC Cancer Research Workbench* with our users' interests in mind. We welcome all requests and feedback from users, as well as suggestions for new features or more general improvements to the program. To contact us via the Workbench, please go to the menu option:

Help | **Contact Support**

1.5.2 Getting help

If you encounter a problem or need help understanding how the *CLC Cancer Research Workbench* works, and the license you are using is covered by our Maintenance, Upgrades and Support (MUS) program (https://www.clcbio.com/support/maintenance-support-program/), you can contact our customer support via the Workbench by going to the menu option:

Help | **Contact Support**

This will open a dialog to enter your contact information and a text field for entering the question or problem you have.

You can also attach small datasets, if this helps explain the problem or you believe it will help in troubleshooting the problem.

When you send a support request this way, it will include technical information about your installation that usually helps when troubleshooting. It also includes your license information so that you do not have to look this up yourself. Our support staff will reply to you by email.

Further information about Maintenance, Upgrades and Support (MUS) program can be found online at https://www.clcbio.com/support/maintenance-support-program/.

Information about how to to find your license information is included in the licenses section of our Frequently Asked Questions (FAQ) area: http://www.clcbio.com/faq

Information about MUS cover on particular licenses can be found by https://secure.clcbio.com/myclc/login.

Users of the freely available CLC Sequence Viewer can make use of any of our online documentation sources, including the manuals (http://www.clcbio.com/manuals), tutorials (http://www.clcbio.com/tutorials) and other entries in our FAQ area (http://helpdesk.clcbio.com/index.php?pg=kb).

Start in safe mode

If the program becomes unstable on start-up, you can start it in **Safe mode**. This is done by pressing and holding down the Shift button while the program starts.

When starting in safe mode, the user settings (e.g. the settings in the **Side Panel**) are deleted and cannot be restored. Your data stored in the **Navigation Area** is not deleted. When started in safe mode, some of the functionalities are missing, and you will have to restart the *CLC Cancer Research Workbench* again (without pressing Shift.

1.6 When the program is installed: Getting started

CLC Cancer Research Workbench includes an extensive **Help** function, which can be found in the **Help** menu of the program's **Menu bar**. The **Help** can also be shown by pressing F1. The help topics are sorted in a table of contents and the topics can be searched.

Tutorials describing hands-on examples of how to use the individual tools and features of the *CLC Cancer Research Workbench* can be found at http://www.clcbio.com/support/tutorials/. We also recommend our **Online presentations** where a product specialist from CLC bio demonstrates our software. This is a very easy way to get started using the program. Read more about video tutorials and other online presentations here: http://www.clcbio.tv/.

1.6.1 Import of example data

It might be easier to understand the logic of the program by trying to do simple operations on existing data. Therefore *CLC Cancer Research Workbench* includes an example data set.

If you would like to download the example data you have three options:

1. You can click **Download Example Data** in the start up table that is visible in the *CLC Cancer Research Workbench* when no datasets have been opened for viewing. This will take you to http://clccancer.com/downloads/ where you can choose to download two different example datasets that can be used for the following purposes:

   - Variant identification in a tumor sample. This dataset is taken from a larger whole exome dataset and includes data from a small fraction of chromosome 5 (Example_data_tumor.zip).
   - Identification of somatic variants in a tumor sample using the matched normal sample for removal of germline variants. This is matched tumor and normal samples from chromosome 22 from a whole exome dataset (Example_data_tumor_normal.zip).

2. You can also go to directly to http://clccancer.com/downloads/ and download the example data from there.

3. Finally, you can use these links to get the data:

   http://download.clcbio.com/testdata/cancer/current/Example_data_tumor.zip or

   http://download.clcbio.com/testdata/cancer/current/Example_data_tumor_normal.zip

When you have downloaded the data from the website, you need to import them into the *CLC Cancer Research Workbench*. How to import data is described in section 6.1.

1.7 Plugins

When you install *CLC Cancer Research Workbench*, it has a standard set of features. However, you can upgrade and customize the program using a variety of plugins.

As the range of plugins is continuously updated and expanded, they will not be listed here. Instead we refer to http://www.clcbio.com/plugins for a full list of plugins with descriptions of their functionalities.

1.7.1 Installing plugins

Plugins are installed using the plugin manager[4]:

[4] In order to install plugins on Windows, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

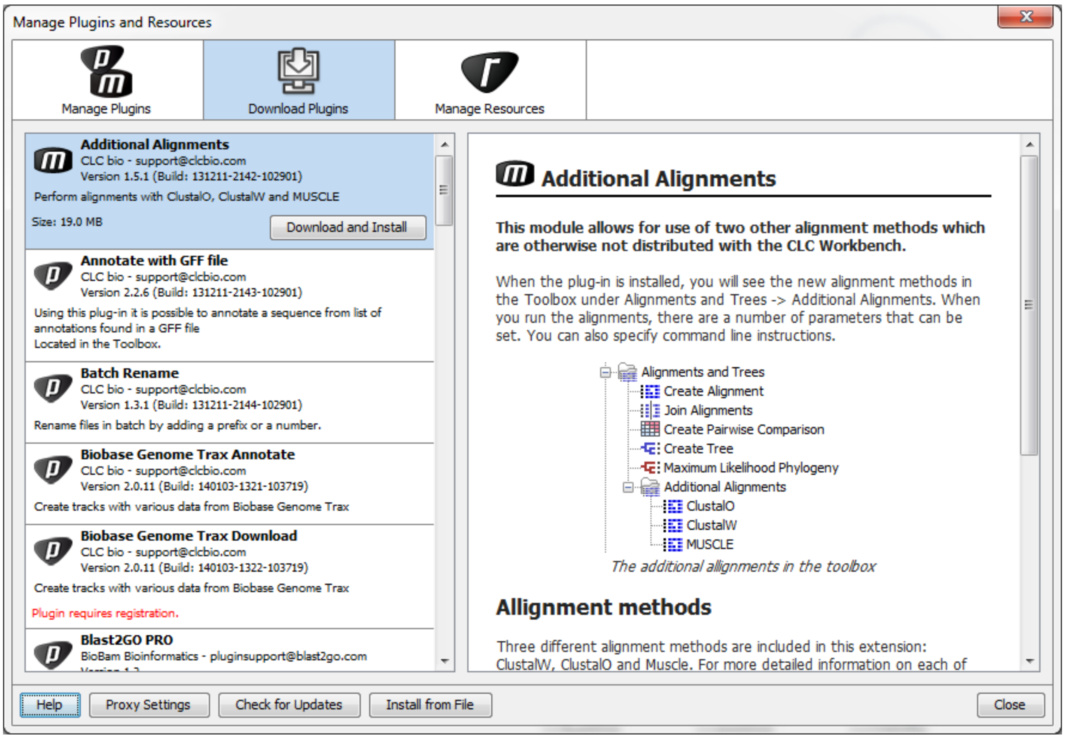
Help in the Menu Bar | Plugins and Resources... ( )

or **Plugins ( ) in the Toolbar**

The plugin manager has three tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins on CLC bio's server.
- **Manage Resources.** This is an overview of resources that are installed.

To install a plugin, click the **Download Plugins** tab. This will display an overview of the plugins that are available for download and installation (see figure 1.24).



Additional Alignments
CLC bio - support@clcbio.com
Version 1.5.1 (Build: 131211-2142-102901)
Perform alignments with ClustalO, ClustalW and MUSCLE
Size: 19.0 MB

Annotate with GFF file
CLC bio - support@clcbio.com
Version 2.2.6 (Build: 131211-2143-102901)
Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.

Batch Rename
CLC bio - support@clcbio.com
Version 1.3.1 (Build: 131211-2144-102901)
Rename files in batch by adding a prefix or a number.

Biobase Genome Trax Annotate
CLC bio - support@clcbio.com
Version 2.0.11 (Build: 140103-1321-103719)
Create tracks with various data from Biobase Genome Trax

Biobase Genome Trax Download
CLC bio - support@clcbio.com
Version 2.0.11 (Build: 140103-1322-103719)
Create tracks with various data from Biobase Genome Trax
Plugin requires registration.

Blast2GO PRO
BioBam Bioinformatics - pluginsupport@blast2go.com

Additional Alignments

This module allows for use of two other alignment methods which are otherwise not distributed with the CLC Workbench.

When the plug-in is installed, you will see the new alignment methods in the Toolbox under Alignments and Trees -> Additional Alignments. When you run the alignments, there are a number of parameters that can be set. You can also specify command line instructions.

```
Alignments and Trees
    Create Alignment
    Join Alignments
    Create Pairwise Comparison
    Create Tree
    Maximum Likelihood Phylogeny
    Additional Alignments
        ClustalO
        ClustalW
        MUSCLE
```

The additional allignments in the toolbox

Allignment methods

Three different alignment methods are included in this extension: ClustalW, ClustalO and Muscle. For more detailed information on each of

Figure 1.24: *The plugins that are available for download.*

Clicking a plugin will display additional information at the right side of the dialog. This will also display a button: **Download and Install**.

Click the plugin and press **Download and Install**. A dialog displaying progress is now shown, and the plugin is downloaded and installed.

If the plugin is not shown on the server, and you have it on your computer (e.g. if you have downloaded it from our web-site), you can install it by clicking the **Install from File** button at the bottom of the dialog. This will open a dialog where you can browse for the plugin. The plugin file should be a file of the type ".cpa".

When you close the dialog, you will be asked whether you wish to restart the *CLC Cancer Research Workbench*. The plugin will not be ready for use until you have restarted.

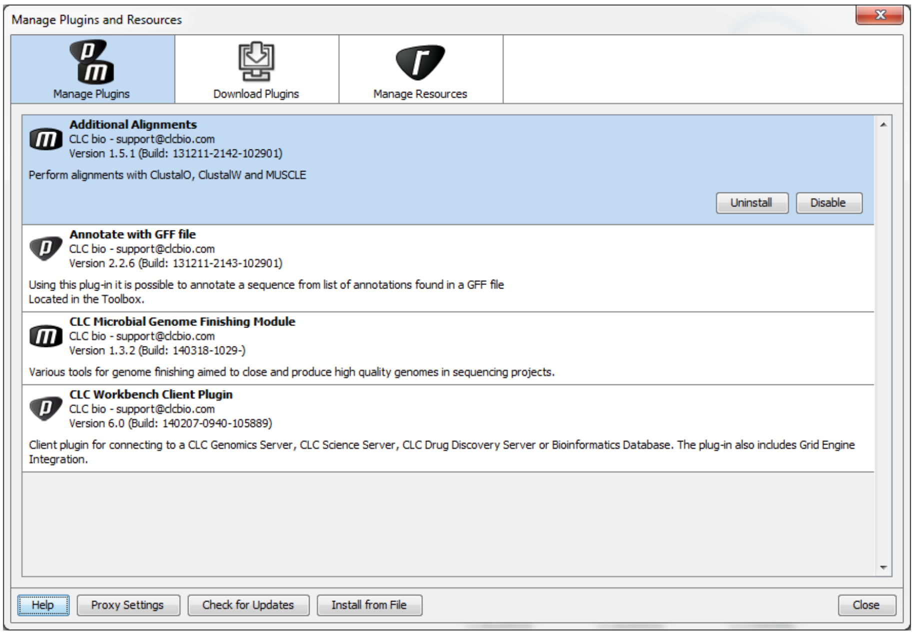
1.7.2 Uninstalling plugins

Plugins are uninstalled using the plugin manager:

Help in the Menu Bar | Plugins and Resources... ( )

or **Plugins ( ) in the Toolbar**

This will open the dialog shown in figure 1.25.



Additional Alignments
CLC bio - support@clcbio.com
Version 1.5.1 (Build: 131211-2142-102901)

Perform alignments with ClustalO, ClustalW and MUSCLE

Annotate with GFF file
CLC bio - support@clcbio.com
Version 2.2.6 (Build: 131211-2143-102901)

Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.

CLC Microbial Genome Finishing Module
CLC bio - support@clcbio.com
Version 1.3.2 (Build: 140318-1029-)

Various tools for genome finishing aimed to close and produce high quality genomes in sequencing projects.

CLC Workbench Client Plugin
CLC bio - support@clcbio.com
Version 6.0 (Build: 140207-0940-105889)

Client plugin for connecting to a CLC Genomics Server, CLC Science Server, CLC Drug Discovery Server or Bioinformatics Database. The plug-in also includes Grid Engine Integration.

Figure 1.25: *The plugin manager with plugins installed.*

The installed plugins are shown in this dialog. To uninstall:

Click the plugin | Uninstall

If you do not wish to completely uninstall the plugin but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be uninstalled until the workbench is restarted.

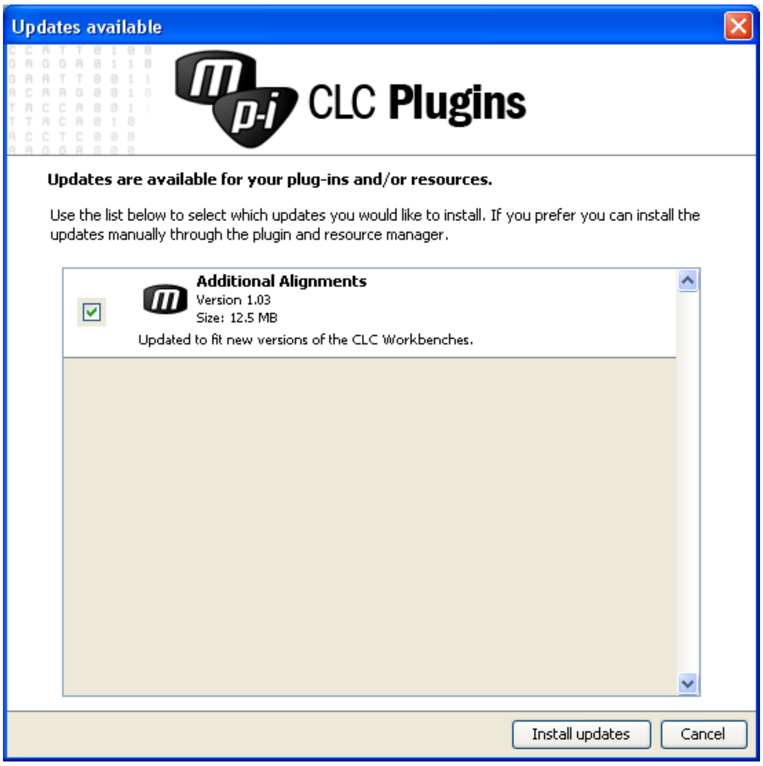
1.7.3 Updating plugins

If a new version of a plugin is available, you will get a notification during start-up as shown in figure 1.26.

In this list, select which plugins you wish to update, and click **Install Updates**. If you press **Cancel** you will be able to install the plugins later by clicking **Check for Updates** in the Plugin manager (see figure 1.25).

1.7.4 Resources

Resources can be manually installed and un-installed the same way as plugins.



Updates are available for your plug-ins and/or resources.

Use the list below to select which updates you would like to install. If you prefer you can install the updates manually through the plugin and resource manager.

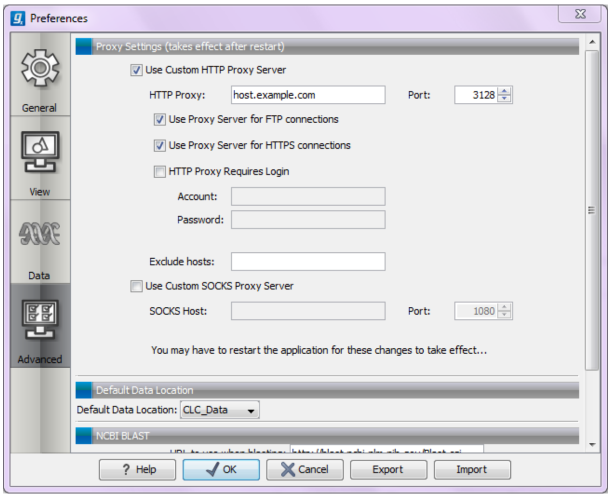
☑ **Additional Alignments**
Version 1.03
Size: 12.5 MB
Updated to fit new versions of the CLC Workbenches.

Figure 1.26: *Plugin updates.*

1.8 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Cancer Research Workbench* to use this. Otherwise you will not be able to perform any online activities (e.g. searching GenBank).

CLC Cancer Research Workbench supports the use of a HTTP-proxy and an anonymous SOCKS-proxy.



Proxy Settings (takes effect after restart)

☑ Use Custom HTTP Proxy Server

HTTP Proxy: host.example.com Port: 3128

☑ Use Proxy Server for FTP connections

☑ Use Proxy Server for HTTPS connections

☐ HTTP Proxy Requires Login

Account:

Password:

Exclude hosts:

☐ Use Custom SOCKS Proxy Server

SOCKS Host: Port: 1080

You may have to restart the application for these changes to take effect...

Default Data Location

Default Data Location: CLC_Data

NCBI BLAST

Figure 1.27: *Adjusting proxy preferences.*

To configure your proxy settings, open *CLC Cancer Research Workbench*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.27) and enter the appropriate information. The **Preferences** dialog is opened from the **Edit** menu.

You have the choice between a HTTP-proxy and a SOCKS-proxy. *CLC Cancer Research Workbench* only supports the use of a SOCKS-proxy that does not require authorization.

You can select whether the proxy should be used also for FTP and HTTPS connections.

Exclude hosts can be used if there are some hosts that should be contacted directly and not through the proxy server. The value can be a list of hosts, each separated by a |, and in addition a wildcard character * can be used for matching. For example: `*.foo.com|localhost`.

If you have any problems with these settings you should contact your systems administrator.

1.9 The format of the user manual

This user manual offers support to Windows, Mac OS X and Linux users. The software is very similar on these operating systems. In areas where differences exist, these will be described separately. However, the term "right-click" is used throughout the manual, but some Mac users may have to use Ctrl+click in order to perform a "right-click" (if they have a single-button mouse).

The most recent version of the user manuals can be downloaded from `http://www.clcbio.com/usermanuals`.

The user manual consists of four parts.

- The **first part** includes the introduction to the *CLC Cancer Research Workbench*.
- The **second part** describes in detail how to operate all the program's basic functionalities.
- The **third part** digs deeper into some of the molecular modeling and bioinformatic features of the program. In this part, you will also find our "Bioinformatics explained" sections. These sections elaborate on the algorithms and analyses of *CLC Cancer Research Workbench* and provide more general knowledge of molecular modeling and bioinformatic concepts.
- The **fourth part** is the Appendix and Index.

Each chapter includes a short table of contents.

1.9.1 Text formats

In order to produce a clearly laid-out content in this manual, different formats are applied:

- A feature in the program is in bold starting with capital letters. (Example: **Navigation Area**)
- An explanation of how a particular function is activated, is illustrated by "|" and bold. (E.g.: **select the element | Edit | Rename**)

Part II

Core functionalities

Chapter 2

User interface

Contents

This chapter provides an overview of the different areas in the user interface of *CLC Cancer Research Workbench*. As can be seen from figure 2.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

A description of the **Navigation Area** is tightly connected to the data management features of *CLC Cancer Research Workbench* and can be found in section 3.1.

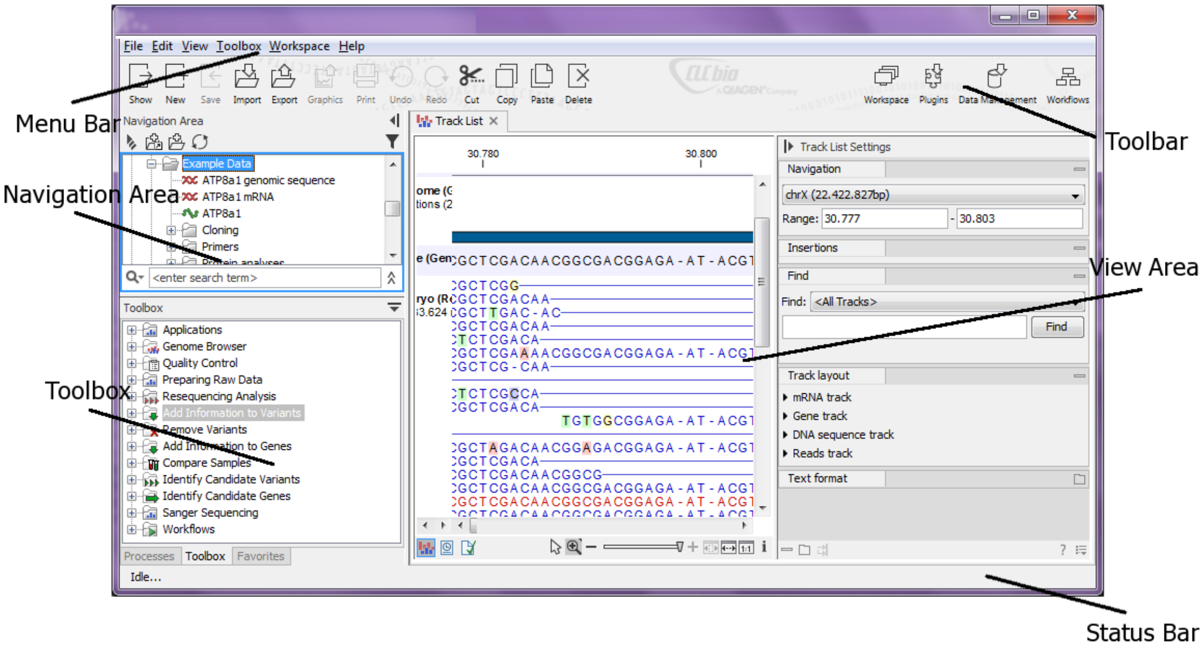


Figure 2.1: *The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.*

2.1 View Area

The **View Area** is the right-hand part of the screen, displaying your current work. The **View Area** may consist of one or more **Views**, represented by tabs at the top of the **View Area**.

This is illustrated in figure 2.2.

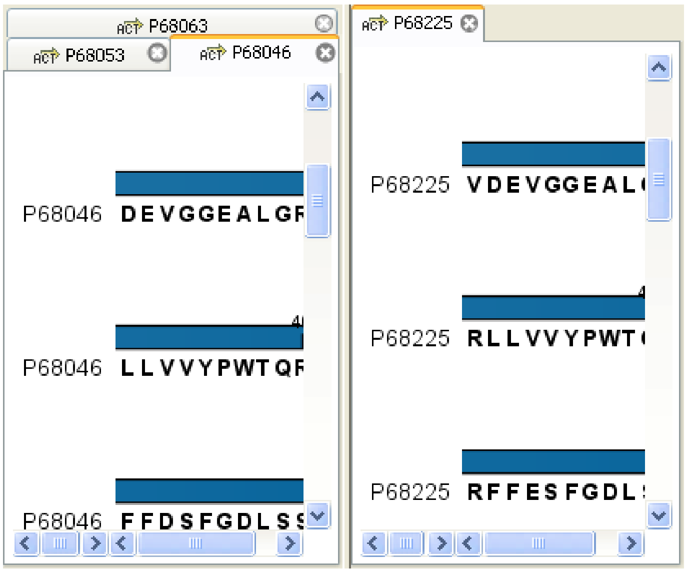


Figure 2.2: *A View Area can enclose several views, each view is indicated with a tab (see right view, which shows protein P68225). Furthermore, several views can be shown at the same time (in this example, four views are displayed).*

The tab concept is central to working with *CLC Cancer Research Workbench*, because several operations can be performed by dragging the tab of a view, and extended right-click menus can be activated from the tabs.

This chapter deals with the handling of views inside a **View Area**. Furthermore, it deals with rearranging the views.

Section 2.2 deals with the zooming and selecting functions.

2.1.1 Open view

Opening a view can be done in a number of ways:

double-click an element in the Navigation Area

or **select an element in the Navigation Area | File | Show | Select the desired way to view the element**

or **select an element in the Navigation Area | Ctrl + O (⌘ + B on Mac)**

Opening a view while another view is already open, will show the new view in front of the other view. The view that was already open can be brought to front by clicking its tab.

Note! If you right-click an open tab of any element, click **Show**, and then choose a different view of the same element, this new view is automatically opened in a split-view, allowing you to see both views.

See section 3.1.5 for instructions on how to open a view using drag and drop.

2.1.2 Show element in another view

Each element can be shown in different ways. A sequence, for example, can be shown as linear, circular, text etc.

In the following example, you want to see a sequence in a circular view. If the sequence is already open in a view, you can change the view to a circular view:

Click Show As Circular (○) at the lower left part of the view

The buttons used for switching views are shown in figure 2.3).

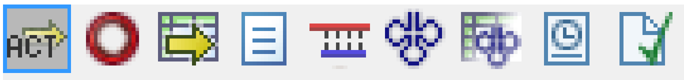


Figure 2.3: *The buttons shown at the bottom of a view of a nucleotide sequence. You can click the buttons to change the view to e.g. a circular view or a history view.*

If the sequence is already open in a linear view (ACT), and you wish to see both a circular and a linear view, you can split the views very easily:

Press Ctrl (⌘ on Mac) while you | Click Show As Circular (○) at the lower left part of the view

This will open a split view with a linear view at the bottom and a circular view at the top (see 9.8).

You can also show a circular view of a sequence without opening the sequence first:

Select the sequence in the Navigation Area | Show (⇥) | As Circular (○)

2.1.3 Close views

When a view is closed, the **View Area** remains open as long as there is at least one open view.

A view is closed by:

right-click the tab of the View | Close

or **select the view | Ctrl + W**

or **hold down the Ctrl-button | Click the tab of the view while the button is pressed**

By right-clicking a tab, the following close options exist. See figure 2.4

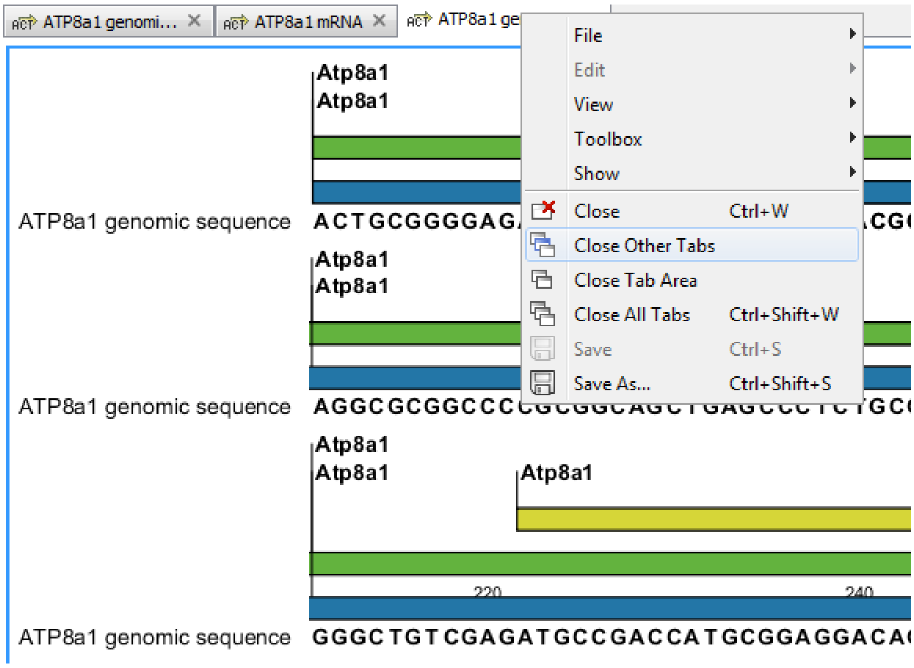


Figure 2.4: *By right-clicking a tab, several close options are available.*

- **Close.** See above.
- **Close Other Tabs.** Closes all other tabs, in all tab areas, except the one that is selected.
- **Close Tab Area.** Closes all tabs in the tab area.
- **Close All Tabs.** Closes all tabs, in all tab areas. Leaves an empty workspace.

2.1.4 Save changes in a view

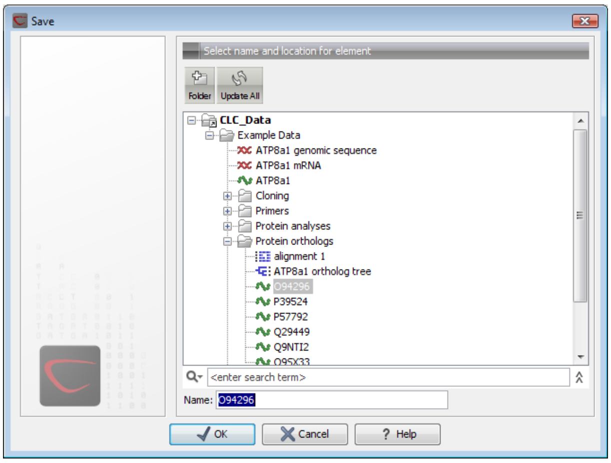
When changes to an element are made in a view, the text on the tab appears *bold and italic* (on Mac it is indicated by an * before the name of the tab). This indicates that the changes are not saved. The **Save** function may be activated in two ways:

Click the tab of the view you want to save | Save ( ) in the toolbar.

or **Click the tab of the view you want to save | Ctrl + S (⌘ + S on Mac)**

If you close a tab of a view containing an element that has been changed since you opened it, you are asked if you want to save.

When saving an element from a new view that has not been opened from the **Navigation Area** (e.g. when opening a sequence from a list of search hits), a save dialog appears (figure 2.5).



Name: O94296

Figure 2.5: *Save dialog.*

In the dialog you select the folder in which you want to save the element.

After naming the element, press **OK**

2.1.5 Undo/Redo

If you make a change to an element in a view, e.g. remove an annotation in a sequence or modify a tree, you can undo the action. In general, **Undo** applies to all changes you can make when right-clicking in a view. **Undo** is done by:

Click undo (↶) in the Toolbar

or **Edit | Undo (↶)**

or **Ctrl + Z**

If you want to undo several actions, just repeat the steps above. To reverse the undo action:

Click the redo icon in the Toolbar

or **Edit | Redo (↷)**

or **Ctrl + Y**

Note! Actions in the **Navigation Area**, e.g. renaming and moving elements, cannot be undone. However, you can restore deleted elements (see section 3.1.7).

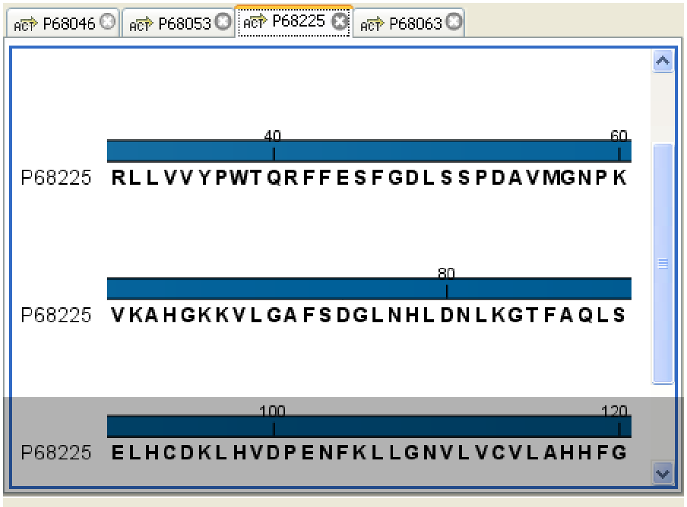
You can set the number of possible undo actions in the Preferences dialog (see section 4).

2.1.6 Arrange views in View Area

To provide more space for viewing data, you can hide **Navigation Area** and the **Toolbox** by clicking the hide icon (◀|) at the top of the **Navigation Area**.

Views are arranged in the **View Area** by their tabs. The order of the **views** can be changed using drag and drop. E.g. drag the tab of one view onto the tab of a another. The tab of the first view is now placed at the right side of the other tab.

If a tab is dragged into a view, an area of the view is made gray (see fig. 2.6) illustrating that the view will be placed in this part of the **View Area**.



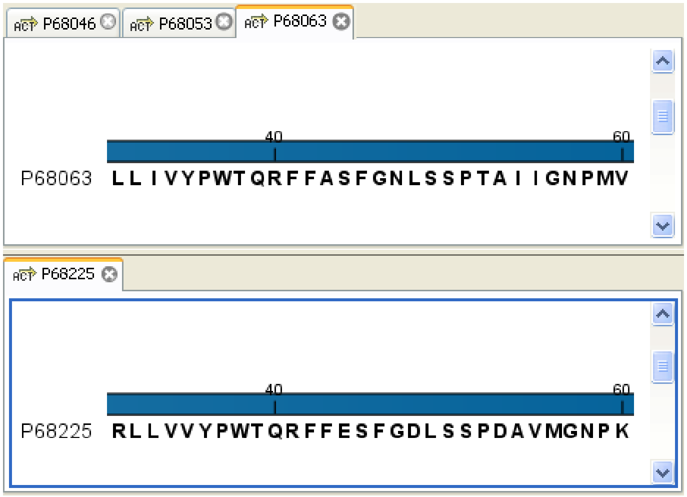
P68225 RLLVVYPWTQRFFESFGDLSSPDAVMGNPK

P68225 VKAHGKKVLGAFSDGLNHLDNLKGTFAQLS

P68225 ELHCDKLHVDPENFKLLGNVLVCVLAHHFG

Figure 2.6: *When dragging a view, a gray area indicates where the view will be shown.*

The results of this action is illustrated in figure 2.7.



P68063 LLIVYPWTQRFFASFGNLSSPTAIIGNPMV

P68225 RLLVVYPWTQRFFESFGDLSSPDAVMGNPK

Figure 2.7: *A horizontal split-screen. The two views split the View Area.*

You can also split a **View Area** horizontally or vertically using the menus.

Splitting horizontally may be done this way:

right-click a tab of the view | View | Split Horizontally (⊟)

This action opens the chosen view below the existing view. (See figure 2.8). When the split is made vertically, the new view opens to the right of the existing view.

Splitting the **View Area** can be undone by dragging e.g. the tab of the bottom view to the tab of the top view. This is marked by a gray area on the top of the view.

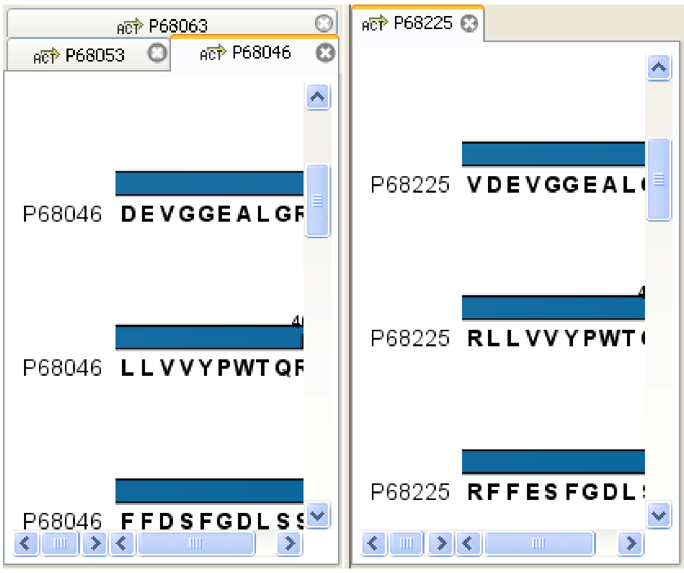
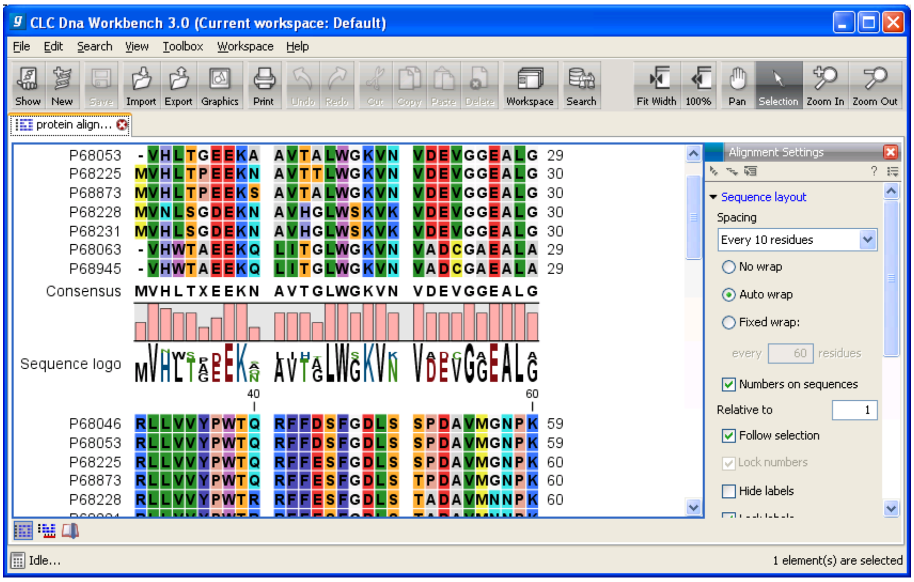


Figure 2.8: *A vertical split-screen.*

Maximize/Restore size of view

The **Maximize/Restore View** function allows you to see a view in maximized mode, meaning a mode where no other **views** nor the **Navigation Area** is shown.


CLC Dna Workbench 3.0 (Current workspace: Default)
File Edit Search View Toolbox Workspace Help
Show New Save Import Export Graphics Print Undo Redo Cut Copy Paste Delete Workspace Search Fit Width 100% Pan Selection Zoom In Zoom Out
protein align...
Alignment Settings
Sequence layout
Consensus
Sequence logo
Idle...
1 element(s) are selected

Spacing: Every 10 residues

○ No wrap

● Auto wrap

○ Fixed wrap:

every 60 residues

☑ Numbers on sequences

Relative to 1

☑ Follow selection

☑ Lock numbers

☐ Hide labels

Figure 2.9: *A maximized view. The function hides the Navigation Area and the Toolbox.*

Maximizing a view can be done in the following ways:

select view | Ctrl + M

or **select view | View | Maximize/restore View (□)**

or **select view | right-click the tab | View | Maximize/restore View (□)**

or **double-click the tab of view**

The following restores the size of the view:

Ctrl + M

or **View | Maximize/restore View (□)**

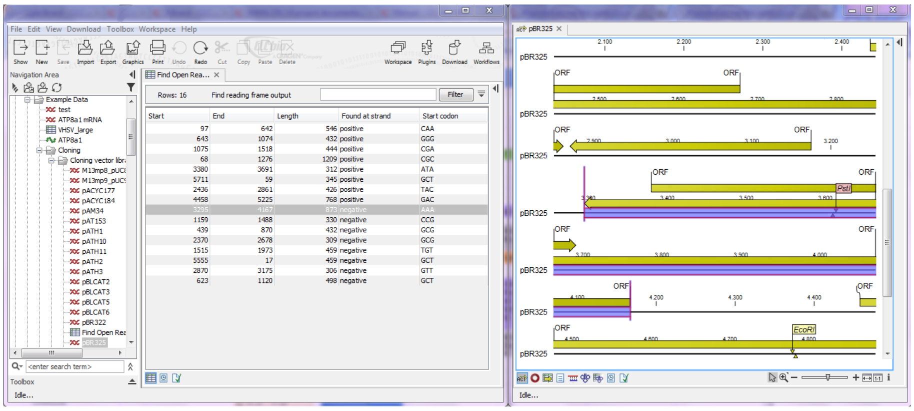
or **double-click title of view**

Please note that you can also hide **Navigation Area** and the **Toolbox** by clicking the hide icon (◀|) at the top of the **Navigation Area**

2.1.7 Moving a view to a different screen

Using multiple screens can be a great benefit when analyzing data with the *CLC Cancer Research Workbench*. You can move a view to another screen by dragging the tab of the view and dropping it outside the workbench window. Alternatively, you can right-click in the view area or on the tab itself and select **View | Move to New Window** from the context menu.

An example is shown in figure 2.10, where the main Workbench window shows a table of open reading frames, and the screen to the right is used to display the sequence and annotations.


File Edit View Download Toolbox Workspace Help
Show New Save Import Export Graphics Print Undo Redo Cut Copy Paste Delete
Workspace Plugins Download Workflows
Navigation Area
Example Data
Cloning
Toolbox
Idle...
Find Open Rea...
pBR325
ORF
PstI
EcoRI

Rows: 16 Find reading frame output Filter

Start	End	Length	Found at strand	Start codon
97	642	546	positive	CAA
643	1074	432	positive	GGG
1075	1518	444	positive	CGA
68	1276	1209	positive	CGC
3380	3691	312	positive	ATA
5711	59	345	positive	GCT
2436	2861	426	positive	TAC
4458	5225	768	positive	GAC
3295	4167	873	negative	AAA
1159	1488	330	negative	CCG
439	870	432	negative	GCG
2370	2678	309	negative	GCG
1515	1973	459	negative	TGT
5555	17	459	negative	GCT
2870	3175	306	negative	GTT
623	1120	498	negative	GCT

Figure 2.10: *Showing the table on one screen while the sequence is displayed on another screen. Clicking the table of open reading frames causes the view on the other screen to follow the selection. Note that the screen resolution in this figure is kept low in order to include it in the manual; in a real scenario, the resolution will be much higher.*

You can make more detached windows, by dropping tabs outside the open workbench windows, or you can drag more tabs to a detached window. To get a tab back to the main workbench window, just drag the detached tab back, and drop it next to the other tabs in the top of the view area. **Note:** You should not drag the detached window header, just the tab itself.

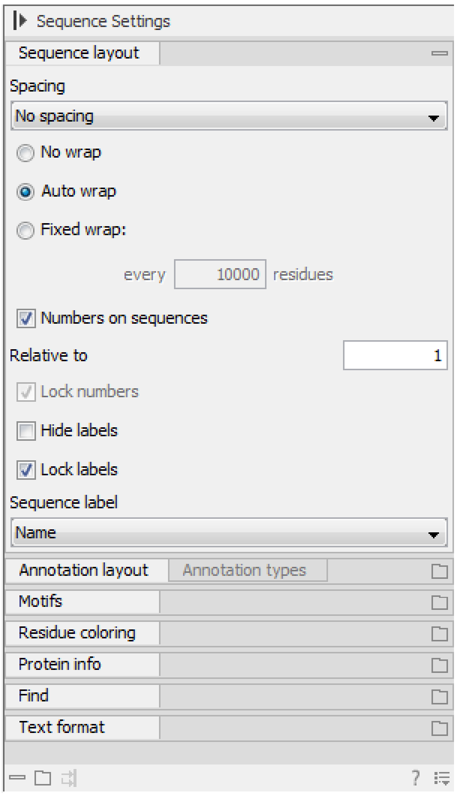
You can also split the view area in the detached windows as described in section 2.1.6.

2.1.8 Side Panel

The **Side Panel** allows you to change the way the contents of a view are displayed. The options in the **Side Panel** depend on the kind of data in the view, and they are described in the relevant sections about sequences, alignments, trees etc.

Figure 2.11 shows the default **Side Panel** for a protein sequence. It is organized into *palettes*.

In this example, there is one for Sequence layout, one for Annotation Layout etc. These palettes can be re-organized by dragging the palette name with the mouse and dropping it where you want



Sequence Settings

Sequence layout

Spacing

No spacing ▾

○ No wrap

● Auto wrap

○ Fixed wrap:

every 10000 residues

☑ Numbers on sequences

Relative to 1

☑ Lock numbers

☐ Hide labels

☑ Lock labels

Sequence label

Name ▾

Annotation layout | Annotation types

Motifs

Residue coloring

Protein info

Find

Text format

Figure 2.11: *The default view of the Side Panel when opening a protein sequence.*

it to be. They can either be situated next to each other, so that you can switch between them, or they can be listed on top of each other, so that expanding one of the palettes will push the palettes below further down.

In addition, they can be moved away from the **Side Panel** and placed anywhere on the screen as shown in figure2.12.

In this example, the **Motifs** palette has been placed on top of the sequence view together with the **Protein info** and the **Residue coloring** palettes. In the **Side Panel** to the right, the **Find** palette has been put on top.

In order to make all palettes dock in the **Side Panel** again, click the **Dock Side Panel** icon (⇥).

You can completely hide the **Side Panel** by clicking the **Hide Side Panel** icon (▶).

At the bottom of the **Side Panel** (see figure 2.13) there are a number of icons used to:

- Expand all settings (□).
- Collapse all settings (▭).
- Dock all palettes (⇥)
- Get **Help** for the particular view and settings (?)
- Save the settings of the **Side Panel** or apply already saved settings. Read more in section 4.5

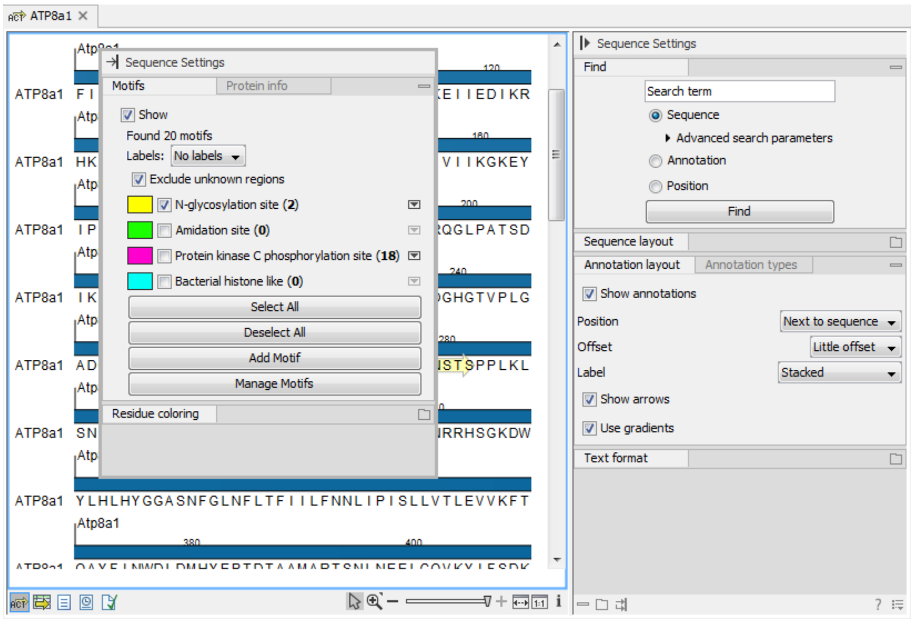


Figure 2.12: *Palettes can be organized in the Side Panel as you like or placed anywhere on the screen.*

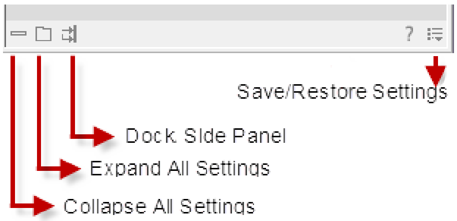


Figure 2.13: *Controlling the Side Panel at the bottom*

Note! Changes made to the **Side Panel**, including the organization of palettes will not be saved when you save the view. See how to save the changes in section 4.5

2.2 Zoom and selection in View Area

All views except tabular and text views support zooming. Figure 2.14 shows the zoom tools, located at the bottom right corner of the view.

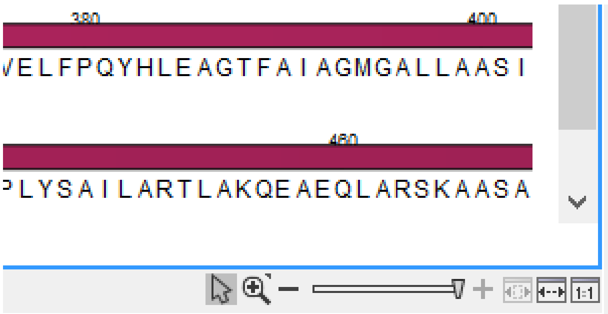


Figure 2.14: *The zoom tools are located at the bottom right corner of the view.*

The zoom tools consist of some shortcuts for zooming to fit the width of the view (↔), zoom to 100 % to see details (1:1), zoom to a selection (⬚), a zoom slider, and two mouse mode buttons (↖) (🔍).

The slider reflects the current zoom level and can be used to quickly adjust this. For more fine-grained control of the zoom level, move the mouse upwards while sliding.

The sections below describes how to use these tools as well as other ways of zooming and navigating data.

2.2.1 Zoom in

There are six ways of **zooming in**:

Click Zoom in mode (🔍) in the zoom tools (or press Ctrl+2) | click the location in. the view that you want to zoom in on

or **Click Zoom in mode (🔍) in the zoom tools | click-and-drag a box around a part of the view | the view now zooms in on the part you selected**

or **Press '+' on your keyboard**

or **Move the zoom slider located in the zoom tools**

or **Click the plus icon in the zoom tools**

The last option for zooming in is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac) | Move the scroll wheel on your mouse forward**

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you press the **Shift** button on your keyboard while in zoom mode, the zoom function is reversed.

If you want to zoom in to 100 % to see all the data, click the **Zoom to base level** (1:1) icon.

2.2.2 Zoom out

It is possible to zoom out in different ways:

Click Zoom out mode (🔍) in the zoom tools (or press Ctrl+3) | click in the view

or **Press '-' on your keyboard**

or **Move the zoom slider located in the zoom tools**

or **Click the minus icon in the zoom tools**

The last option for zooming out is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac) | Move the scroll wheel on your mouse backwards**

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you want to zoom out to see all the data, click the **Zoom to Fit** (↔) icon.

If you press **Shift** while clicking in a **View**, the zoom function is reversed. Hence, clicking on a sequence in this way while the **Zoom out** mode toolbar item is selected, zooms in instead of zooming out.

2.2.3 Selecting, panning and zooming

In the zoom tools, you can control which mouse mode to use. The default is **Selection mode** ( ) which is used for selecting data in a view. Next to the selection mode, you can select the **Zoom in mode** as described in section 2.2.1. If you press and hold this button, two other modes become available as shown in figure 2.15:

- **Panning** ( ) is used for dragging the view with the mouse as a way of scrolling.
- **Zoom out** ( ) is used to change the mouse mode so that whenever you click the view, it zooms out.

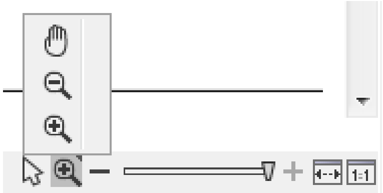


Figure 2.15: *Additional mouse modes can be found in the zoom tools.*

If you hold the mouse over the selection and zoom tools, tooltips will appear that provide further information about how to use the tools.

The mouse modes only apply when the mouse is within the view where they are selected.

The **Selection mode** can also be invoked with the keyboard shortcut Ctrl+1, while the **Panning mode** can be invoked with Ctrl+4.

For some views, if you have made a selection, there is a **Zoom to Selection** ( ) button, which allows you to zoom and scroll directly to fit the view to the selection.

2.3 Toolbox and Status Bar

The **Toolbox** is placed in the left side of the user interface of *CLC Cancer Research Workbench* below the **Navigation Area**.

The **Toolbox** shows a **Processes tab**, **Favorites tab** and a **Toolbox tab**.

The **Toolbox** can be hidden, so that the **Navigation Area** is enlarged and thereby displays more elements:

View | **Show/Hide Toolbox** | **Show/Hide Toolbox**

You can also click the **Hide Toolbox** ( ) button.

2.3.1 Processes

By clicking the **Processes** tab, the **Toolbox** displays previous and running processes, e.g. an NCBI search or a calculation of an alignment. The running processes can be stopped, paused, and resumed by clicking the small icon ( ) next to the process (see figure 2.16).

Running and paused processes are not deleted.

Besides the options to stop, pause and resume processes, there are some extra options for *a selected number* of the tools running from the Toolbox:

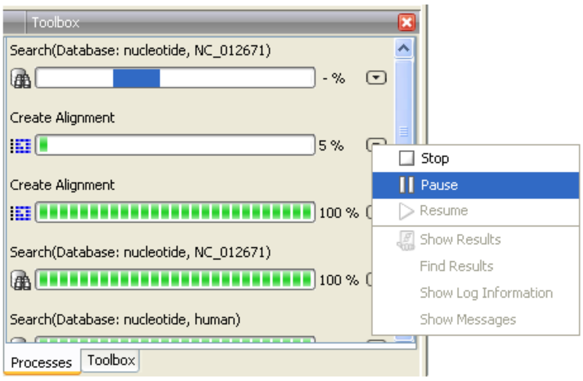


Figure 2.16: *A database search and an alignment calculation are running. Clicking the small icon next to the process allow you to stop, pause and resume processes.*

- **Show results**. If you have chosen to save the results (see section 8.2), you will be able to open the results directly from the process by clicking this option.
- **Find results**. If you have chosen to save the results (see section 8.2), you will be able to high-light the results in the Navigation Area.
- **Show Log Information**. This will display a log file showing progress of the process. The log file can also be shown by clicking **Show Log** in the "handle results" dialog where you choose between saving and opening the results.
- **Show Messages**. Some analyses will give you a message when processing your data. The messages are the black dialogs shown in the lower left corner of the Workbench that disappear after a few seconds. You can reiterate the messages that have been shown by clicking this option.

The terminated processes can be removed by:

View | Remove Finished Processes (✕)

If you close the program while there are running processes, a dialog will ask if you are sure that you want to close the program. Closing the program will stop the process, and it cannot be restarted when you open the program again.

2.3.2 Toolbox

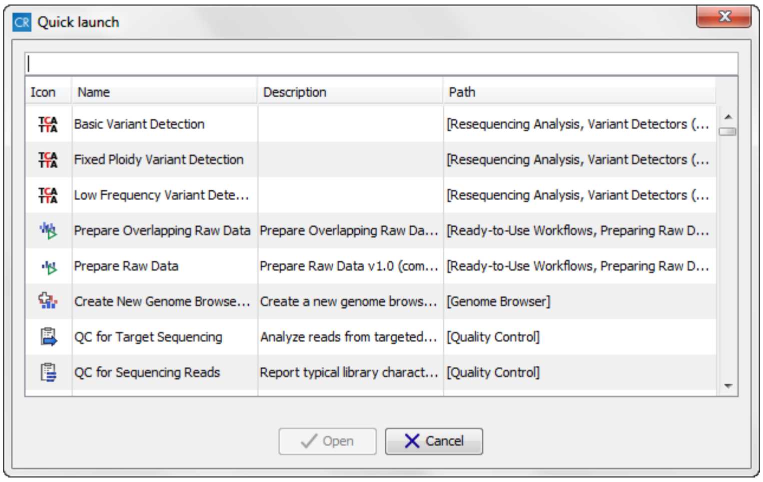
The content of the **Toolbox** tab in the **Toolbox** corresponds to **Toolbox** in the **Menu Bar**.

The tools in the toolbox can be accessed by double-clicking or by dragging elements from the **Navigation Area** to an item in the **Toolbox**.

Quick access to tools

To enable quick launch of tools in *CLC Cancer Research Workbench*, press Ctrl + Shift + T (⌘ + Shift + T on Mac) to show the quick launch dialog (see figure 2.17).

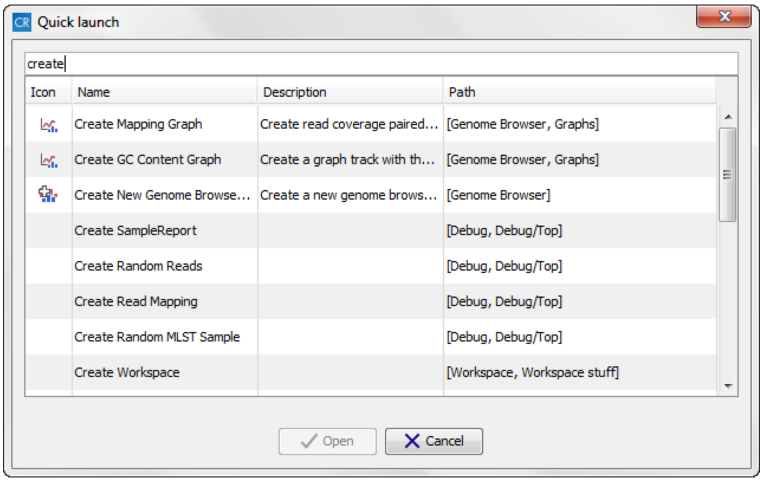
When the dialog is opened, you can start typing search text in the text field at the top. This will bring up the list of tools that match this text either in the name, description or location in the



Icon	Name	Description	Path
	Basic Variant Detection		[Resequencing Analysis, Variant Detectors (...
	Fixed Ploidy Variant Detection		[Resequencing Analysis, Variant Detectors (...
	Low Frequency Variant Dete...		[Resequencing Analysis, Variant Detectors (...
	Prepare Overlapping Raw Data	Prepare Overlapping Raw Da...	[Ready-to-Use Workflows, Preparing Raw D...
	Prepare Raw Data	Prepare Raw Data v1.0 (com...	[Ready-to-Use Workflows, Preparing Raw D...
	Create New Genome Browse...	Create a new genome brows...	[Genome Browser]
	QC for Target Sequencing	Analyze reads from targeted...	[Quality Control]
	QC for Sequencing Reads	Report typical library charact...	[Quality Control]

Figure 2.17: *Quick access to all tools in* **CLC Cancer Research Workbench**.

Toolbox. In the example shown in figure 2.18, typing `create` shows a list of tools involving the word "create", and the arrow keys or mouse can be used for selecting and starting a tool.



Icon	Name	Description	Path
	Create Mapping Graph	Create read coverage paired...	[Genome Browser, Graphs]
	Create GC Content Graph	Create a graph track with th...	[Genome Browser, Graphs]
	Create New Genome Browse...	Create a new genome brows...	[Genome Browser]
	Create SampleReport		[Debug, Debug/Top]
	Create Random Reads		[Debug, Debug/Top]
	Create Read Mapping		[Debug, Debug/Top]
	Create Random MLST Sample		[Debug, Debug/Top]
	Create Workspace		[Workspace, Workspace stuff]

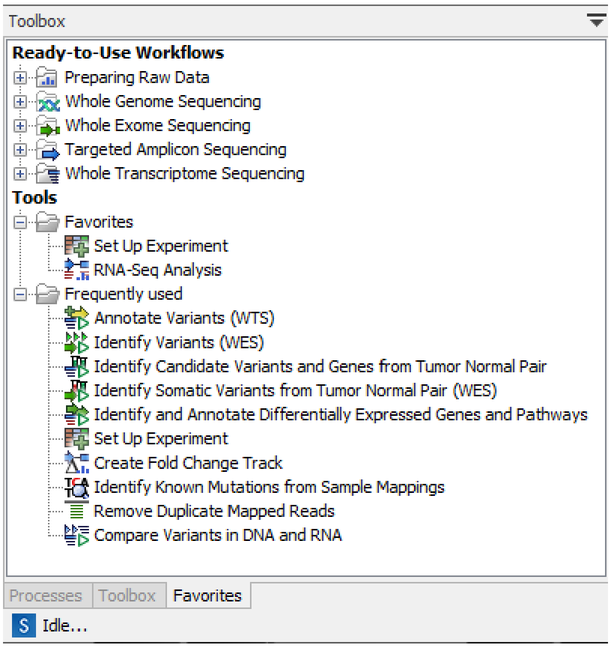
Figure 2.18: *Typing in the search field at the top will filter the list of tools to launch.*

Favorites toolbox

Next to the **Toolbox** tab, you find the **Favorites** tab. This can be used for organizing and getting quick access to the tools you use the most. It consists of two parts as shown in figure 2.19.

Favorites You can manually add tools to the favorites menu simply by right-clicking the tool in the Toolbox. You can also right-click the Favorites folder itself and select **Add Tool**. To remove a tool, right-click and select **Remove from Favorites**. Note that you can also add complete folders to the favorites.

Frequently used The list of tools in this folder is automatically populated as you use the Workbench. The most frequently used tools are listed at the top.



Toolbox
Ready-to-Use Workflows
Preparing Raw Data
Whole Genome Sequencing
Whole Exome Sequencing
Targeted Amplicon Sequencing
Whole Transcriptome Sequencing
Tools
Favorites
Set Up Experiment
RNA-Seq Analysis
Frequently used
Annotate Variants (WTS)
Identify Variants (WES)
Identify Candidate Variants and Genes from Tumor Normal Pair
Identify Somatic Variants from Tumor Normal Pair (WES)
Identify and Annotate Differentially Expressed Genes and Pathways
Set Up Experiment
Create Fold Change Track
Identify Known Mutations from Sample Mappings
Remove Duplicate Mapped Reads
Compare Variants in DNA and RNA
Processes
Toolbox
Favorites
S
Idle...

Figure 2.19: *Favorites toolbox.*

2.3.3 Status Bar

As can be seen from figure 2.1, the **Status Bar** is located at the bottom of the window. In the left side of the bar is an indication of whether the computer is making calculations or whether it is idle. The right side of the **Status Bar** indicates the range of the selection of a sequence. (See chapter 2.2.3 for more about the Selection mode button.)

2.4 Workspace

If you are working on a project and have arranged the views for this project, you can save this arrangement using **Workspaces**. A Workspace remembers the way you have arranged the views, and you can switch between different workspaces.

The **Navigation Area** always contains the same data across **Workspaces**. It is, however, possible to open different folders in the different **Workspaces**. Consequently, the program allows you to display different clusters of the data in separate **Workspaces**.

All **Workspaces** are automatically saved when closing down *CLC Cancer Research Workbench*. The next time you run the program, the **Workspaces** are reopened exactly as you left them.

Note! It is not possible to run more than one version of *CLC Cancer Research Workbench* at a time. Use two or more **Workspaces** instead.

2.4.1 Create Workspace

When working with large amounts of data, it might be a good idea to split the work into two or more **Workspaces**. As default the *CLC Cancer Research Workbench* opens one **Workspace**. Additional **Workspaces** are created in the following way:

Workspace in the Menu Bar) | **Create Workspace** | **enter name of Workspace** | **OK**

When the new **Workspace** is created, the heading of the program frame displays the name of the new **Workspace**. Initially, the selected elements in the **Navigation Area** is collapsed and the **View Area** is empty and ready to work with. (See figure 2.20).

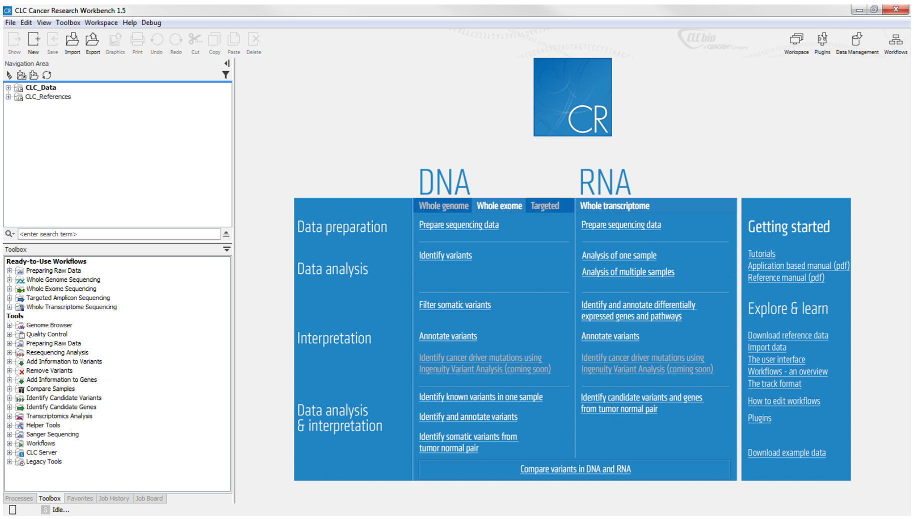

CLC Cancer Research Workbench 1.5
File Edit View Toolbox Workspace Help Debug
Show New Save Import Export Graphics Print Undo Redo Cut Copy Paste Delete
Workspace Plugins Data Management Workflows
Navigation Area
CLC_Data
CLC_References
<enter search term>
Toolbox
Ready-to-Use Workflows
Preparing Raw Data
Whole Genome Sequencing
Whole Exome Sequencing
Targeted Amplicon Sequencing
Whole Transcriptome Sequencing
Tools
Genome Browser
Quality Control
Preparing Raw Data
Resequencing Analysis
Add Information to Variants
Remove Variants
Add Information to Genes
Compare Samples
Identify Candidate Variants
Identify Candidate Genes
Transcriptomics Analysis
Helper Tools
Sanger Sequencing
Workflows
CLC Server
Legacy Tools
Processes Toolbox Favorites Job History Job Board
Idle...
CR
DNA
RNA
Whole genome Whole exome Targeted
Whole transcriptome
Data preparation
Data analysis
Interpretation
Data analysis & interpretation
Prepare sequencing data
Identify variants
Filter somatic variants
Annotate variants
Identify cancer driver mutations using Ingenuity Variant Analysis (coming soon)
Identify known variants in one sample
Identify and annotate variants
Identify somatic variants from tumor normal pair
Prepare sequencing data
Analysis of one sample
Analysis of multiple samples
Identify and annotate differentially expressed genes and pathways
Annotate variants
Identify cancer driver mutations using Ingenuity Variant Analysis (coming soon)
Identify candidate variants and genes from tumor normal pair
Compare variants in DNA and RNA
Getting started
Tutorials
Application based manual (pdf)
Reference manual (pdf)
Explore & learn
Download reference data
Import data
The user interface
Workflows - an overview
The track format
How to edit workflows
Plugins
Download example data

Figure 2.20: *An empty Workspace.*

2.4.2 Select Workspace

When there is more than one **Workspace** in the *CLC Cancer Research Workbench*, there are two ways to switch between them:

Workspace (▣) in the Toolbar | Select the Workspace to activate

or **Workspace in the Menu Bar | Select Workspace (▣) | choose which Workspace to activate | OK**

The name of the selected **Workspace** is shown after "*CLC Cancer Research Workbench*" at the top left corner of the main window, in figure 2.20 it says: (default).

2.4.3 Delete Workspace

Deleting a **Workspace** can be done in the following way:

Workspace in the Menu Bar | Delete Workspace | choose which Workspace to delete | OK

Note! Be careful to select the right **Workspace** when deleting. The delete action cannot be undone. (However, no data is lost, because a workspace is only a representation of data.)

It is not possible to delete the default workspace.

2.5 List of shortcuts

The keyboard shortcuts in *CLC Cancer Research Workbench* are listed below.

Action	**Windows/Linux**	**Mac OS X**
Adjust selection	Shift + arrow keys	Shift + arrow keys
Adjust workflow layout	Shift + Alt + L	⌘ + Shift + Alt + L
Close	Ctrl + W	⌘ + W
Close all views	Ctrl + Shift + W	⌘ + Shift + W
Copy	Ctrl + C	⌘ + C
Create track list	Ctrl + L	⌘ + L
Cut	Ctrl + X	⌘ + X
Delete	Delete	Delete or ⌘ + Backspace
Exit	Alt + F4	⌘ + Q
Export	Ctrl + E	⌘ + E
Export graphics	Ctrl + G	⌘ + G
Find Next Conflict	'.' (dot)	'.' (dot)
Find Previous Conflict	',' (comma)	',' (comma)
Help	F1	F1
Import	Ctrl + I	⌘ + I
Maximize/restore size of View	Ctrl + M	⌘ + M
Move gaps in alignment	Ctrl + arrow keys	⌘ + arrow keys
New Folder	Ctrl + Shift + N	⌘ + Shift + N
Panning Mode	Ctrl + 4	⌘ + 4
Paste	Ctrl + V	⌘ + V
Print	Ctrl + P	⌘ + P
Redo	Ctrl + Y	⌘ + Y
Rename	F2	F2
Save	Ctrl + S	⌘ + S
Save As	Ctrl + Shift + S	⌘ + Shift + S
Scrolling horizontally	Shift + Scroll wheel	Shift + Scroll wheel
Search local data	Ctrl + Shift + F	⌘ + Shift + F
Search via Side Panel	Ctrl + F	⌘ + F
Select All	Ctrl + A	⌘ + A
Select Selection Mode	Ctrl + 1 (one)	⌘ + 1 (one)
Show folder content	Ctrl + O	⌘ + O
Show/hide Side Panel	Ctrl + U	⌘ + U
Sort folder	Ctrl + Shift + R	⌘ + Shift + R
Split Horizontally	Ctrl + T	⌘ + T
Split Vertically	Ctrl + J	⌘ + J
Start Tool Quick Launch	Ctrl + Shift + T	⌘ + Shift + T
Undo	Ctrl + Z	⌘ + Z
Update folder	F5	F5
User Preferences	Ctrl + K	⌘ + ;
Vertical scroll in read tracks	Alt + Scroll wheel	Alt + Scroll wheel
Vertical scroll in reads tracks, fast	Shift+Alt+Scroll wheel	Shift+Alt+Scroll wheel
Vertical zoom in graph tracks	Alt + Scroll wheel	Alt + Scroll wheel

Action	Windows/Linux	Mac OS X
Reverse zoom mode	press and hold Shift	press and hold Shift
Workflow, add element	Alt + Shift + E	Alt + Shift + E
Workflow, collapse if its expanded	Alt + Shift + '-' (minus)	Alt + Shift + '-'
Workflow, create installer	Alt + Shift + I	Alt + Shift + I
Workflow, execute	Ctrl + enter	⌘ + enter
Workflow, expand if its collapsed	Alt + Shift + '+' (plus)	Alt + Shift + '-'
Workflow, highlight used elements	Alt + Shift + U	Alt + Shift + U
Workflow, remove all elements	Alt + Shift + R	Alt + Shift + R
Zoom	Ctrl + Scroll wheel	Ctrl + Scroll wheel
Zoom In Mode	Ctrl + 2	⌘ + 2
Zoom In (without clicking)	'+' (plus)	'+' (plus)
Zoom Out Mode	Ctrl + 3	⌘ + 3
Zoom Out (without clicking)	'-' (minus)	'-' (minus)

Combinations of keys and mouse movements are listed below.

Action	Windows/Linux	Mac OS X	Mouse movement
Maximize View			Double-click the tab of the View
Restore View			Double-click the View title
Reverse zoom mode	Shift	Shift	Click in view
Select multiple elements that are not grouped together	Ctrl	⌘	Click elements
Select multiple elements that are grouped together	Shift	Shift	Click elements

"Elements" in this context refers to elements and folders in the **Navigation Area** selections on sequences, and rows in tables.

Chapter 3

Data organization and management

Contents

This chapter explains the data management features of *CLC Cancer Research Workbench*. The first section explains the basics of the data organization and the **Navigation Area**. The next section explains how to set up custom attributes for the data that can be used for more advanced data management. Finally, there is a section about how to search through local data.

3.1 Navigation Area

The **Navigation Area** is located in the left side of the screen, under the **Toolbar** (see figure 3.1). It is used for organizing and navigating data. Its behavior is similar to the way files and folders are usually displayed on your computer.

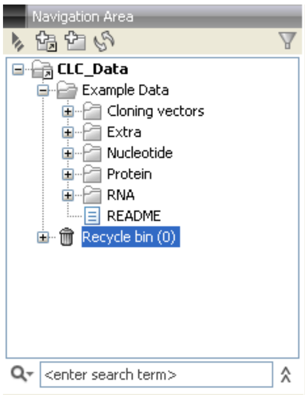


Figure 3.1: *The Navigation Area.*

To provide more space for viewing data, you can hide **Navigation Area** and the **Toolbox** by clicking the hide icon (◀|) at the top.

3.1.1 Data structure

The data in the **Navigation Area** is organized into a number of **Locations**. When the *CLC Cancer Research Workbench* is started for the first time, there is one location called *CLC_Data* (unless your computer administrator has configured the installation otherwise).

A location represents a folder on the computer: The data shown under a location in the **Navigation Area** is stored on the computer in the folder which the location points to.

This is explained visually in figure 3.2. The full path to the system folder can be located by mousing over the data location as shown in figure 3.3.

Adding locations

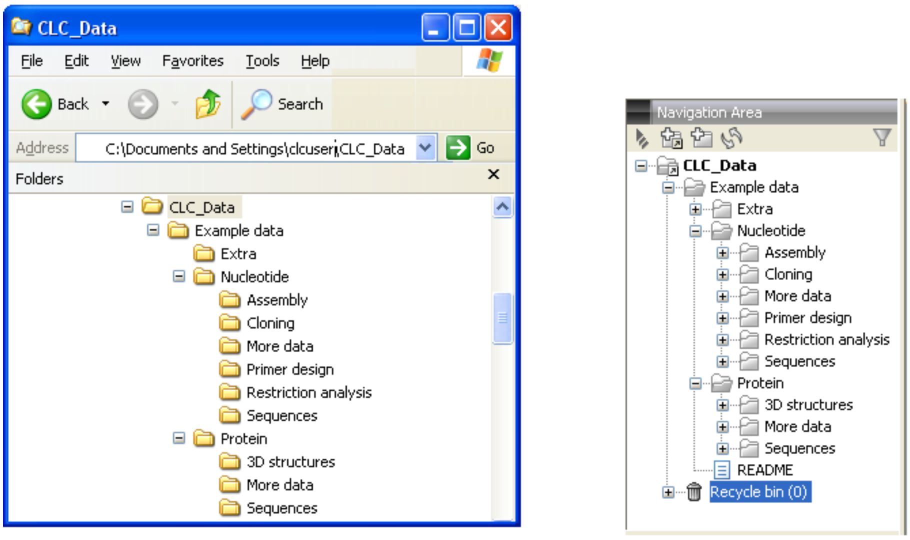
Per default, there is one location in the **Navigation Area** called CLC_Data. It points to the following folder:

- On Windows: C:\Documents and settings\<username>\CLC_Data
- On Mac: ~/CLC_Data
- On Linux: /homefolder/CLC_Data

You can easily add more locations to the **Navigation Area**:

File | New | Location ()

This will bring up a dialog where you can navigate to the folder you wish to use as your new location (see figure 3.4).



Address: C:\Documents and Settings\clcuser\CLC_Data

Figure 3.2: *In this example the location called 'CLC_Data' points to the folder at C:\Documents and settings\clcuser\CLC_Data.*

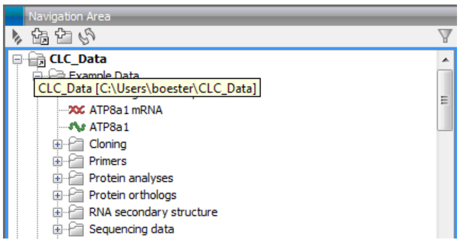
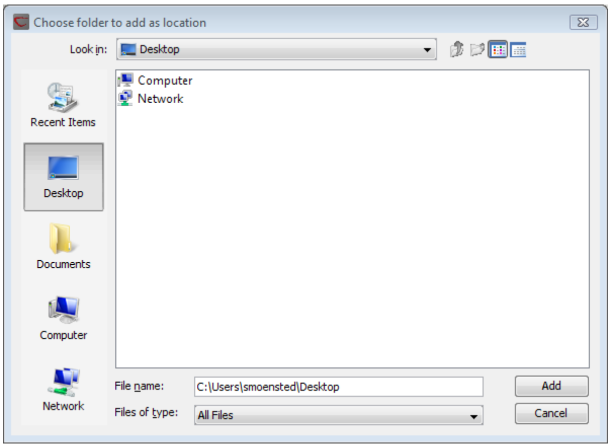


Figure 3.3: *Mousing over the location called 'CLC_Data' shows the full path to the system folder, which in this case is C:\Users\boester\CLC_Data.*



Look in: Desktop

File name: C:\Users\smoensted\Desktop

Files of type: All Files

Figure 3.4: *Navigating to a folder to use as a new location.*

When you click **Open**, the new location is added to the **Navigation Area** as shown in figure 3.5.

The name of the new location will be the name of the folder selected for the location. To see where the folder is located on your computer, place your mouse cursor on the location icon (🗂) for second. This will show the path to the location.

Sharing data is possible of you add a location on a network drive. The procedure is similar to the one described above. When you add a location on a network drive or a removable drive, the location will appear *inactive* when you are not connected. Once you connect to the drive again,

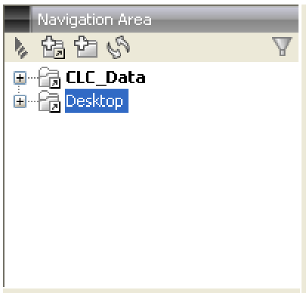


Figure 3.5: *The new location has been added.*

click **Update All** (↻) and it will become active (note that there will be a few seconds' delay from you connect).

Opening data

The elements in the **Navigation Area** are opened by :

Double-click the element

or **Click the element | Show (⇥) in the Toolbar | Select the desired way to view the element**

This will open a view in the **View Area**, which is described in section 2.1.

Adding data

Data can be added to the **Navigation Area** in a number of ways. Files can be imported from the file system (see chapter 6). Furthermore, an element can be added by dragging it into the **Navigation Area**. This could be views that are open, elements on lists, e.g. search hits or sequence lists, and files located on your computer. Finally, you can add data by adding a new location (see section 3.1.1).

If a file or another element is dropped on a folder, it is placed at the bottom of the folder. If it is dropped on another element, it will be placed just below that element.

If the element already exists in the **Navigation Area**, you will be asked whether you wish to create a copy.

3.1.2 Create new folders

In order to organize your files, they can be placed in folders. Creating a new folder can be done in two ways:

right-click an element in the Navigation Area | New | Folder (📁)

or **File | New | Folder (📁)**

If a folder is selected in the **Navigation Area** when adding a new folder, the new folder is added at the bottom of this folder. If an element is selected, the new folder is added right above that element.

You can move the folder manually by selecting it and dragging it to the desired destination.

3.1.3 Sorting folders

You can sort the elements in a folder alphabetically:

right-click the folder | Sort Folder

On Windows, subfolders will be placed at the top of the folder, and the rest of the elements will be listed below in alphabetical order. On Mac, both subfolders and other elements are listed together in alphabetical order.

3.1.4 Multiselecting elements

Multiselecting elements means that you select more than one element at the same time. This can be done in the following ways:

- Holding down the <Ctrl> key (⌘ on Mac) while clicking on multiple elements selects the elements that have been clicked.
- Selecting one element, and selecting another element while holding down the <Shift> key selects all the elements listed between the two locations (the two end locations included).
- Selecting one element, and moving the curser with the arrow-keys while holding down the <Shift> key, enables you to increase the number of elements selected.

3.1.5 Moving and copying elements

Elements can be moved and copied in several ways:

- Using **Copy** (❐), **Cut** (✂) and **Paste** (📄) from the **Edit** menu.
- Using Ctrl + C (⌘ + C on Mac), Ctrl + X (⌘ + X on Mac) and Ctrl + V (⌘ + V on Mac).
- Using **Copy** (❐), **Cut** (✂) and **Paste** (📄) in the **Toolbar**.
- Using drag and drop to move elements.
- Using drag and drop while pressing Ctrl / Command to copy elements.

In the following, all of these possibilities for moving and copying elements are described in further detail.

Copy, cut and paste functions

Copies of elements and folders can be made with the copy/paste function which can be applied in a number of ways:

select the files to copy | right-click one of the selected files | Copy (❐) | right-click the location to insert files into | Paste (📄)

or **select the files to copy | Ctrl + C (⌘ + C on Mac) | select where to insert files | Ctrl + P (⌘ + P on Mac)**

or **select the files to copy | Edit in the Menu Bar | Copy (❐) | select where to insert files | Edit in the Menu Bar | Paste (📄)**

If there is already an element of that name, the pasted element will be renamed by appending a number at the end of the name.

Elements can also be moved instead of copied. This is done with the cut/paste function:

select the files to cut | right-click one of the selected files | Cut (✂) | right-click the location to insert files into | Paste (📋)

or **select the files to cut | Ctrl + X (⌘ + X on Mac) | select where to insert files | Ctrl + V (⌘ + V on Mac)**

When you have cut the element, it is "grayed out" until you activate the paste function. If you change your mind, you can revert the cut command by copying another element.

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Move using drag and drop

Using drag and drop in the **Navigation Area**, as well as in general, is a four-step process:

click the element | click on the element again, and hold left mouse button | drag the element to the desired location | let go of mouse button

This allows you to:

- Move elements between different folders in the **Navigation Area**
- Drag from the **Navigation Area** to the **View Area**: A new view is opened in an existing **View Area** if the element is dragged from the **Navigation Area** and dropped next to the tab(s) in that **View Area**.
- Drag from the **View Area** to the **Navigation Area**: The element, e.g. a sequence, alignment, search report etc. is saved where it is dropped. If the element already exists, you are asked whether you want to save a copy. You drag from the **View Area** by dragging the tab of the desired element.

Use of drag and drop is supported throughout the program, also to open and re-arrange views (see section 2.1.6).

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Copy using drag and drop

To copy instead of move using drag and drop, hold the Ctrl (⌘ on Mac) key while dragging:

click the element | click on the element again, and hold left mouse button | drag the element to the desired location | press Ctrl (⌘ on Mac) while you let go of mouse button release the Ctrl/⌘ button

3.1.6 Change element names

This section describes two ways of changing the names of sequences in the **Navigation Area**. In the first part, the sequences themselves are not changed - it's their representation that changes.

The second part describes how to change the name of the element.

Change how sequences are displayed

Sequence elements can be displayed in the **Navigation Area** with different types of information:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Latin name.
- Latin name (accession).
- Common name.
- Common name (accession).

Whether sequences can be displayed with this information depends on their origin. Sequences that you have created yourself or imported might not include this information, and you will only be able to see them represented by their name. However, sequences downloaded from databases like GenBank will include this information. To change how sequences are displayed:

right-click any element or folder in the Navigation Area | Sequence Representation | select format

This will only affect sequence elements, and the display of other types of elements, e.g. alignments, trees and external files, will be not be changed. If a sequence does not have this information, there will be no text next to the sequence icon.

Rename element

Renaming a folder or an element in the **Navigation Area** can be done in two different ways:

select the element | Edit in the Menu Bar | Rename

or **select the element | F2**

When you can rename the element, you can see that the text is selected and you can move the cursor back and forth in the text. When the editing of the name has finished; press **Enter** or select another element in the **Navigation Area**. If you want to discard the changes instead, press the **Esc**-key.

3.1.7 Delete, restore and remove elements

When one deletes data from a data folder in the Workbench, it is moved to the recycle bin in that data location. Each data location has its own recycle bin. From the recycle bin, data can then be restored, or completely removed. Removal of data from the recycle bin frees disk space.

Deleting a folder or an element from a Workbench data location can be done in two ways:

right-click the element | Delete (☒)

or **select the element | press Delete key**

This will cause the element to be moved to the **Recycle Bin** ( ) where it is kept until the recycle bin is emptied or until you choose to restore the data object to your data location.

Items in a recycle bin can be restored in two ways:

Drag the elements with the mouse into the folder where they used to be.

or **select the element | right click and choose the option Restore**.

Once restored, you can continue to work with that data.

All contents of the recycle bin can be removed by choosing to empty the recycle bin:

Edit in the Menu Bar | Empty Recycle Bin ( )

This deletes the data and frees up disk space.

Note! This cannot be undone. Data is not recoverable after it is removed by emptying the recycle bin.

3.1.8 Show folder elements in a table

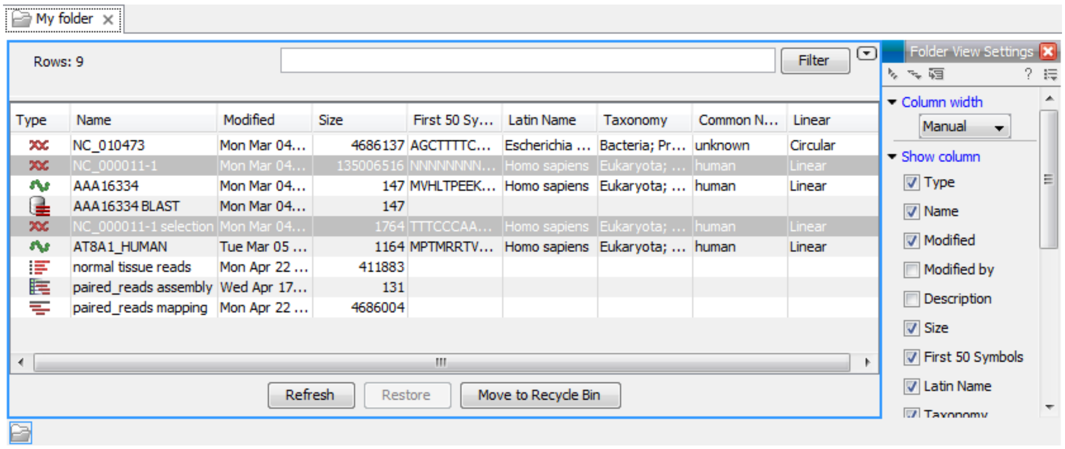
A location or a folder might contain large amounts of elements. It is possible to view their elements in the **View Area**:

select a folder or location | Show ( ) in the Toolbar

or

select a folder or location | right click on the folder and select **Show ( )** | **Contents ( )**

An example is shown in figure 3.6.



Manual

Show column

- ☑ Type
- ☑ Name
- ☑ Modified
- □ Modified by
- □ Description
- ☑ Size
- ☑ First 50 Symbols
- ☑ Latin Name
- ☑ Taxonomy

Type	Name	Modified	Size	First 50 Sy...	Latin Name	Taxonomy	Common N...	Linear
	NC_010473	Mon Mar 04...	4686137	AGCTTTTC...	Escherichia ...	Bacteria; Pr...	unknown	Circular
	NC_000011-1	Mon Mar 04...	135006516	NNNNNNNN...	Homo sapiens	Eukaryota; ...	human	Linear
	AAA16334	Mon Mar 04...	147	MVHLTPEEK...	Homo sapiens	Eukaryota; ...	human	Linear
	AAA16334 BLAST	Mon Mar 04...	147					
	NC_000011-1 selection	Mon Mar 04...	1764	TTTCCCAA...	Homo sapiens	Eukaryota; ...	human	Linear
	AT8A1_HUMAN	Tue Mar 05 ...	1164	MPTMRRTV...	Homo sapiens	Eukaryota; ...	human	Linear
	normal tissue reads	Mon Apr 22 ...	411883					
	paired_reads assembly	Wed Apr 17...	131					
	paired_reads mapping	Mon Apr 22 ...	4686004					

Figure 3.6: *Viewing the elements in a folder.*

When the elements are shown in the view, they can be sorted by clicking the heading of each of the columns. You can further refine the sorting by pressing Ctrl (⌘ on Mac) while clicking the heading of another column.

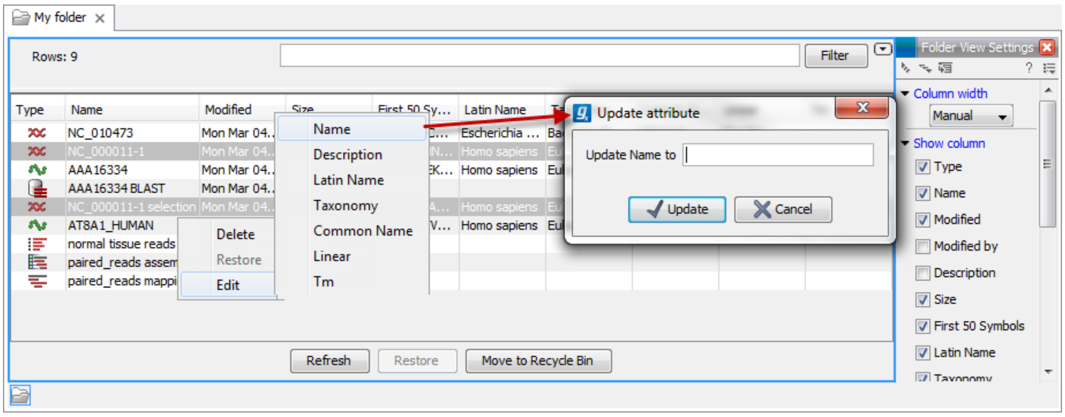
Sorting the elements in a view does not affect the ordering of the elements in the **Navigation Area**.

Note! The view only displays one "layer" at a time: the content of subfolders is not visible in this view. Also note that only sequences have the full span of information like organism etc.

Batch edit folder elements

You can select a number of elements in the table, right-click and choose **Edit** to batch edit the elements. In this way, you can change for example the description or name of several elements in one go.

In figure 3.7 you can see an example where the name of two sequence are renamed in one go. In this example, a dialog with a text field will be shown, letting you enter a new name for these two sequences.



Rows: 9

Type	Name	Modified	Size	First 50 Sy...	Latin Name	Ta
	NC_010473	Mon Mar 04..		C...	Escherichia ...	Ba
	NC_000011-1	Mon Mar 04..		IN...	Homo sapiens	Eu
	AAA16334	Mon Mar 04..		EK...	Homo sapiens	Eu
	AAA16334 BLAST	Mon Mar 04..				
	NC_000011-1 selection	Mon Mar 04..		A...	Homo sapiens	Eu
	AT8A1_HUMAN			V...	Homo sapiens	Eu
	normal tissue reads					
	paired_reads assem					
	paired_reads mappi					

Update Name to

Column width: Manual

Show column
☑ Type
☑ Name
☑ Modified
☐ Modified by
☐ Description
☑ Size
☑ First 50 Symbols
☑ Latin Name
☑ Taxonomy

Figure 3.7: *Changing the common name of two sequences.*

Note! This information is directly saved and you cannot undo.

Drag and drop folder elements

You can drag and drop objects from the folder editor to the **Navigation area**. This will create a copy of the objects at the selected destination. New elements can be included in the folder editor in the view area by dragging and dropping an element from a destination in the **Navigation Area** to the folder in the **Navigation Area** that you have open in the view area. It is not possible to drag elements directly from the **Navigation Area** to the folder editor in the View area.

3.2 Customized attributes on data locations

The *CLC Cancer Research Workbench* makes it possible to define location-specific attributes on all elements stored in a data location. This could be company-specific information such as LIMS id, freezer position etc. Note that the attributes scheme belongs to a location, so if you have added multiple locations, they will have their own separate set of attributes.

Note! For *CLC Genomics Workbench*and *CLC Main Workbench*, a Metadata Import Plugin is available. The plugin consists of two tools: "Import Sequences in Table Format" and "Associate with metadata". These tools allow sequences to be imported from a tabular data source and make it possible to add metadata to existing objects.

3.2.1 Configuring which fields should be available

To configure which fields that should be available[1] go to the Workbench:

right-click the data location | Location | Attribute Manager

This will display the dialog shown in figure 3.8.

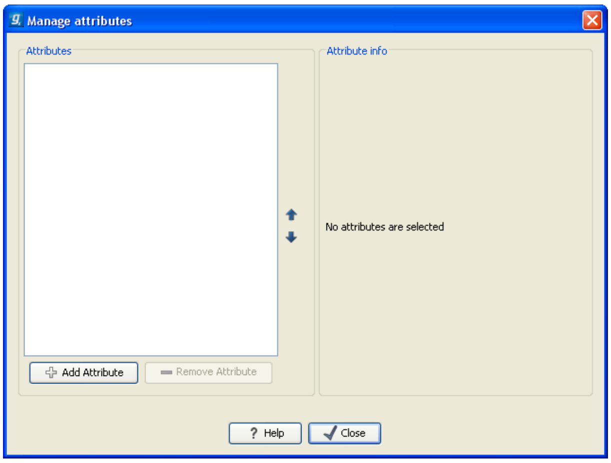


Figure 3.8: *Adding attributes.*

Click the **Add Attribute** (✚) button to create a new attribute. This will display the dialog shown in figure 3.9.

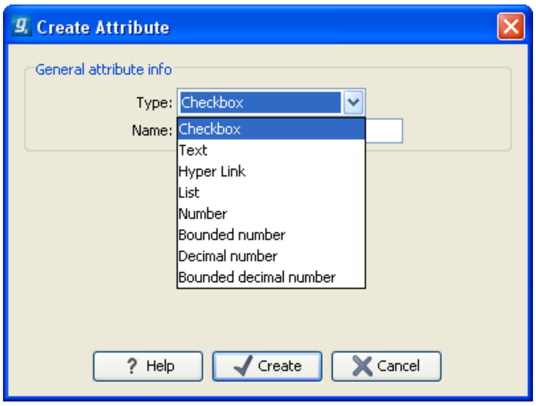


Figure 3.9: *The list of attribute types.*

First, select what kind of attribute you wish to create. This affects the type of information that can be entered by the end users, and it also affects the way the data can be searched. The following types are available:

- **Checkbox**. This is used for attributes that are binary (e.g. true/false, checked/unchecked and yes/no).
- **Text**. For simple text with no constraints on what can be entered.
- **Hyper Link**. This can be used if the attribute is a reference to a web page. A value of this type will appear to the end user as a hyper link that can be clicked. Note that this

[1] If the data location is a server location, you need to be a server administrator to do this

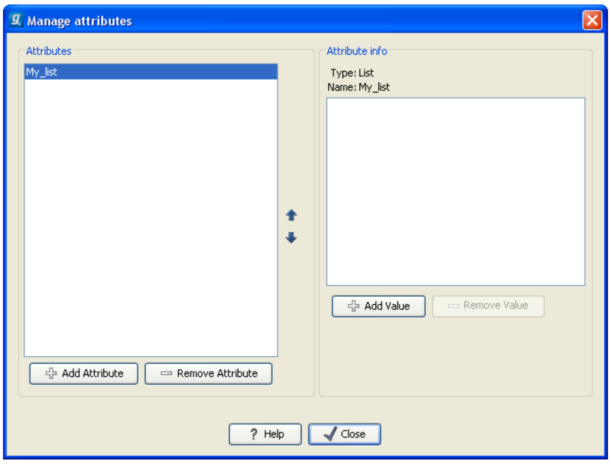
attribute can only contain one hyper link. If you need more, you will have to create additional attributes.

- **List**. Lets you define a list of items that can be selected (explained in further detail below).
- **Number**. Any positive or negative integer.
- **Bounded number**. Same as number, but you can define the minimum and maximum values that should be accepted. If you designate some kind of ID to your sequences, you can use the bounded number to define that it should be at least 1 and max 99999 if that is the range of your IDs.
- **Decimal number**. Same as number, but it will also accept decimal numbers.
- **Bounded decimal number**. Same as bounded number, but it will also accept decimal numbers.

When you click **OK**, the attribute will appear in the list to the left. Clicking the attribute will allow you to see information on its type in the panel to the right.

3.2.2 Editing lists

Lists are a little special, since you have to define the items in the list. When you click a list in the left side of the dialog, you can define the items of the list in the panel to the right by clicking **Add Item** (✚) (see figure 3.10).



Type: List
Name: My_list

Figure 3.10: *Defining items in a list.*

Remove items in the list by pressing **Remove Item** (▭).

3.2.3 Removing attributes

To remove an attribute, select the attribute in the list and click **Remove Attribute** (▭). This can be done without any further implications if the attribute has just been created, but if you remove an attribute where values have already been given for elements in the data location, it will have implications for these elements: The values will not be removed, but they will become static, which means that they cannot be edited anymore.

If you accidentally removed an attribute and wish to restore it, this can be done by creating a new attribute of exactly the same name and type as the one you removed. All the "static" values will now become editable again.

When you remove an attribute, it will no longer be possible to search for it, even if there is "static" information on elements in the data location.

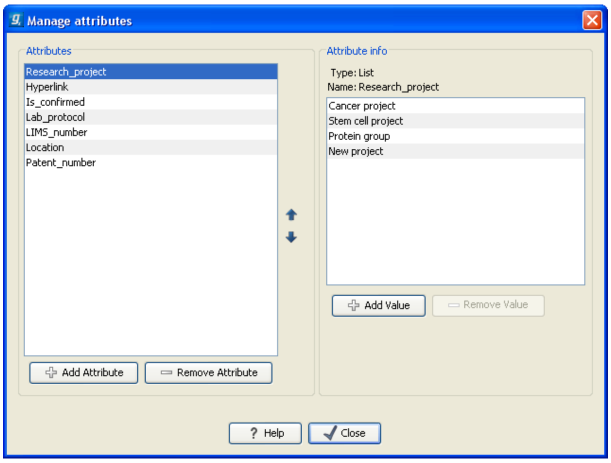
Renaming and changing the type of an attribute is not possible - you will have to create a new one.

3.2.4 Changing the order of the attributes

You can change the order of the attributes by selecting an attribute and click the **Up** and **Down** arrows in the dialog. This will affect the way the attributes are presented for the user.

3.3 Filling in values

When a set of attributes has been created (as shown in figure 3.11), the end users can start filling in information.



Research_project
Hyperlink
Is_confirmed
Lab_protocol
LIMS_number
Location
Patent_number

Type: List
Name: Research_project

Cancer project
Stem cell project
Protein group
New project

Figure 3.11: *A set of attributes defined in the attribute manager.*

This is done in the element info view:

> **right-click a sequence or another element in the Navigation Area | Show ( ) | Element info ( )**

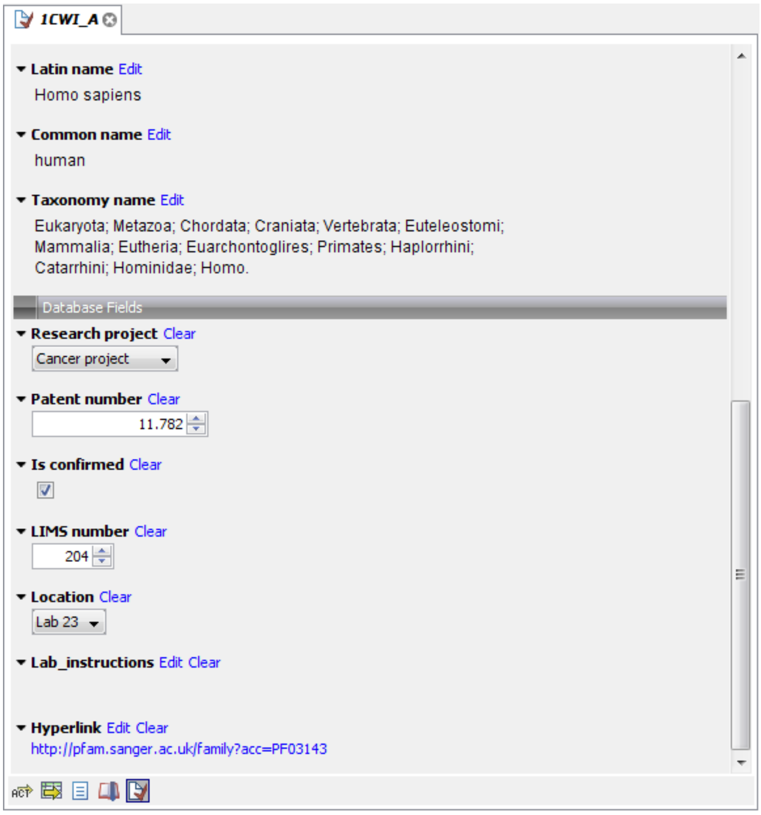
This will open a view similar to the one shown in figure 3.12.

You can now enter the appropriate information and **Save**. When you have saved the information, you will be able to search for it (see below).

Note that the element (e.g. sequence) needs to be saved in the data location before you can edit the attribute values.

When nobody has entered information, the attribute will have a "Not set" written in red next to the attribute (see figure 3.13).

This is particularly useful for attribute types like checkboxes and lists where you cannot tell, from the displayed value, if it has been set or not. Note that when an attribute has not been set, you



1CWI_A

Latin name Edit

Homo sapiens

Common name Edit

human

Taxonomy name Edit

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

Database Fields

Research project Clear

Cancer project

Patent number Clear

11.782

Is confirmed Clear

☑

LIMS number Clear

204

Location Clear

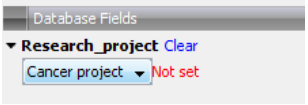
Lab 23

Lab_instructions Edit Clear

Hyperlink Edit Clear

http://pfam.sanger.ac.uk/family?acc=PF03143

Figure 3.12: *Adding values to the attributes.*



Database Fields

Research_project Clear

Cancer project Not set

Figure 3.13: *An attribute which has not been set.*

cannot search for it, even if it looks like it has a value. In figure 3.13, you will *not* be able to find this sequence if you search for research projects with the value "Cancer project", because it has not been set. To set it, simply click in the list and you will see the red "Not set" disappear.

If you wish to reset the information that has been entered for an attribute, press "Clear" (written in blue next to the attribute). This will return it to the "Not set" state.

The **Folder editor**, invoked by pressing **Show** on a given folder from the context menu, provides a quick way of changing the attributes of many elements in one go (see section 3.1.8).

3.3.1 What happens when a clc object is copied to another data location?

The user supplied information, which has been entered in the **Element info**, is attached to the attributes that have been defined in this particular data location. If you copy the sequence to another data location or to a data location containing another attribute set, the information will become fixed, meaning that it is no longer editable and cannot be searched for. Note that attributes that were "Not set" will disappear when you copy data to another location.

If the element (e.g. sequence) is moved back to the original data location, the information will again be editable and searchable.

If the e.g. Molecule Project or Molecule Table is moved back to the original data location, the information will again be editable and searchable.

3.3.2 Searching

When an attribute has been created, it will automatically be available for searching. This means that in the **Local Search** ( ), you can select the attribute in the list of search criteria (see figure 3.14).

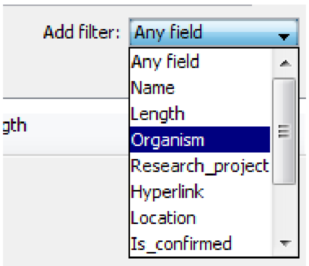


Figure 3.14: *The attributes from figure 3.11 are now listed in the search filter.*

It will also be available in the **Quick Search** below the **Navigation Area** (press Shift+F1 (Fn+Shift+F1 on Mac) and it will be listed - see figure 3.15).

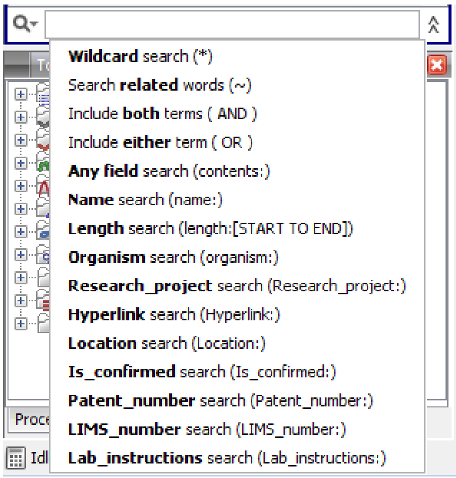

Wildcard search (*)
Search **related** words (~)
Include **both** terms (AND)
Include **either** term (OR)
Any field search (contents:)
Name search (name:)
Length search (length:[START TO END])
Organism search (organism:)
Research_project search (Research_project:)
Hyperlink search (Hyperlink:)
Location search (Location:)
Is_confirmed search (Is_confirmed:)
Patent_number search (Patent_number:)
LIMS_number search (LIMS_number:)
Lab_instructions search (Lab_instructions:)
Proce
Idl

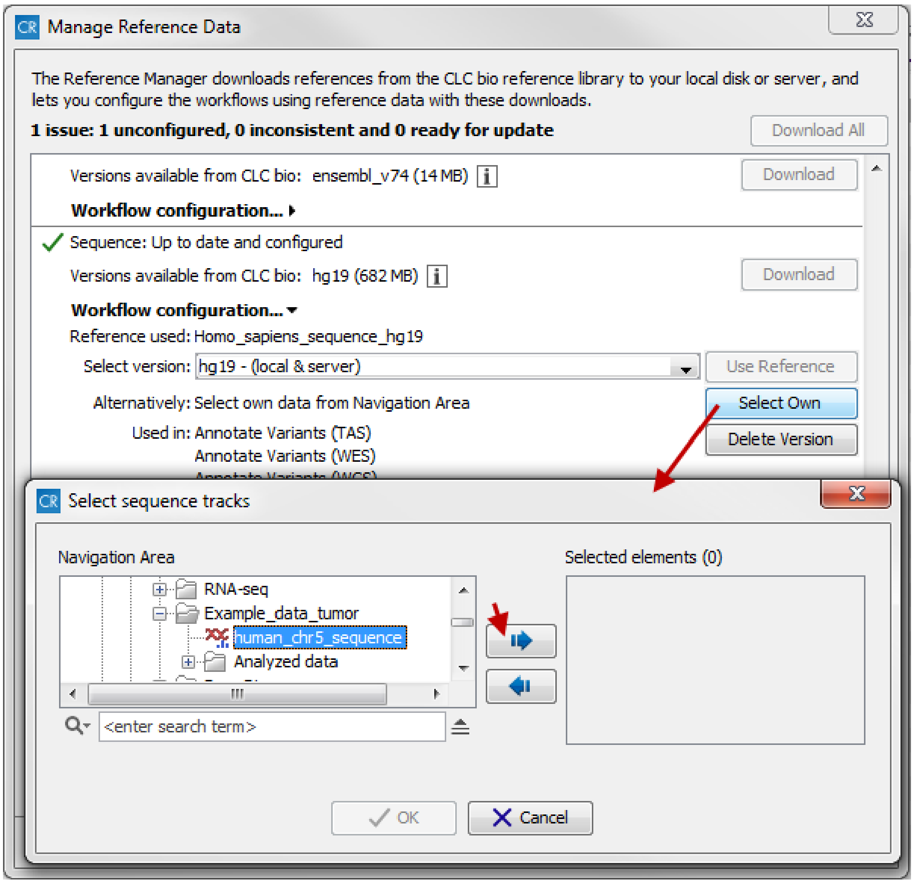
Figure 3.15: *The attributes from figure 3.11 are now available in the Quick Search as well.*

3.3.3 Import your own reference data

There may be situations where you would like to use your own reference data rather than the reference data that is available under **Data management**. You can specify which data you would like to import under **Data management**. To import your own data:

Data Management ( ) | **Workflow configuration** | **Select Own**

When you have clicked on the button labeled **Select Own** that appears when you have clicked on the text **Workflow configuration**, a window opens up that allow you to find and select your own reference data (figure 3.16). Select your reference data by double-clicking on the file name or by clicking once on the file name and pressing once on the arrow in the middle of the window pointing to the right.



Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

1 issue: 1 unconfigured, 0 inconsistent and 0 ready for update

Versions available from CLC bio: ensembl_v74 (14 MB)

Workflow configuration...

✓ Sequence: Up to date and configured

Versions available from CLC bio: hg19 (682 MB)

Workflow configuration...

Reference used: Homo_sapiens_sequence_hg19

Select version: hg19 - (local & server)

Alternatively: Select own data from Navigation Area

Used in: Annotate Variants (TAS)
Annotate Variants (WES)

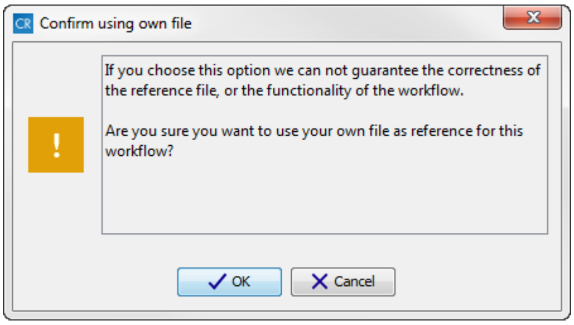
Select sequence tracks

Navigation Area

Selected elements (0)

Figure 3.16: *Click on the bold face text "Workflow configuration" to expand the view and press the button labeled "Select Own". This will create access to the file location where you have stored your own reference data file.*

When you have selected your own reference data and click on the button labeled **Next** a warning will appear that ask you about whether you really want to use your own file rather than the provided reference data (figure 3.17).



Confirm using own file

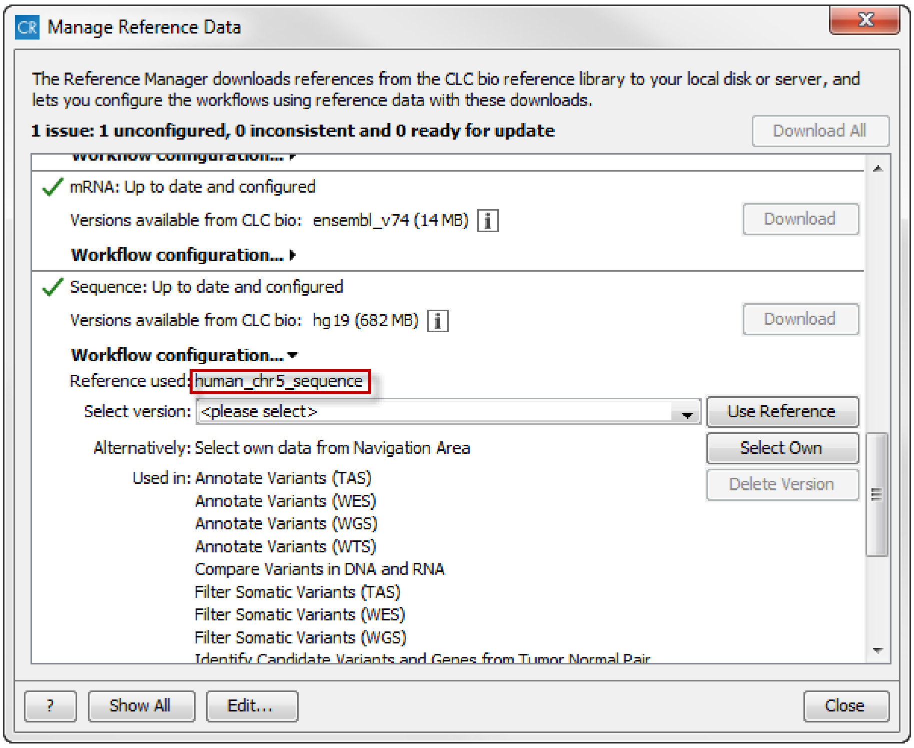
If you choose this option we can not guarantee the correctness of the reference file, or the functionality of the workflow.

Are you sure you want to use your own file as reference for this workflow?

Figure 3.17: *Click on the button labeled "OK" to use your own reference data to replace the previously specified reference data.*

Right under the bold face text "Workflow configuration" you can see which reference that is used. This is shown in figure 3.18.

If you would like to go back to the original reference data file, you can repeat the procedure



Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

1 issue: 1 unconfigured, 0 inconsistent and 0 ready for update

✓ mRNA: Up to date and configured

Versions available from CLC bio: ensembl_v74 (14 MB)

Workflow configuration...

✓ Sequence: Up to date and configured

Versions available from CLC bio: hg19 (682 MB)

Workflow configuration...

Reference used: human_chr5_sequence

Select version: <please select>

Alternatively: Select own data from Navigation Area

Used in: Annotate Variants (TAS)
Annotate Variants (WES)
Annotate Variants (WGS)
Annotate Variants (WTS)
Compare Variants in DNA and RNA
Filter Somatic Variants (TAS)
Filter Somatic Variants (WES)
Filter Somatic Variants (WGS)
Identify Candidate Variants and Genes from Tumor Normal Pair

Figure 3.18: *You can check which reference file that is used.*

and instead of selecting your own reference data file, you must go to the CLC_References data location and select the original reference data file.

3.3.4 Retrieving reference data tracks

For most applications (except de novo sequencing), you will need reference data in the form of a reference genome sequence, annotations, known variants etc. There are three basic ways of obtaining reference data tracks:

1. Import tracks from files (learn more in section 6.2).

   Standard Import ( ) The standard import accepts common data formats like fasta, genbank etc. (learn more in section 6)

2. Use the special plugins that integrate with Biobase's Genome Trax (learn more at `http://www.clcbio.com/clc-plugin/biobase-genome-trax-download/`).

3.4 Sequence web info

CLC Cancer Research Workbench provides direct access to web-based search in various databases and on the Internet using your computer's default browser. You can look up a sequence in the databases of NCBI and UniProt, search for a sequence on the Internet using Google and search for Pubmed references at NCBI. This is useful for quickly obtaining updated and additional information about a sequence.

The procedure for searching is identical for all four search options (see also figure 3.19):

Open a sequence or a sequence list | Right-click the name of the sequence | Web Info (🌍) | select the desired search function

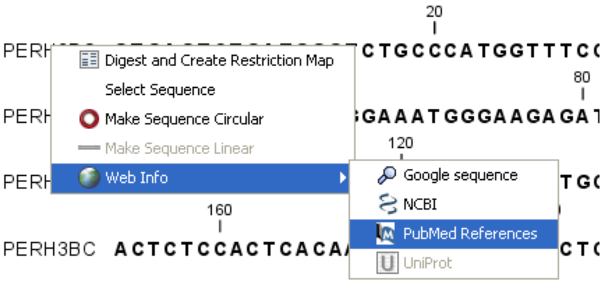


Figure 3.19: *Open webpages with information about this sequence.*

This will open your computer's default browser searching for the sequence that you selected.

3.4.1 Google sequence

The Google search function uses the accession number of the sequence which is used as search term on http://www.google.com. The resulting web page is equivalent to typing the accession number of the sequence into the search field on http://www.google.com.

3.4.2 NCBI

The NCBI search function searches in GenBank at NCBI (http://www.ncbi.nlm.nih.gov) using an identification number (when you view the sequence as text it is the "GI" number). Therefore, the sequence file must contain this number in order to look it up at NCBI. All sequences downloaded from NCBI have this number.

3.4.3 PubMed References

The PubMed references search option lets you look up Pubmed articles based on references contained in the sequence file (when you view the sequence as text it contains a number of "PUBMED" lines). Not all sequence have these PubMed references, but in this case you will se a dialog and the browser will not open.

3.4.4 UniProt

The UniProt search function searches in the UniProt database (http://www.ebi.uniprot.org) using the accession number. Furthermore, it checks whether the sequence was indeed downloaded from UniProt.

3.4.5 Additional annotation information

When sequences are downloaded from GenBank they often link to additional information on taxonomy, conserved domains etc. If such information is available for a sequence it is possible to access additional accurate online information. If the db_xref identifier line is found as part of the annotation information in the downloaded GenBank file, it is possible to easily look up additional information on the NCBI web-site.

To access this feature, simply right click an annotation and see which databases are available. For tracks, these links are also available in the track table.

Chapter 4

User preferences and settings

Contents

The first three sections in this chapter deal with the general preferences that can be set for *CLC Cancer Research Workbench* using the **Preferences** dialog. The next section explains how the settings in the **Side Panel** can be saved and applied to other views. Finally, you can learn how to import and export the preferences.

The **Preferences** dialog offers opportunities for changing the default settings for different features of the program.

The **Preferences** dialog is opened in one of the following ways and can be seen in figure 4.2:

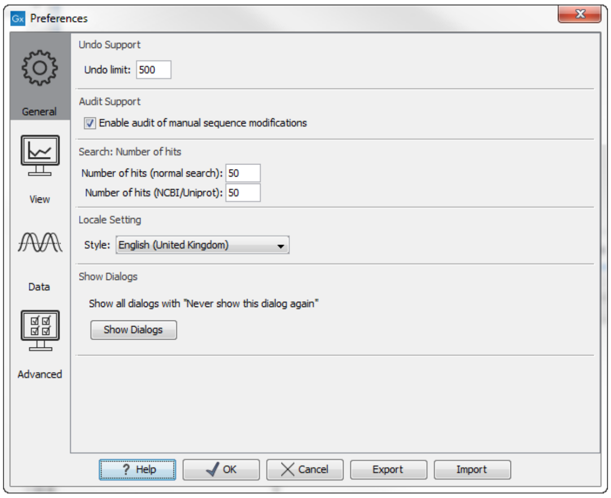
Edit | Preferences ( )

or **Ctrl + K (⌘ + ; on Mac)**

4.1 General preferences

The **General** preferences include:

- **Undo Limit.** As default the undo limit is set to 500. By writing a higher number in this field, more actions can be undone. Undo applies to all changes made on molecules, sequences, alignments or trees. See section 2.1.5 for more on this topic.



Undo Support

Undo limit: 500

Audit Support

☑ Enable audit of manual sequence modifications

Search: Number of hits

Number of hits (normal search): 50

Number of hits (NCBI/Uniprot): 50

Locale Setting

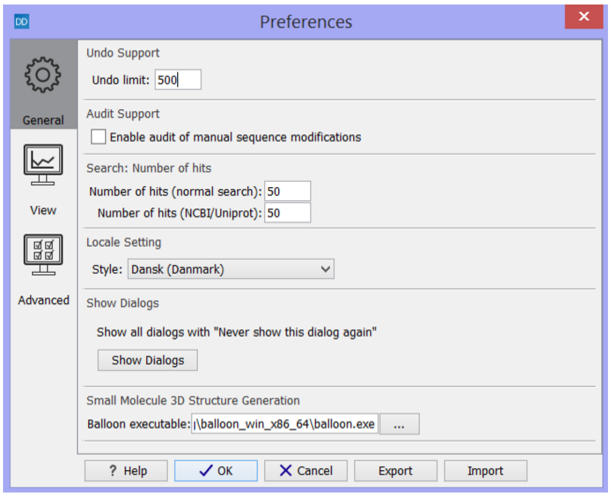
Style: English (United Kingdom)

Show Dialogs

Show all dialogs with "Never show this dialog again"

Show Dialogs

Figure 4.1: *Preferences include General preferences, View preferences, Data preferences, and Advanced settings.*



Undo Support

Undo limit: 500

Audit Support

□ Enable audit of manual sequence modifications

Search: Number of hits

Number of hits (normal search): 50

Number of hits (NCBI/Uniprot): 50

Locale Setting

Style: Dansk (Danmark)

Show Dialogs

Show all dialogs with "Never show this dialog again"

Show Dialogs

Small Molecule 3D Structure Generation

Balloon executable: |\balloon_win_x86_64\balloon.exe ...

Figure 4.2: *Preferences include General preferences, View preferences, Data preferences, and Advanced settings.*

- **Audit Support.** If this option is checked, all manual editing of sequences will be marked with an annotation on the sequence (see figure 4.3). Placing the mouse on the annotation will reveal additional details about the change made to the sequence (see figure 4.4). Note that no matter whether **Audit Support** is checked or not, all changes are also recorded in the **History** (🕓) (see section 7).

- **Number of hits.** The number of hits shown in *CLC Cancer Research Workbench*, when e.g. searching NCBI. (The sequences shown in the program are not downloaded, until they are opened or dragged/saved into the Navigation Area).

- **Locale Setting.** Specify which country you are located in. This determines how punctation is used in numbers all over the program.

- **Show Dialogs.** A lot of information dialogs have a checkbox: "Never show this dialog again". When you see a dialog and check this box in the dialog, the dialog will not be shown again. If you regret and wish to have the dialog displayed again, click the button in the General Preferences: **Show Dialogs**. Then all the dialogs will be shown again.

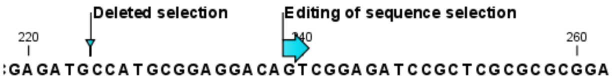
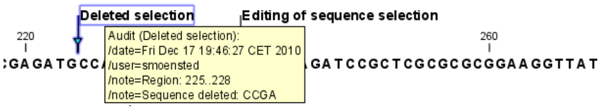

Deleted selection
Editing of sequence selection
220
240
260
:GAGATGCCATGCGGAGGACAGTCGGAGATCCGCTCGCGCGCGGA

Figure 4.3: *Annotations added when the sequence is edited.*



Audit (Deleted selection):
/date=Fri Dec 17 19:46:27 CET 2010
/user=smoensted
/note=Region: 225..228
/note=Sequence deleted: CCGA

Figure 4.4: *Details of the editing.*

4.2 Default view preferences

There are five groups of default **View** settings:

1. **Toolbar**
2. **Show Side Panel**
3. **New View**
4. **Sequence Representation**
5. **User Defined View Settings**

In general, these are default settings for the user interface.

The **Toolbar preferences** let you choose the size of the toolbar icons, and you can choose whether to display names below the icons.

The **Show Side Panel** setting allows you to choose whether to display the side panel.

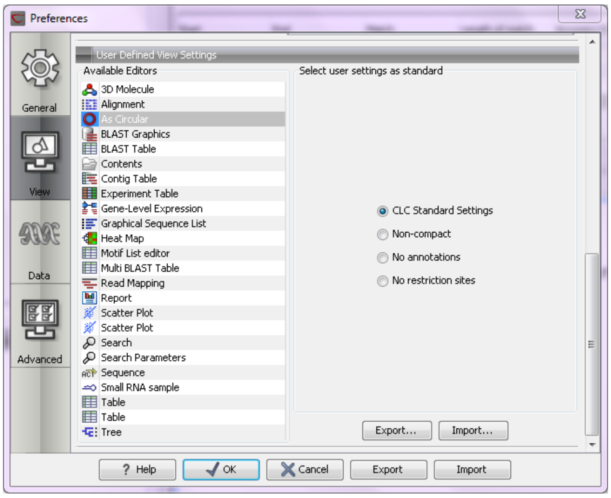
The **New view** setting allows you to choose whether the **View preferences** are to be shown automatically when opening a new view. If this option is not chosen, you can press (Ctrl + U (⌘ + U on Mac)) to see the preferences panels of an open view.

The **Sequence Representation** allows you to change the way the elements appear in the **Navigation Area**. The following text can be used to describe the element:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Latin name.
- Latin name (accession).
- Common name.
- Common name (accession).

The **User Defined View Settings** gives you an overview of the different **Side Panel** settings that are saved for each view. See section 4.5 for more about how to create and save style sheets.

If there are other settings beside **CLC Standard Settings**, you can use this overview to choose which of the settings should be used per default when you open a view (see an example in figure 4.5).



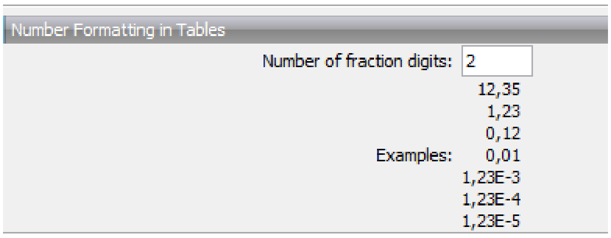
● CLC Standard Settings
○ Non-compact
○ No annotations
○ No restriction sites

Figure 4.5: *Selecting the default view setting.*

In this example, the **CLC Standard Settings** is chosen as default.

4.2.1 Number formatting in tables

In the preferences, you can specify how the numbers should be formatted in tables (see figure 4.6).



Number of fraction digits: 2

Examples:
12,35
1,23
0,12
0,01
1,23E-3
1,23E-4
1,23E-5

Figure 4.6: *Number formatting of tables.*

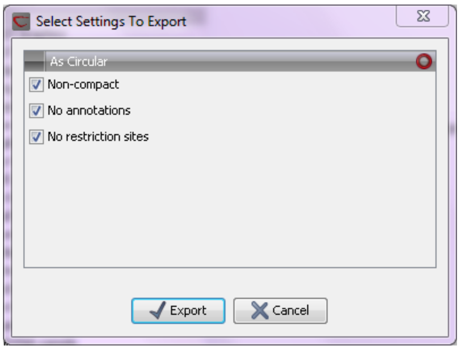
The examples below the text field are updated when you change the value so that you can see the effect. After you have changed the preference, you have to re-open your tables to see the effect.

4.2.2 Import and export Side Panel settings

If you have created a special set of settings in the **Side Panel** that you wish to share with other CLC users, you can export the settings in a file. The other user can then import the settings.

To export the **Side Panel** settings, first select the views that you wish to export settings for. Use Ctrl+click (⌘ + click on Mac) or Shift+click to select multiple views. Next click the **Export...**button. Note that there is also another export button at the very bottom of the dialog, but this will export the other settings of the **Preferences** dialog (see section 4.4).

A dialog will be shown (see figure 4.7) that allows you to select which of the settings you wish to export.



As Circular

☑ Non-compact

☑ No annotations

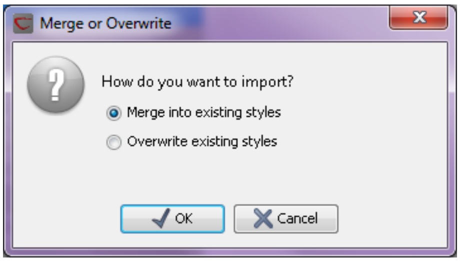
☑ No restriction sites

Figure 4.7: *Exporting all settings for circular views.*

When multiple views are selected for export, all the view settings for the views will be shown in the dialog. Click **Export** and you will now be able to define a save folder and name for the exported file. The settings are saved in a file with a .vsf extension (View Settings File).

To import a **Side Panel** settings file, make sure you are at the bottom of the **View** panel of the **Preferences dialog**, and click the **Import...** button. Note that there is also another import button at the very bottom of the dialog, but this will import the other settings of the **Preferences** dialog (see section 4.4).

The dialog asks if you wish to overwrite existing **Side Panel** settings, or if you wish to merge the imported settings into the existing ones (see figure 4.8).



How do you want to import?

● Merge into existing styles

○ Overwrite existing styles

Figure 4.8: *When you import settings, you are asked if you wish to overwrite existing settings or if you wish to merge the new settings into the old ones.*

Note! If you choose to overwrite the existing settings, you will loose all the **Side Panel** settings that you have previously saved.

To avoid confusion of the different import and export options, here is an overview:

- Import and export of **bioinformatics data** such as sequences, alignments etc. (described in section 6.1).
- **Graphics** export of the views which creates image files in various formats (described in section 6.6).
- Import and export of **Side Panel Settings** as described above.
- Import and export of all the **Preferences** except the Side Panel settings. This is described in the previous section.

- Linkers for importing 454 data (see section 6.3.1).
- Adapter sequences for trimming (see section 19.2.2).

4.3 Advanced preferences

The **Advanced** settings include the possibility to set up a proxy server. This is described in section 1.8.

4.3.1 Default data location

The default location is used when you e.g. import a file without selecting a folder or element in the **Navigation Area** first.

The default data location for CLC Workbenches is, by default, a folder called CLC_Data in a user's home area.

This can be changed to a different location for a particular user of the Workbench by going to

> **Edit | Preferences**

and then choosing the **Advanced** tab. This holds a section called **Default Data Location** and here you can choose a default from a drop down list of data locations you have already added.

Note! The default location cannot be removed. You have to select another location as default first.

If the data area you want as your default is not already available in your Workbench, you need to first add it as a new data location (see section 3.1.1).

4.4 Export/import of preferences

The user preferences of the *CLC Cancer Research Workbench* can be exported to other users of the program, allowing other users to display data with the same preferences as yours. You can also use the export/import preferences function to backup your preferences.

To export preferences, open the **Preferences** dialog (Ctrl + K (⌘ + ; on Mac)) and do the following:

> **Export | Select the relevant preferences | Export | Choose location for the exported file | Enter name of file | Save**

Note! The format of exported preferences is .cpf. This notation must be submitted to the name of the exported file in order for the exported file to work.

Before exporting, you are asked about which of the different settings you want to include in the exported file. One of the items in the list is "User Defined View Settings". If you export this, only the information about which of the settings is the default setting for each view is exported. If you wish to export the **Side Panel Settings** themselves, see section 4.2.2.

The process of importing preferences is similar to exporting:

> **Press Ctrl + K (⌘ + ; on Mac) to open Preferences | Import | Browse to and select the .cpf file | Import and apply preferences**

4.4.1 The different options for export and import

To avoid confusion of the different import and export options, you can find an overview here:

- Import and export of **bioinformatics data** such as molecules, sequences, alignments etc. (described in section 6.1).
- **Graphics** export of the views that create image files in various formats (described in section 6.6).
- Import and export of **Side Panel Settings** as described in the next section.
- Import and export of all the **Preferences** except the Side Panel settings. This is described above.

4.5 View settings for the Side Panel

The **Side Panel** is shown to the right of all views that are opened in *CLC Cancer Research Workbench* and is described in further detail in section 2.1.8.

When you have adjusted a view of e.g. a sequence, your settings in the **Side Panel** can be saved. When you open other sequences, which you want to display in a similar way, the saved settings can be applied. The options for saving and applying are available at the bottom of the **Side Panel** (see figure 4.9).

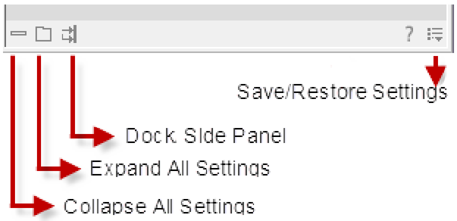


Figure 4.9: *At the bottom of the Side Panel you save the view settings*

4.5.1 Saving, removing and applying saved settings

To save and apply the saved settings, click (☰) seen in figure 4.9. This opens a menu where the following options are available (figure 4.10):

- **Save ... Settings.** (⚙) The settings can be saved in two different ways. When you select either way of saving settings a dialog will open (see figure 4.11) where you can enter a name for your settings.
  - **For ... View in General** (⚙) Will save the currently used settings with all elements of the same type as the one used for adjusting the settings. E.g. if you have selected to save settings "For Track View in General" the settings will be applied each time you open an element of the same type, which in this case means each time one of the saved tracks are opened from the **Navigation Area**. These "general" settings are user specific and will not be saved with or exported with the element.

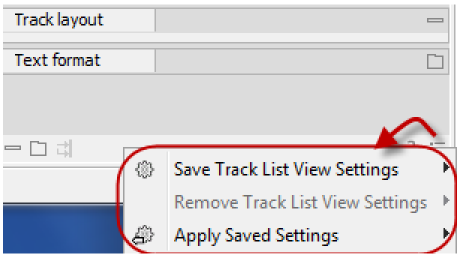

Track layout
Text format
Save Track List View Settings
Remove Track List View Settings
Apply Saved Settings

Figure 4.10: *When you have adjusted the side panel settings and would like to save these, this can be done with the "Save ... Settings" function, where "..." is the element you are working on - e.g. "Track List View", "Sequence View", "Table View", "Alignment View" etc. Saved settings can be deleted again with "Remove ... Settings" and can be applied to other elements with "Apply Saved Settings".*

 - **On This Only** ( ) Settings can be saved with the specific element that you are working on in the View area and will not affect any other elements (neither in the View Area or in the **Navigation Area**). E.g. for a track you would get the option to save settings "On This Track Only". The settings are saved with only this element (and will be exported with the element if you later select to export the element to another destination).

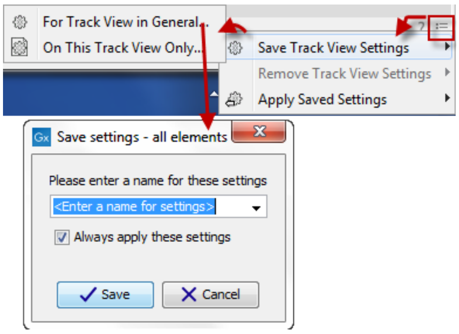

For Track View in General...
On This Track View Only...
Save Track View Settings
Remove Track View Settings
Apply Saved Settings
Save settings - all elements
Please enter a name for these settings
<Enter a name for settings>
☑ Always apply these settings
Save
Cancel

Figure 4.11: *The save settings dialog.Two options exist for saving settings. Click on the relevant option to open the dialog shown at the bottom of the figure.*

- **Remove ... Settings.** ( ) Gives you the option to remove settings specifically for the element that you are working on in the View Area, or on all elements of the same type. When you have selected the relevant option, the dialog shown in figure 4.12 opens and allows you to select which of the saved settings to remove.
  - **From ... View in General** ( ) Will remove the currently used settings on all elements of the same type as the one used for adjusting the settings. E.g. if you have selected to remove settings from all alignments using "From Alignment View in General", all alignments in your **Navigation Area** will be opened with the standard settings in stead.
  - **From This ... Only** ( ) When you select this option, the selected settings will only be removed from the particular element that you are working on in the View area and will not affect any other elements (neither in the View Area or in the **Navigation Area**). The settings for this particular element will be replaced with the CLC standard settings ( ).

- **Apply Saved Settings.** ( ) This is a submenu containing the settings that you have previously saved (figure 4.13). By clicking one of the settings, they will be applied to the current view. You will also see a number of pre-defined view settings in this submenu. They

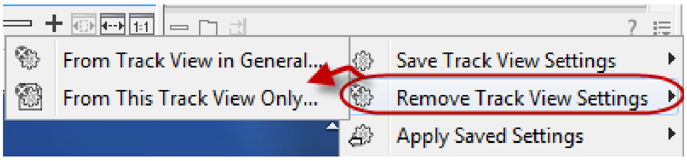

From Track View in General...
From This Track View Only...
Save Track View Settings
Remove Track View Settings
Apply Saved Settings

Figure 4.12: *The remove settings dialog for a track.*

are meant to be examples of how to use the **Side Panel** and provide quick ways of adjusting the view to common usages. At the bottom of the list of settings you will see **CLC Standard Settings** which represent the way the program was set up, when you first launched it. (⚙)

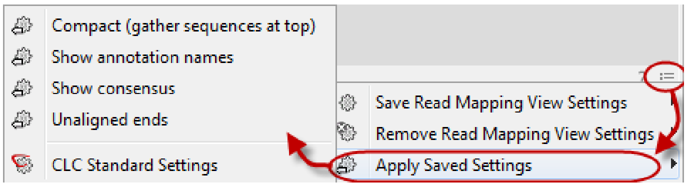

Compact (gather sequences at top)
Show annotation names
Show consensus
Unaligned ends
CLC Standard Settings
Save Read Mapping View Settings
Remove Read Mapping View Settings
Apply Saved Settings

Figure 4.13: *Applying saved settings.*

The settings are specific to the type of view. Hence, when you save settings of a circular view, they will not be available if you open the sequence in a linear view.

If you wish to export the settings that you have saved, this can be done in the **Preferences** dialog under the **View** tab (see section 4.2.2).

Chapter 5

Printing

Contents

CLC Cancer Research Workbench offers different choices of printing the result of your work.

This chapter deals with printing directly from *CLC Cancer Research Workbench*. Another option for using the graphical output of your work, is to export graphics (see chapter 6.6) in a graphic format, and then import it into a document or a presentation.

All the kinds of data that you can view in the **View Area** can be printed. The *CLC Cancer Research Workbench* uses a WYSIWYG principle: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks on the screen. When you print it, it will look exactly the same way on print as on the screen.

For some of the views, the layout will be slightly changed in order to be printer-friendly.

It is not possible to print elements directly from the **Navigation Area**. They must first be opened in a view in order to be printed. To print the contents of a view:

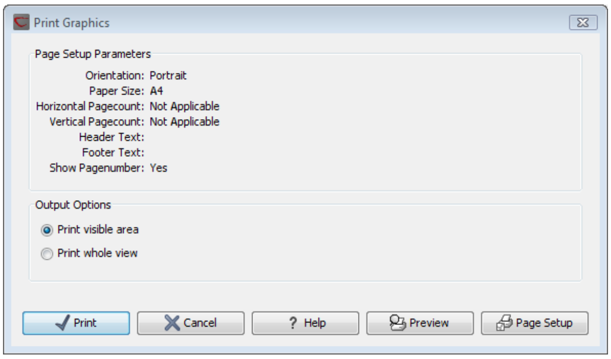
select relevant view | Print (🖨) in the toolbar

This will show a print dialog (see figure 5.1).

In this dialog, you can:

- Select which part of the view you want to print.
- Adjust **Page Setup**.
- See a print **Preview** window.

These three options are described in the three following sections.



Print Graphics

Page Setup Parameters

Orientation: Portrait
Paper Size: A4
Horizontal Pagecount: Not Applicable
Vertical Pagecount: Not Applicable
Header Text:
Footer Text:
Show Pagenumber: Yes

Output Options

● Print visible area
○ Print whole view

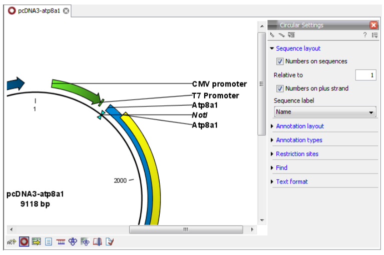
Figure 5.1: *The Print dialog.*

5.1 Selecting which part of the view to print

In the print dialog you can choose to:

- **Print visible area**, or
- **Print whole view**

These options are available for all views that can be zoomed in and out. In figure 5.2 is a view of a circular sequence which is zoomed in so that you can only see a part of it.



Sequence layout

☑ Numbers on sequences
Relative to: 1
☑ Numbers on plus strand
Sequence label: Name

Figure 5.2: *A circular sequence as it looks on the screen.*

When selecting **Print visible area**, your print will reflect the part of the sequence that is *visible* in the view. The result from printing the view from figure 5.2 and choosing **Print visible area** can be seen in figure 5.3.

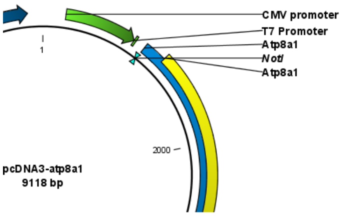


Figure 5.3: *A print of the sequence selecting Print visible area.*

On the other hand, if you select **Print whole view**, you will get a result that looks like figure 5.4. This means that you also print the part of the sequence which is not visible when you have zoomed in.

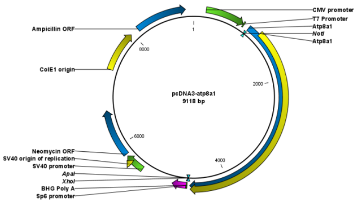
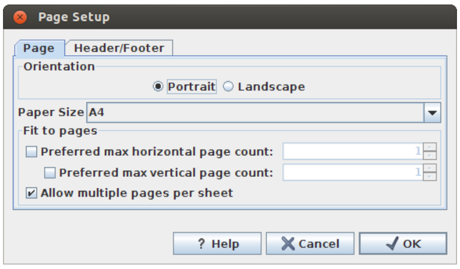


Figure 5.4: *A print of the sequence selecting Print whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

5.2 Page setup

No matter whether you have chosen to print the visible area or the whole view, you can adjust page setup of the print. An example of this can be seen in figure 5.5



Orientation

● Portrait ○ Landscape

Paper Size A4

Fit to pages

□ Preferred max horizontal page count: 1

□ Preferred max vertical page count: 1

☑ Allow multiple pages per sheet

Figure 5.5: *Page Setup.*

In this dialog you can adjust both the setup of the pages and specify a header and a footer by clicking the tab at the top of the dialog.

You can modify the layout of the page using the following options:

- **Orientation**.
  - **Portrait**. Will print with the paper oriented vertically.
  - **Landscape**. Will print with the paper oriented horizontally.
- **Paper size**. Adjust the size to match the paper in your printer.
- **Fit to pages**. Can be used to control how the graphics should be split across pages (see figure 5.6 for an example).
  - **Horizontal pages**. If you set the value to e.g. 2, the printed content will be broken up horizontally and split across 2 pages. This is useful for sequences that are not wrapped
  - **Vertical pages**. If you set the value to e.g. 2, the printed content will be broken up vertically and split across 2 pages.

Note! It is a good idea to consider adjusting view settings (e.g. **Wrap** for sequences), in the **Side Panel** before printing. As explained in the beginning of this chapter, the printed material will look like the view on the screen, and therefore these settings should also be considered when adjusting **Page Setup**.

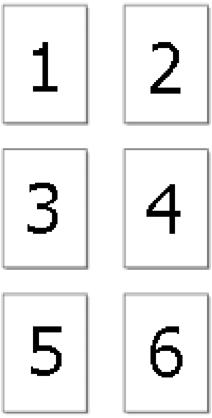


Figure 5.6: *An example where Fit to pages horizontally is set to 2, and Fit to pages vertically is set to 3.*

5.2.1 Header and footer

Click the **Header/Footer** tab to edit the header and footer text. By clicking in the text field for either **Custom header text** or **Custom footer text** you can access the auto formats for header/footer text in **Insert a caret position**. Click either **Date**, **View name**, or **User name** to include the auto format in the header/footer text.

Click **OK** when you have adjusted the **Page Setup**. The settings are saved so that you do not have to adjust them again next time you print. You can also change the **Page Setup** from the **File** menu.

5.3 Print preview

The preview is shown in figure 5.7.

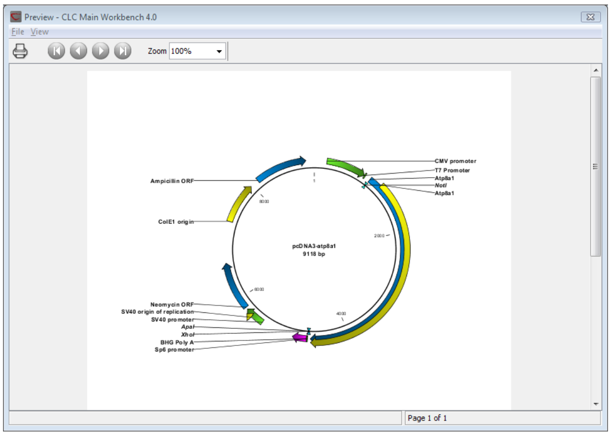


Figure 5.7: *Print preview.*

The **Print preview** window lets you see the layout of the pages that are printed. Use the arrows in the toolbar to navigate between the pages. Click Print (🖶) to show the print dialog, which lets you choose e.g. which pages to print.

The **Print preview** window is for preview only - the layout of the pages must be adjusted in the **Page setup**.

Chapter 6

Import/export of data and graphics

Contents

CLC Cancer Research Workbench handles a large number of different data formats. In order to work with data in the Workbench, it has to be imported ( ). Data types that are not recognized by the Workbench are imported as "external files" which means that when you open these, they will open in the default application for that file type on your computer (e.g. Word documents will open in Word).

This chapter first deals with importing and exporting data in bioinformatic data formats and as external files. Next comes an explanation of how to export graph data points to a file, and how to export graphics.

For **import of NGS data**, please see section 6.3.

6.1 Standard import

CLC Cancer Research Workbench has support for a wide range of bioinformatic data such as molecules, sequences, alignments etc. See a full list of the data formats in section E.1.

These data can be imported through the Import dialog, using drag/drop or copy/paste as explained below.

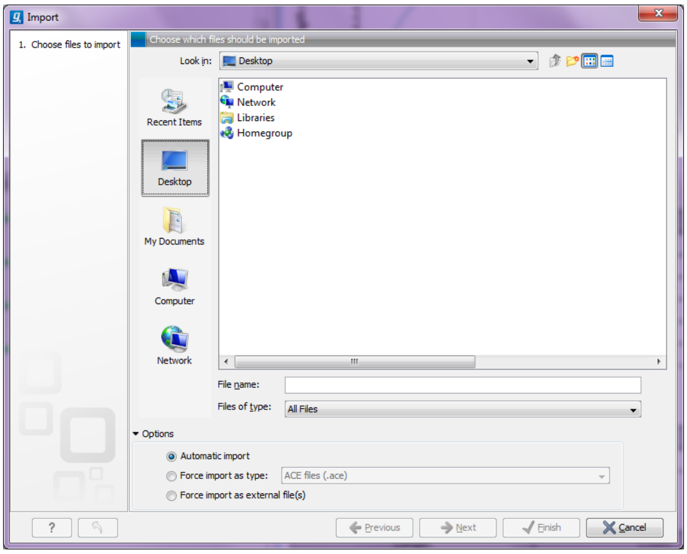
For **import of NGS data**, please see section 6.3 For import of tracks, please see section 6.2.

6.1.1 Import using the import dialog

To start the import using the import dialog:

> **click Import ( ) in the Toolbar | Standard Import**

This will show a dialog similar to figure 6.1. You can change which kind of file types that should be shown by selecting a file format in the **Files of type** box.



Look in: Desktop
File name:
Files of type: All Files
● Automatic import
○ Force import as type: ACE files (.ace)
○ Force import as external file(s)

Figure 6.1: *The import dialog.*

Next, select one or more files or folders to import and click **Next**.

This allows you to select a place for saving the result files.

If you import one or more folders, the contents of the folder is automatically imported and placed in that folder in the **Navigation Area**. If the folder contains subfolders, the whole folder structure is imported.

In the import dialog (figure 6.1), there are three import options:

Automatic import This will import the file and *CLC Cancer Research Workbench* will try to determine the format of the file. The format is determined based on the file extension (e.g. SwissProt files have .swp at the end of the file name) in combination with a detection of elements in the file that are specific to the individual file formats. If the file type is not recognized, it will be imported as an external file. In most cases, automatic import will yield a successful result, but if the import goes wrong, the next option can be helpful:

Force import as type This option should be used if *CLC Cancer Research Workbench* cannot successfully determine the file format. By forcing the import as a specific type, the automatic determination of the file format is bypassed, and the file is imported as the type specified.

Force import as external file This option should be used if a file is imported as a bioinformatics file when it should just have been external file. It could be an ordinary text file which is imported as a sequence.

6.1.2 Import using drag and drop

It is also possible to drag a file from e.g. the desktop into the **Navigation Area** of *CLC Cancer Research Workbench*. This is equivalent to importing the file using the **Automatic import** option described above. If the file type is not recognized, it will be imported as an external file.

6.1.3 Import using copy/paste of text

If you have e.g. a text file or a browser displaying a sequence in one of the formats that can be imported by *CLC Cancer Research Workbench*, there is a very easy way to get this sequence into the **Navigation Area**:

> **Copy the text from the text file or browser | Select a folder in the Navigation Area | Paste (🗋)**

This will create a new sequence based on the text copied. This operation is equivalent to saving the text in a text file and importing it into the *CLC Cancer Research Workbench*.

If the sequence is not formatted, i.e. if you just have a text like this: "ATGACGAATAGGAGTTCTAGCTA" you can also paste this into the **Navigation Area**.

Note! Make sure you copy all the relevant text - otherwise *CLC Cancer Research Workbench* might not be able to interpret the text.

6.1.4 External files

In order to help you organize your research projects, *CLC Cancer Research Workbench* lets you import all kinds of files. E.g. if you have Word, Excel or pdf-files related to your project, you can import them into the **Navigation Area** of *CLC Cancer Research Workbench*. Importing an external file creates a copy of the file which is stored at the location you have chosen for import. The file can now be opened by double-clicking the file in the **Navigation Area**. The file is opened using the default application for this file type (e.g. Microsoft Word for .doc-files and Adobe Reader for .pdf).

External files are imported and exported in the same way as bioinformatics files (see section 6.1). Bioinformatics files not recognized by *CLC Cancer Research Workbench* are also treated as external files.

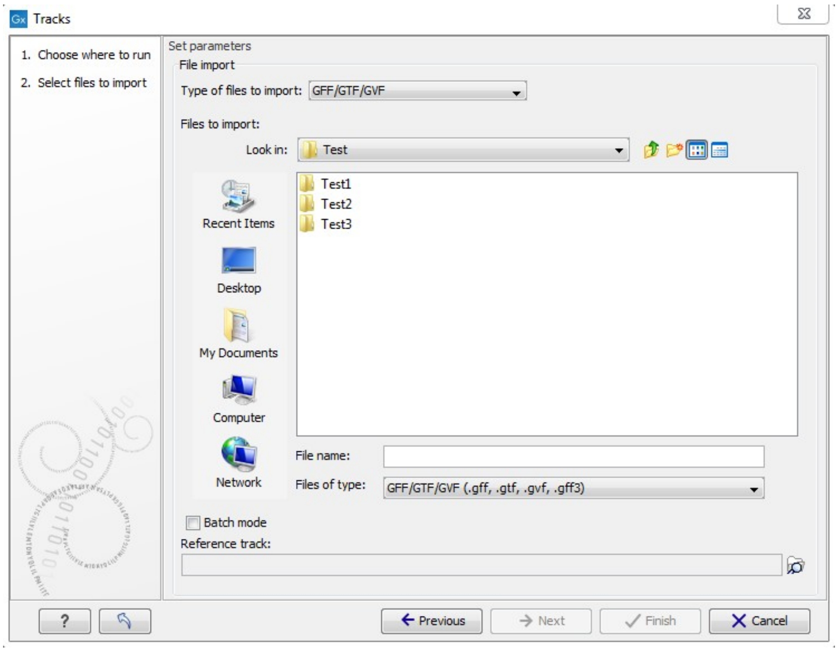
6.2 Import tracks

Tracks (see chapter 17) are imported in a special way, because extra information is needed in order to interpret the files correctly.

Tracks are imported using:

> **click Import ( ) in the Toolbar | Tracks**

This will open a dialog as shown in figure 6.2.



1. Choose where to run
2. Select files to import

Set parameters

File import

Type of files to import: GFF/GTF/GVF

Files to import:

Look in: Test

Files of type: GFF/GTF/GVF (.gff, .gtf, .gvf, .gff3)

□ Batch mode

Reference track:

Figure 6.2: *Select files to import.*

At the top, you select the file type to import. Below, select the files to import. If import is performed with the batch option selected, then each file is processed independently and separate tracks are produced for each file. If the batch option is not selected, then variants for all files will be added to the same track (or tracks in the case VCF files including genotype information).

The formats currently accepted are:

FASTA This is the standard fasta importer that will produce a sequence track rather than a standard fasta sequence.

GFF/GTF/GVF Annotations in gff/gtf/gvf formats. This is explained in detail in the user manual for the GFF annotation plugin:

http://www.clcbio.com/clc-plugin/annotate-sequence-with-gff-file/. This can be particularly useful when working with transcript annotations downloaded from from Ensembl available in gvf format: http://www.ensembl.org/info/data/ftp/index.html.

VCF This is the file format used for variants by the 1000 Genomes Project and it has become a standard format. Read how to access data at http://www.1000genomes.org/data#DataAccess. When importing a single VCF file, you will get a track for each sample contained in the VCF file. In cases where more than one sample is contained in a VCF file, you can choose to import the files together or individually by using the batch mode found in the lower left side of the wizard shown in figure 6.2. The difference between the two import modes is that the batch mode will import the samples individually in separate track files, whereas the non-batch mode will keep variants for one sample in one track, thus merging samples from the different input files (in cases where the same sample is contained in different input files). If you import more than one VCF file that each contain more than one sample, the non-batch mode will generate one track file for each unique sample. The batch mode will generate a track file for each of the original VCF files with the entire content, as if importing each of the VCF files one by one. E.g. VCF file 1 contains sample 1 and sample 2, and VCF file 2 contains sample 2 and sample 3. When VCF file 1 and VCF file 2 are imported in non-batch mode, you will get three individual track files; one for each of the three samples 1, 2, and 3. If VCF file 1 and VCF file 2 were instead imported using the batch function, the result of the import would be four track files: a track from sample 1 from file 1, a track from sample 2 from file 1, a track from sample 2 from file 2, and a track from sample 3 from file 2.

Complete Genomics master var file This is the file format used by Complete Genomics for all kinds of variant data and can be used to analyze and visualize the variant calls made by Complete Genomics. Please note that you can import evidence files with the read alignments into the *CLC Genomics Workbench* as well (refer to the Complete Genomics import section of the Workbench user manual).

BED Simple format for annotations. Read more at http://genome.ucsc.edu/FAQ/FAQformat.html#format1. This format is typically used for very simple annotations, for example target regions for sequence capture methods.

Wiggle The Wiggle format as defined by UCSC (http://genome.ucsc.edu/goldenPath/help/wiggle.html), is used to hold continuous data like conservation scores, GC content etc. When imported into the *CLC Cancer Research Workbench*, a graph track is created. An example of a popular Wiggle file is the conservation scores from UCSC which can be download for human from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/.

UCSC variant database table dump Table dumps of **variant** annotations from the UCSC can be imported using this option. Mainly files ending with .txt.gz on this list can be used: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/. Please note that importer is for variant data and is not a general importer for all annotation types. This is mainly intended to allow you to import the popular *Common SNPs* variant set from UCSC. The file can be downloaded from the UCSC web site here: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/snp138Common.txt.gz. Other sets of variant annotation can also be downloaded in this format using the UCSC Table Browser.

COSMIC variation database This lets you import the COSMIC database, which is a well-known publicly available primary database on somatic mutations in human cancer. The database is already available in the Data Management tab, however to use a different version than that provided, you must import this separately. The file can be downloaded from the UCSC web site here: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/download, and then go to 'Complete COSMIC data' and click on the file link. Users must first register to download the database. Import the file as a track, and then go to Data Management, scroll down to COSMIC and click on **Workflow Configuration**. Click on 'Select own' and select the track that has just been imported. Through Import->Tracks we support the following COSMIC databases in tsv format that can be manually downloaded from the COSMIC ftp site:

- Complete COSMIC data (file: CosmicCompleteExport.tsv)
- Complete mutation data (file: CosmicMutantExport.tsv)
- All Mutation in census genes (file: CosmicMutantExportCensus.tsv)

Please see chapter E.1.7 for more information on how different formats (e.g. VCF and GVF) are interpreted during import in CLC format.

For all of the above, zip files are also supported.

Please note that for human data, there is a difference between the UCSC genome build and Ensembl/NCBI for the mitochondrial genome. This means that for the mitochondrial genome, data from UCSC should not be mixed with data from other sources (see http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19).

Most of the data above is annotation data and if the file includes information about allele variants (like VCF, Complete Genomics and GVF), it will be combined into one **variant** track that can be used for finding known variants in your experimental data. When the data cannot be recognized as variant data, one track is created for each annotation type.

Genome / gene annotation tracks can be automatically imported from relevant databases as described in the http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Selecting_data_types_download.html).

For all types of files except fasta, you need to select a reference track as well. This is because most the annotation files do not contain enough information about chromosome names and lengths which are necessary to create the appropriate data structures.

6.3 Import high-throughput sequencing data

The *CLC Cancer Research Workbench* has dedicated tools for importing data from the following High-throughput sequencing systems.

- The 454 FLX System from Roche

- Illumina's Genome Analyzer, HiSeq and MiSeq
- SOLiD system from Applied Biosystems (read mapping is performed in color space, see section 20.5)
- Ion Torrent from Life Technologies
- Complete Genomics (only processed data - master var and evidence files)

The reason for having dedicated tools for this is to standardize the data so that most downstream analyses and visualization of the data works seamlessly with all sequencing platforms. In addition to these formats, mapped data in SAM/BAM format can also be imported.

This section will describe the various importers in detail.

Clicking on the **Import** () button in the top toolbar will bring up a list of the supported data types as shown in figure 6.3.

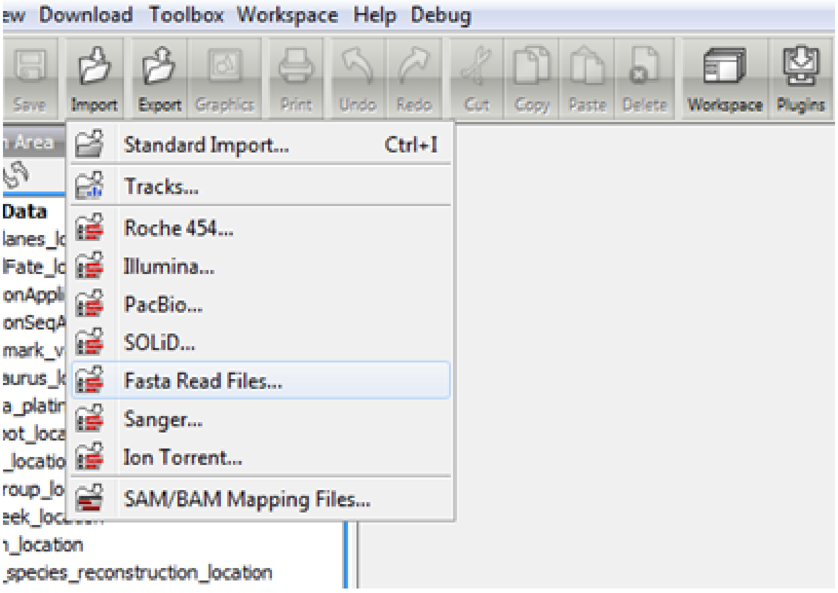


Figure 6.3: *Choosing what kind of data you wish to import.*

Select the appropriate format and then fill in the information as explained in the following sections.

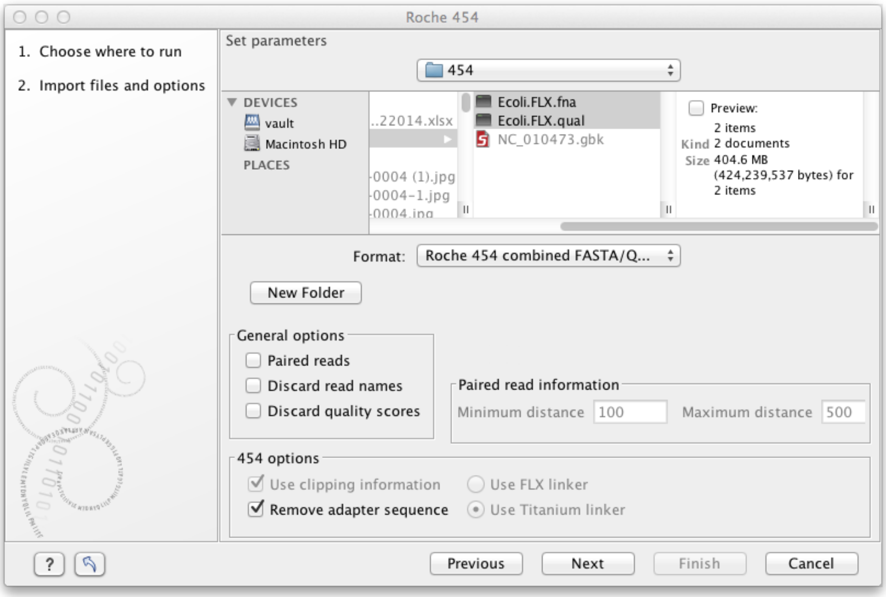
Please note that alignments of *Complete Genomics* data can be imported using the SAM/BAM importer, see section 6.3.7 below.

6.3.1 454 from Roche Applied Science

Choosing the Roche 454 import will open the dialog shown in figure 6.4.

We support import of two kinds of data from 454 GS FLX systems:

- Flowgram files (`.sff`) which contain both sequence data and quality scores amongst others. However, the flowgram information is currently not used by *CLC Cancer Research Workbench*. There is an extra option to make use of clipping information (this will remove parts of the sequence as specified in the .sff file).



Roche 454

1. Choose where to run
2. Import files and options

Set parameters

454

DEVICES
vault
Macintosh HD
PLACES

..22014.xlsx
0004 (1).jpg
0004-1.jpg
0004.jpg

Ecoli.FLX.fna
Ecoli.FLX.qual
NC_010473.gbk

□ Preview:
2 items
Kind 2 documents
Size 404.6 MB (424,239,537 bytes) for 2 items

Format: Roche 454 combined FASTA/Q...

New Folder

General options
□ Paired reads
□ Discard read names
□ Discard quality scores

Paired read information
Minimum distance 100
Maximum distance 500

454 options
☑ Use clipping information
☑ Remove adapter sequence
○ Use FLX linker
● Use Titanium linker

Figure 6.4: *Importing data from Roche 454.*

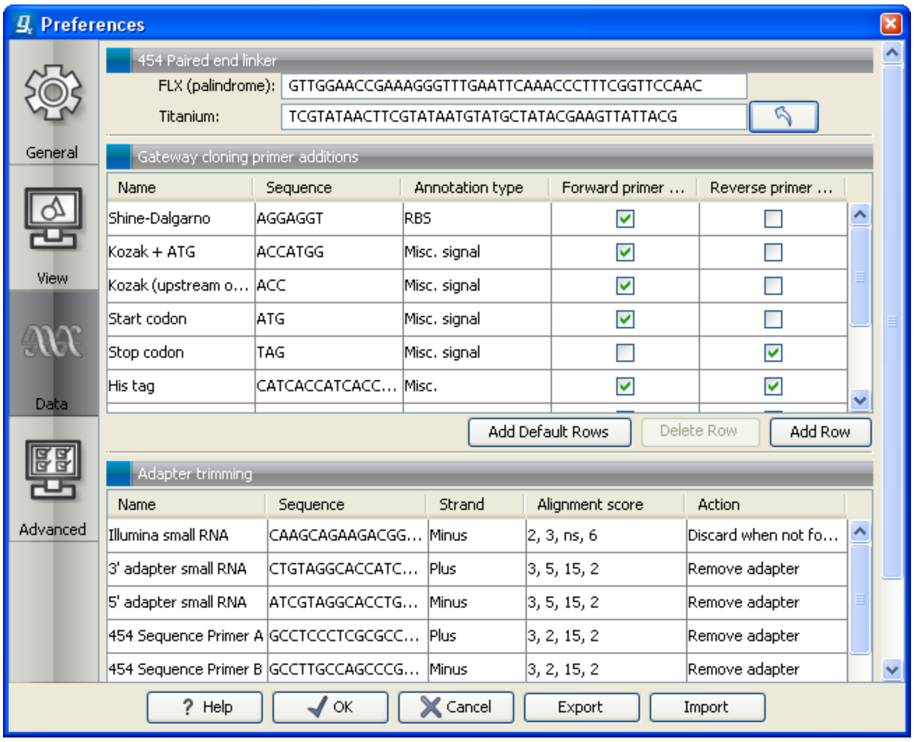
- Fasta/qual files:
  - 454 FASTA files (`.fna`) which contain the sequence data.
  - Quality files (`.qual`) which contain the quality scores.

For all formats, compressed data in gzip format is also supported (.gz).

The **General options** to the left are:

- **Paired reads**. The paired protocol for 454 entails that the forward and reverse reads are separated by a linker sequence. During import of paired data, the linker sequence is removed and the forward and reverse reads are separated and put into the same sequence list (their status as forward and reverse reads is preserved). You can change the linker sequence in the **Preferences** (in the **Edit** menu) under **Data**. Since the linker for the FLX and Titanium versions are different, you can choose the appropriate protocol during import, and in the preferences you can supply a linker for both platforms (see figure 6.5. Note that since the FLX linker is palindromic, it will only be searched on the plus strand, whereas the Titanium linker will be found on both strands. Some of the sequences may not have the linker in the middle of the sequence, and in that case the partial linker sequence is still removed, and the single read is put into a separate sequence list. Thus when you import 454 paired data, you may end up with two sequence lists: one for paired reads and one for single reads. Note that for de novo assembly projects, only the paired list should be used since the single reads list may contain reads where there is still a linker sequence present but only partially due to sequencing errors. Read more about handling paired data in section 6.3.8.
- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard this option to save disk space.

- **Discard quality scores**. Quality scores are visualized in the mapping view and they are used for SNP detection. If this is not relevant for your work, you can choose to **Discard quality scores**. One of the benefits from discarding quality scores is that you will gain a lot in terms of reduced disk space usage and memory consumption. If you have selected the fna/qual option and choose to discard quality scores, you do not need to select a .qual file.



454 Paired end linker

FLX (palindrome): GTTGGAACCGAAAGGGTTTGAATTCAAACCCTTTCGGTTCCAAC

Titanium: TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACG

Gateway cloning primer additions

Name	Sequence	Annotation type	Forward primer ...	Reverse primer ...
Shine-Dalgarno	AGGAGGT	RBS	☑	□
Kozak + ATG	ACCATGG	Misc. signal	☑	□
Kozak (upstream o...	ACC	Misc. signal	☑	□
Start codon	ATG	Misc. signal	☑	□
Stop codon	TAG	Misc. signal	□	☑
His tag	CATCACCATCACC...	Misc.	☑	☑

Adapter trimming

Name	Sequence	Strand	Alignment score	Action
Illumina small RNA	CAAGCAGAAGACGG...	Minus	2, 3, ns, 6	Discard when not fo...
3' adapter small RNA	CTGTAGGCACCATC...	Plus	3, 5, 15, 2	Remove adapter
5' adapter small RNA	ATCGTAGGCACCTG...	Minus	3, 5, 15, 2	Remove adapter
454 Sequence Primer A	GCCTCCCTCGCGCC...	Plus	3, 2, 15, 2	Remove adapter
454 Sequence Primer B	GCCTTGCCAGCCCG...	Minus	3, 2, 15, 2	Remove adapter

Figure 6.5: *Specifying linkers for 454 import.*

Note! During import, partial adapter sequences are removed (`TCAG` and `ATGC`), and if the full sequencing adapters `GCCTTGCCAGCCCGCTCAG`, `GCCTCCCTCGCGCCATCAG` or their reverse complements are found, they are also removed (including tailing Ns). If you do not wish to remove the adapter sequences (e.g. if they have already been removed by other software), please uncheck the **Remove adapter sequence** option.

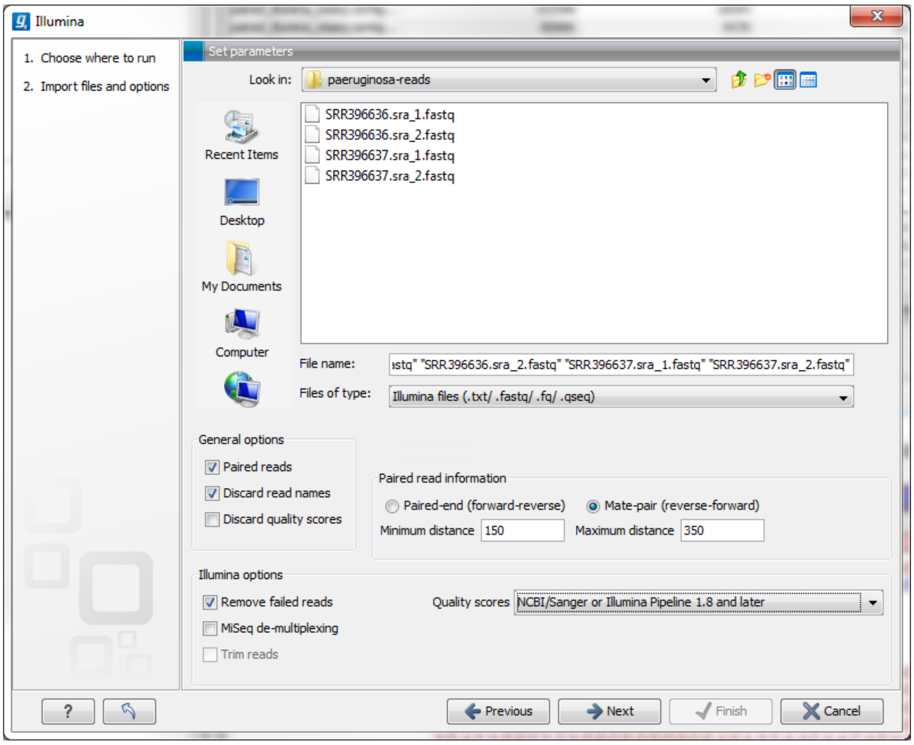
Click **Next** to adjust how to handle the results (see section 8.2). We recommend choosing **Save** in order to save the results directly to a folder, since you probably want to save anyway before proceeding with your analysis. There is an option to put the import data into a separate folder. This can be handy for better organizing subsequent analysis results and for batching (see section 8.1).

6.3.2 Illumina

The *CLC Cancer Research Workbench* supports data from Illumina's Genome Analyzer, HiSeq 2000 and the MiSeq systems. Choosing the Illumina import will open the dialog shown in figure 6.6.

The file formats accepted are:

- Fastq



Illumina

1. Choose where to run
2. Import files and options

Set parameters

Look in: paeruginosa-reads

SRR396636.sra_1.fastq
SRR396636.sra_2.fastq
SRR396637.sra_1.fastq
SRR396637.sra_2.fastq

File name: ıstq" "SRR396636.sra_2.fastq" "SRR396637.sra_1.fastq" "SRR396637.sra_2.fastq"

Files of type: Illumina files (.txt/ .fastq/ .fq/ .qseq)

General options

☑ Paired reads
☑ Discard read names
□ Discard quality scores

Paired read information

○ Paired-end (forward-reverse) ● Mate-pair (reverse-forward)

Minimum distance 150 Maximum distance 350

Illumina options

☑ Remove failed reads
□ MiSeq de-multiplexing
□ Trim reads

Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later

Figure 6.6: *Importing data from Illumina systems.*

- Scarf
- Qseq

Paired data in any of these formats can be imported.

Note that there is information inside qseq and fastq files specifying whether a read has passed a quality filter or not. If you check **Remove failed reads** these reads will be ignored during import. For qseq files there is a flag at the end of each read with values 0 (failed) or 1 (passed). In this example, the read is marked as failed and if Remove failed reads is checked, the read is removed.

```
M10   68   1   1   28680   29475   0   1   CATGGCCGTACAGGAAACACACATCATAGCATCACACGA   BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   0
```

For fastq files, part of the header information for the quality score has a flag where Y means failed and N means passed. In this example, the read has not passed the quality filter:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

Note! In the **Illumina pipeline 1.5-1.7**, the letter B in the quality score has a special meaning. 'B' is used as a trim clipping. This means that when selecting Illumina pipeline 1.5-1.7, the *reads are automatically trimmed* when a B is encountered in the input file. This will happen also if you choose to discard quality scores during import.

If you import paired data and one read in a pair is removed during import, the remaining mate will be saved in a separate sequence list with single reads.

For all formats, compressed data in gzip format is also supported (.gz).

The **General options** to the left are:

- **Paired reads**. For paired import, you can select whether the data is **Paired-end** or **Mate-pair**. For paired data, the Workbench expects the first reads of the pairs to be in one file and the second reads of the pairs to be in another. When importing one pair of files, the first file in a pair will is assumed to contain the first reads of the pair, and the second file is assumed to contain the second read in a pair. So, for example, if you had specified that the pairs were in forward-reverse orientation, then the first file would be assumed to contain the forward reads. The second file would be assumed to contain the reverse reads.

  When loading files containing paired data, the *CLC Cancer Research Workbench* sorts the files selected according to rules based on the file naming scheme:

  - For files coming off the CASAVA1.8 pipeline, we organize pairs according to their identifier and chunk number. Files named with `_R1_` are assumed to contain the first sequences of the pairs, and those with `_R2_` in the name are assumed to contain the second sequence of the pairs.
  - For other files, we sort them all alphanumerically, and then group them two by two. This means that files 1 and 2 in the list are loaded as pairs, files 3 and 4 in the list are seen as pairs, and so on.

  In the simplest case, the files are typically named as shown in figure 6.6. In this case, the data is paired end, and the file containing the forward reads is called `s_1_1_sequence.txt` and the file containing reverse reads is called `s_1_2_sequence.txt`. Other common filenames for paired data, like `_1_sequence.txt`, `_1_qseq.txt`, `_2_sequence.txt` or `_2_qseq.txt` will be sorted alphanumerically. In such cases, files containing the final `_1` should contain the first reads of a pair, and those containing the final `_2` should contain the second reads of a pair.

  For files from CASAVA1.8, files with base names like these: ID_R1_001, ID_R1_002, ID_R2_001, ID_R2_002 would be sorted in this order:

  1. ID_R1_001
  2. ID_R2_001
  3. ID_R1_002
  4. ID_R2_002

  The data in files ID_R1_001 and ID_R2_001 would be loaded as a pair, and ID_R1_002, ID_R2_002 would be loaded as a pair.

  Within each file, the first read of a pair will have a `1` somewhere in the information line. In most cases, this will be a `/1` at the end of the read name. In some cases though (e.g. CASAVA1.8), there will be a `1` elsewhere in the information line for each sequence. Similarly, the second read of a pair will have a `2` somewhere in the information line - either a `/2` at the end of the read name, or a `2` elsewhere in the information line.

  If you do not choose to discard your read names on import (see next parameter setting), you can quickly check that your paired data has imported in the pairs you expect by looking at the first few sequence names in your imported paired data object. The first two sequences should have the same name, except for a `1` or a `2` somewhere in the read name line.

  Paired-end and mate-pair data are handled the same way with regards to sorting on filenames. Their data structure is the same the same once imported into the Workbench. The only difference is that the expected orientation of the reads: reverse-forward in the case of mate pairs, and forward-reverse in the case of paired end data. Read more about handling paired data in section 6.3.8.

- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard read names to save disk space.
- **Discard quality scores**. Quality scores are visualized in the mapping view and they are used for SNP detection. If this is not relevant for your work, you can choose to **Discard quality scores**. One of the benefits from discarding quality scores is that you will gain a lot in terms of reduced disk space usage and memory consumption. Read more about the quality scores of Illumina below.
- **MiSeq de-multiplexing**. For MiSeq multiplexed data, one file includes all the reads containing barcodes/indices from the different samples (in case of paired data it will be two files). Using this option, the data can be divided into groups based on the barcode/index. This is typically the desired behavior, because subsequent analysis can then be executed in batch on all the samples and results can be compared at the end. This is not possible if all samples are in the same file after import. The reads are connected to a group using the last number in the read identifier.
- **Trim reads**. This option applies to Illumina Pipeline 1.5 to 1.7. In this pipeline, the value 2 (B) has special meaning and is used as a trim clipping. This means that when selecting Illumina Pipeline 1.5 and later, the reads are trimmed when a B is encountered in the input file if the **Trim reads** option is checked.

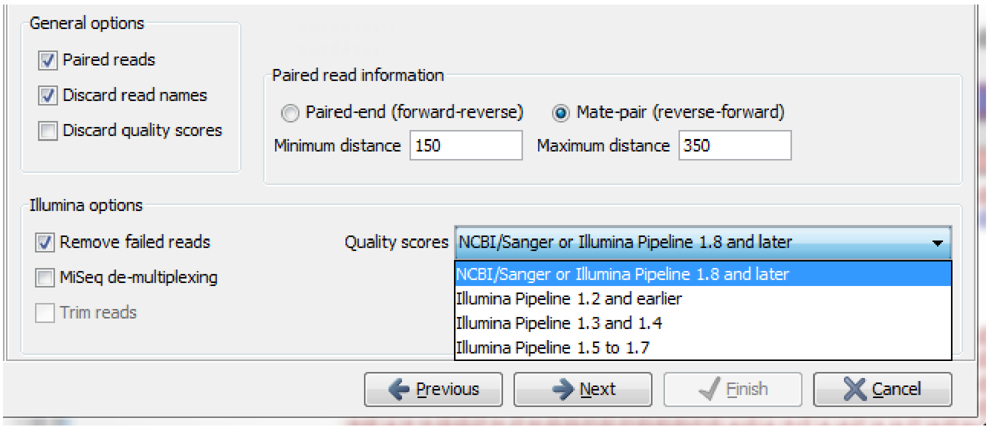
Click **Next** to adjust how to handle the results (see section 8.2). We recommend choosing **Save** in order to save the results directly to a folder, since you probably want to save anyway before proceeding with your analysis. There is an option to put the import data into a separate folder. This can be handy for better organizing subsequent analysis results and for batching (see section 8.1).

Quality scores in the Illumina platform

The quality scores in the FASTQ format come in different versions. You can read more about the FASTQ format at http://en.wikipedia.org/wiki/FASTQ_format. When you select to import Illumina data and click **Next** there is an option to use different quality score schemes at the bottom of the dialog (see figure 6.7).

There are four options:

- **NCBI/Sanger or Illumina 1.8 and later**. Using a Phred scale encoded using ASCII 33 to 93. This is the standard for fastq formats except for the early Illumina data formats (this changed with version 1.8 of the Illumina Pipeline).
- **Illumina Pipeline 1.2 and earlier**. Using a Solexa/Illumina scale (-5 to 40) using ASCII 59 to 104. The Workbench automatically converts these quality scores to the Phred scale on import in order to ensure a common scale for analyses across data sets from different platforms (see details on the conversion next to the sample below).
- **Illumina Pipeline 1.3 and 1.4**. Using a Phred scale using ASCII 64 to 104.



General options

☑ Paired reads

☑ Discard read names

□ Discard quality scores

Paired read information

○ Paired-end (forward-reverse) ● Mate-pair (reverse-forward)

Minimum distance 150 Maximum distance 350

Illumina options

☑ Remove failed reads

□ MiSeq de-multiplexing

□ Trim reads

Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later

NCBI/Sanger or Illumina Pipeline 1.8 and later
Illumina Pipeline 1.2 and earlier
Illumina Pipeline 1.3 and 1.4
Illumina Pipeline 1.5 to 1.7
Previous
Next
Finish
Cancel

Figure 6.7: *Selecting the quality score scheme.*

- **Illumina Pipeline 1.5 to 1.7**. Using a Phred scale using ASCII 64 to 104. Values 0 (@) and 1 (A) are not used anymore. Value 2 (B) has special meaning and is used as a trim clipping. This means that when selecting Illumina Pipeline 1.5 and later, the reads are trimmed when a B is encountered in the input file if the **Trim reads** option is checked.

Small samples of three kinds of files are shown below. The names of the reads have no influence on the quality score format:

NCBI/Sanger Phred scores:

```
@SRR001926.1 FC00002:7:1:111:750 length=36
TTTTTGTAAGGAGGGGGGTCATCAAAATTTGCAAAA
+SRR001926.1 FC00002:7:1:111:750 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIFIIII'IB<IH
@SRR001926.7 FC00002:7:1:110:453 length=36
TTATATGGAGGCTTTAAGAGTCATAGGTTGTTCCCC
+SRR001926.7 FC00002:7:1:110:453 length=36
IIIIIIIIIII:'III?=IIIIII+&III/3I8F/&
```

Illumina Pipeline 1.2 and earlier (note the question mark at the end of line 4 - this is one of the values that are unique to the old Illumina pipeline format):

```
@SLXA-EAS1_89:1:1:672:654/1
GCTACGGAATAAAACCAGGAACAACAGACCCAGCA
+SLXA-EAS1_89:1:1:672:654/1
cccccccccccccccccccc]c``cVcZccbSYb?
@SLXA-EAS1_89:1:1:657:649/1
GCAGAAAATGGGAGTGAAAATCTCCGATGAGCAGC
+SLXA-EAS1_89:1:1:657:649/1
ccccccccccbccbccb``cccbcccZcc`^bR^`
```

The formulas used for converting the special Solexa-scale quality scores to Phred-scale:

$Q_{phred} = -10 \log_{10} p$

$Q_{solexa} = -10 \log_{10} \frac{p}{1-p}$

A sample of the quality scores of the Illumina Pipeline 1.3 and 1.4:

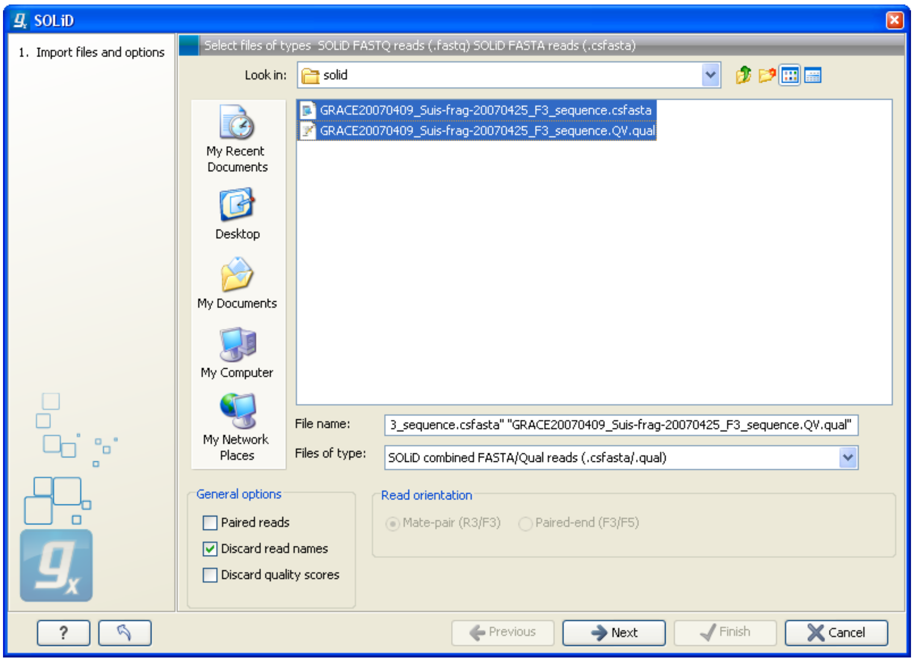
```
@HWI-E4_9_30WAF:1:1:8:178
GCCAGCGGCGCAAAATGNCGGCGGCGATGACCTTC
+HWI-E4_9_30WAF:1:1:8:178
babaaaa\ababaaaaREXabaaaaaaaaaaaaaa
@HWI-E4_9_30WAF:1:1:8:1689
GATGGAGATCTCGACCTNATAGGTGCCCTCATCGG
+HWI-E4_9_30WAF:1:1:8:1689
aab`]_aaaaaaaaaa[ER`abaaa\aaaaaaaa[
```

Note that it is not possible to see from that data itself that it is actually not Illumina Pipeline 1.2 and earlier, since they use the same range of ASCII values.

To learn more about ASCII values, please see http://en.wikipedia.org/wiki/Ascii#ASCII_printable_characters.

6.3.3 SOLiD from Life Technologies

Choosing the SOLiD import will open the dialog shown in figure 6.8.



SOLiD

1. Import files and options

Select files of types SOLiD FASTQ reads (.fastq) SOLiD FASTA reads (.csfasta)

Look in: solid

GRACE20070409_Suis-frag-20070425_F3_sequence.csfasta
GRACE20070409_Suis-frag-20070425_F3_sequence.QV.qual

File name: 3_sequence.csfasta" "GRACE20070409_Suis-frag-20070425_F3_sequence.QV.qual"

Files of type: SOLiD combined FASTA/Qual reads (.csfasta/.qual)

General options

- □ Paired reads
- ☑ Discard read names
- □ Discard quality scores

Read orientation

● Mate-pair (R3/F3) ○ Paired-end (F3/F5)

Figure 6.8: *Importing data from SOLiD from Applied Biosystems.*

There are two formats accepted: the XSQ format which is the native format of newer SOLiD systems, and the csfasta format which is the color space version of fasta format.

The XSQ format

An XSQ file can contain results from multiple libraries produced from the same sequencing run. These are identified by a barcode on each read, and when the XSQ file is produced, each read is placed into its appropriate library based on its barcode. The XSQ importer creates separate sequence lists for each library.

Sometimes when an XSQ file is produced, a barcode cannot be identified accurately enough to place the read into a specific library, or the read is for some other reason not assigned to a library. In this case, the read is placed into an "Unclassified" or "Unassigned" library.

In the case of paired reads, it sometimes happens that one read of a pair could not be read. When the XSQ file is imported in the *CLC Cancer Research Workbench*, the other read of such a pair is placed into a sequence list with " (single)" appended to the name, whereas all intact pairs are placed (alternating) into a sequence list with " (paired)" appended to the name. Thus, two sequence lists are produced for the library.

Hence, when importing data in XSQ format the number of imported files can vary. In the example shown here, where the XSQ file contain a library with the name "Main" (containing paired reads) and an "Unclassified" library (containing reads where e.g. the barcode could not be read), the imported data are segregated into the following sequence lists:

1. Main (single)
2. Main (paired)
3. Unclassified

An XSQ file sometimes contains reads in both base space and color space, and when that is the case, each read library in the XSQ file that contains reads in both formats will result in two files, with " (base space)" and " (color space)" appended to their names, respectively.

The csfasta format

If you want to import quality scores with csfasta files, qual files should also be provided. The reads in a csfasta file look like this:

```
>2_14_26_F3
T011213122200221123032111221021210131332222101
>2_14_192_F3
T110021221100310030120022032222111321022112223
>2_14_233_F3
T011001332311121212312022310203312201132111223
>2_14_294_F3
T213012132300000021323212232.03300033102330332
```

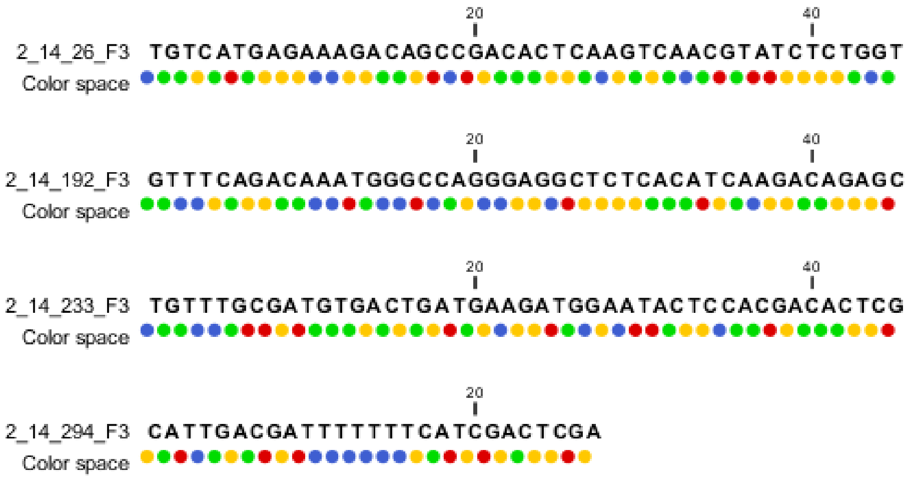
All reads start with a T which specifies the right phasing of the color sequence.

If a reads has a `.` as you can see in the last read in the example above, it means that the color calling was ambiguous (this would have been an `N` if we were in base space). In this case, the Workbench simply cuts off the rest of the read, since there is no way to know the right phase of the rest of the colors in the read. If the read starts with a dot, it is not imported. If all reads start with a dot, a warning dialog will be displayed. The handling of dots is identical for XSQ and csfasta files.

In the quality file, the equivalent value is `-1`, and this will also cause the read to be clipped.

When the example above is imported into the Workbench, it looks as shown in figure 6.9.

For more information about color space, please see section 20.5.



```
                        20                                      40
2_14_26_F3  TGTCATGAGAAAGACAGCCGACACTCAAGTCAACGTATCTCTGGT
Color space
                        20                                      40
2_14_192_F3 GTTTCAGACAAATGGGCCAGGGAGGCTCTCACATCAAGACAGAGC
Color space
                        20                                      40
2_14_233_F3 TGTTTGCGATGTGACTGATGAAGATGGAATACTCCACGACACTCG
Color space
                        20
2_14_294_F3 CATTGACGATTTTTTTCATCGACTCGA
Color space
```

Figure 6.9: *Importing data from SOLiD from Applied Biosystems. Note that the fourth read is cut off so that the color following the dot are not included*

In addition to the csfasta and XSQ formats used by SOLiD, you can also input data in fastq format. This is particularly useful for data downloaded from the Sequence Read Archive at NCBI (http://www.ncbi.nlm.nih.gov/Traces/sra/). An example of a SOLiD fastq file is shown here with both quality scores and the color space encoding:

```
@SRR016056.1.1 AMELIA_20071210_2_YorubanCGB_Frag_16bit_2_51_130.1 length=50
T3100031312131021102231222331121211302212120133221 3
+SRR016056.1.1 AMELIA_20071210_2_YorubanCGB_Frag_16bit_2_51_130.1 length=50
!*%;2'%%050%'0'3%%5*.%%%),%%%%&%%%%%%'%%%%%'%%3+%%%
@SRR016056.2.1 AMELIA_20071210_2_YorubanCGB_Frag_16bit_2_51_223.1 length=50
T2000220112002121101201033221112213321233122130222 2
+SRR016056.2.1 AMELIA_20071210_2_YorubanCGB_Frag_16bit_2_51_223.1 length=50
!%%)%'))'&'%(((&%/&)%+(%%%&%%%%%%%%%%%%%%%+%%%%%%+'
```

For all formats, compressed data in gzip format is also supported (.gz).

The **General options** to the left are:

- **Paired reads**. When you import paired data, two different protocols are supported:
  - **Mate-pair**. For mate-pair data, the reads should be in two files with _F3 and _R3 in front of the file extension. The orientation of the reads is expected to be forward-forward.
  - **Paired-end**. For paired-end data, the reads should be in two files with _F3 and _F5-P2 or _F5-BC. The orientation is expected to be forward-reverse.

  Read more about handling paired data in section 6.3.8. Please note that for XSQ files, the pairing protocol is defined in the file itself, which means that the choices of protocol will be ignored.

  An example of a complete list of the four files needed for a SOLiD mate-paired data set including quality scores:

```
dataset_F3.csfasta   dataset_F3.qual
dataset_R3.csfasta   dataset_R3.qual
```

or

```
dataset_F3.csfasta   dataset_F3_.QV.qual
dataset_R3.csfasta   dataset_R3_.QV.qual
```

- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard this option to save disk space.
- **Discard quality scores**. Quality scores are visualized in the mapping view and they are used for SNP detection. If this is not relevant for your work, you can choose to **Discard quality scores**. One of the benefits from discarding quality scores is that you will gain a lot in terms of reduced disk space usage and memory consumption. If you choose to discard quality scores, you do not need to select a .qual file when importing csfasta files.

Click **Next** to adjust how to handle the results (see section 8.2). We recommend choosing **Save** in order to save the results directly to a folder, since you probably want to save anyway before proceeding with your analysis. There is an option to put the import data into a separate folder. This can be handy for better organizing subsequent analysis results and for batching (see section 8.1).

6.3.4 Fasta read files

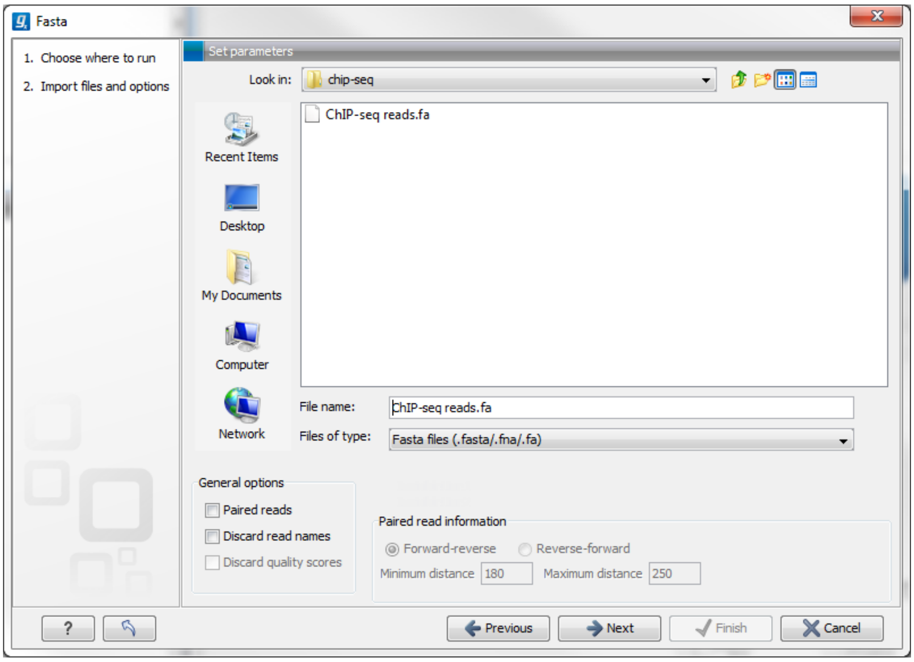
The **Fasta** importer is designed for high volumes of read data such as high-throughput sequencing data (NGS reads). When using this import option the read names can be included but the descriptions from the fasta files are ignored. For import of other fasta format data, such as reference sequences, please use the ()**Standard Import**, described in section 6 as this import format also includes the descriptions.

The dialog for importing data in fasta format is shown in figure 6.10.

Compressed data in gzip format is also supported (.gz).

The **General options** to the left are:

- **Paired reads**. For paired import, the Workbench expects the forward reads to be in one file and the reverse reads in another. The Workbench will sort the files before import and then assume that the first and second file belong together, and that the third and fourth file belong together etc. At the bottom of the dialog, you can choose whether the ordering of the files is **Forward-reverse** or **Reverse-forward**. As an example, you could have a data set with two files: `sample1_fwd` containing all the forward reads and `sample1_rev` containing all the reverse reads. In each file, the reads have to match each other, so that the first read in the `fwd` list should be paired with the first read in the `rev` list. Note that you can specify the insert sizes when running mapping and assembly. If you have data sets with different insert sizes, you should import each data set individually in order to be able to specify different insert sizes. Read more about handling paired data in section 6.3.8.



Look in: chip-seq

ChIP-seq reads.fa

File name: ChIP-seq reads.fa

Files of type: Fasta files (.fasta/.fna/.fa)

General options

□ Paired reads

□ Discard read names

□ Discard quality scores

Paired read information

● Forward-reverse ○ Reverse-forward

Minimum distance 180 Maximum distance 250

Figure 6.10: *Importing data in fasta format.*

- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard this option to save disk space.
- **Discard quality scores**. This option is not relevant for fasta import, since quality scores are not supported.

Click **Next** to adjust how to handle the results (see section 8.2). We recommend choosing **Save** in order to save the results directly to a folder, since you probably want to save anyway before proceeding with your analysis. There is an option to put the import data into a separate folder. This can be handy for better organizing subsequent analysis results and for batching (see section 8.1).

6.3.5 Sanger sequencing data

Although traditional sequencing data (with chromatogram traces like abi files) is usually imported using the standard **Import** (), see section 6, this option has also been included in the High-Throughput Sequencing Data import. It is designed to handle import of large amounts of sequences, and there are three differences from the standard import:

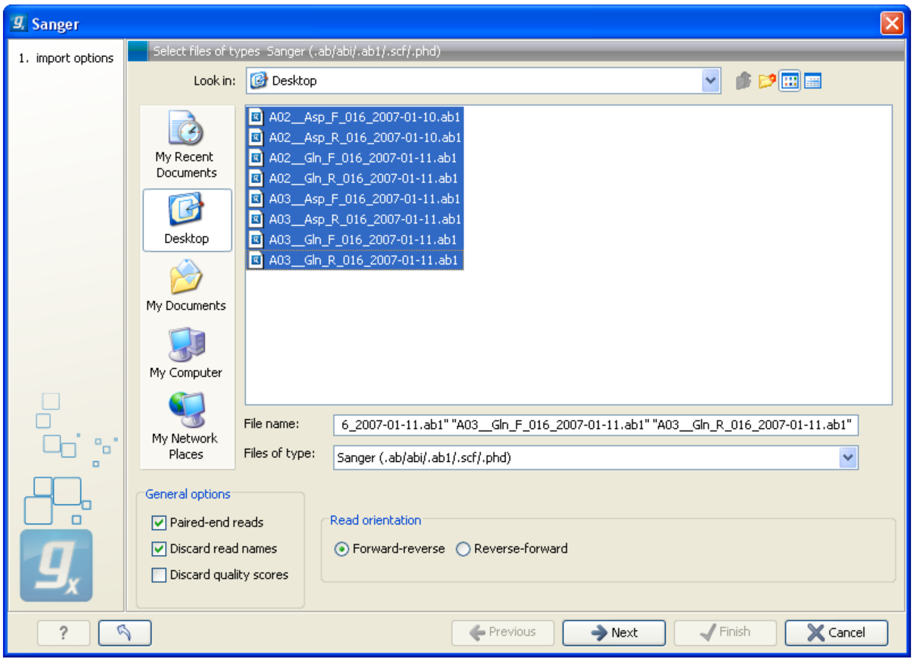
- All the sequences will be put in one sequence list (instead of single sequences).
- The chromatogram traces will be removed (quality scores remain). This is done to improve performance, since the trace data takes up a lot of disk space and significantly impacts speed and memory consumption for further analysis.
- Paired data is supported.

With the standard import, it is practically impossible to import up to thousands of trace files and use them in an assembly. With this special High-Throughput Sequencing import, there is no limit.

The import formats supported are the same: ab, abi, ab1, scf and phd.

For all formats, compressed data in gzip format is also supported (.gz).

The dialog for importing data Sanger sequencing data is shown in figure 6.11.



Select files of types Sanger (.ab/abi/.ab1/.scf/.phd)

Look in: Desktop

A02__Asp_F_016_2007-01-10.ab1
A02__Asp_R_016_2007-01-10.ab1
A02__Gln_F_016_2007-01-11.ab1
A02__Gln_R_016_2007-01-11.ab1
A03__Asp_F_016_2007-01-11.ab1
A03__Asp_R_016_2007-01-11.ab1
A03__Gln_F_016_2007-01-11.ab1
A03__Gln_R_016_2007-01-11.ab1

File name: 6_2007-01-11.ab1" "A03__Gln_F_016_2007-01-11.ab1" "A03__Gln_R_016_2007-01-11.ab1"

Files of type: Sanger (.ab/abi/.ab1/.scf/.phd)

General options

☑ Paired-end reads
☑ Discard read names
□ Discard quality scores

Read orientation

● Forward-reverse ○ Reverse-forward

Figure 6.11: *Importing data from Sanger sequencing.*

The **General options** to the left are:

- **Paired reads**. The Workbench will sort the files before import and then assume that the first and second file belong together, and that the third and fourth file belong together etc. At the bottom of the dialog, you can choose whether the ordering of the files is **Forward-reverse** or **Reverse-forward**. As an example, you could have a data set with two files: `sample1_fwd` for the forward read and `sample1_rev` for the reverse reads. Note that you can specify the insert sizes when running the mapping and the assembly. If you have data sets with different insert sizes, you should import each data set individually in order to be able to specify different insert sizes. Read more about handling paired data in section 6.3.8.

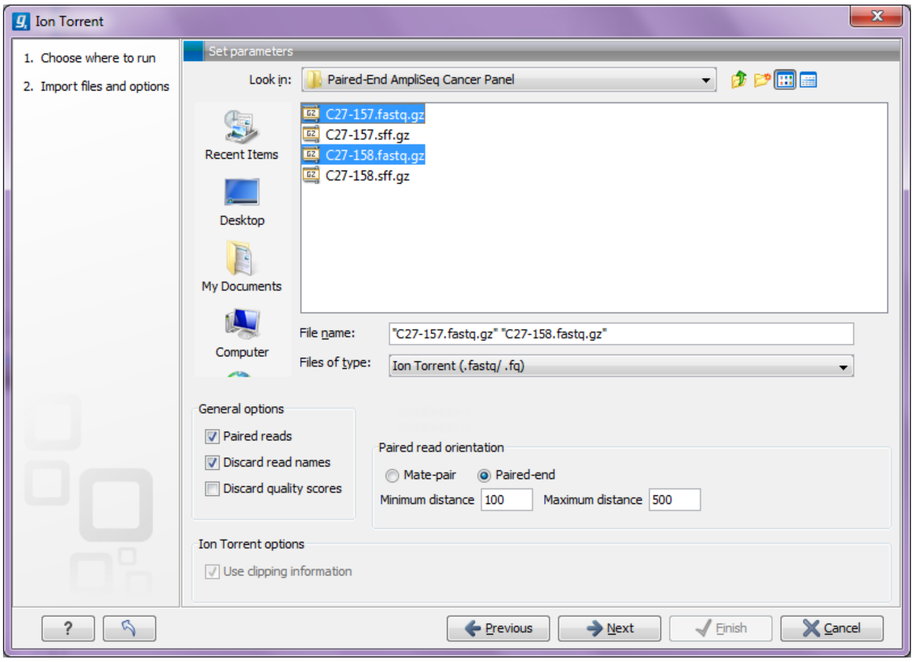
- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard this option to save disk space.

- **Discard quality scores**. Quality scores are visualized in the mapping view and they are used for SNP detection. If this is not relevant for your work, you can choose to **Discard quality scores**. One of the benefits from discarding quality scores is that you will gain a lot in terms of reduced disk space usage and memory consumption.

Click **Next** to adjust how to handle the results (see section 8.2). We recommend choosing **Save** in order to save the results directly to a folder, since you probably want to save anyway before proceeding with your analysis. There is an option to put the import data into a separate folder. This can be handy for better organizing subsequent analysis results and for batching (see section 8.1).

6.3.6 Ion Torrent PGM from Life Technologies

Choosing the Ion Torrent import will open the dialog shown in figure 6.12.



Look in: Paired-End AmpliSeq Cancer Panel

C27-157.fastq.gz
C27-157.sff.gz
C27-158.fastq.gz
C27-158.sff.gz

File name: "C27-157.fastq.gz" "C27-158.fastq.gz"

Files of type: Ion Torrent (.fastq/ .fq)

General options

☑ Paired reads
☑ Discard read names
□ Discard quality scores

Paired read orientation

○ Mate-pair ● Paired-end

Minimum distance 100 Maximum distance 500

Ion Torrent options

☑ Use clipping information

Figure 6.12: *Importing data from Ion Torrent.*

We support import of two kinds of data from the Ion Torrent system:

- SFF files (`.sff`)
- Fastq files (`.fastq`). Quality scores are expected to be in the NCBI/Sanger format (see section 6.3.2)

For all formats, compressed data in gzip format is also supported (.gz).

The **General options** to the left are:

- **Paired reads**. The *CLC Cancer Research Workbench* supports both paired end and mate pair protocols.

  Paired end Paired end data from Ion Torrent comes in two files per data set. The first file in is assumed to contain the first reads of the pair, and the second file is assumed to contain the second read in a pair. On import, the orientation of the reads is set to forward - reverse. When the reads have been imported, there will be one file with intact pairs, and one file where one part of the pair is missing (in this case, "single" is appended to the file name). The Workbench connects the right sequences together in the pair based on the read name. Read more about handling paired data in section 6.3.8.

  Mate pair The mate pair protocol for Ion Torrent entails that the two reads are separated by a linker sequence. During import of paired data, the linker sequence is removed and the two reads are separated and put into the same sequence list. You can change the linker sequence in the **Preferences** (in the **Edit** menu) under **Data**. When looking for the linker sequence, the Workbench requires 80 % of the maximum alignment score, using the following scoring scheme: matches = 1, mismatches = -2 and indels = -3. Some of the sequences may not have the linker in the middle of the sequence, and in that case the partial linker sequence is still removed, and the single read is put into a separate sequence list. Thus when you import Ion Torrent mate pair data, you may end up with two sequence lists: one for paired reads and one for single reads. Note that for de novo assembly projects, only the paired list should be used since the single reads list may contain reads where there is still a linker sequence present but only partially due to sequencing errors. Read more about handling paired data in section 6.3.8.

- **Discard read names**. For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the huge amount of reads. This option allows you to discard this option to save disk space.
- **Discard quality scores**. Quality scores are visualized in the mapping view and they are used for SNP detection. If this is not relevant for your work, you can choose to **Discard quality scores**. One of the benefits from discarding quality scores is that you will gain a lot in terms of reduced disk space usage and memory consumption. If you have selected the fna/qual option and choose to discard quality scores, you do not need to select a .qual file.

For sff files, you can also decide whether to use the clipping information in the file or not.

6.3.7 Complete Genomics

With *CLC Cancer Research Workbench 1.5* you can import *evidence* and *variation* files from Complete Genomics.

The variation files can be imported as tracks (see section 6.2).

The evidence files can be imported using the SAM/BAM importer, see section 6.3.9.

In order to import the evidence data file it need to be converted first. This is achieved using the CGA tools that can be downloaded from `http://www.completegenomics.com/sequence-data/cgatools/`.

The procedure for converting the data is the following.

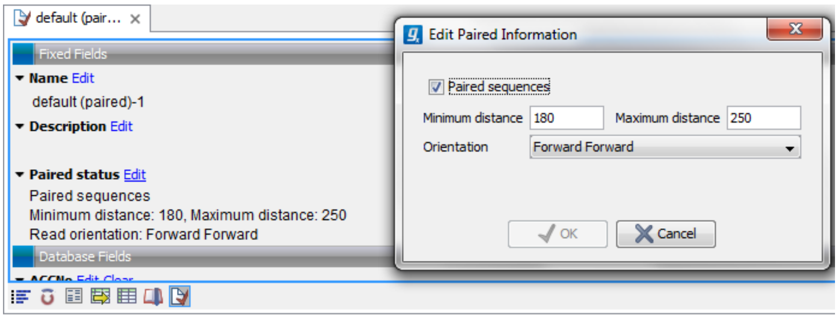
1. Download the human genome in fasta format and make sure the chromosomes are named `chr<number>.fa`, e.g. `chr9.fa`.
2. Run the **fasta2crr** tool with a command like this:
   ```
   cgatools fasta2crr --input chr9.fa --output chr9.crr
   ```
3. Run the **evidence2sam** tool with a command like this:
   ```
   cgatools evidence2sam --beta -e evidenceDnbs-chr9-.tsv -o chr9.sam -s chr9.crr
   ```
   where the .tsv file is the evidence file provided by Complete Genomics (you can find sample data sets on their ftp server: `ftp://ftp2.completegenomics.com/`).
4. **Import** (📥) the fasta file from 1. into the Workbench.
5. Use the SAM/BAM importer (section 6.3.9) to import the file created by the evidence2sam tool.

Please refer to the CGA documentation for a description about these tools. Note that this is not software supported by CLC bio.

6.3.8 General notes on handling paired data

During import, information about paired data (distances and orientation) can be specified (see figure 6.3 and figure 6.6 for data import examples of Roche 454 and Illumina reads, respectively) and is stored by the *CLC Cancer Research Workbench*. All subsequent analyses automatically take differences in orientation into account. Once imported, both reads of a pair will be stored in the same sequence list. The forward and reverse reads (e.g. for paired-end data) simply alternate so that the first read is forward, the second read is the mate reverse read; the third is again forward and the fourth read is the mate reverse read and so on. When manipulating sequence lists with paired data, be careful not break this order.

You can view and edit the orientation of the reads after they have been imported by opening the read list in the Element information view (), see section 9.4 as shown in figure 6.13.



Fixed Fields

Name Edit
default (paired)-1

Description Edit

Paired status Edit
Paired sequences
Minimum distance: 180, Maximum distance: 250
Read orientation: Forward Forward

Database Fields

Edit Paired Information

☑ Paired sequences

Minimum distance 180 Maximum distance 250

Orientation Forward Forward

Figure 6.13: *The paired orientation and distance.*

In the **Paired status** part, you can specify whether the *CLC Cancer Research Workbench* should treat the data as paired data, what the orientation is and what the preferred distance is. The orientation and preferred distance is specified during import and can be changed in this view.

Note that the **paired distance** measure that is used throughout the *CLC Cancer Research Workbench* is always *including the full read sequence*. For paired-end libraries it means from the beginning of the forward read to the beginning of the reverse read.

6.3.9 SAM and BAM mapping files

The *CLC Cancer Research Workbench* supports import and export of files in SAM (Sequence Alignment/Map) and BAM format, which are designed for storing large nucleotide sequence alignments. Read more and see the format specification at `http://samtools.sourceforge.net/`

The *CLC Cancer Research Workbench* includes support for importing SAM and BAM files from **Complete Genomics**.

Note! If you wish to import the reads in a SAM/BAM file as a sequence list, disregarding any mapping information, please use the Standard import tool instead (see section 6.1).

For a detailed explanation of the SAM and BAM files exported from *CLC Cancer Research Workbench*, please see Appendix F.

Input data for importing a mapping from a SAM/BAM file

To import a mapping from a SAM/BAM file containing mapping data into the Workbench, you need to:

- Provide the SAM/BAM file
- Specify the reference sequences that are referred to within that file. The references can either be sequences already imported into the Workbench, or, if appropriately recorded in the SAM/BAM file, can be fetched from URLs specified in the SAM/BAM file.

The mapping is built up within the Workbench using the reference sequence data, the reads and the information from the SAM/BAM file about how the reads are associated with a particular reference.

Data created in the Workbench after importing a SAM/BAM mapping file

- Reads recorded as mapping to a particular reference that is known inside the Workbench are imported as part of the mapping for that reference.
- Reads recorded as not mapping to any reference are imported into a sequence list.
  - If they are part of an intact pair, they are imported into a sequence list of paired data.
  - If they are single reads or a member of a pair that did not map while its mate did, they are imported into a sequence list containing single reads.

  One list is made per read group, with the potential that several such lists could be produced from a single mapping import. The sequence lists are given names of this form for single reads "<read group id> [read group sample] (single) un-mapped reads" and this form for paired reads "<read group id> [read group sample] (paired) un-mapped reads".

  If you do not wish to import the unmapped reads, deselect the **Import unmapped reads** option in the final step of the tool dialog.
- Reads recorded as mapping to a reference sequence that is **not** known within the Workbench are not imported.

When setting up the import, you are given the option of creating a track-based mapping, or a stand-alone mapping. In the latter case, if there is only one reference sequence, the result will be a single read mapping ( ). When there is more than one reference sequence, a multi- mapping object ( ) is created.

Please note that mappings within the *CLC Cancer Research Workbench* do not allow for an individual read sequence to map to more than one location. In cases where a SAM/BAM file contains multiple alignment records for a single read, only one such record will be used to build the mapping.

Running the SAM/BAM Mapping Files importer

Click on the Import button on the toolbar or go to:

File | Import ( ) | SAM/BAM Mapping Files ( )

This will open a dialog where you select the SAM/BAM file to import as well as the reference sequences to be used (Figure 6.14).

When you select the reference sequence(s) two options exist:

1. Select a matching reference sequence that has already been imported into the Workbench. Click on the "Find in folder" icon (🔍) to localize the reference sequence.

2. If the SAM/BAM file already contains information about where to find the reference sequence, tick the "Download references" box to automatically download the reference sequence.

The selected reference sequence(s) will be listed under "References in files" with "Name", "Length", and "Status". Whenever the correct reference sequence (with the correct name and sequence length) has been selected the "Status" field will indicate this with an "OK". The length of your reference sequence must **match exactly** the length of the reference specified in the SAM/BAM file. The name is more flexible as it allows a range of different "synonyms" (with no distinction between capital and lowercase letters). E.g. for chromosome 1 the allowed synonyms would be: 1, chr1, chromosome_1, nc_000001, for chromosome M: m, mt, chrm, chrmt, chromosome_m, chromosome_mt, nc_001807, for chromosome X: x, chrx, chromosome_x, nc_000023, and for chr Y: y, chry, chromosome_y, nc_000024.

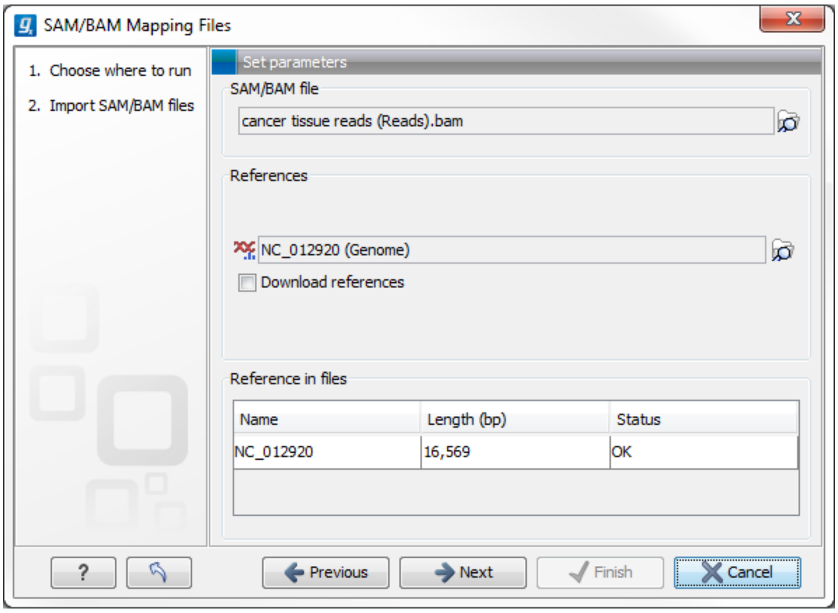
If there are inconsistencies in the names or lengths of the reference sequences being chosen and those recorded in the SAM/BAM file, an entry will appear in the "Status" column indicating this. E.g "Length differs" or "Input missing"[1].

Unmatched reads (reads that are mapped to an unmatched reference e.g. a SAM reference for which there is no CLC reference counterpart) are not imported. The same is the case whenever inconsistencies have occurred with respect to name or length. The log lists all mapping data or unmatched reads that were not imported and marks whether import failed because of unmatched reads being present in the SAM/BAM file or because of inconsistencies in name/length.

Some notes regarding reference sequence naming Reference sequences in a SAM/BAM file **cannot contain spaces**. If the name of a reference sequence in the Workbench contains spaces, the Workbench assume that the names of the references in the SAM file will be the same as the names of the References within the Workbench, but with all spaces removed. For exapmple, if your reference sequence in the Workbench was called `my reference sequence`, the Workbench would recognize a reference in the SAM file as the appropriate reference if it was of the same length and had the name `myreferencesequence`.

Neither the `@` character nor the `=` character are allowed within reference sequence names in SAM files. Any instances of these characters in the name of a reference sequence in the Workbench will be replaced with a _ for the sake of identifying the appropriate reference when importing a SAM or BAM file. For example, if a reference sequence in the Workbench was called `my=reference@sequence`, the Workbench would recognize a reference in the SAM file as the appropriate reference if it was of the same length and had the name `my_reference_sequence`.

[1] If you are using a CLC Genomics Server to import files located on the Server (rather than locally), then checks for corresponding reference names and lengths cannot be carried out, so nothing will be reported in this section of the Wizard. This means you will be able to continue to launch the import with correct or incorrect reference sets specified. However, any inconsistencies in these will lead to the import task failing with an error related to this.



SAM/BAM file

cancer tissue reads (Reads).bam

References

NC_012920 (Genome)

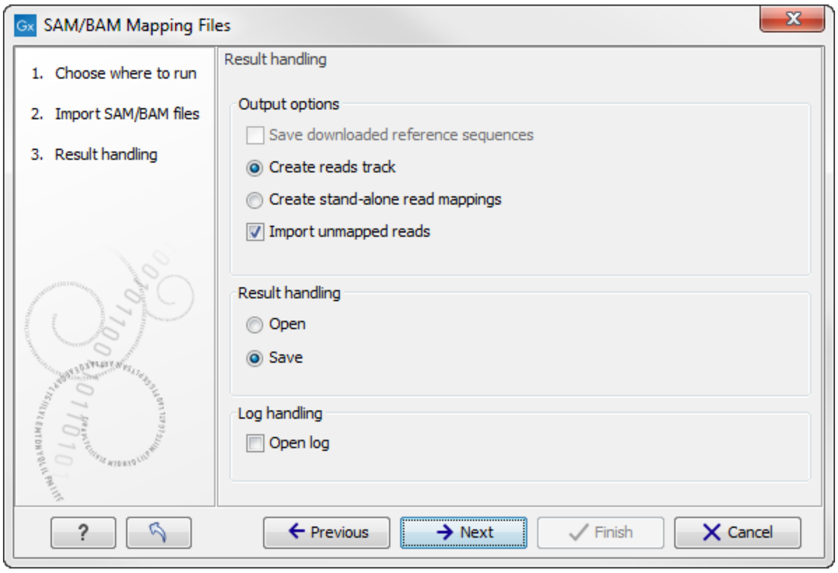
□ Download references

Reference in files

Name	Length (bp)	Status
NC_012920	16,569	OK

Figure 6.14: *Defining SAM/BAM file and reference sequence(s).*

Click **Next** to specify how to handle the results (Figure 6.15). Under **Output options** the "Save downloaded reference sequence" will be enabled if the "Download references" box was ticked in the previous step (which would be the case when the SAM/BAM file contained information about where to find the reference sequence e.g. if the SAM/BAM file came from an external provider).



Output options

□ Save downloaded reference sequences

● Create reads track

○ Create stand-alone read mappings

☑ Import unmapped reads

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 6.15: *Specify the result handling.*

Ticking the "Create Reads Track" box results in the generation of a track-based mapping. Alternatively, the "Create Stand-Alone Read Mapping" results in a normal read mapping file. By ticking the "Import unmapped reads" box, a sequence list of the unmapped reads will be created. To avoid importing unmapped reads, untick this box.

We recommend choosing **Save** in order to save the results directly to a folder, as you will probably wish to save the data anyway before proceeding with your analysis. For further information about how to handle the results, (see section 8.2).

Note that this import operation is very memory-consuming for large data sets, and particularly those with many reads marked as members of broken pairs in the mapping.

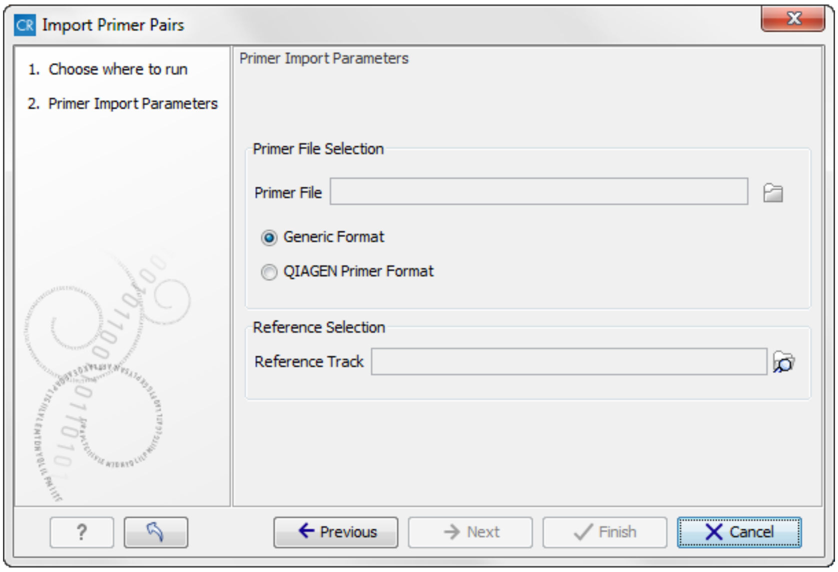
6.4 Import Primer Pairs

The **Import Primer Pairs** importer can import descriptions of primer locations from a generic text format file or from a QIAGEN gene panel primer file. The primer location file describes the location of primers used for targeted resequencing and is used for primer trimming by the tool **Trim Primers of Mapped Reads**. This tool is particularly useful for trimming off primers when you have targeted data with overlapping reads.

The **Import Primer Pairs** can be found in the toolbar:

Import ( ) | **Import Primer Pairs** ( )

This will open the wizard shown in figure 6.16. The first step is to select the data to import.



Import Primer Pairs

1. Choose where to run
2. Primer Import Parameters

Primer Import Parameters

Primer File Selection

Primer File

● Generic Format

○ QIAGEN Primer Format

Reference Selection

Reference Track

Figure 6.16: *Select files to import.*

- **Primer File** Click on the folder icon in the right side to select your primer pair location file.

  There are two primer pair formats that can be imported by the Workbench.

  – **Generic Format** Select this option for primer location files with the exception of QIAGEN gene panel primers. Provide your primer location information in a tab delimited text file with the following columns

    * Column 1: reference name
    * Column 2: primer1 first position on reference
    * Column 3: primer1 last position on reference
    * Column 4: primer2 first position on reference
    * Column 5: primer2 last position on reference
    * Column 6: amplicon name

    An example of the format expected for each row is:

```
chr1     42     65     142     106     Amplicon1
```

  - **QIAGEN Primer Format** Use this option for importing information about QIAGEN gene panel primers.

- **Reference Track** Use folder icon in the right side to select the relevant reference track.

Click on the button labeled **Next** to go to the wizard step choose to save the imported primer location file.

6.5 Data export

The exporter can be used to:

- Export bioinformatic data in most of the formats that can be imported. There are a few exceptions (see section E.1).
- Export one or more data elements at a time to a given format. When multiple data elements are selected, each is written out to an individual file, unless compression is turned on, or "Output as single file" is selected.

The standard export functionality can be launched using the Export button on the toolbar, or by going to the menu:

File | Export (📤)

An additional export tool is available from under the File menu:

File | Export with Dependent Elements

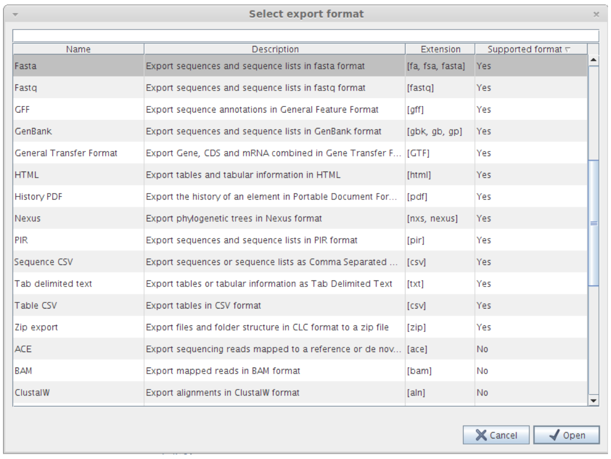
This tool is described further in section 6.5.2.

The general steps when configuring a standard export job are:

- (Optional) Select the data to export in the **Navigation Area**.
- Start up the exporter tool via the Export button in the toolbar or using the **Export** option under the File menu.
- Select the format the data should be exported to.
- Select the data to export, or confirm the data to export if it was already selected via the **Navigation Area**.
- Configure the parameters. This includes compression, multiple or single outputs, and naming of the output files, along with other format-specific settings where relevant.
- Select where the data should be exported to.
- Click on the button labeled **Finish**.

Selecting data for export - part I. You can select the data elements to export **before** you run the export tool **or after** the format to export to has been selected. If you are not certain which formats are supported for the data being exported, we recommend selecting the data in the **Navigation Area** before launching the export tool.

Selecting a format to export to. When data is pre-selected in the **Navigation Area** before launching the export tool you will see a column in the export interface called **Supported formats**. Formats that the selected data elements can be exported to are indicated by a "Yes" in this column. Supported formats will appear at the top of the list of formats. See figure 6.17.



Name	Description	Extension	Supported format
Fasta	Export sequences and sequence lists in fasta format	[fa, fsa, fasta]	Yes
Fastq	Export sequences and sequence lists in fastq format	[fastq]	Yes
GFF	Export sequence annotations in General Feature Format	[gff]	Yes
GenBank	Export sequences and sequence lists in GenBank format	[gbk, gb, gp]	Yes
General Transfer Format	Export Gene, CDS and mRNA combined in Gene Transfer F...	[GTF]	Yes
HTML	Export tables and tabular information in HTML	[html]	Yes
History PDF	Export the history of an element in Portable Document For...	[pdf]	Yes
Nexus	Export phylogenetic trees in Nexus format	[nxs, nexus]	Yes
PIR	Export sequences and sequence lists in PIR format	[pir]	Yes
Sequence CSV	Export sequences or sequence lists as Comma Separated ...	[csv]	Yes
Tab delimited text	Export tables or tabular information as Tab Delimited Text	[txt]	Yes
Table CSV	Export tables in CSV format	[csv]	Yes
Zip export	Export files and folder structure in CLC format to a zip file	[zip]	Yes
ACE	Export sequencing reads mapped to a reference or de nov...	[ace]	No
BAM	Export mapped reads in BAM format	[bam]	No
ClustalW	Export alignments in ClustalW format	[aln]	No

Figure 6.17: *The Select exporter dialog where sequence lists were pre-selected in the Navigation Area before launching the export tool. Here, the formats sequence lists can be exported to are listed at the top, with a Yes in the Selected formats column. Other formats are found below, with No in this column.*

Formats that cannot be used for export of the selected data have a "No" listed in the **Supported formats** column. If you have selected multiple data elements of different types, then formats which can be used for some of the selected data elements but not all of them are indicated by the text "For some elements" in this column.

Please note that the information in the **Supported formats** column only refers to the data already selected in the **Navigation Area**. If you are going to choose your data later in the export process, then the information in this column will not be pertinent.

Only one export format is available if you select a folder to be exported. This is described in more detail in section 6.5.1.

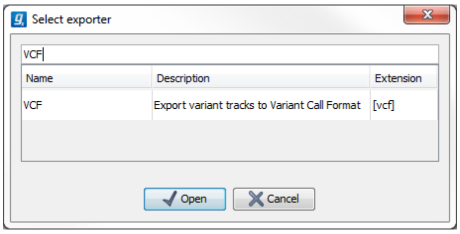
Finding a particular format in the list. You can quickly find a particular format by using the text box at the top of the exporter window as shown in figure 6.18, where formats that include the term VCF are searched for. This search term will remain in place the next time the Export tool is launched. Just delete the text from the search box if you no longer wish only the formats with that term to be listed.

When the desired export format has been identified, click on the button labeled **Open**.

Selecting data for export - part II. A dialog appears, with a name reflecting the format you have chosen. For example if the "Variant Call Format" (VCF format) was selected, the window is labeled "Export VCF".

If you are logged into a CLC Server, you will be asked whether to run the export job using the Workbench or the Server. After this, you are provided with the opportunity to select or de-select data to be exported.

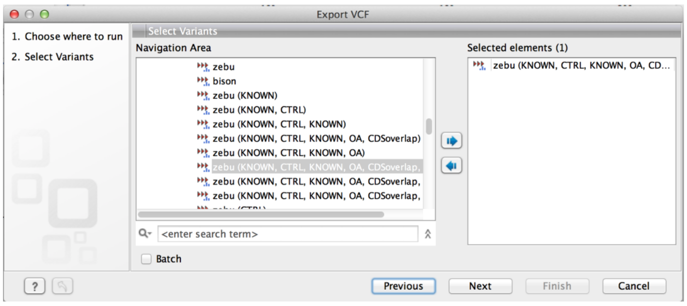
In figure 6.19 we show the selection of a variant track for export to VCF format.



VCF

Name	Description	Extension
VCF	Export variant tracks to Variant Call Format	[vcf]

Figure 6.18: *The text field has been used to search for VCF format in the Select exporter dialog.*



zebu
bison
zebu (KNOWN)
zebu (KNOWN, CTRL)
zebu (KNOWN, CTRL, KNOWN)
zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap)
zebu (KNOWN, CTRL, KNOWN, OA)
zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap,
zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap,
zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap,

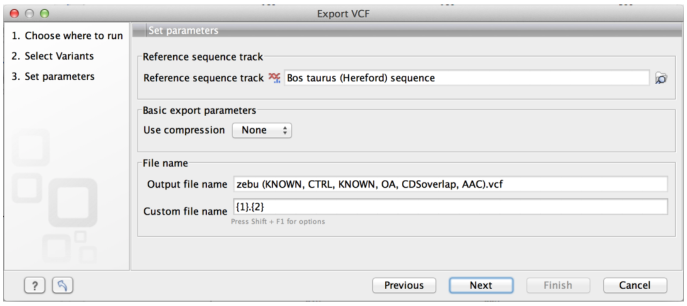
<enter search term>

Selected elements (1): zebu (KNOWN, CTRL, KNOWN, OA, CD...

□ Batch

Figure 6.19: *The Select exporter dialog. Select the data element(s) to export.*

The parameters under **Basic export parameters** and **File name** are offered when exporting to any format. There may be additional parameters for particular export formats. This is illustrated here with the VCF exporter, where a reference sequence track must be selected. See figure 6.20.



Reference sequence track
Reference sequence track: Bos taurus (Hereford) sequence

Basic export parameters
Use compression: None

File name
Output file name: zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap, AAC).vcf
Custom file name: {1}.{2}
Press Shift + F1 for options

Figure 6.20: *Set the export parameters. When exporting in VCF format, a reference sequence track must be selected.*

Compression options. Within the **Basic export parameters** section, you can choose to compress the exported files. The options are no compression (None), gzip or zip format. Choosing zip format results in all data files being compressed into a single file. Choosing gzip compresses the exported file for each data element individually.

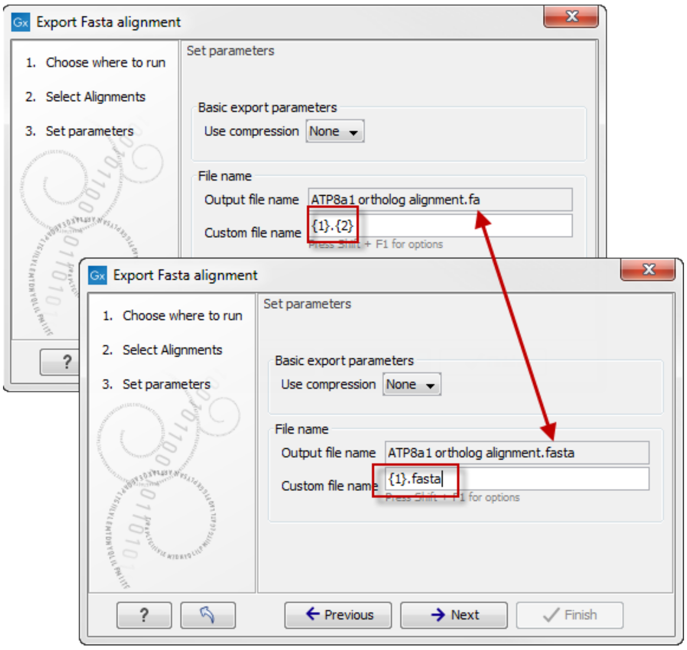
Exporting multiple files. If you have selected multiple files of the same type, you can choose to export them in one single file (only for certain file formats) by selecting "Output as single file" in the **Basic export parameters** section. If you wish to keep the files separate after export, make sure this box is not ticked. **Note:** Exporting in zip format will export only one zipped file, but the files will be separated again when unzipped.

Choosing the exported file name(s) The default setting for the **File name** is to use the original data element name as the basename and the export format as the suffix.

When exporting just one data element, or exporting to a zip file, the desired filename could just be typed in the Custom file name box.

When working with the export of multiple files, using some combination of the terms shown by default in this field and in figure 6.22 are recommended. Clicking in the **Custome file name** field with the mouse and then simultaneously pressing the Shift + F1 keys bring up a list of the available terms that can be included in this field. You can see that "{1}" is the name of the input element and "{2}" is the file name extension. It is possible to change the input file name and the file extension name. We will look at an example to illustrate this:

In this example we would like to change the export file format to .fasta in a situation where .fa was the default format that would be used if you kept the default file extension suggestion ("{2}"). To do this replace "{2}" with ".fasta" in the "Custom file name field". You can see that when changing "{2}" to ".fasta" , the file name extension in the "Output file name" field automatically changes to the new format (see figure 6.21).



Use compression None

Output file name ATP8a1 ortholog alignment.fa

Custom file name {1}.{2}

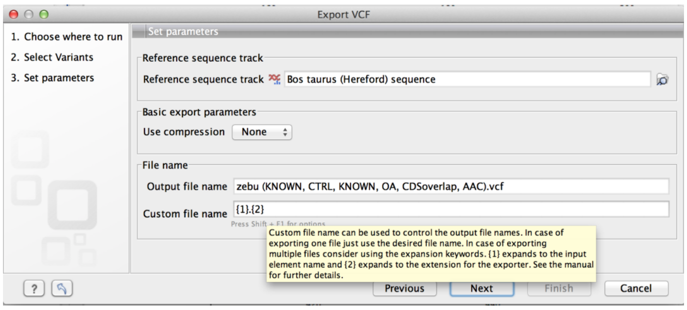
Use compression None

Output file name ATP8a1 ortholog alignment.fasta

Custom file name {1}.fasta

Figure 6.21: *The file name extension can be changed by typing in the preferred file name format.*

As you add or remove text and terms in the **Custome file name** field, the text in the **Output file name** field will change so you can see what the result of your naming choice will be for your data. When working with multiple files, only the name of the first one is shown. Just move the mouse cursor over the name shown in the **Output file name** field to show a listing of the all the filenames.



Reference sequence track Bos taurus (Hereford) sequence

Use compression None

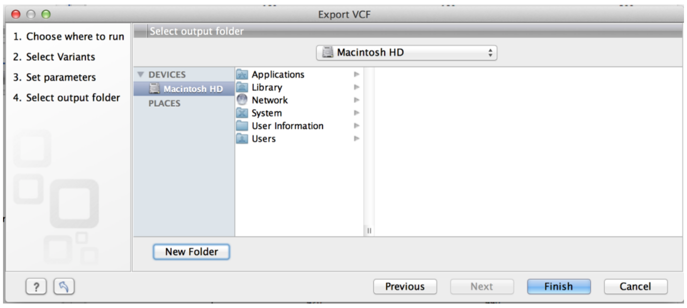
Output file name zebu (KNOWN, CTRL, KNOWN, OA, CDSoverlap, AAC).vcf

Custom file name {1}.{2}

Custom file name can be used to control the output file names. In case of exporting one file just use the desired file name. In case of exporting multiple files consider using the expansion keywords. {1} expands to the input element name and {2} expands to the extension for the exporter. See the manual for further details.

Figure 6.22: *Use the custom file name pattern text field to make custom names.*

The last step is to specify the exported data should be saved (figure 6.23).


Export VCF

1. Choose where to run
2. Select Variants
3. Set parameters
4. Select output folder

Select output folder

Figure 6.23: *Select where to save the exported data.*

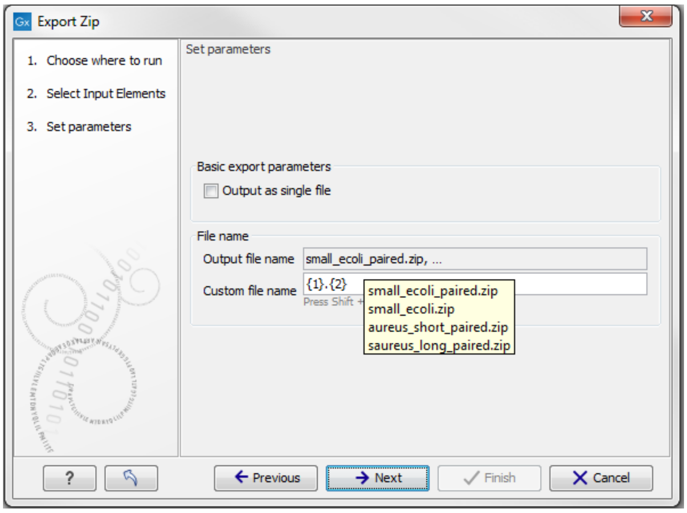
A note about decimals and Locale settings. When exporting to CSV and tab delimited files, decimal numbers are formatted according to the Locale setting of the Workbench (see section 4.1). If you open the CSV or tab delimited file with spreadsheet software like Excel, you should make sure that both the Workbench and the spreadsheet software are using the same Locale.

6.5.1 Export of folders and multiple elements in CLC format

In the list of export formats presented is one called zip format. Choosing this format means that you wish to export the selected data element(s) or folders to a single, compressed CLC format file. This is useful in cases where you wish to exchange data between workbenches or as part of a simple backup procedure.

A zip file generated this way can be imported directly into a CLC Workbench using the Standard Import tool and leaving the import type as Automatic.

Note! When exporting multiple files, the names will be listed in the "Output file name" text field with only the first file name being visible and the rest being substituted by "...", but will appear in a tool tip if you hover the mouse over that field (figure 6.24).


Export Zip

1. Choose where to run
2. Select Input Elements
3. Set parameters

Set parameters

Basic export parameters

☐ Output as single file

File name

Output file name: small_ecoli_paired.zip, ...

Custom file name: {1}.{2}

Press Shift +

small_ecoli_paired.zip
small_ecoli.zip
aureus_short_paired.zip
saureus_long_paired.zip

Figure 6.24: *The output file names are listed in the "Output file name" text field.*

6.5.2 Export of dependent elements

Sometimes it can be useful to export the results of an analysis and its dependent elements. That is, the results along with the data that was used in the analysis. For example, one might wish to export an alignment along with all the sequences that were used in generating that alignment.

To export a data element with its dependent elements:

- Select the parent data element (like an alignment) in the **Navigation Area**.
- Start up the exporter tool by going to **File** | **Export with Dependent Elements**.
- Edit the output name if desired and select where the resulting zip format file should be exported to.

The file you export contains compressed CLC format files containing the data element you chose and all its dependent data elements.

A zip file created this way can be imported directly into a CLC workbench by going to

File | **Import** | **Standard Import**

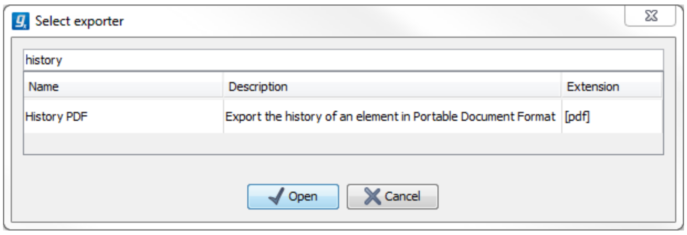
In this case, the import type can be left as Automatic.

6.5.3 Export history

Each data element in the Workbench has a history. The history information includes things like the date and time data was imported or an analysis was run, the parameters and values set, and where the data came from. For example, in the case of an alignment, one would see the sequence data used for that alignment listed. You can view this information for each data element by clicking on the Show History view (🕒) at the bottom of the viewing area when a data element is open in the Workbench.

This history information can be exported to a pdf document. To do this:

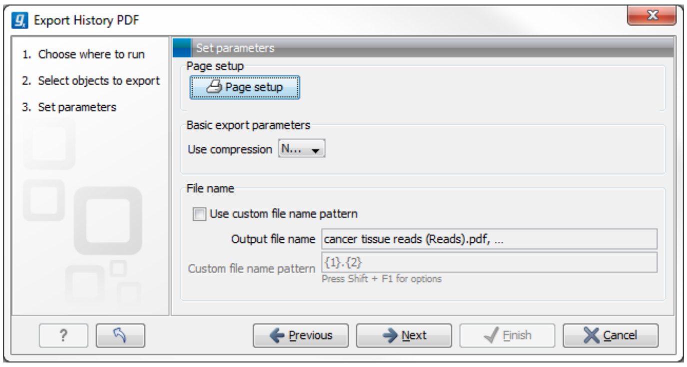
- (Optional, but preferred) Select the data element (like an alignment) in the **Navigation Area**.
- Start up the exporter tool via the Export button in the toolbar or using the **Export** option under the File menu.
- Select the **History PDF** as the format to export to. See figure 6.25.
- Select the data to export, or confirm the data to export if it was already selected via the **Navigation Area**.
- Edit any parameters of interest, such as the Page Setup details, the output filename(s) and whether or not compression should be applied. See figure 6.26.
- Select where the data should be exported to.
- Click on the button labeled **Finish**.



history

Name	Description	Extension
History PDF	Export the history of an element in Portable Document Format	[pdf]

Figure 6.25: *Select "History PDF" for exporting the history of an element.*



1. Choose where to run
2. Select objects to export
3. Set parameters

Page setup

Basic export parameters

Use compression N...

File name

☐ Use custom file name pattern

Output file name cancer tissue reads (Reads).pdf, ...

Custom file name pattern {1}.{2}

Press Shift + F1 for options

Figure 6.26: *When exporting the history in PDF, it is possible to adjust the page setup.*

6.5.4 The CLC format

The *CLC Cancer Research Workbench* stores bioinformatic data in CLC format. The CLC format contains data, as well as information about that data like history information and comments you may have added.

A given data element in the Workbench can contain different types of data. This is reflected when exporting data, as the choice of different export formats can lead to the extraction of some parts of that data object rather than others. The part of the data exported reflects the type of data a given format can support. As a simple example, if you export the results of an alignment to Annotation CSV format, you will get just the annotation information. If you exported to Fasta alignment format, you would get the aligned sequences in fasta format, but no annotations.

The CLC format holds all the information for a given data object. Thus if you plan to share the data with colleagues who also have a CLC Workbench or you are communicating with the CLC Support team and you wish to share the data from within the Workbench, exporting to CLC format is usually the best choice as all information associated with that data object in your Workbench will then be available to the other person who imports that data.

If you are planning to share your data with someone who does not have access to a CLC Workbench, then you will wish to export to another data format. Specifically, one they can use with the software they are working with.

6.5.5 Backing up data from the CLC Workbench

Regular backups of your data are advisable.

The data stored in your CLC Workbench is in the areas defined as CLC Data Locations. Whole data locations can be backed up directly (option 1) or, for smaller amounts of data, you could export the selected data elements to a zip file (option 2).

Option 1: Backing up each CLC Data Location

The easiest way for most people to find out where their data is stored is to put the mouse cursor over the top level directories, that is, the ones that have an icon like (), in the **Navigation Area** of the Workbench. This brings up a tool tip with the system location for that data location.

To back up all your CLC data, please ensure that all your CLC Data Locations are backed up.

Here, if you needed to recover the data later, you could put add the data folder from backup as a data location in your Workbench. If the original data location is not present, then the data should be usable directly. If the original data location is still present, the Workbench will re-index the (new) data location. For large volumes of data, re-indexing can take some time.

Information about your data locations can also be found in an xml file called model_settings_300.xml This file is located in the settings folder in the user home area. Further details about this file and how it pertains to data locations in the Workbench can be found in the Deployment Manual:

http://www.clcsupport.com/workbenchdeployment/current/index.php?manual=Changing_default_location.html

Option 2: Export a folder of data or individual data elements to a CLC zip file

This option is for backing up smaller amounts of data, for example, certain results files or a whole data location, where that location contains smaller amounts of data. For data that takes up many gigabases of space, this method can be used, but it can be very demanding on space, as well as time.

Select the data items, including any folders, in the Navigation area of your Workbench and choose to export by going to:

File | **Export** ()

and choosing ZIP format.

The zip file created will contain all the data you selected. You can later re-import the zip file into the Workbench by going to:

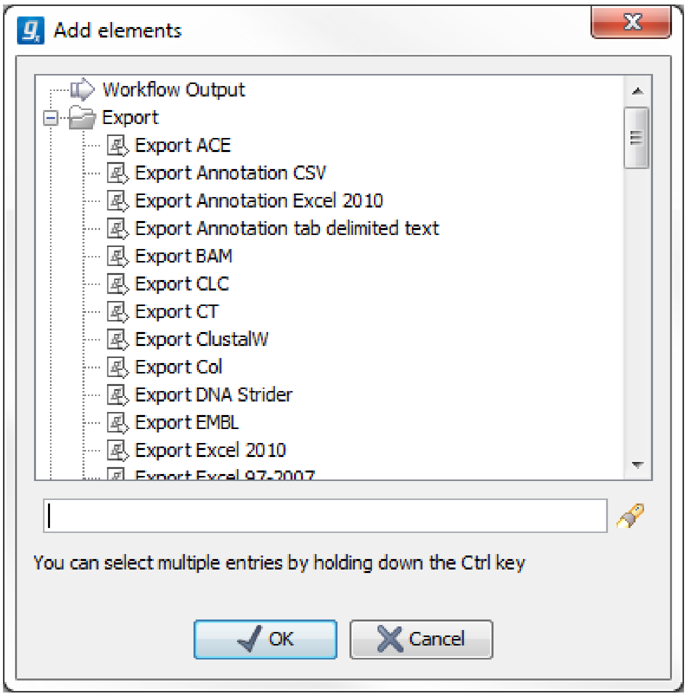
File | **Import** ()

The only data files associated with the *CLC Cancer Research Workbench* not within a specified data location are BLAST databases. It is unusual to back up BLAST databases as they are usually updated relatively frequently and in many cases can be easily re-created from the original files or re-downloaded from public resources. If you do wish to backup your BLAST database files, they can be found in the folders specified in the BLAST Database Manager, which is started by going to: **Toolbox | BLAST | Manage BLAST databases** .

6.5.6 Export of workflow output

The output from a workflow can be exported by adding one or more workflow export elements (figure 6.27). Multiple elements can be selected by holding down the Ctrl key while clicking on the desired elements.

When the workflow has been created, you can set the export parameters and the location to export data to by double clicking on each export element or leave fields empty and unlocked if you wish users of the Workflow to enter this information when the Workflow is launched.



You can select multiple entries by holding down the Ctrl key

Figure 6.27: *Pressing "Add element" enables addition of workflow export elements.*

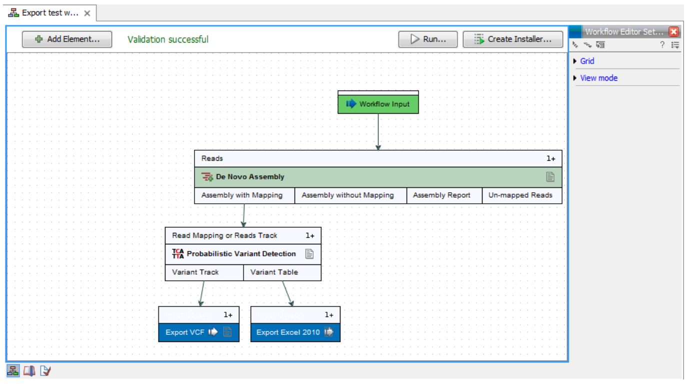


Figure 6.28: *A simple workflow with two export elements. The variant track will be exported in VCF format and the variant table in Excel format.*

6.5.7 Export of tables

Tables can be exported in four different formats; CSV, tab-separated, Excel, or html. When exporting a table in CSV, tab-separated, or Excel format, numbers with many decimals are printed in the exported file with 10 decimals, or in 1.123E-5 format when the number is close to zero.

When exporting a table in html format, data are exported with the number of decimals that have been defined in the workbench preference settings. When tables are exported in html format from the server or using command line tools, the default number of exported decimals is 3.

6.6 Export graphics to files

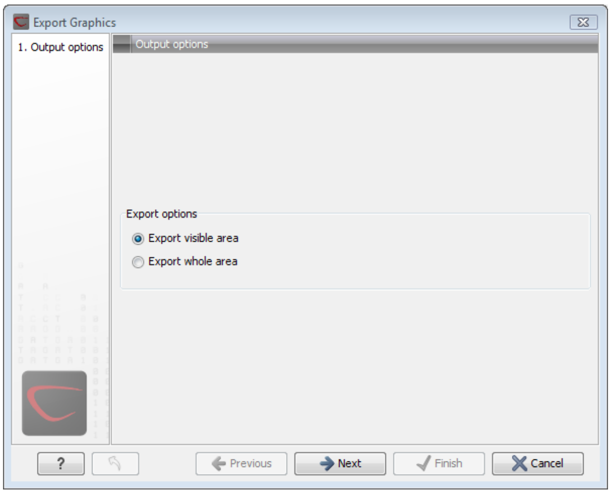
CLC Cancer Research Workbench supports export of graphics into a number of formats. This way, the visible output of your work can easily be saved and used in presentations, reports etc. The **Export Graphics** function () is found in the **Toolbar**.

CLC Cancer Research Workbench uses a WYSIWYG principle for graphics export: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks in the program. When you export it, the graphics file will look exactly the same way.

It is not possible to export graphics of elements directly from the **Navigation Area**. They must first be opened in a view in order to be exported. To export graphics of the contents of a view:

select tab of View | Graphics ( ) on Toolbar

This will display the dialog shown in figure 6.29.



Export Graphics

1. Output options

Output options

Export options

● Export visible area

○ Export whole area

Figure 6.29: *Selecting to export whole view or to export only the visible area.*

6.6.1 Which part of the view to export

In this dialog you can choose to:

- **Export visible area**, or
- **Export whole view**

These options are available for all views that can be zoomed in and out. In figure 6.30 is a view of a circular sequence which is zoomed in so that you can only see a part of it.

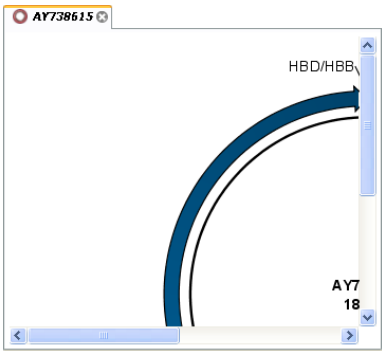


Figure 6.30: *A circular sequence as it looks on the screen.*

When selecting **Export visible area**, the exported file will only contain the part of the sequence that is *visible* in the view. The result from exporting the view from figure 6.30 and choosing **Export visible area** can be seen in figure 6.31.

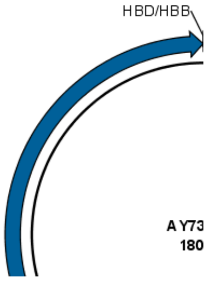


Figure 6.31: *The exported graphics file when selecting Export visible area.*

On the other hand, if you select **Export whole view**, you will get a result that looks like figure 6.32. This means that the graphics file will also include the part of the sequence which is not visible when you have zoomed in.

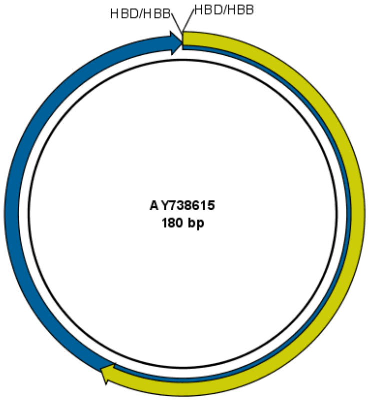


Figure 6.32: *The exported graphics file when selecting Export whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

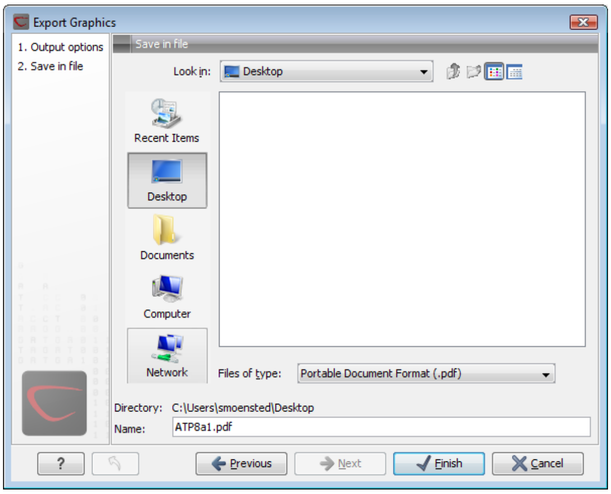
For 3D structures, this first step is omitted and you will always export what is shown in the view (equivalent to selecting **Export visible area**).

Click **Next** when you have chosen which part of the view to export.

6.6.2 Save location and file formats

In this step, you can choose name and save location for the graphics file (see figure 6.33).

CLC Cancer Research Workbench supports the following file formats for graphics export:



Export Graphics

1. Output options
2. Save in file

Save in file

Look in: Desktop

Files of type: Portable Document Format (.pdf)

Directory: C:\Users\smoensted\Desktop

Name: ATP8a1.pdf

Figure 6.33: *Location and name for the graphics file.*

Format	**Suffix**	**Type**
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

These formats can be divided into bitmap and vector graphics. The difference between these two categories is described below:

Bitmap images

In a bitmap image, each dot in the image has a specified color. This implies, that if you zoom in on the image there will not be enough dots, and if you zoom out there will be too many. In these cases the image viewer has to interpolate the colors to fit what is actually looked at. A bitmap image needs to have a high resolution if you want to zoom in. This format is a good choice for storing images without large shapes (e.g. dot plots). It is also appropriate if you don't have the need for resizing and editing the image after export.

Vector graphics

Vector graphic is a collection of shapes. Thus what is stored is e.g. information about where a line starts and ends, and the color of the line and its width. This enables a given viewer to decide how to draw the line, no matter what the zoom factor is, thereby always giving a correct image. This format is good for e.g. graphs and reports, but less usable for e.g. dot plots. If the image is to be resized or edited, vector graphics are by far the best format to store graphics. If you open a vector graphics file in an application like e.g. Adobe Illustrator, you will be able to manipulate the image in great detail.

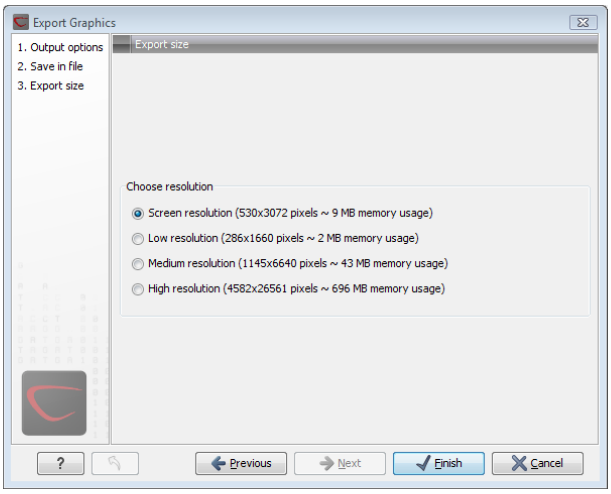
Graphics files can also be imported into the **Navigation Area**. However, no kinds of graphics files can be displayed in *CLC Cancer Research Workbench*. See section 6.1.4 for more about importing external files into *CLC Cancer Research Workbench*.

6.6.3 Graphics export parameters

When you have specified the name and location to save the graphics file, you can either click **Next** or **Finish**. Clicking **Next** allows you to set further parameters for the graphics export, whereas clicking **Finish** will export using the parameters that you have set last time you made a graphics export in that file format (if it is the first time, it will use default parameters).

Parameters for bitmap formats

For bitmap files, clicking **Next** will display the dialog shown in figure 6.34.



Choose resolution

● Screen resolution (530x3072 pixels ~ 9 MB memory usage)
○ Low resolution (286x1660 pixels ~ 2 MB memory usage)
○ Medium resolution (1145x6640 pixels ~ 43 MB memory usage)
○ High resolution (4582x26561 pixels ~ 696 MB memory usage)

Figure 6.34: *Parameters for bitmap formats: size of the graphics file.*

You can adjust the size (the resolution) of the file to four standard sizes:

- Screen resolution
- Low resolution
- Medium resolution
- High resolution

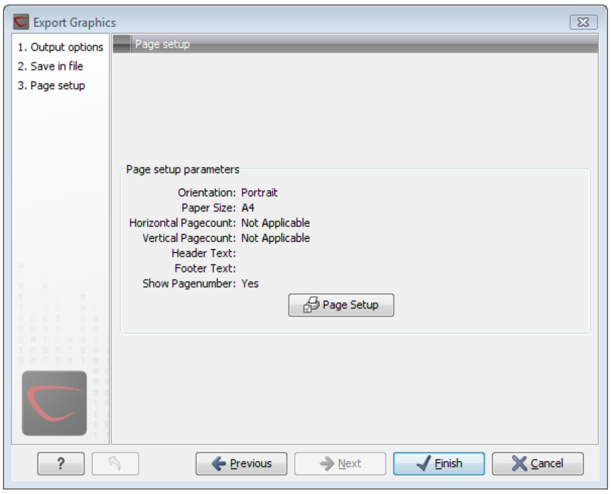
The actual size in pixels is displayed in parentheses. An estimate of the memory usage for exporting the file is also shown. If the image is to be used on computer screens only, a low resolution is sufficient. If the image is going to be used on printed material, a higher resolution is necessary to produce a good result.

Parameters for vector formats

For pdf format, clicking **Next** will display the dialog shown in figure 6.35 (this is only the case if the graphics is using more than one page).

The settings for the page setup are shown, and clicking the **Page Setup** button will display a dialog where these settings can ba adjusted. This dialog is described in section 5.2.

The page setup is only available if you have selected to export the whole view - if you have chosen to export the visible area only, the graphics file will be on one page with no headers or footers.



Page setup parameters

Orientation: Portrait
Paper Size: A4
Horizontal Pagecount: Not Applicable
Vertical Pagecount: Not Applicable
Header Text:
Footer Text:
Show Pagenumber: Yes

Page Setup

Figure 6.35: *Page setup parameters for vector formats.*

6.6.4 Exporting protein reports

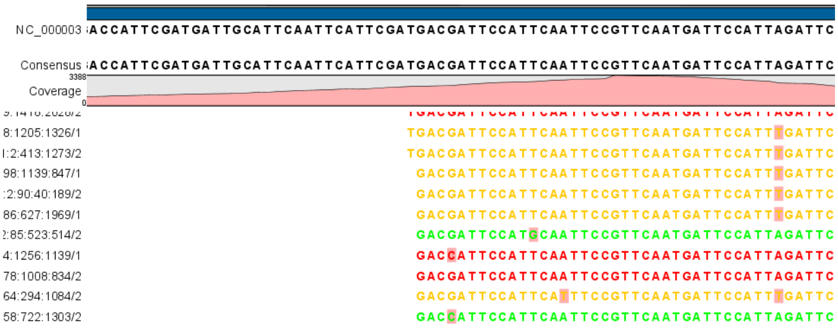
It is possible to export a protein report using the normal **Export** function (📤) which will generate a pdf file with a table of contents:

> **Click the report in the Navigation Area | Export (📤) in the Toolbar | select pdf**

You can also choose to export a protein report using the **Export graphics** function (📤), but in this way you will not get the table of contents.

6.7 Export graph data points to a file

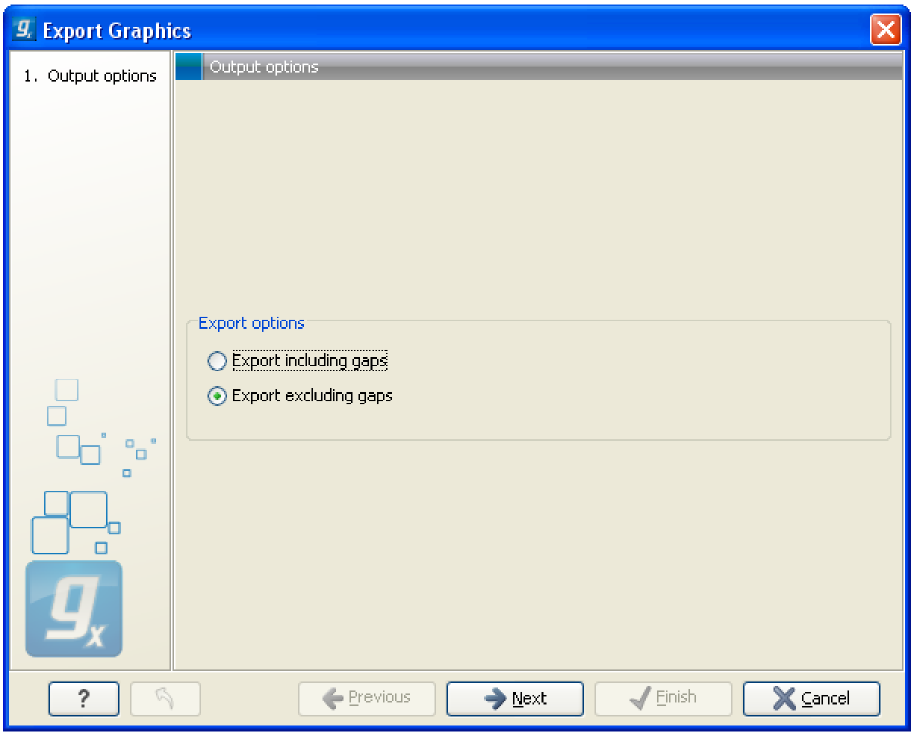
Data points for graphs displayed along the sequence or along an alignment, mapping or BLAST result, can be exported to a semicolon-separated text file (csv format). An example of such a graph is shown in figure 6.36. This graph shows the coverage of reads of a read mapping (produced with *CLC Genomics Workbench*).



NC_000003 ACCATTCGATGATTGCATTCAATTCATTCGATGACGATTCCATTCAATTCCGTTCAATGATTCCATTAGATTC
Consensus ACCATTCGATGATTGCATTCAATTCATTCGATGACGATTCCATTCAATTCCGTTCAATGATTCCATTAGATTC
3388
Coverage
0
8:1205:1326/1 TGACGATTCCATTCAATTCCGTTCAATGATTCCATTTGATTC
I:2:413:1273/2 TGACGATTCCATTCAATTCCGTTCAATGATTCCATTTGATTC
98:1139:847/1 GACGATTCCATTCAATTCCGTTCAATGATTCCATTTGATTC
:2:90:40:189/2 GACGATTCCATTCAATTCCGTTCAATGATTCCATTTGATTC
86:627:1969/1 GACGATTCCATTCAATTCCGTTCAATGATTCCATTTGATTC
2:85:523:514/2 GACGATTCCATGCAATTCCGTTCAATGATTCCATTAGATTC
4:1256:1139/1 GACCATTCCATTCAATTCCGTTCAATGATTCCATTAGATTC
78:1008:834/2 GACGATTCCATTCAATTCCGTTCAATGATTCCATTAGATTC
64:294:1084/2 GACGATTCCATTCATTTCCGTTCAATGATTCCATTTGATTC
58:722:1303/2 GACCATTCCATTCAATTCCGTTCAATGATTCCATTAGATTC

Figure 6.36: *A graph displayed along the mapped reads. Right-click the graph to export the data points to a file.*

To export the data points for the graph, right-click the graph and choose **Export Graph to Comma-separated File**. Depending on what kind of graph you have selected, different options will be shown: If the graph is covering a set of aligned sequences with a main sequence, such as read mappings and BLAST results, the dialog shown in figure 6.37 will be displayed. These kinds of graphs are located under **Alignment info** in the Side Panel. In all other cases, a normal file dialog will be shown letting you specify name and location for the file.



Export options

○ Export including gaps

● Export excluding gaps

Figure 6.37: *Choosing to include data points with gaps*

In this dialog, select whether you wish to include positions where the main sequence (the reference sequence for read mappings and the query sequence for BLAST results) has gaps. If you are exporting e.g. coverage information from a read mapping, you would probably want to exclude gaps, if you want the positions in the exported file to match the reference (i.e. chromosome) coordinates. If you export including gaps, the data points in the file no longer corresponds to the reference coordinates, because each gap will shift the coordinates.

Clicking **Next** will present a file dialog letting you specify name and location for the file.

The output format of the file is like this:

```
"Position";"Value";
"1";"13";
"2";"16";
"3";"23";
"4";"17";
...
```

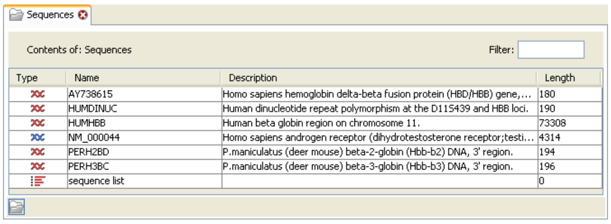
6.8 Copy/paste view output

The content of tables, e.g. in reports, folder lists, and sequence lists can be copy/pasted into different programs, where it can be edited. *CLC Cancer Research Workbench* pastes the data in tabulator separated format which is useful if you use programs like Microsoft Word and Excel. There is a huge number of programs in which the copy/paste can be applied. For simplicity, we include one example of the copy/paste function from a **Folder Content** view to Microsoft Excel.

First step is to select the desired elements in the view:

> **click a line in the Folder Content view | hold Shift-button | press arrow down/up key**

See figure 6.38.



Sequences

Contents of: Sequences

Filter:

Type	Name	Description	Length
	AY738615	Homo sapiens hemoglobin delta-beta fusion protein (HBD/HBB) gene,...	180
	HUMDINUC	Human dinucleotide repeat polymorphism at the D11S439 and HBB loci.	190
	HUMHBB	Human beta globin region on chromosome 11.	73308
	NM_000044	Homo sapiens androgen receptor (dihydrotestosterone receptor;testi...	4314
	PERH2BD	P.maniculatus (deer mouse) beta-2-globin (Hbb-b2) DNA, 3' region.	194
	PERH3BC	P.maniculatus (deer mouse) beta-3-globin (Hbb-b3) DNA, 3' region.	196
	sequence list		0

Figure 6.38: *Selected elements in a Folder Content view.*

When the elements are selected, do the following to copy the selected elements:

> **right-click one of the selected elements | Edit | Copy ( )**

Then:

> **right-click in the cell A1 | Paste ( )**

The outcome might appear unorganized, but with a few operations the structure of the view in *CLC Cancer Research Workbench* can be produced. (Except the icons which are replaced by file references in Excel.)

Note that all tables can also be **Exported** ( ) directly in Excel format.

Chapter 7

History log

Contents

CLC Cancer Research Workbench keeps a log of all operations you make in the program. If e.g. you rename a sequence, align sequences, create a phylogenetic tree or translate a sequence, dock a ligand, align sequences, or create a phylogenetic tree, you can always go back and check what you have done. In this way, you are able to document and reproduce previous operations.

This can be useful in several situations: It can be used for documentation purposes, where you can specify exactly how your data has been created and modified. It can also be useful if you return to a project after some time and want to refresh your memory on how the data was created. Also, if you have performed an analysis and you want to reproduce the analysis on another element, you can check the history of the analysis which will give you all parameters you set.

This chapter will describe how to use the **History** functionality of *CLC Cancer Research Workbench*.

7.1 Element history

You can view the history of all elements in the **Navigation Area** except files that are opened in other programs (e.g. Word and pdf-files). The history starts when the element appears for the first time in *CLC Cancer Research Workbench*. To view the history of an element:

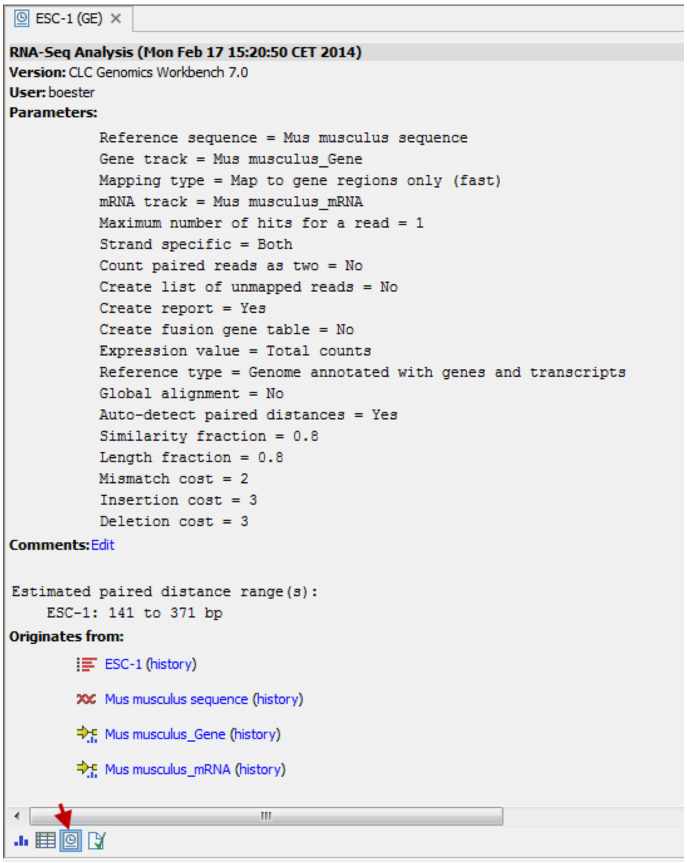
Select the element in the Navigation Area | Show ( ) in the Toolbar | History ( )

or **If the element is already open | History ( ) at the bottom left part of the view**

This opens a view that looks like the one in figure 7.1.

When an element's history is opened, the newest change is submitted in the top of the view. The following information is available:

- **Title**. The action that the user performed.
- **Date and time**. Date and time for the operation. The date and time are displayed according



RNA-Seq Analysis (Mon Feb 17 15:20:50 CET 2014)
Version: CLC Genomics Workbench 7.0
User: boester
Parameters:

```
          Reference sequence = Mus musculus sequence
          Gene track = Mus musculus_Gene
          Mapping type = Map to gene regions only (fast)
          mRNA track = Mus musculus_mRNA
          Maximum number of hits for a read = 1
          Strand specific = Both
          Count paired reads as two = No
          Create list of unmapped reads = No
          Create report = Yes
          Create fusion gene table = No
          Expression value = Total counts
          Reference type = Genome annotated with genes and transcripts
          Global alignment = No
          Auto-detect paired distances = Yes
          Similarity fraction = 0.8
          Length fraction = 0.8
          Mismatch cost = 2
          Insertion cost = 3
          Deletion cost = 3
```

Comments: Edit

```
Estimated paired distance range(s):
    ESC-1: 141 to 371 bp
```

Originates from:

ESC-1 (history)
Mus musculus sequence (history)
Mus musculus_Gene (history)
Mus musculus_mRNA (history)

Figure 7.1: *An element's history.*

to your locale settings (see section 4.1).

- **User**. The user who performed the operation. If you import some data created by another person in a CLC Workbench, that persons name will be shown.
- **Parameters**. Details about the action performed. This could be the parameters that was chosen for an analysis.
- **Origins from**. This information is usually shown at the bottom of an element's history. Here, you can see which elements the current element origins from. If you have e.g. created an alignment of three sequences, the three sequences are shown here. Clicking the element selects it in the **Navigation Area**, and clicking the 'history' link opens the element's own history.
- **Comments**. By clicking **Edit** you can enter your own comments regarding this entry in the history. These comments are saved.

7.1.1 Sharing data with history

The history of an element is attached to that element, which means that exporting an element in CLC format (*.clc) will export the history too. In this way, you can share folders and files with others while preserving the history. If an element's history includes source elements (i.e. if there are elements listed in 'Origins from'), they must also be exported in order to see the full history. Otherwise, the history will have entries named "Element deleted". An easy way to export an element with all its source elements is to use the **Export Dependent Elements** function described in section 6.5.2.

The history view can be printed. To do so, click the **Print** icon (🖨). The history can also be exported as a pdf file:

> **Select the element in the Navigation Area | Export (📤) | in "File of type" choose History PDF | Save**

Chapter 8

Batching and result handling

Contents

8.1 Batch processing

Most of the analyses in the **Toolbox** are able to perform the same analysis on several elements in one batch. This means that analyzing large amounts of data is very easily accomplished. As an example, if you use the **Find Binding Sites and Create Fragments ()** tool available in *CLC Cancer Research Workbench* and supply five sequences as shown in figure 8.1, the result table will present an overview of the results for all five sequences.

This is because the input sequences are pooled before running the analysis. If you want individual outputs for each sequence, you would need to run the tool five times, or alternatively use the **Batching mode**.

Batching mode is activated by clicking the **Batch** checkbox in the dialog where the input data is selected. Batching simply means that each data set is run separately, just as if the tool has been run manually for each one. For some analyses, this simply means that each input sequence should be run separately, but in other cases it is desirable to pool sets of files together in one run. This selection of data for a batch run is defined as a **batch unit**.

When batching is selected, the data to be added is the folder containing the data you want to batch. The content of the folder is assigned into batch units based on this concept:

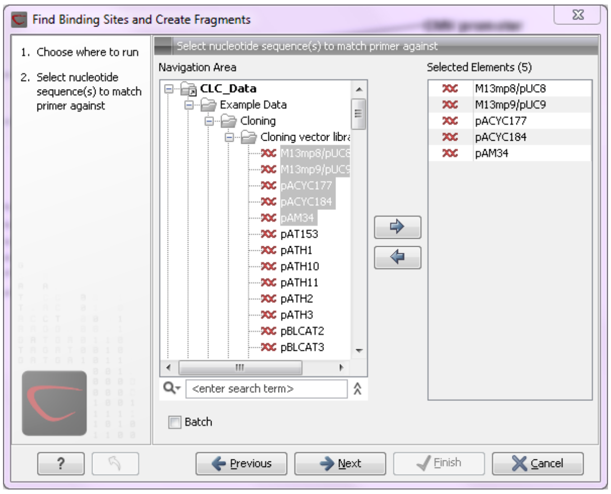


Figure 8.1: *Inputting five sequences to Find Binding Sites and Create Fragments.*

- All subfolders are treated as individual batch units. This means that if the subfolder contains several input files, they will be pooled as one batch unit. Nested subfolders (i.e. subfolders within the subfolder) are ignored.
- All files that are not in subfolders are treated as individual batch units.

An example of a batch run is shown in figure 8.2.

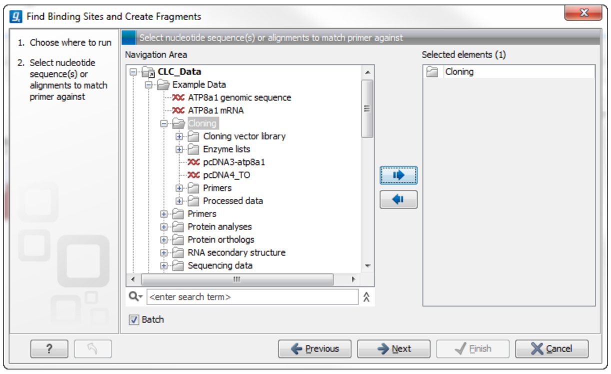


Figure 8.2: *The Cloning folder includes both folders and sequences.*

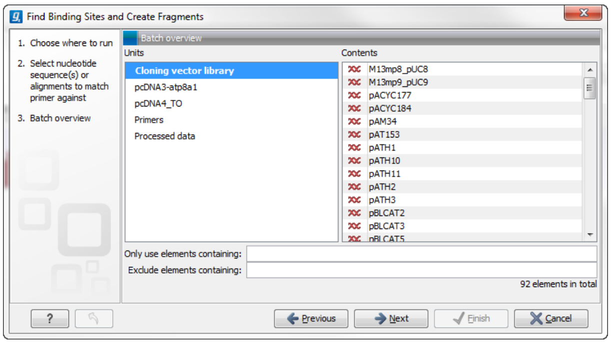
The `Cloning` folder that is found in the example data (see section 1.6.1) contains two sequences (🧬) and four folders (📁). If you click **Batch**, only folders can be added to the list of selected elements in the right-hand side of the dialog. To run the contents of the Cloning folder in batch, double-click to select it.

When the Cloning folder is selected and you click **Next**, a batch overview is shown.

8.1.1 Batch overview

The batch overview lists the batch units to the left and the contents of the selected unit to the right (see figure 8.3).

In this example, the two sequences are defined as separate batch units because they are located at the top level of the Cloning folder. There were also four folders in the Cloning folder (see



Find Binding Sites and Create Fragments

1. Choose where to run
2. Select nucleotide sequence(s) or alignments to match primer against
3. Batch overview

Batch overview

Units

- **Cloning vector library**
- pcDNA3-atp8a1
- pcDNA4_TO
- Primers
- Processed data

Contents

- M13mp8_pUC8
- M13mp9_pUC9
- pACYC177
- pACYC184
- pAM34
- pAT153
- pATH1
- pATH10
- pATH11
- pATH2
- pATH3
- pBLCAT2
- pBLCAT3
- pBLCAT5

Only use elements containing:

Exclude elements containing:

92 elements in total

Figure 8.3: *Overview of the batch run.*

figure 8.2), and three of them are listed as well. This means that the contents of these folders are pooled in one batch run (you can see the contents of the `Cloning vector library` batch run in the panel at the right-hand side of the dialog). The reason why the `Enzyme lists` folder is not listed as a batch unit is that it does not contain any sequences.

In this overview dialog, the Workbench has filtered the data so that only the types of data accepted by the tool is shown (DNA sequences in the example above).

8.1.2 Batch filtering and counting

At the bottom of the dialog shown in figure 8.3, the Workbench counts the number of files that will be run in total (92 in this case). This is counted across all the batch units.

In some situations it is useful to filter the input for the batching based on names. As an example, this could be to include only paired reads for a mapping, by only allowing names where "paired" is part of the name.

This is achieved using the **Only use elements containing** and **Exclude elements containing** text fields. Note that the count is dynamically updated to reflect the number of input files based on the filtering.

If a complete batch unit should be removed, you can select it, right-click and choose **Remove Batch Unit**. You can also remove items from the contents of each batch unit using right-click and **Remove Element**.

8.1.3 Setting parameters for batch runs

For some tools, the subsequent dialogs depend on the input data. In this case, one of the units is specified as parameter prototype and will be used to guide the choices in the dialogs. Per default, this will be the first batch unit (marked in bold), but this can be changed by right-clicking another batch unit and click **Set as Parameter Prototype**.

Note that the Workbench is validating a lot of the input and parameters when running in normal "non-batch" mode. When running in batch, this validation is not performed, and this means that some analyses will fail if combinations of input data and parameters are not right. Therefore batching should only be used when the batch units are very homogenous in terms of the type and size of data.

8.1.4 Running the analysis and organizing the results

At the last dialog before clicking **Finish**, it is only possible to use the **Save** option. When a tool is run in batch mode, the default behavior is to place the result files in the same folder as the input files. In the example shown in figure 8.3, the result of the two single sequences will be placed in the Cloning folder, whereas the results for the `Cloning vector library` and `Processed data` runs will be placed inside these folders.

However, there is an option to save the results in a separate folder structure by checking **Into separate folders**. This will allow you to specify a new save destination, and the *CLC Cancer Research Workbench* will create a subfolder for each batch unit where the results are saved..

When the batch run is started, there will be one "master" process representing the overall batch job, and there will then be a separate process for each batch unit. The behavior of this is different between Workbench and Server:

- When running the batch job in the Workbench, only one batch unit is run at a time. So when the first batch unit is done, the second will be started and so on. This is done in order to avoid many parallel analyses that would draw on the same compute resources and slow down the computer.
- When this is run on a CLC Server (see `http://clcbio.com/server`), all the processes are placed in the queue, and the queue is then taking care of distributing the jobs. This means that if the server set-up includes multiple nodes, the jobs can be run in parallel.

If you need to stop the whole batch run, you need to stop the "master" process.

8.2 How to handle results of analyses

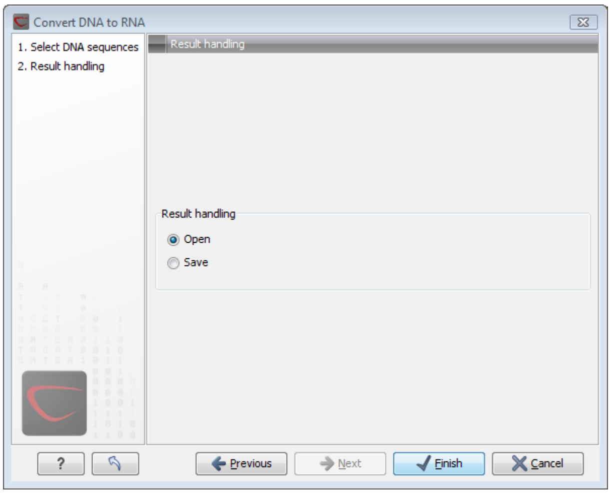
This section will explain how results generated from tools in the Toolbox are handled by *CLC Cancer Research Workbench*. Note that this also applies to tools not running in batch mode (see above). All the analyses in the **Toolbox** are performed in a step-by-step procedure. First, you select elements for analyses, and then there are a number of steps where you can specify parameters (some of the analyses have no parameters, e.g. when translating DNA to RNA). The final step concerns the handling of the results of the analysis, and it is almost identical for all the analyses so we explain it in this section in general.

In this step, shown in figure 8.4, you have two options:

- **Open.** This will open the result of the analysis in a view. This is the default setting.
- **Save.** This means that the result will not be opened but saved to a folder in the **Navigation Area**. If you select this option, click **Next** and you will see one more step where you can specify where to save the results (see figure 8.5). In this step, you also have the option of creating a new folder or adding a location by clicking the buttons ()/ () at the top of the dialog.

8.2.1 Table outputs

Some analyses also generate a table with results, and for these analyses the last step looks like figure 8.6.



Convert DNA to RNA

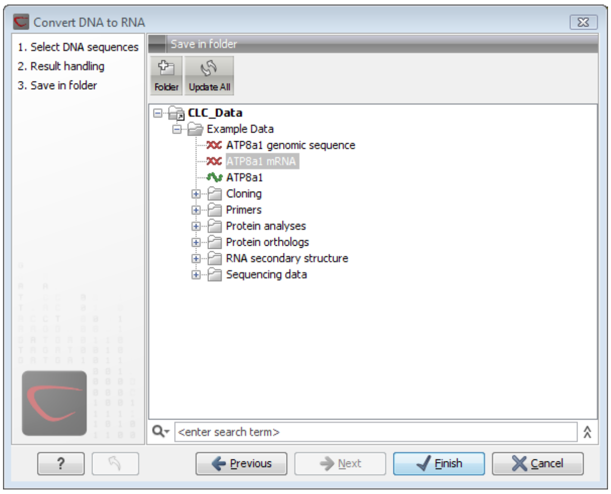
1. Select DNA sequences
2. Result handling

Result handling

Result handling

● Open
○ Save

Figure 8.4: *The last step of the analyses exemplified by Translate DNA to RNA.*



Convert DNA to RNA

1. Select DNA sequences
2. Result handling
3. Save in folder

Save in folder

- CLC_Data
  - Example Data
    - ATP8a1 genomic sequence
    - ATP8a1 mRNA
    - ATP8a1
    - Cloning
    - Primers
    - Protein analyses
    - Protein orthologs
    - RNA secondary structure
    - Sequencing data

Figure 8.5: *Specify a folder for the results of the analysis.*

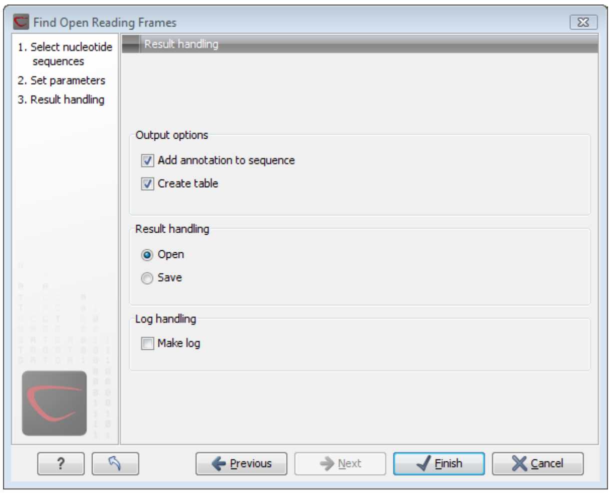
In addition to the **Open** and **Save** options you can also choose whether the result of the analysis should be added as annotations on the sequence or shown on a table. If both options are selected, you will be able to click the results in the table and the corresponding region on the sequence will be selected.

If you choose to add annotations to the sequence, they can be removed afterwards by clicking **Undo** (↶) in the **Toolbar**.

8.2.2 Batch log

For some analyses, there is an extra option in the final step to create a log of the batch process (see e.g. figure 8.6). This log will be created in the beginning of the process and continually updated with information about the results. See an example of a log in figure 8.7. In this example, the log displays information about how many open reading frames were found.

The log will either be saved with the results of the analysis or opened in a view with the results, depending on how you chose to handle the results.



Find Open Reading Frames

1. Select nucleotide sequences
2. Set parameters
3. Result handling

Result handling

Output options

☑ Add annotation to sequence

☑ Create table

Result handling

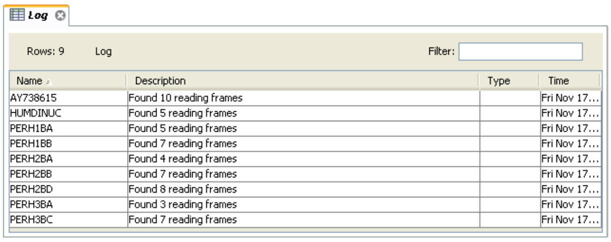
● Open

○ Save

Log handling

□ Make log

Figure 8.6: *Analyses which also generate tables.*



Log

Rows: 9 Log Filter:

Name	Description	Type	Time
AY738615	Found 10 reading frames		Fri Nov 17...
HUMDINUC	Found 5 reading frames		Fri Nov 17...
PERH1BA	Found 5 reading frames		Fri Nov 17...
PERH1BB	Found 7 reading frames		Fri Nov 17...
PERH2BA	Found 4 reading frames		Fri Nov 17...
PERH2BB	Found 7 reading frames		Fri Nov 17...
PERH2BD	Found 8 reading frames		Fri Nov 17...
PERH3BA	Found 3 reading frames		Fri Nov 17...
PERH3BC	Found 7 reading frames		Fri Nov 17...

Figure 8.7: *An example of a batch log when finding open reading frames.*

8.3 Working with tables

Tables are used in a lot of places in the *CLC Cancer Research Workbench*. There are some general features for all tables, irrespective of their contents, that are described here.

Figure 8.8 shows an example of a typical table. This is the table result of **Find Open Reading Frames** (). We use this table as an example to illustrate concepts relevant to all kinds of tables.

Table viewing options

Options relevant to the view of the table can be configured in the **Side Panel** on the right.

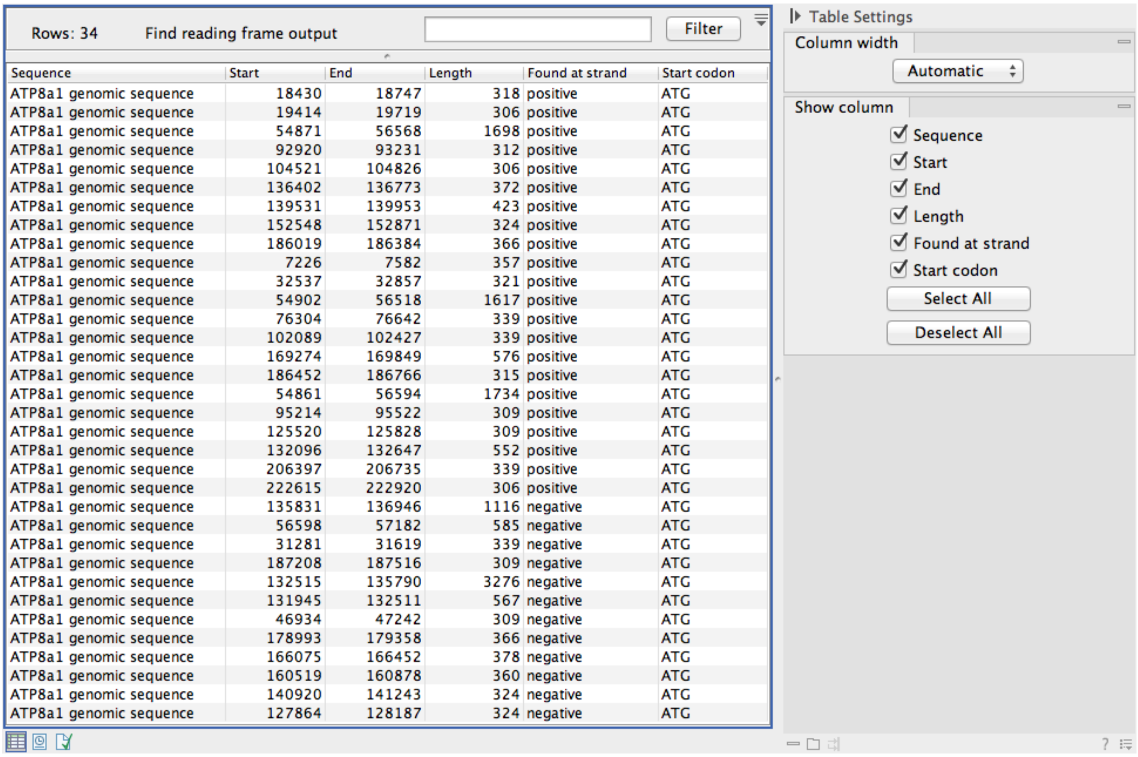
For example, the columns that can be dispalyed in the table are listed in the section called **Show column**. The checkboxes allow you to see or hide any of the available columns for that table.

The Column width can be set to **Automatic** or **Manual**. By default, the first time you open a table, it will be set to **Automatic**. The default selected columns are hereby resized to fit the width of the viewing area. When changing to the **Manual** option, column widths will adjust to the actual header size, and each column size can subsequently by adjusted manually. When the table content exceeds the size of the viewing area, a horizontal scroll becomes available for navigation across the columns.

Sorting tables

You can **sort** table according to the values of a particular column by clicking a column header. (Pressing Ctrl - ⌘ on Mac - while you click will refine the existing sorting).

Clicking once will sort in ascending order. A second click will change the order to descending. A



Rows: 34 Find reading frame output

Sequence	Start	End	Length	Found at strand	Start codon
ATP8a1 genomic sequence	18430	18747	318	positive	ATG
ATP8a1 genomic sequence	19414	19719	306	positive	ATG
ATP8a1 genomic sequence	54871	56568	1698	positive	ATG
ATP8a1 genomic sequence	92920	93231	312	positive	ATG
ATP8a1 genomic sequence	104521	104826	306	positive	ATG
ATP8a1 genomic sequence	136402	136773	372	positive	ATG
ATP8a1 genomic sequence	139531	139953	423	positive	ATG
ATP8a1 genomic sequence	152548	152871	324	positive	ATG
ATP8a1 genomic sequence	186019	186384	366	positive	ATG
ATP8a1 genomic sequence	7226	7582	357	positive	ATG
ATP8a1 genomic sequence	32537	32857	321	positive	ATG
ATP8a1 genomic sequence	54902	56518	1617	positive	ATG
ATP8a1 genomic sequence	76304	76642	339	positive	ATG
ATP8a1 genomic sequence	102089	102427	339	positive	ATG
ATP8a1 genomic sequence	169274	169849	576	positive	ATG
ATP8a1 genomic sequence	186452	186766	315	positive	ATG
ATP8a1 genomic sequence	54861	56594	1734	positive	ATG
ATP8a1 genomic sequence	95214	95522	309	positive	ATG
ATP8a1 genomic sequence	125520	125828	309	positive	ATG
ATP8a1 genomic sequence	132096	132647	552	positive	ATG
ATP8a1 genomic sequence	206397	206735	339	positive	ATG
ATP8a1 genomic sequence	222615	222920	306	positive	ATG
ATP8a1 genomic sequence	135831	136946	1116	negative	ATG
ATP8a1 genomic sequence	56598	57182	585	negative	ATG
ATP8a1 genomic sequence	31281	31619	339	negative	ATG
ATP8a1 genomic sequence	187208	187516	309	negative	ATG
ATP8a1 genomic sequence	132515	135790	3276	negative	ATG
ATP8a1 genomic sequence	131945	132511	567	negative	ATG
ATP8a1 genomic sequence	46934	47242	309	negative	ATG
ATP8a1 genomic sequence	178993	179358	366	negative	ATG
ATP8a1 genomic sequence	166075	166452	378	negative	ATG
ATP8a1 genomic sequence	160519	160878	360	negative	ATG
ATP8a1 genomic sequence	140920	141243	324	negative	ATG
ATP8a1 genomic sequence	127864	128187	324	negative	ATG

Table Settings

Column width: Automatic

Show column

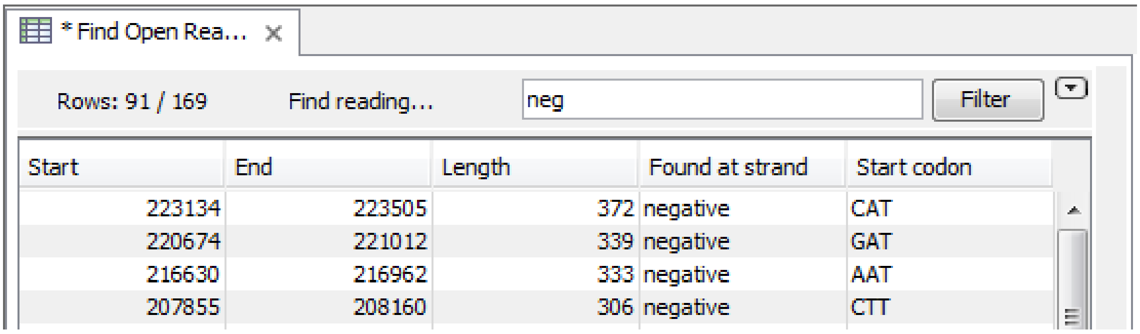
☑ Sequence
☑ Start
☑ End
☑ Length
☑ Found at strand
☑ Start codon

Figure 8.8: *A table showing the results of an open reading frames analysis.*

third click will set the order back its original order.

8.3.1 Filtering tables

The final concept to introduce is **Filtering**. The table filter as an advanced and a simple mode. The simple mode is the default and is applied simply by typing text or numbers (see an example in figure 8.9).[1]



Rows: 91 / 169 Find reading... neg

Start	End	Length	Found at strand	Start codon
223134	223505	372	negative	CAT
220674	221012	339	negative	GAT
216630	216962	333	negative	AAT
207855	208160	306	negative	CTT

Figure 8.9: *Typing "neg" in the filter in simple mode.*

Typing "neg" in the filter will only show the rows where "neg" is part of the text in any of the columns (also the ones that are not shown). The text does not have to be in the beginning, thus "ega" would give the same result. This simple filter works fine for fast, textual and non-complicated filtering and searching.

[1] Note that for tables with more than 10000 rows, you have to actually click the **Filter** button for the table to take effect.

However, if you wish to make use of numerical information or make more complex filters, you can switch to the advanced mode by clicking the **Advanced filter** (▾) button. The advanced filter is structure in a different way: First of all, you can have more than one criterion in the filter. Criteria can be added or removed by clicking the **Add** (➕) or **Remove** (✖) buttons. At the top, you can choose whether all the criteria should be fulfilled (**Match all**), or if just one of the needs to be fulfilled (**Match any**).

For each filter criterion, you first have to select which column it should apply to. Next, you choose an operator. For numbers, you can choose between:

- **=** (equal to)
- **<** (smaller than)
- **>** (greater than)
- **<>** (not equal to)
- **abs. value <** (absolute value smaller than. This is useful if it doesn't matter whether the number is negative or positive)
- **abs. value >** (absolute value greater than. This is useful if it doesn't matter whether the number is negative or positive)

Note, that the number of digits displayed is a formatting option which can be set in the View Preferences. The true number may well be (slightly) larger. This behaviour can lead to problems when filtering on exact matches using the **=** (equal to) operator on numbers. Instead, users are advised to use two filters of inequalities (**<** (smaller than) and **>** (greater than)) delimiting a (small) interval around the target value.

For text-based columns, you can choose between:

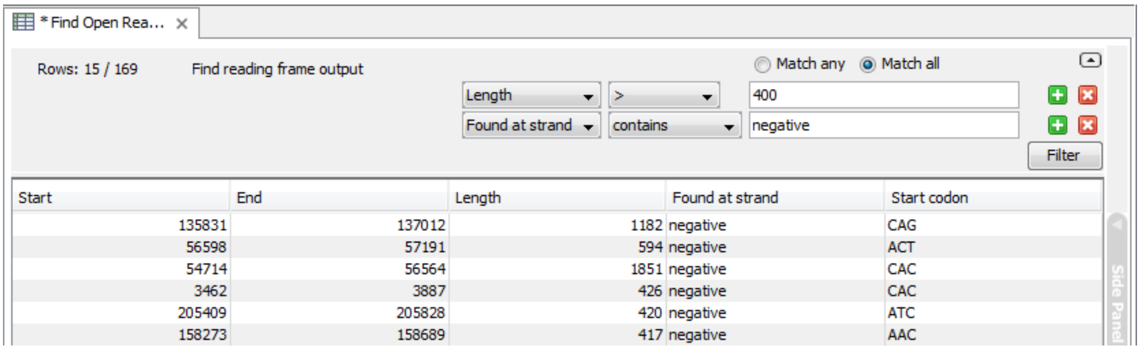
- **starts with** (the text starts with your search term)
- **contains** (the text does not have to be in the beginning)
- **doesn't contain**
- **=** (the whole text in the table cell has to match, also lower/upper case)
- ≠ (the text in the table cell has to not match)
- **is in list** (The text in the table cell has to match one of the items of the list. Items are separated by comma, semicolon, or space. This filter is case-insensitive)

Once you have chosen an operator, you can enter the text or numerical value to use.

If you wish to reset the filter, simply remove (✖) all the search criteria. Note that the last one will not disappear - it will be reset and allow you to start over.

Figure 8.10 shows an example of an advanced filter which displays the open reading frames larger than 400 that are placed on the negative strand.

Both for the simple and the advanced filter, there is a counter at the upper left corner which tells you the number of rows that pass the filter (91 in figure 8.9 and 15 in figure 8.10).



Rows: 15 / 169 Find reading frame output

○ Match any ● Match all

Length > 400

Found at strand contains negative

Start	End	Length	Found at strand	Start codon
135831	137012	1182	negative	CAG
56598	57191	594	negative	ACT
54714	56564	1851	negative	CAC
3462	3887	426	negative	CAC
205409	205828	420	negative	ATC
158273	158689	417	negative	AAC

Figure 8.10: *The advanced filter showing open reading frames larger than 400 that are placed on the negative strand.*

Chapter 9

Viewing and editing sequences

Contents

CLC Cancer Research Workbench offers five different ways of viewing and editing single sequences as described in the first five sections of this chapter. Furthermore, this chapter also explains how to create a new sequence and how to gather several sequences in a sequence list.

9.1 View sequence

When you double-click a sequence in the **Navigation Area**, the sequence will open automatically, and you will see the nucleotides or amino acids. The zoom options described in section 2.2 allow you to e.g. zoom out in order to see more of the sequence in one view. There are a number of options for viewing and editing the sequence which are all described in this section.

9.1.1 Sequence settings in Side Panel

Each view of a sequence has a **Side Panel** located at the right side of the view (see figure 9.1.

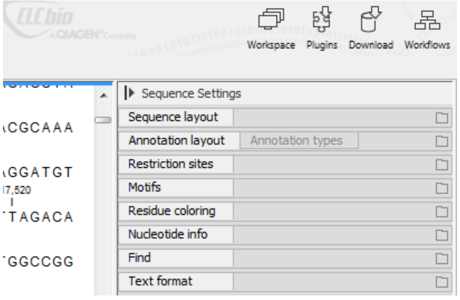


Figure 9.1: *Overview of the Side Panel which is always shown to the right of a view.*

When you make changes in the **Side Panel** the view of the sequence is instantly updated. To show or hide the **Side Panel**:

select the View | Ctrl + U

or **Click the (|▸) at the top right corner of the Side Panel to hide | Click the (◂|) to the right to show**

Below, each group of settings will be explained. Some of the preferences are not the same for nucleotide and protein sequences, but the differences will be explained for each group of settings.

Note! When you make changes to the settings in the **Side Panel**, they are not automatically saved when you save the sequence. Click **Save/restore Settings** (☰) to save the settings (see section 4.5 for more information).

Sequence Layout

These preferences determine the overall layout of the sequence:

- **Spacing.** Inserts a space at a specified interval:
  - **No spacing.** The sequence is shown with no spaces.
  - **Every 10 residues.** There is a space every 10 residues, starting from the beginning of the sequence.
  - **Every 3 residues, frame 1.** There is a space every 3 residues, corresponding to the reading frame starting at the first residue.
  - **Every 3 residues, frame 2.** There is a space every 3 residues, corresponding to the reading frame starting at the second residue.
  - **Every 3 residues, frame 3.** There is a space every 3 residues, corresponding to the reading frame starting at the third residue.
- **Wrap sequences.** Shows the sequence on more than one line.
  - **No wrap.** The sequence is displayed on one line.
  - **Auto wrap.** Wraps the sequence to fit the width of the view, not matter if it is zoomed in our out (displays minimum 10 nucleotides on each line).

  - **Fixed wrap.** Makes it possible to specify when the sequence should be wrapped. In the text field below, you can choose the number of residues to display on each line.

- **Double stranded.** Shows both strands of a sequence (only applies to DNA sequences).
- **Numbers on sequences.** Shows residue positions along the sequence. The starting point can be changed by setting the number in the field below. If you set it to e.g. 101, the first residue will have the position of -100. This can also be done by right-clicking an annotation and choosing **Set Numbers Relative to This Annotation**.
- **Numbers on plus strand.** Whether to set the numbers relative to the positive or the negative strand in a nucleotide sequence (only applies to DNA sequences).
- **Lock numbers.** When you scroll vertically, the position numbers remain visible. (Only possible when the sequence is not wrapped.)
- **Lock labels.** When you scroll horizontally, the label of the sequence remains visible.
- **Sequence label.** Defines the label to the left of the sequence.
  - Name (this is the default information to be shown).
  - Accession (sequences downloaded from databases like GenBank have an accession number).
  - Latin name.
  - Latin name (accession).
  - Common name.
  - Common name (accession).
- **Matching residues as dots** Residues in aligned sequences identical to residues in the first (reference) sequence will be presented as dots. An option that is only available for "Alignments" and "Read mappings".

Annotation Layout and Annotation Types

See section 9.3.1.

Restriction sites

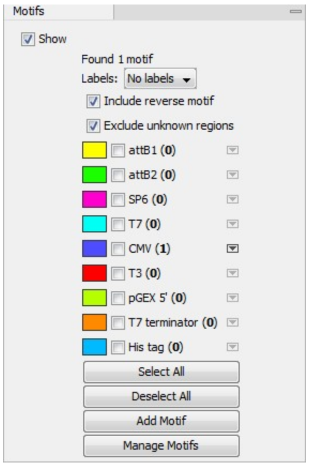
See section 9.1.2.

Motifs

In the **Side Panel** of sequence views, there is a group called **Motifs** (see figure 9.2).

The Workbench will look for the listed motifs in the sequence that is open and by clicking the check box next to the motif it will be shown in the view as illustrated in figure 9.3.

This case shows the CMV promoter primer sequence which is one of the pre-defined motifs in *CLC Cancer Research Workbench*. The motif is per default shown as a faded arrow with no text. The direction of the arrow indicates the strand of the motif.



☑ Show

Found 1 motif

Labels: No labels

☑ Include reverse motif

☑ Exclude unknown regions

🟨 □ attB1 (0)

🟩 □ attB2 (0)

🟪 □ SP6 (0)

🟦 □ T7 (0)

🟦 □ CMV (1)

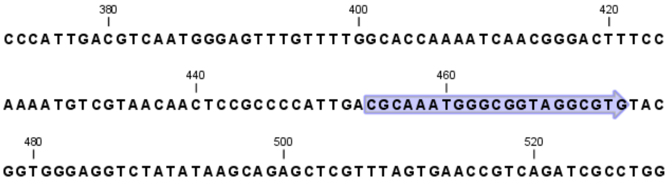
🟥 □ T3 (0)

🟩 □ pGEX 5' (0)

🟧 □ T7 terminator (0)

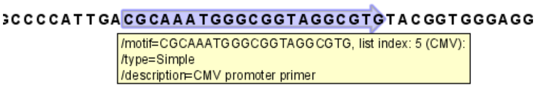
🟦 □ His tag (0)

Figure 9.2: *Dynamic motifs in the Side Panel.*



```
CCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCC
AAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTAC
GGTGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATCGCCTGG
```

Figure 9.3: *Showing dynamic motifs on the sequence.*



```
GCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGG
/motif=CGCAAATGGGCGGTAGGCGTG, list index: 5 (CMV):
/type=Simple
/description=CMV promoter primer
```

Figure 9.4: *Showing dynamic motifs on the sequence.*

Placing the mouse cursor on the arrow will display additional information about the motif as illustrated in figure 9.4.

To add **Labels** to the motif, select the **Flag** or **Stacked** option. They will put the name of the motif as a flag above the sequence. The stacked option will stack the labels when there is more than one motif so that all labels are shown.

Residue coloring

These preferences make it possible to color both the residue letter and set a background color for the residue.

- **Non-standard residues.** For nucleotide sequences this will color the residues that are not C, G, A, T or U. For amino acids only B, Z, and X are colored as non-standard residues.
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Rasmol colors.** Colors the residues according to the Rasmol color scheme.
  See `http://www.openrasmol.org/doc/rasmol.html`
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.

  - **Background color.** Sets the background color of the residues. Click the color box to change the color.

- **Polarity colors (only protein).** Colors the residues according to the following categories:
  - **Green** neutral, polar
  - **Black** neutral, nonpolar
  - **Red** acidic, polar
  - **Blue** basic ,polar
  - As with other options, you can choose to set or change the coloring for either the residue letter or its background:
    - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
    - **Background color.** Sets the background color of the residues. Click the color box to change the color.

- **Trace colors (only DNA).** Colors the residues according to the color conventions of chromatogram traces: A=green, C=blue, G=black, and T=red.
  - **Foreground color.** Sets the color of the letter.
  - **Background color.** Sets the background color of the residues.

Nucleotide info

These preferences only apply to nucleotide sequences.

- **Color space encoding.** Lets you define a few settings for how the colors should appear.

  Infer encoding This is used if you want to display the colors for non-color space sequence (e.g. a reference sequence). The colors are then simply inferred from the sequence.

  Show corrections This is only relevant for mapping results - it will show where the mapping process has detected color errors. An example of a color error is shown in figure 20.18.

  Hide unaligned This option determines whether color for the unaligned ends of reads should be displayed. It also controls whether colors should be shown for gaps. The idea behind this is that these color dots will interfere with the color alignment, so it is possible to turn them off.

- **Translation.** Displays a translation into protein just below the nucleotide sequence. Depending on the zoom level, the amino acids are displayed with three letters or one letter.
  - **Frame.** Determines where to start the translation.
    - **ORF/CDS**. If the sequence is annotated, the translation will follow the CDS or ORF annotations. If annotations overlap, only one translation will be shown. If only one annotation is visible, the Workbench will attempt to use this annotation to mark the start and stop for the translation. In cases where this is not possible, the first annotation will be used (i.e. the one closest to the 5' end of the sequence).

        - **Selection.** This option will only take effect when you make a selection on the sequence. The translation will start from the first nucleotide selected. Making a new selection will automatically display the corresponding translation. Read more about selecting in section 9.1.3.
        - **+1 to -1.** Select one of the six reading frames.
        - **All forward/All reverse.** Shows either all forward or all reverse reading frames.
        - **All.** Select all reading frames at once. The translations will be displayed on top of each other.
    - **Table.** The translation table to use in the translation.
    - **Only AUG start codons.** For most genetic codes, a number of codons can be start codons (TTG, CTG, or ATG). These will be colored green, unless selecting the "Only AUG start codons" option, which will result in only the AUG codons colored in green.
    - **Single letter codes.** Choose to represent the amino acids with a single letter instead of three letters.

- **Trace data.** See section 29.1.

- **Quality scores.** For sequencing data containing quality scores, the quality score information can be displayed along the sequence.

    - **Show as probabilities.** Converts quality scores to error probabilities on a 0-1 scale, i.e. not log-transformed.
    - **Foreground color.** Colors the letter using a gradient, where the left side color is used for low quality and the right side color is used for high quality. The sliders just above the gradient color box can be dragged to highlight relevant levels. The colors can be changed by clicking the box. This will show a list of gradients to choose from.
    - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
    - **Graph.** The quality score is displayed on a graph (Learn how to export the data behind the graph in section 6.7).
        - **Height.** Specifies the height of the graph.
        - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
        - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. For Colors, the color box is replaced by a gradient color box as described under Foreground color.

- **G/C content.** Calculates the G/C content of a part of the sequence and shows it as a gradient of colors or as a graph below the sequence.

    - **Window length.** Determines the length of the part of the sequence to calculate. A window length of 9 will calculate the G/C content for the nucleotide in question plus the 4 nucleotides to the left and the 4 nucleotides to the right. A narrow window will focus on small fluctuations in the G/C content level, whereas a wider window will show fluctuations between larger parts of the sequence.
    - **Foreground color.** Colors the letter using a gradient, where the left side color is used for low levels of G/C content and the right side color is used for high levels of G/C content. The sliders just above the gradient color box can be dragged to highlight relevant levels of G/C content. The colors can be changed by clicking the box. This will show a list of gradients to choose from.

  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** The G/C content level is displayed on a graph (Learn how to export the data behind the graph in section 6.7).
    - **Height.** Specifies the height of the graph.
    - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
    - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. For Colors, the color box is replaced by a gradient color box as described under Foreground color.

- **Secondary structure.** Allows you to choose how to display a symbolic representation of the secondary structure along the sequence.

Protein info

These preferences only apply to proteins. The first nine items are different hydrophobicity scales.

- **Kyte-Doolittle.** The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size used. Short window sizes of 5-7 generally work well for predicting putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.

- **Cornette.** Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.

- **Engelman.** The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.

- **Eisenberg.** The Eisenberg scale is a normalized consensus hydrophobicity scale which shares many features with the other hydrophobicity scales [Eisenberg et al., 1984].

- **Rose.** The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.

- **Janin.** This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].

- **Hopp-Woods.** Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].

- **Welling.** [Welling et al., 1985] Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.
- **Kolaskar-Tongaonkar.** A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.
- **Surface Probability.** Display of surface probability based on the algorithm by [Emini et al., 1985]. This algorithm has been used to identify antigenic determinants on the surface of proteins.
- **Chain Flexibility.** Display of backbone chain flexibility based on the algorithm by [Karplus and Schulz, 1985]. It is known that chain flexibility is an indication of a putative antigenic determinant.

Find

The Find function can be used for searching the sequence and is invoked by pressing Ctrl + Shift + F (⌘ + Shift + F on Mac). Initially, specify the 'search term' to be found, select the type of search (see various options in the following) and finally click on the Find button. The first occurrence of the search term will then be highlighted. Clicking the find button again will find the next occurrence and so on. If the search string is found, the corresponding part of the sequence will be selected.

- **Search term.** Enter the text or number to search for. The search function does not discriminate between lower and upper case characters.
- **Sequence search.** Search the nucleotides or amino acids. For amino acids, the single letter abbreviations should be used for searching. The sequence search also has a set of advanced search parameters:
  - Include negative strand. This will search on the negative strand as well.
  - Treat ambiguous characters as wildcards in search term. If you search for e.g. ATN, you will find both ATG and ATC. If you wish to find literally exact matches for ATN (i.e. only find ATN - not ATG), this option should not be selected.
  - Treat ambiguous characters as wildcards in sequence. If you search for e.g. ATG, you will find both ATG and ATN. If you have large regions of Ns, this option should not be selected.

  Note that if you enter a position instead of a sequence, it will automatically switch to position search.
- **Annotation search.** Searches the annotations on the sequence. The search is performed both on the labels of the annotations, but also on the text appearing in the tooltip that you see when you keep the mouse cursor fixed. If the search term is found, the part of the sequence corresponding to the matching annotation is selected. Below this option you can choose to search for translations as well. Sequences annotated with coding regions often have the translation specified which can lead to undesired results.

- **Position search.** Finds a specific position on the sequence. In order to find an interval, e.g. from position 500 to 570, enter "500..570" in the search field. This will make a selection from position 500 to 570 (both included). Notice the two periods (..) between the start an end number (see section 9.3.2). If you enter positions including thousands separators like 123,345, the comma will just be ignored and it would be equivalent to entering 123345.
- **Include negative strand.** When searching the sequence for nucleotides or amino acids, you can search on both strands.
- **Name search.** Searches for sequence names. This is useful for searching sequence lists, mapping results and BLAST results.

This concludes the description of the **View Preferences**. Next, the options for selecting and editing sequences are described.

Text format

These preferences allow you to adjust the format of all the text in the view (both residue letters, sequence name and translations if they are shown).

- **Text size.** Five different sizes.
- **Font.** Shows a list of Fonts available on your computer.
- **Bold residues.** Makes the residues bold.

9.1.2 Restriction sites in the Side Panel

Please see the CLC Genomics Workbench manual in the "Cloning and cutting" chapter, section "Restriction site analysis": http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Dynamic_restriction_sites.html.

9.1.3 Selecting parts of the sequence

You can select parts of a sequence:

> **Click Selection ( ) in Toolbar | Press and hold down the mouse button on the sequence where you want the selection to start | move the mouse to the end of the selection while holding the button | release the mouse button**

Alternatively, you can search for a specific interval using the find function described above.

If you have made a selection and wish to adjust it:

> **drag the edge of the selection (you can see the mouse cursor change to a horizontal arrow**
>
> or **press and hold the Shift key while using the right and left arrow keys to adjust the right side of the selection.**

If you wish to select the entire sequence:

> **double-click the sequence name to the left**

Selecting several parts at the same time (multiselect)

You can select several parts of sequence by holding down the **Ctrl** button while making selections. Holding down the **Shift** button lets you extend or reduce an existing selection to the position you clicked.

To select a part of a sequence covered by an annotation:

right-click the annotation | Select annotation

or **double-click the annotation**

To select a fragment between two restriction sites that are shown on the sequence:

double-click the sequence between the two restriction sites

(Read more about restriction sites in section 9.1.2.)

Open a selection in a new view

A selection can be opened in a new view and saved as a new sequence:

right-click the selection | Open selection in New View ()

This opens the annotated part of the sequence in a new view. The new sequence can be saved by dragging the tab of the sequence view into the **Navigation Area**.

The process described above is also the way to manually translate coding parts of sequences (CDS) into protein. You simply translate the new sequence into protein. This is done by:

right-click the tab of the new sequence | Toolbox | Nucleotide Analysis ()| Translate to Protein ()

A selection can also be copied to the clipboard and pasted into another program:

make a selection | Ctrl + C (⌘ + C on Mac)

Note! The annotations covering the selection will not be copied.

A selection of a sequence can be edited as described in the following section.

9.1.4 Editing the sequence

When you make a selection, it can be edited by:

right-click the selection | Edit Selection ()

A dialog appears displaying the sequence. You can add, remove or change the text and click **OK**. The original selected part of the sequence is now replaced by the sequence entered in the dialog. This dialog also allows you to paste text into the sequence using Ctrl + V (⌘ + V on Mac).

If you delete the text in the dialog and press **OK**, the selected text on the sequence will also be deleted. Another way to delete a part of the sequence is to:

right-click the selection | Delete Selection ()

If you wish to correct only one residue, this is possible by simply making the selection cover only one residue and then type the new residue.

Note When editing annotated nucleotide sequences, the annotation content is not updated automatically (but its position is). Please refer to section 9.3.3 for details on annotation editing. Before exporting annotated nucleotide sequences in GenBank format, ensure that the annotations in the Annotations Table reflect the edits that have been made to the sequence.

9.1.5 Sequence region types

The various annotations on sequences cover parts of the sequence. Some cover an interval, some cover intervals with unknown endpoints, some cover more than one interval etc. In the following, all of these will be referred to as *regions*. Regions are generally illustrated by markings (often arrows) on the sequences. An arrow pointing to the right indicates that the corresponding region is located on the positive strand of the sequence. Figure 9.5 is an example of three regions with separate colors.

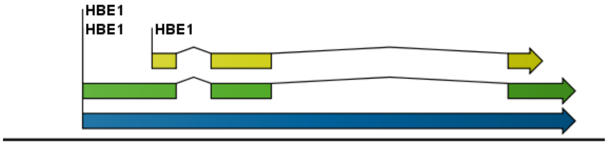


Figure 9.5: *Three regions on a human beta globin DNA sequence (HUMHBB).*

Figure 9.6 shows an artificial sequence with all the different kinds of regions.

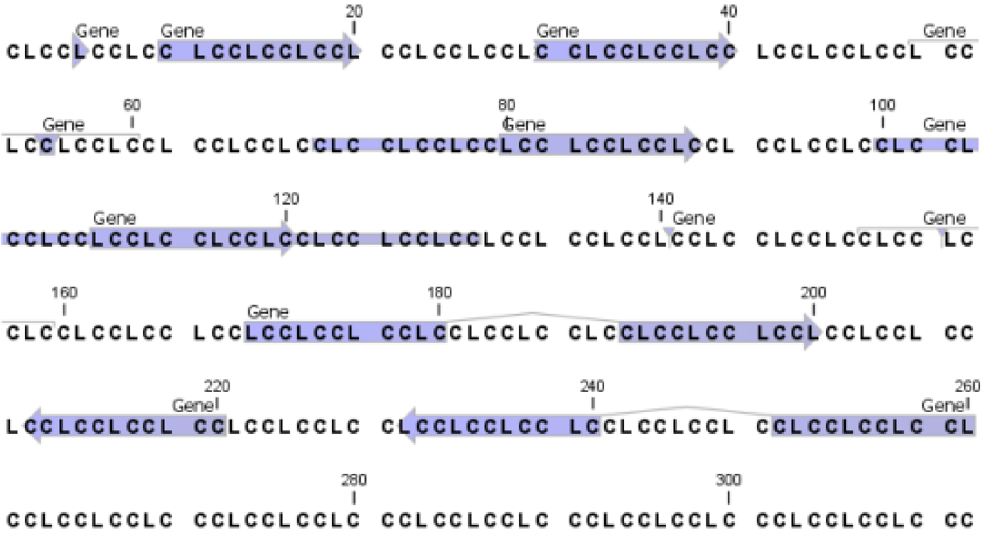


Figure 9.6: *Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning somewhere between 70 and 80 inclusive and ending at 90 inclusive, Region #6: A range of residues beginning somewhere between 100 and 110 inclusive and ending somewhere between 120 and 130 inclusive, Region #7: A site between residues 140 and 141, Region #8: A site between two residues somewhere between 150 and 160 inclusive, Region #9: A region that covers ranges from 170 to 180 inclusive and 190 to 200 inclusive, Region #10: A region on negative strand that covers ranges from 210 to 220 inclusive, Region #11: A region on negative strand that covers ranges from 230 to 240 inclusive and 250 to 260 inclusive.*

9.2 Circular DNA

A sequence can be shown as a circular molecule:

> **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Select "Circular View" ( ) **

or **If the sequence is already open | Click "Show Circular View" ( ) at the lower left part of the view**

This will open a view of the molecule similar to the one in figure 9.7.

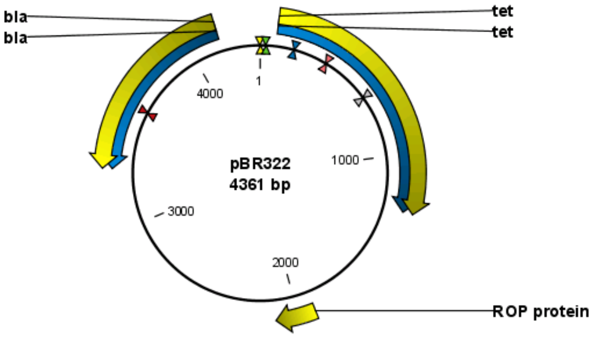


Figure 9.7: *A molecule shown in a circular view.*

This view of the sequence shares some of the properties of the linear view of sequences as described in section 9.1, but there are some differences. The similarities and differences are listed below:

- **Similarities**:
  - The editing options.
  - Options for adding, editing and removing annotations.
  - **Restriction Sites**, **Annotation Types**, **Find** and **Text Format** preferences groups.
- **Differences**:
  - In the **Sequence Layout** preferences, only the following options are available in the circular view: **Numbers on plus strand**, **Numbers on sequence** and **Sequence label**.
  - You cannot zoom in to see the residues in the circular molecule. If you wish to see these details, split the view with a linear view of the sequence
  - In the **Annotation Layout**, you also have the option of showing the labels as **Stacked**. This means that there are no overlapping labels and that all labels of both annotations and restriction sites are adjusted along the left and right edges of the view.

9.2.1 Using split views to see details of the circular molecule

In order to see the nucleotides of a circular molecule you can open a new view displaying a circular view of the molecule:

> **Press and hold the Ctrl button (⌘ on Mac) | click Show Sequence ( ) at the bottom of the view**

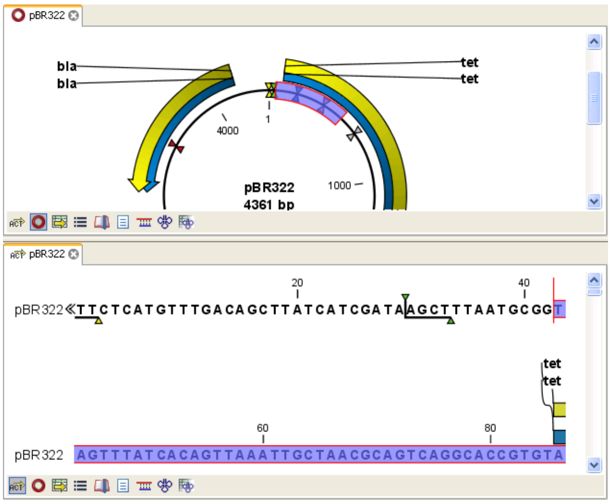

pBR322
bla
bla
tet
tet
4000
1
1000
pBR322
4361 bp
pBR322
20
40
pBR322
TTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATGCGGT
tet
tet
60
80
pBR322
AGTTTATCACAGTTAAATTGCTAACGCAGTCAGGCACCGTGTA

Figure 9.8: *Two views showing the same sequence. The bottom view is zoomed in.*

This will open a linear view of the sequence below the circular view. When you zoom in on the linear view you can see the residues as shown in figure 9.8.

Note! If you make a selection in one of the views, the other view will also make the corresponding selection, providing an easy way for you to focus on the same region in both views.

9.2.2 Mark molecule as circular and specify starting point

You can mark a DNA molecule as circular by right-clicking its name in either the sequence view or the circular view. In the right-click menu you can also make a circular molecule linear. A circular molecule displayed in the normal sequence view, will have the sequence ends marked with a ï¿½.

The starting point of a circular sequence can be changed by:

> **make a selection starting at the position that you want to be the new starting point | right-click the selection | Move Starting Point to Selection Start**

Note! This can only be done for sequence that have been marked as circular.

9.3 Working with annotations

Note! This section only applies to sequences that is not in track format e.g. sequences from Sanger sequencing.

Annotations provide information about specific regions of a sequence. A typical example is the annotation of a gene on a genomic DNA sequence.

Annotations derive from different sources:

- Sequences downloaded from databases like GenBank are annotated.
- In some of the data formats that can be imported into *CLC Cancer Research Workbench*, sequences can have annotations (GenBank, EMBL and Swiss-Prot format).
- The result of a number of analyses in *CLC Cancer Research Workbench* are annotations on the sequence (e.g. finding open reading frames and restriction map analysis).

- You can manually add annotations to a sequence (described in the section 9.3.2).

Note! Annotations are included if you export the sequence in GenBank, Swiss-Prot, EMBL or CLC format. When exporting in other formats, annotations are not preserved in the exported file.

9.3.1 Viewing annotations

Annotations can be viewed in a number of different ways:

- As arrows or boxes in all views displaying sequences (sequence lists, alignments etc)
- In the table of annotations ( ).
- In the text view of sequences ( )

In the following sections, these view options will be described in more detail. In all the views except the text view ( ), annotations can be added, modified and deleted. This is described in the following sections.

View Annotations in sequence views

Figure 9.9 shows an annotation displayed on a sequence.

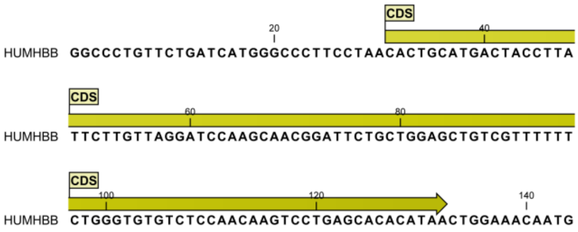

CDS
20
40
HUMHBB GGCCCTGTTCTGATCATGGGCCCTTCCTAACACTGCATGACTACCTTA
CDS
60
80
HUMHBB TTCTTGTTAGGATCCAAGCAACGGATTCTGCTGGAGCTGTCGTTTTTT
CDS
100
120
140
HUMHBB CTGGGTGTGTCTCCAACAAGTCCTGAGCACACATAACTGGAAACAATG

Figure 9.9: *An annotation showing a coding region on a genomic dna sequence.*

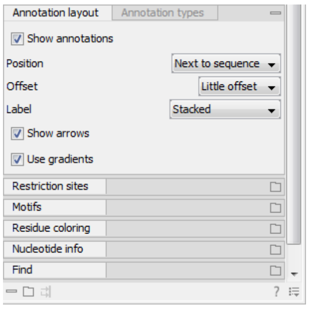
The various sequence views listed in section 9.3.1 have different default settings for showing annotations. However, they all have two groups in the **Side Panel** in common:

- **Annotation Layout**
- **Annotation Types**

The two groups are shown in figure 9.10.

In the **Annotation layout** group, you can specify how the annotations should be displayed (notice that there are some minor differences between the different sequence views):

- **Show annotations.** Determines whether the annotations are shown.
- **Position.**
  - **On sequence.** The annotations are placed on the sequence. The residues are visible through the annotations (if you have zoomed in to 100%).



☑ Show annotations

Position: Next to sequence

Offset: Little offset

Label: Stacked

☑ Show arrows

☑ Use gradients

Figure 9.10: *The annotation layout in the Side Panel. The annotation types can be shown by clicking on the "Annotation types" tab.*

  - **Next to sequence.** The annotations are placed above the sequence.
  - **Separate layer.** The annotations are placed above the sequence and above restriction sites (only applicable for nucleotide sequences).

- **Offset.** If several annotations cover the same part of a sequence, they can be spread out.
  - **Piled.** The annotations are piled on top of each other. Only the one at front is visible.
  - **Little offset.** The annotations are piled on top of each other, but they have been offset a little.
  - **More offset.** Same as above, but with more spreading.
  - **Most offset.** The annotations are placed above each other with a little space between. This can take up a lot of space on the screen.

- **Label.** The name of the annotation can shown as a label. Additional information about the sequence is shown if you place the mouse cursor on the annotation and keep it still.
  - **No labels.** No labels are displayed.
  - **On annotation.** The labels are displayed in the annotation's box.
  - **Over annotation.** The labels are displayed above the annotations.
  - **Before annotation.** The labels are placed just to the left of the annotation.
  - **Flag.** The labels are displayed as flags at the beginning of the annotation.
  - **Stacked.** The labels are offset so that the text of all labels is visible. This means that there is varying distance between each sequence line to make room for the labels.

- **Show arrows.** Displays the end of the annotation as an arrow. This can be useful to see the orientation of the annotation (for DNA sequences). Annotations on the negative strand will have an arrow pointing to the left.

- **Use gradients.** Fills the boxes with gradient color.

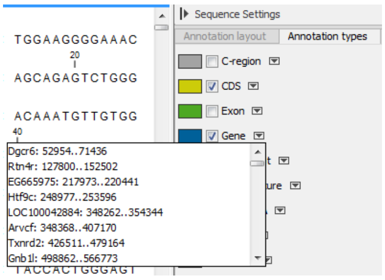
In the **Annotation types** group, you can choose which kinds of annotations that should be displayed. This group lists all the types of annotations that are attached to the sequence(s) in the view. For sequences with many annotations, it can be easier to get an overview if you deselect the annotation types that are not relevant.

Unchecking the checkboxes in the **Annotation layout** will not remove this type of annotations them from the sequence - it will just hide them from the view.

Besides selecting which types of annotations that should be displayed, the **Annotation types** group is also used to change the color of the annotations on the sequence. Click the colored square next to the relevant annotation type to change the color.

This will display a dialog with three tabs: Swatches, HSB, and RGB. They represent three different ways of specifying colors. Apply your settings and click **OK**. When you click **OK**, the color settings cannot be reset. The **Reset** function only works for changes made before pressing **OK**.

Furthermore, the **Annotation types** can be used to easily browse the annotations by clicking the small button (▾) next to the type. This will display a list of the annotations of that type (see figure 9.11).



🟫 □ C-region ▾
🟨 ☑ CDS ▾
🟩 □ Exon ▾
🟦 ☑ Gene ▾

Dgcr6: 52954..71436
Rtn4r: 127800..152502
EG665975: 217973..220441
Htf9c: 248977..253596
LOC100042884: 348262..354344
Arvcf: 348368..407170
Txnrd2: 426511..479164
Gnb1l: 498862..566773

Figure 9.11: *Browsing the gene annotations on a sequence.*

Clicking an annotation in the list will select this region on the sequence. In this way, you can quickly find a specific annotation on a long sequence.

View Annotations in a table

Annotations can also be viewed in a table:

> **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Annotation Table ( )**

or **If the sequence is already open | Click Show Annotation Table ( ) at the lower left part of the view**

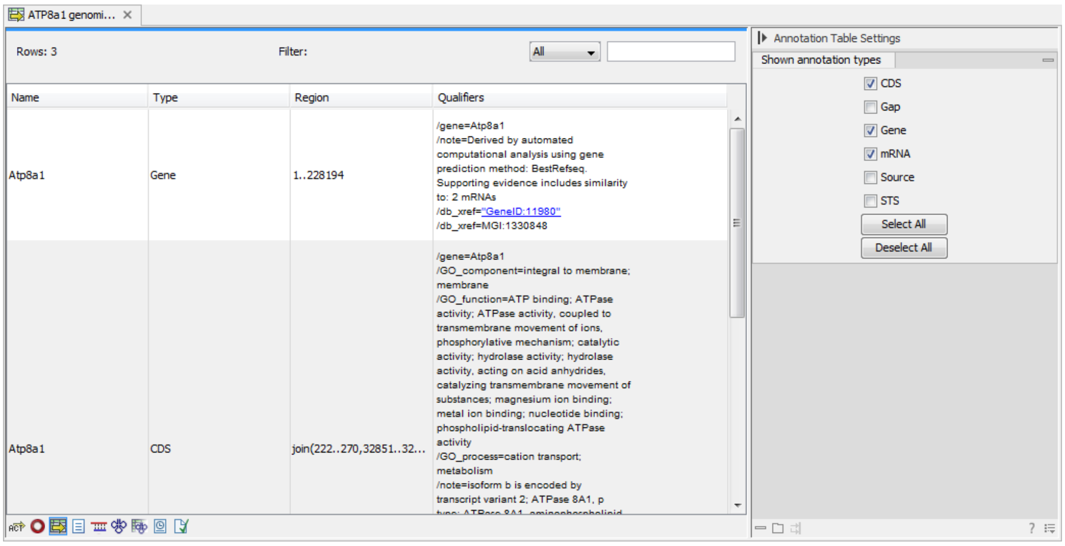
This will open a view similar to the one in figure 9.12).

In the **Side Panel** you can show or hide individual annotation types in the table. E.g. if you only wish to see "gene" annotations, de-select the other annotation types so that only "gene" is selected.

Each row in the table is an annotation which is represented with the following information:

- **Name.**
- **Type.**
- **Region.**
- **Qualifiers.**

The Name, Type and Region for each annotation can be edited simply by double-clicking, typing the change directly, and pressing **Enter**.



Rows: 3

Filter: All

Name	Type	Region	Qualifiers
Atp8a1	Gene	1..228194	/gene=Atp8a1 /note=Derived by automated computational analysis using gene prediction method: BestRefseq. Supporting evidence includes similarity to: 2 mRNAs /db_xref="GeneID:11980" /db_xref=MGI:1330848
Atp8a1	CDS	join(222..270,32851..32...	/gene=Atp8a1 /GO_component=integral to membrane; membrane /GO_function=ATP binding; ATPase activity; ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism; catalytic activity; hydrolase activity; hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances; magnesium ion binding; metal ion binding; nucleotide binding; phospholipid-translocating ATPase activity /GO_process=cation transport; metabolism /note=isoform b is encoded by transcript variant 2; ATPase 8A1, p

Annotation Table Settings

Shown annotation types

- ☑ CDS
- ☐ Gap
- ☑ Gene
- ☑ mRNA
- ☐ Source
- ☐ STS

Figure 9.12: *A table showing annotations on the sequence.*

This information corresponds to the information in the dialog when you edit and add annotations (see section 9.3.2).

You can benefit from this table in several ways:

- It provides an intelligible overview of all the annotations on the sequence.
- You can use the filter at the top to search the annotations. Type e.g. "UCP" into the filter and you will find all annotations which have "UCP" in either the name, the type, the region or the qualifiers. Combined with showing or hiding the annotation types in the **Side Panel**, this makes it easy to find annotations or a subset of annotations.
- You can copy and paste annotations, e.g. from one sequence to another.
- If you wish to edit many annotations consecutively, the double-click editing makes this very fast (see section 9.3.2).

9.3.2 Adding annotations

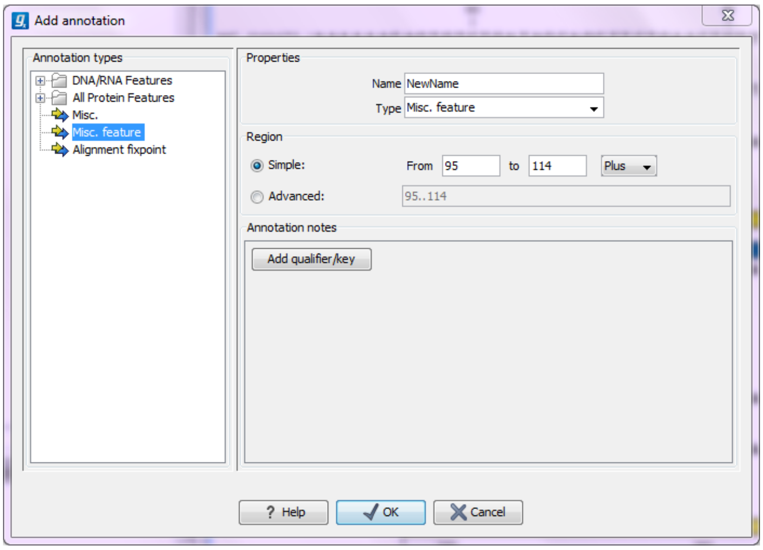
Adding annotations to a sequence can be done in two ways:

Open the sequence in a sequence view (double-click in the Navigation Area) | make a selection covering the part of the sequence you want to annotate[1] **| right-click the selection | Add Annotation ( )**

or **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Annotation table ( ) | right click anywhere in the annotation table | select Add Annotation ( )**

This will display a dialog like the one in figure 9.13.

The left-hand part of the dialog lists a number of **Annotation types**. When you have selected an annotation type, it appears in **Type** to the right. You can also select an annotation directly in this list. Choosing an annotation type is mandatory. If you wish to use an annotation type which is



Name: NewName

Type: Misc. feature

● Simple: From 95 to 114 Plus

○ Advanced: 95..114

Figure 9.13: *The Add Annotation dialog.*

not present in the list, simply enter this type into the **Type** field [2].

The right-hand part of the dialog contains the following text fields:

- **Name.** The name of the annotation which can be shown on the label in the sequence views. (Whether the name is actually shown depends on the **Annotation Layout** preferences, see section 9.3.1).
- **Type.** Reflects the left-hand part of the dialog as described above. You can also choose directly in this list or type your own annotation type.
- **Region.** If you have already made a selection, this field will show the positions of the selection. You can modify the region further using the conventions of DDBJ, EMBL and GenBank. The following are examples of how to use the syntax (based on `http://www.ncbi.nlm.nih.gov/collab/FT/`):
  - **467**. Points to a single residue in the presented sequence.
  - **340..565**. Points to a continuous range of residues bounded by and including the starting and ending residues.
  - **<345..500**. Indicates that the exact lower boundary point of a region is unknown. The location begins at some residue previous to the first residue specified (which is not necessarily contained in the presented sequence) and continues up to and including the ending residue.
  - **<1..888**. The region starts before the first sequenced residue and continues up to and including residue 888.
  - **1..>888**. The region starts at the first sequenced residue and continues beyond residue 888.
  - **(102.110)**. Indicates that the exact location is unknown, but that it is one of the residues between residues 102 and 110, inclusive.

[2] Note that your own annotation types will be converted to "unsure" when exporting in GenBank format. As long as you use the sequence in CLC format, you own annotation type will be preserved

  - **123ˆ124**. Points to a site between residues 123 and 124.
  - **join(12..78,134..202)**. Regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.
  - **complement(34..126)** Start at the residue complementary to 126 and finish at the residue complementary to residue 34 (the region is on the strand complementary to the presented strand).
  - **complement(join(2691..4571,4918..5163))**. Joins regions 2691 to 4571 and 4918 to 5163, then complements the joined segments (the region is on the strand complementary to the presented strand).
  - **join(complement(4918..5163),complement(2691..4571))**. Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the region is on the strand complementary to the presented strand).

- **Annotations.** In this field, you can add more information about the annotation like comments and links. Click the **Add qualifier/key** button to enter information. Select a qualifier which describes the kind of information you wish to add. If an appropriate qualifier is not present in the list, you can type your own qualifier. The pre-defined qualifiers are derived from the GenBank format. You can add as many qualifier/key lines as you wish by clicking the button. Redundant lines can be removed by clicking the delete icon (✖). The information entered on these lines is shown in the annotation table (see section 9.3.1) and in the yellow box which appears when you place the mouse cursor on the annotation. If you write a hyperlink in the **Key** text field, like e.g. "www.clcbio.com", it will be recognized as a hyperlink. Clicking the link in the annotation table will open a web browser.

Click **OK** to add the annotation.

Note! The annotation will be included if you export the sequence in GenBank, Swiss-Prot or CLC format. When exporting in other formats, annotations are not preserved in the exported file.

9.3.3 Edit annotations

To edit an existing annotation from within a sequence view:

right-click the annotation | Edit Annotation (✎)

This will show the same dialog as in figure 9.13, with the exception that some of the fields are filled out depending on how much information the annotation contains.

There is another way of quickly editing annotations which is particularly useful when you wish to edit several annotations.

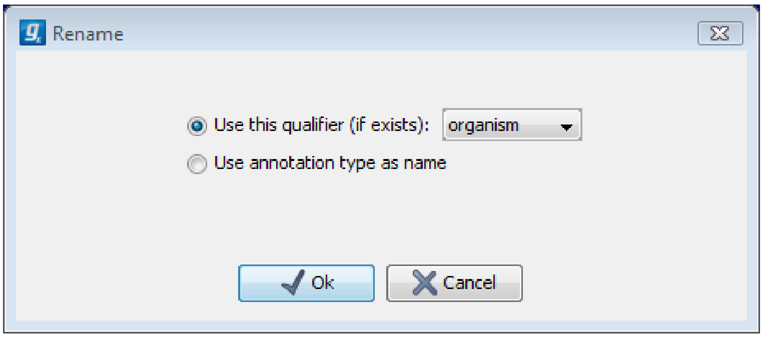
To edit the information, simply double-click and you will be able to edit e.g. the name or the annotation type. If you wish to edit the qualifiers and double-click in this column, you will see the dialog for editing annotations.

Advanced editing of annotations

Sometimes you end up with annotations which do not have a meaningful name. In that case there is an advanced batch rename functionality:

Open the Annotation Table (☰) | select the annotations that you want to rename | right-click the selection | Advanced Rename

This will bring up the dialog shown in figure 9.14.



Rename

● Use this qualifier (if exists): organism
○ Use annotation type as name

Figure 9.14: *The Advanced Rename dialog.*

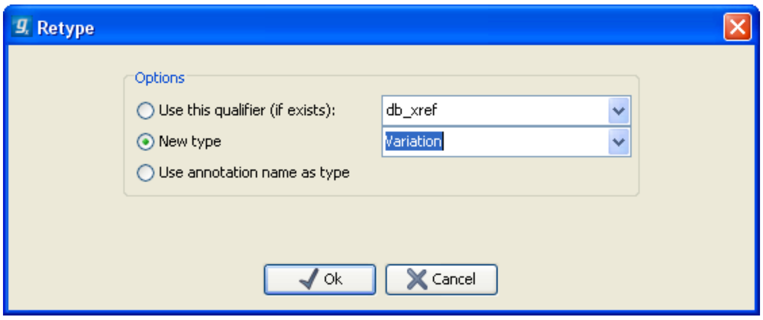
In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as name. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be renamed. If an annotation has multiple qualifiers of the same type, the first is used for naming.
- **Use annotation type as name.** The annotation's type will be used as name (e.g. if you have an annotation of type "Promoter", it will get "Promoter" as its name by using this option).

A similar functionality for batch re-typing annotations is available in the right-click menu as well, in case your annotations are not typed correctly:

Open the Annotation Table (☰) | select the annotations that you want to retype | right-click the selection | Advanced Retype

This will bring up the dialog shown in figure 9.15.



Retype

Options
○ Use this qualifier (if exists): db_xref
● New type Variation
○ Use annotation name as type

Figure 9.15: *The Advanced Retype dialog.*

In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as type. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be retyped. If an annotation has multiple qualifiers of the same type, the first is used for the new type.

- **New type**. You can select from a list of all the pre-defined types as well as enter your own annotation type. All the selected annotations will then get this type.
- **Use annotation name as type.** The annotation's name will be used as type (e.g. if you have an annotation named "Promoter", it will get "Promoter" as its type by using this option).

9.3.4 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 9.3.1). In order to completely remove the annotation:

> **right-click the annotation | Delete Annotation ( )**

If you want to remove all annotations of one type:

> **right-click an annotation of the type you want to remove | Delete | Delete Annotations of Type "type"**

If you want to remove all annotations from a sequence:

> **right-click an annotation | Delete | Delete All Annotations**

The removal of annotations can be undone using Ctrl + Z or Undo ( ) in the Toolbar.

If you have more sequences (e.g. in a sequence list, alignment or contig), you have two additional options:

> **right-click an annotation | Delete | Delete All Annotations from All Sequences**

> **right-click an annotation | Delete | Delete Annotations of Type "type" from All Sequences**

9.4 Element information

The normal view of a sequence (by double-clicking) shows the annotations as boxes along the sequence, but often there is more information available about sequences. This information is available through the **Element info** view.

To view the sequence information:

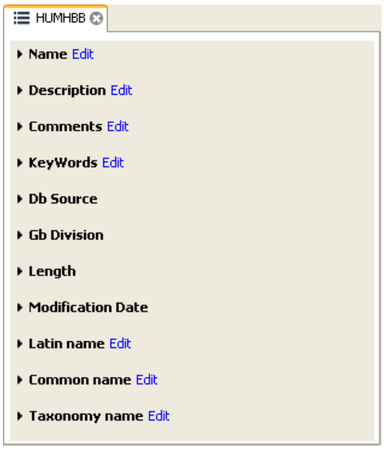
> **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Element Info ( )**

Another way to show the text view is to open the sequence in the **View Area** and click on the "Show Element Info" icon ( ) found at the bottom of the window.

This will display a view similar to fig 9.16.

All the lines in the view are headings, and the corresponding text can be shown by clicking the text.

- **Name.** The name of the sequence which is also shown in sequence views and in the **Navigation Area**.
- **Description.** A description of the sequence.
- **Comments.** The author's comments about the sequence.



HUMHBB

- Name Edit
- Description Edit
- Comments Edit
- KeyWords Edit
- Db Source
- Gb Division
- Length
- Modification Date
- Latin name Edit
- Common name Edit
- Taxonomy name Edit

Figure 9.16: *The initial display of sequence info for the HUMHBB DNA sequence from the Example data.*

- **Keywords.** Keywords describing the sequence.
- **Db source.** Accession numbers in other databases concerning the same sequence.
- **Gb Division.** Abbreviation of GenBank divisions. See section 3.3 in the GenBank release notes for a full list of GenBank divisions.
- **Length.** The length of the sequence.
- **Modification date.** Modification date from the database. This means that this date does not reflect your own changes to the sequence. See the history (section 7) for information about the latest changes to the sequence after it was downloaded from the database.
- **Latin name.** Latin name of the organism.
- **Common name.** Scientific name of the organism.
- **Taxonomy name.** Taxonomic classification levels.

The information available depends on the origin of the sequence.

Some of the information can be edited by clicking the blue **Edit** text. This means that you can add your own information to sequences that do not derive from databases.

Note that for other kinds of data, the **Element info** will only have **Name** and **Description**.

9.5 View as text

A sequence can be viewed as text without any layout and text formatting. This displays all the information about the sequence in the GenBank file format. To view a sequence as text:

> **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Select "Text View" (☰)**

Another way to show the text view is to open the sequence in the **View Area** and click on the "Show Text View" icon (☰) found at the bottom of the window.

This makes it possible to see background information about e.g. the authors and the origin of DNA and protein sequences. Selections or the entire text of the **Sequence Text View** can be copied and pasted into other programs:

Much of the information is also displayed in the **Sequence info**, where it is easier to get an overview (see section 9.4.)

In the **Side Panel**, you find a search field for searching the text in the view.

9.6 Sequence Lists

The **Sequence List** shows a number of sequences in a tabular format or it can show the sequences together in a normal sequence view.

Having sequences in a sequence list can help organizing sequence data.

Sequence lists are generated automatically when you import files containing more than one sequence. Sequence lists may also be created as the output from particular Workbench tool including database searches.

Sequence List can also be created from single sequences or by merging already existing sequence lists with the Workbench. To do this:

> **select two or more sequences or sequence lists | right-click the elements | New | Sequence List ()**

Alternatively, you can launch this took via the menu system:

> **File | New | Sequence List ()**

This opens the **Sequence List** Wizard:

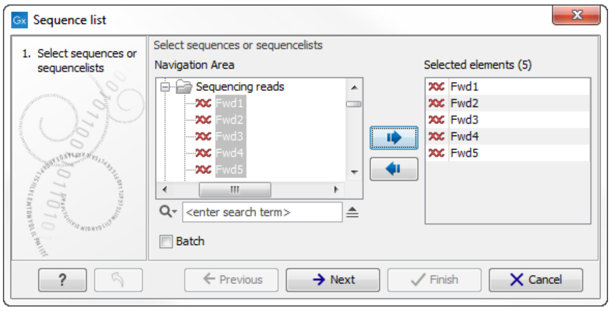


Figure 9.17: *A Sequence List dialog.*

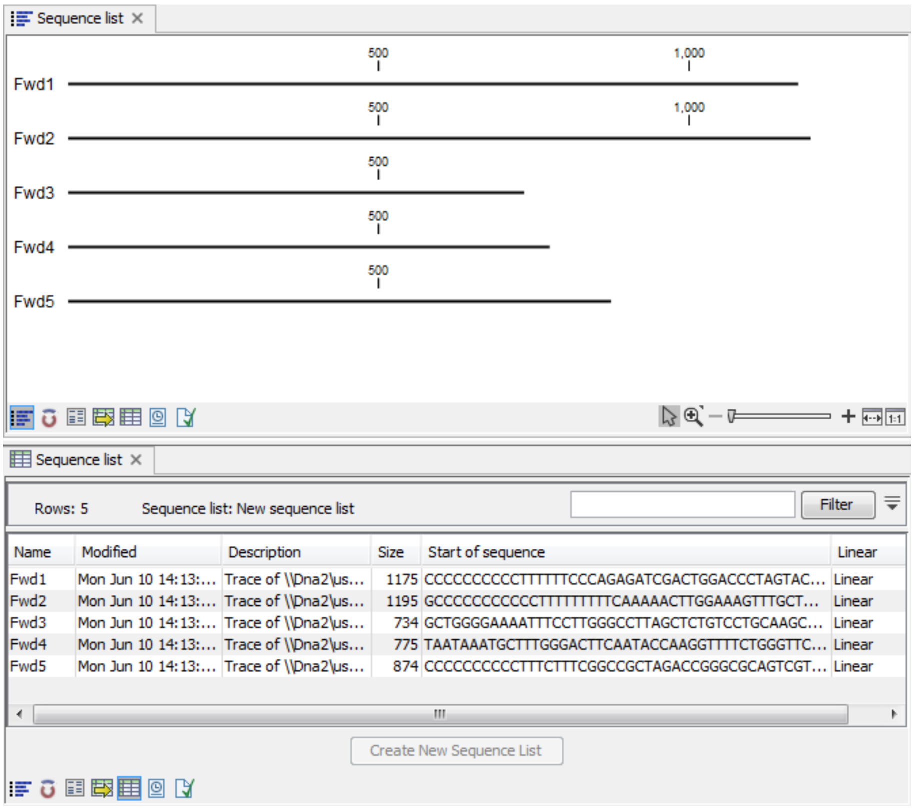
The dialog allows you to select more sequences to include in the list, or to remove already chosen sequences from the list.

Clicking **Finish** opens the sequence list. It can be saved by clicking **Save** () or by dragging the tab of the view into the **Navigation Area**.

Opening a Sequence list is done by:

> **right-click the sequence list in the Navigation Area | Show () | Graphical Sequence List () OR Table ()**

The two different views of the same sequence list are shown in split screen in figure 9.18.



Rows: 5 Sequence list: New sequence list

Name	Modified	Description	Size	Start of sequence	Linear
Fwd1	Mon Jun 10 14:13:...	Trace of \\Dna2\us...	1175	CCCCCCCCCCTTTTTTCCCAGAGATCGACTGGACCCTAGTAC...	Linear
Fwd2	Mon Jun 10 14:13:...	Trace of \\Dna2\us...	1195	GCCCCCCCCCCCTTTTTTTTTCAAAAACTTGGAAAGTTTGCT...	Linear
Fwd3	Mon Jun 10 14:13:...	Trace of \\Dna2\us...	734	GCTGGGGAAAATTTCCTTGGGCCTTAGCTCTGTCCTGCAAGC...	Linear
Fwd4	Mon Jun 10 14:13:...	Trace of \\Dna2\us...	775	TAATAAATGCTTTGGGACTTCAATACCAAGGTTTTCTGGGTTC...	Linear
Fwd5	Mon Jun 10 14:13:...	Trace of \\Dna2\us...	874	CCCCCCCCCCTTTCTTTCGGCCGCTAGACCGGGCGCAGTCGT...	Linear

Figure 9.18: *A sequence list containing multiple sequences can be viewed in either a table or in a graphical sequence list. The graphical view is useful for viewing annotations and the sequence itself, while the table view provides other information like sequence lengths, and the number of sequences in the list (number of Rows reported).*

9.6.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 9.1). The main difference is that you now can see more than one sequence in the same view.

However, you also have a few extra options for sorting, deleting and adding sequences:

- To add extra sequences to the list, right-click an empty (white) space in the view, and select **Add Sequences**.
- To delete a sequence from the list, right-click the sequence's name and select **Delete Sequence**.
- To sort the sequences in the list, right-click the name of one of the sequences and select **Sort Sequence List by Name** or **Sort Sequence List by Length**.
- To rename a sequence, right-click the name of the sequence and select **Rename Sequence**.

9.6.2 Sequence list table

Each sequence in the table sequence list is displayed with:

- Name.
- Accession.
- Description.
- Modification date.
- Length.
- First 50 residues.

The number of sequences in the list is reported as the number of Rows at the top of the table view.

Adding and removing sequences from the list is easy: adding is done by dragging the sequence from another list or from the **Navigation Area** and drop it in the table. To delete sequences, simply select them and press **Delete** ().

You can also create a subset of the sequence list:

select the relevant sequences | **right-click** | **Create New Sequence List**

This will create a new sequence list, which only includes the selected sequences.

Learn more about tables in Appendix 8.3.

9.6.3 Extract sequences from sequence list

Sequences can be extracted from a sequence list when the sequence list is opened in tabular view. One or more sequences can be dragged (with the mouse) directly from the table into the **Navigation Area**. This allows you to extract specific sequences from the entire list. Another option is to extract all sequences found in the list. This can be done with the **Extract Sequences** tool:

Toolbox | **General Sequence Analysis ()**| **Extract Sequences** () .

A description of how to use the **Extract Sequences** tool can be found in section 28.1.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Part III

Applications - ready-to-use workflows

Chapter 10

Getting started

Contents

10.1 Reference data

The ready-to-use workflows rely on the presence of particular reference datasets. This reference data must be downloaded and configured before these workflows can be used. The tools in the Workbench make it easy to download the necessary data such that the workflows can find and use it. This section covers the download and configurations needed to make available the reference data relevant to the *CLC Cancer Research Workbench*, including the human genome reference, annotations and variants made available by a variety of databases.

10.1.1 The Workbench Reference data location

Reference data must be stored in a folder called **CLC_References**. When the *CLC Cancer Research Workbench* is installed, such a folder is created on your file system under your home area. This folder is specified within the Workbench as a **reference location**.

You can specify a different location to download reference data to. This is recommended if you do not have enough space in the area the Workbench designates as the reference data location by default. To change the reference data location from within the **Navigation Area**:

> **Right-click on the folder "CLC_References" | Choose "Location" | Choose "Specify Reference Location"**

The new folder will also be called CLC_References, but will be located where you specify.

In more detail, this action results in the following:

- A folder called CLC_References is created in the location you specified, if a folder of this name did not already exist.
- The Workbench sets this new location as the place to download reference data to and the place the ready-to-use workflows should look for reference data.

This action does **not**:

- Remove the old CLC_References folder.
- Remove the contents of the old CLC_References folder, such as previously downloaded data.

If you have previously downloaded data into the CLC_References folder with the old location, you will need to use standard system tools to delete this folder and/or its contents. If you would like to keep the reference data from the old location, you can move it, using standard system tools, into the new CLC_References folder that you just specified. This would save you needing to download it again.

Note! If you run out of space, and realize that the CLC_References should be stored somewhere else, you can do this by choosing a new location, then manually moving the already downloaded files to that new location, and restarting the workbench. The "downloaded references" file will then be updated with all the new references.

10.1.2 Space requirements

The total size of the complete reference data set you can download is approximately 12 GB[1]. It is in a zipped format, and the total size after the data is unzipped is substantially larger. The amount of time it will take to download this amount of data depends on your network connection. It can take several hours, or longer on slower connections. When unzipped the size of the full reference dataset is about 75 GB[2].

For reference, in April, 2014, the size of each individual reference data file was approximately:

[1] Size as estimated in April, 2014

[2] Size as estimated in April, 2014

Database	**Size**
1000 Genomes	10 GB
CDS	49 MB
ClinVar	41 MB
PhastConc	5 GB
COSMIC	372 MB
dbSNP	44 GB
dbSNP Common	12 GB
Genes	3 MB
Gene Ontology	33 MB
HapMap	3 GB
mRNA	62 MB
Sequence	683 MB

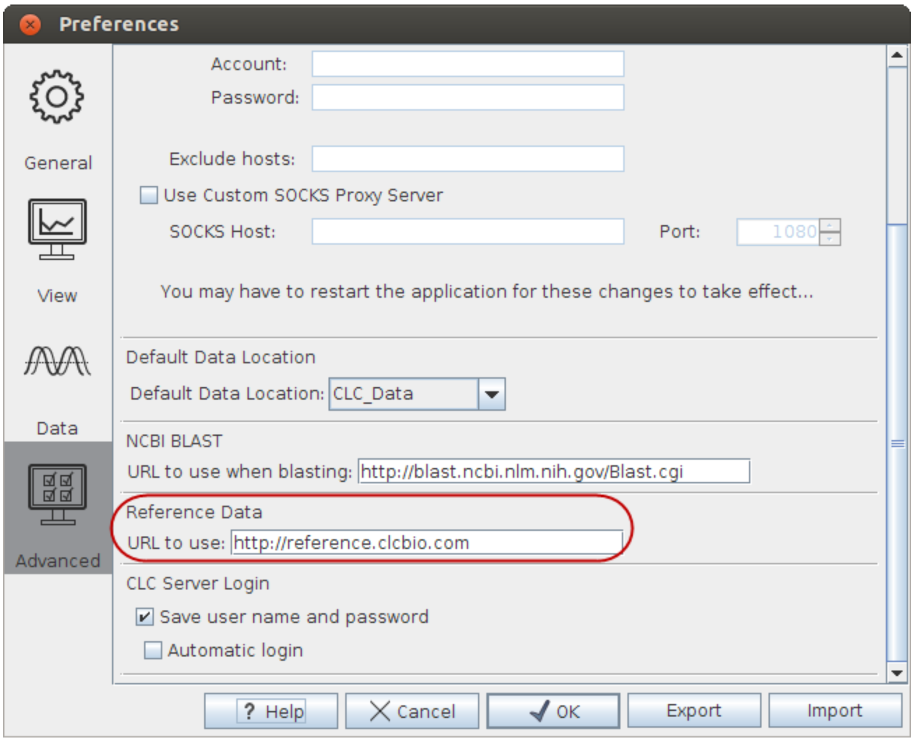
10.1.3 Where reference data is downloaded from

Reference data must be downloaded and configured manually before you can start using the ready-to-use workflows in the *CLC Cancer Research Workbench*. You only have to do this once. When all necessary reference data have been downloaded and configured, you will be automatically notified whenever updated reference data are available.

Data is provided by CLC bio and the Workbench is configured to download from CLC bio by default. The location to download the data from can be seen in the Workbench Preferences as shown in figure 10.1).

Edit | **Preferences** | **Advanced**

Unless you are in the special circumstance that your system administrator has decided to mirror this data locally and wishes you to use that mirror of the data, you should **not** change this setting.



Account:

Password:

Exclude hosts:

□ Use Custom SOCKS Proxy Server

SOCKS Host: Port: 1080

You may have to restart the application for these changes to take effect...

Default Data Location

Default Data Location: CLC_Data

NCBI BLAST

URL to use when blasting: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Reference Data

URL to use: http://reference.clcbio.com

CLC Server Login

☑ Save user name and password

□ Automatic login

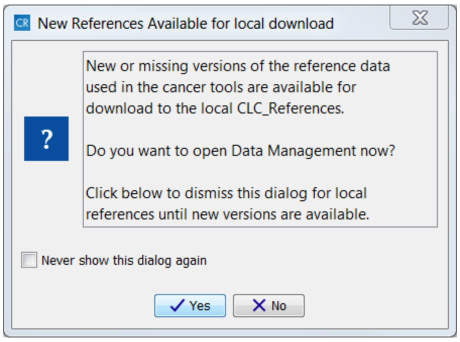
Figure 10.1: *The location where reference data is downloaded from can be seen in the Workbench Preferences. Generally this should not be altered except in the special case that the data from CLC bio is being mirrored locally.*

10.1.4 Download and configure reference data

The first time you open *CLC Cancer Research Workbench* you will be presented with the dialog box shown in figure 10.2, which informs you that data are available for download for either to the local or server CLC_References repository. If you check the "Never show this dialog again" then subsequently you will only be presented with the dialog box when updated versions of the reference data are available.

Click on the button labeled **Yes**. This will take you to the wizard shown in figure 10.3.

This wizard can also be accessed from the upper right corner of the *CLC Cancer Research*


New References Available for local download

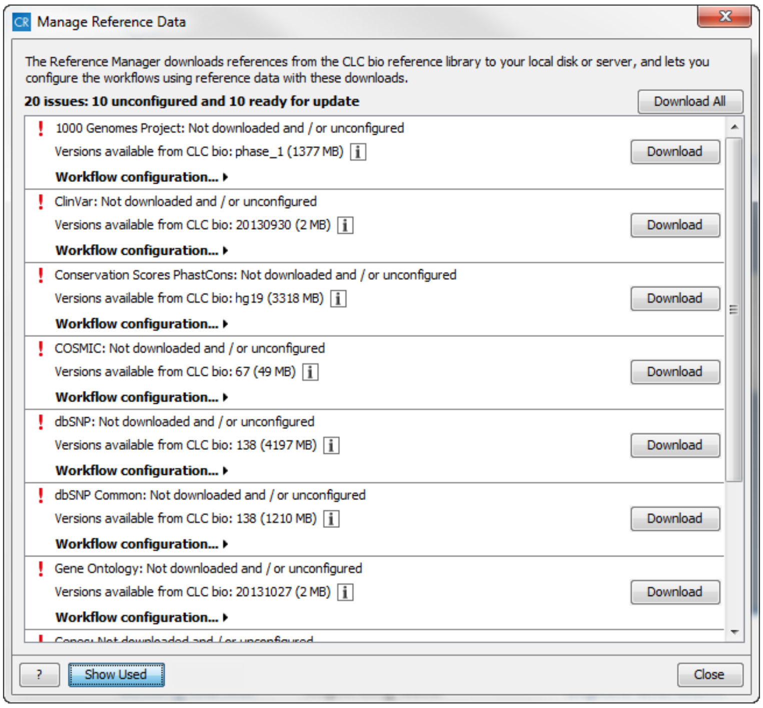
New or missing versions of the reference data used in the cancer tools are available for download to the local CLC_References.

Do you want to open Data Management now?

Click below to dismiss this dialog for local references until new versions are available.

□ Never show this dialog again

Figure 10.2: *Notification that new versions of the reference data are available.*


Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

20 issues: 10 unconfigured and 10 ready for update

! 1000 Genomes Project: Not downloaded and / or unconfigured
Versions available from CLC bio: phase_1 (1377 MB)
Workflow configuration...

! ClinVar: Not downloaded and / or unconfigured
Versions available from CLC bio: 20130930 (2 MB)
Workflow configuration...

! Conservation Scores PhastCons: Not downloaded and / or unconfigured
Versions available from CLC bio: hg19 (3318 MB)
Workflow configuration...

! COSMIC: Not downloaded and / or unconfigured
Versions available from CLC bio: 67 (49 MB)
Workflow configuration...

! dbSNP: Not downloaded and / or unconfigured
Versions available from CLC bio: 138 (4197 MB)
Workflow configuration...

! dbSNP Common: Not downloaded and / or unconfigured
Versions available from CLC bio: 138 (1210 MB)
Workflow configuration...

! Gene Ontology: Not downloaded and / or unconfigured
Versions available from CLC bio: 20131027 (2 MB)
Workflow configuration...

! Genes: Not downloaded and / or unconfigured

Figure 10.3: *The Manage Reference Data wizard gives access to the reference data that are required to be able to run the ready-to-use workflows. The default view shows the references that are used in the workflows. With the "Show All" button the reference list can be expanded with additional (optional) reference data that you may find useful.*

Workbench by clicking on **Data Management** ( ) figure 10.4.

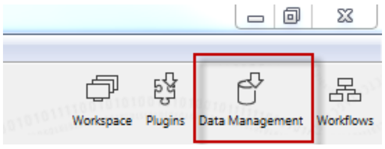


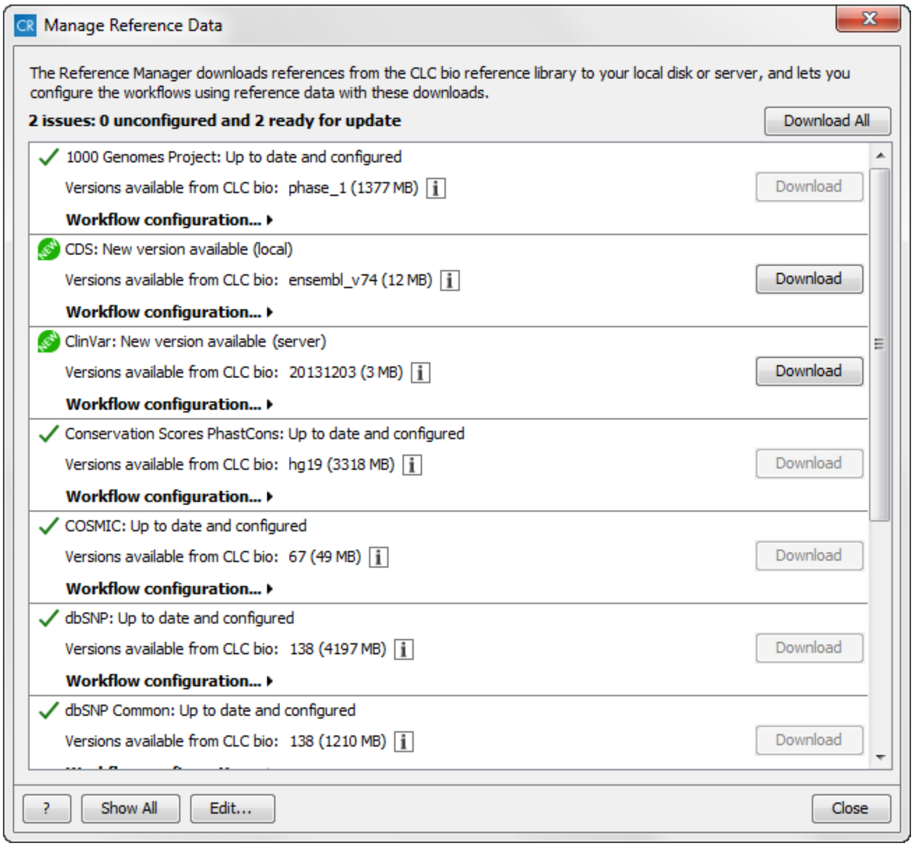
Figure 10.4: *Click on the button labeled "Data management" to open the "Manage Reference Data" dialog where you can download and configure the reference data that are necessary to be able to run the ready-to-use-workflows.*

The "Manage Reference Data" wizard gives access to all the reference data that are used in the ready-to-use workflows. From the wizard you can download and configure the reference data. A button labeled "Show All" at the bottom of the dialog can be used to expand the list with additional reference data that are not required for any of the workflows (Gene Ontology). Rather these extra reference data have been provided as an extra service for those of our users who would like to include information from these databases in the data analyses.

Icons are used in the "Manage Reference Data" wizard to give a quick overview of the current status of each reference: "Not downloaded and / or unconfigured", "Workflows use different versions" or "Selected version is inconsistent / not fully downloaded" references are marked with a red exclamation mark (!), references that are "Up to date and configured" are marked with a green check mark (✓), and when a new version of a reference data set is available, you will see a green mark labeled "New" (NEW).

Guide to the "Manage Reference Data" wizard:

- *In the upper part of the wizard you can find:*
  - A small descriptive text
  - An indication of how many issues you have, how many of these are "unconfigured issues", and how many are reference data that are "ready for update".
  - The button labeled **Download All**, which can be used to download all reference data that are shown in the wizard. This is the case the first time you use the "Download All" button. Subsequently, only reference data where a newer version is available, will be downloaded. If you have selected "Show All" (the "Show All" button is found at the bottom of the wizard), all reference data will be downloaded (including "Gene Ontology"). If you have selected "Show Used", only the reference data that are used in the ready-to-use workflows will be downloaded.
- *The central area of the wizard:*
  - Lists all available references data. After the reference name, a small note shows the status of the reference (see figure 10.5), which can be:
    - Not downloaded and / or unconfigured (!)
    - Workflows use different versions (!)
    - Selected version is inconsistent / not fully downloaded (!)
    - Up to date and configured (✓)
    - New version available (NEW)
  - When a *new version is available* (NEW), it is stated in a parenthesis whether it is for your local disc, for the server, or local and server (see figure 10.5).
  - If a version is *inconsistent / not fully downloaded* (!), it will be stated in parenthesis whether it is the local or server version (or both). Check the process tab for running or suspended download processes. Please wait for all of these to finish. If the data is inconsistent, even after all downloads have finished, it is likely that you ran out of disk space, or the download or import was somehow stopped prematurely.
    In this case, you can ''Delete'' the reference, and try downloading it again.
  - In the unlikely event that a reference has the mark *Workflows use different versions* (!), the Workbench has discovered that two or more installed workflows use different versions of a reference, and is unable to determine which should be used. Please select the correct version from the drop down menu and click ''Use Reference'' to



Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

2 issues: 0 unconfigured and 2 ready for update

✓ 1000 Genomes Project: Up to date and configured
Versions available from CLC bio: phase_1 (1377 MB)

CDS: New version available (local)
Versions available from CLC bio: ensembl_v74 (12 MB)

ClinVar: New version available (server)
Versions available from CLC bio: 20131203 (3 MB)

✓ Conservation Scores PhastCons: Up to date and configured
Versions available from CLC bio: hg19 (3318 MB)

✓ COSMIC: Up to date and configured
Versions available from CLC bio: 67 (49 MB)

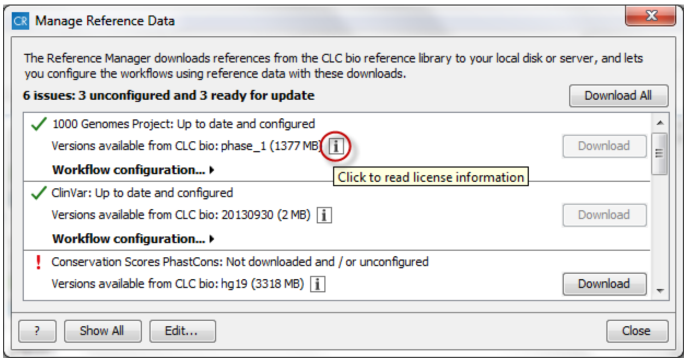
✓ dbSNP: Up to date and configured
Versions available from CLC bio: 138 (4197 MB)

✓ dbSNP Common: Up to date and configured
Versions available from CLC bio: 138 (1210 MB)

Figure 10.5: *The Manage Reference Data wizard lists the reference data. Three different icons are used to mark the status of the reference.*

solve this. See **Workflow configuration** below for more information on configuring workflows.

- Under the reference used, you can find info about the reference version (**Versions available from CLC bio**) and the size of the reference data. By clicking on (ℹ) you can see the legal notice and license information for this particular reference data set (see figure 10.6).



Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

6 issues: 3 unconfigured and 3 ready for update

✓ 1000 Genomes Project: Up to date and configured
Versions available from CLC bio: phase_1 (1377 MB)

Click to read license information

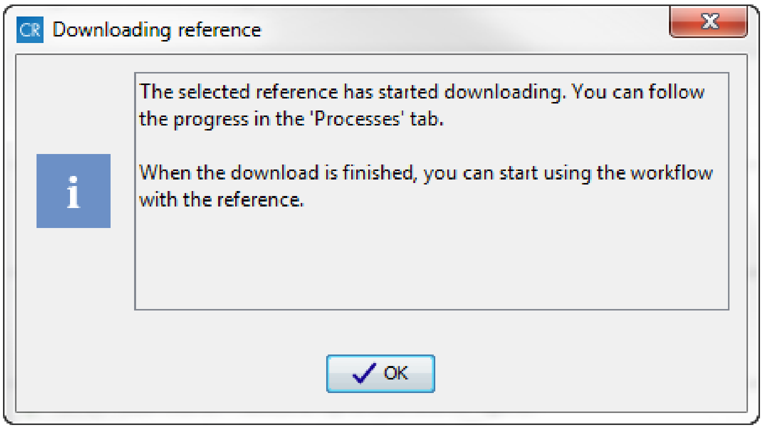
✓ ClinVar: Up to date and configured
Versions available from CLC bio: 20130930 (2 MB)

! Conservation Scores PhastCons: Not downloaded and / or unconfigured
Versions available from CLC bio: hg19 (3318 MB)

Figure 10.6: *Click on the info button to see the legal notice and license information.*

- The button labeled **Download** can be used to download the reference data individually. When you click on the button labeled **Download**, a wizard appears with a message informing you that the selected reference data are now being downloaded (figure 10.7).

  After the reference data have been downloaded the icon changes to a green check mark for those of the databases that only contain one reference data file.



Downloading reference

The selected reference has started downloading. You can follow the progress in the 'Processes' tab.

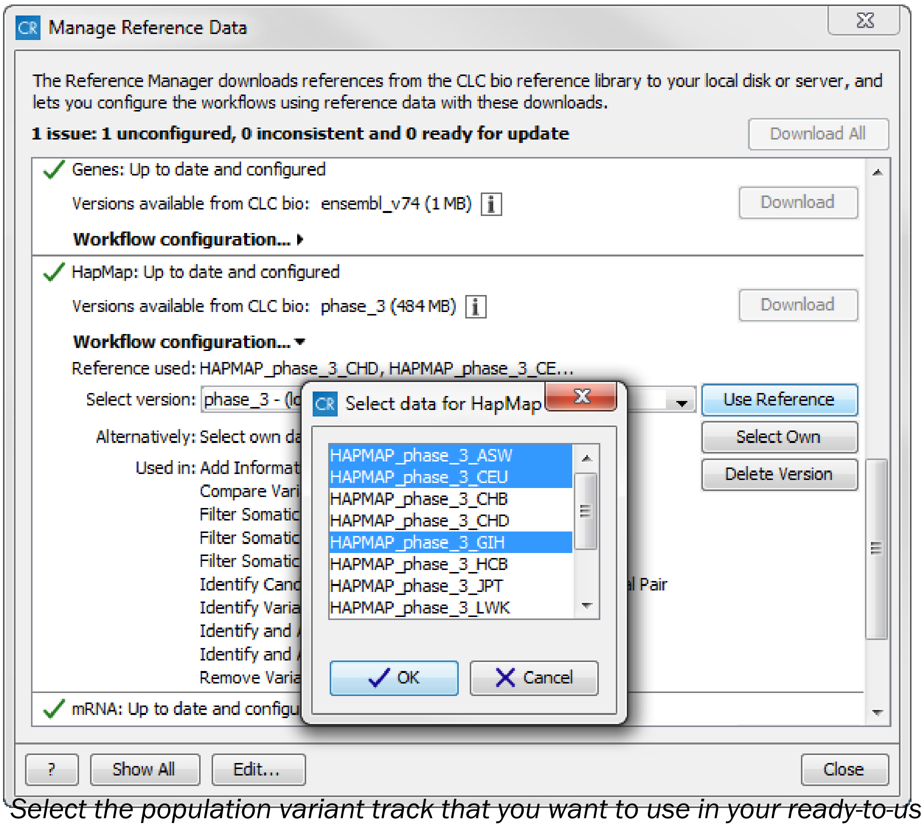
When the download is finished, you can start using the workflow with the reference.

Figure 10.7: *The Downloading Reference wizard informs you about that data is being downloaded.*

- The boldface text **Workflow configuration** can be expanded to reveal additional options. When unfolded, you can see which version of the reference is being used, and which of the ready-to-use workflows use this reference. In addition, three buttons appear:
  - **Use Reference** When the reference data have been downloaded, the workflows will automatically be configured with all the reference data available. The drop down ''Select Version'' allows you to change between the downloaded versions, and pressing ''Use Reference'' will update the installed workflows to use this specific version for the selected reference. However, references like the ''1000 Genomes Project'' and ''HapMap'' databases, which contain more than one reference data file[3], you have to specify which reference data to use. This is what the "Use Reference" option allows you to do. Select the reference data by clicking on the data you want to use. If you want to select more than one population, hold down the Ctrl key while selecting data files.

    When you have selected the population that you want to use for your data analyses, click on the button labeled **OK**. Your workflow will now be configured with the reference data for the population(s) that you have selected. Please note that you have to do this for both the "1000 Genomes Project" and for the "HapMap" reference data. See figure 10.8.
  - **Delete Version** With this button all users are capable of deleting locally installed reference data, whereas only administrators are capable of deleting reference data installed on the server. This can be used if you suspect that a downloaded reference is corrupt, and needs to be re-downloaded, or if you need to clean up space, e.g. locally.
  - **Use Own File** allows you to use your own reference data. The data type and number of files to select will be restricted to match the reference. This is useful when you have your own version of the reference data that you would like to use rather than the data made available to download directly into the Workbench. If you want to switch back to using the downloaded references, you must use the

[3] In some cases, reference data are available from different population subgroups. This is the case for HapMap and the 1000 Genomes Project. Three letter codes are used to specify the population that the different reference data origin from (e.g. ASW = American's of African Ancestry in SW USA). For the phase 3 HapMap population codes, please see `http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html` and for the 1000 Genomes Project see `http://www.ensembl.org/Help/Faq?id=328`. Figure 10.10 shows the CLC_References folder. You can see that different populations are available for HapMap and the 1000 Genomes Project.



Manage Reference Data

The Reference Manager downloads references from the CLC bio reference library to your local disk or server, and lets you configure the workflows using reference data with these downloads.

1 issue: 1 unconfigured, 0 inconsistent and 0 ready for update

✓ Genes: Up to date and configured

Versions available from CLC bio: ensembl_v74 (1 MB)

Workflow configuration...

✓ HapMap: Up to date and configured

Versions available from CLC bio: phase_3 (484 MB)

Workflow configuration...

Reference used: HAPMAP_phase_3_CHD, HAPMAP_phase_3_CE...

Select version: phase_3 - (lo

Alternatively: Select own da

Used in: Add Informat

Compare Vari

Filter Somatic

Filter Somatic

Filter Somatic

Identify Cand

Identify Varia

Identify and

Identify and

Remove Varia

l Pair

✓ mRNA: Up to date and configu

Select data for HapMap

HAPMAP_phase_3_ASW

HAPMAP_phase_3_CEU

HAPMAP_phase_3_CHB

HAPMAP_phase_3_CHD

HAPMAP_phase_3_GIH

HAPMAP_phase_3_HCB

HAPMAP_phase_3_JPT

HAPMAP_phase_3_LWK

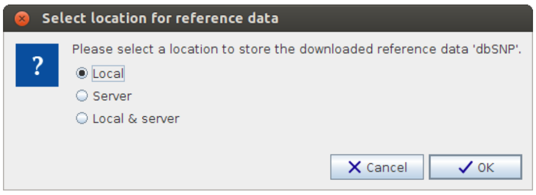
Figure 10.8: *Select the population variant track that you want to use in your ready-to-use workflows.*

"Use Reference" again.

- *At the bottom of the wizard you can find:*
  - A button with a question mark. This is the "help" button that links to the section in the *CLC Cancer Research Workbench* reference manual that describes the "Manage Reference Data" button.
  - A button labeled "Show All" (or "Show Used"). With this button you can choose whether you only want to see the reference data that is being used in the ready-to-use workflows, or if you want to see all available reference data. Please note that if you choose to use the "Download All" function, you will download the references that are shown in the wizard. This means that if you have selected "Show Used" you will only download the reference data that is being used in the workflows.
  - A button labeled "Close". Click on this to close the wizard.

If you are connected to a CLC Server you will be asked where you want to save the downloaded reference data, to your Workbench or your Server when you click on the button labeled **Download** or **Download All**. See figure 10.9. You will see this dialog the first time you download data. After this the dialog will appear only in situations where both the Local and Server version need updating. If a new version is found with respect to only Local or Server, the data will automatically be downloaded to that location.

When the reference data have been downloaded, the workflows will automatically be configured with the reference data. However, in some cases reference data are available from different population subgroups. This is the case for HapMap and the 1000 Genomes Project. Three letter codes are used to specify the population that the different reference data origin from (e.g. ASW =



Select location for reference data

Please select a location to store the downloaded reference data 'dbSNP'.

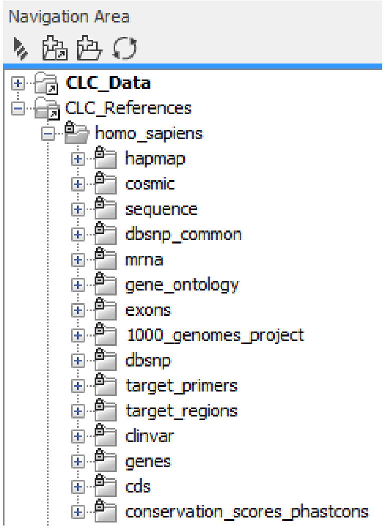
● Local
○ Server
○ Local & server

Figure 10.9: *Select where to save the downloaded reference data. Please be aware that the total size of all reference data (in April 2014) is about 12 GB when compressed. It can take some time to download all reference data. When unzipped the size of all the reference data, when the compressed size was about 12 GB is about 75 GB.*

American's of African Ancestry in SW USA). For the phase 3 HapMap population codes, please see http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html and for the 1000 Genomes Project see http://www.ensembl.org/Help/Faq?id=328.

Whenever workflows use reference data that are available from more than one population, the workflow will initially be automatically configured with all the populations being available, and which population to use in the workflow will then need to be specified by the user in one of the wizard steps that appear when starting the workflow. How to configure your workflow with the right population is described in section 10.1.4.

Figure 10.10 shows the CLC_References folder. If you open the folders holding the reference data, you can see that different populations are available for HapMap and the 1000 Genomes Project.



Navigation Area

- CLC_Data
- CLC_References
  - homo_sapiens
    - hapmap
    - cosmic
    - sequence
    - dbsnp_common
    - mrna
    - gene_ontology
    - exons
    - 1000_genomes_project
    - dbsnp
    - target_primers
    - target_regions
    - clinvar
    - genes
    - cds
    - conservation_scores_phastcons

Figure 10.10: *For the 1000 Genomes Project and HapMap reference data, data are available from different populations. For these two databases the user must manually specify the relevant population to be used in the workflows. If the user choose not to select a population manually, the workflow will use a randomly selected population.*

10.1.5 Troubleshooting reference data downloads

Network connection errors can occur when downloading reference data. If this happens, you can try to resume the download when the network connection has been restored (see figure 10.11). Alternatively, you can simply press stop to cancel the download process and clean up any temporary data.

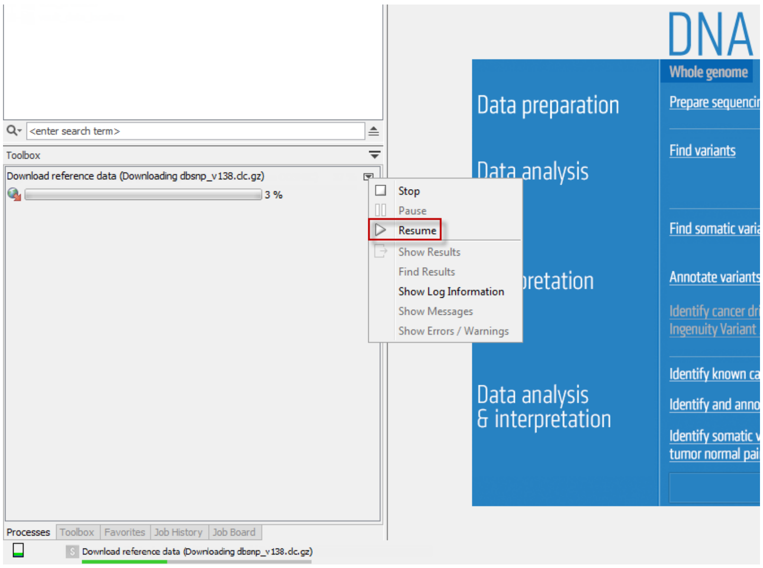


Figure 10.11: *It is possible to resume the download of data if you have encountered e.g. network connection errors.*

Please note that it is possible to specify the reference data location. E.g. if you do not have enough disc space on (C:) you can specify another location for your reference data by right-clicking on "CLC_References" in the **Navigation Area** (see figure 10.12).

10.2 Create new folder

To get started you need some data to work with. However, before looking into how you can import your data into the *CLC Cancer Research Workbench* we will first create a new folder in the **Navigation area** that can be used to hold all data that are relevant for the analysis you are about to perform. You can see how to do this in figure 10.13.

The folder that you have just created will be placed in the CLC_Data location as shown in figure 10.14.

10.3 Import data

We are now ready to start importing the data. The simplistic diagram shown in figure 10.15 will be used throughout the rest of the manual to provide an overview as we step by step move through the different steps from data import to analysis of your sequencing data.

Below you can find a short guide on how to import data into the *CLC Cancer Research Workbench*. If you wish to learn more about the import options in the *CLC Cancer Research Workbench*, you can find a more detailed description in the *CLC Cancer Research Workbench* reference manual (`http://clccancer.com/software/#downloads`).

10.3.1 How to import data

1. Use the **Import** tool in the toolbar (see figure 10.16) to import your sequencing data into the *CLC Cancer Research Workbench*.

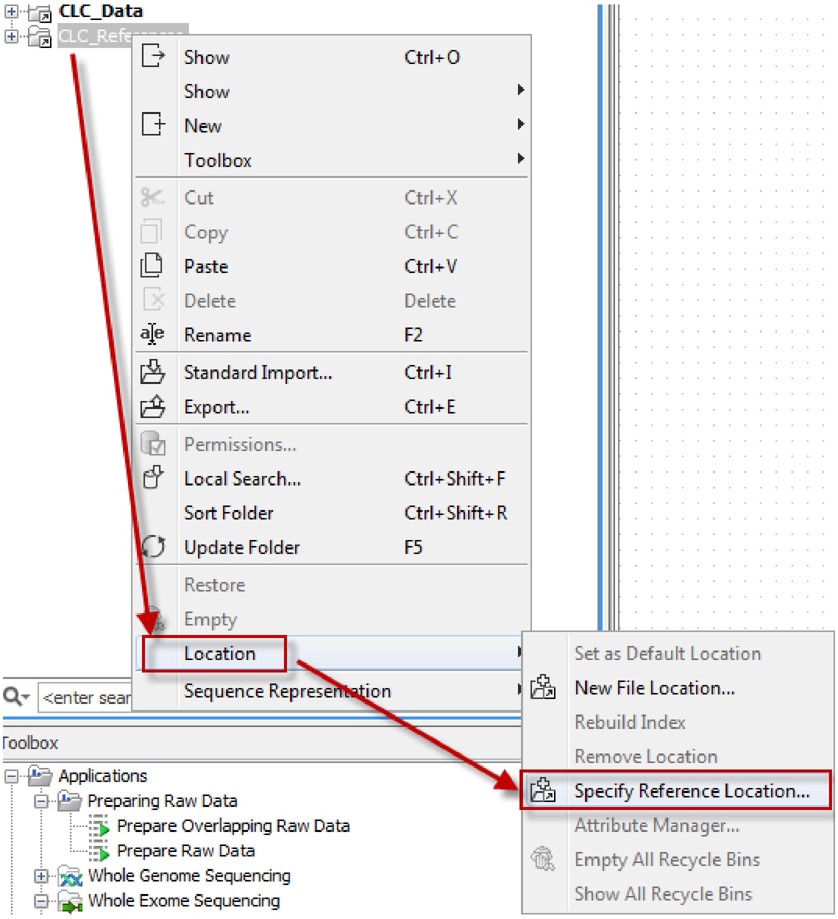
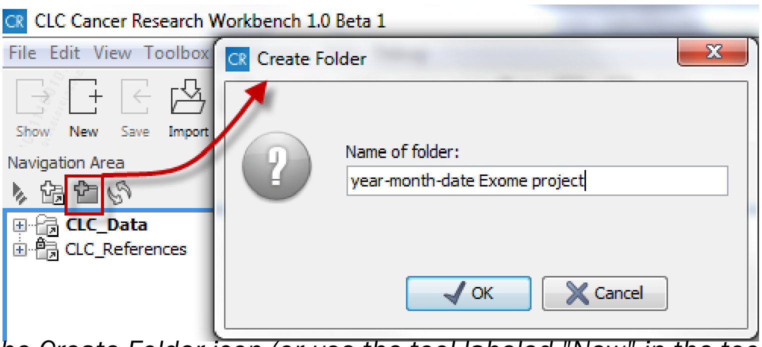


Figure 10.12: *It is possible to change the reference data location by right-clicking on "CLC_References". Select "Location" and "Specify Reference Location".*



Name of folder: year-month-date Exome project

Figure 10.13: *Click on the Create Folder icon (or use the tool labeled "New" in the toolbar) to create a new folder. Provide a name that will make it easy to keep track of your data.*

2. Click on one of the import options e.g. "Illumina". This will make a wizard appear as shown in figure 10.17.

3. Locate and select the files to import. Note that you can select all sequence files and import them simultaneously. If you take a closer look at the different options in this wizard, you can see that it is possible to choose different import options. We recommend to import data with the standard settings. If you wish to make your own adjustments, you can find further details about the import options in the CLC Cancer Research Workbench reference manual (`http://clccancer.com/software/#downloads`).

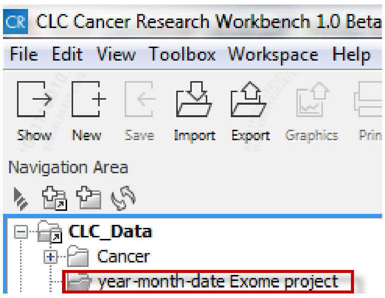

CLC Cancer Research Workbench 1.0 Beta
File Edit View Toolbox Workspace Help
Show New Save Import Export Graphics Prin
Navigation Area
CLC_Data
Cancer
year-month-date Exome project

Figure 10.14: *The folder that you have just created will be placed in the CLC_Data location.*

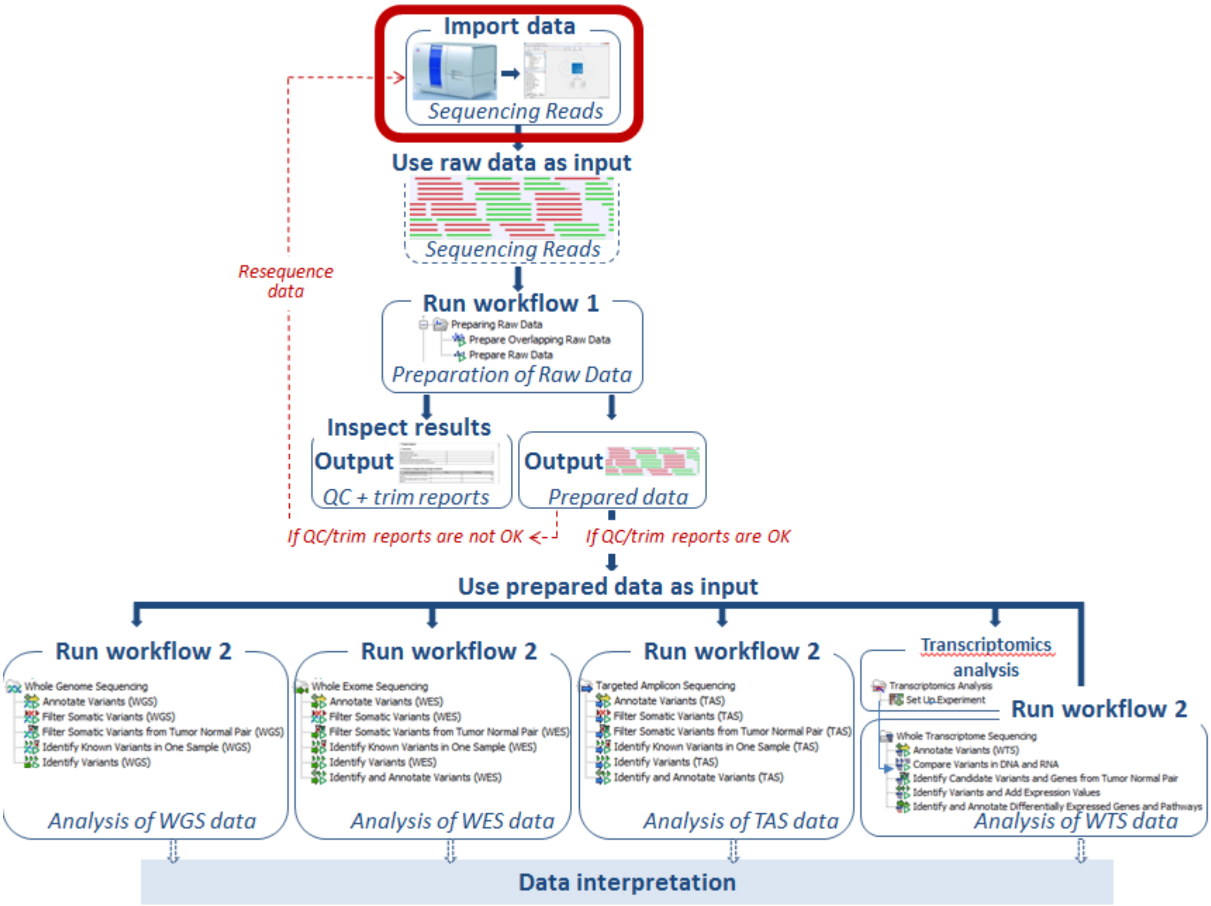

Import data
Sequencing Reads
Use raw data as input
Sequencing Reads
Resequence data
Run workflow 1
Preparing Raw Data
Prepare Overlapping Raw Data
Prepare Raw Data
Preparation of Raw Data
Inspect results
Output
QC + trim reports
Output
Prepared data
If QC/trim reports are not OK
If QC/trim reports are OK
Use prepared data as input
Run workflow 2
Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Filter Somatic Variants from Tumor Normal Pair (WGS)
Identify Known Variants in One Sample (WGS)
Identify Variants (WGS)
Analysis of WGS data
Run workflow 2
Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Filter Somatic Variants from Tumor Normal Pair (WES)
Identify Known Variants in One Sample (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Analysis of WES data
Run workflow 2
Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Filter Somatic Variants from Tumor Normal Pair (TAS)
Identify Known Variants in One Sample (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)
Analysis of TAS data
Transcriptomics analysis
Transcriptomics Analysis
Set Up Experiment
Run workflow 2
Whole Transcriptome Sequencing
Annotate Variants (WTS)
Compare Variants in DNA and RNA
Identify Candidate Variants and Genes from Tumor Normal Pair
Identify Variants and Add Expression Values
Identify and Annotate Differentially Expressed Genes and Pathways
Analysis of WTS data
Data interpretation

Figure 10.15: *The first thing to do is to import your sequencing data.*

4. Click on the button labeled **Next**. This will take you to the next wizard step (see figure 10.18).

5. Choose the default settings to save the sequence data and click on the button labeled **Next**. This will take you to the wizard step shown in figure 10.19.

6. Locate the folder in the Navigation Area that you have created for the purpose.

7. Click on the button labeled **Finish**. It can take some seconds or even minutes before all data have been imported and saved.

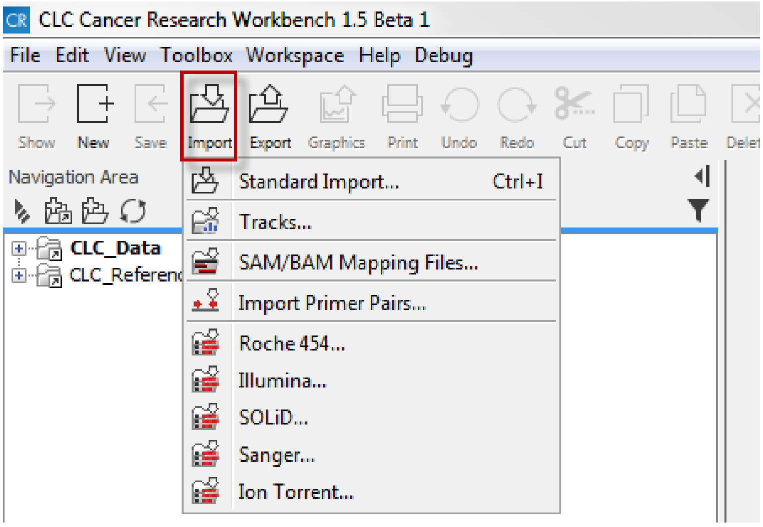
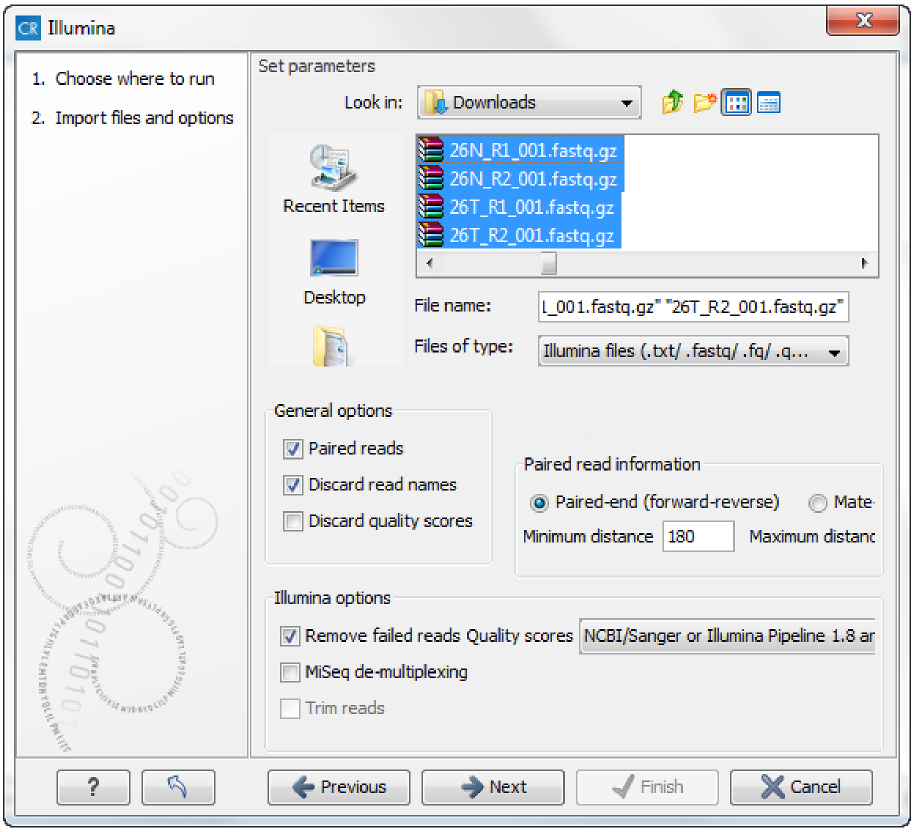

CLC Cancer Research Workbench 1.5 Beta 1
File Edit View Toolbox Workspace Help Debug
Show New Save Import Export Graphics Print Undo Redo Cut Copy Paste Delet
Navigation Area
CLC_Data
CLC_Referenc
Standard Import... Ctrl+I
Tracks...
SAM/BAM Mapping Files...
Import Primer Pairs...
Roche 454...
Illumina...
SOLiD...
Sanger...
Ion Torrent...

Figure 10.16: *Click on the tool labeled "Import" in the toolbar to import data. Select importer according to the data type you wish to import.*


Illumina

1. Choose where to run
2. Import files and options

Set parameters

Look in: Downloads

26N_R1_001.fastq.gz
26N_R2_001.fastq.gz
26T_R1_001.fastq.gz
26T_R2_001.fastq.gz

File name: L_001.fastq.gz" "26T_R2_001.fastq.gz"

Files of type: Illumina files (.txt/ .fastq/ .fq/ .q...

General options

☑ Paired reads
☑ Discard read names
□ Discard quality scores

Paired read information

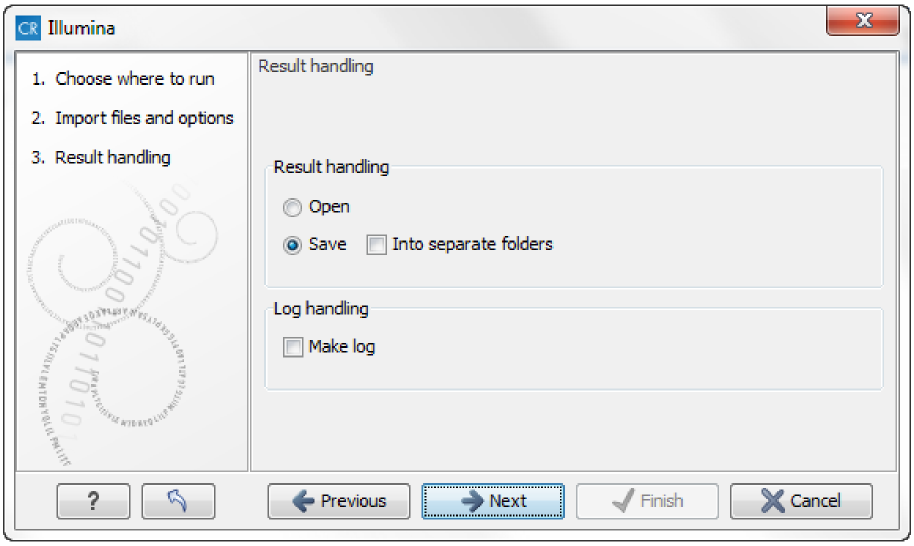
● Paired-end (forward-reverse) ○ Mate

Minimum distance 180 Maximum distanc

Illumina options

☑ Remove failed reads Quality scores NCBI/Sanger or Illumina Pipeline 1.8 ar
□ MiSeq de-multiplexing
□ Trim reads

Figure 10.17: *Locate and select the files to import. Tick "Paired reads" if you, as in this example, are importing paired reads.*


Illumina

1. Choose where to run
2. Import files and options
3. Result handling

Result handling

Result handling

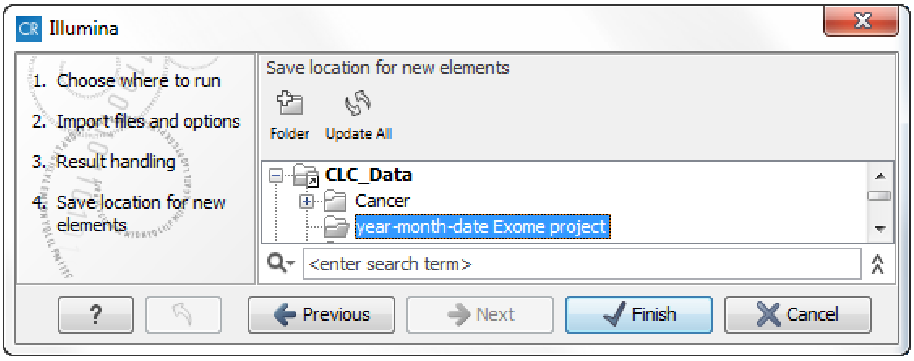
○ Open

● Save □ Into separate folders

Log handling

□ Make log

Figure 10.18: *You now have the option to choose whether you wish to open or save the imported reads. If you select to open the reads, they will not be saved unless you do it manually at a later point. Select "Save" and click on the button labeled "Next".*


Illumina

1. Choose where to run
2. Import files and options
3. Result handling
4. Save location for new elements

Save location for new elements

CLC_Data
Cancer
year-month-date Exome project

<enter search term>

Figure 10.19: *Locate the folder in the Navigation Area that you have just created and save your imported reads in the folder.*

Chapter 11

Preparing Raw Data

Contents

11.1 Prepare sequencing data - all application types

The first thing to do after data import is to check the quality of the sequencing reads and perform the necessary trimming. This applies no matter whether you are working with Whole Genome Sequencing, Exome Sequencing, or Targeted Amplicon Sequencing. In the toolbox you can choose between the two different ready-to-use workflows for data preparation that are shown in the "Run workflow 1" box in figure 11.1.

The "Preparing Raw Data" ready-to-use workflows are universal and can be used for all applications; Whole Genome Sequencing, Exome Sequencing, and Targeted Amplicon Sequencing.

Choosing between "Prepare Raw Data" and "Prepare Overlapping Raw Data" workflows:

Many whole genome sequencing, exome sequencing using capture technology, and targeted amplicon sequencing strategies produce overlapping reads. Downstream stages of the Cancer Research Workbench (e.g. Variant calling) take the frequencies of observed alleles into consideration as well as the forward-reverse strand balance. When merging overlapping reads these two parameters will be affected: 1) the frequency of observed alleles in overlapping regions will be corrected (a variant found both on the forward and the reverse read of the same fragment should only be counted once), and 2) in the merged fragments the information on forward-reverse strand origin has become meaningless. These effects have to be taken into consideration when filtering variants on these statistics. As the forward-reverse strand balance statistic is used as a variant filter (i.e. the Read direction filter), we recommend using the "Prepare Overlapping Raw

Data" workflow on targeted amplicon sequencing data with overlapping read sequencing strategy, whereas we recommend the "Prepare Raw Data" workflow for other sequencing protocols (e.g. whole genome sequencing, whole exome-sequencing, also if making use of overlapping read sequencing).

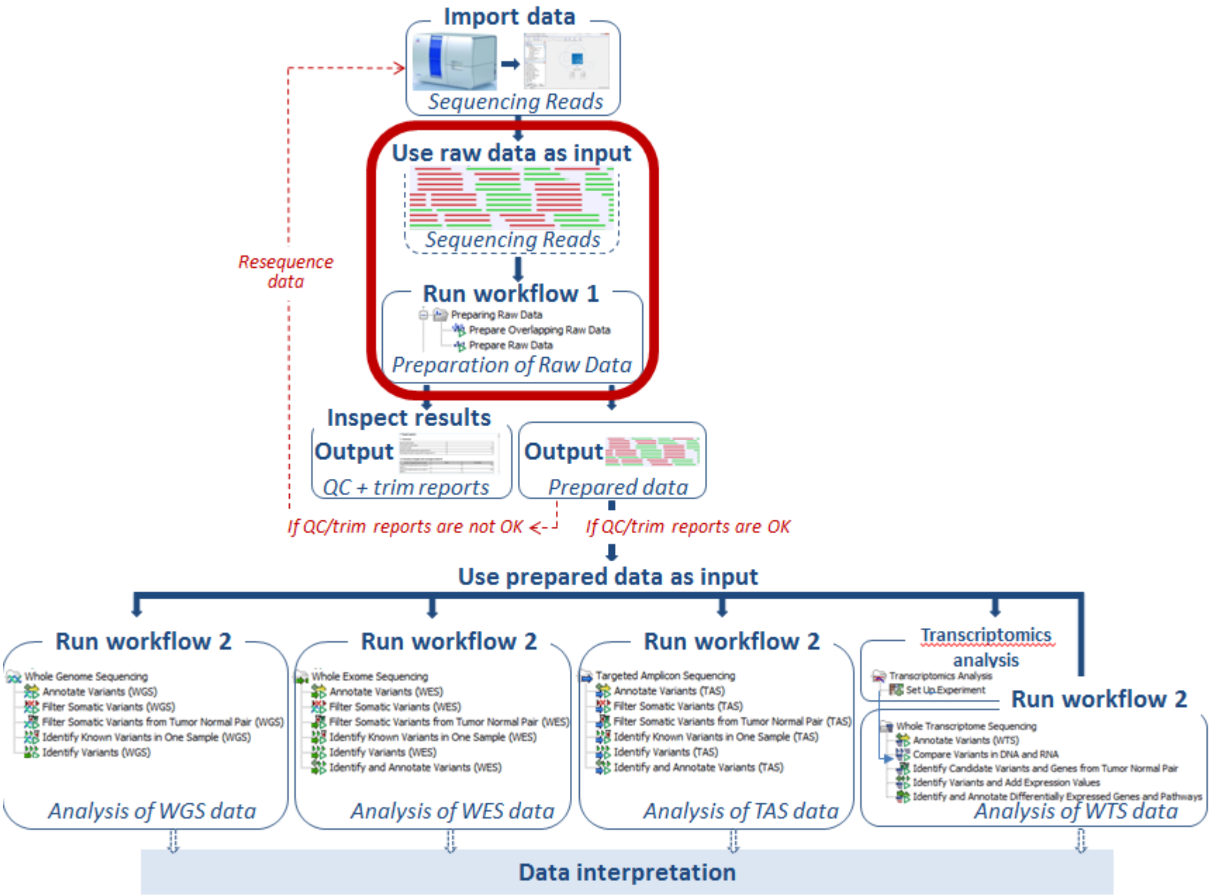
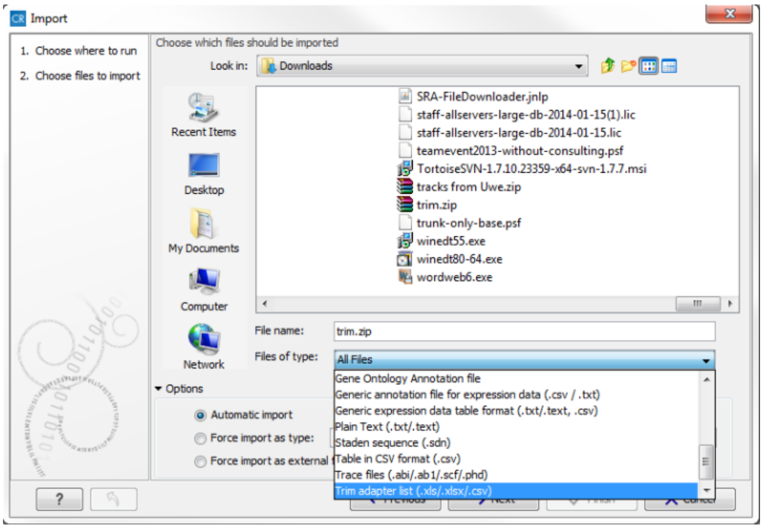

Import data
Sequencing Reads
Use raw data as input
Sequencing Reads
Resequence data
Run workflow 1
Preparing Raw Data
Prepare Overlapping Raw Data
Prepare Raw Data
Preparation of Raw Data
Inspect results
Output
QC + trim reports
Output
Prepared data
If QC/trim reports are not OK
If QC/trim reports are OK
Use prepared data as input
Run workflow 2
Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Filter Somatic Variants from Tumor Normal Pair (WGS)
Identify Known Variants in One Sample (WGS)
Identify Variants (WGS)
Analysis of WGS data
Run workflow 2
Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Filter Somatic Variants from Tumor Normal Pair (WES)
Identify Known Variants in One Sample (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Analysis of WES data
Run workflow 2
Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Filter Somatic Variants from Tumor Normal Pair (TAS)
Identify Known Variants in One Sample (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)
Analysis of TAS data
Transcriptomics analysis
Transcriptomics Analysis
Set Up Experiment
Run workflow 2
Whole Transcriptome Sequencing
Annotate Variants (WTS)
Compare Variants in DNA and RNA
Identify Candidate Variants and Genes from Tumor Normal Pair
Identify Variants and Add Expression Values
Identify and Annotate Differentially Expressed Genes and Pathways
Analysis of WTS data
Data interpretation

Figure 11.1: *Two ready-to-use workflows are available for data preparation; "Prepare Overlapping Raw Data" and "Prepare Raw data".*

11.1.1 Import adapter trim list

One important part of the preparation of raw data is adapter trimming. To be able to trim off the adaptors, an adapter trim list is required. To obtain this file you will have to get in contact with the vendor and ask them to send this adapter trim list file to you. As the adapter trim list has been supplied by the vendor of the enrichment kit and sequencing machine, the adapter trim list must be imported into the *CLC Cancer Research Workbench*. The adapter trim list can be imported by clicking on the button labeled "Import" in the **Toolbar**. Select standard import (figure 11.2) and find the adapter trim list you want to import.

Select "Trim adapter list (.xls, .xlsx/.csv)" in the "Files of type" drop-down list in the Import wizard. Click on the button labeled **Next** and select where you wish to save the adapter trim list.



Import

1. Choose where to run
2. Choose files to import

Choose which files should be imported

Look in: Downloads

- SRA-FileDownloader.jnlp
- staff-allservers-large-db-2014-01-15(1).lic
- staff-allservers-large-db-2014-01-15.lic
- teamevent2013-without-consulting.psf
- TortoiseSVN-1.7.10.23359-x64-svn-1.7.7.msi
- tracks from Uwe.zip
- trim.zip
- trunk-only-base.psf
- winedt55.exe
- winedt80-64.exe
- wordweb6.exe

File name: trim.zip

Files of type: All Files

- Gene Ontology Annotation file
- Generic annotation file for expression data (.csv / .txt)
- Generic expression data table format (.txt/.text, .csv)
- Plain Text (.txt/.text)
- Staden sequence (.sdn)
- Table in CSV format (.csv)
- Trace files (.abi/.ab1/.scf/.phd)
- Trim adapter list (.xls/.xlsx/.csv)

Options

● Automatic import
○ Force import as type:
○ Force import as external

Figure 11.2: *After you have identified the trim list that you want to import, select "Trim adapter list (.xls, .xlsx/.csv)" in the "Files of type" drop-down list in the Import wizard.*

11.1.2 How to run the "Prepare Overlapping Raw Data" ready-to-use workflow

If your sequencing reads contain overlapping pairs you can use the "Prepare Overlapping Raw Data" ready-to-use workflow for preparation of your sequences before you proceed to data analysis such as variant calling.

1. Go to the toolbox and double-click on the "Prepare Overlapping Raw Data" ready-to-use workflow (figure 11.3).

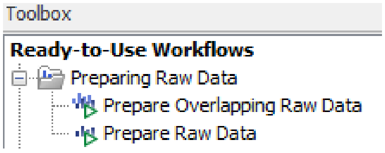
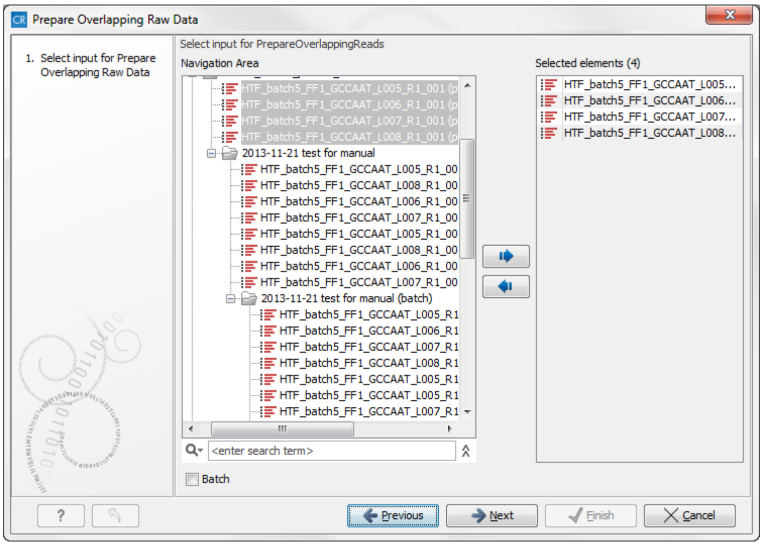


Figure 11.3: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard shown in figure 11.4 where you can select the reads that you wish to prepare for further analyses.

At this step you can choose to prepare one sample at the time or you can select several samples and prepare them simultaneously. If you choose to select more than one sample you can choose to select multiple samples and use the small arrow pointing to the right side in the middle of the wizard to send them to "Selected elements" in the right side of the wizard. Alternatively you can run the samples in "Batch" mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 11.4) and select the **folder** that holds the data you wish to analyze. If your sequencing data are found in separate folders, you should choose to run the analysis in batch mode.

The difference between analyzing multiple samples in batch mode versus in non-batch mode is the reporting. If you use batch mode, you will get an individual report for every single sample whereas you will get one combined report for all samples if you do not run in batch mode.


Prepare Overlapping Raw Data
1. Select input for Prepare Overlapping Raw Data
Select input for PrepareOverlappingReads
Navigation Area
Selected elements (4)
<enter search term>
Previous
Next
Finish
Cancel

HTF_batch5_FF1_GCCAAT_L005_R1_001 (p
HTF_batch5_FF1_GCCAAT_L006_R1_001 (p
HTF_batch5_FF1_GCCAAT_L007_R1_001 (p
HTF_batch5_FF1_GCCAAT_L008_R1_001 (p
2013-11-21 test for manual
HTF_batch5_FF1_GCCAAT_L005_R1_00
HTF_batch5_FF1_GCCAAT_L008_R1_00
HTF_batch5_FF1_GCCAAT_L006_R1_00
HTF_batch5_FF1_GCCAAT_L007_R1_00
HTF_batch5_FF1_GCCAAT_L005_R1_00
HTF_batch5_FF1_GCCAAT_L008_R1_00
HTF_batch5_FF1_GCCAAT_L006_R1_00
HTF_batch5_FF1_GCCAAT_L007_R1_00
2013-11-21 test for manual (batch)
HTF_batch5_FF1_GCCAAT_L005_R1
HTF_batch5_FF1_GCCAAT_L006_R1
HTF_batch5_FF1_GCCAAT_L007_R1
HTF_batch5_FF1_GCCAAT_L008_R1
HTF_batch5_FF1_GCCAAT_L005_R1
HTF_batch5_FF1_GCCAAT_L005_R1
HTF_batch5_FF1_GCCAAT_L007_R1

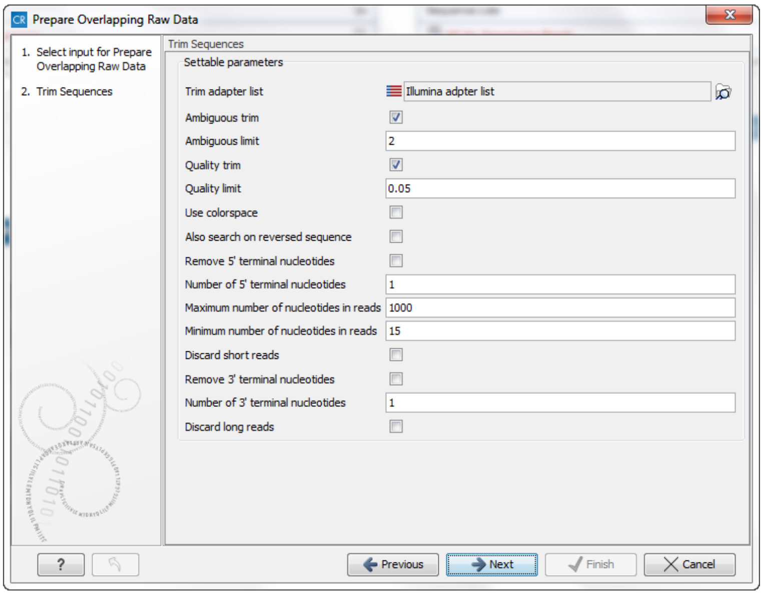
Selected elements (4):
HTF_batch5_FF1_GCCAAT_L005...
HTF_batch5_FF1_GCCAAT_L006...
HTF_batch5_FF1_GCCAAT_L007...
HTF_batch5_FF1_GCCAAT_L008...

□ Batch

Figure 11.4: *Select the sequencing raw data that should be prepared for further analysis. At this step you can also choose to prepare several reads in batch mode.*

When you have selected the sample(s) you want to prepare, click on the button labeled **Next**.

2. As part of the data preparation, the sequences are trimmed. In the wizard shown in figure 11.5 you can specify different trimming parameters and select the adapter trim list that should be used for adapter trimming by clicking on the folder icon ( ).

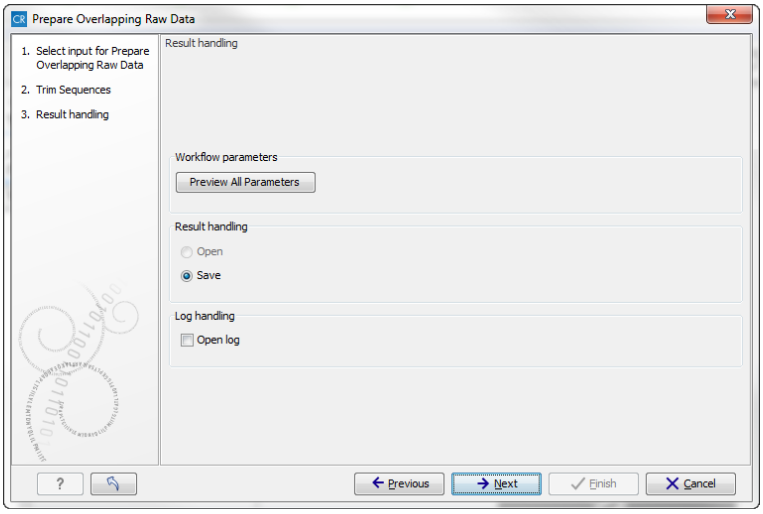

Prepare Overlapping Raw Data
1. Select input for Prepare Overlapping Raw Data
2. Trim Sequences
Trim Sequences
Settable parameters
Previous
Next
Finish
Cancel

Trim adapter list	Illumina adpter list
Ambiguous trim	☑
Ambiguous limit	2
Quality trim	☑
Quality limit	0.05
Use colorspace	□
Also search on reversed sequence	□
Remove 5' terminal nucleotides	□
Number of 5' terminal nucleotides	1
Maximum number of nucleotides in reads	1000
Minimum number of nucleotides in reads	15
Discard short reads	□
Remove 3' terminal nucleotides	□
Number of 3' terminal nucleotides	1
Discard long reads	□

Figure 11.5: *Select your adapter trim list. You can use the default trim parameters or adjust them if necessary.*

3. Click on the button labeled **Next**. This will take you to the next wizard step (figure 11.6).

   At this step you get the chance to check the selected settings by clicking on the button labeled **Preview All Parameters** (figure 11.7).



Prepare Overlapping Raw Data

1. Select input for Prepare Overlapping Raw Data
2. Trim Sequences
3. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

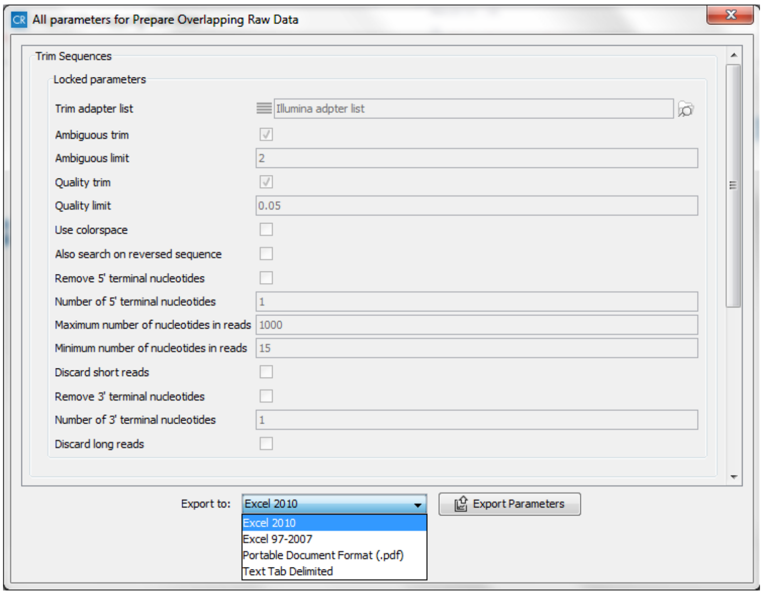
○ Open

● Save

Log handling

□ Open log

Figure 11.6: *Check the settings and save your results.*



All parameters for Prepare Overlapping Raw Data

Trim Sequences

Locked parameters

Trim adapter list	Illumina adpter list
Ambiguous trim	☑
Ambiguous limit	2
Quality trim	☑
Quality limit	0.05
Use colorspace	□
Also search on reversed sequence	□
Remove 5' terminal nucleotides	□
Number of 5' terminal nucleotides	1
Maximum number of nucleotides in reads	1000
Minimum number of nucleotides in reads	15
Discard short reads	□
Remove 3' terminal nucleotides	□
Number of 3' terminal nucleotides	1
Discard long reads	□

Export to: Excel 2010

Figure 11.7: *In this wizard you can check the parameter settings. It is also possible to export the settings to a file format that can be specified using the "Export to" drop-down list.*

In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of the wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

4. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

11.1.3 How to run the "Prepare Raw Data" ready-to-use workflow

If you have sequencing reads without overlapping pairs, you can use the "Prepare Raw Data" ready-to-use workflow for preparation of your sequences before you proceed to data analysis such as variant calling.

1. Go to the toolbox and double-click on the "Prepare Raw Data" ready-to-use workflow (figure 11.8).

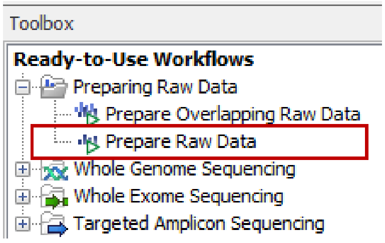


Figure 11.8: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard shown in figure 11.9 where you can select the reads that you wish to prepare for further analyses.

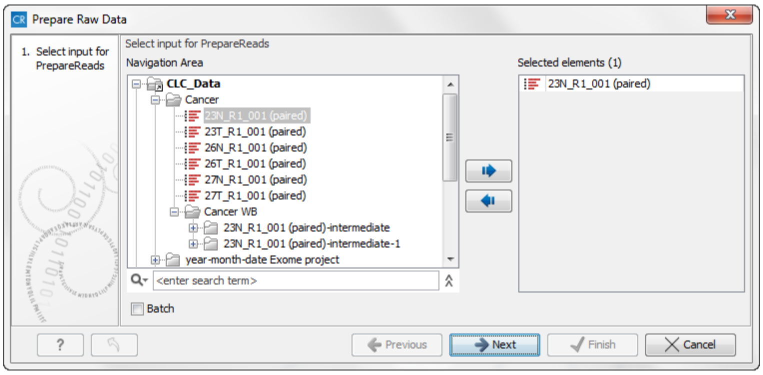


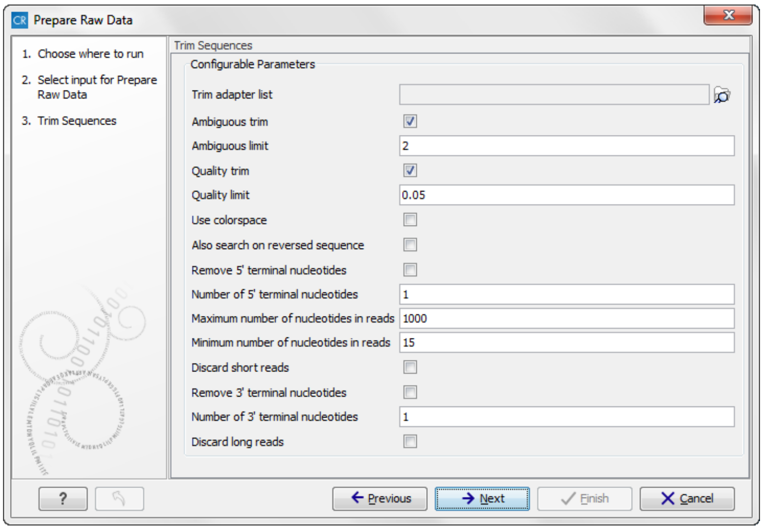
Figure 11.9: *Select the sequencing raw data that you wish to prepare before further analysis. At this step you can also choose whether you wish to prepare several reads in batch mode.*

At this step you can choose to prepare one sample at the time or you can select several samples and prepare them simultaneously. If you choose to select more than one sample you can choose to either select multiple samples and use the small arrow to send them to the "Selected elements" in the right side of the wizard. Alternatively you can run the samples in "batch mode". This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 11.4) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

The difference between analyzing multiple samples in batch mode versus in non-batch mode is the reporting. If you use batch mode, you will get an individual report for every single sample whereas you will get one combined report for all samples if you do not run in batch mode.

2. When you have selected the sample(s) you want to prepare, click on the button labeled **Next**.

As part of the data preparation, the sequences are trimmed. In the next wizard (figure 11.10) you can specify different trimming parameters and select the adapter trim list that should be used for adapter trimming by clicking on the folder icon (🔍). To obtain this file you will have to get in contact with the vendor and ask them to send this adapter trim list file to you. The adapter trim list has been supplied by the vendor of the enrichment kit and sequencing machine. See section 11.1.1 for a description of how to import the adapter trim list.


Prepare Raw Data

1. Choose where to run
2. Select input for Prepare Raw Data
3. Trim Sequences

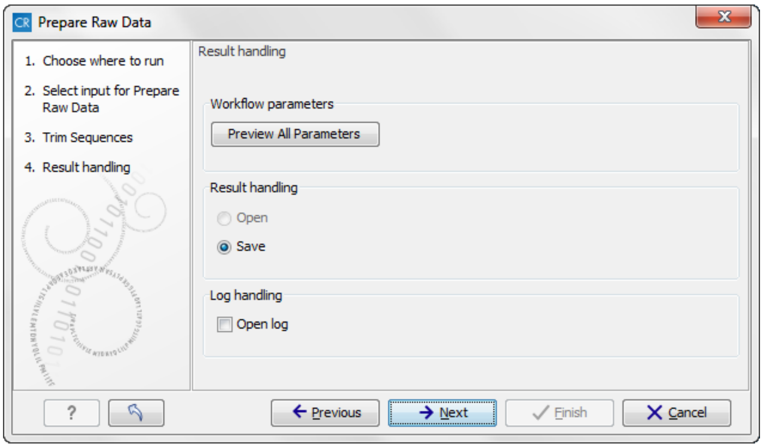
Trim Sequences

Configurable Parameters

Trim adapter list	
Ambiguous trim	☑
Ambiguous limit	2
Quality trim	☑
Quality limit	0.05
Use colorspace	□
Also search on reversed sequence	□
Remove 5' terminal nucleotides	□
Number of 5' terminal nucleotides	1
Maximum number of nucleotides in reads	1000
Minimum number of nucleotides in reads	15
Discard short reads	□
Remove 3' terminal nucleotides	□
Number of 3' terminal nucleotides	1
Discard long reads	□

Figure 11.10: *Select your adapter trim list. You can use the default trim parameters or adjust them if necessary.*

3. Click on the button labeled **Next**, which will take you to the next wizard (figure 11.11).


Prepare Raw Data

1. Choose where to run
2. Select input for Prepare Raw Data
3. Trim Sequences
4. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open
● Save

Log handling

□ Open log

Figure 11.11: *Check the settings and save your results.*

If you click on the button labeled **Preview All Parameters** you get the chance to check the selected settings. At this step you can only check the settings, it is not possible to make any changes at this point.

The settings can be exported with the two buttons found at the bottom of this wizard; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

4. Click on the button labeled **OK** to go back to the previous wizard and choose **Save**.

11.1.4 Output from the Prepare Overlapping Raw Data and Prepare Raw Data workflows

Different outputs are generated from the "Prepare Overlapping Raw Data" and "Prepare Raw Data" workflows.

Prepare Overlapping Raw Data. Performs quality control and trimming of the sequencing reads and merging of overlapping read pairs and generates five different outputs:

1. QC graphic report. The report should be checked by the user.
2. QC supplementary report. The report should be checked by the user.
3. Trimming report (the trimmed sequences are automatically used as input in the merging of paired reads step). The report should be checked by the user.
4. Merged reads output. Use as input together with the "Not merged reads output" in the next ready-to-use workflow (e.g. "Identify Variants WES").
5. Not merged reads output. These should be used as input together with the "Merged reads output" in the next ready-to-use workflow (e.g. "Identify Variants WES").

Prepare Raw Data. Performs quality control and trimming of the sequencing reads and generates five different outputs:

1. QC graphic report. The report should be checked by the user.
2. QC supplementary report. The report should be checked by the user.
3. Trimming report. The report should be checked by the user.
4. Trimmed sequences output. Use as input together with the "Trimmed sequences (broken pairs) output" in the next ready-to-use workflow (e.g. "Identify Variants WES").
5. Trimmed sequences (broken pairs) output. Use as input together with the "Trimmed sequences output" in the next ready-to-use workflow (e.g. "Identify Variants WES").

11.1.5 How to check the output reports

Three different reports are generated, and all of these should be inspected in order to determine whether the quality of the sequencing reads and the trimming is acceptable. We are now at the "Inspect results" step in figure 11.12. The interpretation of the reports is not always completely straightforward, but as you gain experience it becomes easier.

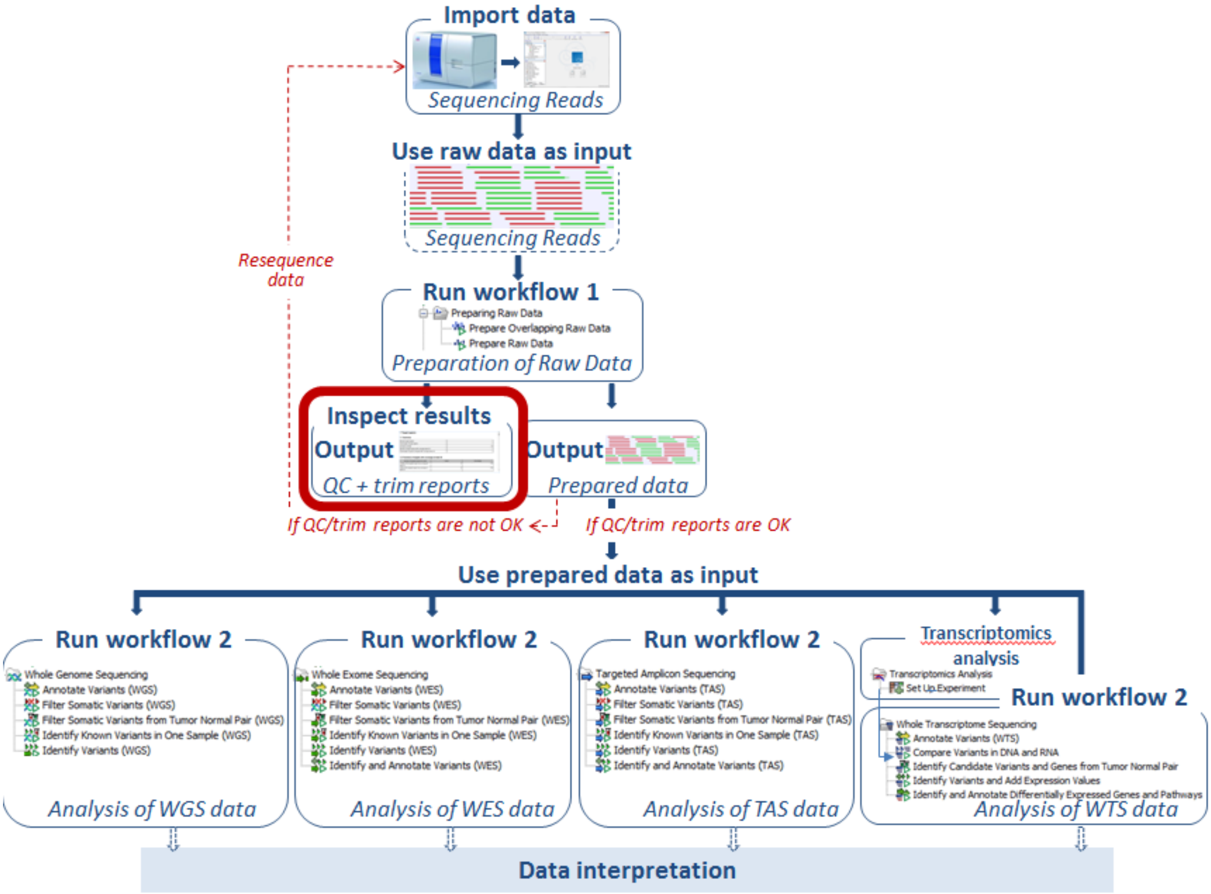

Import data
Sequencing Reads
Use raw data as input
Sequencing Reads
Resequence data
Run workflow 1
Preparing Raw Data
Prepare Overlapping Raw Data
Prepare Raw Data
Preparation of Raw Data
Inspect results
Output
QC + trim reports
Output
Prepared data
If QC/trim reports are not OK
If QC/trim reports are OK
Use prepared data as input
Run workflow 2
Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Filter Somatic Variants from Tumor Normal Pair (WGS)
Identify Known Variants in One Sample (WGS)
Identify Variants (WGS)
Analysis of WGS data
Run workflow 2
Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Filter Somatic Variants from Tumor Normal Pair (WES)
Identify Known Variants in One Sample (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Analysis of WES data
Run workflow 2
Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Filter Somatic Variants from Tumor Normal Pair (TAS)
Identify Known Variants in One Sample (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)
Analysis of TAS data
Transcriptomics analysis
Transcriptomics Analysis
Set Up Experiment
Run workflow 2
Whole Transcriptome Sequencing
Annotate Variants (WTS)
Compare Variants in DNA and RNA
Identify Candidate Variants and Genes from Tumor Normal Pair
Identify Variants and Add Expression Values
Identify and Annotate Differentially Expressed Genes and Pathways
Analysis of WTS data
Data interpretation

Figure 11.12: *Inspect the quality and trimming reports and determine whether you can proceed with the data analysis or if you have to resequence some of the samples.*

Graphical QC Report

- **1 Summary**
- **2 Per-sequence analysis**
- *2.1 Lengths distribution*
- *2.2 GC-content*
- *2.3 Ambiguous base-content*
- *2.4 Quality distribution*
- **3 Per-base analysis**
- *3.1 Coverage*
- *3.2 Nucleotide distributions*
- *3.3 GC-content*
- *3.4 Ambiguous base-content*

- *3.5 Quality distribution*
- **4 Over-representation analyses**
- *4.1 Enriched 5mers*
- *4.2 Sequence duplication levels*
- *4.3 Duplicated sequences*

Supplementary QC Report

- **1 Summary**
- **2 Per-sequence analysis**
- *2.1 Lengths distribution*
- *2.2 GC-content*
- *2.3 Ambiguous base-content*
- *2.4 Quality distribution*
- **3 Per-base analysis**
- *3.1 Coverage*
- *3.2 Nucleotide distributions*
- *3.3 GC-content*
- *3.4 Ambiguous base-content*
- *3.5 Quality distribution*
- **4 Over-representation analyses**
- *4.1 Enriched 5mers*
- *4.2 Sequence duplication levels*
- *4.3 Duplicated sequences*

The majority of the reads should have a PHRED score above 30 when looking at the "Quality distribution" graph.

If you can accept the read quality you can now proceed to the next step and use the prepared reads output as input in the next ready-to-use workflow. If the quality of your reads is poor and cannot be accepted for further analysis, the best solution to the problem is to go back to start and resequence the sample.

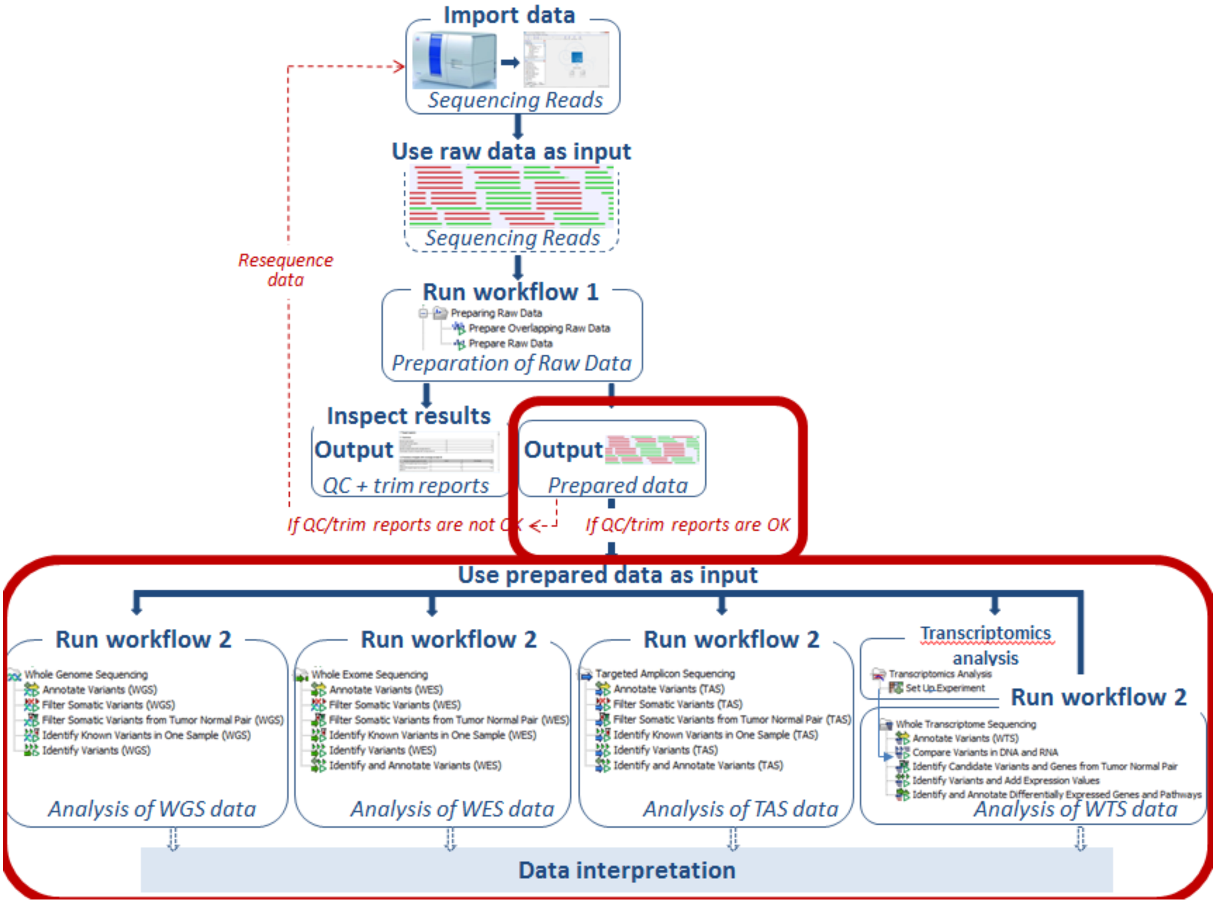

Import data
Sequencing Reads
Use raw data as input
Sequencing Reads
Resequence data
Run workflow 1
Preparing Raw Data
Prepare Overlapping Raw Data
Prepare Raw Data
Preparation of Raw Data
Inspect results
Output
QC + trim reports
Output
Prepared data
If QC/trim reports are not OK
If QC/trim reports are OK
Use prepared data as input
Run workflow 2
Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Filter Somatic Variants from Tumor Normal Pair (WGS)
Identify Known Variants in One Sample (WGS)
Identify Variants (WGS)
Analysis of WGS data
Run workflow 2
Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Filter Somatic Variants from Tumor Normal Pair (WES)
Identify Known Variants in One Sample (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Analysis of WES data
Run workflow 2
Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Filter Somatic Variants from Tumor Normal Pair (TAS)
Identify Known Variants in One Sample (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)
Analysis of TAS data
Transcriptomics analysis
Transcriptomics Analysis
Set Up Experiment
Run workflow 2
Whole Transcriptome Sequencing
Annotate Variants (WTS)
Compare Variants in DNA and RNA
Identify Candidate Variants and Genes from Tumor Normal Pair
Identify Variants and Add Expression Values
Identify and Annotate Differentially Expressed Genes and Pathways
Analysis of WTS data
Data interpretation

Figure 11.13: *Use the prepared data as input in the relevant ready-to-use workflow, which we here for the sake of simplicity call "Workflow 2".*

11.2 Analysis of sequencing data

You are now ready to perform the actual analysis of your sequencing data (see figure 11.13).

For each application six different ready-to-use workflows are available. These can be divided into three different categories; "Data analysis", "Interpretation", and "Data analysis and Interpretation".

Note! The ready-to-use workflows found under each of the three application types have similar names (with the only difference that "WGS", "WES", or "TAS" have been added after the name). However, some of the workflows have been tailored to the individual applications. Therefore, we recommend that you use the ready-to-use workflow that is found under the relevant application heading.

- **Data analysis** The data analysis includes read mapping and variant calling. One ready-to-use workflow is available in this category; the **Identify Variants** ready to use workflow.
- **Interpretation** At this step you can annotate, filter and compare the variants, that were identified in the data analysis step.

  The available tools for interpretation are:

  - *Annotate Variants*: Annotates variants with gene names, conservation scores, amino acid changes, and information from clinically relevant databases.
  - *Filter Somatic Variants*: Removes variants outside the target region (only targeted experiments) and common variants present in publicly available databases. Annotates with gene names, conservation scores, and information from clinically relevant databases.
  - *Identify Somatic Variants from Tumor Normal Pair*: Removes germline variants by referring to the control sample read mapping, removes variants outside the target region (in case of a targeted experiment), and annotates with gene names, conservation scores, amino acid changes, and information from clinically relevant databases.

- **Data analysis and Interpretation** With these ready-to-use workflows you can perform the variant calling, annotation, filtering, and/or comparison of variants in one go.

  The available tools for Data analysis and Interpretation are:

  - *Identify and Annotate Variants*: Maps reads to the human reference sequence, does a local realignment, runs quality control for targeted regions, calls variants, removes false positives, and annotates variants with gene names, amino acid changes, conservation scores, and information from different external databases.
  - *Identify Known Variants in One Sample*: Maps sequencing reads and looks for the presence or absence of user-specified variants in the mapping.

Chapter 12

Whole genome sequencing (WGS)

Contents

The most comprehensive sequencing method is whole genome sequencing that allows for identification of genetic variations and somatic mutations across the entire human genome. This type of sequencing encompasses both chromosomal and mitochondrial DNA. The advantage of sequencing the entire genome is that not only the protein-coding regions are sequenced, but information is also provided for regulatory and non-protein-coding regions.

12.1 Automatic analysis of sequencing data (WGS)

Five ready-to-use workflows are available for analysis of whole genome sequencing data. The concept of the pre-installed ready-to-use workflows is that read data are used as input in one end of the workflow and in the other end of the workflow you get a track based genome browser view and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual ready-to-use workflows can be used for and go through step by step how to run the workflows.

Note! Often you will have to prepare data with one of the two **Preparing Raw Data** workflows described in section 11 before you proceed to **Automatic analysis of sequencing data (WGS)**.

12.2 Identify Variants (WGS)

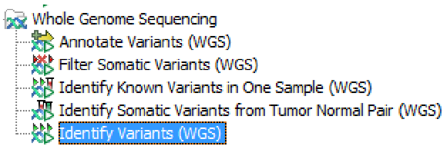
The "Identify Variants" tool takes sequencing reads as input and returns identified variants in a Genome Browser View.

The tool runs an internal workflow that first maps the sequencing reads to the human reference sequence. Next, it runs a local realignment that is used to improve the variant detection that comes after the local realignment. Two different variant callers are used; the "Low Frequency Variant Detection" caller that is used to call small insertions, deletions, SNVs, MNV, and replacements, and the "InDel and Structural Variants" caller that calls larger insertions, deletions, translocations, and replacements. By the end of the variant detection, variants that have been detected by the "Low Frequency Variant Detection" caller with an average base quality smaller than 20 are filtered away.

A detailed mapping report is created to inspect the overall coverage and mapping specificity in the targeted regions.

12.2.1 How to run the "Identify Variants" ready-to-use workflow

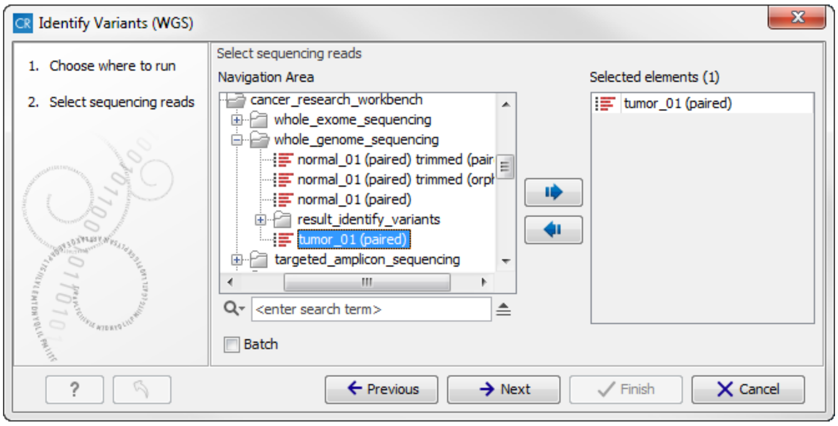
1. Go to the toolbox and double-click on the "Identify Variants" ready-to-use workflow (figure 12.1).



Whole Genome Sequencing
- Annotate Variants (WGS)
- Filter Somatic Variants (WGS)
- Identify Known Variants in One Sample (WGS)
- Identify Somatic Variants from Tumor Normal Pair (WGS)
- Identify Variants (WGS)

Figure 12.1: *Find the "Identify Variants" ready-to-use workflows in the toolbox from the folder that has the name of the application you are using.*

This will open the wizard shown in figure 12.2 where you can select the sequencing reads from the sample that should be analyzed.



Identify Variants (WGS)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

Navigation Area

- cancer_research_workbench
  - whole_exome_sequencing
  - whole_genome_sequencing
    - normal_01 (paired) trimmed (pair
    - normal_01 (paired) trimmed (orpl
    - normal_01 (paired)
    - result_identify_variants
    - tumor_01 (paired)
  - targeted_amplicon_sequencing

<enter search term>

□ Batch

Selected elements (1)

tumor_01 (paired)

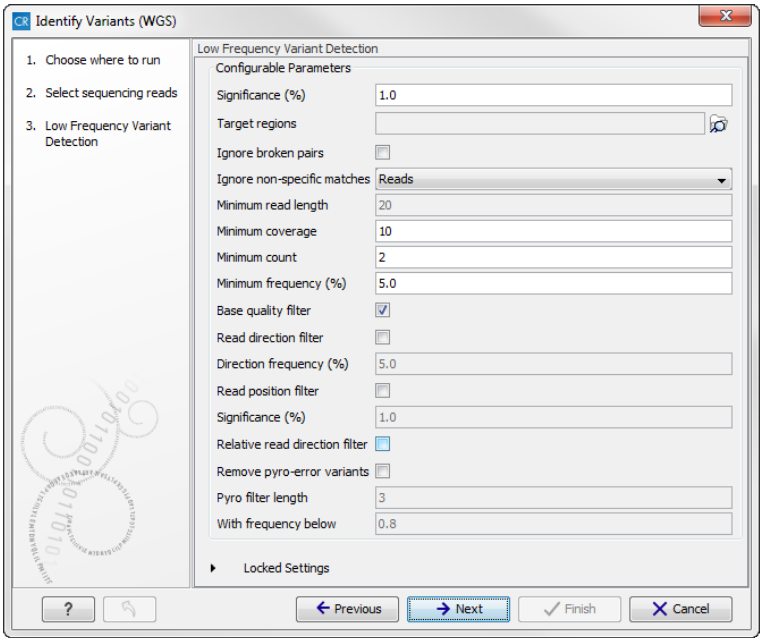
Figure 12.2: *Please select all sequencing reads from the sample to be analyzed.*

Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. To do this, tick "Batch" at the bottom of the wizard and select the **folder** that holds the data you wish to analyze.

If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

When you have selected the sample(s) that you want to prepare, click on the button labeled **Next**.

2. In the next wizard step (figure 12.3) you can specify the parameters for variant detection.



Identify Variants (WGS)

1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection

Low Frequency Variant Detection

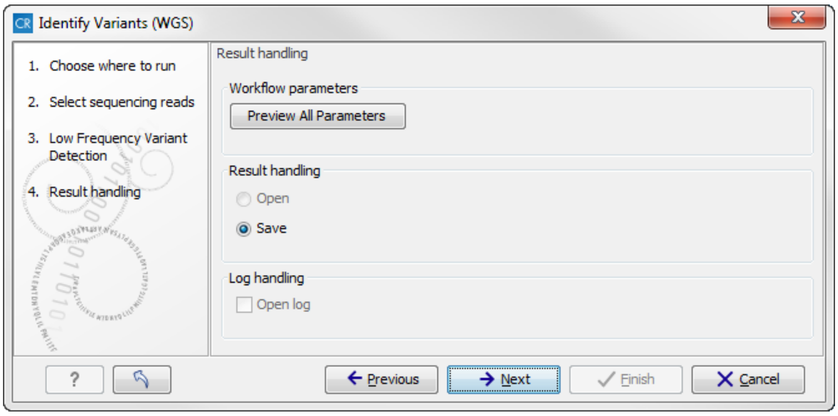
Configurable Parameters

Significance (%) 1.0
Target regions
Ignore broken pairs □
Ignore non-specific matches Reads
Minimum read length 20
Minimum coverage 10
Minimum count 2
Minimum frequency (%) 5.0
Base quality filter ☑
Read direction filter □
Direction frequency (%) 5.0
Read position filter □
Significance (%) 1.0
Relative read direction filter □
Remove pyro-error variants □
Pyro filter length 3
With frequency below 0.8

Locked Settings

Figure 12.3: *The next thing to do is to specify the parameters that should be used to detect variants.*

3. Click on the button labeled **Next**. This will take you to the next wizard step (figure 12.4).



Identify Variants (WGS)

1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection
4. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open
● Save

Log handling

□ Open log

Figure 12.4: *Check the settings and save your results.*

In this wizard you can check the selected settings by clicking on the button labeled **Preview All Parameters**.

In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

4. Click on the button labeled **OK** to go back to the previous wizard and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

12.2.2 Output from the Identify Variants workflow

The "Identify Variants" tool produces six different types of output:

1. **Structural Variants** () Variant track showing the structural variants; insertions, deletions, replacements. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant. The structural variants can also be viewed in table format by switching to the table view. This is done by pressing the table icon found in the lower left corner of the **View Area**.
2. **Structural Variant Report** () The report consists of a number of tables and graphs that in different ways provide information about the structural variants.
3. **Read Mapping** () The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: `http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html`.
4. **Read Mapping Report** () The report consists of a number of tables and graphs that in different ways provide information about the mapped reads.
5. **Structural Variants** () A variant track holding the identified variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
6. **Genome Browser View Identify Variants** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 12.10).

Before looking at the identified variants, we recommend that you first take a look at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Furthermore, please check that at least 90% of the reads map to the human reference sequence. In case of a targeted experiment, please also check that the majority of reads map to the targeted region.

Next, open the Genome Browser View file (see figure 12.5).

The Genome Browser View lists the track of the identified variants in context to the human reference sequence, genes, transcripts, coding regions, and mapped sequencing reads.

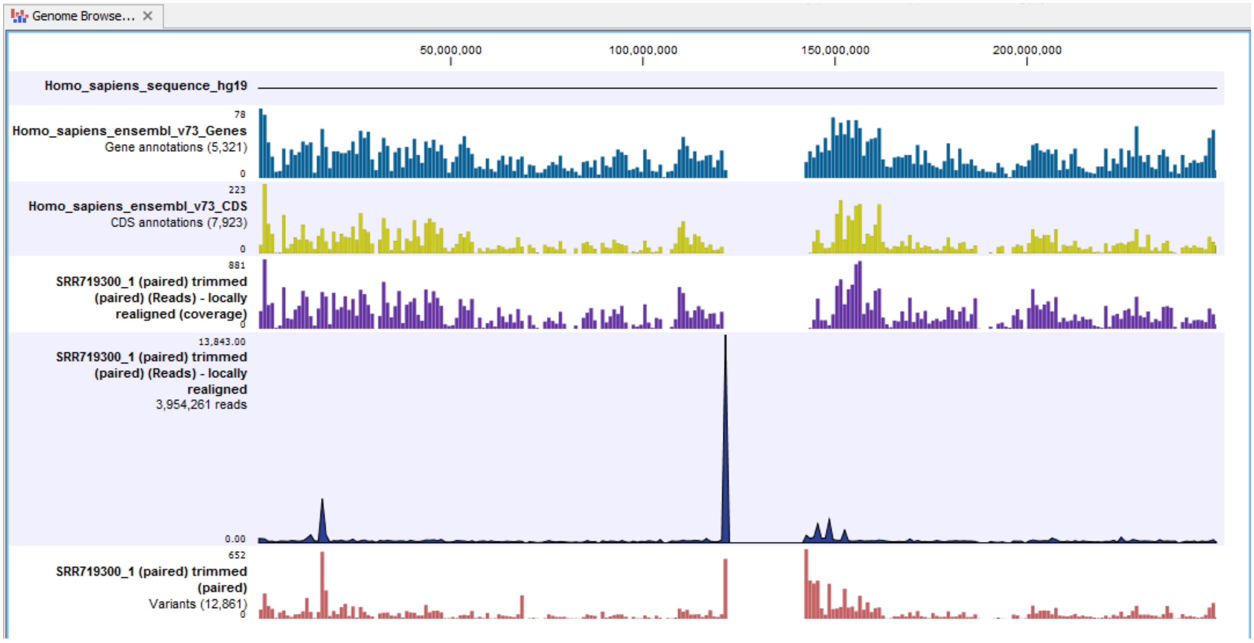

Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
78
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
0
223
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
0
881
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned (coverage)
0
13,843.00
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned
3,954,261 reads
0.00
652
SRR719300_1 (paired) trimmed (paired)
Variants (12,861)
0

Figure 12.5: *The Genome Browser View allows easy inspection of the identified smaller variants, larger insertions and deletions, and structural variants in the context of the human genome.*

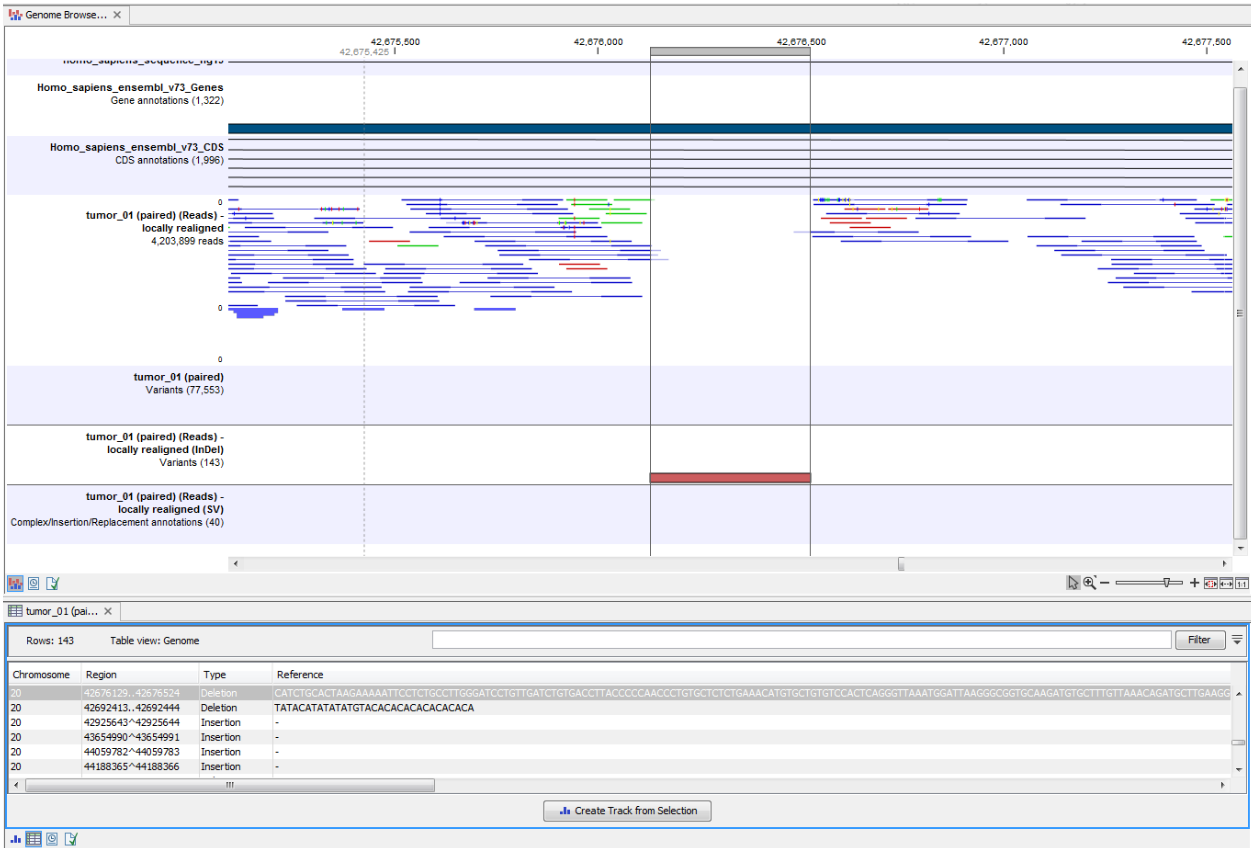
By double-clicking on the InDel variant track in the Genome Browser View, a table will be shown that lists all identified larger insertions and deletions (see figure 12.6).

In case you would like to change the reference sequence used for read mapping or the human genes, please use the "Data Management" (see section 10.1.4).

12.3 Annotate Variants (WGS)

Using a variant track () (e.g. the output from the Identify Variants ready-to-use workflow) the **Annotate Variants (WGS)** ready-to-use workflow runs an internal workflow that adds the following annotations to the variant track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.
- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from COSMIC**. Adds information from the "Catalogue of Somatic Mutations in Cancer" database.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.


Genome Browse...
42,675,500
42,675,425
42,676,000
42,676,500
42,677,000
42,677,500
Homo_sapiens_ensembl_v73_Genes
Gene annotations (1,322)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (1,996)
tumor_01 (paired) (Reads) - locally realigned
4,203,899 reads
tumor_01 (paired)
Variants (77,553)
tumor_01 (paired) (Reads) - locally realigned (InDel)
Variants (143)
tumor_01 (paired) (Reads) - locally realigned (SV)
Complex/Insertion/Replacement annotations (40)
tumor_01 (pai...
Filter
Create Track from Selection

Rows: 143 Table view: Genome

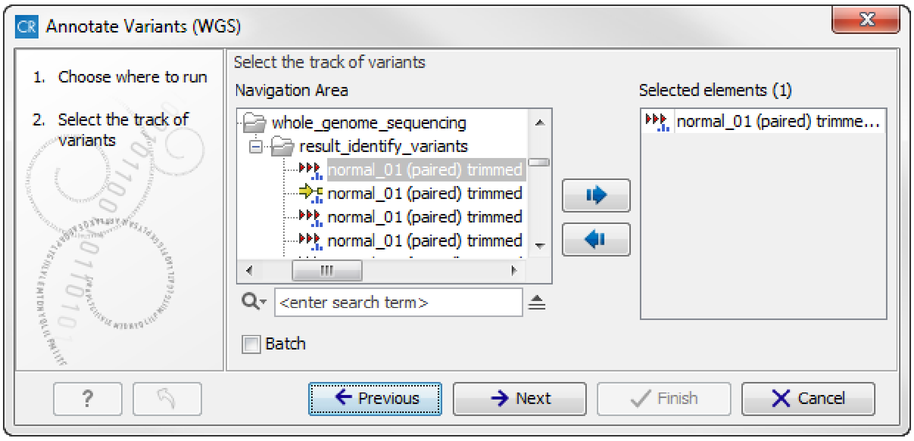
Chromosome	Region	Type	Reference
20	42676129..42676524	Deletion	CATCTGCACTAAGAAAAATTCCTCTGCCTTGGGATCCTGTTGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGAAACATGTGCTGTGTCCACTCAGGGTTAAATGGATTAAGGGCGGTGCAAGATGTGCTTTGTTAAACAGATGCTTGAAGG
20	42692413..42692444	Deletion	TATACATATATATGTACACACACACACACA
20	42925643^42925644	Insertion	-
20	43654990^43654991	Insertion	-
20	44059782^44059783	Insertion	-
20	44188365^44188366	Insertion	-

Figure 12.6: *This figure shows a Genome Browser View with an open track table. The table allows deeper inspection of the identified variants.*

- **Information from dbSNP** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (InDels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

1. Go to the toolbox and select the **Annotate Variants (WGS)** workflow. In the first wizard step, select the input variant track (figure 12.7).
2. Click on the button labeled **Next**. The only parameter that should be specified by the user is which 1000 Genomes population you use (figure 12.8). This can be done using the drop-down list found in this wizard step. Please note that the populations available from the drop-down list can be specified with the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).
3. Click on the button labeled **Next** to go to the last wizard step (figure 12.9).

   In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point.


Annotate Variants (WGS)

1. Choose where to run
2. Select the track of variants

Select the track of variants

Navigation Area

- whole_genome_sequencing
  - result_identify_variants
    - normal_01 (paired) trimmed
    - normal_01 (paired) trimmed
    - normal_01 (paired) trimmed
    - normal_01 (paired) trimmed

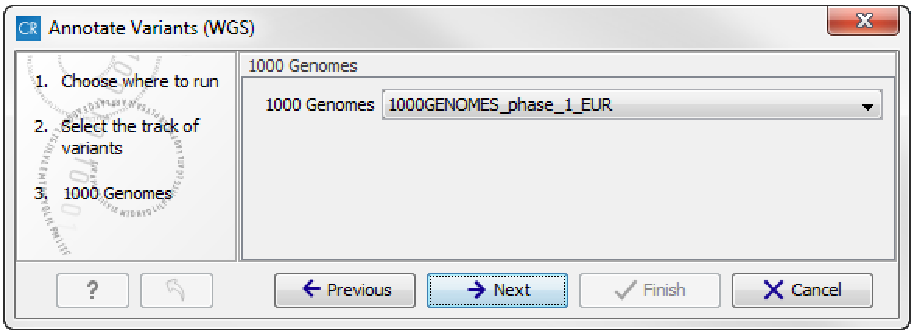
Selected elements (1)

normal_01 (paired) trimme...

<enter search term>

□ Batch

Figure 12.7: *Select the variant track to annotate.*

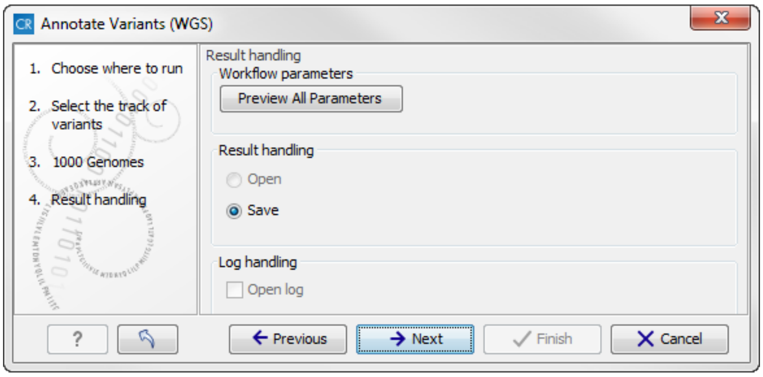

Annotate Variants (WGS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes

1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 12.8: *Select the relevant 1000 Genomes population(s).*


Annotate Variants (WGS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes
4. Result handling

Result handling

Workflow parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 12.9: *Check the settings and save your results.*

4. Choose to **Save** your results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Annotated Variants** ( ) Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. **Genome Browser View Annotated Variants** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes,

transcripts, coding regions, and variants detected in dbSNP, ClinVar, COSMIC, 1000 Genomes, and PhastCons conservation scores (see figure 12.10).

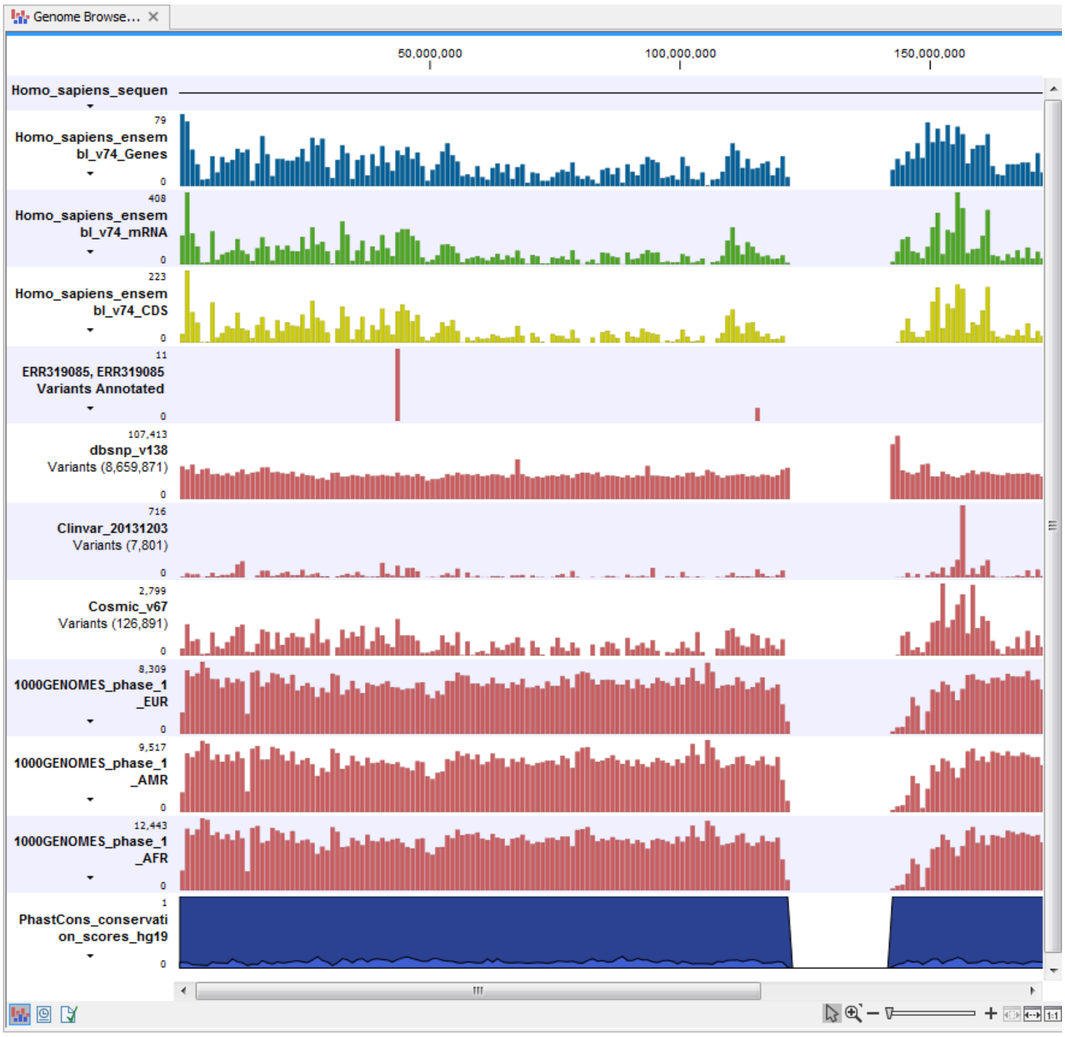

Genome Browse...
50,000,000
100,000,000
150,000,000
Homo_sapiens_sequen
79
Homo_sapiens_ensembl_v74_Genes
0
408
Homo_sapiens_ensembl_v74_mRNA
0
223
Homo_sapiens_ensembl_v74_CDS
0
11
ERR319085, ERR319085 Variants Annotated
0
107,413
dbsnp_v138 Variants (8,659,871)
0
716
Clinvar_20131203 Variants (7,801)
0
2,799
Cosmic_v67 Variants (126,891)
0
8,309
1000GENOMES_phase_1_EUR
0
9,517
1000GENOMES_phase_1_AMR
0
12,443
1000GENOMES_phase_1_AFR
0
1
PhastCons_conservation_scores_hg19
0

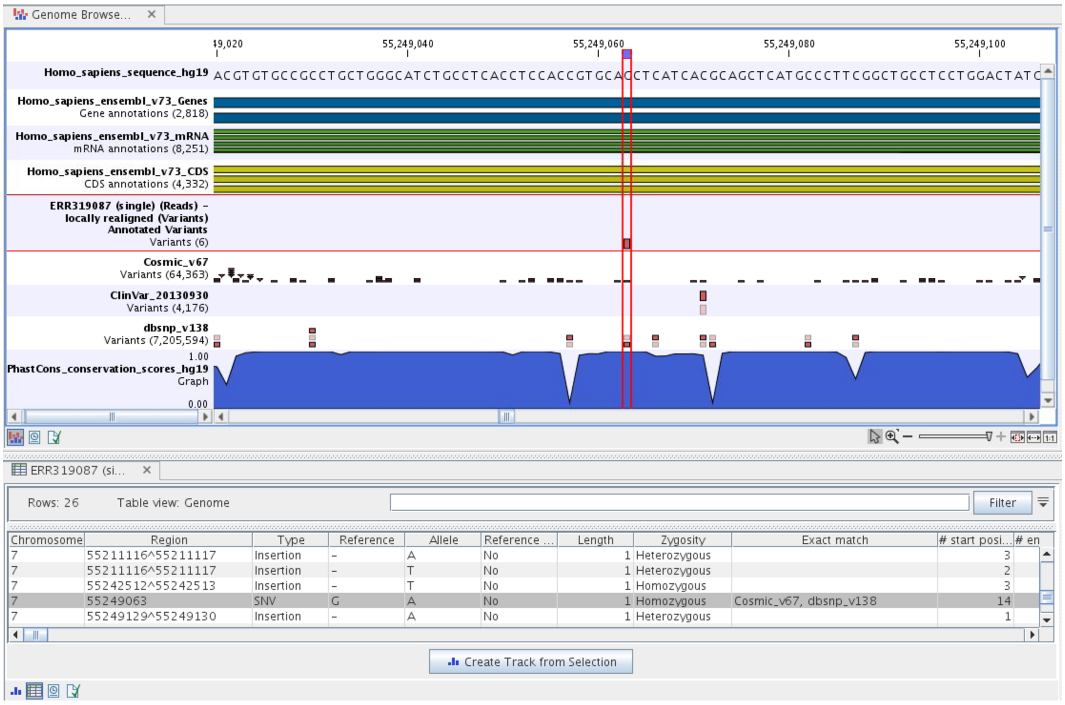
Figure 12.10: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list) containing individual tracks for all added annotations.*

Note! Please be aware, that if you delete the annotated variant track, this track will also disappear from the genome browser view.

It is possible to add tracks to the Genome Browser View such as mapped sequencing reads as well as other tracks. This can be done by dragging the track directly from the **Navigation Area** to the Genome Browser View.

If you double-click on the name of the annotated variant track in the left hand side of the Genome Browser View, a table that includes all variants and the added information/annotations will open (see figure 12.11). The table and the Genome Browser View are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Genome Browser View.

You may be met with a warning as shown in figure 12.12. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts



Rows: 26 Table view: Genome

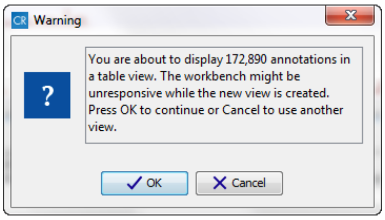
Filter

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# en
7	55211116^55211117	Insertion	-	A	No	1	Heterozygous		3	
7	55211116^55211117	Insertion	-	T	No	1	Heterozygous		2	
7	55242512^55242513	Insertion	-	T	No	1	Homozygous		3	
7	55249063	SNV	G	A	No	1	Homozygous	Cosmic_v67, dbsnp_v138	14	
7	55249129^55249130	Insertion	-	A	No	1	Heterozygous		1	

Create Track from Selection

Figure 12.11: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list). The information is also available in table view. Click on the small table icon to open the table view. If you hold down the "Ctrl" key while clicking on the table icon, you will open a split view showing both the genome browser view and the table view.*

of annotations. Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.



You are about to display 172,890 annotations in a table view. The workbench might be unresponsive while the new view is created. Press OK to continue or Cancel to use another view.

Figure 12.12: *Warning that appears when you work with tracks containing many annotations.*

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals in the region containing the variant can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) are prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:

> **Toolbox** | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (`http://clccancer.com/software/#downloads`, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

12.4 Filter Somatic Variants (WGS)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same patient, you can use the "Filter Somatic Variants (WGS)" ready-to-use workflow to identify potential somatic variants. The purpose of this ready-to-use workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same patient is available.

The "Filter Somatic Variants (WGS)" ready-to-use workflow accepts variant tracks ( ) (e.g. the output from the Identify Variants ready-to-use workflow) as input. Variants that are identical to the human reference sequence are first filtered away and then variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

Please note that this tool will likely also remove inherited cancer variants that are present at a low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes, conservation scores and information from COSMIC (database with known variants in cancer), ClinVar (known variants with medical impact) and dbSNP (all known variants).

To run the **Filter Somatic Variants** tool, go to:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Genome Sequencing** ( ) | **Filter Somatic Variants** ( )

1. Double-click on the **Filter Somatic Variants** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the variant track you would like to use for filtering somatic variants. The panel in the left side of the wizard shows the kind of input that should be provided (figure 12.13). Select by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard.

   Click on the button labeled **Next**.

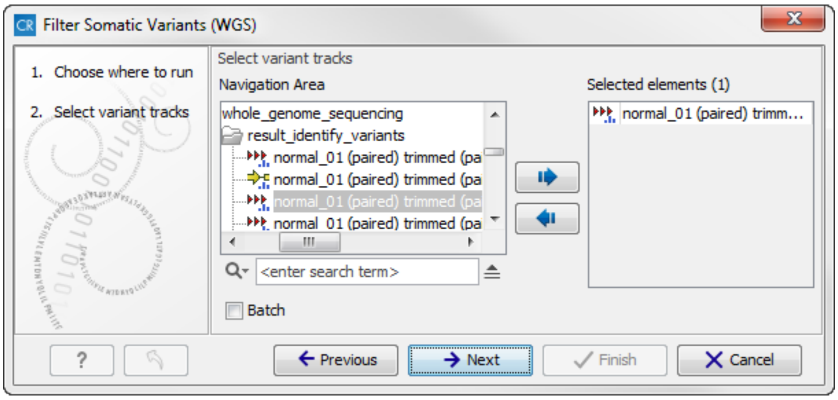
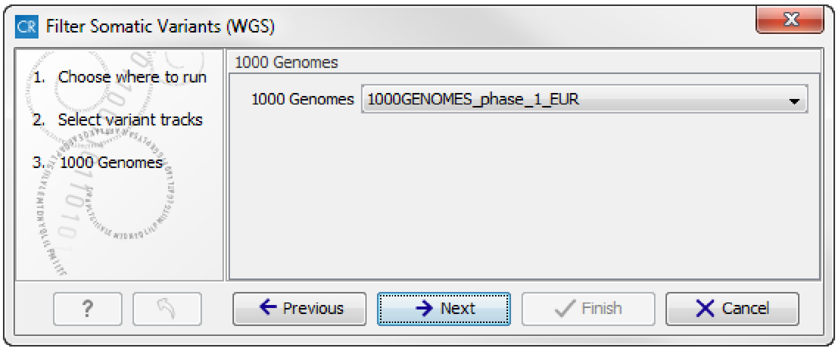


Figure 12.13: *Select the variant track from which you would like to filter somatic variants.*

2. In the next step you will be asked to specify which of the 1000 Genomes populations that should be used for annotation (figure 12.14).

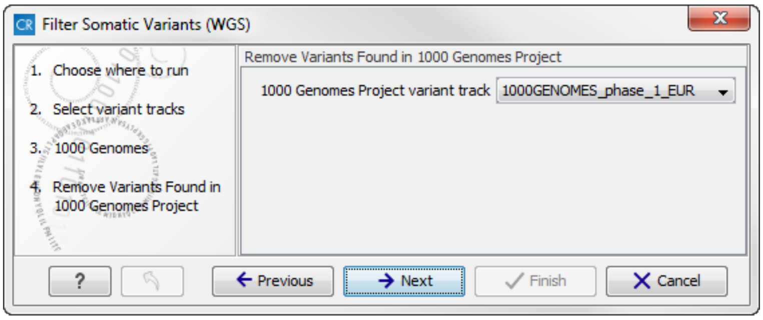


1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 12.14: *Specify which 1000 Genomes population to use for annotation.*

Click on the button labeled **Next**.

3. The next wizard step will once again allow you to specify the 1000 Genomes population that should be used, this time for filtering out variants found in the 1000 Genomes project (figure 12.15).



1000 Genomes Project variant track: 1000GENOMES_phase_1_EUR

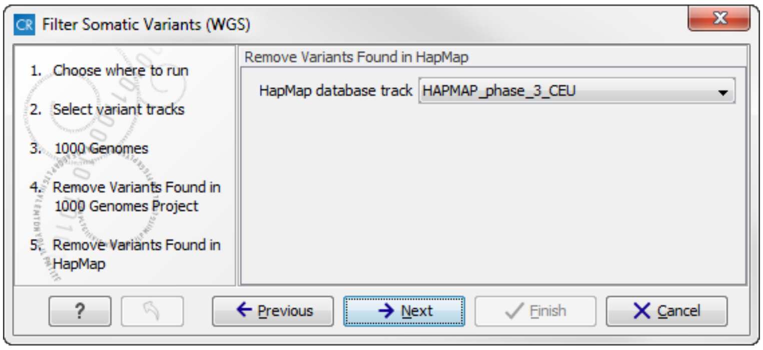
Filter Somatic Variants (WGS)
1. Choose where to run
2. Select variant tracks
3. 1000 Genomes
4. Remove Variants Found in 1000 Genomes Project
Remove Variants Found in 1000 Genomes Project
?
Previous
Next
Finish
Cancel

Figure 12.15: *Specify which 1000 Genomes population to use for filtering out known variants.*

Click on the button labeled **Next**.

4. The next wizard step (figure 12.16) concerns removal of variants found in the HapMap database. Select the population you would like to use from the drop-down list. Please

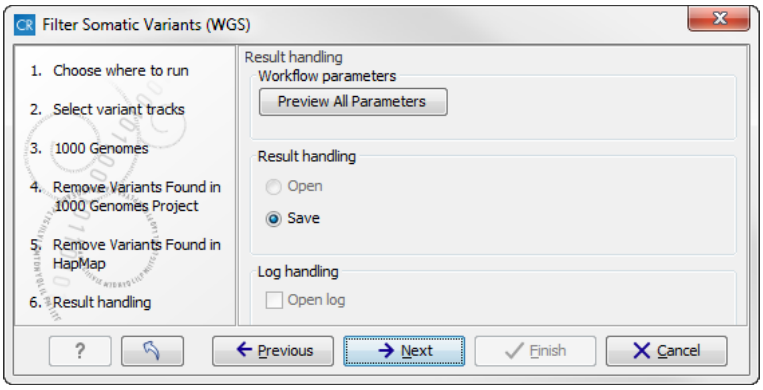
note that the populations available from the drop-down list can be specified with the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).


Filter Somatic Variants (WGS)
1. Choose where to run
2. Select variant tracks
3. 1000 Genomes
4. Remove Variants Found in 1000 Genomes Project
5. Remove Variants Found in HapMap
Remove Variants Found in HapMap
?
Previous
Next
Finish
Cancel

HapMap database track HAPMAP_phase_3_CEU

Figure 12.16: *Specify which HapMap population to use for filtering out known variants.*

5. Click on the button labeled **Next** to go to the last wizard step (shown in figure 12.17).



Result handling
○ Open
● Save

Log handling
□ Open log

Figure 12.17: *Check the selected parametes by pressing "Preview All Parameters".*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes. Choose to save the results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Genome Browser View. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.
2. **Genome Browser View Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, COSMIC, 1000 Genomes, and the PhastCons conservation scores (see figure 12.18).

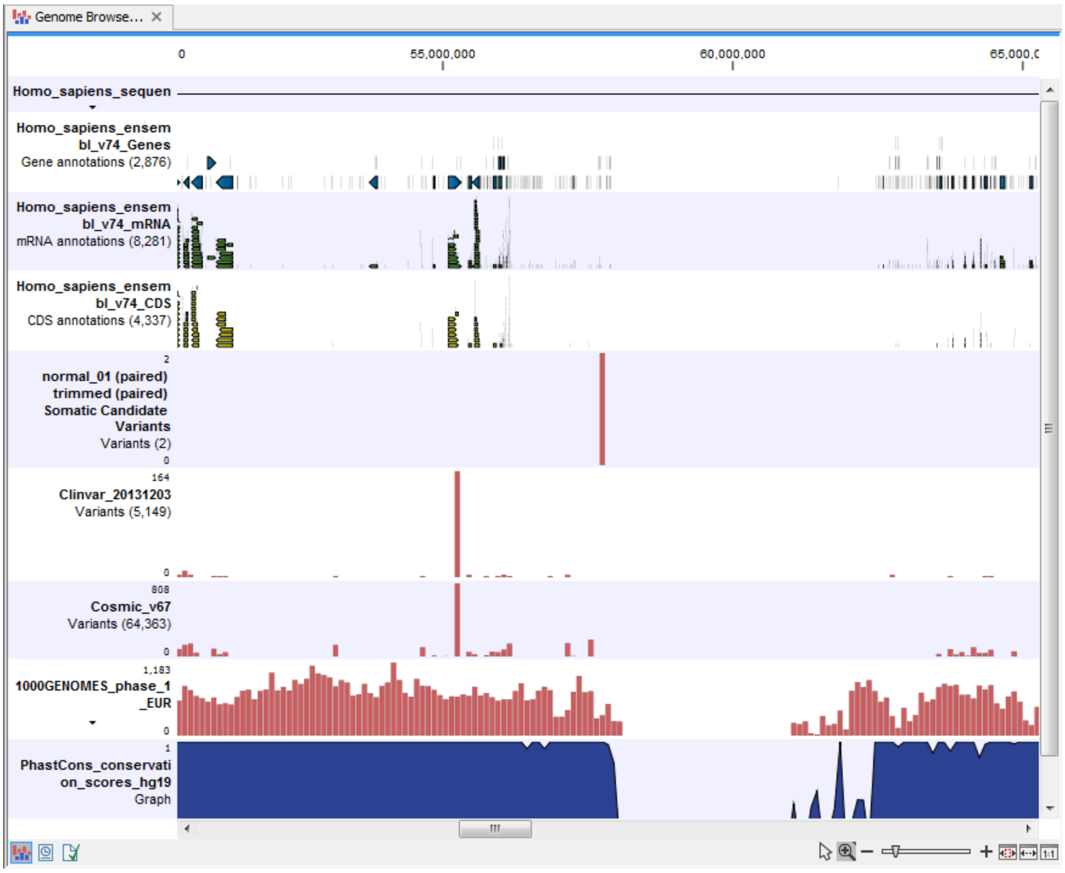


Figure 12.18: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

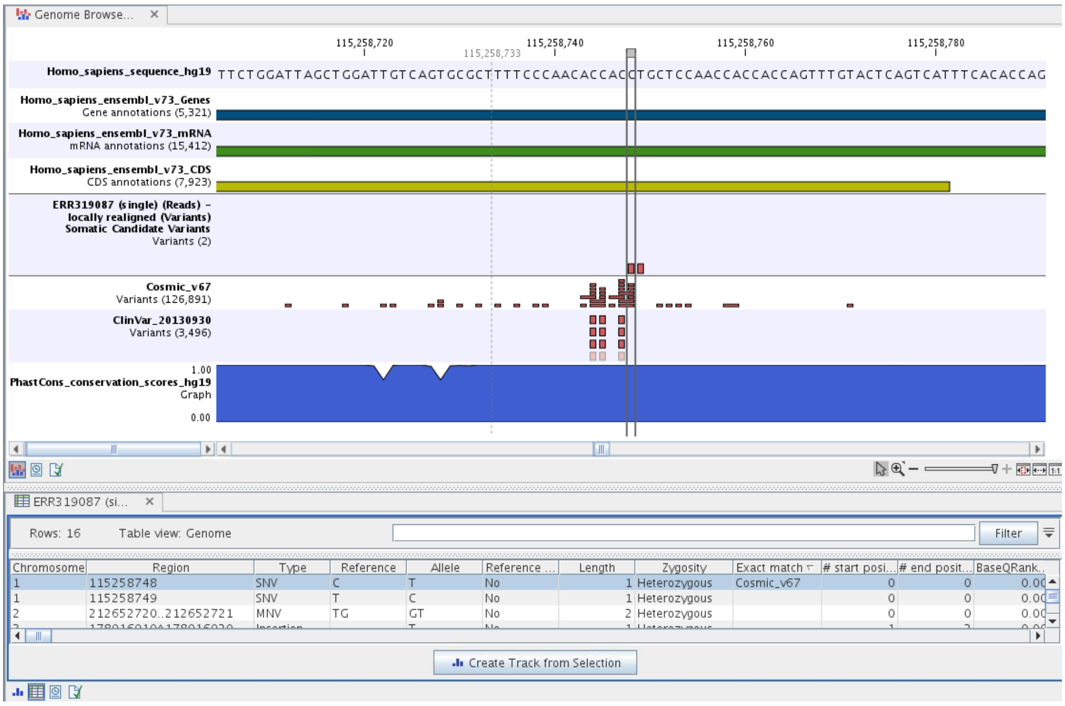
The track with the conservation scores allows you to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant. Mapped sequencing reads as well as other tracks can be easily added to the Genome Browser View.

If you click on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations. This is shown in figure 12.19.

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level, between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments


Genome Browse...
115,258,720
115,258,733
115,258,740
115,258,760
115,258,780
Homo_sapiens_sequence_hg19
TTCTGGATTAGCTGGATTGTCAGTGCGCTTTTCCCAACACCACCTGCTCCAACCACCACCAGTTTGTACTCAGTCATTTCACACCAG
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (15,412)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
ERR319087 (single) (Reads) - locally realigned (Variants) Somatic Candidate Variants
Variants (2)
Cosmic_v67
Variants (126,891)
ClinVar_20130930
Variants (3,496)
PhastCons_conservation_scores_hg19
Graph
1.00
0.00
ERR319087 (si...
Filter
Create Track from Selection

Rows: 16 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# end posit...	BaseQRank..
1	115258748	SNV	C	T	No	1	Heterozygous	Cosmic_v67	0	0	0.00
1	115258749	SNV	T	C	No	1	Heterozygous		0	0	0.00
2	212652720..212652721	MNV	TG	GT	No	2	Heterozygous		0	0	0.00

Figure 12.19: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (http://clccancer.com/software/#downloads, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

12.5 Identify Somatic Variants from Tumor Normal Pair (WGS)

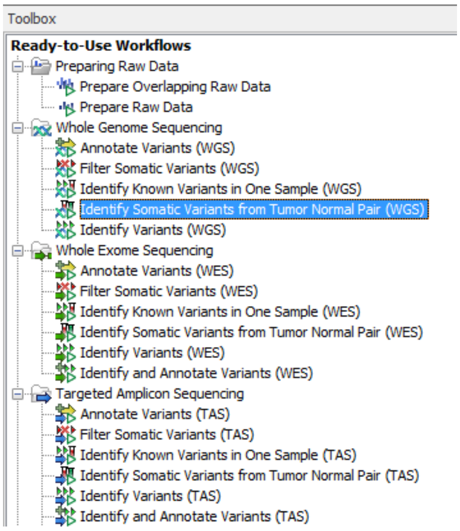
The "Identify Somatic Variants from Tumor Normal Pair (WGS)" ready-to-use workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same patient.

When running the "Identify Somatic Variants from Tumor Normal Pair (WGS)" the reads are mapped and the variants identified. An internal workflow removes germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads.

Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from clinically relevant databases like COSMIC (known cancer associated variants) and ClinVar (variants with clinically relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

How to run the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow (figure 12.20).



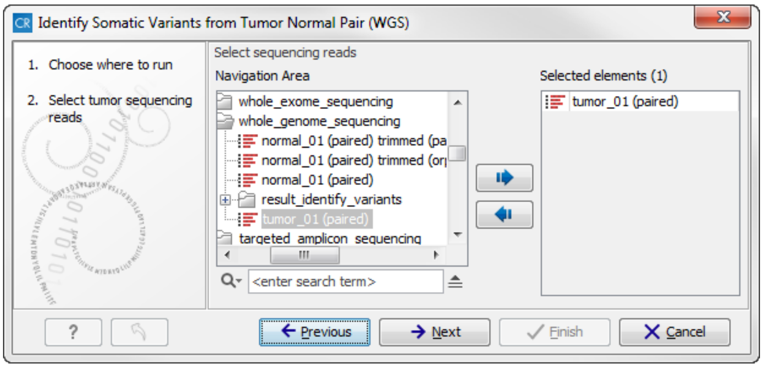
Toolbox

Ready-to-Use Workflows

- Preparing Raw Data
  - Prepare Overlapping Raw Data
  - Prepare Raw Data
- Whole Genome Sequencing
  - Annotate Variants (WGS)
  - Filter Somatic Variants (WGS)
  - Identify Known Variants in One Sample (WGS)
  - Identify Somatic Variants from Tumor Normal Pair (WGS)
  - Identify Variants (WGS)
- Whole Exome Sequencing
  - Annotate Variants (WES)
  - Filter Somatic Variants (WES)
  - Identify Known Variants in One Sample (WES)
  - Identify Somatic Variants from Tumor Normal Pair (WES)
  - Identify Variants (WES)
  - Identify and Annotate Variants (WES)
- Targeted Amplicon Sequencing
  - Annotate Variants (TAS)
  - Filter Somatic Variants (TAS)
  - Identify Known Variants in One Sample (TAS)
  - Identify Somatic Variants from Tumor Normal Pair (TAS)
  - Identify Variants (TAS)
  - Identify and Annotate Variants (TAS)

Figure 12.20: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard shown in figure 12.21 where you can select the tumor sample reads.



Identify Somatic Variants from Tumor Normal Pair (WGS)

1. Choose where to run
2. Select tumor sequencing reads

Select sequencing reads

Navigation Area

- whole_exome_sequencing
- whole_genome_sequencing
  - normal_01 (paired) trimmed (pa
  - normal_01 (paired) trimmed (or
  - normal_01 (paired)
  - result_identify_variants
  - tumor_01 (paired)
- targeted_amplicon_sequencing

<enter search term>

Selected elements (1)

- tumor_01 (paired)

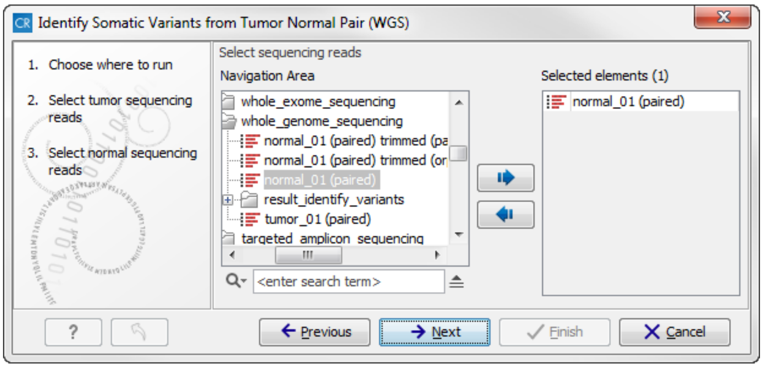
Figure 12.21: *Select the tumor sample reads.*

When you have selected the tumor sample reads click on the button labeled **Next**.

2. In the next wizard step (figure 12.22), please specify the normal sample reads.

3. Click on the button labeled **Next**, which will take you to the next wizard step (figure 12.23).

   In this wizard step you can adjust the settings used for variant detection. For a description of the different parameters that can be adjusted in the variant detection



Identify Somatic Variants from Tumor Normal Pair (WGS)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads

Select sequencing reads

Navigation Area

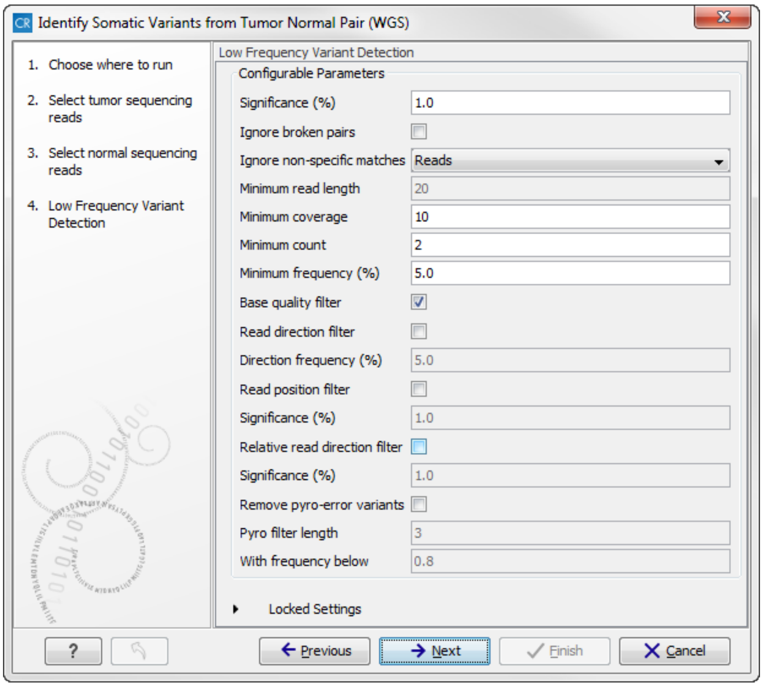
- whole_exome_sequencing
- whole_genome_sequencing
  - normal_01 (paired) trimmed (pa
  - normal_01 (paired) trimmed (or
  - normal_01 (paired)
  - result_identify_variants
  - tumor_01 (paired)
- targeted_amplicon_sequencing

<enter search term>

Selected elements (1)

- normal_01 (paired)

Figure 12.22: *Select the normal sample reads.*



Identify Somatic Variants from Tumor Normal Pair (WGS)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection

Low Frequency Variant Detection

Configurable Parameters

Significance (%)	1.0
Ignore broken pairs	□
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	2
Minimum frequency (%)	5.0
Base quality filter	☑
Read direction filter	□
Direction frequency (%)	5.0
Read position filter	□
Significance (%)	1.0
Relative read direction filter	□
Significance (%)	1.0
Remove pyro-error variants	□
Pyro filter length	3
With frequency below	0.8

Locked Settings

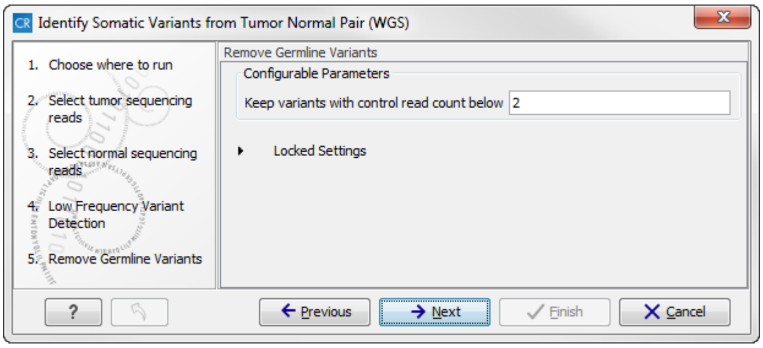
Figure 12.23: *Specify the settings for the variant detection.*

step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html). As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the description of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Filters.html). If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the ready-to-use workflow.

4. Click on the button labeled **Next** to go to the step where you can adjust the settings for removal of germline variants (figure 12.24)..

5. Click on the button labeled **Next**.

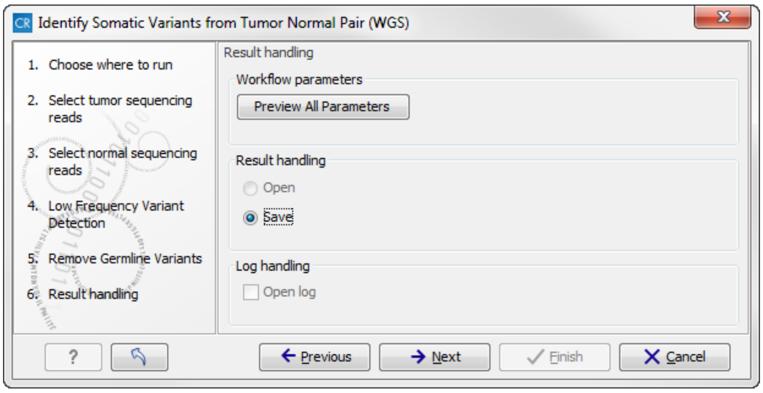
   In the next wizard step you can check the selected settings by clicking on the button labeled


Identify Somatic Variants from Tumor Normal Pair (WGS)
1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. Remove Germline Variants
Remove Germline Variants
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Keep variants with control read count below: 2

Figure 12.24: *Specify setting for removal of germline variants.*

Preview All Parameters (figure 12.25).


Identify Somatic Variants from Tumor Normal Pair (WGS)
1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. Remove Germline Variants
6. Result handling
Result handling
Workflow parameters
Preview All Parameters
Result handling
Log handling
?
Previous
Next
Finish
Cancel

○ Open
● Save
□ Open log

Figure 12.25: *Check the parameters and save the results.*

In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

6. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Six different outputs are generated:

1. **Read Mapping Tumor** (☰) The mapped sequencing reads for the tumor sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **Read Mapping Normal** (☰) The mapped sequencing reads for the normal sample. The reads are shown in different colors depending on their orientation, whether they are single

reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

3. **Mapping Report Tumor** (📊) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.

4. **Mapping Report Normal** (📊) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the normal sample.

5. **Annotated Somatic Variants** (📊) A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

6. **Genome Browser View Tumor Normal Comparison** (📊) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar and COSMIC databases, and finally a track showing the conservation score (see figure 12.26).

12.6 Identify Known Variants in One Sample (WGS)

The "Identify Known Variants in One Sample" ready-to-use workflow is a combined data analysis and interpretation ready-to-use workflow.

It should be used to identify known variants, specified by the user (e.g. known breast cancer associated variants), for their presence or absence in a sample.

Please note that the ready-to-use workflow will not identify new variants.

The Identify Known Variants in One Sample ready-to-use workflow runs an internal workflow that maps the sequencing reads to the human genome sequence and does a local realignment of the mapped reads to improve the following variant detection. Next, specified variants by the user are identified in the read mapping. At the end, information present on the known variants before, are added to the results.

12.6.1 Import your known variants

To make an import into the Cancer Research Workbench, you should have your variants in GVF or VCF 4.1 format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

12.6.2 Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit will be provided by the vendor. To obtain this file you will have to get in contact with the vendor and ask them to

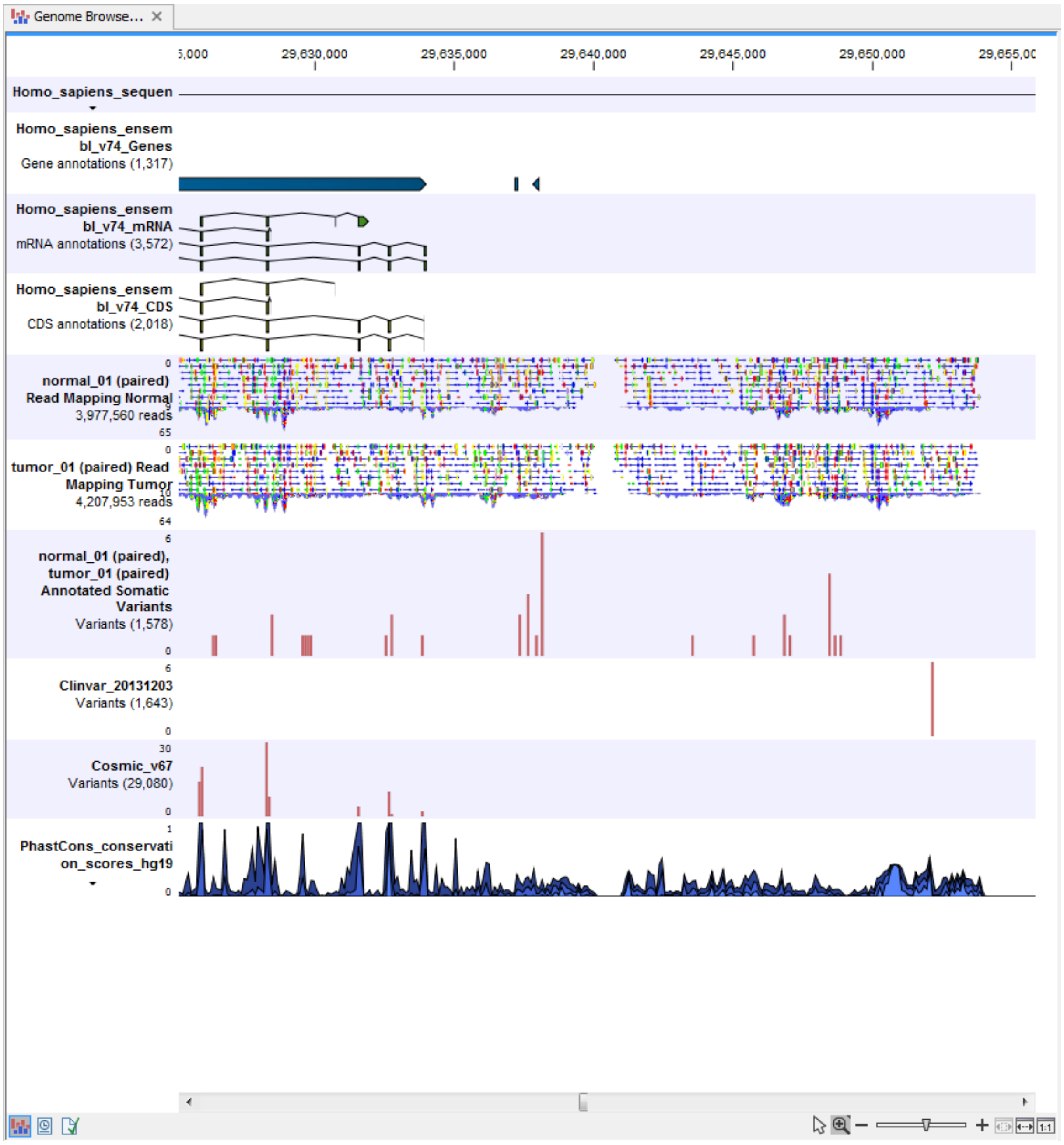


Figure 12.26: *The Genome Browser View presents all the different data tracks together and makes it easy to compare different tracks.*

send this target regions file to you. You will get it in either .bad or .gff format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

12.6.3 How to run the "Identify Known Variants in One Sample" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Known Variants from One Sample" ready-to-use workflow (figure 12.27).

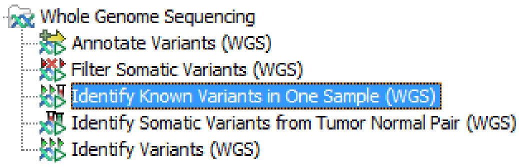

Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Identify Known Variants in One Sample (WGS)
Identify Somatic Variants from Tumor Normal Pair (WGS)
Identify Variants (WGS)

Figure 12.27: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard step shown in figure 12.28 where you can select the reads of the sample, which should be tested for presence or absence of your known variants.

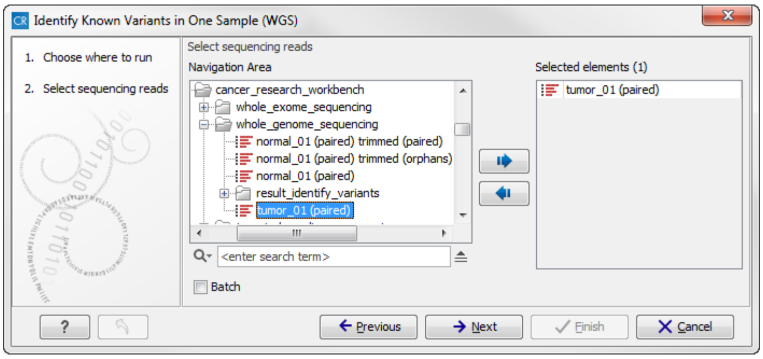


Figure 12.28: *Select the sequencing reads from the sample you would like to test for your known variants.*

Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 12.28) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

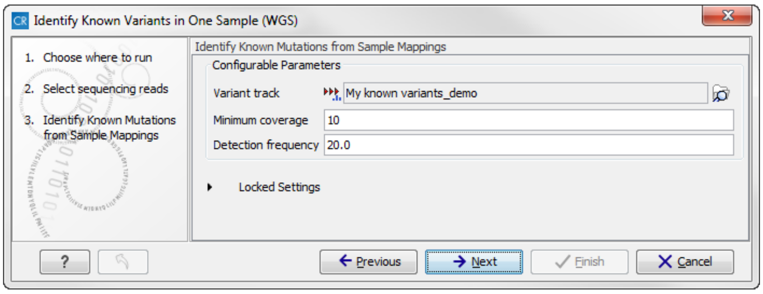
When you have selected the sample(s) you wish to analyze, click on the button labeled **Next** and specify the track with the known variants that should be identified in your sample (figure 12.29). Furthermore, in this wizard step you can specify the minimum read coverage for the position of the variant that should be identified. If the coverage at the position of the variant is below this, the result will show this.

The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or not (observed frequency > specified frequency). Moreover, it will determine if a variant should be labeled as heterozygote (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygote (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).

2. Click on the button labeled **Next**, which will take you to the next wizard step (figure 12.30). In this and the next dialog, you will be asked about which of the annotations/informations added to variants should be included in the results.

   Please specify your track with known variants.

3. Click on the button labeled **Next** and once again select the same track with known variants (figure 12.31).



1. Choose where to run
2. Select sequencing reads
3. Identify Known Mutations from Sample Mappings

Identify Known Mutations from Sample Mappings

Configurable Parameters

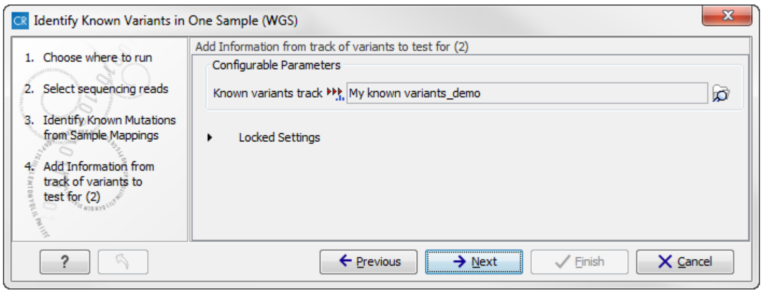
Variant track: My known variants_demo

Minimum coverage: 10

Detection frequency: 20.0

Locked Settings

Figure 12.29: *Specify the track with the known variants that should be identified.*



1. Choose where to run
2. Select sequencing reads
3. Identify Known Mutations from Sample Mappings
4. Add Information from track of variants to test for (2)

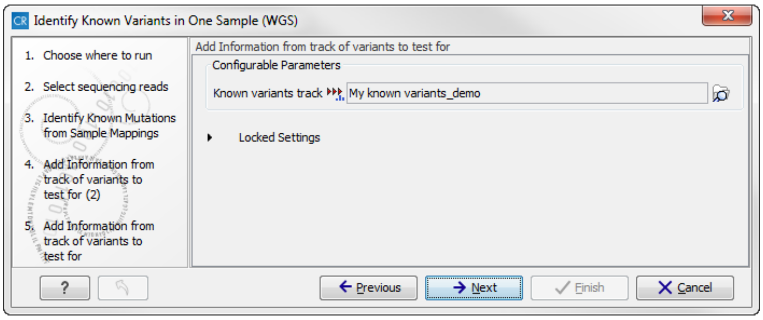
Add Information from track of variants to test for (2)

Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

Figure 12.30: *Please select the track with your known variants again. Annotations/Informations from this track will be added to the overview mutation track.*



1. Choose where to run
2. Select sequencing reads
3. Identify Known Mutations from Sample Mappings
4. Add Information from track of variants to test for (2)
5. Add Information from track of variants to test for

Add Information from track of variants to test for

Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

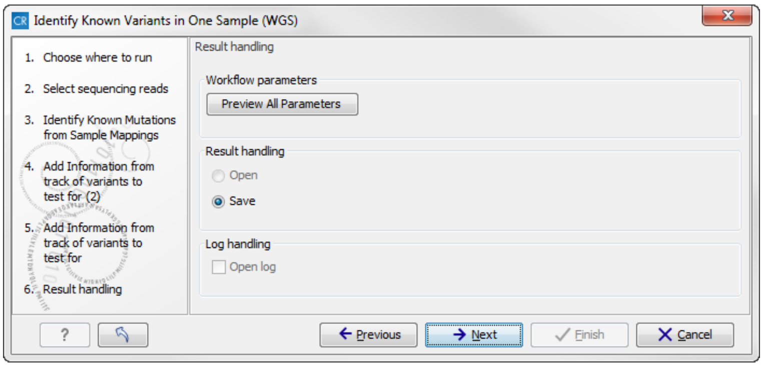
Figure 12.31: *Once again select the track with known variants. This time the track is used to add information to the detailed mutation track.*

4. Click on the button labeled **Next** to go to the last wizard step (figure 12.32).

   In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

5. Click on the button labeled **OK** to go back to the previous dialog box and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.


Identify Known Variants in One Sample (WGS)

1. Choose where to run
2. Select sequencing reads
3. Identify Known Mutations from Sample Mappings
4. Add Information from track of variants to test for (2)
5. Add Information from track of variants to test for
6. Result handling

Result handling

Workflow parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 12.32: *Check the settings and save your results.*

12.6.4 Output from the Identify Known Variants in One Sample

The "Identify Known Variants in One Sample" tool produces six different output types.

1. **Read Mapping Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads.

2. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

3. **Overview Variants Detected** ( ) Annotation track showing the known variants. The table view provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads.

4. **Variants Detected in Detail** ( ) Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. The difference between the two tables is that the "Variants Detected in Detail" table includes detailed information about the most frequent alternative allele (MFAA).

5. **Genome Browser View Identify Known Variants** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

6. **Log** ( ) A log of the workflow execution.

It is a good idea to start looking at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, we also recommend that you check that the majority of the reads are mapping to the targeted region.

When this has been done you can open the Genome Browser View file (see 12.33).

The Genome Browser View includes the overview track of known variants and the detailed result track in the context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, and clinically relevant variants in the COSMIC databases.

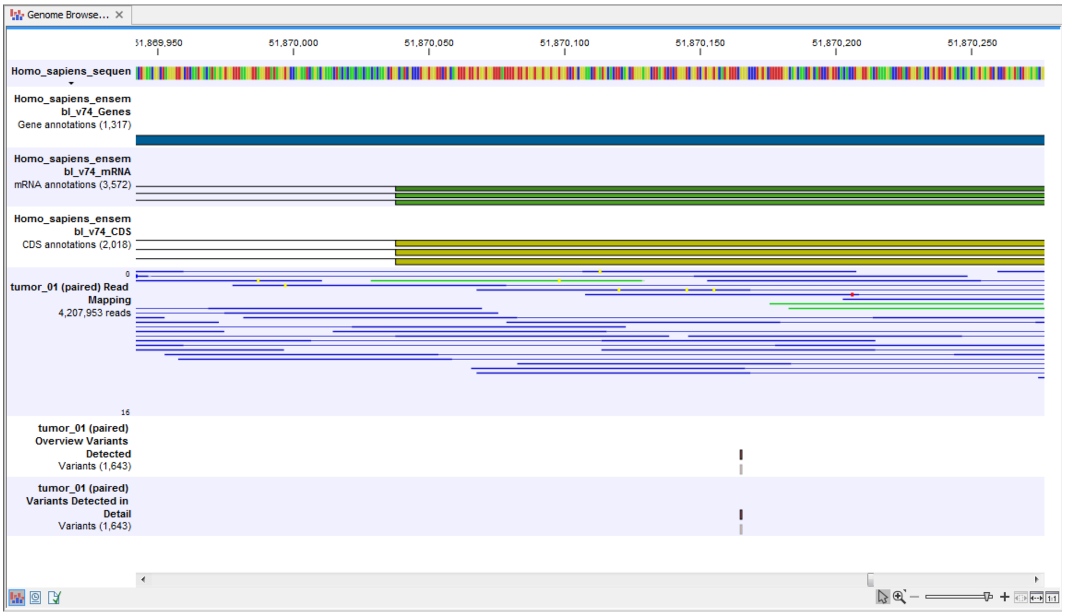


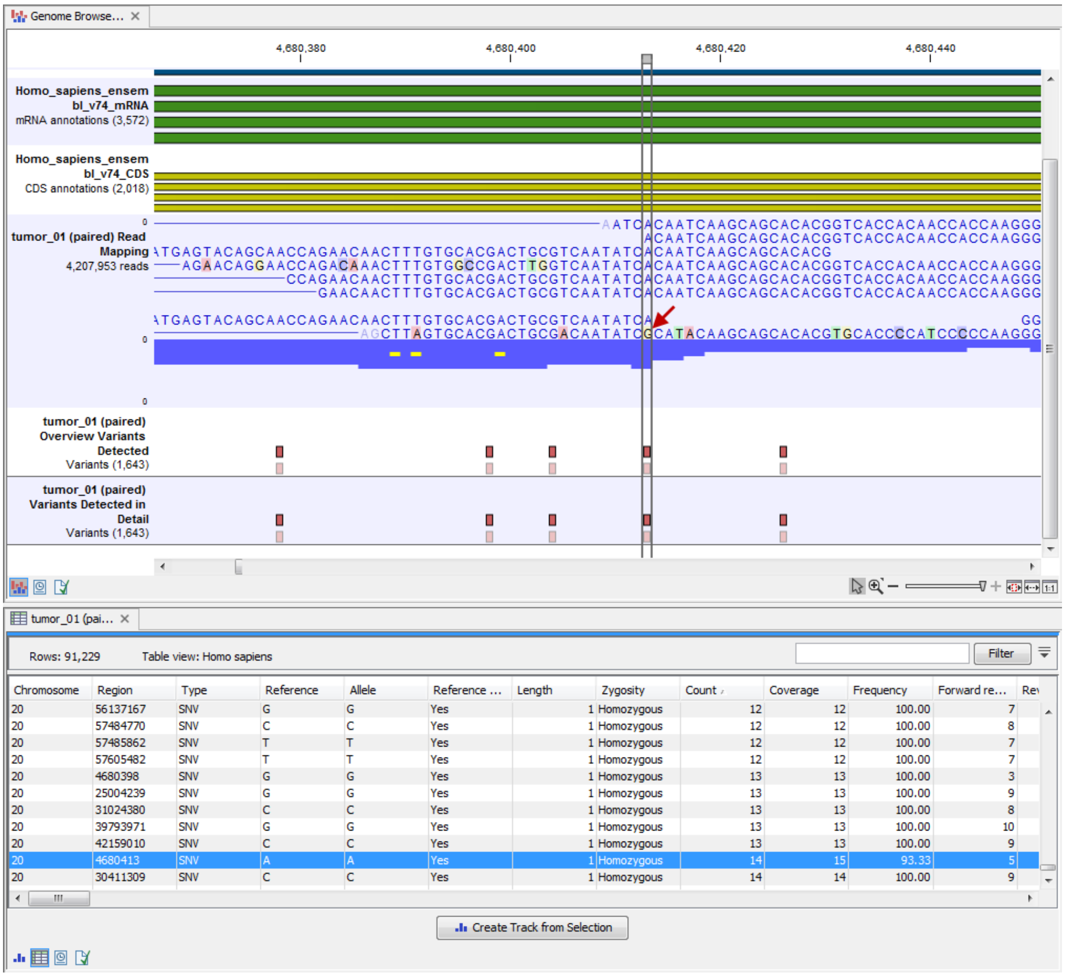
Figure 12.33: *Genome Browser View that allows inspection of the identified variants in the context of the human genome and external databases.*

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

The difference between the overview variant track and the detailed variant track is the annotations added to the variants.

By double clicking on one of the annotated variant tracks in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see 12.34).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.


Genome Browse...
4,680,380
4,680,400
4,680,420
4,680,440
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (3,572)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (2,018)
tumor_01 (paired) Read Mapping
4,207,953 reads
tumor_01 (paired) Overview Variants Detected
Variants (1,643)
tumor_01 (paired) Variants Detected in Detail
Variants (1,643)
tumor_01 (pai...
Filter
Create Track from Selection

Rows: 91,229 Table view: Homo sapiens

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Forward re...	Rev
20	56137167	SNV	G	G	Yes	1	Homozygous	12	12	100.00	7	
20	57484770	SNV	C	C	Yes	1	Homozygous	12	12	100.00	8	
20	57485862	SNV	T	T	Yes	1	Homozygous	12	12	100.00	7	
20	57605482	SNV	T	T	Yes	1	Homozygous	12	12	100.00	7	
20	4680398	SNV	G	G	Yes	1	Homozygous	13	13	100.00	3	
20	25004239	SNV	G	G	Yes	1	Homozygous	13	13	100.00	9	
20	31024380	SNV	C	C	Yes	1	Homozygous	13	13	100.00	8	
20	39793971	SNV	G	G	Yes	1	Homozygous	13	13	100.00	10	
20	42159010	SNV	C	C	Yes	1	Homozygous	13	13	100.00	9	
20	4680413	SNV	A	A	Yes	1	Homozygous	14	15	93.33	5	
20	30411309	SNV	C	C	Yes	1	Homozygous	14	14	100.00	9	

Figure 12.34: *Genome Browser View with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

Chapter 13

Whole exome sequencing (WES)

Contents

The protein coding part of the human genome accounts for around 1 % of the genome and consists of around 180,000 exons covering an area of 3̃0 megabases (Mb) [Ng et al., 2009]. By targeting sequencing to only the protein coding parts of the genome, exome sequencing is a cost efficient way of generating sequencing data that is believed to harbor the vast majority of the disease-causing mutations [Choi et al., 2009].

13.1 Automatic analysis of sequencing data (WES)

Six ready-to-use workflows are available for analysis of whole genome sequencing data. The concept of the pre-installed ready-to-use workflows is that read data are used as input in one end of the workflow and in the other end of the workflow you get a track based genome browser view and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual ready-to-use workflows can be used for and go through step by step how to run the workflows.

Note! Often you will have to prepare data with one of the two **Preparing Raw Data** workflows described in section 11 before you proceed to **Analysis of sequencing data (WES)**.

13.2 Identify Variants (WES)

The "Identify Variants" tool takes sequencing reads as input and returns identified variants as part of a Genome Browser View.

The tool runs an internal workflow, which starts with mapping the sequencing reads to the human reference sequence. Then it runs a local realignment to improve the variant detection, which is run afterwards. At the end, variants with an average base quality smaller than 20 are filtered away.

In addition, a targeted region report is created to inspect the overall coverage and mapping specificity in the targeted regions.

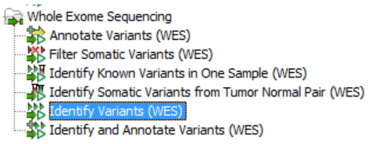
Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit will be provided by the vendor. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get it in either .bed or .gff format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

How to run the "Identify Variants" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Variants" ready-to-use workflow (figure 13.1).



Whole Exome Sequencing
- Annotate Variants (WES)
- Filter Somatic Variants (WES)
- Identify Known Variants in One Sample (WES)
- Identify Somatic Variants from Tumor Normal Pair (WES)
- Identify Variants (WES)
- Identify and Annotate Variants (WES)

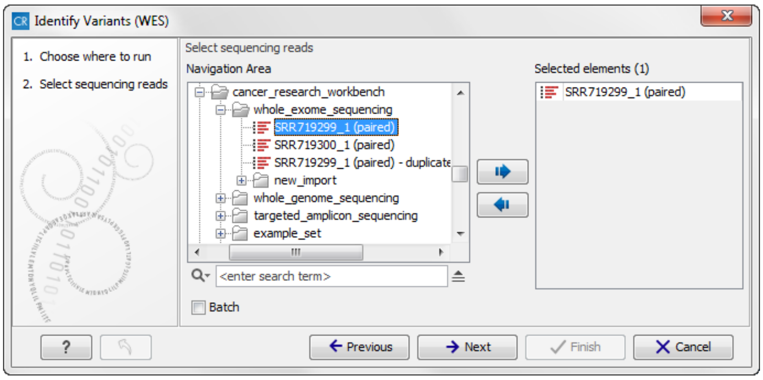
Figure 13.1: *The ready-to-use workflows are found in the toolbox.*

   This will open the wizard shown in figure 13.2 where you can select the sequencing reads from the sample, which should be analyzed.

   Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 13.43) and select the **folder** that holds the data you wish to analyze. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

   When you have selected the sample(s) you wish to prepare, click on the button labeled **Next**.

2. In the next wizard step (figure 13.3) you have to specify the track with the targeted regions from the experiment. You can also specify the minimum read coverage, which should be present in the targeted regions.



Identify Variants (WES)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

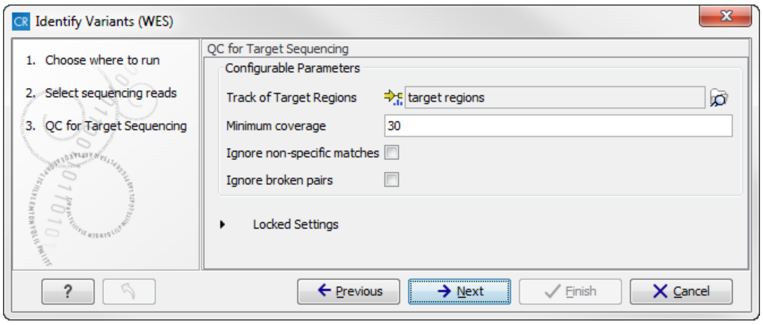
Navigation Area

- cancer_research_workbench
  - whole_exome_sequencing
    - SRR719299_1 (paired)
    - SRR719300_1 (paired)
    - SRR719299_1 (paired) - duplicate
    - new_import
  - whole_genome_sequencing
  - targeted_amplicon_sequencing
  - example_set

Selected elements (1)

- SRR719299_1 (paired)

Figure 13.2: *Please select all sequencing reads from the sample to be analyzed.*



Identify Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing

QC for Target Sequencing

Configurable Parameters

Track of Target Regions: target regions

Minimum coverage: 30

Ignore non-specific matches □

Ignore broken pairs □

Figure 13.3: *Select the track with the targeted regions from your experiment.*

3. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.4). In this wizard you can specify the parameter for detecting variants.

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.5).

5. Click on the button labeled **Next** to go to the last wizard step (figure 13.6).

   In this wizard you get the chance to check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard step you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

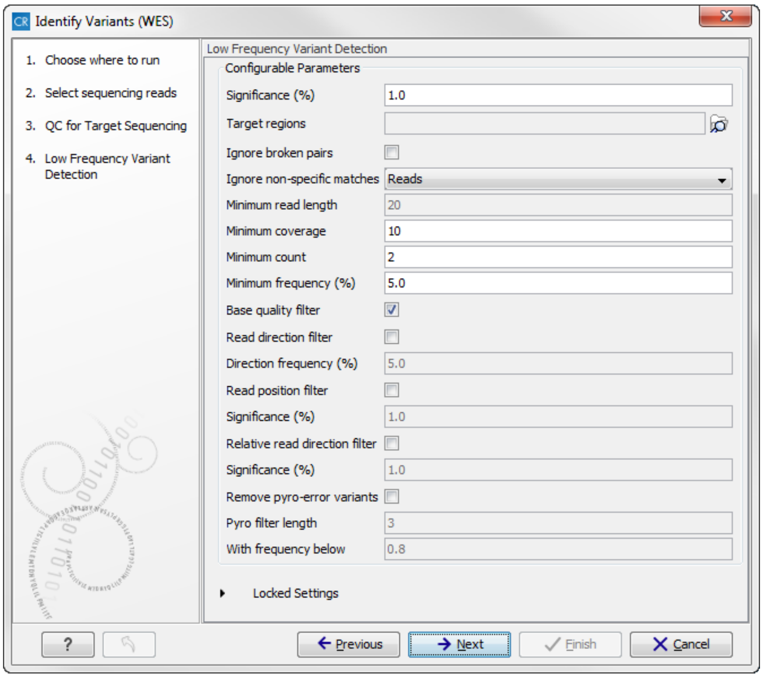
6. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Output from the Identify Variants workflow

The "Identify Variants" tool produces six different types of output:

1. **Read Mapping** (☰) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads,



Identify Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection

Low Frequency Variant Detection

Configurable Parameters

Significance (%): 1.0

Target regions:

Ignore broken pairs: ☐

Ignore non-specific matches: Reads

Minimum read length: 20

Minimum coverage: 10

Minimum count: 2

Minimum frequency (%): 5.0

Base quality filter: ☑

Read direction filter: ☐

Direction frequency (%): 5.0

Read position filter: ☐

Significance (%): 1.0

Relative read direction filter: ☐

Significance (%): 1.0

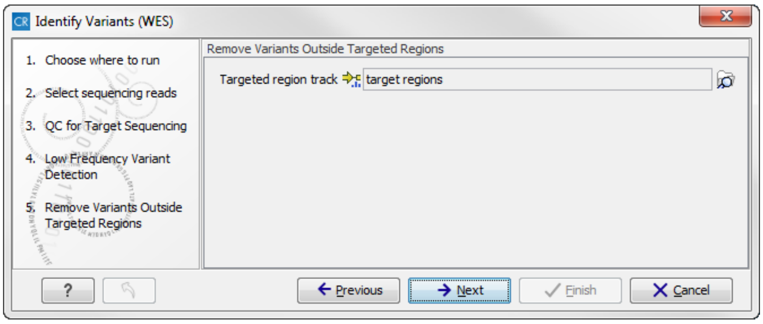
Remove pyro-error variants: ☐

Pyro filter length: 3

With frequency below: 0.8

▸ Locked Settings

Figure 13.4: *Please specify the parameters for variant detection.*



Identify Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection
5. Remove Variants Outside Targeted Regions

Remove Variants Outside Targeted Regions

Targeted region track: target regions

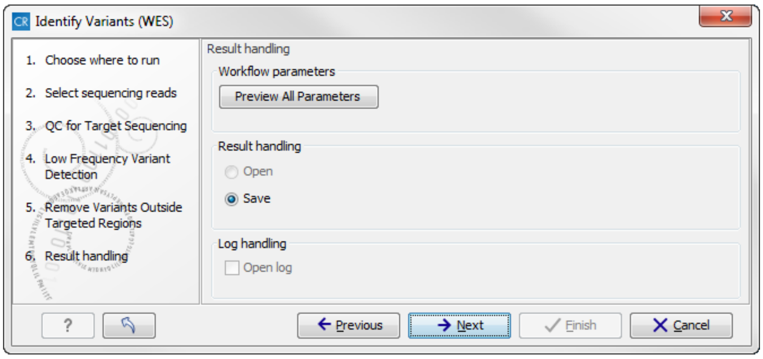
Figure 13.5: *Select the targeted region track. Variants found outside the targeted region will be removed.*

and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **Target Regions Coverage** () The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

3. **Target Regions Coverage Report** () The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

4. **Identified Variants** () A variant track holding the identified variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual



Identify Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection
5. Remove Variants Outside Targeted Regions
6. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 13.6: *Choose to save the results. In this wizard step you get the chance to preview the settings used in the ready-to-use workflow.*

variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

5. **Genome Browser View Identify Variants** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 13.12).

It is important that you do not delete any of the produced files individually as some of the outputs are linked to other outputs. If you would like to delete the outputs, please always delete all of them at the same time.

Please have first a look at the mapping report to see if the coverage is sufficient in regions of interest (e.g. > 30). Furthermore, please check that at least 90% of reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of reads are mapping to the targeted region.

Afterwards please open the Genome Browser View file (see 13.7).

The Genome Browser View includes the track of identified variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions and mapped sequencing reads.

By double clicking on the variant track in the Genome Browser View, a table will be shown which includes information about all identified variants (see 13.8).

In case you like to change the reference sequence used for mapping as well as the human genes, please use the "Data Management".

13.3 Annotate Variants (WES)

Using a variant track () (e.g. the output from the Identify Variants ready-to-use workflow) the **Annotate Variants (WGS)** ready-to-use workflow runs an internal workflow that adds the following annotations to the variant track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.

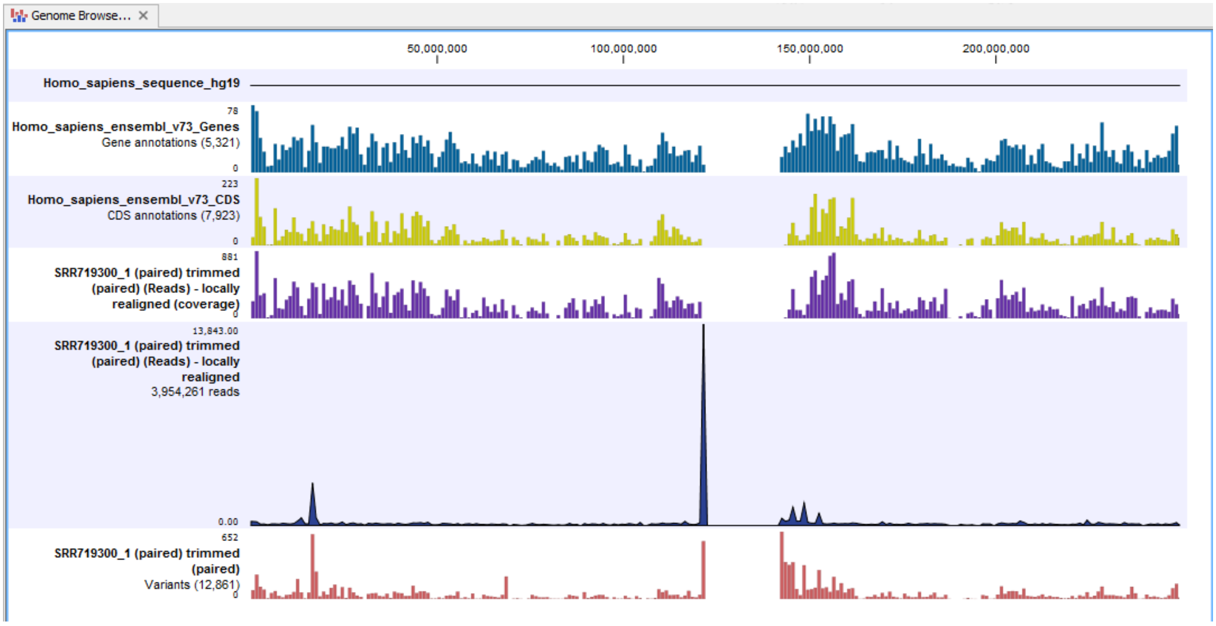
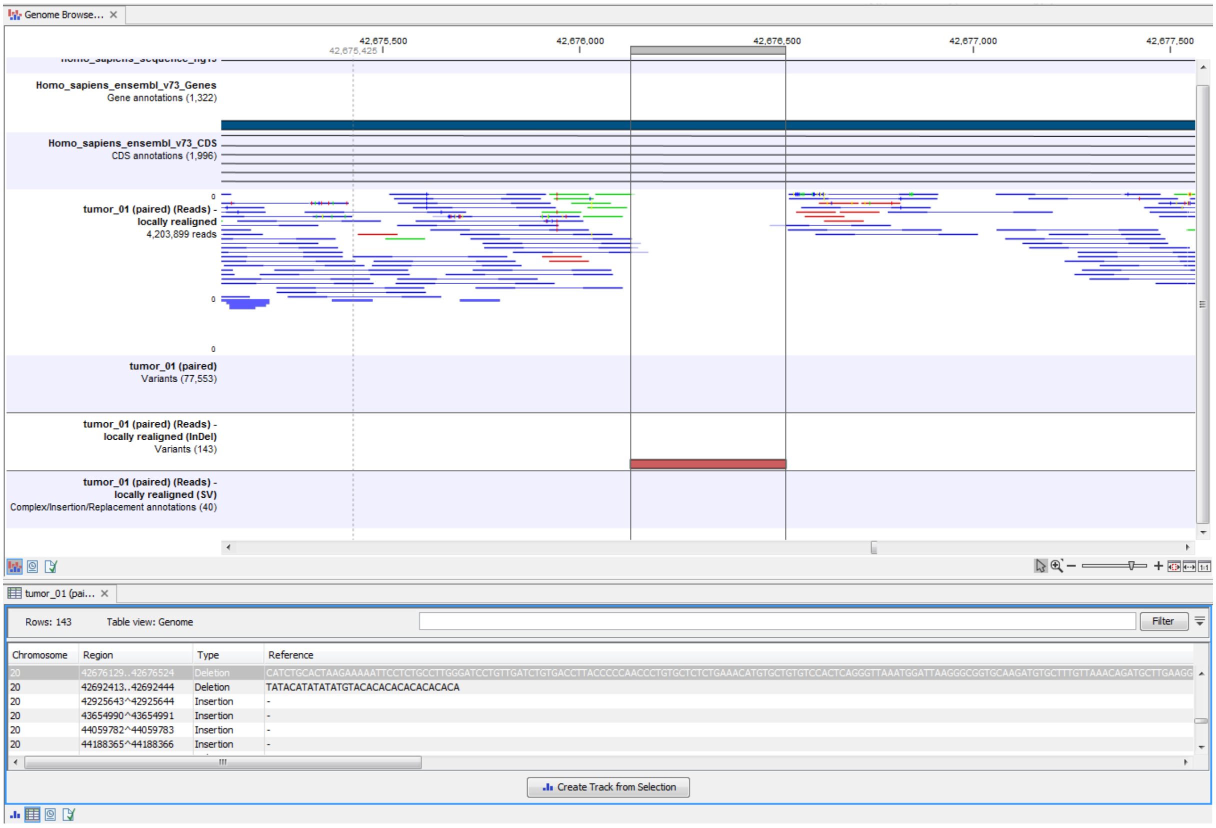

Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
78
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
0
223
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
0
881
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned (coverage)
0
13,843.00
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned
3,954,261 reads
0.00
652
SRR719300_1 (paired) trimmed (paired)
Variants (12,861)
0

Figure 13.7: *The Genome Browser View allows you to inspect the identified variants in the context of the human genome.*

- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from COSMIC**. Adds information from the "Catalogue of Somatic Mutations in Cancer" database.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (InDels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

1. Go to the toolbox and select the **Annotate Variants (WES)** workflow. In the first wizard step, select the input variant track (figure 13.9).
2. Click on the button labeled **Next**. The only parameter that should be specified by the user is which 1000 Genomes population yo use (figure 13.10). This can be done using the drop-down list found in this wizard step. Please note that the populations available from the drop-down list can be specified with the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).
3. Click on the button labeled **Next** to go to the last wizard step (figure 13.11).


Genome Browse...
42,675,500
42,675,425
42,676,000
42,676,500
42,677,000
42,677,500
Homo_sapiens_ensembl_v73_Genes
Gene annotations (1,322)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (1,996)
tumor_01 (paired) (Reads) - locally realigned
4,203,899 reads
tumor_01 (paired)
Variants (77,553)
tumor_01 (paired) (Reads) - locally realigned (InDel)
Variants (143)
tumor_01 (paired) (Reads) - locally realigned (SV)
Complex/Insertion/Replacement annotations (40)
tumor_01 (pai...
Filter
Create Track from Selection

Rows: 143 Table view: Genome

Chromosome	Region	Type	Reference
20	42676129..42676524	Deletion	CATCTGCACTAAGAAAAATTCCTCTGCCTTGGGATCCTGTTGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGAAACATGTGCTGTGTCCACTCAGGGTTAAATGGATTAAGGGCGGTGCAAGATGTGCTTTGTTAAACAGATGCTTGAAGG
20	42692413..42692444	Deletion	TATACATATATATGTACACACACACACACA
20	42925643^42925644	Insertion	-
20	43654990^43654991	Insertion	-
20	44059782^44059783	Insertion	-
20	44188365^44188366	Insertion	-

Figure 13.8: *Genome Browser View with an open track table to inspect identified variants more closely in the context of the human genome.*

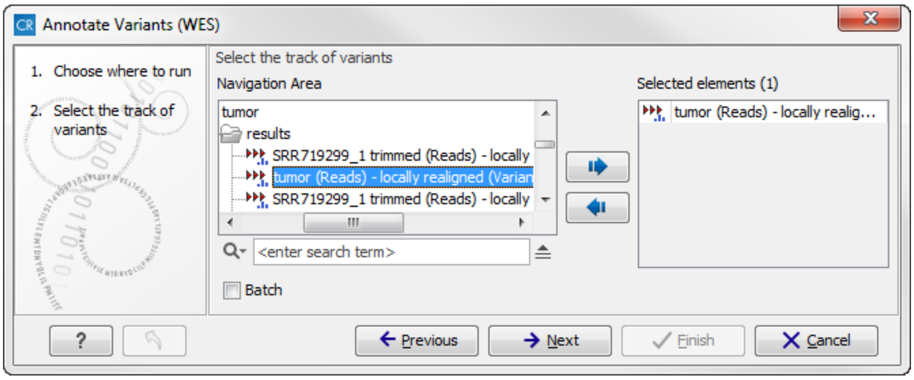

Annotate Variants (WES)
1. Choose where to run
2. Select the track of variants
Select the track of variants
Navigation Area
tumor
results
SRR719299_1 trimmed (Reads) - locally
tumor (Reads) - locally realigned (Varian
SRR719299_1 trimmed (Reads) - locally
<enter search term>
Batch
Selected elements (1)
tumor (Reads) - locally realig...
?
Previous
Next
Finish
Cancel

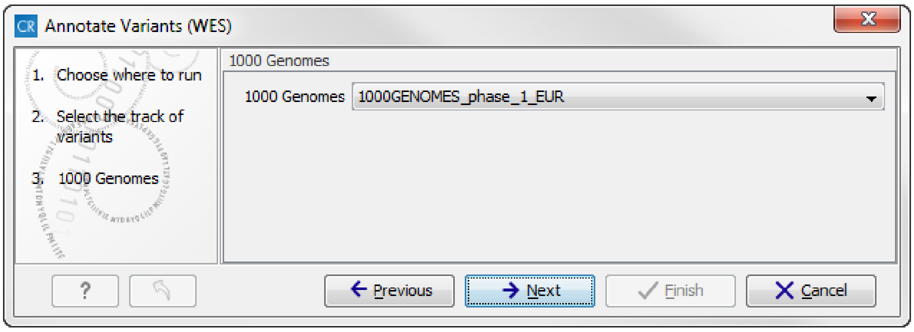
Figure 13.9: *Select the variant track to annotate.*

In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point.

4. Choose to **Save** your results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Annotated Variants** (▸▸▸) Annotation track showing the variants. Hold the mouse over one

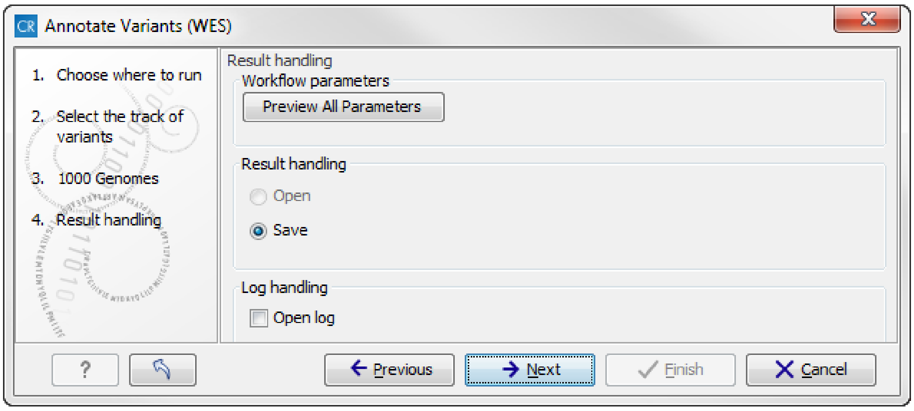

Annotate Variants (WES)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes

1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 13.10: *Select the relevant 1000 Genomes popultaion(s).*


Annotate Variants (WES)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes
4. Result handling

Result handling

Workflow parameters

Result handling
○ Open
● Save

Log handling
□ Open log

Figure 13.11: *Check the settings and save your results.*

   of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. **Genome Browser View Annotated Variants** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP, ClinVar, COSMIC, 1000 Genomes, and PhastCons conservation scores (see figure 13.12).

Note! Please be aware, that if you delete the annotated variant track, this track will also disappear from the genome browser view.

It is possible to add tracks to the **Genome Browser View** such as mapped sequencing reads as well as other tracks. This can be done by dragging the track directly from the **Navigation Area** to the **Genome Browser View**.

If you double-click on the name of the annotated variant track in the left hand side of the Genome Browser View, a table that includes all variants and the added information/annotations will open (see figure 13.13). The table and the Genome Browser View are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Genome Browser View.

You may be met with a warning as shown in figure 13.14. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

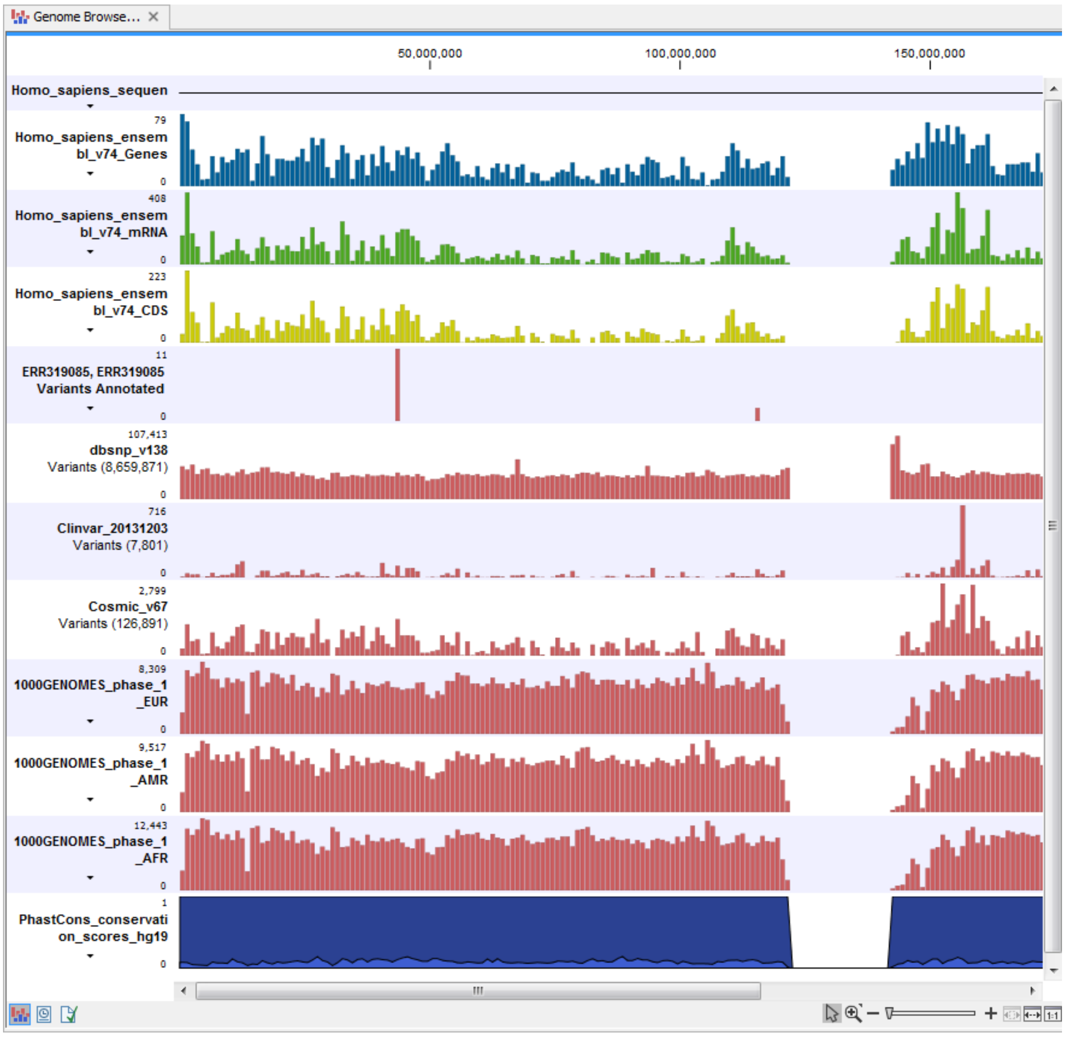
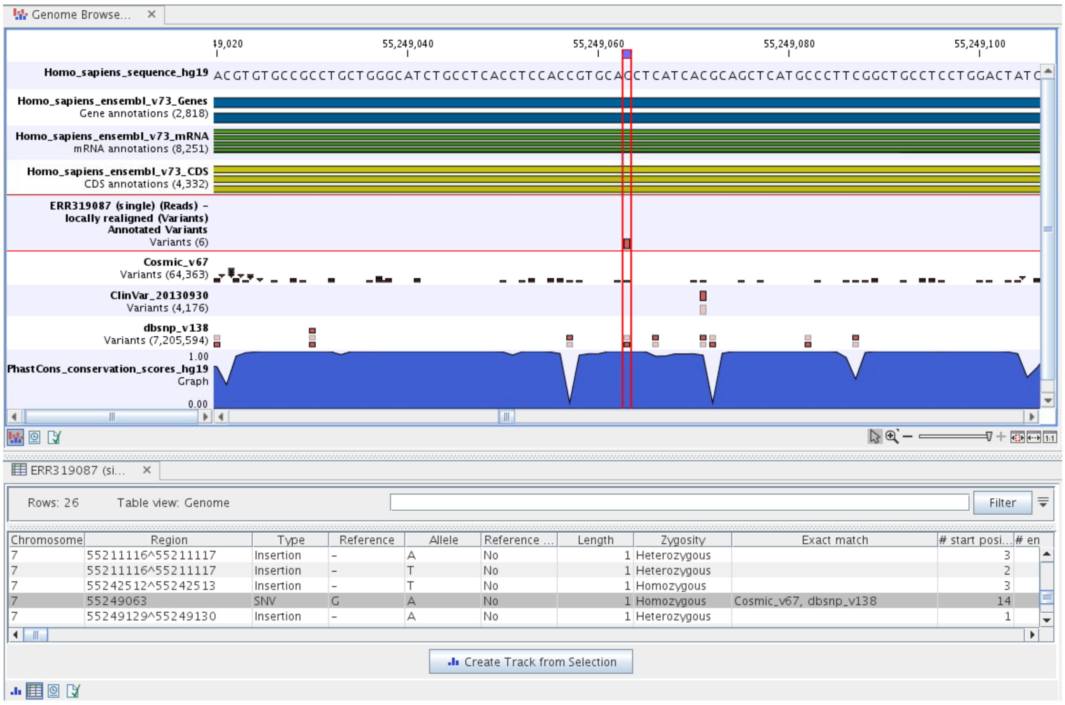

Genome Browse...
50,000,000
100,000,000
150,000,000
Homo_sapiens_sequen
79
Homo_sapiens_ensembl_v74_Genes
0
408
Homo_sapiens_ensembl_v74_mRNA
0
223
Homo_sapiens_ensembl_v74_CDS
0
11
ERR319085, ERR319085 Variants Annotated
0
107,413
dbsnp_v138 Variants (8,659,871)
0
716
Clinvar_20131203 Variants (7,801)
0
2,799
Cosmic_v67 Variants (126,891)
0
8,309
1000GENOMES_phase_1_EUR
0
9,517
1000GENOMES_phase_1_AMR
0
12,443
1000GENOMES_phase_1_AFR
0
1
PhastCons_conservation_scores_hg19
0

Figure 13.12: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list) containing individual tracks for all added annotations.*

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

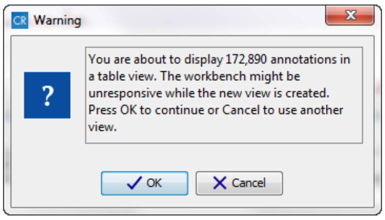
It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:


Genome Browse...
19,020
55,249,040
55,249,060
55,249,080
55,249,100
Homo_sapiens_sequence_hg19
ACGTGTGCCGCCTGCTGGGCATCTGCCTCACCTCCACCGTGCAGCTCATCACGCAGCTCATGCCCTTCGGCTGCCTCCTGGACTATC
Homo_sapiens_ensembl_v73_Genes
Gene annotations (2,818)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (8,251)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (4,332)
ERR319087 (single) (Reads) - locally realigned (Variants) Annotated Variants
Variants (6)
Cosmic_v67
Variants (64,363)
ClinVar_20130930
Variants (4,176)
dbsnp_v138
Variants (7,205,594)
1.00
PhastCons_conservation_scores_hg19
Graph
0.00
ERR319087 (si...
Filter
Create Track from Selection

Rows: 26 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# en
7	55211116^55211117	Insertion	-	A	No	1	Heterozygous		3	
7	55211116^55211117	Insertion	-	T	No	1	Heterozygous		2	
7	55242512^55242513	Insertion	-	T	No	1	Homozygous		3	
7	55249063	SNV	G	A	No	1	Homozygous	Cosmic_v67, dbsnp_v138	14	
7	55249129^55249130	Insertion	-	A	No	1	Heterozygous		1	

Figure 13.13: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list). The information is also available in table view. Click on the small table icon to open the table view. If you hold down the "Ctrl" key while clicking on the table icon, you will open a split view showing both the genome browser view and the table view.*



You are about to display 172,890 annotations in a table view. The workbench might be unresponsive while the new view is created. Press OK to continue or Cancel to use another view.

Figure 13.14: *Warning that appears when you work with tracks containing many annotations.*

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (`http://clccancer.com/software/#downloads`, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

13.4 Filter Somatic Variants (WES)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same patient, you can use the "Filter Somatic Variants (WES)" ready-to-use workflow to identify potential somatic variants. The purpose of this ready-to-use workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same patient is available.

The "Filter Somatic Variants (WES)" ready-to-use workflow accepts variant tracks ( ) (e.g. the output from the Identify Variants ready-to-use workflow) as input. In cases with heterozygous variants, the reference allele is first filtered away, then variants outside the targeted region are removed, and lastly, variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

Please note that this tool will likely also remove inherited cancer variants that are present at a low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes, conservation scores and information from COSMIC (database with known variants in cancer), ClinVar (known variants with medical impact) and dbSNP (all known variants).

To run the **Filter Somatic Variants** tool, go to:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Exome Sequencing** ( ) | **Filter Somatic Variants** ( )

1. Double-click on the **Filter Somatic Variants** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the variant track you would like to use for filtering somatic variants. The panel in the left side of the wizard shows the kind of input that should be provided (figure 13.15). Select by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard.

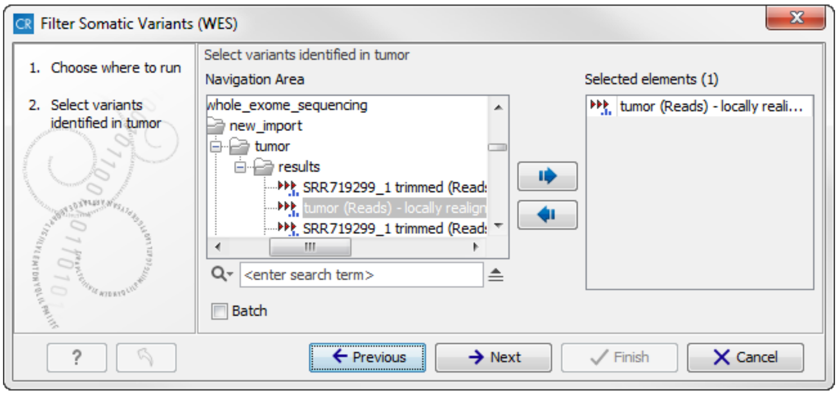
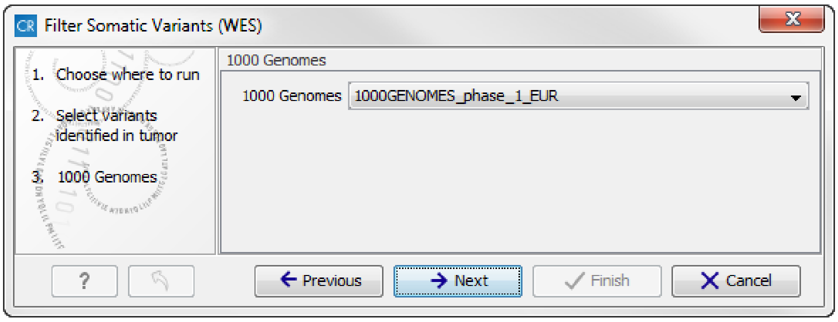

Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
Select variants identified in tumor
Navigation Area
whole_exome_sequencing
new_import
tumor
results
SRR719299_1 trimmed (Read:
tumor (Reads) - locally realign
SRR719299_1 trimmed (Read:
<enter search term>
Batch
Selected elements (1)
tumor (Reads) - locally reali...
?
Previous
Next
Finish
Cancel

Figure 13.15: *Select the variant track from which you would like to filter somatic variants.*

   Click on the button labeled **Next**.

2. In the next step you will be asked to specify which of the 1000 Genomes populations that should be used for annotation (figure 13.16).


Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
?
Previous
Next
Finish
Cancel

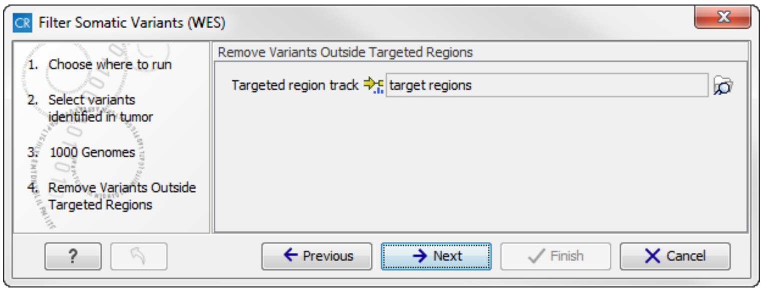
1000 Genomes

1000 Genomes 1000GENOMES_phase_1_EUR

Figure 13.16: *Specify which 1000 Genomes population to use for annotation.*

Click on the button labeled **Next**.

3. In this wizard step, you are asked to supply a track containing the targeted regions (figure 13.17). Select the track by clicking on the folder icon () in the wizard.


Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
?
Previous
Next
Finish
Cancel

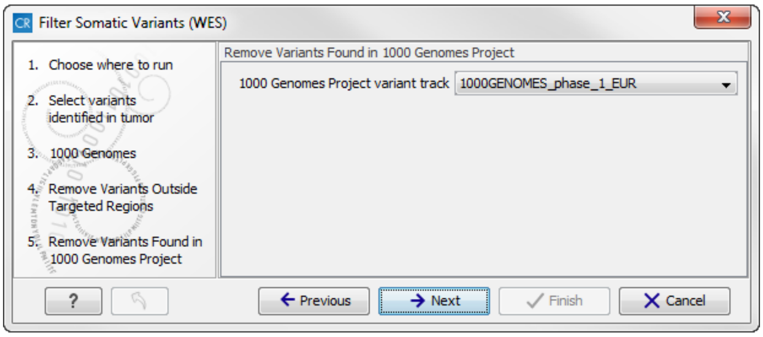
Remove Variants Outside Targeted Regions

Targeted region track target regions

Figure 13.17: *Select your target regions track.*

Click on the button labeled **Next**.

4. The next wizard step will once again allow you to specify the 1000 Genomes population that should be used, this time for filtering out variants found in the 1000 Genomes project (figure 13.18).


Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
?
Previous
Next
Finish
Cancel

Remove Variants Found in 1000 Genomes Project

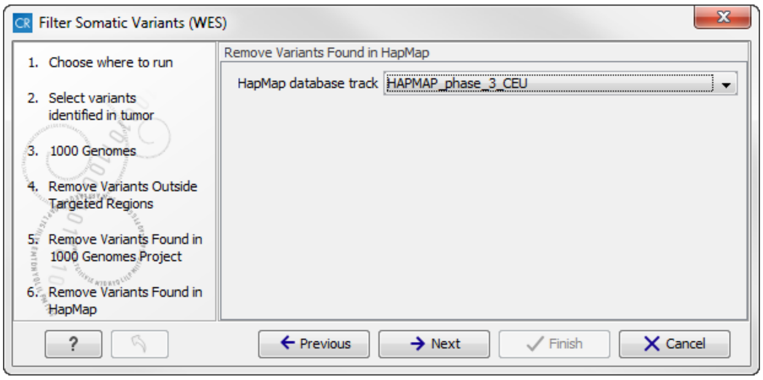
1000 Genomes Project variant track 1000GENOMES_phase_1_EUR

Figure 13.18: *Specify which 1000 Genomes population to use for filtering out known variants.*

Click on the button labeled **Next**.

5. The next wizard step (figure 13.19) concerns removal of variants found in the HapMap database. Select the population you would like to use from the drop-down list. Please note that the populations available from the drop-down list can be specified with the **Data**

Management ( ) function found in the top right corner of the Workbench (see section 10.1.4).

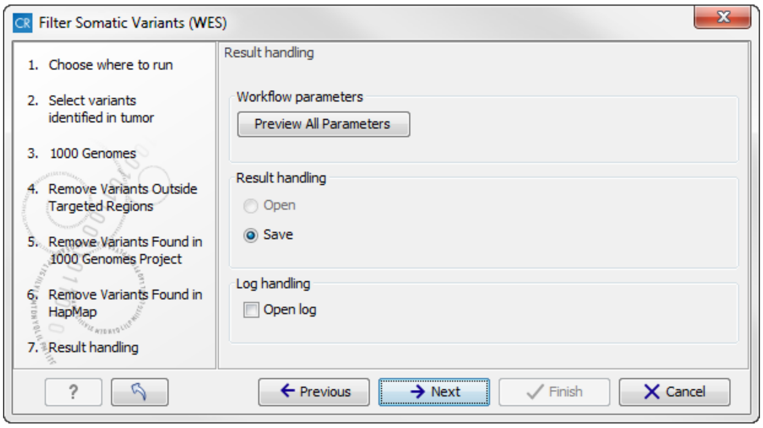

Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
6. Remove Variants Found in HapMap
?
Previous
Next
Finish
Cancel

Remove Variants Found in HapMap

HapMap database track HAPMAP_phase_3_CEU

Figure 13.19: *Specify which HapMap population to use for filtering out known variants.*

6. Click on the button labeled **Next** to go to the last wizard step (shown in figure 13.20).


Filter Somatic Variants (WES)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
6. Remove Variants Found in HapMap
7. Result handling
Preview All Parameters
?
Previous
Next
Finish
Cancel

Result handling

Workflow parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 13.20: *Check the selected parametes by pressing "Preview All Parameters".*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes. Choose to save the results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Genome Browser View. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.

2. **Genome Browser View Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, COSMIC, 1000 Genomes, and the PhastCons conservation scores (see figure 13.21).

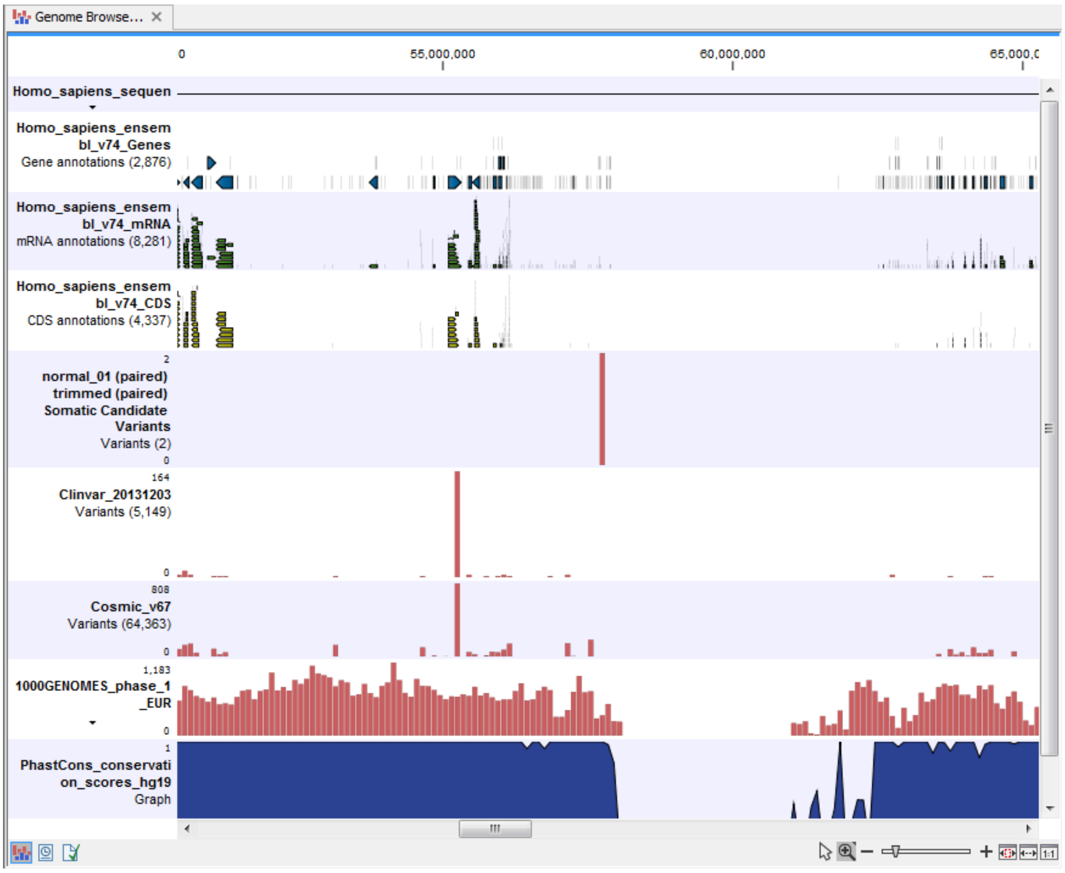


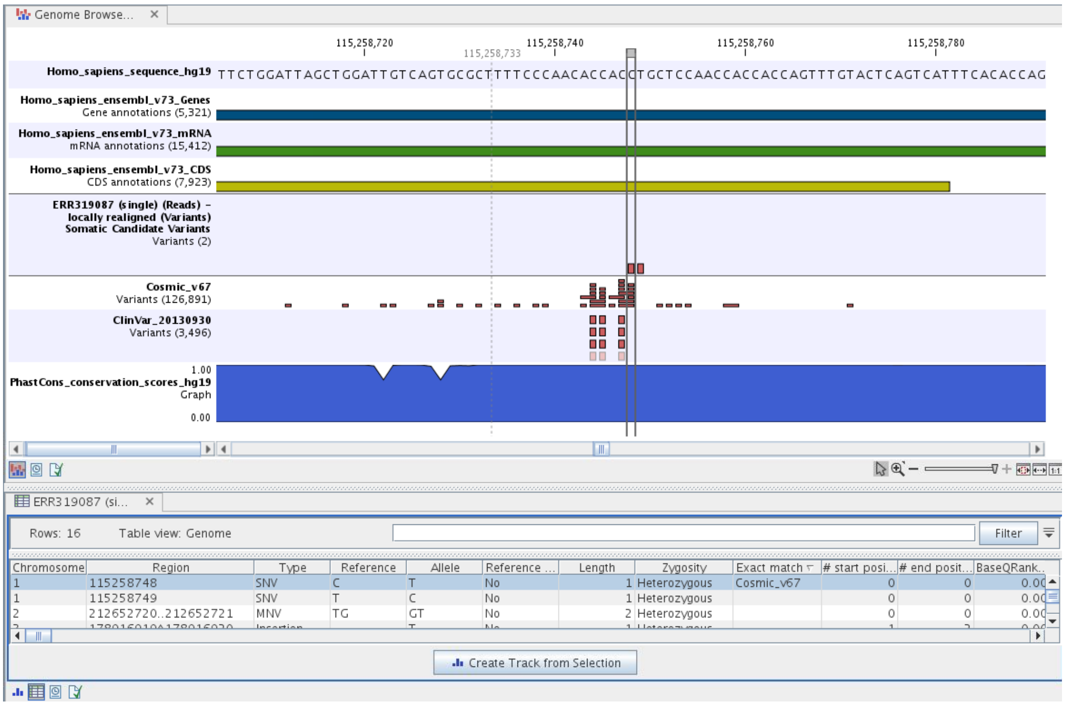
Figure 13.21: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well. Mapped sequencing reads as well as other tracks can be easily added to this Genome Browser View. By double clicking on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see figure 13.22).

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level, between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be


Genome Browse...
115,258,720
115,258,733
115,258,740
115,258,760
115,258,780
Homo_sapiens_sequence_hg19
TTCTGGATTAGCTGGATTGTCAGTGCGCTTTTCCCAACACCACCTGCTCCAACCACCACCAGTTTGTACTCAGTCATTTCACACCAG
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (15,412)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
ERR319087 (single) (Reads) - locally realigned (Variants) Somatic Candidate Variants
Variants (2)
Cosmic_v67
Variants (126,891)
ClinVar_20130930
Variants (3,496)
1.00
PhastCons_conservation_scores_hg19
Graph
0.00
ERR319087 (si...
Filter
Create Track from Selection

Rows: 16 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# end posit...	BaseQRank..
1	115258748	SNV	C	T	No	1	Heterozygous	Cosmic_v67	0	0	0.00
1	115258749	SNV	T	C	No	1	Heterozygous		0	0	0.00
2	212652720..212652721	MNV	TG	GT	No	2	Heterozygous		0	0	0.00

Figure 13.22: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

saved and reused. To do this:

Toolbox | **Identify Candidate Variants** (![]) | **Create Filter Criteria** (![])

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (`http://clccancer.com/software/#downloads`, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

13.5 Identify Somatic Variants from Tumor Normal Pair (WES)

The "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same patient.

When running the "Identify Somatic Variants from Tumor Normal Pair" the reads are mapped and the variants identified. An internal workflow removes germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads. Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from clinically relevant databases like COSMIC (known cancer associated variants) and ClinVar (variants with clinically relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

13.5.1 Import your targeted regions

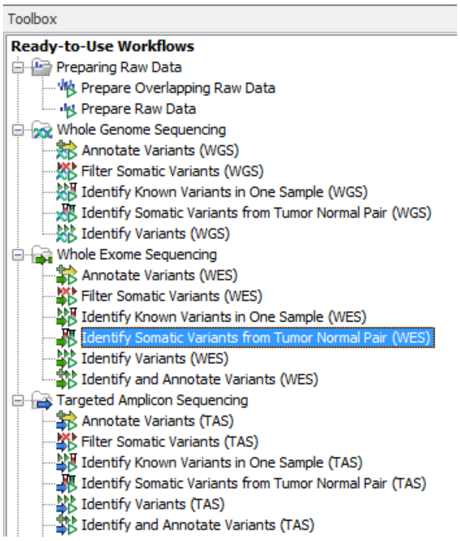
A file with the genomic regions targeted by the amplicon or hybridization kit is available from the vendor of the enrichment kit and sequencing machine. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get the file in either .bed or .gff format.

To import the file:

Go to the toolbar | Import () | Tracks ()

13.5.2 How to run the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow (figure 13.23).



Toolbox

Ready-to-Use Workflows
- Preparing Raw Data
  - Prepare Overlapping Raw Data
  - Prepare Raw Data
- Whole Genome Sequencing
  - Annotate Variants (WGS)
  - Filter Somatic Variants (WGS)
  - Identify Known Variants in One Sample (WGS)
  - Identify Somatic Variants from Tumor Normal Pair (WGS)
  - Identify Variants (WGS)
- Whole Exome Sequencing
  - Annotate Variants (WES)
  - Filter Somatic Variants (WES)
  - Identify Known Variants in One Sample (WES)
  - Identify Somatic Variants from Tumor Normal Pair (WES)
  - Identify Variants (WES)
  - Identify and Annotate Variants (WES)
- Targeted Amplicon Sequencing
  - Annotate Variants (TAS)
  - Filter Somatic Variants (TAS)
  - Identify Known Variants in One Sample (TAS)
  - Identify Somatic Variants from Tumor Normal Pair (TAS)
  - Identify Variants (TAS)
  - Identify and Annotate Variants (TAS)

Figure 13.23: *The ready-to-use workflows are found in the toolbox.*

   This will open the wizard shown in figure 13.24 where you can select the tumor sample reads.

   When you have selected the tumor sample reads click on the button labeled **Next**.

2. In the next wizard step (figure 13.25), please specify the normal sample reads.

3. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.26).

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.27). In this wizard step you can select your target regions track.

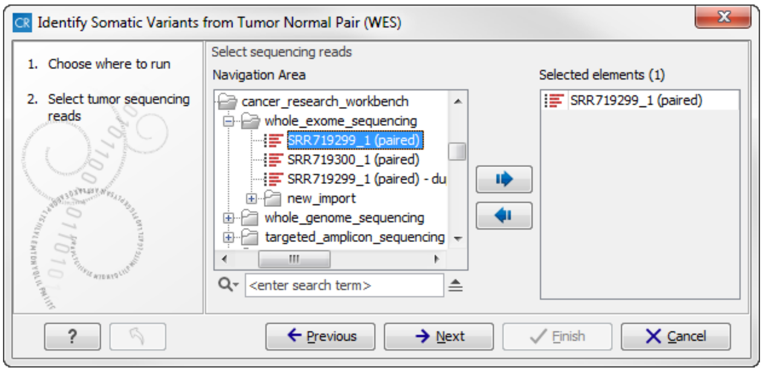


Figure 13.24: *Select the tumor sample reads.*

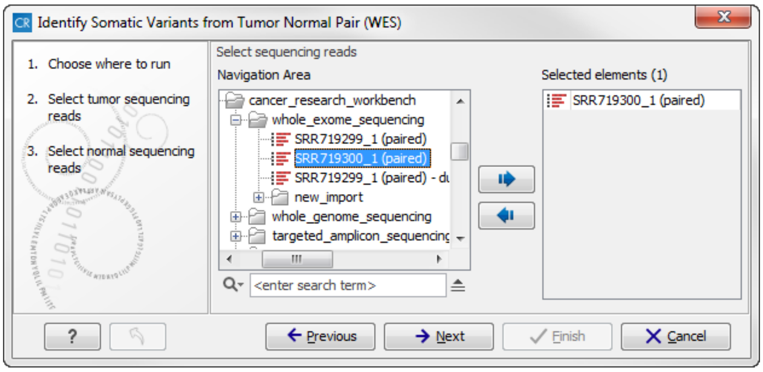


Figure 13.25: *Select the normal sample reads.*

5. Click on the button labeled **Next** to specify the target regions track to be used in the "Remove Variants Outside Targeted Regions" step (figure 13.28). The targeted region track should be the same as the track you selected in the previous wizard step. Variants found outside the targeted regions will not be included in the output that is generated with the ready-to-use workflow.

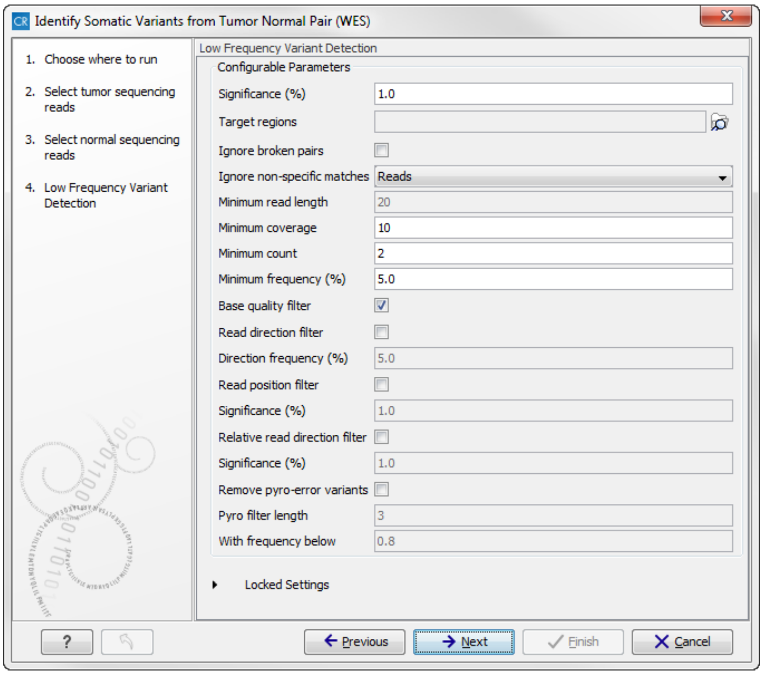
   Click on the button labeled **Next**.

6. Click on the button labeled **Next** to go to the step where you can adjust the settings for removal of germline variants (figure 13.29)..

7. Click on the button labeled **Next** and once again select the target region track (the same track as you have already selected in previous wizard steps). This time you specify the track to be used for quality control of the targeted sequencing as this tool reports the performance (enrichment and specificity) of a targeted re-sequencing experiment(figure 13.30).

   In the next wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters** (figure 13.31).

   In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.


Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection

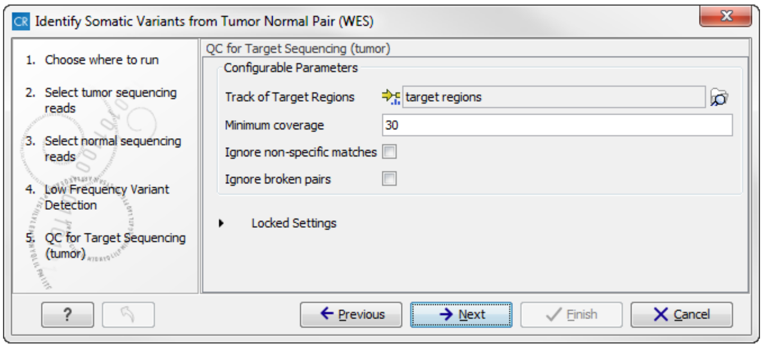
Low Frequency Variant Detection

Configurable Parameters

Significance (%) 1.0
Target regions
Ignore broken pairs □
Ignore non-specific matches Reads
Minimum read length 20
Minimum coverage 10
Minimum count 2
Minimum frequency (%) 5.0
Base quality filter ☑
Read direction filter □
Direction frequency (%) 5.0
Read position filter □
Significance (%) 1.0
Relative read direction filter □
Significance (%) 1.0
Remove pyro-error variants □
Pyro filter length 3
With frequency below 0.8

Locked Settings

Figure 13.26: *Specify the settings for the variant detection.*


Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. QC for Target Sequencing (tumor)

QC for Target Sequencing (tumor)

Configurable Parameters

Track of Target Regions target regions
Minimum coverage 30
Ignore non-specific matches □
Ignore broken pairs □

Locked Settings

Figure 13.27: *Select your target region track.*

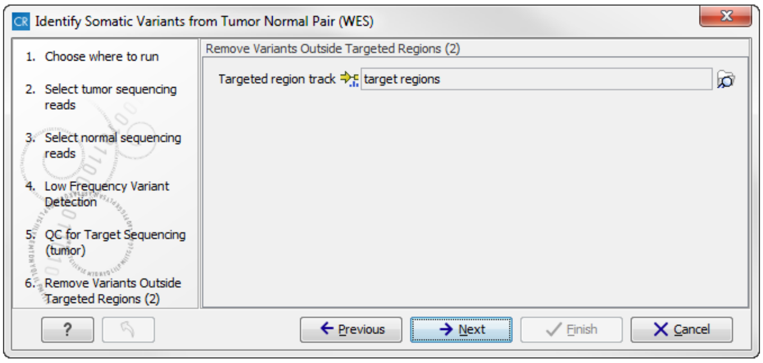
8. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Eight different outputs are generated:

1. **Read Mapping Normal** (≡) The mapped sequencing reads for the normal sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html

2. **Read Mapping Tumor** (≡) The mapped sequencing reads for the tumor sample. The reads are shown in different colors depending on their orientation, whether they are single



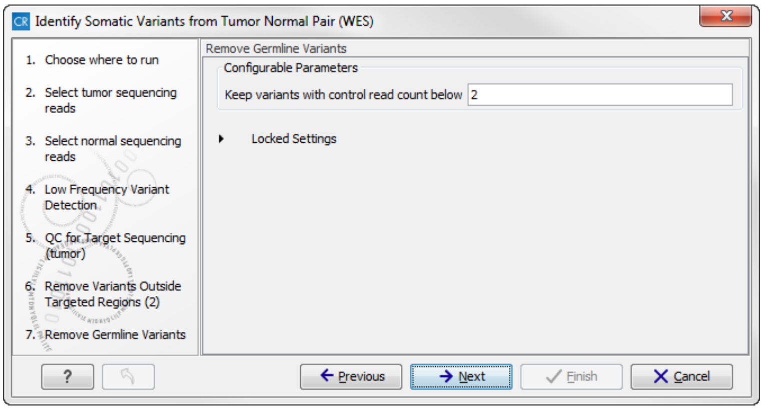
Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. QC for Target Sequencing (tumor)
6. Remove Variants Outside Targeted Regions (2)

Remove Variants Outside Targeted Regions (2)

Targeted region track: target regions

Figure 13.28: *Select your target region track.*



Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. QC for Target Sequencing (tumor)
6. Remove Variants Outside Targeted Regions (2)
7. Remove Germline Variants

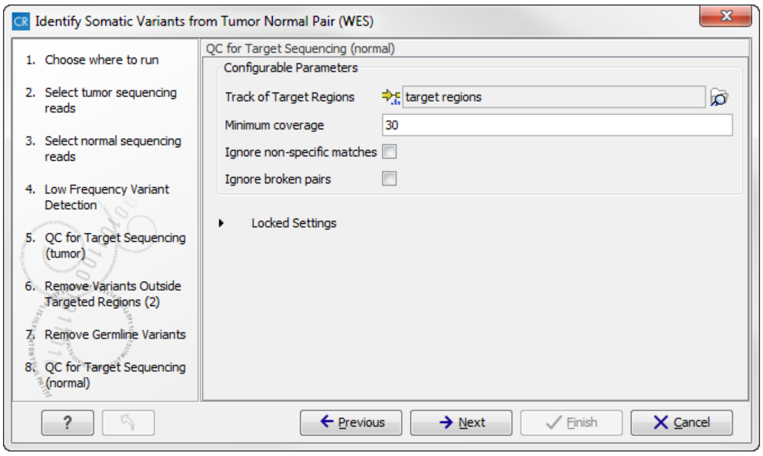
Remove Germline Variants

Configurable Parameters

Keep variants with control read count below: 2

Locked Settings

Figure 13.29: *Specify setting for removal of germline variants.*



Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. QC for Target Sequencing (tumor)
6. Remove Variants Outside Targeted Regions (2)
7. Remove Germline Variants
8. QC for Target Sequencing (normal)

QC for Target Sequencing (normal)

Configurable Parameters

Track of Target Regions: target regions

Minimum coverage: 30

Ignore non-specific matches ☐

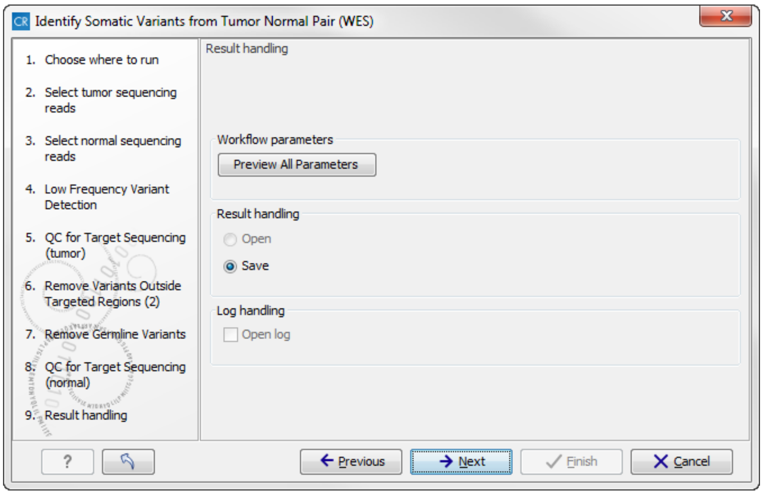
Ignore broken pairs ☐

Locked Settings

Figure 13.30: *Select target region track.*

reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@View_settings_in_Side_Panel.html.

3. **Target Region Coverage Report Normal** (📊) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the normal



Identify Somatic Variants from Tumor Normal Pair (WES)

1. Choose where to run
2. Select tumor sequencing reads
3. Select normal sequencing reads
4. Low Frequency Variant Detection
5. QC for Target Sequencing (tumor)
6. Remove Variants Outside Targeted Regions (2)
7. Remove Germline Variants
8. QC for Target Sequencing (normal)
9. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 13.31: *Check the parameters and save the results.*

sample.

4. **Target Region Coverage Tumor** () A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.

5. **Target Region Coverage Report Tumor** () The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.

6. **Variants** () A variant track holding the identified variants that are found in the targeted resions. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

7. **Annotated Somatic Variants** () A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

8. **Genome Browser View Tumor Normal Comparison** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar and COSMIC databases, and finally a track showing the conservation score (see figure 13.32).

13.6 Identify Known Variants in One Sample (WES)

The "Identify Known Variants in One Sample" ready-to-use workflow is a combined data analysis and interpretation ready-to-use workflow.

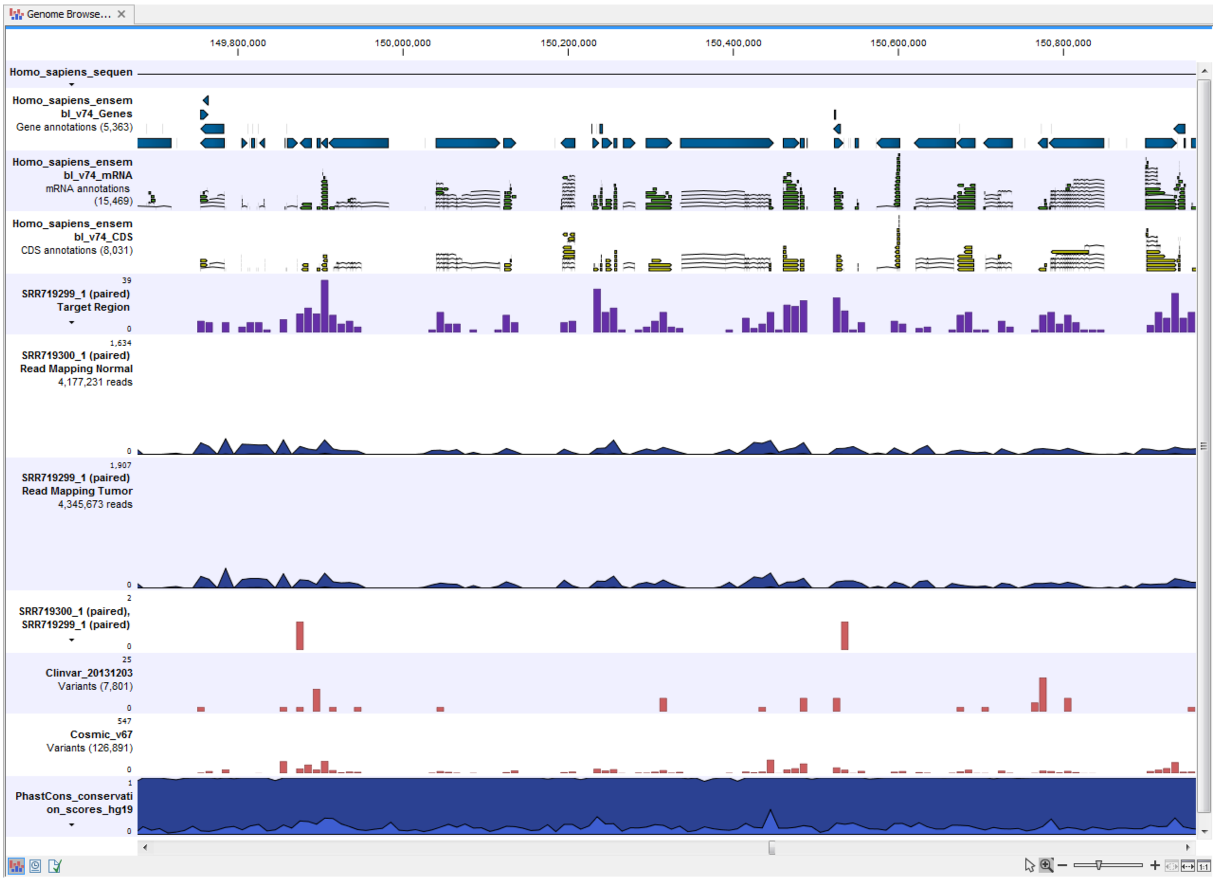

Genome Browse...
149,800,000
150,000,000
150,200,000
150,400,000
150,600,000
150,800,000
Homo_sapiens_sequen
Homo_sapiens_ensembl_v74_Genes
Gene annotations (5,363)
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (15,469)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (8,031)
39
SRR719299_1 (paired) Target Region
0
1,634
SRR719300_1 (paired) Read Mapping Normal
4,177,231 reads
0
1,907
SRR719299_1 (paired) Read Mapping Tumor
4,345,673 reads
0
2
SRR719300_1 (paired), SRR719299_1 (paired)
0
25
Clinvar_20131203
Variants (7,801)
0
547
Cosmic_v67
Variants (126,891)
0
1
PhastCons_conservation_scores_hg19
0

Figure 13.32: *The Genome Browser View presents all the different data tracks together and makes it easy to compare different tracks.*

It should be used to identify known variants, specified by the user (e.g. known breast cancer associated variants), for their presence or absence in a sample.

Please note that the ready-to-use workflow will not identify new variants.

The Identify Known Variants in One Sample ready-to-use workflow runs an internal workflow that maps the sequencing reads to the human genome sequence and does a local realignment of the mapped reads to improve the following variant detection. Next, specified variants by the user are identified in the read mapping. At the end, information present on the known variants before, are added to the results.

13.6.1 Import your known variants

To make an import into the Cancer Research Workbench, you should have your variants in GVF or VCF 4.1 format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

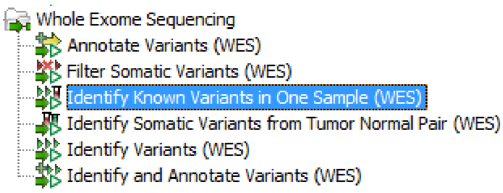
13.6.2 Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit will be provided by the vendor. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get it in either .bed or .gff format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

13.6.3 How to run the "Identify Known Variants in One Sample" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Known Variants from One Sample" ready-to-use workflow (figure 13.33).



Whole Exome Sequencing
- Annotate Variants (WES)
- Filter Somatic Variants (WES)
- Identify Known Variants in One Sample (WES)
- Identify Somatic Variants from Tumor Normal Pair (WES)
- Identify Variants (WES)
- Identify and Annotate Variants (WES)

Figure 13.33: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard step shown in figure 13.34 where you can select the reads of the sample, which should be tested for presence or absence of your known variants.

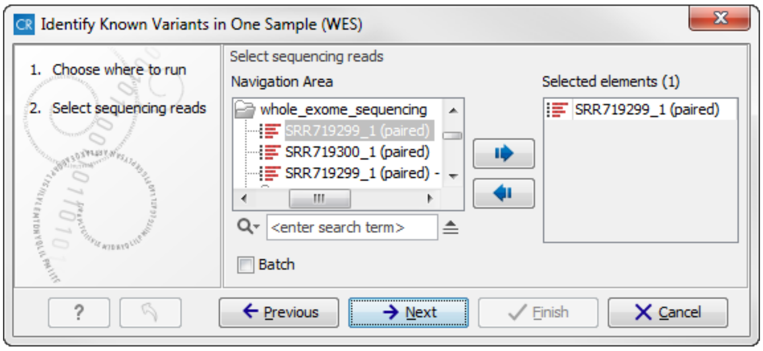


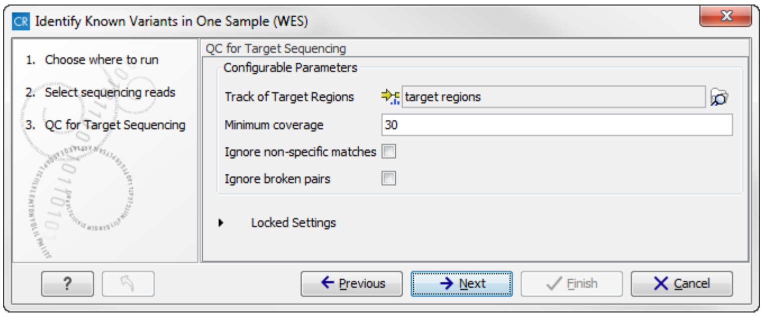
Figure 13.34: *Select the sequencing reads from the sample you would like to test for your known variants.*

Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 13.34) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

When you have selected the sample(s) you wish to analyze, click on the button labeled **Next**.

2. In the next wizard step you can select your target regions track and specify the minimum coverage to be used when checking the quality of the targeted sequencing. The minimum coverage will be used to provide the length of each target region that has at least this coverage. You can also specify whether or not to ignore non-specific matches and broken

pairs. When these are applied, reads that are non-specifically mapped or belong to broken pairs will be ignored (figure 13.35).



Track of Target Regions: target regions

Minimum coverage: 30

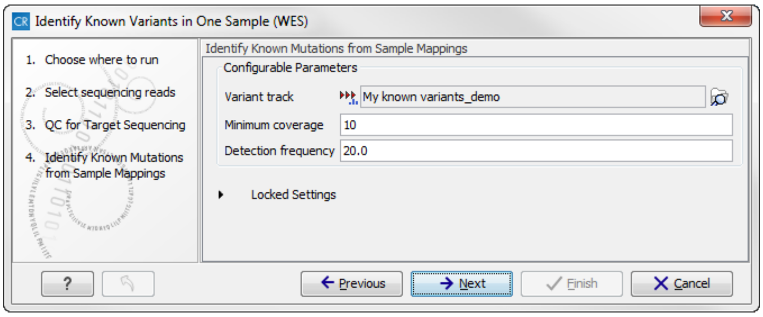
Ignore non-specific matches □

Ignore broken pairs □

Figure 13.35: *Select your target regions track and specify the parameters to be used for checking the quality of the targeted sequecing.*

3. Click on the button labeled **Next** and in specify the track with the known variants that should be identified in your sample (figure 13.36). Furthermore, in this wizard step you can specify the minimum read coverage for the position of the variant that should be identified. If the coverage at the position of the variant is below this, the result will show this.

   The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or not (observed frequency <= specified frequency). Moreover, it will determine if a variant should be labeled as heterozygous (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygous (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).


Identify Known Variants in One Sample (WES)
1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
Identify Known Mutations from Sample Mappings
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Variant track: My known variants_demo

Minimum coverage: 10

Detection frequency: 20.0

Figure 13.36: *Specify the track with the known variants that should be identified.*

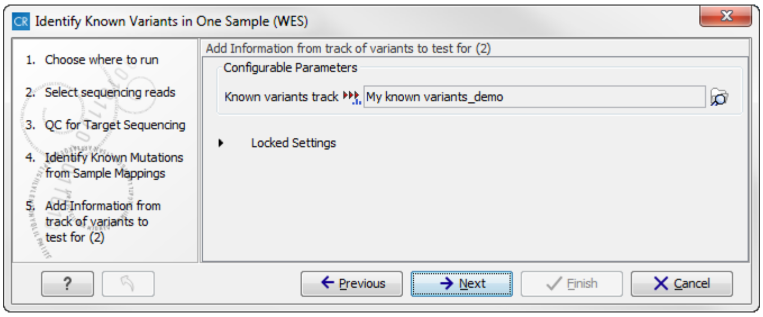
4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.37). In this and the next dialog, you will be asked about which of the annotations/informations added to variants should be included in the results.

   Please specify your track with known variants.

5. Click on the button labeled **Next** and once again select the same track with known variants (figure 13.38).

6. Click on the button labeled **Next** to go to the last wizard step (figure 13.39).

   In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the


Identify Known Variants in One Sample (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)

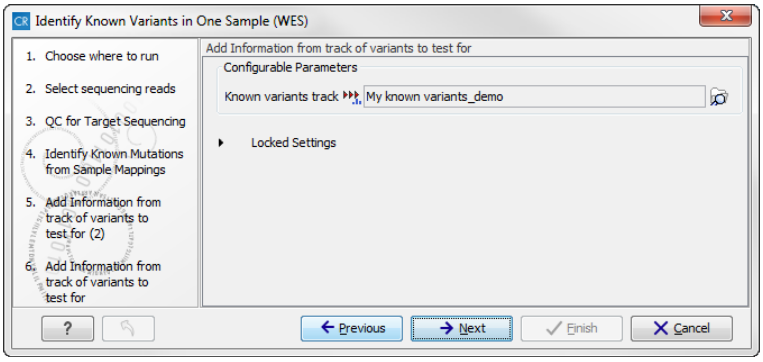
Add Information from track of variants to test for (2)

Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

Figure 13.37: *Please select the track with your known variants again. Annotations/Informations from this track will be added to the overview mutation track.*


Identify Known Variants in One Sample (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)
6. Add Information from track of variants to test for

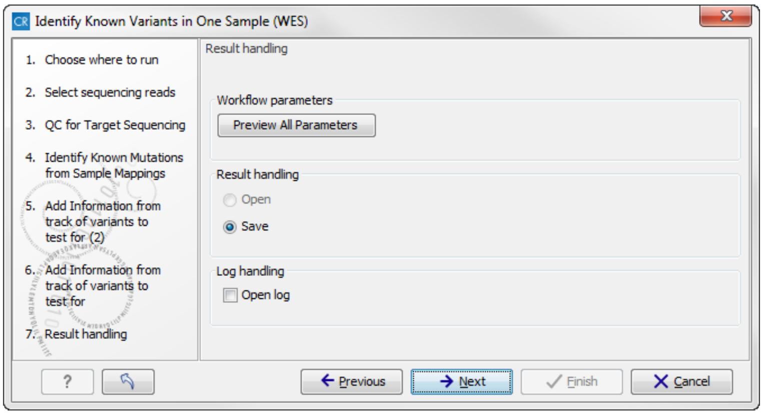
Add Information from track of variants to test for

Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

Figure 13.38: *Once again select the track with known variants. This time the track is used to add information to the detailed mutation track.*


Identify Known Variants in One Sample (WES)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)
6. Add Information from track of variants to test for
7. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 13.39: *Check the settings and save your results.*

settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

7. Click on the button labeled **OK** to go back to the previous dialog box and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

13.6.4 Output from the Identify Known Variants in One Sample

The "Identify Known Variants in One Sample" tool produces seven different output types:

1. **Read Mapping** () The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **Target Regions Coverage** () A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.

3. **Target Regions Coverage Report** () The report consists of a number of tables and graphs that in different ways show e.g. the number, length, and coverage of the target regions and provides information about the read count per GC%.

4. **Overview Variants Detected** () Annotation track showing the known variants. The table view provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads.

5. **Variants Detected in Detail** () Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. The difference between the two tables is that the "Variants Detected in Detail" table includes detailed information about the most frequent alternative allele (MFAA).

6. **Genome Browser View Identify Known Variants** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

7. **Log** () A log of the workflow execution.

It is a good idea to start looking at the Target Regions Coverage Report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, we also recommend that you check that the majority of the reads are mapping to the targeted region.

When you have inspected the target regions coverage report you can open the Genome Browser View Identify Known Variants file (see 13.40).

The Genome Browser View includes an overview track of the known variants and a detailed result track presented in the context of the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, and clinically relevant variants in the COSMIC databases.

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

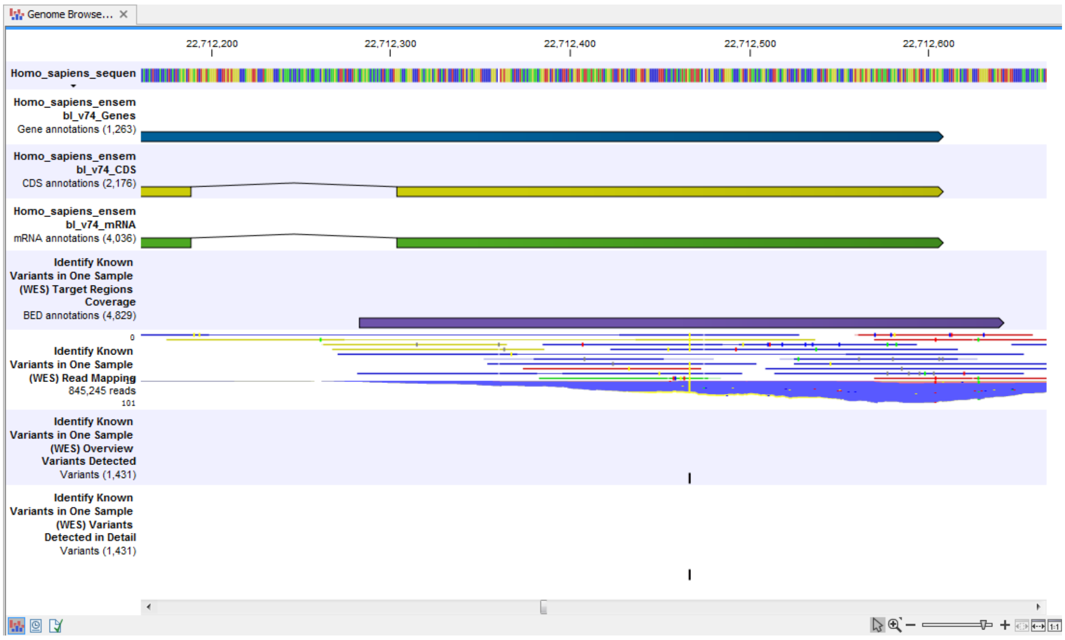

Genome Browse...
22,712,200
22,712,300
22,712,400
22,712,500
22,712,600
Homo_sapiens_sequen
Homo_sapiens_ensembl_v74_Genes
Gene annotations (1,263)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (2,176)
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (4,036)
Identify Known Variants in One Sample (WES) Target Regions Coverage
BED annotations (4,829)
0
Identify Known Variants in One Sample (WES) Read Mapping
845,245 reads
101
Identify Known Variants in One Sample (WES) Overview Variants Detected
Variants (1,431)
Identify Known Variants in One Sample (WES) Variants Detected in Detail
Variants (1,431)

Figure 13.40: *Genome Browser View that allows inspection of the identified variants in the context of the human genome and external databases.*

The difference between the overview variant track and the detailed variant track is the annotations added to the variants.

By double clicking on one of the annotated variant tracks in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see 13.41).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.

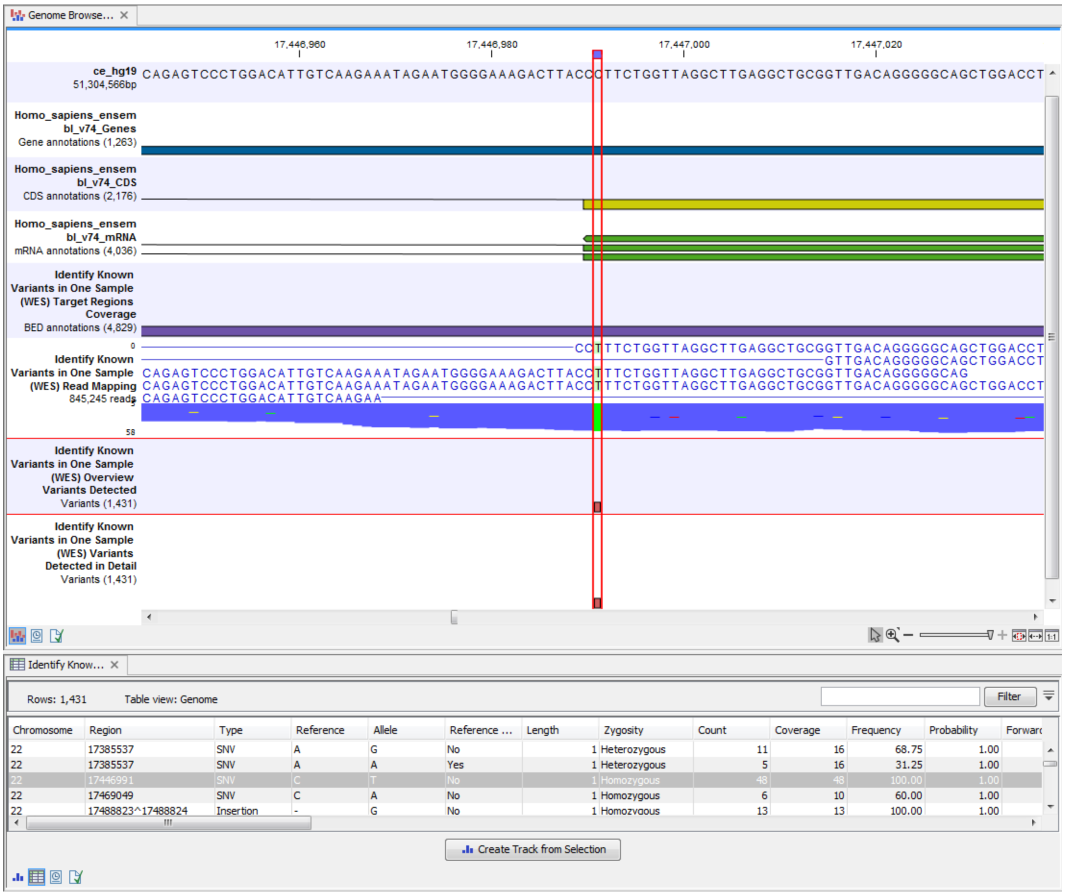
13.7 Identify and Annotate Variants (WES)

The "Identify and Annotate Variants" tool should be used to identify and annotate variants in one sample. The tool consists of a workflow that is a combination of the "Identify Variants" and the "Annotate Variants" workflows.

The tool runs an internal workflow, which starts with mapping the sequencing reads to the human reference sequence. Then it runs a local realignment to improve the variant detection, which is run afterwards. After the variants have been detected, they are annotated with gene names, amino acid changes, conservation scores, information from clinically relevant variants present in the COSMIC and ClinVar database, and information from common variants present in the common dbSNP, HapMap, and 1000 Genomes database. Furthermore, a detailed mapping report or a targeted region report (whole exome and targeted amplicon analysis) is created to inspect the overall coverage and mapping specificity.

Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit is available from the vendor of the enrichment kit and sequencing machine. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get the file


Genome Browse...
17,446,960
17,446,980
17,447,000
17,447,020
ce_hg19
51,304,566bp
Homo_sapiens_ensembl_v74_Genes
Gene annotations (1,263)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (2,176)
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (4,036)
Identify Known Variants in One Sample (WES) Target Regions Coverage
BED annotations (4,829)
Identify Known Variants in One Sample (WES) Read Mapping
845,245 reads
Identify Known Variants in One Sample (WES) Overview Variants Detected
Variants (1,431)
Identify Known Variants in One Sample (WES) Variants Detected in Detail
Variants (1,431)
Identify Know...
Filter
Create Track from Selection

Rows: 1,431 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Probability	Forwarc
22	17385537	SNV	A	G	No	1	Heterozygous	11	16	68.75	1.00	
22	17385537	SNV	A	A	Yes	1	Heterozygous	5	16	31.25	1.00	
22	17446991	SNV	C	T	No	1	Homozygous	48	48	100.00	1.00	
22	17469049	SNV	C	A	No	1	Homozygous	6	10	60.00	1.00	
22	17488823^17488824	Insertion	-	G	No	1	Homozygous	13	13	100.00	1.00	

Figure 13.41: *Genome Browser View with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

in either .bed or .gff format.

To import the file:

Go to the toolbar | Import (📥) | Tracks (📊)

How to run the "Identify and Annotate Variants" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify and Annotate Variants" ready-to-use workflow (figure 13.42).

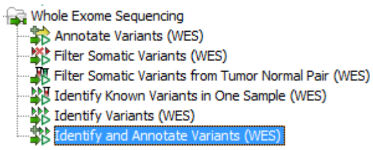
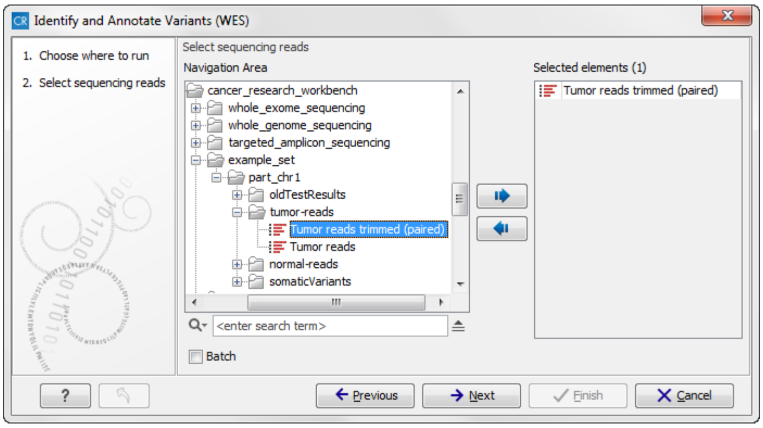

Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Filter Somatic Variants from Tumor Normal Pair (WES)
Identify Known Variants in One Sample (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)

Figure 13.42: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard shown in figure 13.43 where you can select the sequencing reads from the sample that should be analyzed.



Identify and Annotate Variants (WES)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

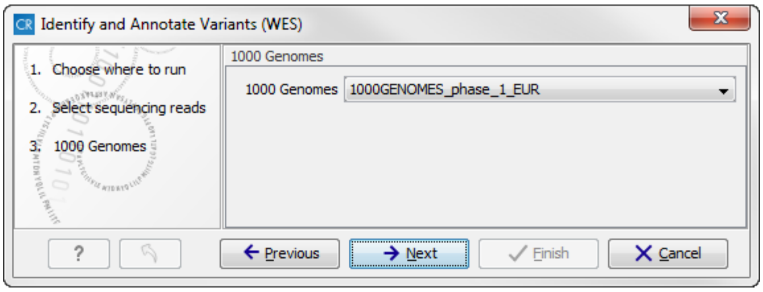
Selected elements (1): Tumor reads trimmed (paired)

Figure 13.43: *Please select all sequencing reads from the sample to be analyzed.*

If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 13.43) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

When you have selected the sample(s) you wish to prepare, click on the button labeled **Next**.

2. In the next wizard step (figure 13.44) you can select the population from the 1000 Genomes project that you would like to use for annotation.



Identify and Annotate Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes

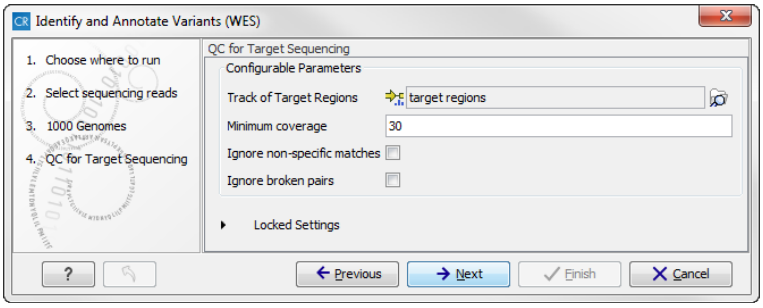
1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 13.44: *Select the population from the 1000 Genomes project that you would like to use for annotation.*

3. In the next wizard (figure 13.45) you can select the target region track and specify the minimum read coverage that should be present in the targeted regions.

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.46). In this dialog, you have to specify the parameters for the variant detection. For a description of the different parameters that can be adjusted in the variant detection step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html). As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the description of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/


Identify and Annotate Variants (WES)
1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
?
Previous
Next
Finish
Cancel

QC for Target Sequencing

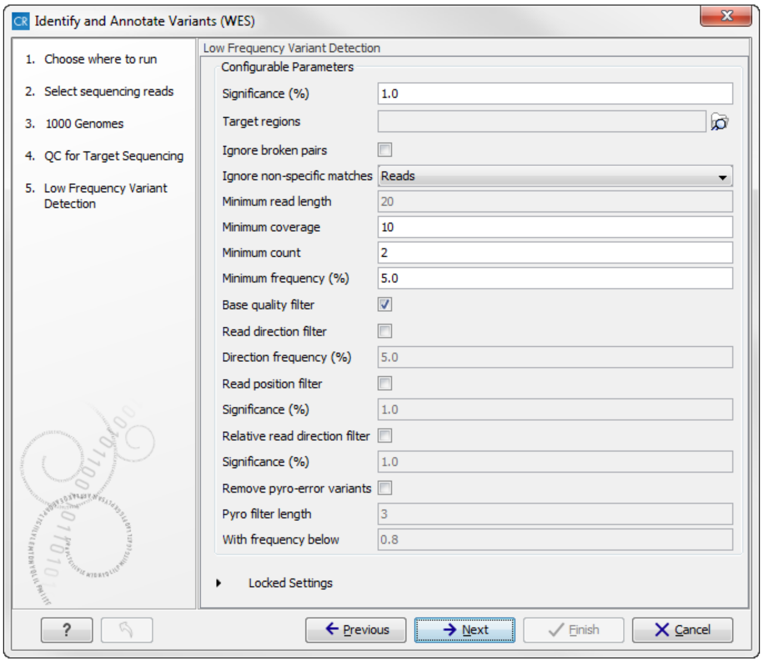
Configurable Parameters

Track of Target Regions	target regions
Minimum coverage	30
Ignore non-specific matches	☐
Ignore broken pairs	☐

Locked Settings

Figure 13.45: *Select the track with targeted regions from your experiment.*

clccancerresearchworkbench/current/index.php?manual=Filters.html). If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the ready-to-use workflow.


Identify and Annotate Variants (WES)
1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
?
Previous
Next
Finish
Cancel

Low Frequency Variant Detection

Configurable Parameters

Significance (%)	1.0
Target regions	
Ignore broken pairs	☐
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	2
Minimum frequency (%)	5.0
Base quality filter	☑
Read direction filter	☐
Direction frequency (%)	5.0
Read position filter	☐
Significance (%)	1.0
Relative read direction filter	☐
Significance (%)	1.0
Remove pyro-error variants	☐
Pyro filter length	3
With frequency below	0.8

Locked Settings

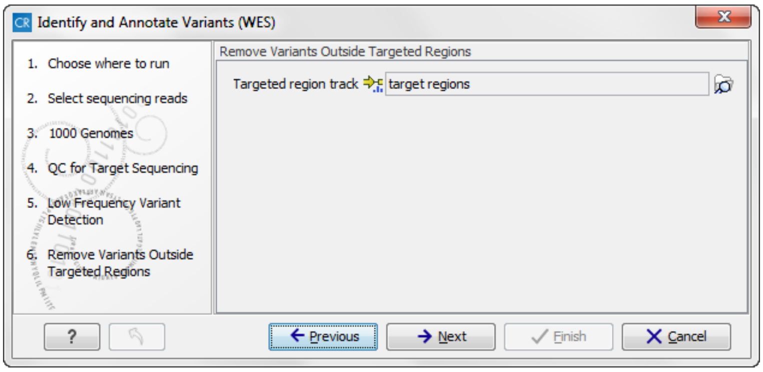
Figure 13.46: *Specify the parameters for variant calling.*

5. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.47). In this dialog you can specify the target regions track. The variants found outside the targeted region will be removed at this step in the workflow.

6. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.48). Once again, select the relevant population from the 1000 Genomes project. This will add information from the 1000 Genomes project to your variants.

7. Click on the button labeled **Next**, which will take you to the next wizard step (figure 13.49). At this step you can select a population from the HapMap database. This will add information from the Hapmap database to your variants.

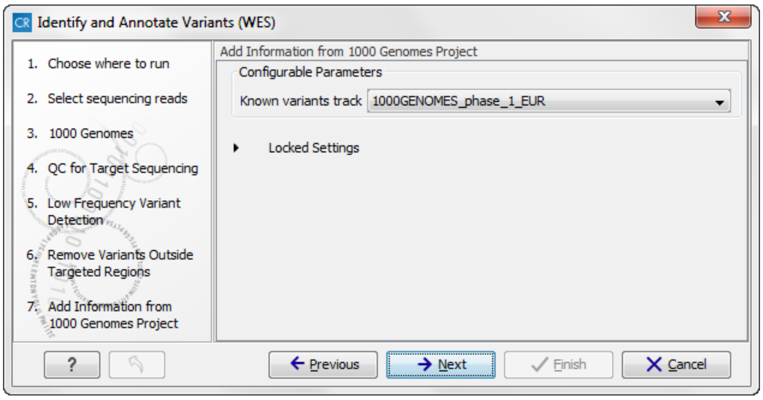
8. In this wizard step (figure 13.50) you get the chance to check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters**


Identify and Annotate Variants (WES)
1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
?
Previous
Next
Finish
Cancel

Remove Variants Outside Targeted Regions

Targeted region track: target regions

Figure 13.47: *In this wizard step you can specify the target regions track. Variants found outside these regions will be removed.*


Identify and Annotate Variants (WES)
1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project
Locked Settings
?
Previous
Next
Finish
Cancel

Add Information from 1000 Genomes Project

Configurable Parameters

Known variants track: 1000GENOMES_phase_1_EUR

Figure 13.48: *Select the relevant population from the 1000 Genomes project. This will add information from the 1000 Genomes project to your variants.*

wizard you can only check the settings, it is not possible to make any changes at this point.

9. Choose to **Save** your results and press **Finish**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Output from the Identify and Annotate Variants workflow

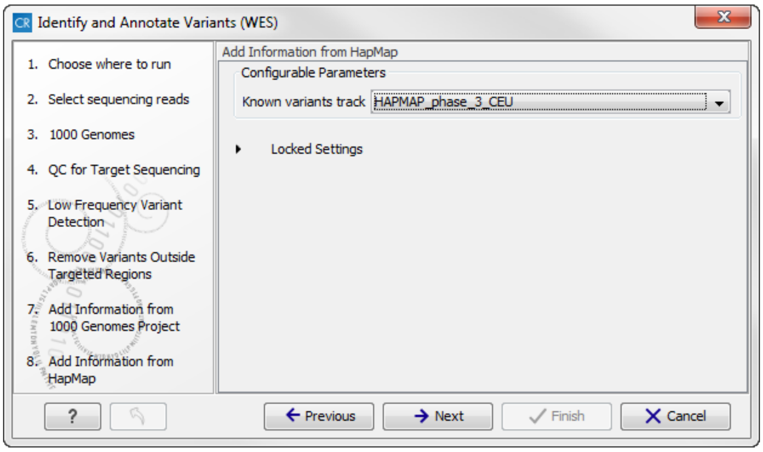
The "Identify and Annotate Variants" tool produces several outputs.

Please do not delete any of the produced files alone as some of them are linked to other outputs. Please always delete all of them at the same time.

A good place to start is to take a look at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Furthermore, please check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of the reads are mapping to the targeted region.

Next, open the Genome Browser View file (see figure 13.51).

The Genome Browser View includes a track of the identified annotated variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped



Identify and Annotate Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project
8. Add Information from HapMap

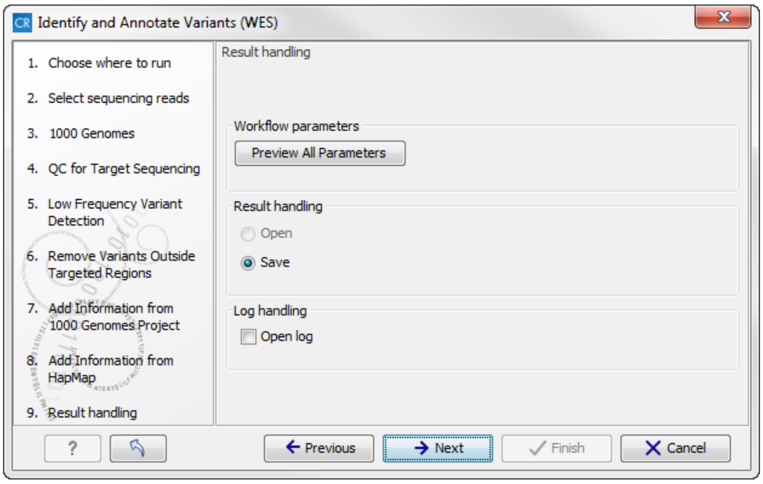
Add Information from HapMap

Configurable Parameters

Known variants track: HAPMAP_phase_3_CEU

▸ Locked Settings

Figure 13.49: *Select a population from the HapMap database. This will add information from the Hapmap database to your variants.*



Identify and Annotate Variants (WES)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project
8. Add Information from HapMap
9. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

☐ Open log

Figure 13.50: *Check the settings and save your results.*

sequencing reads, clinically relevant variants in the COSMIC and ClinVar database as well as common variants in common dbSNP, HapMap, and 1000 Genomes databases.

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well.

By double-clicking on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see 13.52).

The added information will help you to identify candidate variants for further research. For example can known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) easily be seen.

Not identified variants in COSMIC and ClinVar, can for example be prioritized based on amino acid changes (do they cause any changes on the amino acid level?). A high conservation level on the position of the variant between many vertebrates or mammals can also be a hint that this region could have an important functional role and variants with a conservation score of more

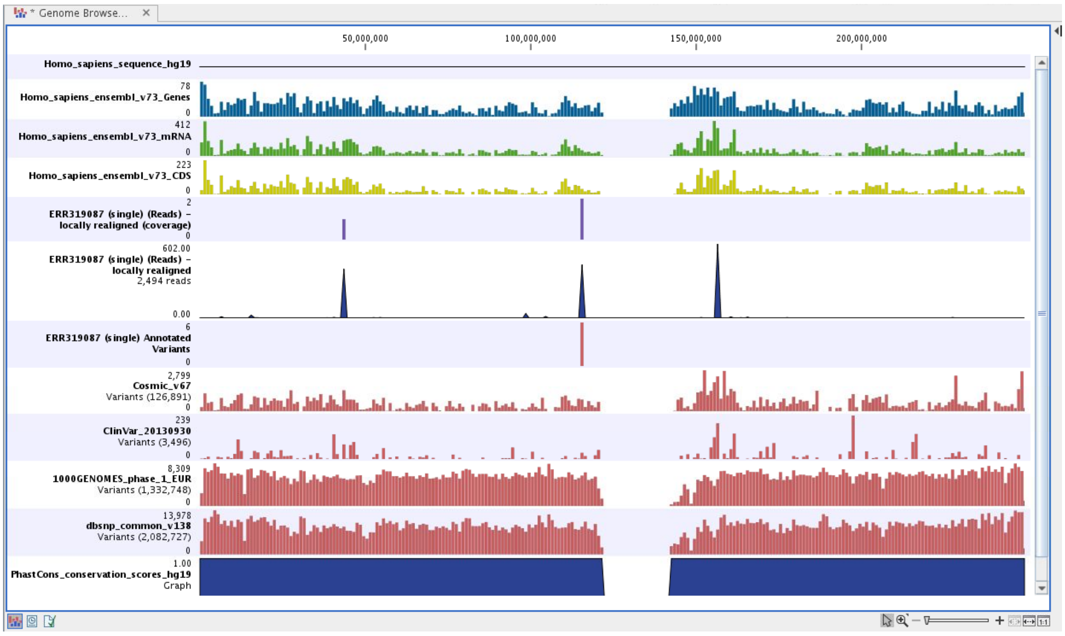
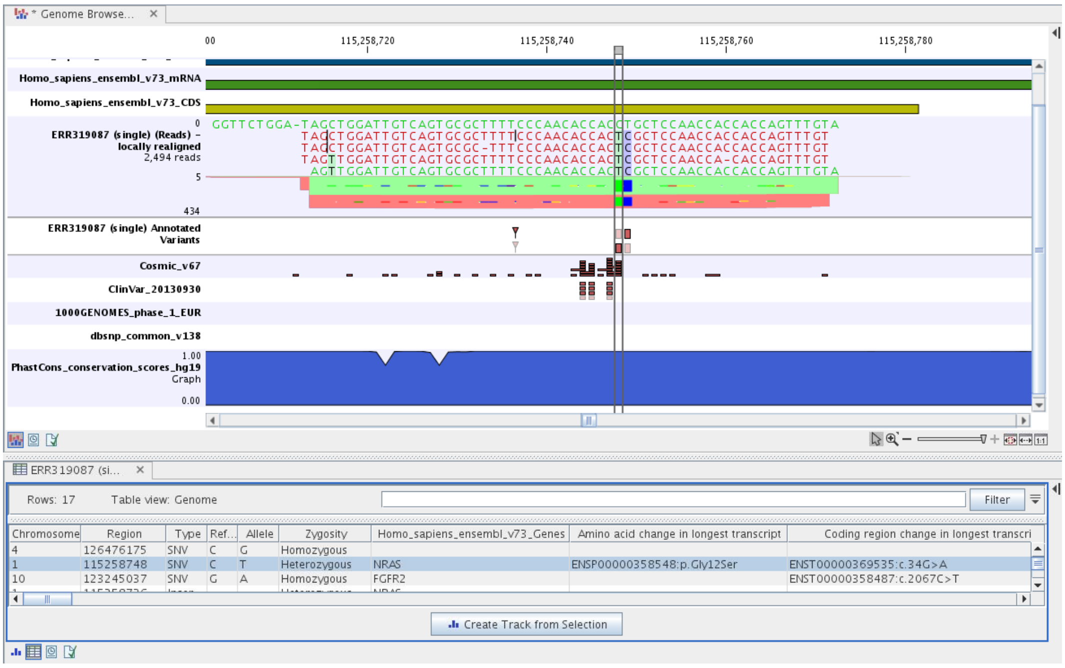

Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
Homo_sapiens_ensembl_v73_Genes
Homo_sapiens_ensembl_v73_mRNA
Homo_sapiens_ensembl_v73_CDS
ERR319087 (single) (Reads) – locally realigned (coverage)
ERR319087 (single) (Reads) – locally realigned 2,494 reads
ERR319087 (single) Annotated Variants
Cosmic_v67 Variants (126,891)
ClinVar_20130930 Variants (3,496)
1000GENOMES_phase_1_EUR Variants (1,332,748)
dbsnp_common_v138 Variants (2,082,727)
PhastCons_conservation_scores_hg19 Graph

Figure 13.51: *Genome Browser View to inspect identified variants in the context of the human genome and external databases.*


Genome Browse...
115,258,720
115,258,740
115,258,760
115,258,780
Homo_sapiens_ensembl_v73_mRNA
Homo_sapiens_ensembl_v73_CDS
ERR319087 (single) (Reads) – locally realigned 2,494 reads
ERR319087 (single) Annotated Variants
Cosmic_v67
ClinVar_20130930
1000GENOMES_phase_1_EUR
dbsnp_common_v138
PhastCons_conservation_scores_hg19 Graph
ERR319087 (si...
Filter
Create Track from Selection

Rows: 17 Table view: Genome

Chromosome	Region	Type	Ref...	Allele	Zygosity	Homo_sapiens_ensembl_v73_Genes	Amino acid change in longest transcript	Coding region change in longest transcri
4	126476175	SNV	C	G	Homozygous			
1	115258748	SNV	C	T	Heterozygous	NRAS	ENSP00000358548:p.Gly12Ser	ENST00000369535:c.34G>A
10	123245037	SNV	G	A	Homozygous	FGFR2		ENST00000358487:c.2067C>T

Figure 13.52: *Genome Browser View with an open track table to inspect identified somatic variants more closely in the context of the human genome and external databases.*

than 0.9 (PhastCons score) should be prioritized higher. A further filtering of the variants based on their annotations can be facilitated using the table filter on top of the table.

If you wish to always apply the same filter criteria, the "Create new Filter Criteria" tool should be used to specify this filter and the "Identify and Annotate" workflow should be extended by the "Identify Candidate Tool" (configured with the Filter Criterion). See the reference manual for more information on how preinstalled workflows can be edited.

Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

In case you like to change the databases as well as the used database version, please use the "Data Management".

Chapter 14

Targeted amplicon sequencing (TAS)

Contents

Targeted sequencing, also known as "targeted resequencing" or "amplicon sequencing" is a focused approach to genome sequencing with only selected areas of the genome being sequenced. In cancer research and diagnostics, targeted sequencing is usually based on sequencing panels that target a number of known cancer-associated genes.

14.1 Automatic analysis of sequencing data (TAS)

Six ready-to-use workflows are available for analysis of whole genome sequencing data. The concept of the pre-installed ready-to-use workflows is that read data are used as input in one end of the workflow and in the other end of the workflow you get a track based genome browser view and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual ready-to-use workflows can be used for and go through step by step how to run the workflows.

Note! Often you will have to prepare data with one of the two **Preparing Raw Data** workflows described in section 11 before you proceed to **Automatic analysis of sequencing data (TAS)**.

14.2 Identify Variants (TAS)

The "Identify Variants" tool takes sequencing reads as input and returns identified variants as part of a Genome Browser View.

The tool runs an internal workflow, which starts with mapping the sequencing reads to the human reference sequence. Then it runs a local realignment to improve the variant detection, which is run afterwards. At the end, variants with an average base quality smaller than 20 are filtered away.

In addition, a targeted region report is created to inspect the overall coverage and mapping specificity in the targeted regions.

Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit will be provided by the vendor. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get it in either .bed or .gff format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

How to run the "Identify Variants" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Variants" ready-to-use workflow (figure 14.1).

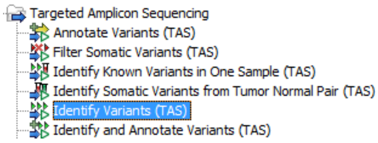

Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Identify Known Variants in One Sample (TAS)
Identify Somatic Variants from Tumor Normal Pair (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)

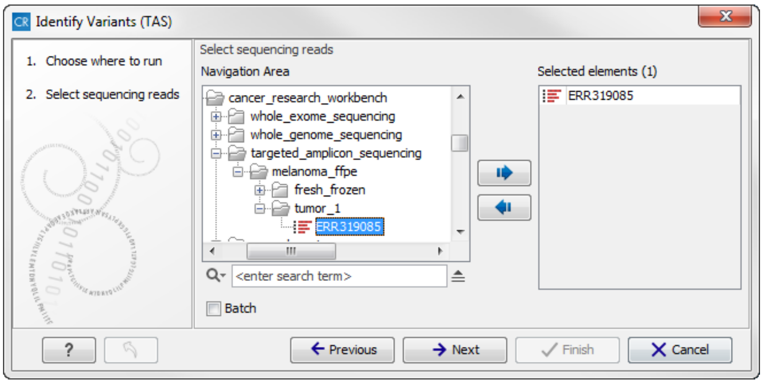
Figure 14.1: *The ready-to-use workflows are found in the toolbox.*

   This will open the wizard shown in figure 14.2 where you can select the sequencing reads from the sample, which should be analyzed.

   Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 14.43) and select the **folder** that holds the data you wish to analyze. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

   When you have selected the sample(s) you wish to prepare, click on the button labeled **Next**.

2. In the next wizard step (figure 14.3) you have to specify the track with the targeted regions from the experiment. You can also specify the minimum read coverage, which should be present in the targeted regions.



Identify Variants (TAS)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

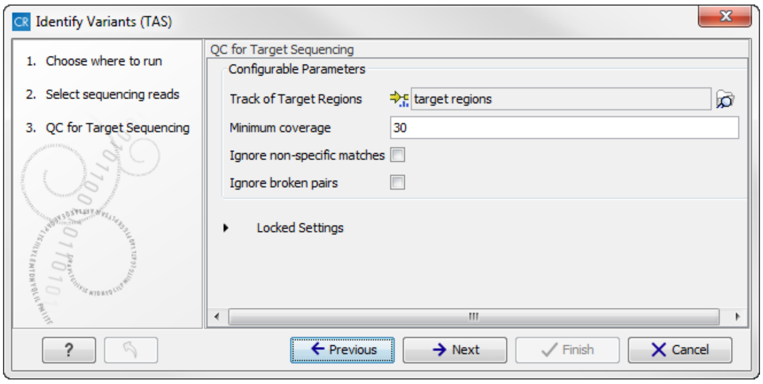
Navigation Area

- cancer_research_workbench
  - whole_exome_sequencing
  - whole_genome_sequencing
  - targeted_amplicon_sequencing
    - melanoma_ffpe
      - fresh_frozen
      - tumor_1
        - ERR319085

Selected elements (1)

- ERR319085

Figure 14.2: *Please select all sequencing reads from the sample to be analyzed.*



Identify Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing

QC for Target Sequencing

Configurable Parameters

Track of Target Regions: target regions

Minimum coverage: 30

Ignore non-specific matches □

Ignore broken pairs □

Figure 14.3: *Select the track with the targeted regions from your experiment.*

3. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.4). In this wizard you can specify the parameter for detecting variants.

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.5).

5. Click on the button labeled **Next** to go to the last wizard step (figure 14.6).

   In this wizard you get the chance to check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard step you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

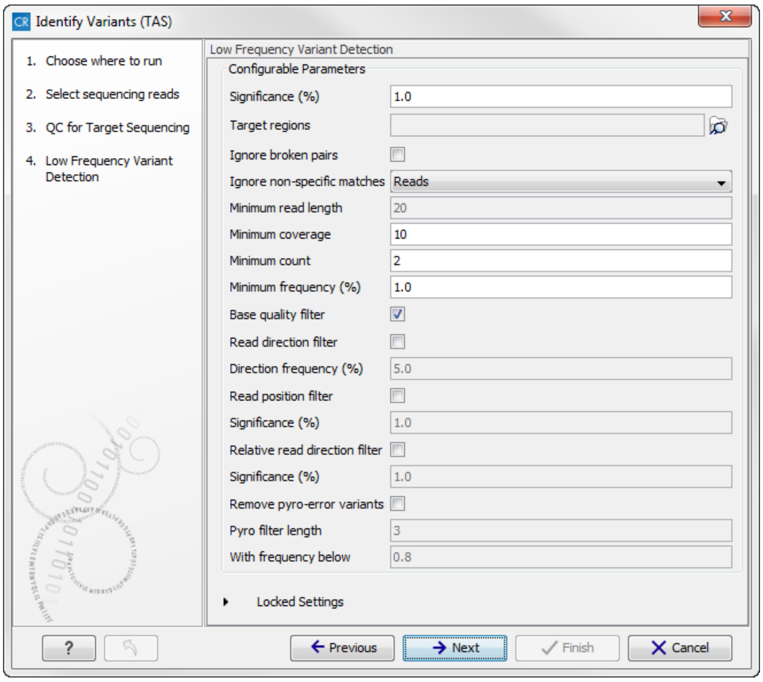
6. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Output from the Identify Variants workflow

The "Identify Variants" tool produces six different types of output:

1. **Read Mapping** (☰) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads,



Identify Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection

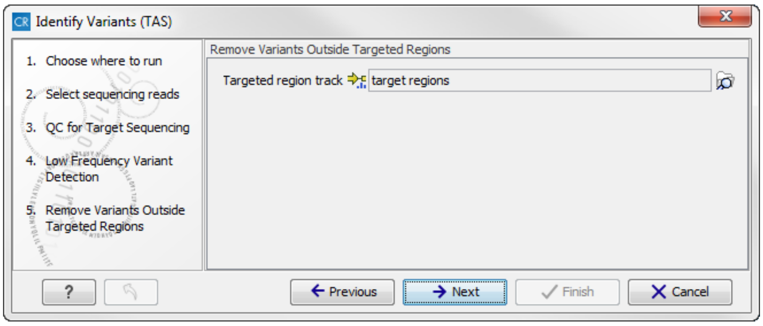
Low Frequency Variant Detection

Configurable Parameters

Significance (%)	1.0
Target regions	
Ignore broken pairs	☐
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	2
Minimum frequency (%)	1.0
Base quality filter	☑
Read direction filter	☐
Direction frequency (%)	5.0
Read position filter	☐
Significance (%)	1.0
Relative read direction filter	☐
Significance (%)	1.0
Remove pyro-error variants	☐
Pyro filter length	3
With frequency below	0.8

Locked Settings

Figure 14.4: *Please specify the parameters for variant detection.*



Identify Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection
5. Remove Variants Outside Targeted Regions

Remove Variants Outside Targeted Regions

Targeted region track: target regions

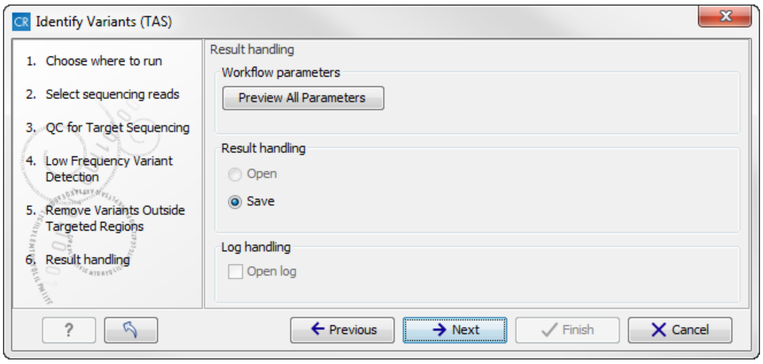
Figure 14.5: *Select the targeted region track. Variants found outside the targeted region will be removed.*

and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **Target Regions Coverage** () The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

3. **Target Regions Coverage Report** () The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

4. **Identified Variants** () A variant track holding the identified variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual



Identify Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Low Frequency Variant Detection
5. Remove Variants Outside Targeted Regions
6. Result handling

Result handling

Workflow parameters

Result handling
○ Open
● Save

Log handling
□ Open log

Figure 14.6: *Choose to save the results. In this wizard step you get the chance to preview the settings used in the ready-to-use workflow.*

variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

5. **Genome Browser View Identify Variants** () A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 14.12).

It is important that you do not delete any of the produced files individually as some of the outputs are linked to other outputs. If you would like to delete the outputs, please always delete all of them at the same time.

Please have first a look at the mapping report to see if the coverage is sufficient in regions of interest (e.g. > 30). Furthermore, please check that at least 90% of reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of reads are mapping to the targeted region.

Afterwards please open the Genome Browser View file (see 14.7).

The Genome Browser View includes the track of identified variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions and mapped sequencing reads.

By double clicking on the variant track in the Genome Browser View, a table will be shown which includes information about all identified variants (see 14.8).

In case you like to change the reference sequence used for mapping as well as the human genes, please use the "Data Management".

14.3 Annotate Variants (TAS)

Using a variant track () (e.g. the output from the Identify Variants ready-to-use workflow) the **Annotate Variants (WGS)** ready-to-use workflow runs an "internal" workflow that adds the following annotations to the variant track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.

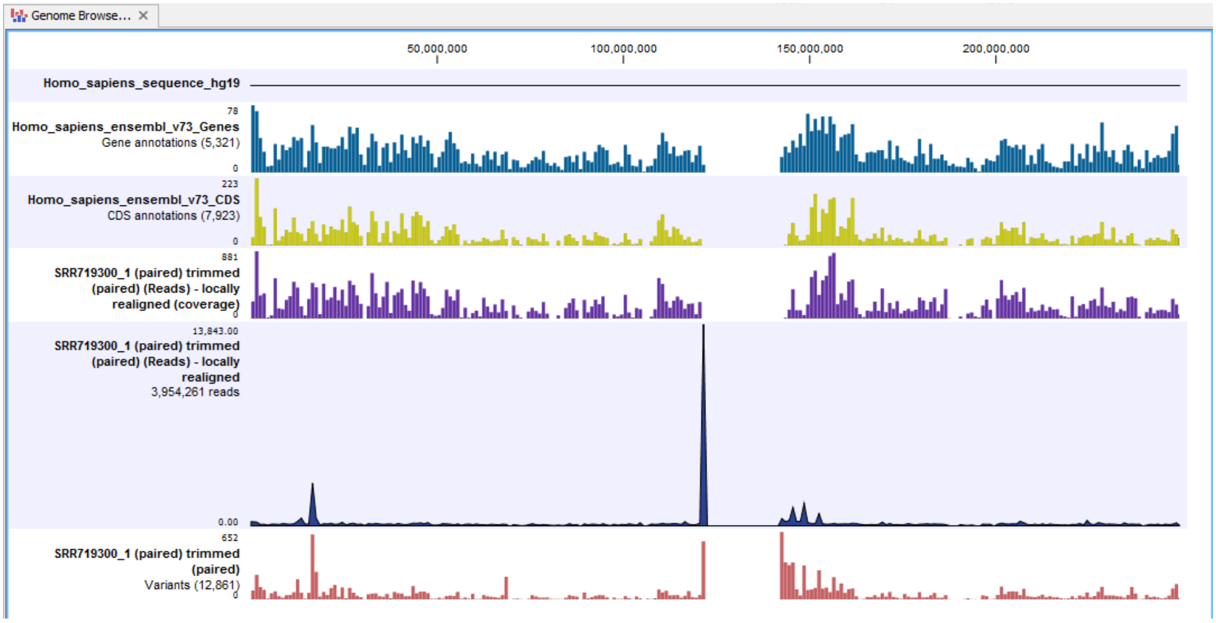
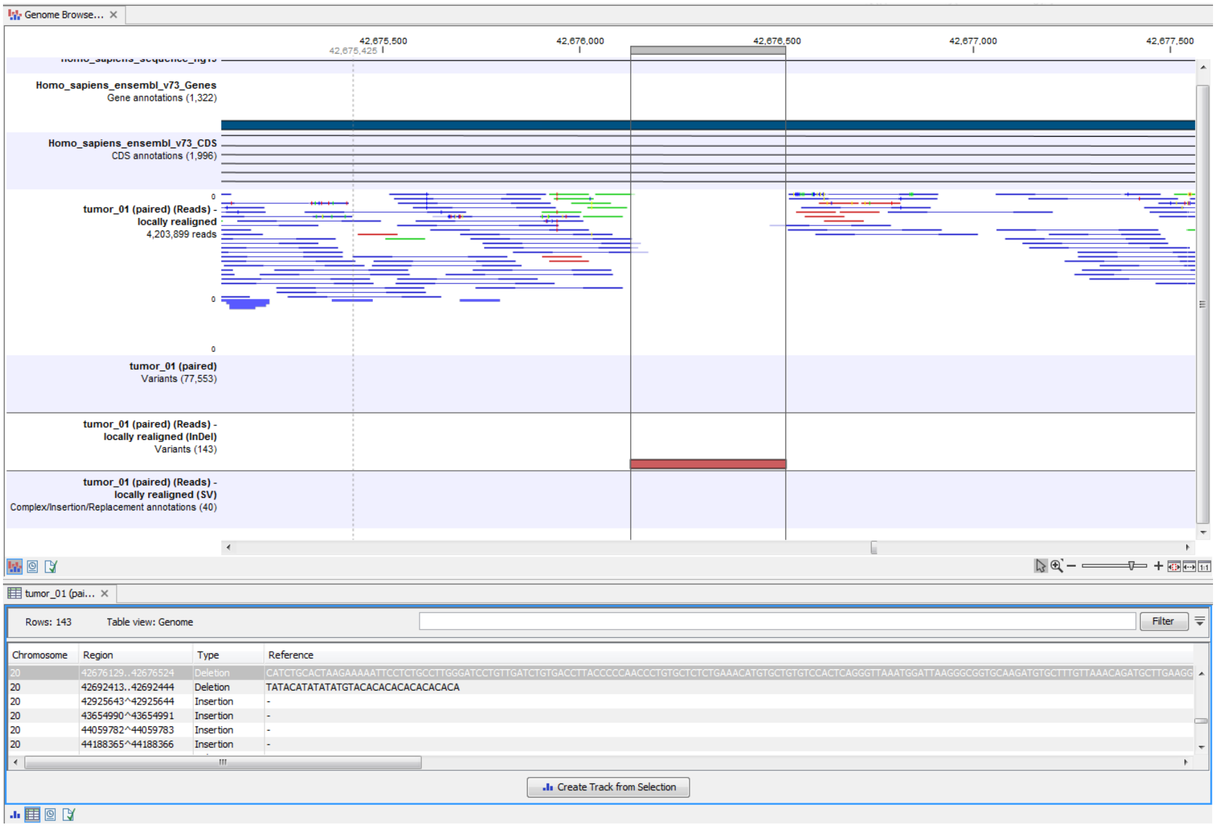

Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
78
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
0
223
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
0
881
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned (coverage)
0
13,843.00
SRR719300_1 (paired) trimmed (paired) (Reads) - locally realigned
3,954,261 reads
0.00
652
SRR719300_1 (paired) trimmed (paired)
Variants (12,861)
0

Figure 14.7: *The Genome Browser View allows you to inspect the identified variants in the context of the human genome.*

- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from COSMIC**. Adds information from the "Catalogue of Somatic Mutations in Cancer" database.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (InDels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

1. Go to the toolbox and select the **Annotate Variants (TAS)** workflow. In the first wizard step, select the input variant track (figure 14.9).
2. Click on the button labeled **Next**. The only parameter that should be specified by the user is which 1000 Genomes population yo use (figure 14.10). This can be done using the drop-down list found in this wizard step. Please note that the populations available from the drop-down list can be specified with the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).
3. Click on the button labeled **Next** to go to the last wizard step (figure 14.11).


Genome Browse...
42,675,500
42,675,425
42,676,000
42,676,500
42,677,000
42,677,500
Homo_sapiens_ensembl_v73_Genes
Gene annotations (1,322)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (1,996)
tumor_01 (paired) (Reads) - locally realigned
4,203,899 reads
tumor_01 (paired)
Variants (77,553)
tumor_01 (paired) (Reads) - locally realigned (InDel)
Variants (143)
tumor_01 (paired) (Reads) - locally realigned (SV)
Complex/Insertion/Replacement annotations (40)
tumor_01 (pai...
Filter
Create Track from Selection

Rows: 143 Table view: Genome

Chromosome	Region	Type	Reference
20	42676129..42676524	Deletion	CATCTGCACTAAGAAAAATTCCTCTGCCTTGGGATCCTGTTGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGAAACATGTGCTGTGTCCACTCAGGGTTAAATGGATTAAGGGCGGTGCAAGATGTGCTTTGTTAAACAGATGCTTGAAGG
20	42692413..42692444	Deletion	TATACATATATATGTACACACACACACACACA
20	42925643^42925644	Insertion	-
20	43654990^43654991	Insertion	-
20	44059782^44059783	Insertion	-
20	44188365^44188366	Insertion	-

Figure 14.8: *Genome Browser View with an open track table to inspect identified variants more closely in the context of the human genome.*

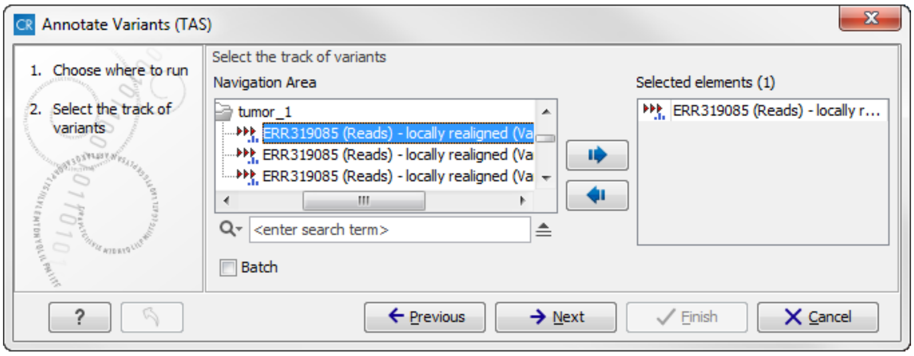

Annotate Variants (TAS)
1. Choose where to run
2. Select the track of variants
Select the track of variants
Navigation Area
tumor_1
ERR319085 (Reads) - locally realigned (Va
ERR319085 (Reads) - locally realigned (Va
ERR319085 (Reads) - locally realigned (Va
<enter search term>
□ Batch
Selected elements (1)
ERR319085 (Reads) - locally r...
?
Previous
Next
Finish
Cancel

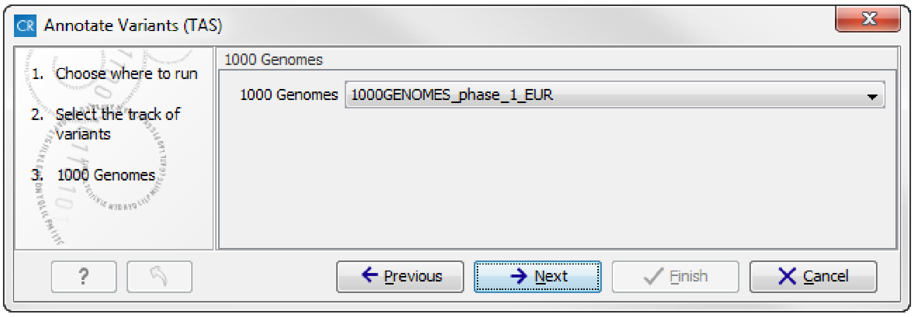
Figure 14.9: *Select the variant track to annotate.*

In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point.

4. Choose to **Save** your results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Annotated Variants** (▸▸▸) Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information



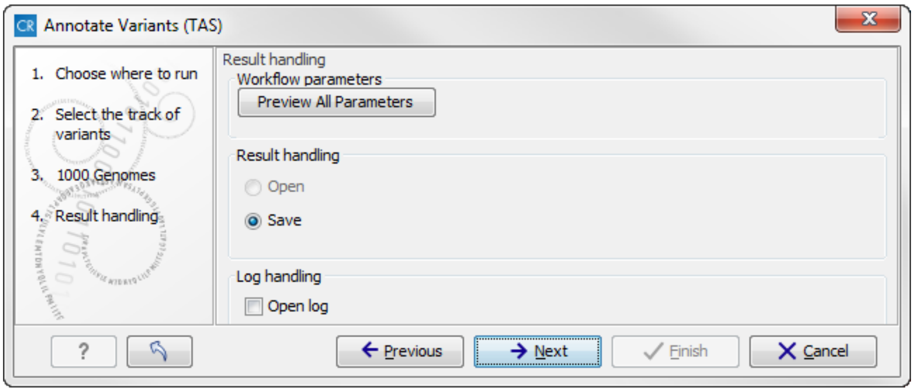
Annotate Variants (TAS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes

1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 14.10: *Select the relevant 1000 Genomes popultaion(s).*



Annotate Variants (TAS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes
4. Result handling

Result handling

Workflow parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 14.11: *Check the settings and save your results.*

about the variant.

2. **Genome Browser View Annotated Variants** (📊) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP, ClinVar, COSMIC, 1000 Genomes, and PhastCons conservation scores (see figure 14.12).

Note! Please be aware, that if you delete the annotated variant track, this track will also disappear from the genome browser view.

It is possible to add tracks to the Genome Browser View such as mapped sequencing reads as well as other tracks. This can be done by dragging the track directly from the **Navigation Area** to the Genome Browser View.

If you double-click on the name of the annotated variant track in the left hand side of the Genome Browser View, a table that includes all variants and the added information/annotations will open (see figure 14.13). The table and the Genome Browser View are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Genome Browser View.

You may be met with a warning as shown in figure 14.14. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

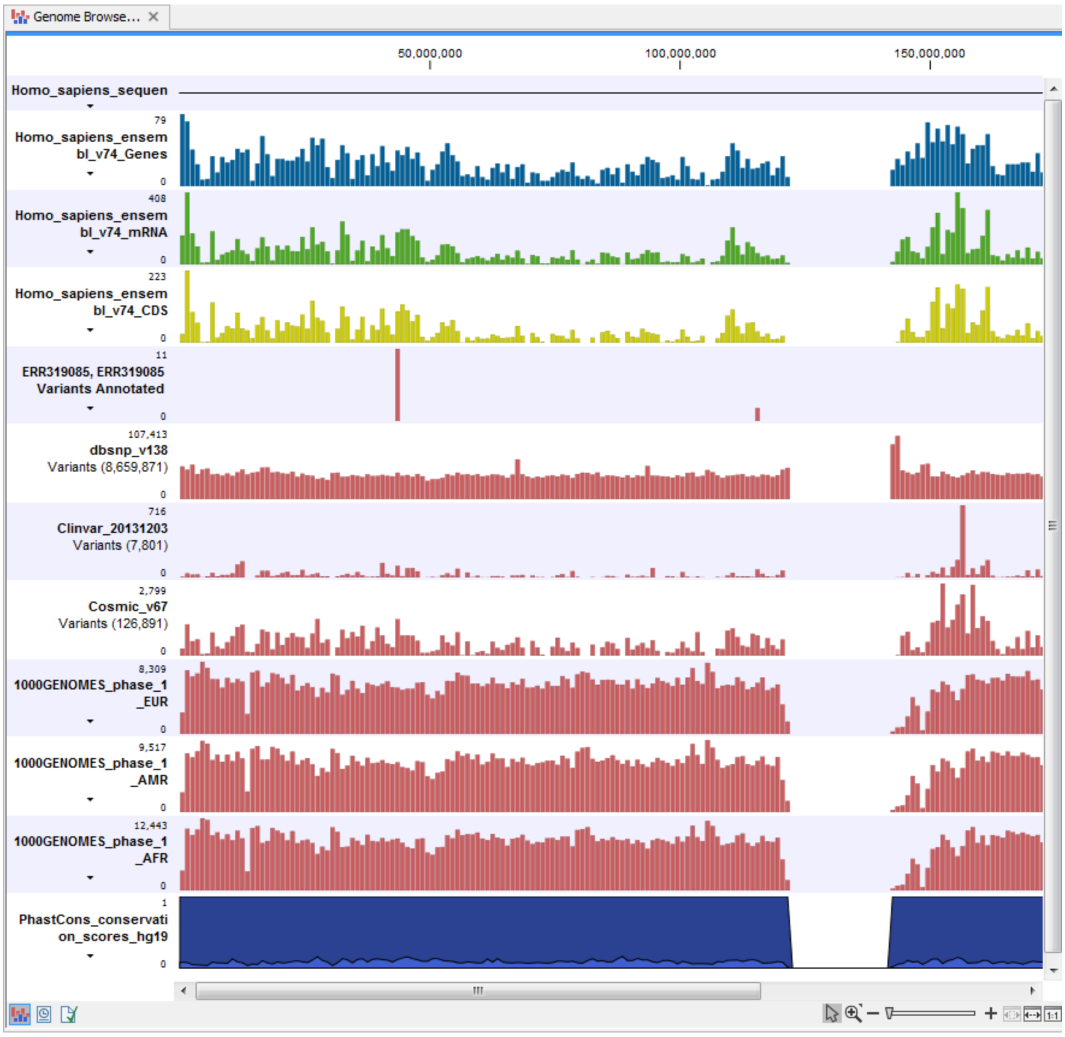
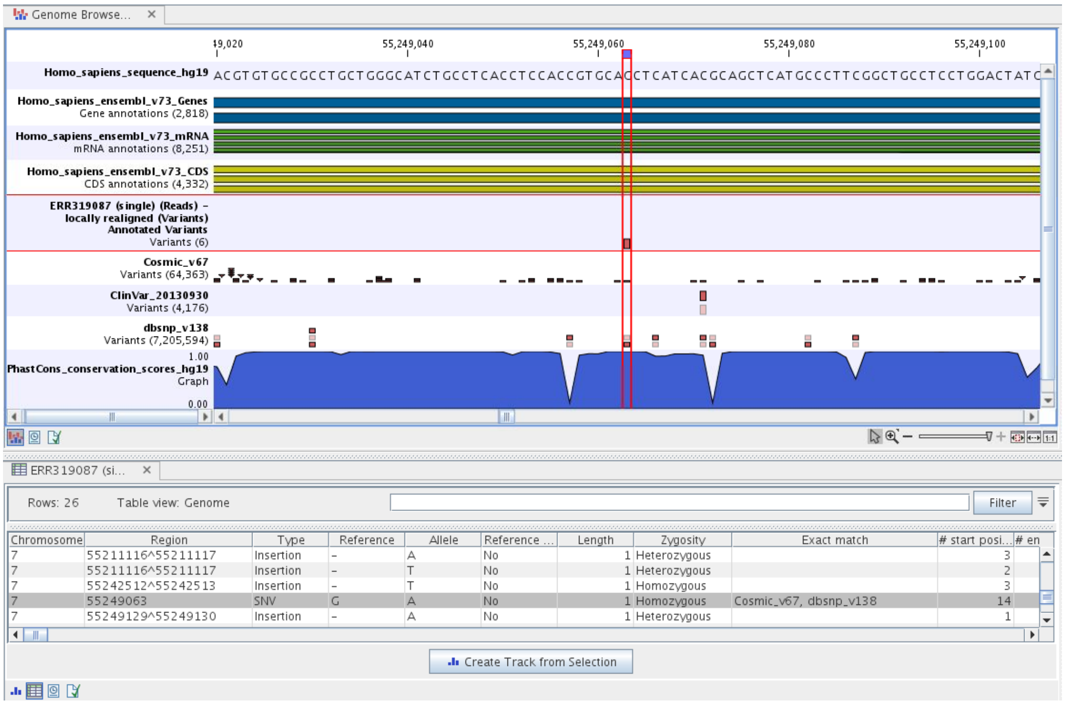

Genome Browse...
50,000,000
100,000,000
150,000,000
Homo_sapiens_sequen
79
Homo_sapiens_ensembl_v74_Genes
0
408
Homo_sapiens_ensembl_v74_mRNA
0
223
Homo_sapiens_ensembl_v74_CDS
0
11
ERR319085, ERR319085 Variants Annotated
0
107,413
dbsnp_v138 Variants (8,659,871)
0
716
Clinvar_20131203 Variants (7,801)
0
2,799
Cosmic_v67 Variants (126,891)
0
8,309
1000GENOMES_phase_1_EUR
0
9,517
1000GENOMES_phase_1_AMR
0
12,443
1000GENOMES_phase_1_AFR
0
1
PhastCons_conservation_scores_hg19
0
1:1

Figure 14.12: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list) containing individual tracks for all added annotations.*

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

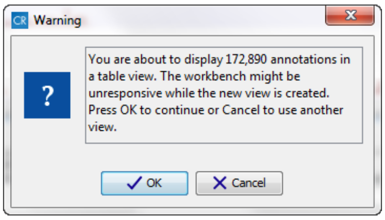
A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:


Genome Browse...
19,020
55,249,040
55,249,060
55,249,080
55,249,100
Homo_sapiens_sequence_hg19
ACGTGTGCCGCCTGCTGGGCATCTGCCTCACCTCCACCGTGCAGCTCATCACGCAGCTCATGCCCTTCGGCTGCCTCCTGGACTATC
Homo_sapiens_ensembl_v73_Genes
Gene annotations (2,818)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (8,251)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (4,332)
ERR319087 (single) (Reads) - locally realigned (Variants) Annotated Variants
Variants (6)
Cosmic_v67
Variants (64,363)
ClinVar_20130930
Variants (4,176)
dbsnp_v138
Variants (7,205,594)
1.00
PhastCons_conservation_scores_hg19
Graph
0.00
ERR319087 (si...
Rows: 26
Table view: Genome
Filter
Create Track from Selection

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# en
7	55211116^55211117	Insertion	-	A	No	1	Heterozygous		3	
7	55211116^55211117	Insertion	-	T	No	1	Heterozygous		2	
7	55242512^55242513	Insertion	-	T	No	1	Homozygous		3	
7	55249063	SNV	G	A	No	1	Homozygous	Cosmic_v67, dbsnp_v138	14	
7	55249129^55249130	Insertion	-	A	No	1	Heterozygous		1	

Figure 14.13: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list). The information is also available in table view. Click on the small table icon to open the table view. If you hold down the "Ctrl" key while clicking on the table icon, you will open a split view showing both the genome browser view and the table view.*



You are about to display 172,890 annotations in a table view. The workbench might be unresponsive while the new view is created. Press OK to continue or Cancel to use another view.

Figure 14.14: *Warning that appears when you work with tracks containing many annotations.*

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (http://clccancer.com/software/#downloads, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

14.4 Filter Somatic Variants (TAS)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same patient, you can use the "Filter Somatic Variants (TAS)" ready-to-use workflow to identify potential somatic variants. The purpose of this ready-to-use workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same patient is available.

The "Filter Somatic Variants (TAS)" ready-to-use workflow accepts variant tracks () (e.g. the output from the Identify Variants ready-to-use workflow) as input. Variants that are identical to the human reference sequence are first filtered away, then variants outside the targeted region are removed, and lastly, variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

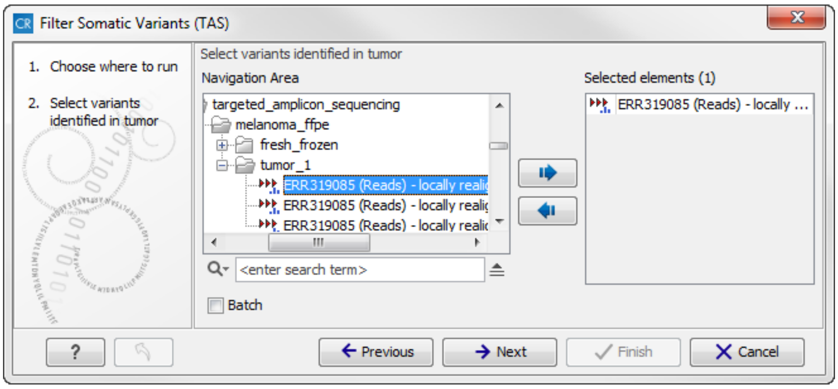
Please note that this tool will likely also remove inherited cancer variants that are present at a low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes, conservation scores and information from COSMIC (database with known variants in cancer), ClinVar (known variants with medical impact) and dbSNP (all known variants).

To run the **Filter Somatic Variants** tool, go to:

Toolbox | **Ready-to-Use Workflows** | **Targeted Amplicon Sequencing** () | **Filter Somatic Variants** ()

1. Double-click on the **Filter Somatic Variants** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the variant track you would like to use for filtering somatic variants. The panel in the left side of the wizard shows the kind of input that should be provided (figure 14.15). Select by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard.



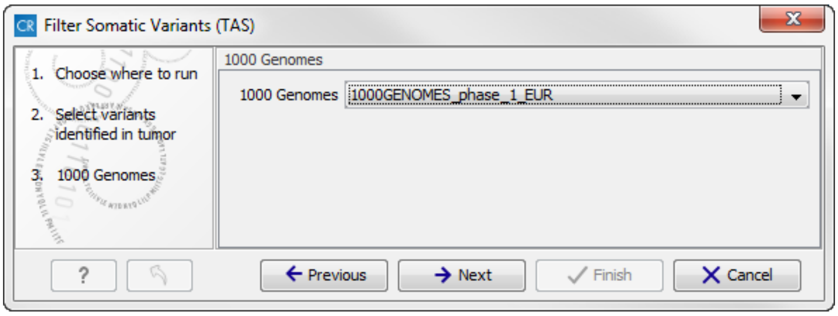
Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
Select variants identified in tumor
Navigation Area
targeted_amplicon_sequencing
melanoma_ffpe
fresh_frozen
tumor_1
ERR319085 (Reads) - locally realig
ERR319085 (Reads) - locally realig
ERR319085 (Reads) - locally realig
<enter search term>
Batch
Selected elements (1)
?
Previous
Next
Finish
Cancel

Selected elements (1): ERR319085 (Reads) - locally ...

Figure 14.15: *Select the variant track from which you would like to filter somatic variants.*

Click on the button labeled **Next**.

2. In the next step you will be asked to specify which of the 1000 Genomes populations that should be used for annotation (figure 14.16).


Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
? Previous Next Finish Cancel

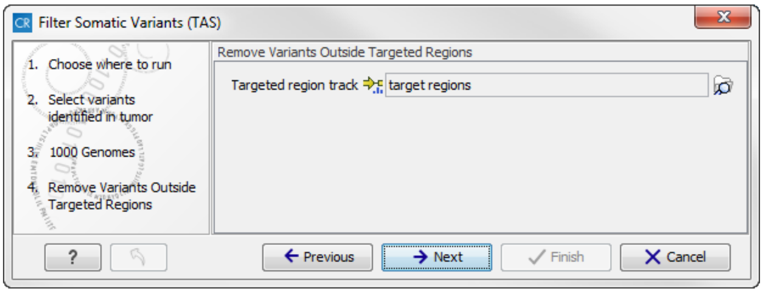
1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 14.16: *Specify which 1000 Genomes population to use for annotation.*

Click on the button labeled **Next**.

3. In this wizard step, you are asked to supply a track containing the targeted regions (figure 14.17). Select the track by clicking on the folder icon (🔍) in the wizard.


Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
? Previous Next Finish Cancel

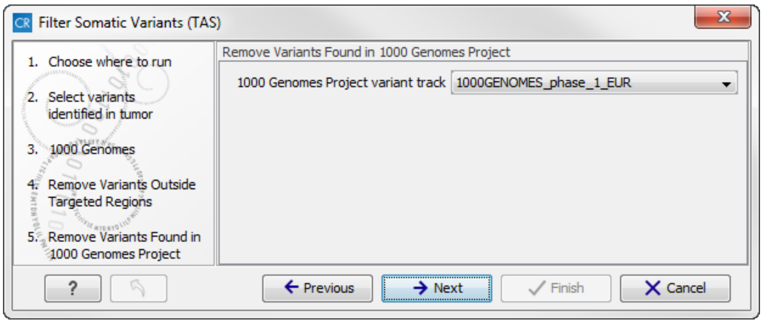
Remove Variants Outside Targeted Regions

Targeted region track: target regions

Figure 14.17: *Select your target regions track.*

Click on the button labeled **Next**.

4. The next wizard step will once again allow you to specify the 1000 Genomes population that should be used, this time for filtering out variants found in the 1000 Genomes project (figure 14.18).


Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
? Previous Next Finish Cancel

Remove Variants Found in 1000 Genomes Project

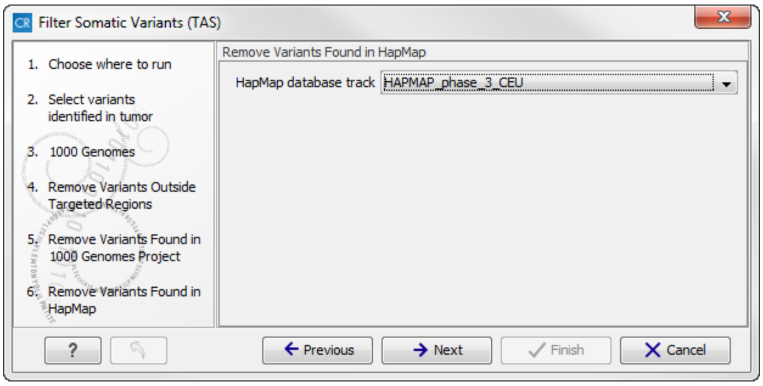
1000 Genomes Project variant track: 1000GENOMES_phase_1_EUR

Figure 14.18: *Specify which 1000 Genomes population to use for filtering out known variants.*

Click on the button labeled **Next**.

5. The next wizard step (figure 14.19) concerns removal of variants found in the HapMap database. Select the population you would like to use from the drop-down list. Please note that the populations available from the drop-down list can be specified with the **Data**

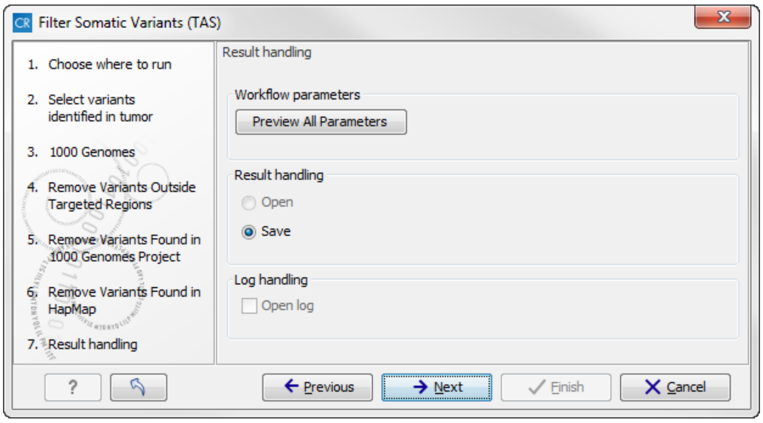
Management (⎘) function found in the top right corner of the Workbench (see section 10.1.4).


Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
6. Remove Variants Found in HapMap
Remove Variants Found in HapMap
?
Previous
Next
Finish
Cancel

HapMap database track: HAPMAP_phase_3_CEU

Figure 14.19: *Specify which HapMap population to use for filtering out known variants.*

6. Click on the button labeled **Next** to go to the last wizard step (shown in figure 14.20).


Filter Somatic Variants (TAS)
1. Choose where to run
2. Select variants identified in tumor
3. 1000 Genomes
4. Remove Variants Outside Targeted Regions
5. Remove Variants Found in 1000 Genomes Project
6. Remove Variants Found in HapMap
7. Result handling
Result handling
Workflow parameters
Preview All Parameters
Result handling
Log handling
?
Previous
Next
Finish
Cancel

○ Open
● Save
□ Open log

Figure 14.20: *Check the selected parametes by pressing "Preview All Parameters".*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes. Choose to save the results and click on the button labeled **Finish**.

Two types of output are generated:

1. **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Genome Browser View. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.

2. **Genome Browser View Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, COSMIC, 1000 Genomes, and the PhastCons conservation scores (see figure 14.21).

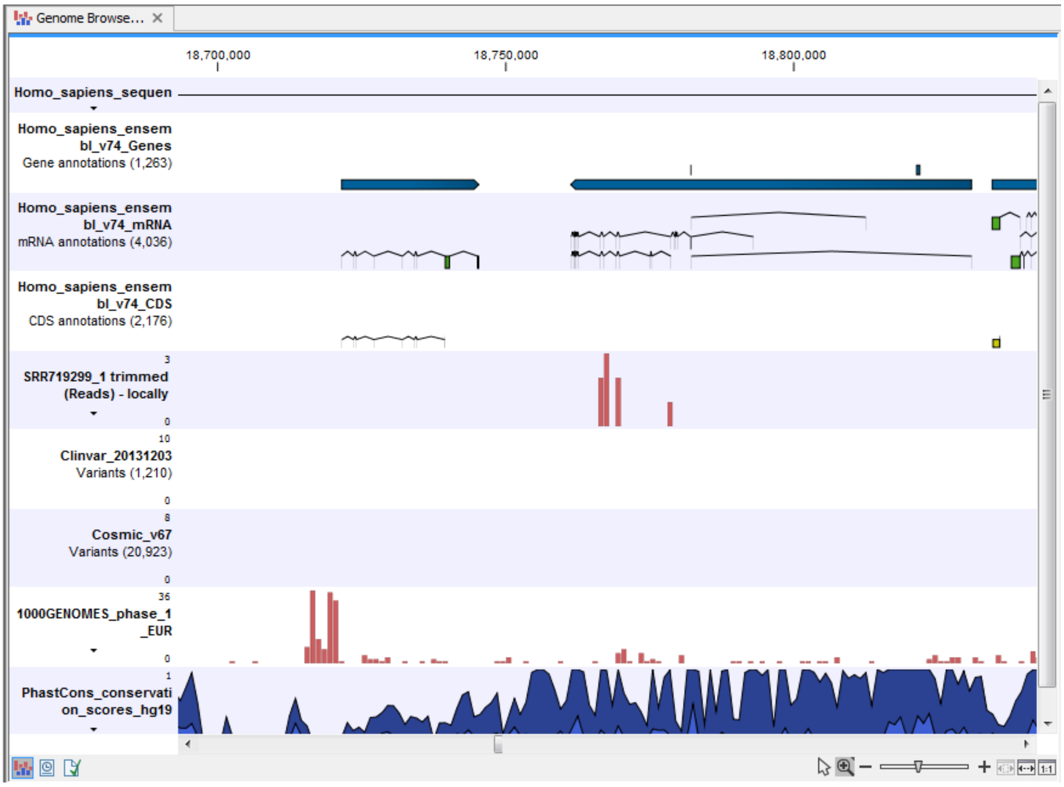


Figure 14.21: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

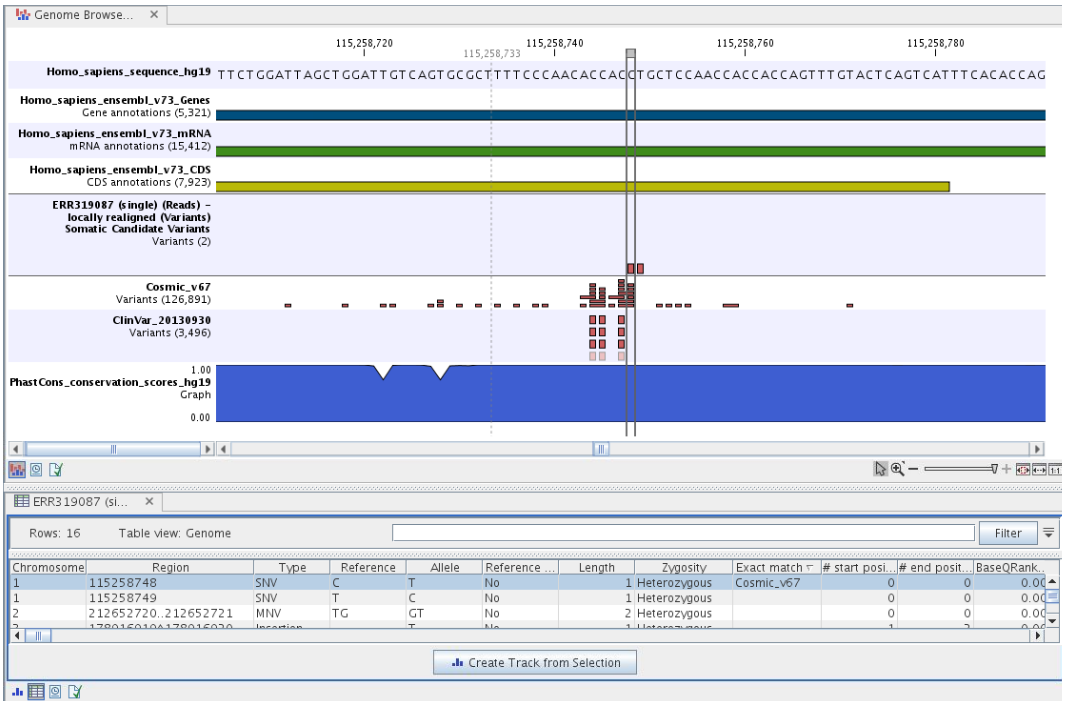
To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well. Mapped sequencing reads as well as other tracks can be easily added to this Genome Browser View. By double clicking on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see figure 14.22).

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )


Genome Browse...
115,258,720
115,258,733
115,258,740
115,258,760
115,258,780
Homo_sapiens_sequence_hg19
TTCTGGATTAGCTGGATTGTCAGTGCGCTTTTCCCAACACCACCTGCTCCAACCACCACCAGTTTGTACTCAGTCATTTCACACCAG
Homo_sapiens_ensembl_v73_Genes
Gene annotations (5,321)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (15,412)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (7,923)
ERR319087 (single) (Reads) – locally realigned (Variants) Somatic Candidate Variants
Variants (2)
Cosmic_v67
Variants (126,891)
ClinVar_20130930
Variants (3,496)
1.00
PhastCons_conservation_scores_hg19
Graph
0.00
ERR319087 (si...
Filter
Create Track from Selection

Rows: 16 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# end posit...	BaseQRank...
1	115258748	SNV	C	T	No	1	Heterozygous	Cosmic_v67	0	0	0.00
1	115258749	SNV	T	C	No	1	Heterozygous		0	0	0.00
2	212652720..212652721	MNV	TG	GT	No	2	Heterozygous		0	0	0.00

Figure 14.22: *The Genome Browser View showing the annotated somatic variants together with a range of other tracks.*

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (`http://clccancer.com/software/#downloads`, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

14.5 Identify Somatic Variants from Tumor Normal Pair (TAS)

The "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same patient.

When running the "Identify Somatic Variants from Tumor Normal Pair" the reads are mapped and the variants identified. An internal workflow removes germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads. Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from clinically relevant databases like COSMIC (known cancer associated variants) and ClinVar (variants with clinically relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

14.5.1 Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit is available from the vendor of the enrichment kit and sequencing machine. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get the file in either .bed or .gff format.

To import the file:

Go to the toolbar | **Import ( )** | **Tracks ( )**

14.5.2 How to run the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow

1. Go to the toolbox and double-click on the "Identify Somatic Variants from Tumor Normal Pair" ready-to-use workflow (figure 14.23).

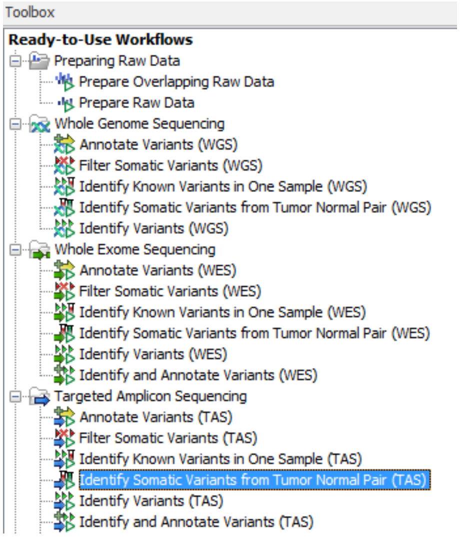

Toolbox
Ready-to-Use Workflows
Preparing Raw Data
Prepare Overlapping Raw Data
Prepare Raw Data
Whole Genome Sequencing
Annotate Variants (WGS)
Filter Somatic Variants (WGS)
Identify Known Variants in One Sample (WGS)
Identify Somatic Variants from Tumor Normal Pair (WGS)
Identify Variants (WGS)
Whole Exome Sequencing
Annotate Variants (WES)
Filter Somatic Variants (WES)
Identify Known Variants in One Sample (WES)
Identify Somatic Variants from Tumor Normal Pair (WES)
Identify Variants (WES)
Identify and Annotate Variants (WES)
Targeted Amplicon Sequencing
Annotate Variants (TAS)
Filter Somatic Variants (TAS)
Identify Known Variants in One Sample (TAS)
Identify Somatic Variants from Tumor Normal Pair (TAS)
Identify Variants (TAS)
Identify and Annotate Variants (TAS)

Figure 14.23: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard shown in figure 14.24 where you can select the tumor sample reads.

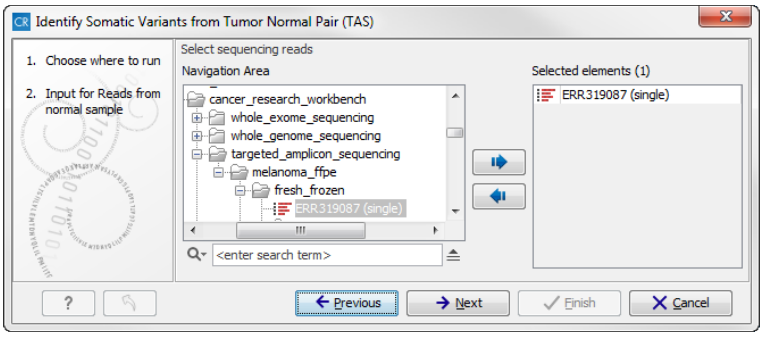
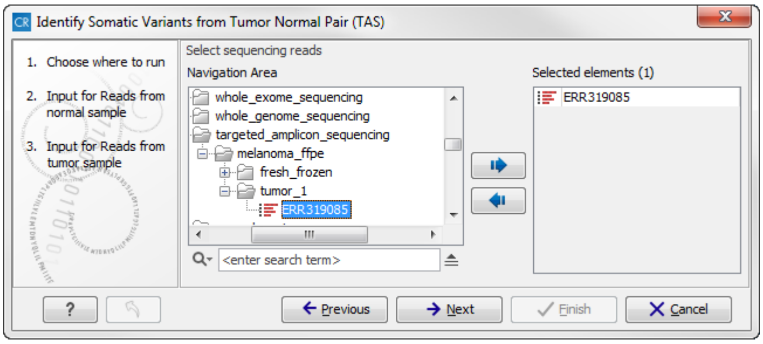


Figure 14.24: *Select the tumor sample reads.*

When you have selected the tumor sample reads click on the button labeled **Next**.

2. In the next wizard step (figure 14.25), please specify the normal sample reads.



Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample

Select sequencing reads

Navigation Area

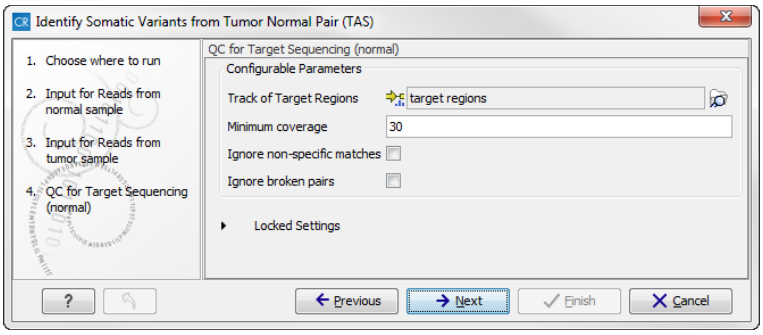
- whole_exome_sequencing
- whole_genome_sequencing
- targeted_amplicon_sequencing
  - melanoma_ffpe
    - fresh_frozen
    - tumor_1
      - ERR319085

Selected elements (1)

- ERR319085

Figure 14.25: *Select the normal sample reads.*

3. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.26).



Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)

QC for Target Sequencing (normal)

Configurable Parameters

Track of Target Regions: target regions

Minimum coverage: 30

Ignore non-specific matches ☐

Ignore broken pairs ☐

Locked Settings

Figure 14.26: *Specify the settings for the variant detection.*

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.27). In this wizard step you can select your target regions track.

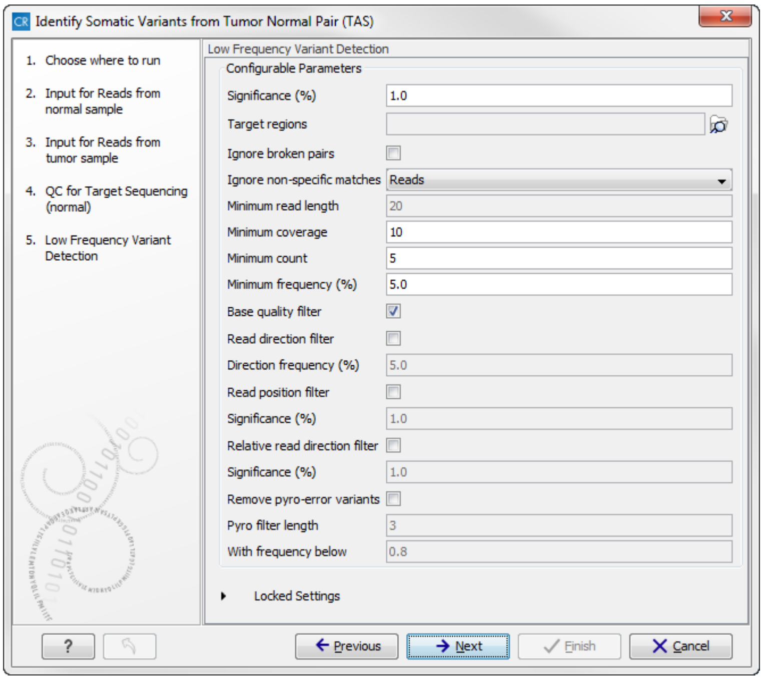
5. Click on the button labeled **Next** to specify the target regions track to be used in the "Remove Variants Outside Targeted Regions" step (figure 14.28). The targeted region track should be the same as the track you selected in the previous wizard step. Variants found outside the targeted regions will not be included in the output that is generated with the ready-to-use workflow.

   Click on the button labeled **Next**.

6. Click on the button labeled **Next** to go to the step where you can adjust the settings for removal of germline variants (figure 14.29)..

7. Click on the button labeled **Next** and once again select the target region track (the same track as you have already selected in previous wizard steps). This time you specify the track to be used for quality control of the targeted sequencing as this tool reports the performance (enrichment and specificity) of a targeted re-sequencing experiment(figure 14.30).

   In the next wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters** (figure 14.31).



Identify Somatic Variants from Tumor Normal Pair (TAS)

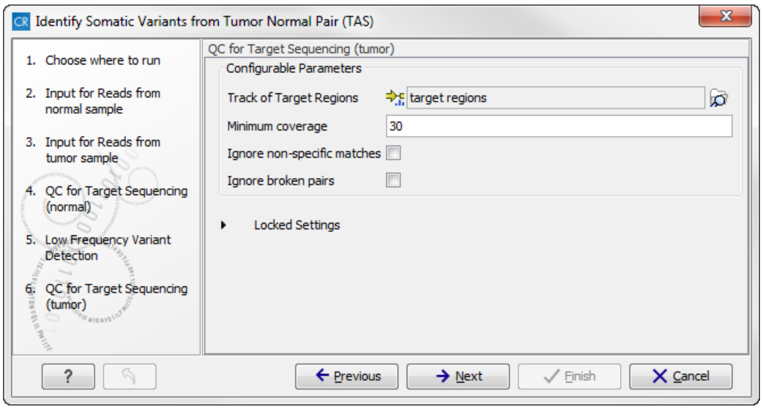
1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)
5. Low Frequency Variant Detection

Low Frequency Variant Detection

Configurable Parameters

Significance (%)	1.0
Target regions	
Ignore broken pairs	□
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	5
Minimum frequency (%)	5.0
Base quality filter	☑
Read direction filter	□
Direction frequency (%)	5.0
Read position filter	□
Significance (%)	1.0
Relative read direction filter	□
Significance (%)	1.0
Remove pyro-error variants	□
Pyro filter length	3
With frequency below	0.8

Figure 14.27: *Select your target region track.*



Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)
5. Low Frequency Variant Detection
6. QC for Target Sequencing (tumor)

QC for Target Sequencing (tumor)

Configurable Parameters

Track of Target Regions	target regions
Minimum coverage	30
Ignore non-specific matches	□
Ignore broken pairs	□

Figure 14.28: *Select your target region track.*

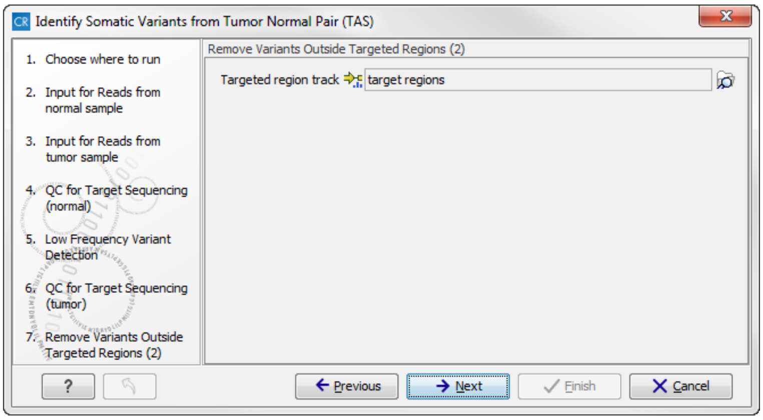
In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

8. Click on the button labeled **OK** to go back to the previous wizard step and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Eight different outputs are generated:

1. **Read Mapping Normal** (☰) The mapped sequencing reads for the normal sample. The



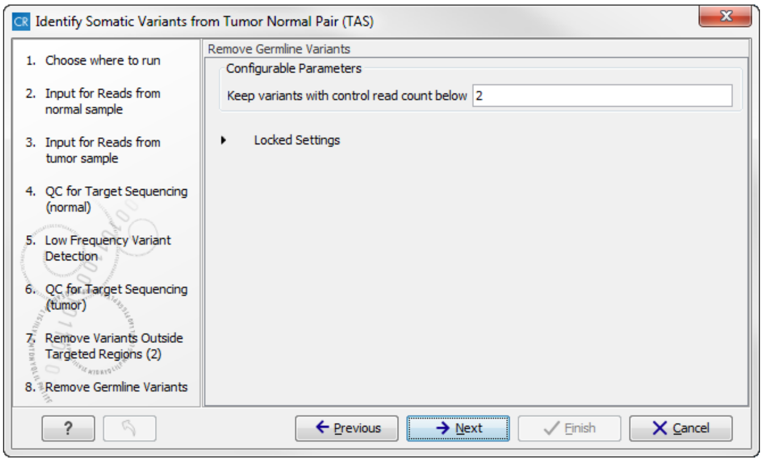
Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)
5. Low Frequency Variant Detection
6. QC for Target Sequencing (tumor)
7. Remove Variants Outside Targeted Regions (2)

Remove Variants Outside Targeted Regions (2)

Targeted region track: target regions

Figure 14.29: *Specify setting for removal of germline variants.*



Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)
5. Low Frequency Variant Detection
6. QC for Target Sequencing (tumor)
7. Remove Variants Outside Targeted Regions (2)
8. Remove Germline Variants

Remove Germline Variants

Configurable Parameters

Keep variants with control read count below: 2

Locked Settings

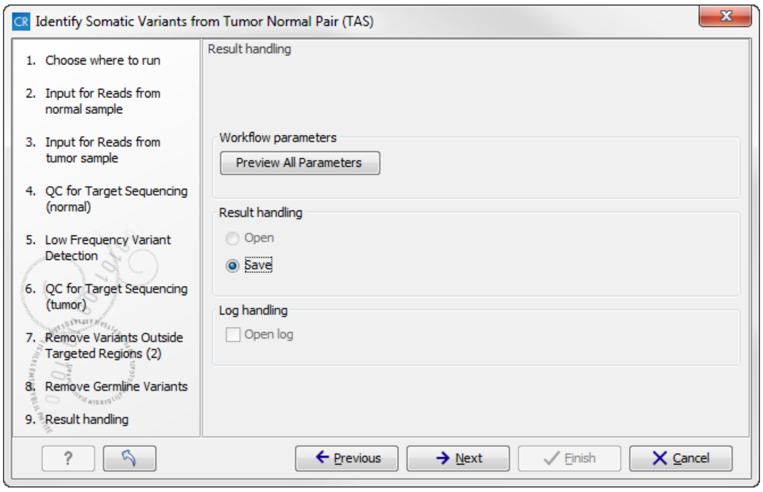
Figure 14.30: *Select target region track.*

reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html

2. **Read Mapping Tumor** () The mapped sequencing reads for the tumor sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@View_settings_in_Side_Panel.html.

3. **Target Region Coverage Report Normal** () The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the normal sample.

4. **Target Region Coverage Tumor** () A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.



Identify Somatic Variants from Tumor Normal Pair (TAS)

1. Choose where to run
2. Input for Reads from normal sample
3. Input for Reads from tumor sample
4. QC for Target Sequencing (normal)
5. Low Frequency Variant Detection
6. QC for Target Sequencing (tumor)
7. Remove Variants Outside Targeted Regions (2)
8. Remove Germline Variants
9. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 14.31: *Check the parameters and save the results.*

5. **Target Region Coverage Report Tumor** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.

6. **Variants** ( ) A variant track holding the identified variants that are found in the targeted resions. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

7. **Annotated Somatic Variants** ( ) A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Genome Browser view a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

8. **Genome Browser View Tumor Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar and COSMIC databases, and finally a track showing the conservation score (see figure 14.32).

14.6 Identify Known Variants in One Sample (TAS)

The "Identify Known Variants in One Sample" ready-to-use workflow is a combined data analysis and interpretation ready-to-use workflow.

It should be used to identify known variants, specified by the user (e.g. known breast cancer associated variants), for their presence or absence in a sample.

Please note that the ready-to-use workflow will not identify new variants.

The Identify Known Variants in One Sample ready-to-use workflow runs an internal workflow that maps the sequencing reads to the human genome sequence and does a local realignment of the

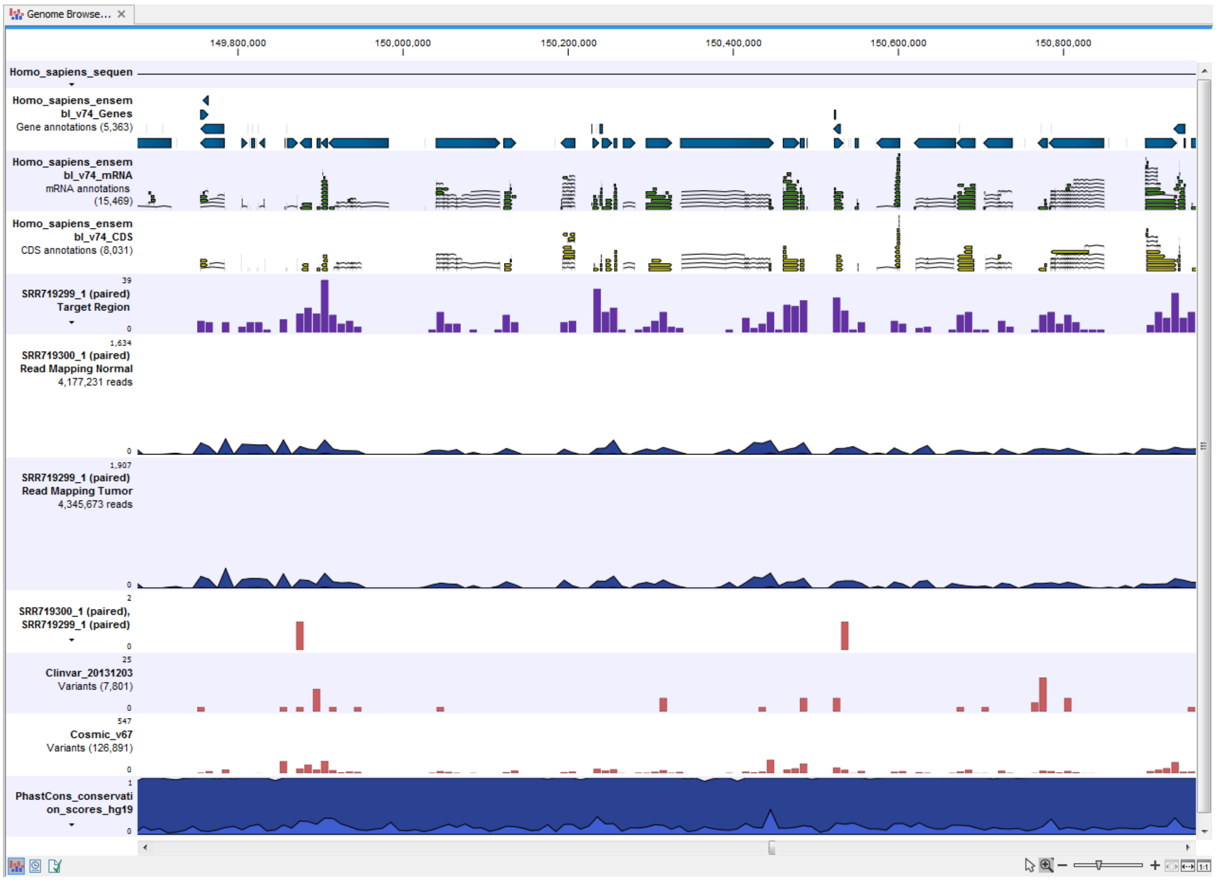

Genome Browse...
149,800,000
150,000,000
150,200,000
150,400,000
150,600,000
150,800,000
Homo_sapiens_sequen
Homo_sapiens_ensembl_v74_Genes
Gene annotations (5,363)
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (15,469)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (8,031)
39
SRR719299_1 (paired) Target Region
0
1,634
SRR719300_1 (paired) Read Mapping Normal
4,177,231 reads
0
1,907
SRR719299_1 (paired) Read Mapping Tumor
4,345,673 reads
0
2
SRR719300_1 (paired), SRR719299_1 (paired)
0
25
Clinvar_20131203
Variants (7,801)
0
547
Cosmic_v67
Variants (126,891)
0
1
PhastCons_conservation_scores_hg19
0

Figure 14.32: *The Genome Browser View presents all the different data tracks together and makes it easy to compare different tracks.*

mapped reads to improve the following variant detection. Next, specified variants by the user are identified in the read mapping. At the end, information present on the known variants before, are added to the results.

14.6.1 Import your known variants

To make an import into the Cancer Research Workbench, you should have your variants in GVF or VCF 4.1 format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

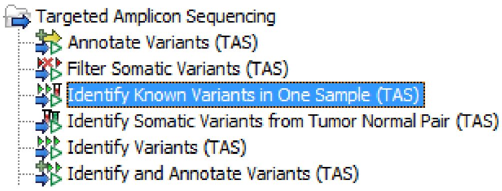
14.6.2 Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit will be provided by the vendor. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get it in either .bed or .gff format.

Please use the Tracks import as part of the Import tool in the toolbar to import your file into the Cancer Research Workbench.

14.6.3 How to run the "Identify Known Variants in One Sample" ready-to-use workflow

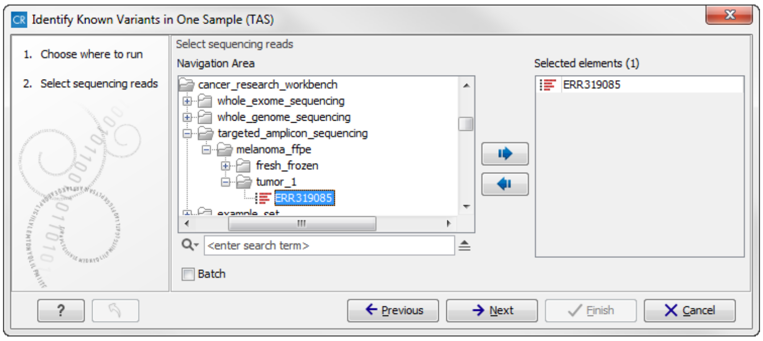
1. Go to the toolbox and double-click on the "Identify Known Variants from One Sample" ready-to-use workflow (figure 14.33).



Targeted Amplicon Sequencing
- Annotate Variants (TAS)
- Filter Somatic Variants (TAS)
- Identify Known Variants in One Sample (TAS)
- Identify Somatic Variants from Tumor Normal Pair (TAS)
- Identify Variants (TAS)
- Identify and Annotate Variants (TAS)

Figure 14.33: *The ready-to-use workflows are found in the toolbox.*

This will open the wizard step shown in figure 14.34 where you can select the reads of the sample, which should be tested for presence or absence of your known variants.



Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

Navigation Area
- cancer_research_workbench
  - whole_exome_sequencing
  - whole_genome_sequencing
  - targeted_amplicon_sequencing
    - melanoma_ffpe
      - fresh_frozen
      - tumor_1
        - ERR319085
  - example_set

Selected elements (1)
- ERR319085

Figure 14.34: *Select the sequencing reads from the sample you would like to test for your known variants.*

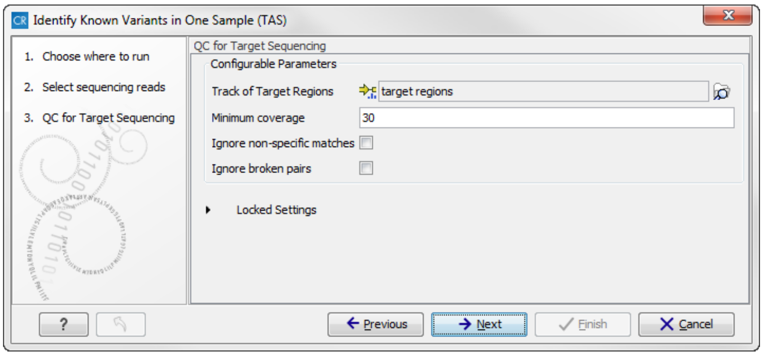
Please select all sequencing reads from your sample. If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 14.34) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

When you have selected the sample(s) you wish to analyze, click on the button labeled **Next**.

2. In the next wizard step you can select your target regions track and specify the minimum coverage to be used when checking the quality of the targeted sequencing. The minimum coverage will be used to provide the length of each target region that has at least this coverage. You can also specify whether or not to ignore non-specific matches and broken pairs. When these are applied, reads that are non-specifically mapped or belong to broken pairs will be ignored (figure 14.35).

3. Click on the button labeled **Next** and in specify the track with the known variants that should be identified in your sample (figure 14.36). Furthermore, in this wizard step you can specify the minimum read coverage for the position of the variant that should be identified. If the coverage at the position of the variant is below this, the result will show this.

   The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or


Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing

QC for Target Sequencing

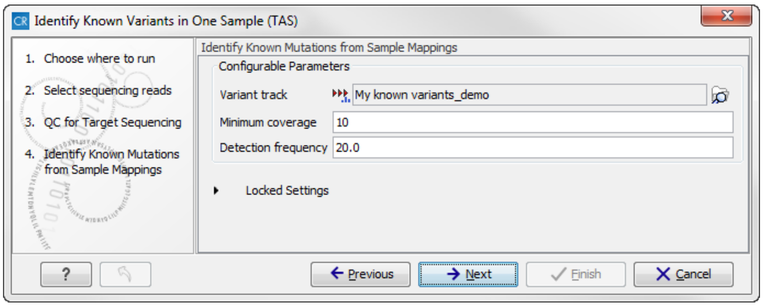
Configurable Parameters

Track of Target Regions: target regions
Minimum coverage: 30
Ignore non-specific matches: □
Ignore broken pairs: □

Locked Settings

Figure 14.35: *Select your target regions track and specify the parameters to be used for checking the quality of the targeted sequecing.*

not (observed frequency <= specified frequency). Moreover, it will determine if a variant should be labeled as heterozygous (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygous (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).


Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings

Identify Known Mutations from Sample Mappings

Configurable Parameters

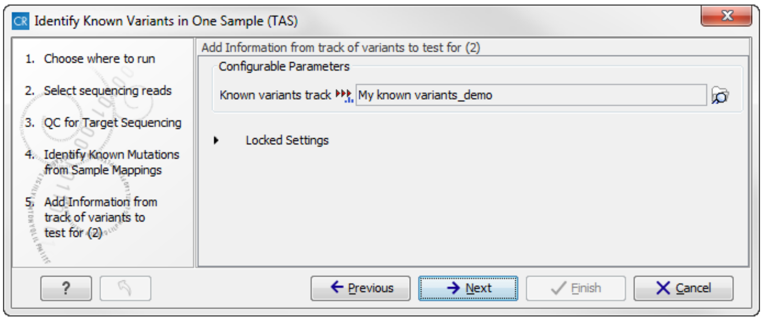
Variant track: My known variants_demo
Minimum coverage: 10
Detection frequency: 20.0

Locked Settings

Figure 14.36: *Specify the track with the known variants that should be identified.*

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.37). In this and the next dialog, you will be asked about which of the annotations/informations added to variants should be included in the results.

   Please specify your track with known variants.


Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)

Add Information from track of variants to test for (2)

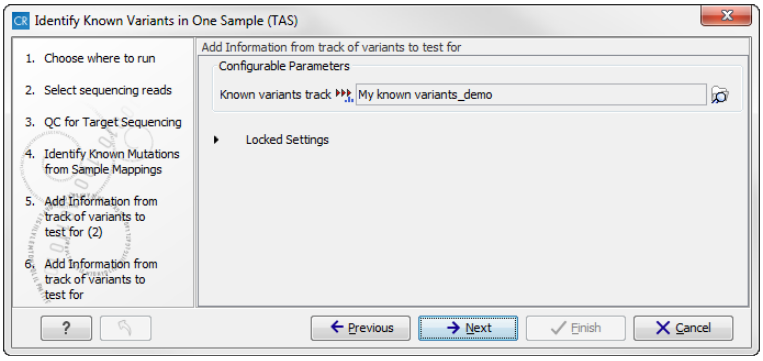
Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

Figure 14.37: *Please select the track with your known variants again. Annotations/Informations from this track will be added to the overview mutation track.*

5. Click on the button labeled **Next** and once again select the same track with known variants (figure 14.38).


Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)
6. Add Information from track of variants to test for

Add Information from track of variants to test for

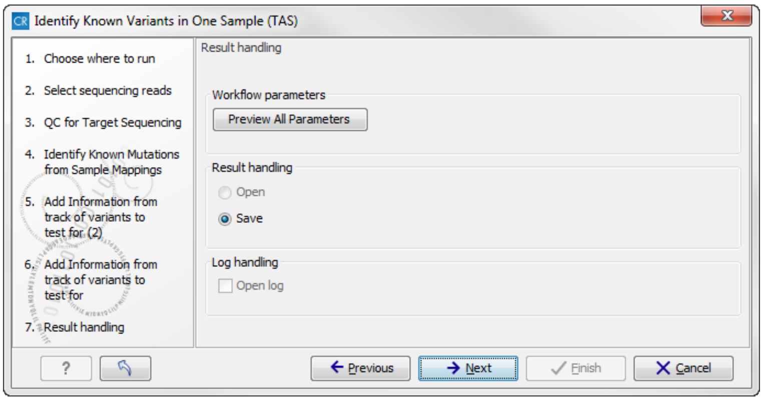
Configurable Parameters

Known variants track: My known variants_demo

Locked Settings

Figure 14.38: *Once again select the track with known variants. This time the track is used to add information to the detailed mutation track.*

6. Click on the button labeled **Next** to go to the last wizard step (figure 14.39).


Identify Known Variants in One Sample (TAS)

1. Choose where to run
2. Select sequencing reads
3. QC for Target Sequencing
4. Identify Known Mutations from Sample Mappings
5. Add Information from track of variants to test for (2)
6. Add Information from track of variants to test for
7. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 14.39: *Check the settings and save your results.*

In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point. At the bottom of this wizard there are two buttons regarding export functions; one button allows specification of the export format, and the other button (the one labeled "Export Parameters") allows specification of the export destination. When selecting an export location, you will export the analysis parameter settings that were specified for this specific experiment.

7. Click on the button labeled **OK** to go back to the previous dialog box and choose **Save**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

14.6.4 Output from the Identify Known Variants in One Sample

The "Identify Known Variants in One Sample" tool produces seven different output types:

1. **Read Mapping** (☰) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found

here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **Target Regions Coverage** ( ) A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.

3. **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways show e.g. the number, length, and coverage of the target regions and provides information about the read count per GC%.

4. **Overview Variants Detected** ( ) Annotation track showing the known variants. The table view provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads.

5. **Variants Detected in Detail** ( ) Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. The difference between the two tables is that the "Variants Detected in Detail" table includes detailed information about the most frequent alternative allele (MFAA).

6. **Genome Browser View Identify Known Variants** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

7. **Log** ( ) A log of the workflow execution.

It is a good idea to start looking at the Target Regions Coverage Report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, we also recommend that you check that the majority of the reads are mapping to the targeted region.

When you have inspected the target regions coverage report you can open the Genome Browser View Identify Known Variants file (see 14.40).

The Genome Browser View includes an overview track of the known variants and a detailed result track presented in the context of the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, and clinically relevant variants in the COSMIC databases.

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

The difference between the overview variant track and the detailed variant track is the annotations added to the variants.

By double clicking on one of the annotated variant tracks in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see 14.41).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.

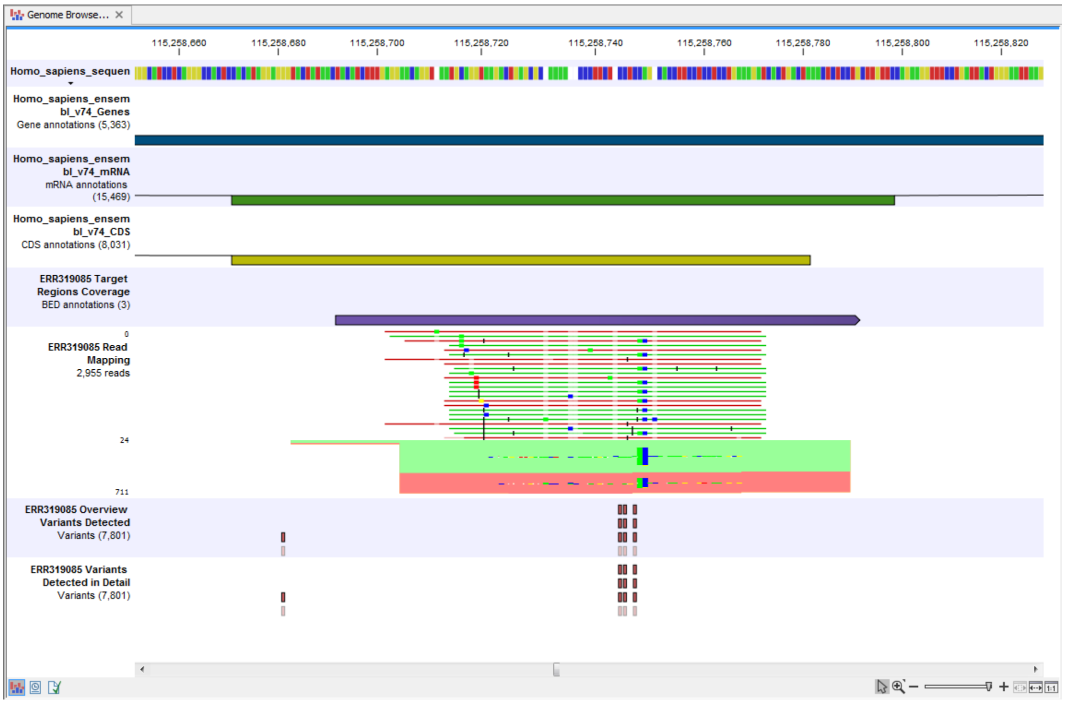


Figure 14.40: *Genome Browser View that allows inspection of the identified variants in the context of the human genome and external databases.*

14.7 Identify and Annotate Variants (TAS)

The "Identify and Annotate Variants" tool should be used to identify and annotate variants in one sample. The tool consists of a workflow that is a combination of the "Identify Variants" and the "Annotate Variants" workflows.

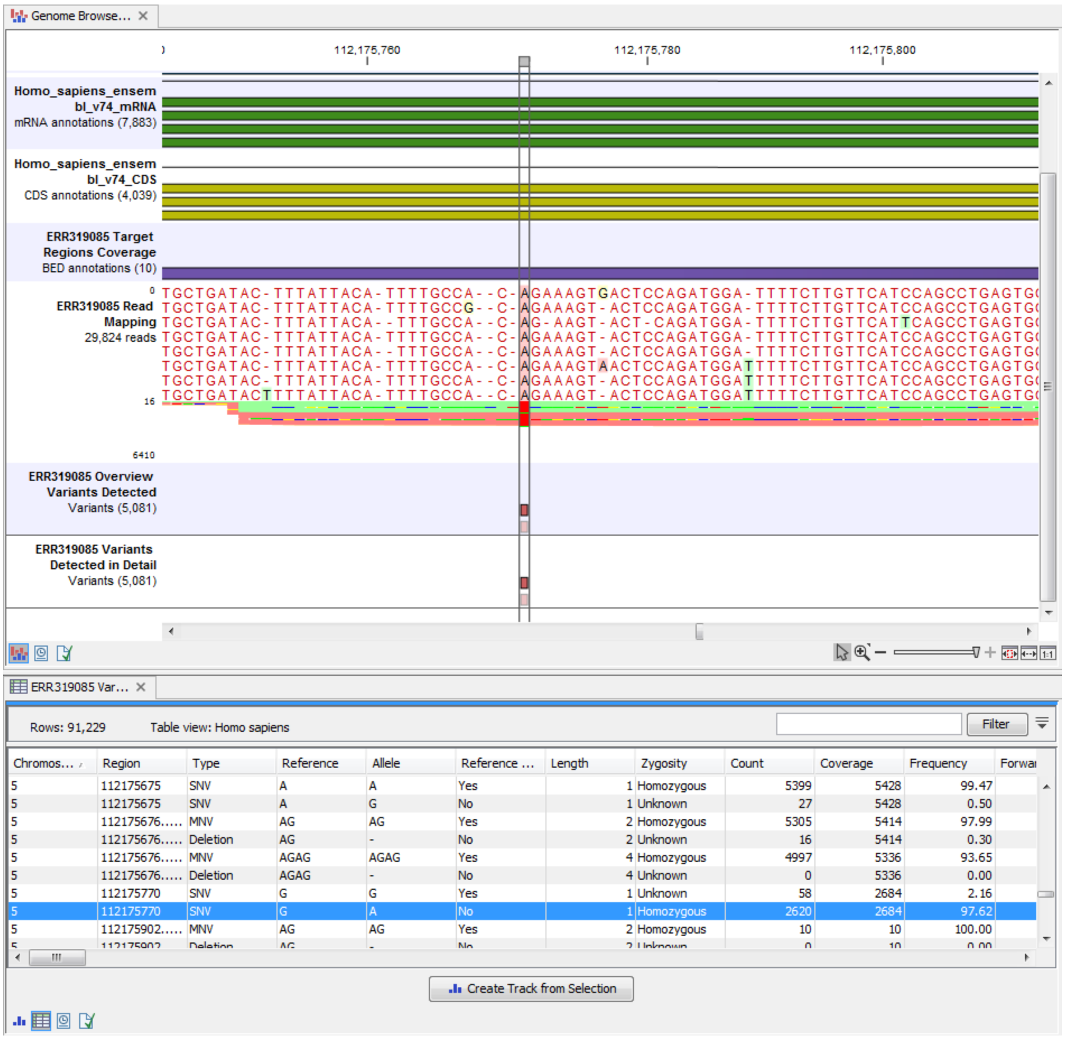
The tool runs an internal workflow, which starts with mapping the sequencing reads to the human reference sequence. Then it runs a local realignment to improve the variant detection, which is run afterwards. After the variants have been detected, they are annotated with gene names, amino acid changes, conservation scores, information from clinically relevant variants present in the COSMIC and ClinVar database, and information from common variants present in the common dbSNP, HapMap, and 1000 Genomes database. Furthermore, a detailed mapping report or a targeted region report (whole exome and targeted amplicon analysis) is created to inspect the overall coverage and mapping specificity.

Import your targeted regions

A file with the genomic regions targeted by the amplicon or hybridization kit is available from the vendor of the enrichment kit and sequencing machine. To obtain this file you will have to get in contact with the vendor and ask them to send this target regions file to you. You will get the file in either .bed or .gff format.

To import the file:

Go to the toolbar | **Import** () | **Tracks** ()



Rows: 91,229 Table view: Homo sapiens

Chromos...	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Forwa
5	112175675	SNV	A	A	Yes	1	Homozygous	5399	5428	99.47	
5	112175675	SNV	A	G	No	1	Unknown	27	5428	0.50	
5	112175676.....	MNV	AG	AG	Yes	2	Homozygous	5305	5414	97.99	
5	112175676.....	Deletion	AG	-	No	2	Unknown	16	5414	0.30	
5	112175676.....	MNV	AGAG	AGAG	Yes	4	Homozygous	4997	5336	93.65	
5	112175676.....	Deletion	AGAG	-	No	4	Unknown	0	5336	0.00	
5	112175770	SNV	G	G	Yes	1	Unknown	58	2684	2.16	
5	112175770	SNV	G	A	No	1	Homozygous	2620	2684	97.62	
5	112175902.....	MNV	AG	AG	Yes	2	Homozygous	10	10	100.00	
5	112175902	Deletion	AG	-	No	2	Unknown	0	10	0.00	

Figure 14.41: *Genome Browser View with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

How to run the "Identify and Annotate Variants" ready-to-use workflow

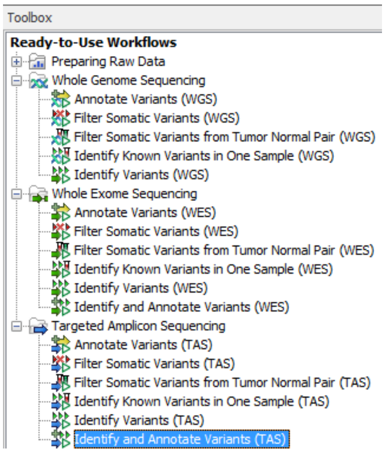
1. Go to the toolbox and double-click on the "Identify and Annotate Variants" ready-to-use workflow (figure 14.42).

   This will open the wizard shown in figure 14.43 where you can select the sequencing reads from the sample that should be analyzed.

   If several samples should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" (tick "Batch" at the bottom of the wizard as shown in figure 14.43) and select the **folder** that holds the data you wish to analyse. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode.

   When you have selected the sample(s) you wish to prepare, click on the button labeled **Next**.

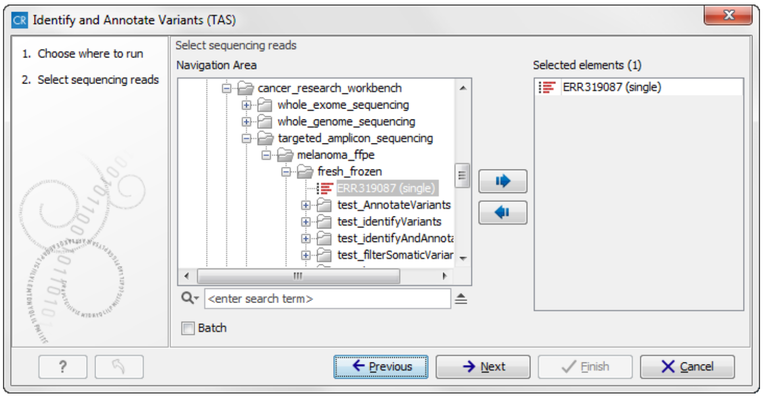
2. In the next wizard step (figure 14.44) you can select the population from the 1000 Genomes project that you would like to use for annotation.



Toolbox

Ready-to-Use Workflows
- Preparing Raw Data
- Whole Genome Sequencing
  - Annotate Variants (WGS)
  - Filter Somatic Variants (WGS)
  - Filter Somatic Variants from Tumor Normal Pair (WGS)
  - Identify Known Variants in One Sample (WGS)
  - Identify Variants (WGS)
- Whole Exome Sequencing
  - Annotate Variants (WES)
  - Filter Somatic Variants (WES)
  - Filter Somatic Variants from Tumor Normal Pair (WES)
  - Identify Known Variants in One Sample (WES)
  - Identify Variants (WES)
  - Identify and Annotate Variants (WES)
- Targeted Amplicon Sequencing
  - Annotate Variants (TAS)
  - Filter Somatic Variants (TAS)
  - Filter Somatic Variants from Tumor Normal Pair (TAS)
  - Identify Known Variants in One Sample (TAS)
  - Identify Variants (TAS)
  - Identify and Annotate Variants (TAS)

Figure 14.42: *The ready-to-use workflows are found in the toolbox.*



Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

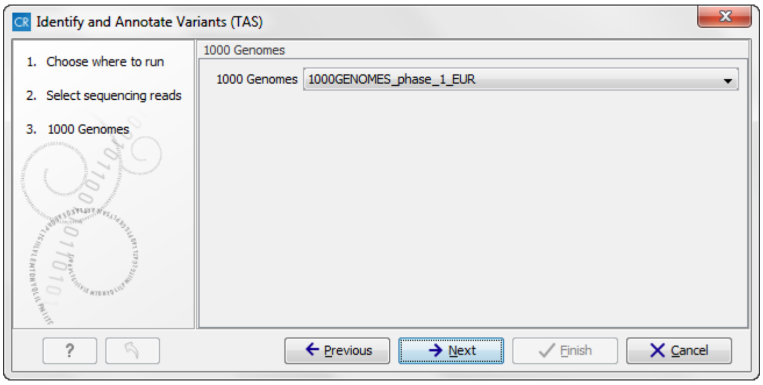
Navigation Area
- cancer_research_workbench
  - whole_exome_sequencing
  - whole_genome_sequencing
  - targeted_amplicon_sequencing
    - melanoma_ffpe
      - fresh_frozen
        - ERR319087 (single)
        - test_AnnotateVariants
        - test_identifyVariants
        - test_identifyAndAnnota
        - test_filterSomaticVariar

<enter search term>

□ Batch

Selected elements (1)
- ERR319087 (single)

Figure 14.43: *Please select all sequencing reads from the sample to be analyzed.*



Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes

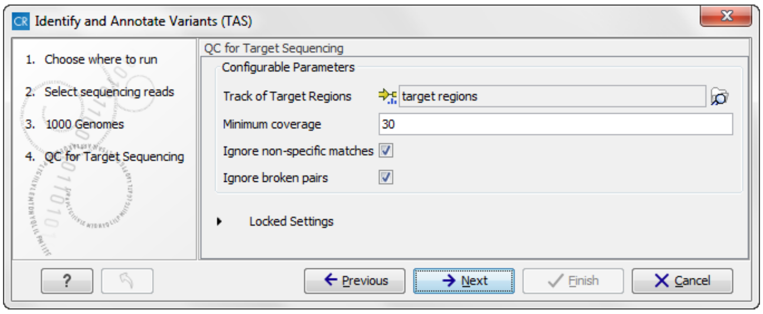
1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 14.44: *Select the population from the 1000 Genomes project that you would like to use for annotation.*

3. In the next wizard (figure 14.45) you can select the target region track and specify the minimum read coverage that should be present in the targeted regions.

4. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.46). In this dialog, you have to specify the parameters for the variant detection. For a description of the different parameters that can be adjusted in the variant


Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing

QC for Target Sequencing

Configurable Parameters

Track of Target Regions: target regions

Minimum coverage: 30

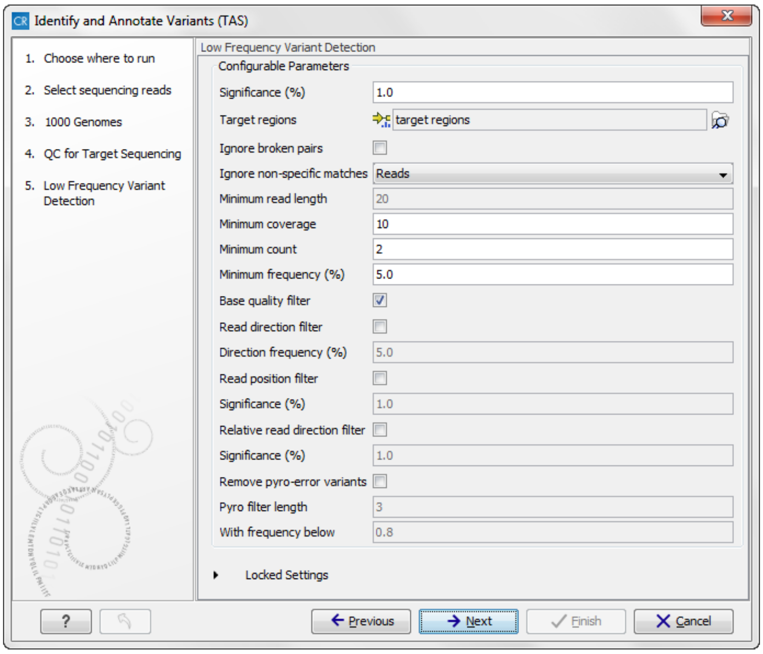
Ignore non-specific matches: ☑

Ignore broken pairs: ☑

Locked Settings

Figure 14.45: *Select the track with targeted regions from your experiment.*

detection step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html). As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the description of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Filters.html). If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the ready-to-use workflow.


Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection

Low Frequency Variant Detection

Configurable Parameters

Significance (%): 1.0

Target regions: target regions

Ignore broken pairs: □

Ignore non-specific matches: Reads

Minimum read length: 20

Minimum coverage: 10

Minimum count: 2

Minimum frequency (%): 5.0

Base quality filter: ☑

Read direction filter: □

Direction frequency (%): 5.0

Read position filter: □

Significance (%): 1.0

Relative read direction filter: □

Significance (%): 1.0

Remove pyro-error variants: □

Pyro filter length: 3

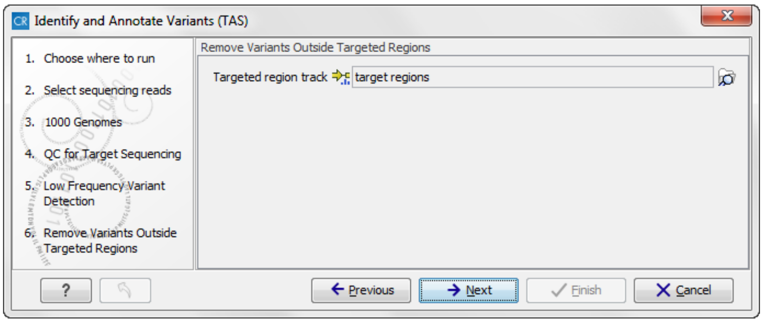
With frequency below: 0.8

Locked Settings

Figure 14.46: *Specify the parameters for variant calling.*

5. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.47). In this dialog you can specify the target regions track. The variants found outside the targeted region will be removed at this step in the workflow.

6. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.48). Once again, select the relevant population from the 1000 Genomes project. This will add information from the 1000 Genomes project to your variants.



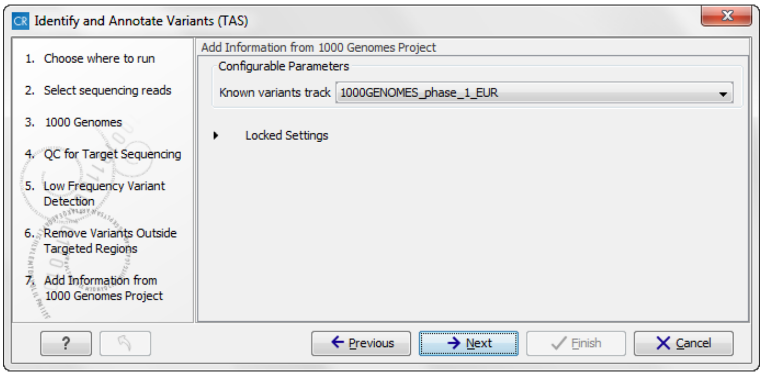
Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions

Remove Variants Outside Targeted Regions

Targeted region track: target regions

Figure 14.47: *In this wizard step you can specify the target regions track. Variants found outside these regions will be removed.*



Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project

Add Information from 1000 Genomes Project

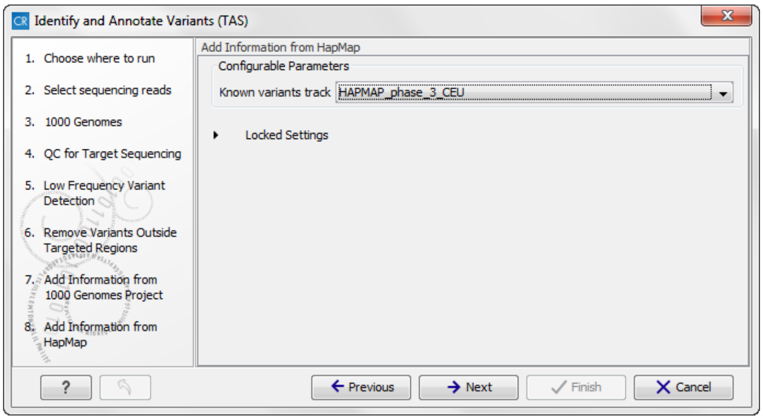
Configurable Parameters

Known variants track: 1000GENOMES_phase_1_EUR

Locked Settings

Figure 14.48: *Select the relevant population from the 1000 Genomes project. This will add information from the 1000 Genomes project to your variants.*

7. Click on the button labeled **Next**, which will take you to the next wizard step (figure 14.49). At this step you can select a population from the HapMap database. This will add information from the Hapmap database to your variants.



Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project
8. Add Information from HapMap

Add Information from HapMap

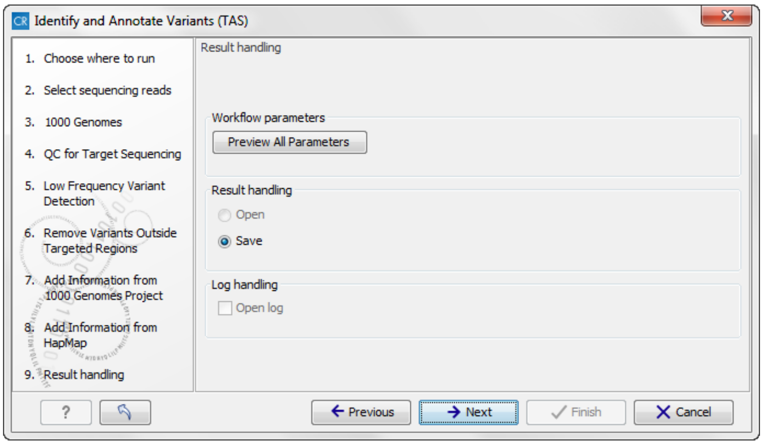
Configurable Parameters

Known variants track: HAPMAP_phase_3_CEU

Locked Settings

Figure 14.49: *Select a population from the HapMap database. This will add information from the Hapmap database to your variants.*

8. In this wizard step (figure 14.50) you get the chance to check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point.



Identify and Annotate Variants (TAS)

1. Choose where to run
2. Select sequencing reads
3. 1000 Genomes
4. QC for Target Sequencing
5. Low Frequency Variant Detection
6. Remove Variants Outside Targeted Regions
7. Add Information from 1000 Genomes Project
8. Add Information from HapMap
9. Result handling

Result handling

Workflow parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 14.50: *Check the settings and save your results.*

9. Choose to **Save** your results and press **Finish**.

   Note! If you choose to open the results, the results will not be saved automatically. You can always save the results at a later point.

Output from the Identify and Annotate Variants workflow

The "Identify and Annotate Variants" tool produces several outputs.

Please do not delete any of the produced files alone as some of them are linked to other outputs. Please always delete all of them at the same time.

A good place to start is to take a look at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Furthermore, please check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of the reads are mapping to the targeted region.

Next, open the Genome Browser View file (see figure 14.51).

The Genome Browser View includes a track of the identified annotated variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, clinically relevant variants in the COSMIC and ClinVar database as well as common variants in common dbSNP, HapMap, and 1000 Genomes databases.

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well.

By double-clicking on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see 14.52).

The added information will help you to identify candidate variants for further research. For example can known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) easily be seen.

Not identified variants in COSMIC and ClinVar, can for example be prioritized based on amino acid changes (do they cause any changes on the amino acid level?). A high conservation level on the position of the variant between many vertebrates or mammals can also be a hint that this

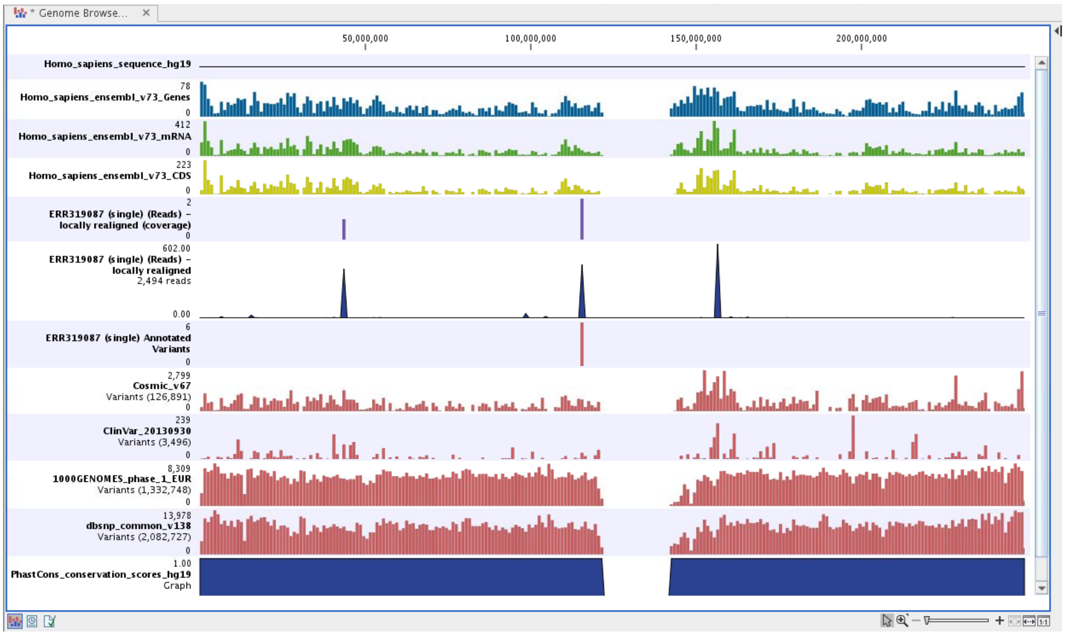
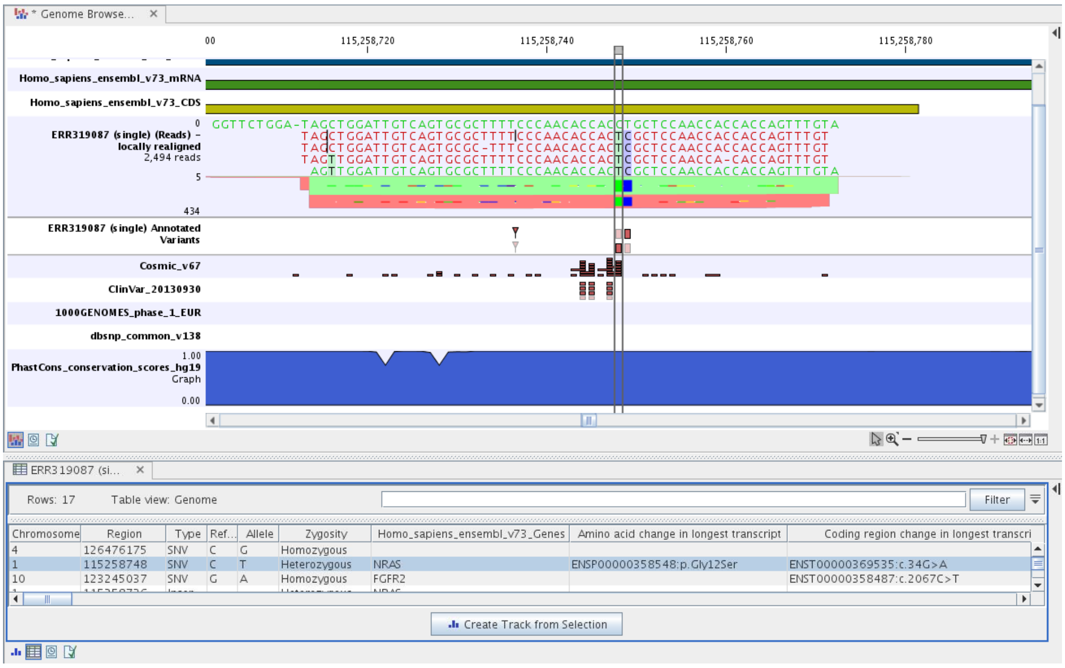

Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
Homo_sapiens_ensembl_v73_Genes
Homo_sapiens_ensembl_v73_mRNA
Homo_sapiens_ensembl_v73_CDS
ERR319087 (single) (Reads) – locally realigned (coverage)
ERR319087 (single) (Reads) – locally realigned 2,494 reads
ERR319087 (single) Annotated Variants
Cosmic_v67 Variants (126,891)
ClinVar_20130930 Variants (3,496)
1000GENOMES_phase_1_EUR Variants (1,332,748)
dbsnp_common_v138 Variants (2,082,727)
PhastCons_conservation_scores_hg19 Graph

Figure 14.51: *Genome Browser View to inspect identified variants in the context of the human genome and external databases.*


Genome Browse...
115,258,720
115,258,740
115,258,760
115,258,780
Homo_sapiens_ensembl_v73_mRNA
Homo_sapiens_ensembl_v73_CDS
ERR319087 (single) (Reads) – locally realigned 2,494 reads
ERR319087 (single) Annotated Variants
Cosmic_v67
ClinVar_20130930
1000GENOMES_phase_1_EUR
dbsnp_common_v138
PhastCons_conservation_scores_hg19 Graph
ERR319087 (si...
Filter
Create Track from Selection

Rows: 17 Table view: Genome

Chromosome	Region	Type	Ref...	Allele	Zygosity	Homo_sapiens_ensembl_v73_Genes	Amino acid change in longest transcript	Coding region change in longest transcri
4	126476175	SNV	C	G	Homozygous			
1	115258748	SNV	C	T	Heterozygous	NRAS	ENSP00000358548:p.Gly12Ser	ENST00000369535:c.34G>A
10	123245037	SNV	G	A	Homozygous	FGFR2		ENST00000358487:c.2067C>T

Figure 14.52: *Genome Browser View with an open track table to inspect identified somatic variants more closely in the context of the human genome and external databases.*

region could have an important functional role and variants with a conservation score of more than 0.9 (PhastCons score) should be prioritized higher. A further filtering of the variants based on their annotations can be facilitated using the table filter on top of the table.

If you wish to always apply the same filter criteria, the "Create new Filter Criteria" tool should be used to specify this filter and the "Identify and Annotate" workflow should be extended by the

"Identify Candidate Tool" (configured with the Filter Criterion). See the reference manual for more information on how preinstalled workflows can be edited.

Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

In case you like to change the databases as well as the used database version, please use the "Data Management".

Chapter 15

Whole Transcriptome Sequencing (WTS)

Contents

The technologies originally developed for next-generation DNA sequencing can also be applied to deep sequencing of the transcriptome. This is done through cDNA sequencing and is called RNA sequencing or simply RNA-seq.

One of the key advantages of RNA-seq is that the method is independent of prior knowledge of the corresponding genomic sequences and therefore can be used to identify transcripts from unannotated genes, novel splicing isoforms, and gene-fusion transcripts [Wang et al., 2009, Martin and Wang, 2011]. Another strength is that it opens up for studies of transcriptomic complexities such as deciphering allele-specific transcription by the use of SNPs present in the transcribed regions [Heap et al., 2010].

RNA-seq-based transcriptomic studies have the potential to increase the overall understanding of the transcriptome. However, the key to get access to the hidden information and be able to make a meaningful interpretation of the sequencing data highly relies on the downstream bioinformatic analysis.

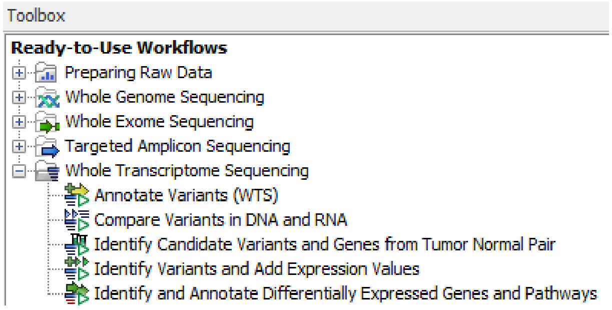
In this chapter we will first discuss the initial steps in the data analysis that lie upstream of the analysis using ready-to-use workflows. Next, we will look at what the individual ready-to-use workflows can be used for and go through step by step how to run the workflows.

15.1 Automatic analysis of RNA-seq data

The *CLC Cancer Research Workbench* offers a range of different tools for RNA-seq analysis. Currently four different ready-to-use workflows are available for analysis of RNA-seq data:

- Annotate Variants (WTS)
- Compare Variants in DNA and RNA
- Identify Candidate Variants and Genes from Tumor Normal Pair
- Identify and Annotate Differentially Expressed Genes and Pathways
- Identify Variants and Add Expression Values

The ready-to-use workflows can be found in the toolbox under Whole Transcriptome Sequencing as shown in figure 15.1.



Toolbox
Ready-to-Use Workflows
- Preparing Raw Data
- Whole Genome Sequencing
- Whole Exome Sequencing
- Targeted Amplicon Sequencing
- Whole Transcriptome Sequencing
  - Annotate Variants (WTS)
  - Compare Variants in DNA and RNA
  - Identify Candidate Variants and Genes from Tumor Normal Pair
  - Identify Variants and Add Expression Values
  - Identify and Annotate Differentially Expressed Genes and Pathways

Figure 15.1: *The RNA-seq ready-to-use workflows.*

Note! Often you will have to prepare data with one of the two **Preparing Raw Data** workflows described in section 11 before you proceed to the analysis of the sequencing data **RNA-Seq**.

15.2 Analysis of multiple samples

To analyze differential expression in multiple samples, you need to tell the workbench how the samples are related. This is done by setting up an experiment. The tool that can be used to do this can be found here:

Toolbox | **Tools** | **Transcriptomics Analysis** ()| **Set Up Experiment** ()

The output from the tool is an experiment, which essentially is a set of samples that are grouped. When setting up the experiment, you define the relationship between the samples. This makes it possible to do statistical analysis to investigate the differential expression between the groups. The experiment is also used to accumulate calculations like t-tests and clustering because this information is closely related to the grouping of the samples.

How to set up an experiment is described in detail in the CLC Cancer Research Workbench reference manual under "Setting up an experiment" in Chapter "Transcriptomics Analysis".

15.3 Annotate Variants (WTS)

Using a variant track () (e.g. the output from the Identify Variants and Add Expression Values ready-to-use workflow) the **Annotate Variants (WGS)** ready-to-use workflow runs an "internal" workflow that adds the following annotations to the variant track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.

- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.
- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from COSMIC**. Adds information from the "Catalogue of Somatic Mutations in Cancer" database.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (InDels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

1. Go to the toolbox and select the **Annotate Variants (WTS)** workflow. In the first wizard step, select the input variant track (figure 15.2).

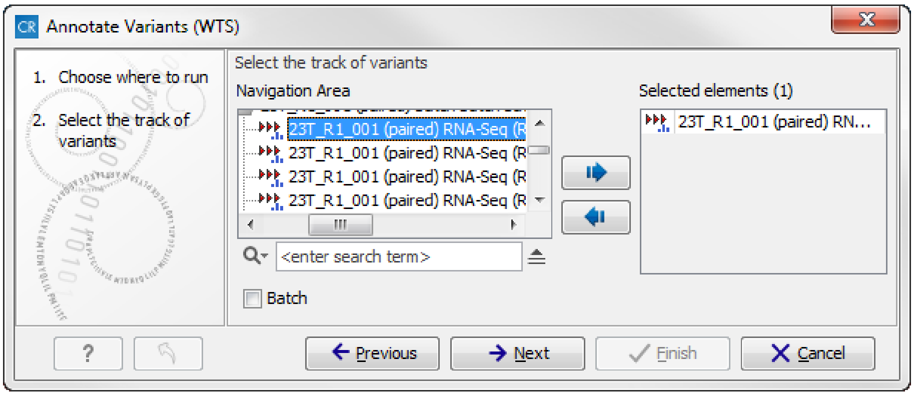

Annotate Variants (WTS)
1. Choose where to run
2. Select the track of variants
Select the track of variants
Navigation Area
23T_R1_001 (paired) RNA-Seq (R
23T_R1_001 (paired) RNA-Seq (R
23T_R1_001 (paired) RNA-Seq (R
23T_R1_001 (paired) RNA-Seq (R
<enter search term>
Batch
Selected elements (1)
23T_R1_001 (paired) RN...
?
Previous
Next
Finish
Cancel

Figure 15.2: *Select the variant track to annotate.*

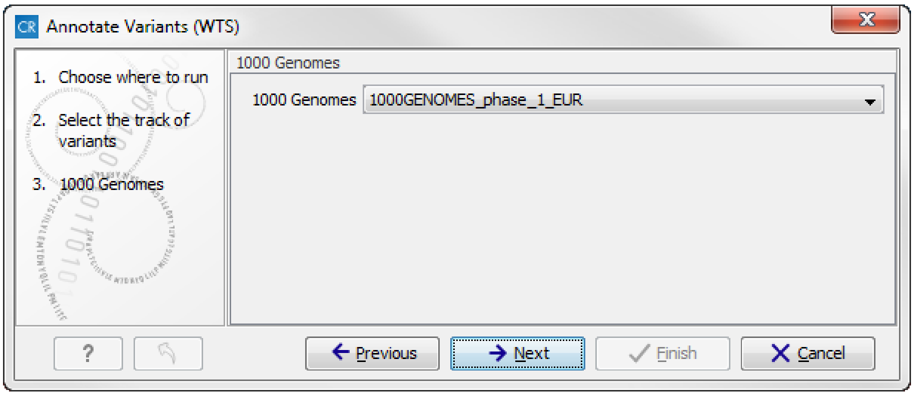
2. Click on the button labeled **Next**. The only parameter that should be specified by the user is which 1000 Genomes population yo use (figure 15.3). This can be done using the drop-down list found in this wizard step. Please note that the populations available from the drop-down list can be specified with the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).

3. Click on the button labeled **Next** to go to the last wizard step (figure 15.4).

   In this wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, it is not possible to make any changes at this point.

4. Choose to **Save** your results and click on the button labeled **Finish**.

Two types of output are generated:



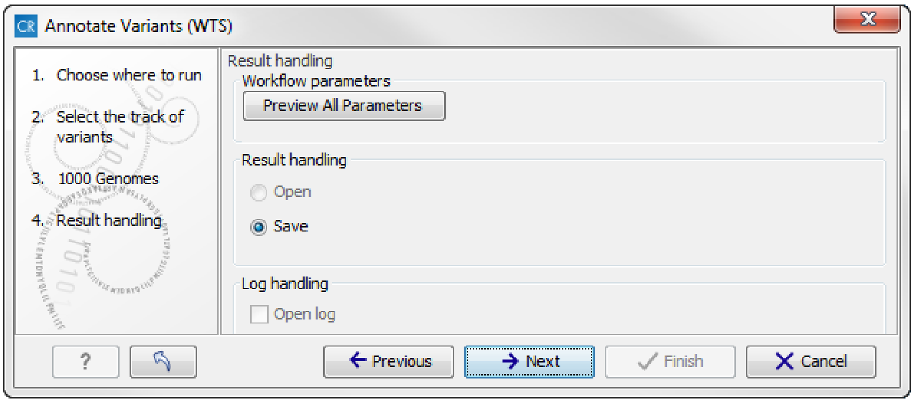
Annotate Variants (WTS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes

1000 Genomes

1000 Genomes: 1000GENOMES_phase_1_EUR

Figure 15.3: *Select the relevant 1000 Genomes popultaion(s).*



Annotate Variants (WTS)

1. Choose where to run
2. Select the track of variants
3. 1000 Genomes
4. Result handling

Result handling

Workflow parameters

Preview All Parameters

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 15.4: *Check the settings and save your results.*

1. **Annotated Variants** ( ) Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. **Genome Browser View Annotated Variants** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP, ClinVar, COSMIC, 1000 Genomes, and PhastCons conservation scores (see figure 15.5).

Note! Please be aware that if you delete the annotated variant track, this track will also disappear from the genome browser view.

It is possible to add tracks to the Genome Browser View such as mapped sequencing reads as well as other tracks. This can be done by dragging the track directly from the **Navigation Area** to the Genome Browser View.

If you double-click on the name of the annotated variant track in the left hand side of the Genome Browser View, a table that includes all variants and the added information/annotations will open (see figure 15.6). The table and the Genome Browser View are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Genome Browser View.

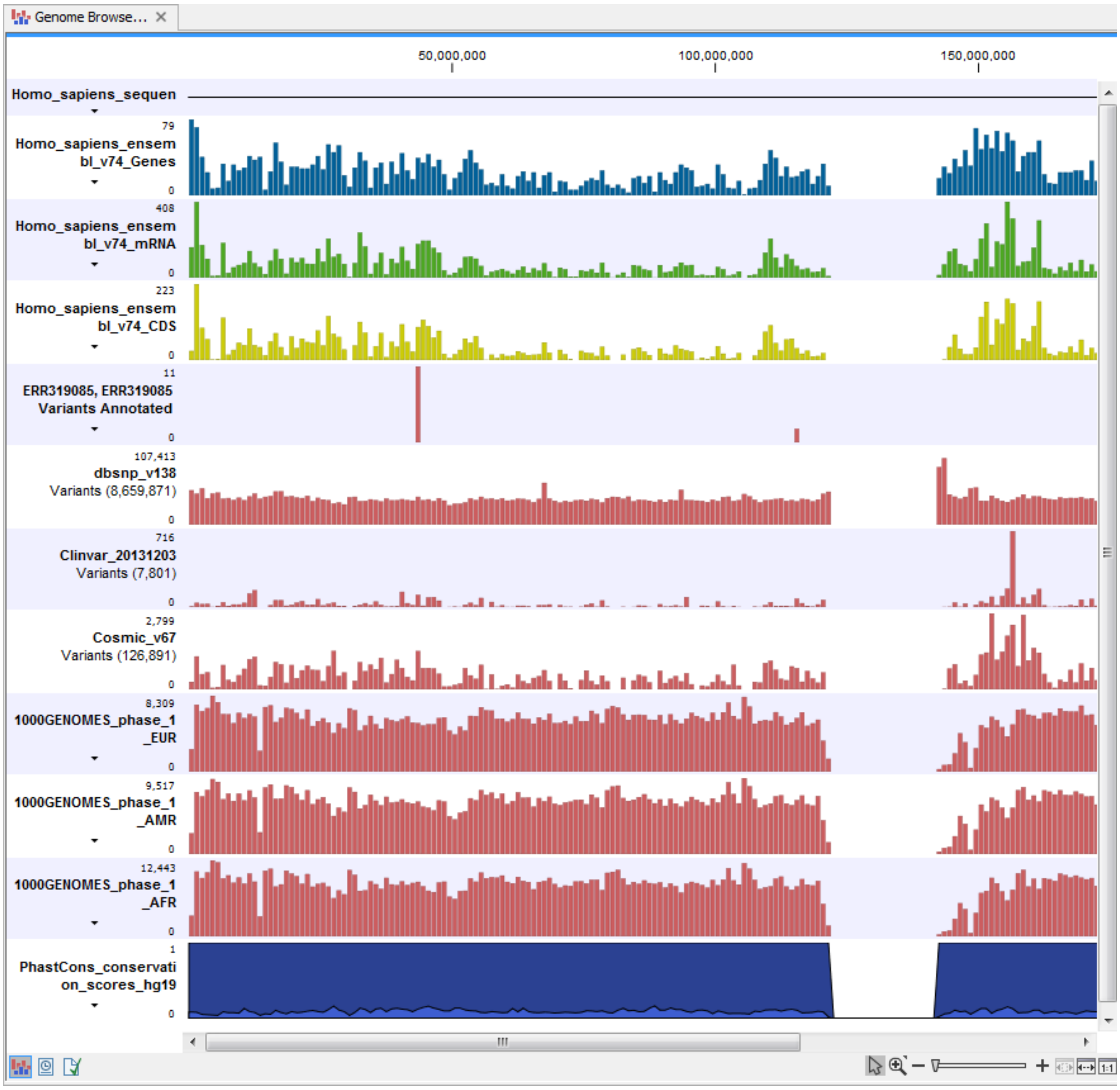
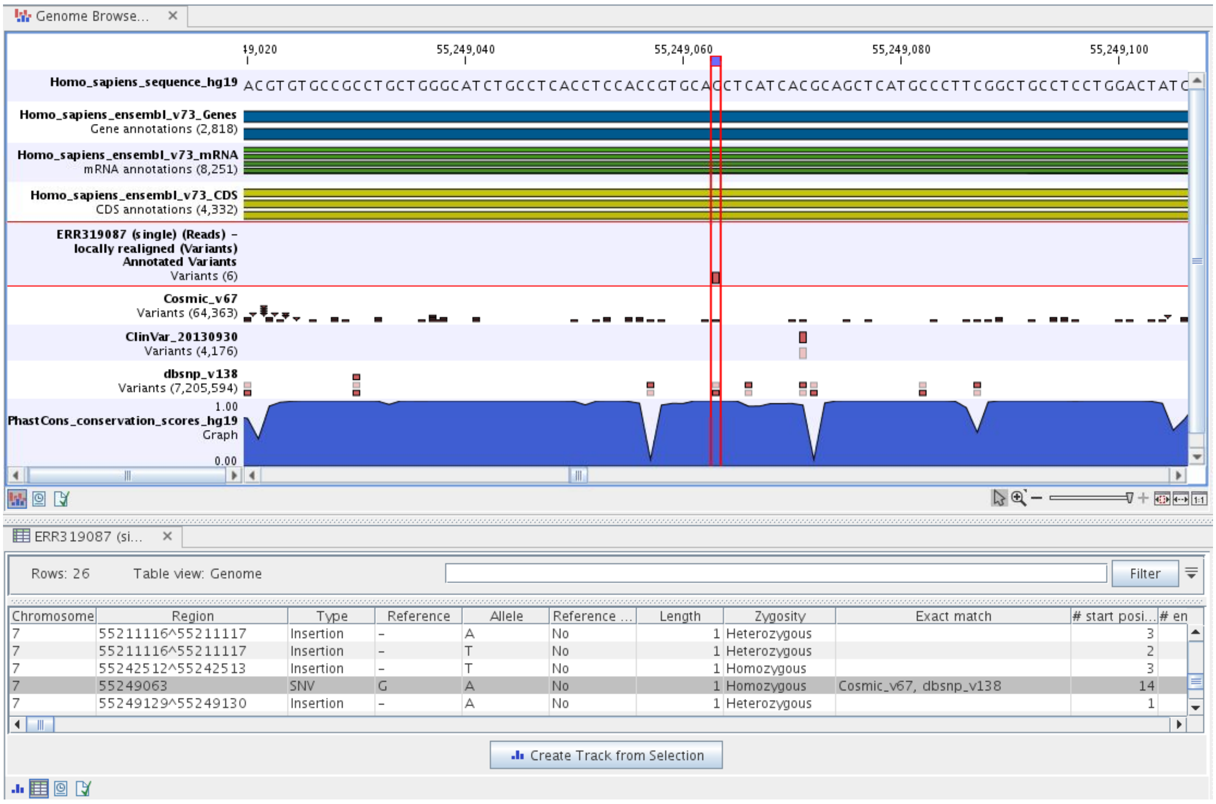


Figure 15.5: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list) containing individual tracks for all added annotations.*

You may be met with a warning as shown in figure 15.7. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in COSMIC, ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

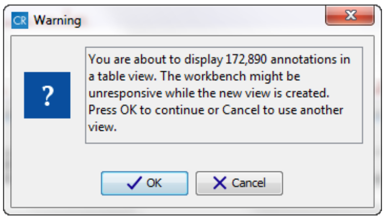
Adding information from other sources may help you identify interesting candidate variants for further research. E.g. known cancer associated variants (present in the COSMIC database) or variants known to play a role in drug response or other clinical relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the COSMIC and/or ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.


Genome Browse...
49,020
55,249,040
55,249,060
55,249,080
55,249,100
Homo_sapiens_sequence_hg19
ACGTGTGCCGCCTGCTGGGCATCTGCCTCACCTCCACCGTGCAGCTCATCACGCAGCTCATGCCCTTCGGCTGCCTCCTGGACTATC
Homo_sapiens_ensembl_v73_Genes
Gene annotations (2,818)
Homo_sapiens_ensembl_v73_mRNA
mRNA annotations (8,251)
Homo_sapiens_ensembl_v73_CDS
CDS annotations (4,332)
ERR319087 (single) (Reads) - locally realigned (Variants) Annotated Variants
Variants (6)
Cosmic_v67
Variants (64,363)
ClinVar_20130930
Variants (4,176)
dbsnp_v138
Variants (7,205,594)
1.00
PhastCons_conservation_scores_hg19
Graph
0.00
ERR319087 (si...
Filter
Create Track from Selection

Rows: 26 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Exact match	# start posi...	# en
7	55211116^55211117	Insertion	-	A	No	1	Heterozygous		3	
7	55211116^55211117	Insertion	-	T	No	1	Heterozygous		2	
7	55242512^55242513	Insertion	-	T	No	1	Homozygous		3	
7	55249063	SNV	G	A	No	1	Homozygous	Cosmic_v67, dbsnp_v138	14	
7	55249129^55249130	Insertion	-	A	No	1	Heterozygous		1	

Figure 15.6: *The output from the "Annotate Variants" ready-to-use workflow is a genome browser view (a track list). The information is also available in table view. Click on the small table icon to open the table view. If you hold down the "Ctrl" key while clicking on the table icon, you will open a split view showing both the genome browser view and the table view.*



You are about to display 172,890 annotations in a table view. The workbench might be unresponsive while the new view is created. Press OK to continue or Cancel to use another view.

Figure 15.7: *Warning that appears when you work with tracks containing many annotations.*

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this:

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

This tool can be used to specify the filter and the **Annotate Variants** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). The *CLC Cancer Research Workbench* reference manual has a chapter that describes this in detail (`http://clccancer.com/software/#downloads`, see chapter: "Workflows" for more information on how pre-installed workflows can be extended and/or edited).

Note! Sometimes the databases (e.g. COSMIC) are updated with a newer version, or maybe you have your own version of the database. In such cases you may wish to change one of the used databases. This can be done with "Data Management" function, which is described in section 10.1.4.

15.4 Compare variants in DNA and RNA

Integrated analysis of genomic and transcriptomic sequencing data is a powerful tool that can help increase our current understanding of human genomic variants. The **Compare variants in DNA and RNA** ready-to-use workflow identifies variants in DNA and RNA and studies the relationship between the identified genomic and transcriptomic variants.

To run the ready-to-use workflow:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Transcriptome Sequencing** () | **Compare variants in DNA and RNA** ()

1. Double-click on the **Compare variants in DNA and RNA** ready-to-use workflow to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the DNA reads that you would like to analyze (figure 15.8). To select the DNA reads, double-click on the reads file name or click once on the file and then on the arrow pointing to the right side in the middle of the wizard. Click on the button labeled **Next**.

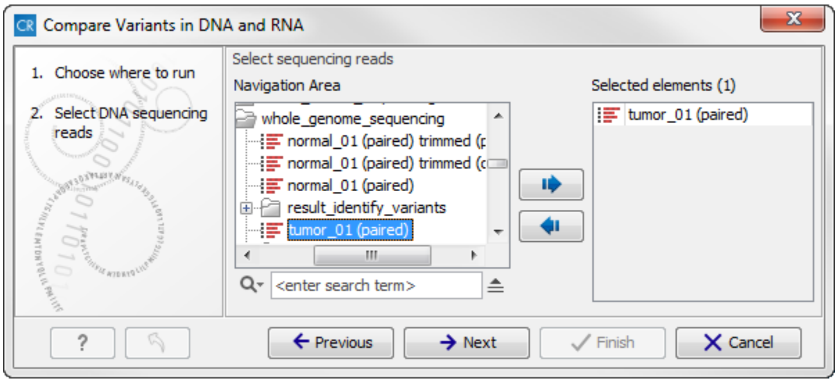
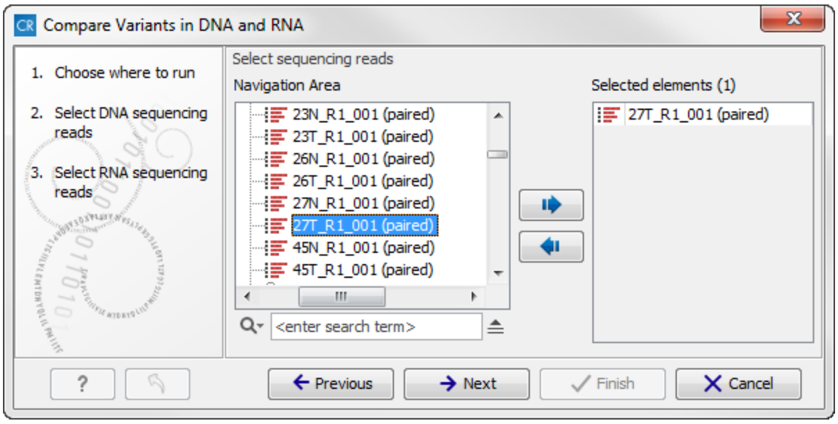


Figure 15.8: *Select the DNA reads to analyze.*

2. In the next step you can choose the RNA reads to analyze (see figure 15.9).

3. Click on the button labeled **Next** to go to the transcriptomic variant detection step (see figure 15.10). For a description of the different parameters that can be adjusted in the variant detection step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (`http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_`



Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads

Select sequencing reads

Navigation Area

- 23N_R1_001 (paired)
- 23T_R1_001 (paired)
- 26N_R1_001 (paired)
- 26T_R1_001 (paired)
- 27N_R1_001 (paired)
- 27T_R1_001 (paired)
- 45N_R1_001 (paired)
- 45T_R1_001 (paired)

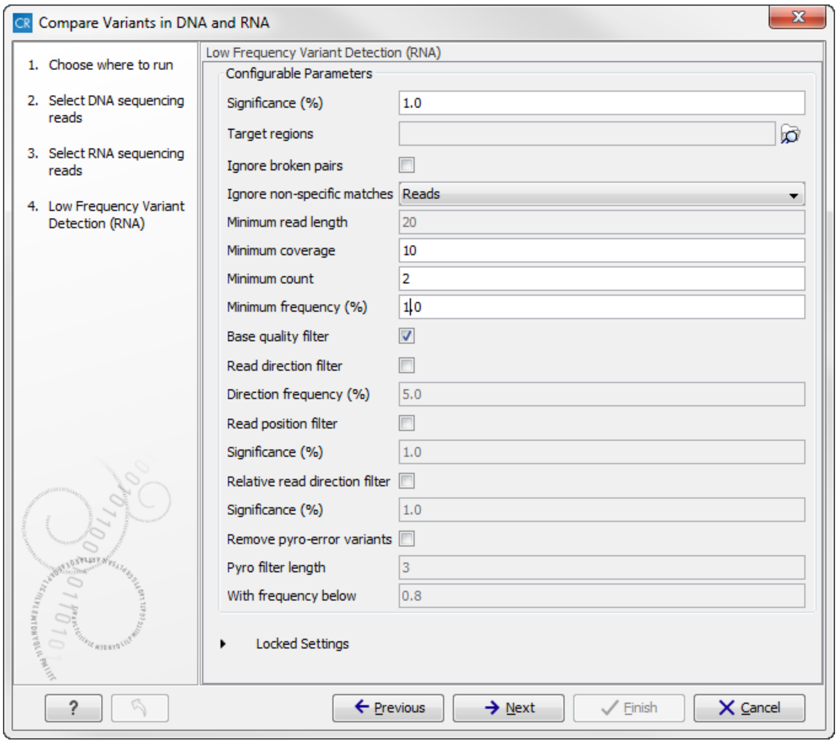
<enter search term>

Selected elements (1)

- 27T_R1_001 (paired)

Figure 15.9: *Select the RNA reads to analyze.*

Variant_Detection.html). As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the description of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Filters.html). If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the ready-to-use workflow.



Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)

Low Frequency Variant Detection (RNA)

Configurable Parameters

Significance (%)	1.0
Target regions	
Ignore broken pairs	□
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	2
Minimum frequency (%)	1.0
Base quality filter	☑
Read direction filter	□
Direction frequency (%)	5.0
Read position filter	□
Significance (%)	1.0
Relative read direction filter	□
Significance (%)	1.0
Remove pyro-error variants	□
Pyro filter length	3
With frequency below	0.8

Locked Settings

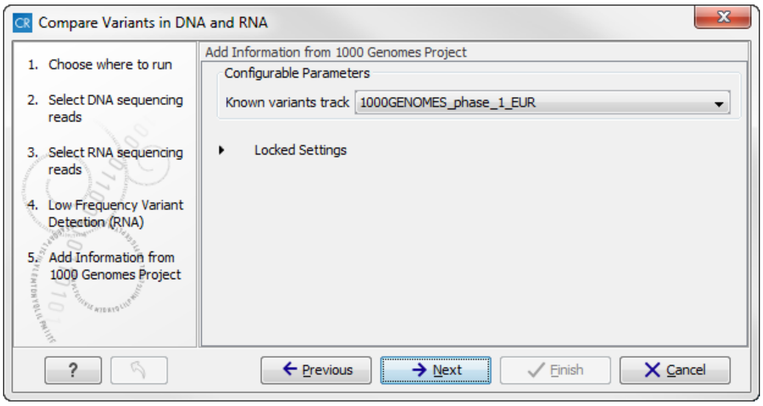
Figure 15.10: *Specify the parametes for transcriptomic variant detection.*

4. The next two wizard steps are annotation steps where the transcriptomic variants are annotated with information from known databases. Actually the variants are annotated with a range of different data in this ready-to-use workflow, but only databases that provide data from more than one population needs to be specified by the user. This is the case for

HapMap and the 1000 Genomes Project. First, the variants are annotated with information from the 1000 Genomes Project (see figure 15.11). From the drop-down list you can choose the population that matches the population your samples are derived from. The drop-down list shows the populations that were selected under "Data Management" as described in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Download_configure_reference_data.html).

Under "Locked settings" you can see that "Automatically join adjacent MNVs and SNVs" has been selected. The reason for this is that many databases do not report a succession of SNVs as one MNV as is the case for the *CLC Cancer Research Workbench*, and as a consequence it is not possible to directly compare variants called with *CLC Cancer Research Workbench* with these databases. In order to support filtering against these databases anyway, the option to **Automatically join adjacent SNVs and MNVs** is enabled. This means that an MNV in the experimental data will get an exact match, if a set of SNVs and MNVs in the database can be combined to provide the same allele.

Note! This assumes that SNVs and MNVs in the track of known variants represent the same allele, although there is no evidence for this in the track of known variants.


Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project

Add Information from 1000 Genomes Project

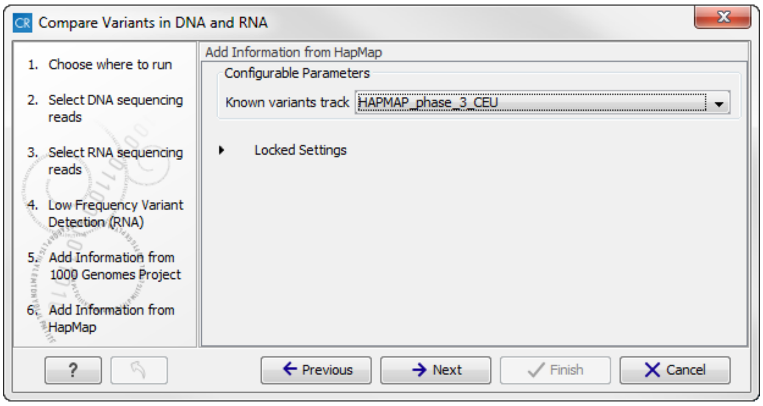
Configurable Parameters

Known variants track 1000GENOMES_phase_1_EUR

Locked Settings

Figure 15.11: *Select the relevant population from the drop-down list.*

5. Click on the button labeled **Next** and do the same to annotate with information from HapMap (figure 15.12).


Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project
6. Add Information from HapMap

Add Information from HapMap

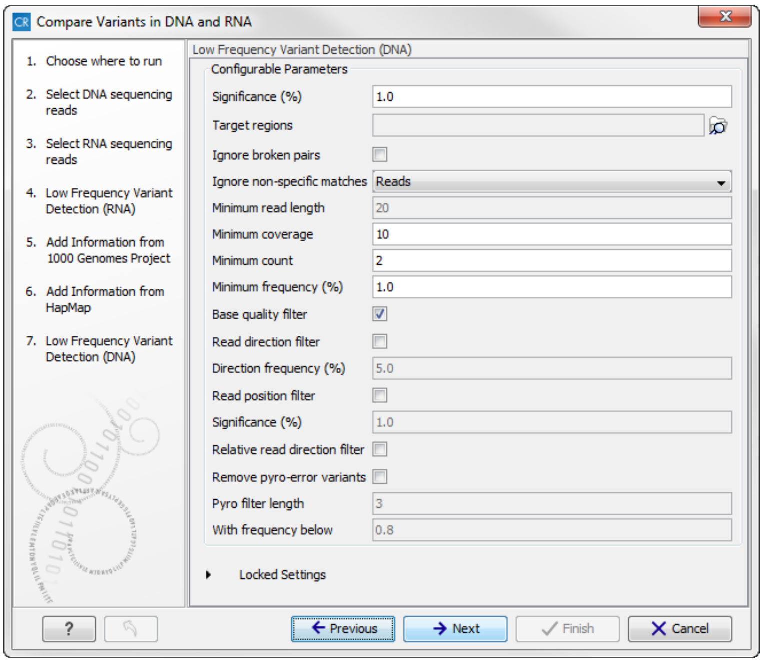
Configurable Parameters

Known variants track HAPMAP_phase_3_CEU

Locked Settings

Figure 15.12: *Select the relevant population from the drop-down list.*

6. Click on the button labeled **Next** to go to the genomic variant detection step (shown in figure 15.13).



Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project
6. Add Information from HapMap
7. Low Frequency Variant Detection (DNA)

Low Frequency Variant Detection (DNA)

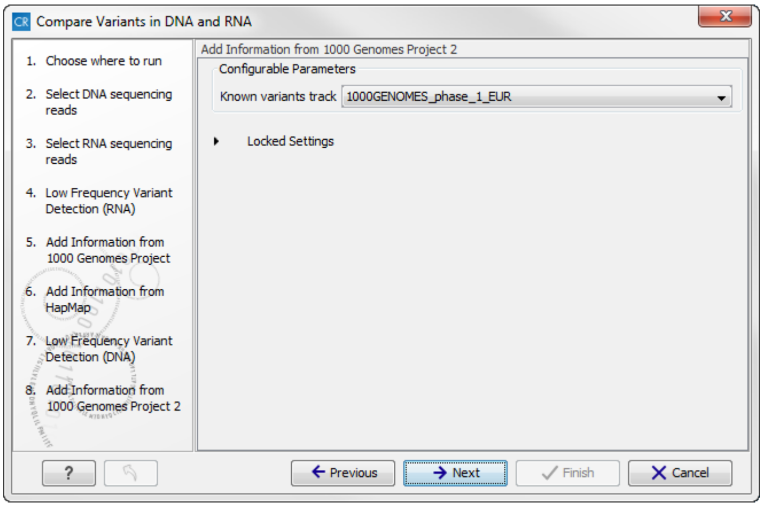
Configurable Parameters

Significance (%) 1.0
Target regions
Ignore broken pairs ☐
Ignore non-specific matches Reads
Minimum read length 20
Minimum coverage 10
Minimum count 2
Minimum frequency (%) 1.0
Base quality filter ☑
Read direction filter ☐
Direction frequency (%) 5.0
Read position filter ☐
Significance (%) 1.0
Relative read direction filter ☐
Remove pyro-error variants ☐
Pyro filter length 3
With frequency below 0.8

Locked Settings

Figure 15.13: *Specify the parametes for genomic variant detection.*

7. Again, the two next wizard steps are annotation steps. This time the genomic variants are annotated with information from known databases. First, the variants are annotated with information from the 1000 Genomes Project (see figure 15.14).



Compare Variants in DNA and RNA

1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project
6. Add Information from HapMap
7. Low Frequency Variant Detection (DNA)
8. Add Information from 1000 Genomes Project 2

Add Information from 1000 Genomes Project 2

Configurable Parameters

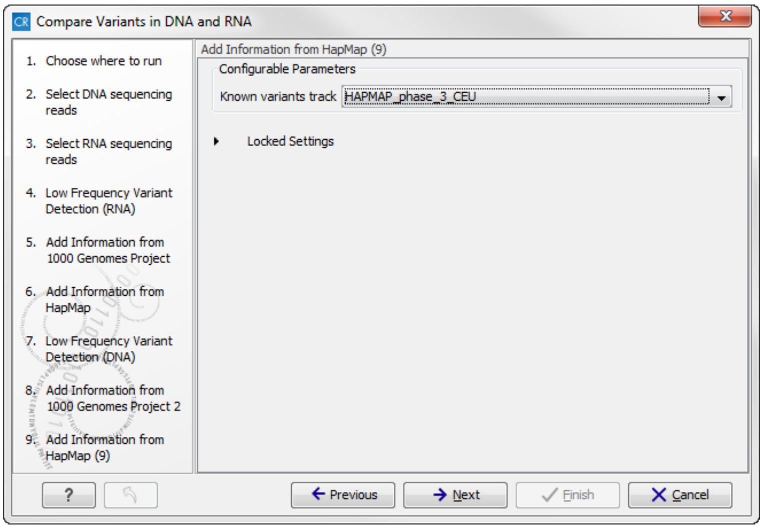
Known variants track 1000GENOMES_phase_1_EUR

Locked Settings

Figure 15.14: *Select the relevant population from the drop-down list.*

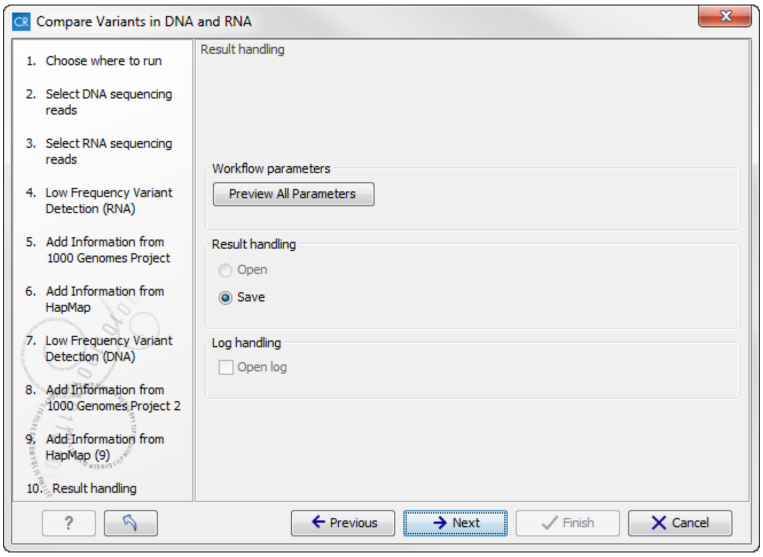
8. Click on the button labeled **Next** and do the same to annotate the genomic variants with information from HapMap (figure 15.15).

9. Click on the button labeled **Next** to go to the result handling step (figure 15.16).


Compare Variants in DNA and RNA
1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project
6. Add Information from HapMap
7. Low Frequency Variant Detection (DNA)
8. Add Information from 1000 Genomes Project 2
9. Add Information from HapMap (9)
Add Information from HapMap (9)
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Known variants track HAPMAP_phase_3_CEU

Figure 15.15: *Select the relevant population from the drop-down list.*


Compare Variants in DNA and RNA
1. Choose where to run
2. Select DNA sequencing reads
3. Select RNA sequencing reads
4. Low Frequency Variant Detection (RNA)
5. Add Information from 1000 Genomes Project
6. Add Information from HapMap
7. Low Frequency Variant Detection (DNA)
8. Add Information from 1000 Genomes Project 2
9. Add Information from HapMap (9)
10. Result handling
Result handling
Workflow parameters
Preview All Parameters
?
Previous
Next
Finish
Cancel

Result handling
○ Open
● Save

Log handling
□ Open log

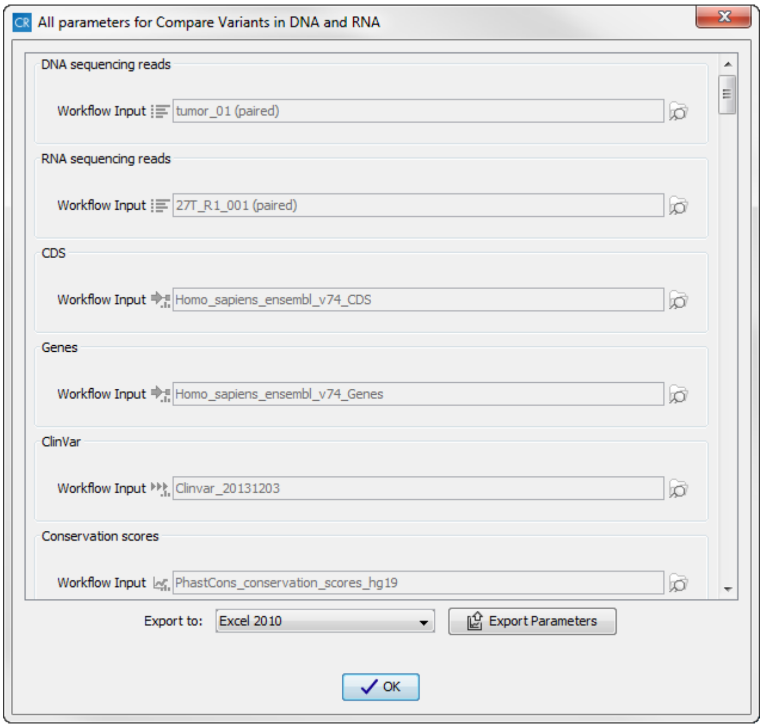
Figure 15.16: *Select the relevant population from the drop-down list.*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes (see figure 15.17). Choose to save the results and click on the button labeled **Finish**.

10. Press **OK**, specify where to save the results, and then click on the button labeled **Finish** to run the analysis.

Ten different output types are generated:

1. **DNA Read Mapping** () The mapped DNA sequencing reads. The DNA sequencing reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can



All parameters for Compare Variants in DNA and RNA

DNA sequencing reads

Workflow Input: tumor_01 (paired)

RNA sequencing reads

Workflow Input: 27T_R1_001 (paired)

CDS

Workflow Input: Homo_sapiens_ensembl_v74_CDS

Genes

Workflow Input: Homo_sapiens_ensembl_v74_Genes

ClinVar

Workflow Input: Clinvar_20131203

Conservation scores

Workflow Input: PhastCons_conservation_scores_hg19

Export to: Excel 2010

Figure 15.17: *Preview all parameters. At this step it is not possible to introduce any changes, it is only possible to view the settings.*

be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

2. **DNA Mapping Report** () This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Cancer Research Workbench reference manual in section **RNA-Seq report** (http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=RNA_Seq_report.html).

3. **RNA Gene Expression** () A track showing gene expression annotations. Hold the mouse over or right-clicking on the track. If you have zoomed in to nucleotide level, a tooltip will appear with information about e.g. gene name and expression values.

4. **RNA Transcript Expression** () A track showing transcript expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and expression values.

5. **RNA Mapping Report** () This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Cancer Research Workbench reference manual in section **RNA-Seq report** (http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=RNA_Seq_report.html).

6. **RNA Read Mapping** () The mapped RNA-seq reads. The RNA-seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the

description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

7. **Variants Found in Both DNA and RNA** ( ) This track shows only the variants that are present in both DNA and RNA. With the table icon ( ) found in the lower left part of the **View Area** it is possible to switch to table view. The table view provides details about the variants such as type, zygosity, and information from a range of different databases.

8. **All Variants Found in DNA or RNA** ( ) This track shows all variants that have been detected in either RNA, DNA or both.

9. **Genome Browser View Variants Found in DNA and RNA** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in COSMIC, ClinVar and dbSNP (see figure 15.18).

10. **Log** ( ) A log of the workflow execution.

The three most important tracks of the ten generated are the **Variants found in both DNA and RNA track**, **All variants found in DNA or RNA track**, and the **Genome Browser View**. The Genome Browser View makes it easy to get an overview in the context of a reference sequence, and compare variant and expression tracks with information from different databases. The two other tracks (**Variants found in both DNA and RNA track** and **All variants found in DNA or RNA track**) provides detailed information about the detected variants when opened in table view.

15.5 Identify Candidate Variants and Genes from Tumor Normal Pair

The **Identify Candidate Variants and Genes from Tumor Normal Pair** tool identifies somatic variants and differentially expressed genes in a tumor normal pair. One tumor normal pair can be compared at the time. If you would like to compare more than one pair you must repeat the analysis with the next tumor normal pair.

To run the ready-to-use workflow:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Transcriptome Sequencing** ( ) | **Identify Candidate Variants and Genes from Tumor Normal Pair** ( )

1. Double-click on the **Identify Candidate Variants and Genes from Tumor Normal Pair** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the RNA-seq reads from the normal sample. The panel in the left side of the wizard shows the kind of input that should be provided (figure 15.19). Select by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard. Click on the button labeled **Next**.

2. In the next step you will be asked to select the RNA-seq reads from the tumor sample (see figure 15.20).

3. Click on the button labeled **Next**. In this wizard step (figure 15.21) you can adjust the settings for the **Create fold change track** tool. In brief, what the tool does is,

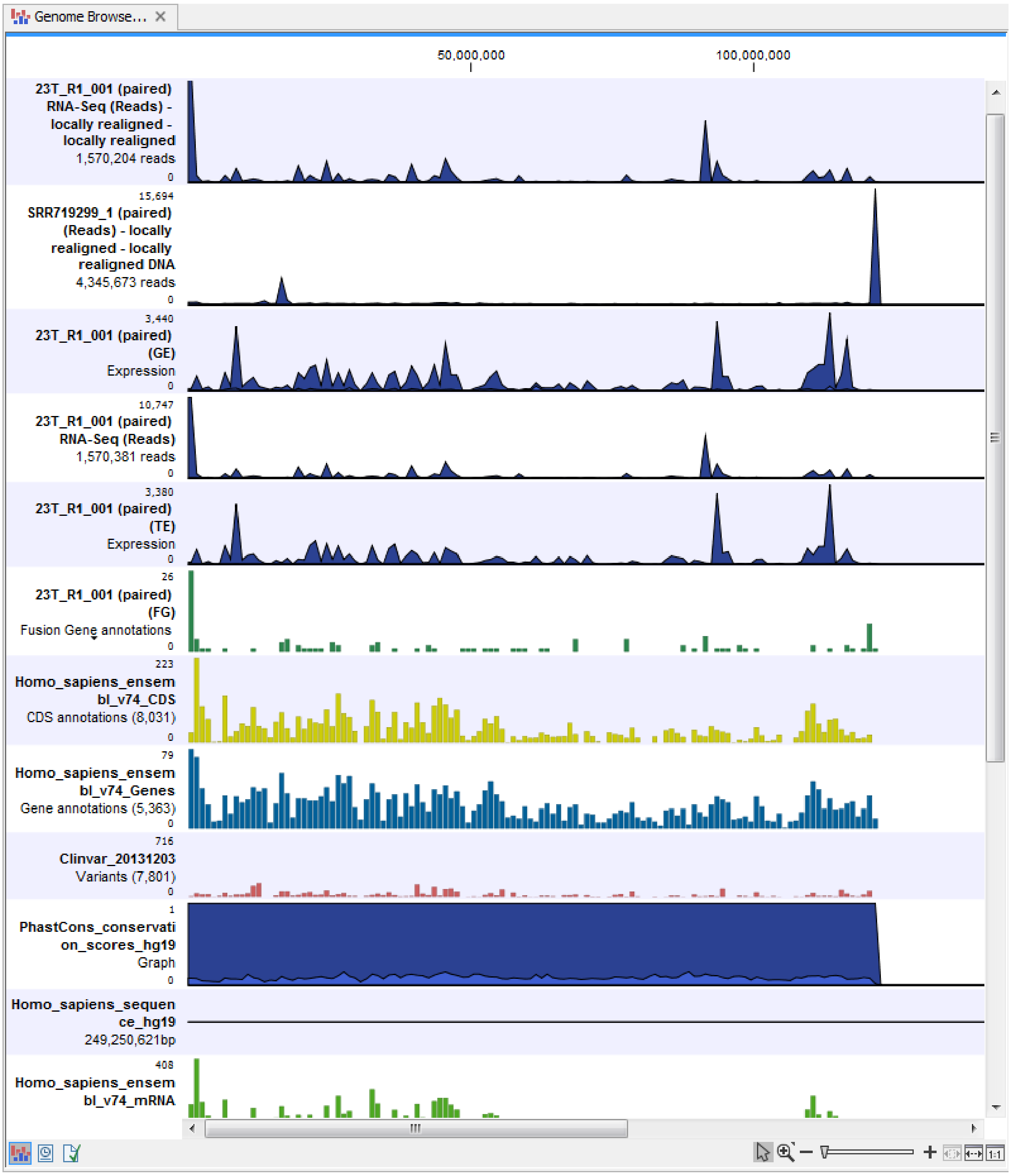


Figure 15.18: *The genome browser view makes it easy to compare a range of different data.*

for each transcript or gene, to calculate the ratio between the expression values in the normal and the tumor sample. This makes it possible to filter on fold changes and expression values, which makes it easy to identify differentially expressed transcripts or genes. The parameters that can be adjusted in this wizard step are described in detail in the *CLC Cancer Research Workbench* user manual (see http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Create_fold_change_

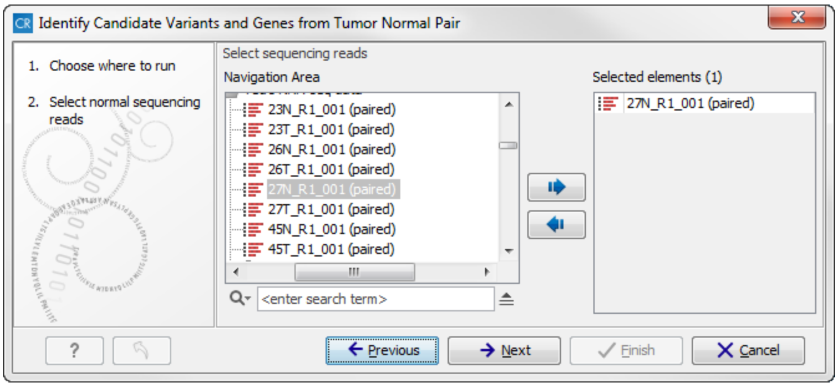


Figure 15.19: *Select the RNA-seq reads from the normal sample.*

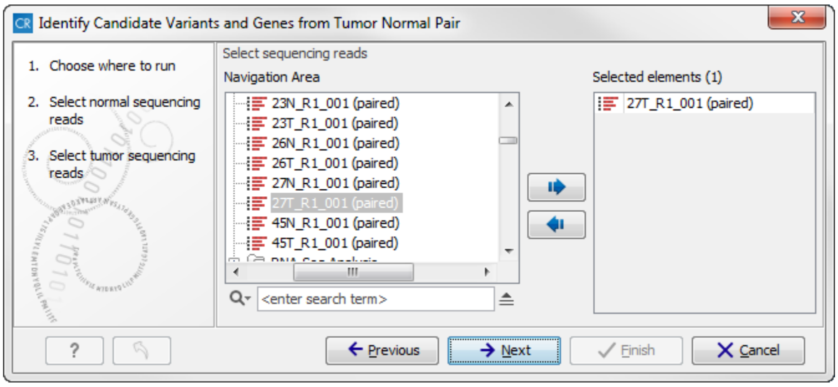
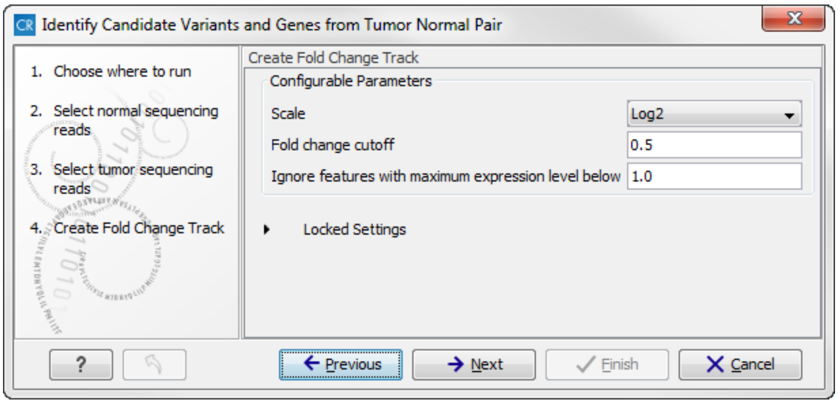


Figure 15.20: *Select the RNA-seq reads from the tumor sample.*

track.html).

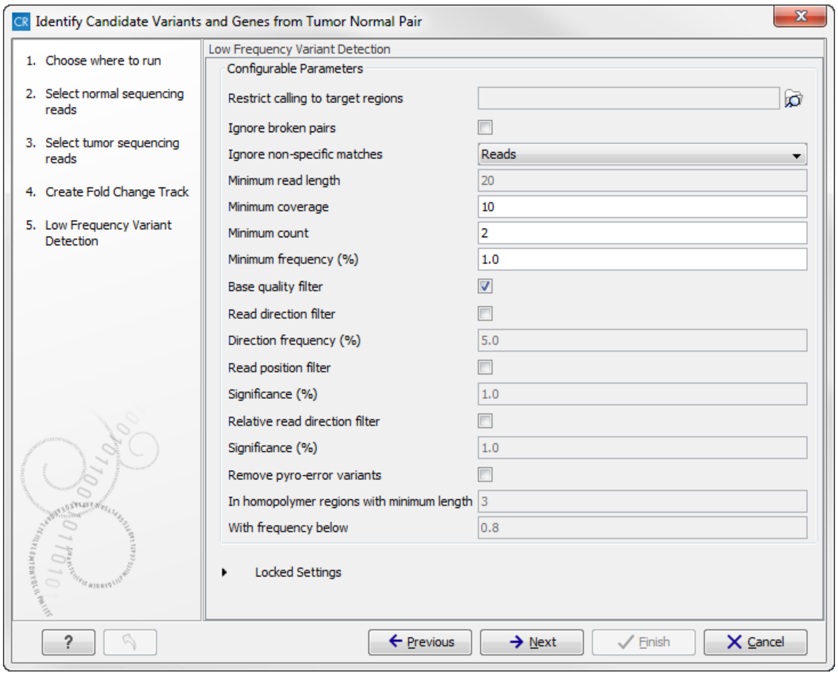

Identify Candidate Variants and Genes from Tumor Normal Pair
1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
Create Fold Change Track
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Scale: Log2
Fold change cutoff: 0.5
Ignore features with maximum expression level below: 1.0

Figure 15.21: *Specify the parameters for variant calling.*

4. Click on the button labeled **Next**. This will allow you to specify the parameters for the variant detection (figure 15.22). For a description of the different parameters that can be adjusted in the variant detection step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html). . As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the de-

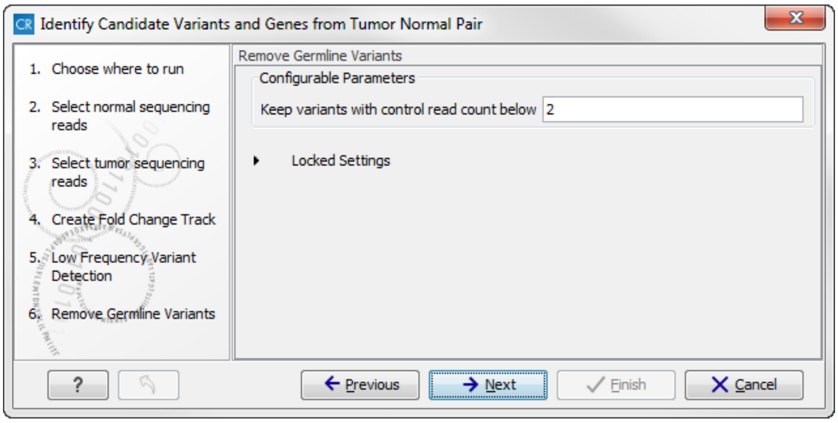
scription of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Filters.html). If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the ready-to-use workflow.


Identify Candidate Variants and Genes from Tumor Normal Pair
1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
5. Low Frequency Variant Detection
Low Frequency Variant Detection
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Restrict calling to target regions
Ignore broken pairs □
Ignore non-specific matches Reads
Minimum read length 20
Minimum coverage 10
Minimum count 2
Minimum frequency (%) 1.0
Base quality filter ☑
Read direction filter □
Direction frequency (%) 5.0
Read position filter □
Significance (%) 1.0
Relative read direction filter □
Significance (%) 1.0
Remove pyro-error variants □
In homopolymer regions with minimum length 3
With frequency below 0.8

Figure 15.22: *Specify the parameters for variant calling.*

5. The next wizard step (figure 15.23) concerns removal of germline variants. You are asked to supply the number of reads in the control data set that should support the variant allele in order to include it as a match. All the variants where at least this number of control reads show the particular allele will be filtered away in the result track.

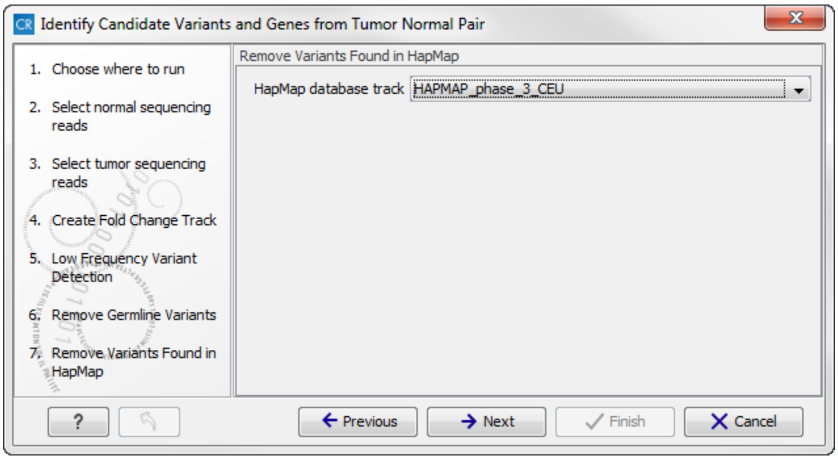

Identify Candidate Variants and Genes from Tumor Normal Pair
1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
5. Low Frequency Variant Detection
6. Remove Germline Variants
Remove Germline Variants
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Keep variants with control read count below 2

Figure 15.23: *Specify the number of reads to use as cutoff for removal of germline variants.*

6. In the next wizard step variants found in known databases are removed. Actually the variants from a range of different databases are removed in this ready-to-use workflow, but only databases that provide data from more than one population needs to be specified by the

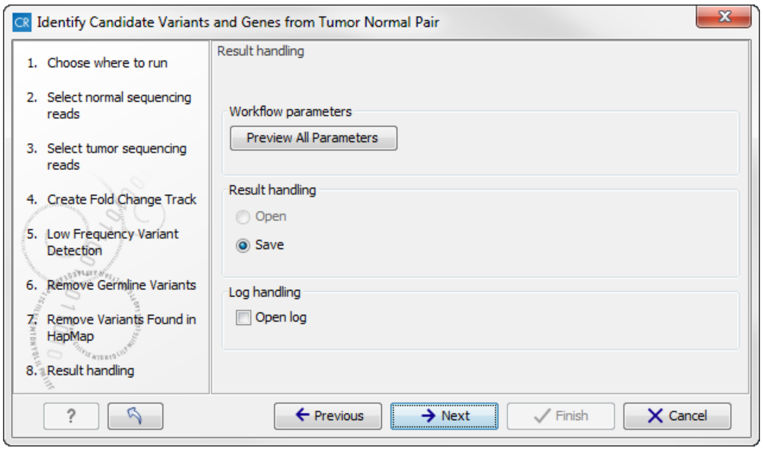
user. This is the case for the HapMap database. From the drop-down list you can choose the population that matches the population your samples are derived from (figure 15.24). The drop-down list shows the populations that were selected under "Data Management" as described in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Download_configure_reference_data.html).


Identify Candidate Variants and Genes from Tumor Normal Pair
1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
5. Low Frequency Variant Detection
6. Remove Germline Variants
7. Remove Variants Found in HapMap
Remove Variants Found in HapMap
?
Previous
Next
Finish
Cancel

HapMap database track HAPMAP_phase_3_CEU

Figure 15.24: *Select the relevant population from the drop-down list.*

7. Click on the button labeled **Next** to go to the last wizard step (shown in figure 15.25).


Identify Candidate Variants and Genes from Tumor Normal Pair
1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
5. Low Frequency Variant Detection
6. Remove Germline Variants
7. Remove Variants Found in HapMap
8. Result handling
Result handling
Workflow parameters
Preview All Parameters
?
Previous
Next
Finish
Cancel

Result handling

○ Open

● Save

Log handling

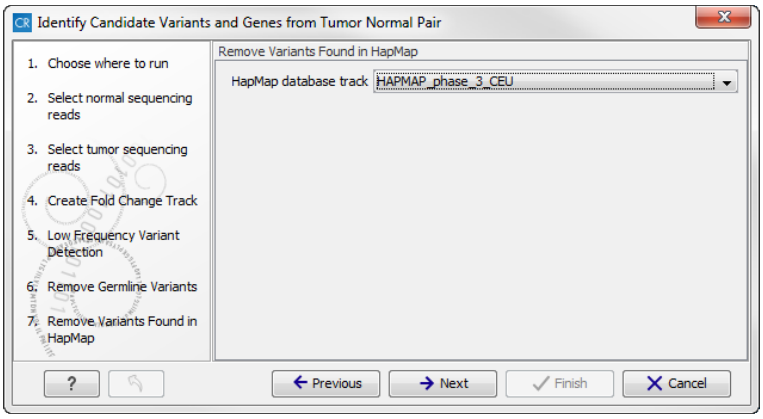
□ Open log

Figure 15.25: *Check the selected parametes by pressing "Preview All Parameters".*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes (see figure 15.26). Choose to save the results and click on the button labeled **Finish**.

Thirteen types of output are generated:

1. **Gene Expression Normal** () A track showing gene expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and gene expression values.



Identify Candidate Variants and Genes from Tumor Normal Pair

1. Choose where to run
2. Select normal sequencing reads
3. Select tumor sequencing reads
4. Create Fold Change Track
5. Low Frequency Variant Detection
6. Remove Germline Variants
7. Remove Variants Found in HapMap

Remove Variants Found in HapMap

HapMap database track HAPMAP_phase_3_CEU

Figure 15.26: *Preview all parameters. At this step it is not possible to introduce any changes, it is only possible to view the settings.*

2. **Transcript Expression Normal** () A track showing transcript expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and transcript expression values.

3. **RNA-Seq Mapping Report Normal** () This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Cancer Research Workbench reference manual in section **RNA-Seq report** (http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=RNA_Seq_report.html).

4. **Gene Expression Tumor** () A track showing gene expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and gene expression values.

5. **Transcript Expression Tumor** () A track showing transcript expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and transcript expression values.

6. **RNA-Seq Mapping Report Tumor** () This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Cancer Research Workbench reference manual in section **RNA-Seq report** (http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=RNA_Seq_report.html).

7. **Differentially Expressed Genes** () A track showing the differentially expressed genes. The table view provides information about fold change, difference in expression, the maximum expression (observed in either the case or the control), the expression in the case, and the expression in the control.

8. **Read Mapping Tumor** () The mapped RNA-seq reads. The RNA-seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

9. **Read Mapping Normal** ( ) The mapped RNA-seq reads. The RNA-seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

10. **Variant Calling Report Tumor** ( ) Report showing error rates for quality categories, quality of examined sites, and estimated frequencies of actual to called bases for different quality score ranges.

11. **Annotated Somatic Variants with Expression Values** ( ) A variant track showing the somatic variants. When mousing over a variant, a tooltip will appear with information about the variant.

12. **Genome Browser View RNA-Seq Tumor_Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in COSMIC, ClinVar and dbSNP (see figure 15.27).

13. **Log** ( ) A log of the workflow execution.

15.6 Identify variants and add expression values

The **Identify Variants and Add Expression Values** ready-to-use workflows can be used to identify novel and known mutations in RNA-seq data, automatically map, quantify, and annotate the transcriptomes, and compare the mutational patterns in the samples with the expression values of the corresponding transcripts and genes.

To run the ready-to-use workflow:

> **Toolbox | Ready-to-Use Workflows | Whole Transcriptome Sequencing ( ) | Identify Variants and Add Expression Values ( )**

1. Double-click on the **Identify Variants and Add Expression Values** tool to start the analysis. If you are connected to a server, you will first be asked, where you would like to run the analysis. Next, you will be asked to select the RNA-seq reads. The reads can be selected by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard (figure 15.30).

   Click on the button labeled **Next**.

2. In the next wizard step (figure 15.29) you can specify the parameters for variant detection. For a description of the different parameters that can be adjusted in the variant detection step, we refer to the description of the "Low Frequency Variant Detection" tool in the *CLC Cancer Research Workbench* user manual (http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html). As general filters are applied to the different variant detectors that are available in *CLC Cancer Research Workbench*, the description of the filters are found in a separate section called "Filters" (see http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Filters.html).

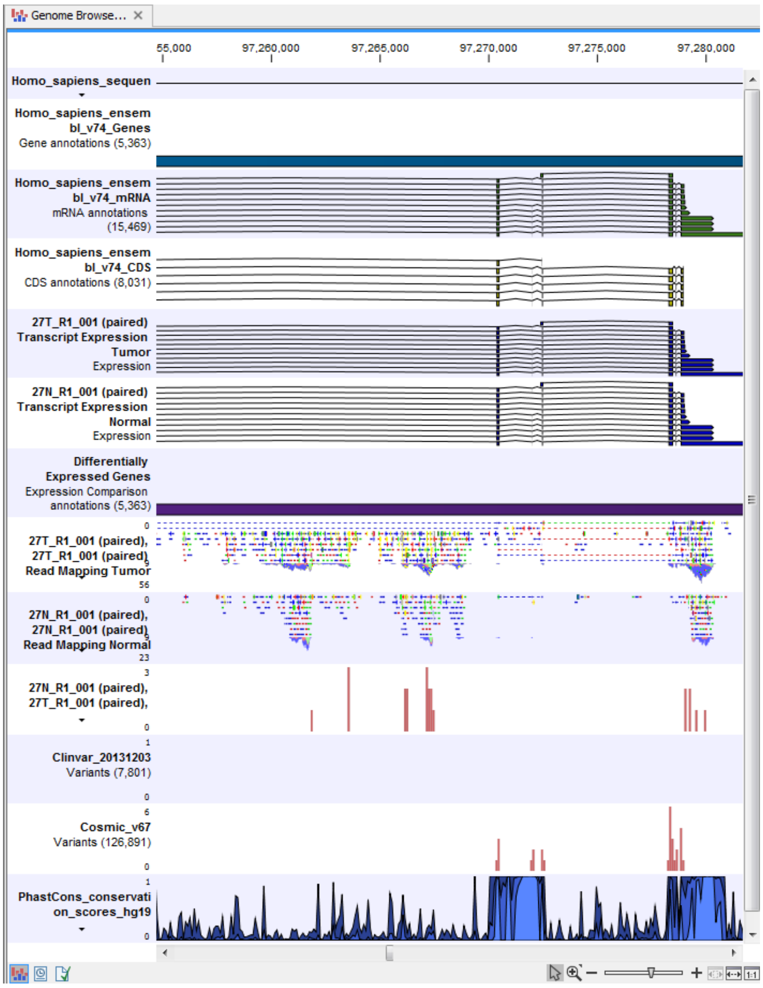
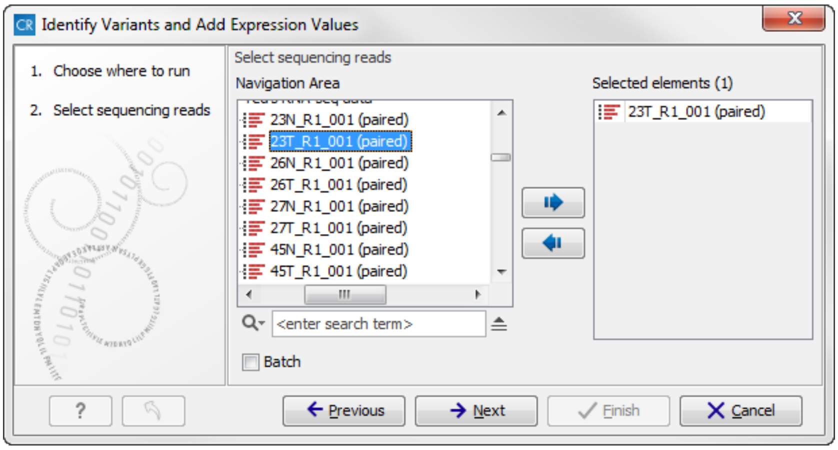

Genome Browse...
55,000
97,260,000
97,265,000
97,270,000
97,275,000
97,280,000
Homo_sapiens_sequen
Homo_sapiens_ensembl_v74_Genes
Gene annotations (5,363)
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (15,469)
Homo_sapiens_ensembl_v74_CDS
CDS annotations (8,031)
27T_R1_001 (paired) Transcript Expression Tumor
Expression
27N_R1_001 (paired) Transcript Expression Normal
Expression
Differentially Expressed Genes
Expression Comparison annotations (5,363)
0
27T_R1_001 (paired), 27T_R1_001 (paired) Read Mapping Tumor
56
0
27N_R1_001 (paired), 27N_R1_001 (paired) Read Mapping Normal
23
3
27N_R1_001 (paired), 27T_R1_001 (paired),
0
1
Clinvar_20131203
Variants (7,801)
0
6
Cosmic_v67
Variants (126,891)
0
1
PhastCons_conservation_scores_hg19
0

Figure 15.27: *The Genome Browser View is a collection of a number of tracks. The Genome Browser View makes it easy to compare the different tracks. Each track kan be opened individually by double-clicking on the track name in the left side of the View Area.*

3. The next two wizard steps are annotation steps where the detected variants are annotated with information from known databases. Actually the variants are annotated with a range of different data in this ready-to-use workflow, but only databases that provide data from more than one population needs to be specified by the user. This is the case for HapMap and the 1000 Genomes Project. First, the variants are annotated with information from the 1000 Genomes Project (see figure 15.30). From the drop-down list you can choose the population that matches the population your samples are derived from. The drop-down list shows the populations that were selected under "Data Management" as described in the *CLC Cancer Research Workbench* user manual (`http://www.clcsupport.com/clccancerresearchworkbench/current/index.php?manual=Download_configure_reference_data.html`).

   Under "Locked settings" you can see that "Automatically join adjacent MNVs and SNVs" has been selected. The reason for this is that many databases do not report a succession of SNVs as one MNV as is the case for the *CLC Cancer Research Workbench*, and as a consequence it is not possible to directly compare variants called with *CLC Cancer Research Workbench* with these databases. In order to support filtering against these databases



Identify Variants and Add Expression Values

1. Choose where to run
2. Select sequencing reads

Select sequencing reads

Navigation Area

- 23N_R1_001 (paired)
- 23T_R1_001 (paired)
- 26N_R1_001 (paired)
- 26T_R1_001 (paired)
- 27N_R1_001 (paired)
- 27T_R1_001 (paired)
- 45N_R1_001 (paired)
- 45T_R1_001 (paired)

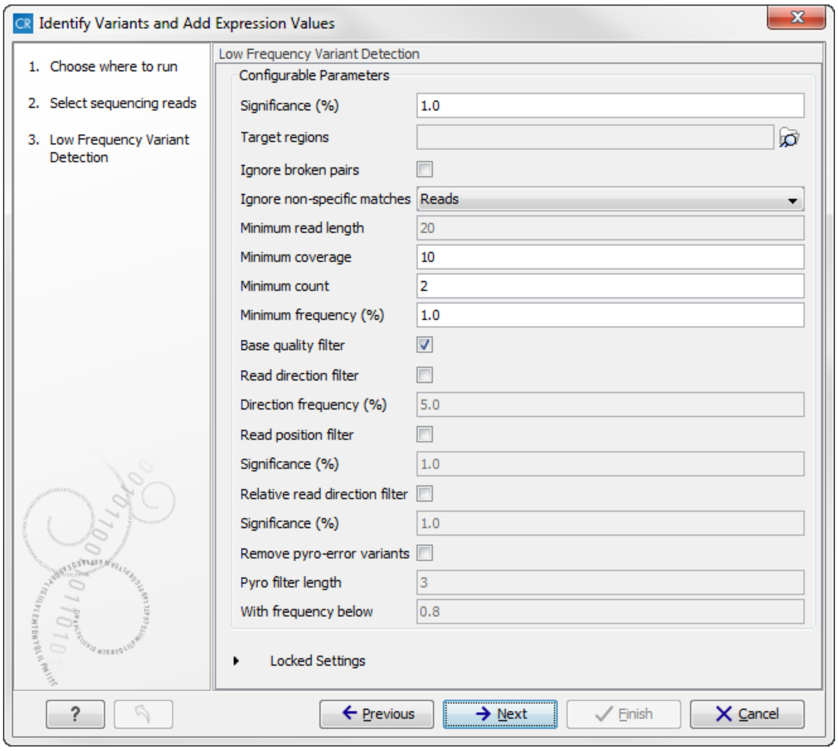
<enter search term>

□ Batch

Selected elements (1)

- 23T_R1_001 (paired)

Figure 15.28: *Select the sequencing reads to analyze.*



Identify Variants and Add Expression Values

1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection

Low Frequency Variant Detection

Configurable Parameters

Significance (%)	1.0
Target regions	
Ignore broken pairs	□
Ignore non-specific matches	Reads
Minimum read length	20
Minimum coverage	10
Minimum count	2
Minimum frequency (%)	1.0
Base quality filter	☑
Read direction filter	□
Direction frequency (%)	5.0
Read position filter	□
Significance (%)	1.0
Relative read direction filter	□
Significance (%)	1.0
Remove pyro-error variants	□
Pyro filter length	3
With frequency below	0.8

Locked Settings

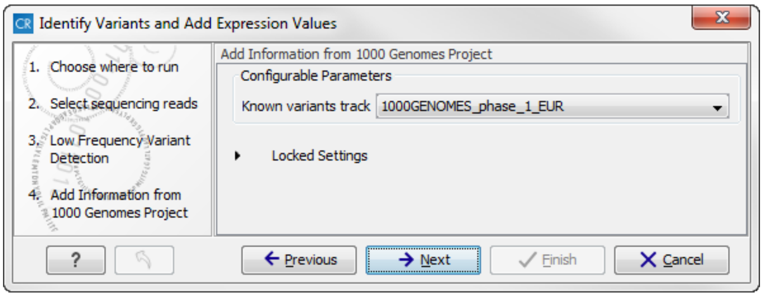
Figure 15.29: *Specify the parameters for variant calling.*

anyway, the option to **Automatically join adjacent SNVs and MNVs** is enabled. This means that an MNV in the experimental data will get an exact match, if a set of SNVs and MNVs in the database can be combined to provide the same allele.

Note! This assumes that SNVs and MNVs in the track of known variants represent the same allele, although there is no evidence for this in the track of known variants.

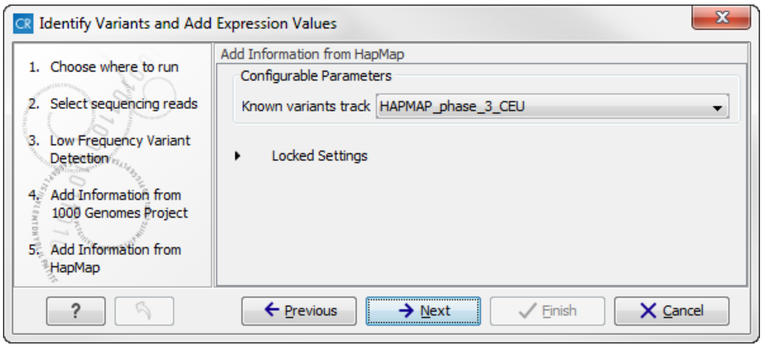
4. Click on the button labeled **Next** and do the same to annotate with information from HapMap (figure 15.31).

5. Click on the button labeled **Next** to go to the last wizard step (shown in figure 15.32).


Identify Variants and Add Expression Values
1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection
4. Add Information from 1000 Genomes Project
Add Information from 1000 Genomes Project
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

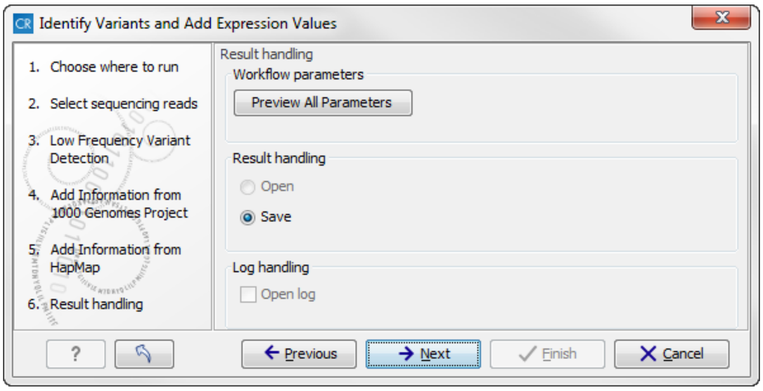
Known variants track: 1000GENOMES_phase_1_EUR

Figure 15.30: *Select the relevant population from the drop-down list.*


Identify Variants and Add Expression Values
1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection
4. Add Information from 1000 Genomes Project
5. Add Information from HapMap
Add Information from HapMap
Configurable Parameters
Locked Settings
?
Previous
Next
Finish
Cancel

Known variants track: HAPMAP_phase_3_CEU

Figure 15.31: *Select the relevant population from the drop-down list.*


Identify Variants and Add Expression Values
1. Choose where to run
2. Select sequencing reads
3. Low Frequency Variant Detection
4. Add Information from 1000 Genomes Project
5. Add Information from HapMap
6. Result handling
Result handling
Workflow parameters
Preview All Parameters
?
Previous
Next
Finish
Cancel

Result handling
○ Open
● Save

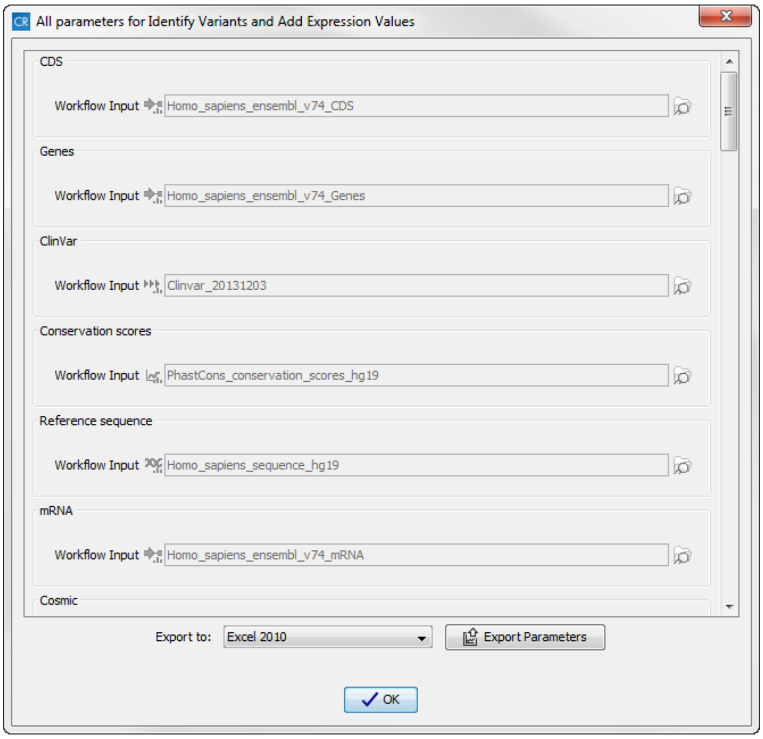
Log handling
□ Open log

Figure 15.32: *Check the selected parametes by pressing "Preview All Parameters".*

Pressing the button **Preview All Parameters** allows you to preview all parameters. At this step you can only view the parameters, it is not possible to make any changes (see figure 15.33). Choose to save the results and click on the button labeled **Finish**.

Seven different output types are generated:

1. **Gene expression** () A track showing gene expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and expression values.
2. **Transcript expression** () A track showing transcript expression annotations. Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name and expression values.



All parameters for Identify Variants and Add Expression Values

CDS

Workflow Input Homo_sapiens_ensembl_v74_CDS

Genes

Workflow Input Homo_sapiens_ensembl_v74_Genes

ClinVar

Workflow Input Clinvar_20131203

Conservation scores

Workflow Input PhastCons_conservation_scores_hg19

Reference sequence

Workflow Input Homo_sapiens_sequence_hg19

mRNA

Workflow Input Homo_sapiens_ensembl_v74_mRNA

Cosmic

Export to: Excel 2010

Figure 15.33: *Preview all parameters. At this step it is not possible to introduce any changes, it is only possible to view the settings.*

3. **RNA-Seq Mapping Report** ( ) This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Cancer Research Workbench reference manual in section **RNA-Seq report** (http://clcsupport.com/clccancerresearchworkbench/current/index.php?manual=RNA_Seq_report.html).

4. **Read Mapping** ( ) The mapped RNA-seq reads. The RNA-seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously. For the color codes please see the description of sequence colors in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

5. **Annotated Variants with Expression Values** ( ) Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

6. **RNA-Seq Genome Browser View** ( ) A collection of tracks presented together. Shows the annotated variants track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in COSMIC, ClinVar and dbSNP (see figure 15.18).

7. **Log** ( ) A log of the workflow execution.

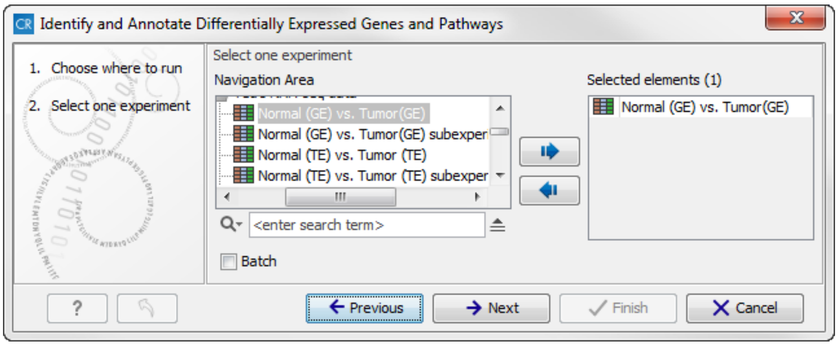
15.7 Identify and Annotate Differentially Expressed Genes and Pathways

The **Identify and Annotate Differentially Expressed Genes and Pathways** compares the gene expression in different groups of samples using an empirical analysis and performs a gene ontology (GO) enrichment analysis on the differentially expressed genes to identify affected pathways.

To run the ready-to-use workflow:

> **Toolbox** | **Ready-to-Use Workflows** | **Whole Transcriptome Sequencing** ( ) | **Identify and Annotate Differentially Expressed Genes and Pathways** ( )

1. Double-click on the **Identify and Annotate Differentially Expressed Genes and Pathways** ready-to-use workflow to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the experiment to analyze (figure 15.34). To select an experiment ( ), double-click on the experiment file name or click once on the file and then on the arrow pointing to the right side in the middle of the wizard. Click on the button labeled **Next**.



Identify and Annotate Differentially Expressed Genes and Pathways

1. Choose where to run
2. Select one experiment

Select one experiment

Navigation Area

- Normal (GE) vs. Tumor(GE)
- Normal (GE) vs. Tumor(GE) subexper
- Normal (TE) vs. Tumor (TE)
- Normal (TE) vs. Tumor (TE) subexper

<enter search term>

□ Batch

Selected elements (1)

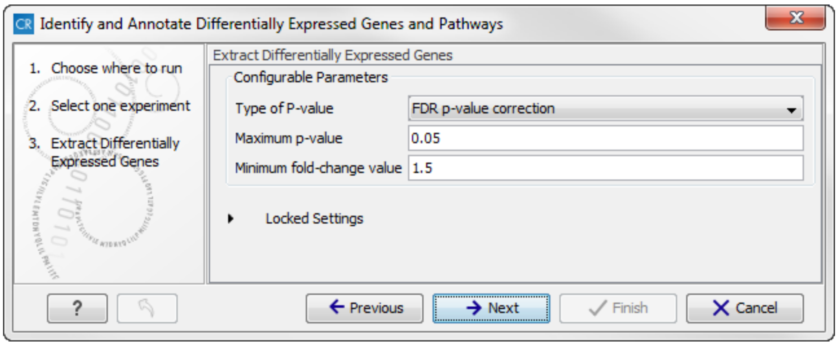
- Normal (GE) vs. Tumor(GE)

Figure 15.34: *Select the experiment to analyze.*

2. In the next wizard step you can specify the parameters to be used for extraction of differentially expressed genes.

   Configurable Parameters

   - **Type of p-value** This drop-down menu allows you to select between raw and corrected p-values. For a description of these, please see the Transcriptomics Chapter, section "Corrected p-values" in the CLC Genomics Workbench manual that can be found here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Corrected_p_values.html. Only the types of p-values available for the given statistical analysis will be present in the drop-down menu.
   - **Maximum p-value** In this input field, you can enter the maximum allowed p-value, as a number between 0 and 1. If you do not want any filtering based on p-value, enter 1.
   - **Minimum fold-change value** You can also specify the minimum allowed fold-change value as a number greater than zero. If you do not want any filtering based on fold-change, enter 0.


Identify and Annotate Differentially Expressed Genes and Pathways

Extract Differentially Expressed Genes

Configurable Parameters

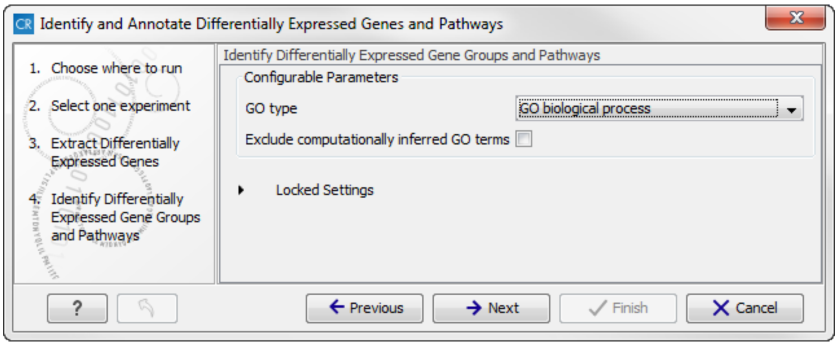
Type of P-value: FDR p-value correction

Maximum p-value: 0.05

Minimum fold-change value: 1.5

Figure 15.35: *Select the parameters for extraction of differentially expressed genes.*

3. Click on the button labeled **Next** to go to the next step where you can choose the gene ontology type you wish to use.


Identify and Annotate Differentially Expressed Genes and Pathways

Identify Differentially Expressed Gene Groups and Pathways

Configurable Parameters

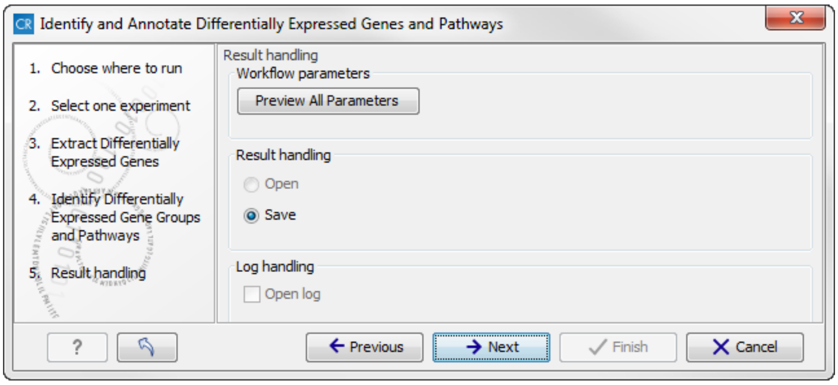
GO type: GO biological process

Exclude computationally inferred GO terms ☐

1. Choose where to run
2. Select one experiment
3. Extract Differentially Expressed Genes
4. Identify Differentially Expressed Gene Groups and Pathways
Locked Settings
?
Previous
Next
Finish
Cancel

Figure 15.36: *Select which gene ontology type to use.*

4. In the next step you can choose to preview the settings and save the results (see figure 15.37).


Identify and Annotate Differentially Expressed Genes and Pathways

Result handling

Workflow parameters

Result handling: ○ Open ● Save

Log handling: ☐ Open log

Figure 15.37: *The results handling step.*

5. Click on the button labeled "Preview All Parameters" if you would like to preview the settings. The parameters settings can be viewed but not edited in this view.

6. Press **OK**, specify where to save the results, and then click on the button labeled **Finish** to run the analysis.

Three different types of output are generated:

1. **Annotated Differentially Expressed Genes** ( ) This is an annotation track that gives access to the expression values and other information. This information can be accessed in two different ways:
    - Hold the mouse over or right-clicking on the track. A tooltip will appear with information about e.g. gene name, results of statistical tests, expression values, and GO information.
    - Open the track in table format by clicking on the table icon in the lower left side of the View Area.
2. **Enriched Gene Groups and Pathways** ( ) A table showing the results of the GO enrichment analysis. The table includes GO terms, a description of the affected function/pathway, the number of genes in each function/pathway, the number of affected genes within the function/pathway, and p-values.
3. **Genome Browser View Differentially Expressed Genes and Pathways** ( ) A collection of tracks presented together. Shows the human reference sequence, annotation tracks for genes, coding regions, transcripts, and expression comparison with GO information, and a conservation score track (see figure 15.38).

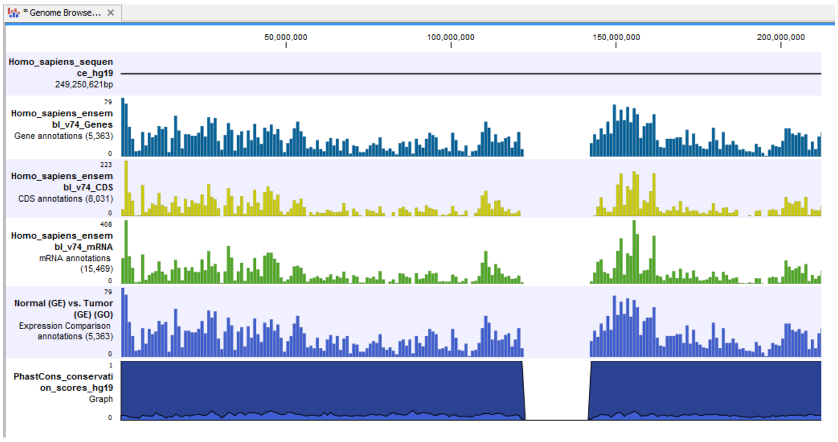

* Genome Browse...
50,000,000
100,000,000
150,000,000
200,000,000
Homo_sapiens_sequence_hg19
249,250,621bp
79
Homo_sapiens_ensembl_v74_Genes
Gene annotations (5,363)
0
223
Homo_sapiens_ensembl_v74_CDS
CDS annotations (8,031)
0
408
Homo_sapiens_ensembl_v74_mRNA
mRNA annotations (15,469)
0
79
Normal (GE) vs. Tumor (GE) (GO)
Expression Comparison annotations (5,363)
0
1
PhastCons_conservation_scores_hg19
Graph
0

Figure 15.38: *The genome browser view allows comparison of the expression comparison tracks with the reference sequence and different annotation tracks.*

Chapter 16

Using data from other workbenches

Contents

16.1 Open outputs from other workbenches

Please note that if you also have access to CLC Genomics Workbench, CLC Main Workbench, or CLC Sequence Viewer you may have generated different types of output that you would like to view in the *CLC Cancer Research Workbench*. All types of output that have been created in CLC Genomics Workbench, CLC Main Workbench, or CLC Sequence Viewer can be opened in the *CLC Cancer Research Workbench*. This means that you are capable of opening certain output types that cannot be generated from within the *CLC Cancer Research Workbench*. In such cases we refer to our other manuals e.g. the CLC Genomics Workbench manual that can be found here: `http://www.clcbio.com/support/downloads/#manuals` for further information about the output types that are not described in the *CLC Cancer Research Workbench* manual.

Output files from other workbenches can be imported as described in section 10.3.1 using **Standard Import**.

Part IV

CLC Genome Browser

Chapter 17

Genome browser tools

Contents

This chapter explains how to visualize tracks, how to retrieve reference data and finally how to perform generic comparisons between tracks.

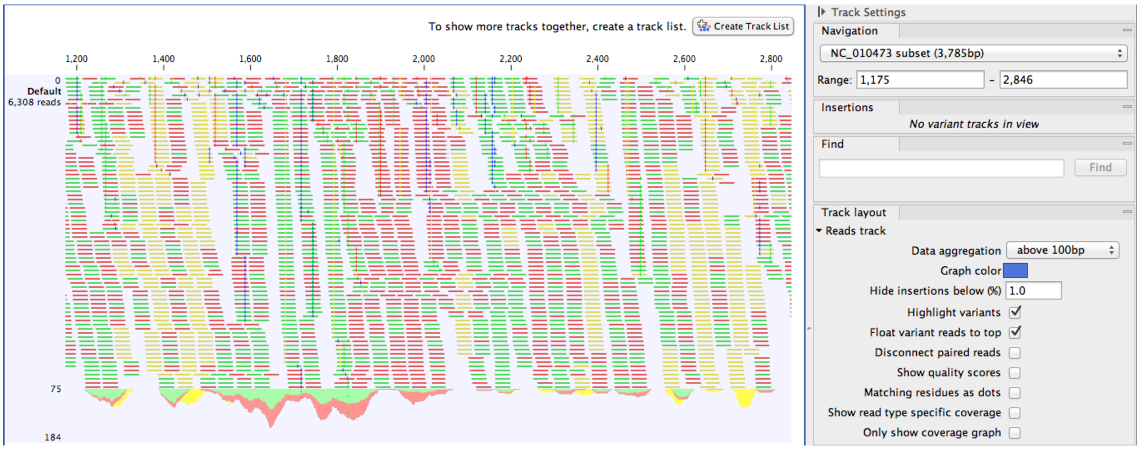
The genome browser is the graphical interface where tracks can be presented alone or together with other tracks. Tracks are the fundamental building blocks for data analysis in the *CLC Cancer Research Workbench* and provide a unified framework for the visualization, comparison and analysis of genome-scale studies.

In tracks, all information is tied to genomic positions. A central coordinate-system is provided by a reference genome, which allows that different types of data or results for different samples can be seen and analyzed together.

Different types of data are represented in different types of tracks, and each type of track has its own particular editors. An example of a single mapping read-track displaying reads and coverage is shown in figure 17.1.

The different track types in the *CLC Cancer Research Workbench* are:

A sequence () This is the track type that is used for holding the reference genome. The sequence track contains the single reference sequences of the genome (e.g. the chromosomes or the consensus sequences of de novo assembled contigs).



Track Settings

Navigation

NC_010473 subset (3,785bp)

Range: 1,175 – 2,846

Insertions

No variant tracks in view

Find

Find

Track layout

Reads track

Data aggregation: above 100bp

Graph color: 🟦

Hide insertions below (%): 1.0

Highlight variants ☑

Float variant reads to top ☑

Disconnect paired reads □

Show quality scores □

Matching residues as dots □

Show read type specific coverage □

Only show coverage graph □

Figure 17.1: *A single mapping read-track opened, displaying reads and coverage. On the top right, the button for creating a Track List is visible. On the right is the SidePanel.*

A reads track (☰) This is the track type that is used for holding a read mapping e.g. as produced by the Map Reads to Reference (see section 20.3) or Local Realignment (see section 20.7) tools. The reads track contains all the reads that have been mapped at their mapped positions, and you can zoom in all the way to base resolution.

A variant track (▸▸▸) A variant track is a particular kind of track that is used to store features that fulfill the requirements for being a variant. A particular requirement for being a variant is that it refers to a particular region of the reference, and it is possible to describe exactly how the sample "Allele" sequence looks in this region, as compared to what the "reference allele" sequence looks like in this region.

Variants may be of type SNV, MNV, replacement, insertion or deletion. A variant track may be produced either by running a Variant detection analysis (e.g using the Probabilistic or Quality-based variant callers or by importing a variant format file (such as a "vcf" or a "gvf" file) or downloading it from a database (e.g. COSMIC or dbSNP).

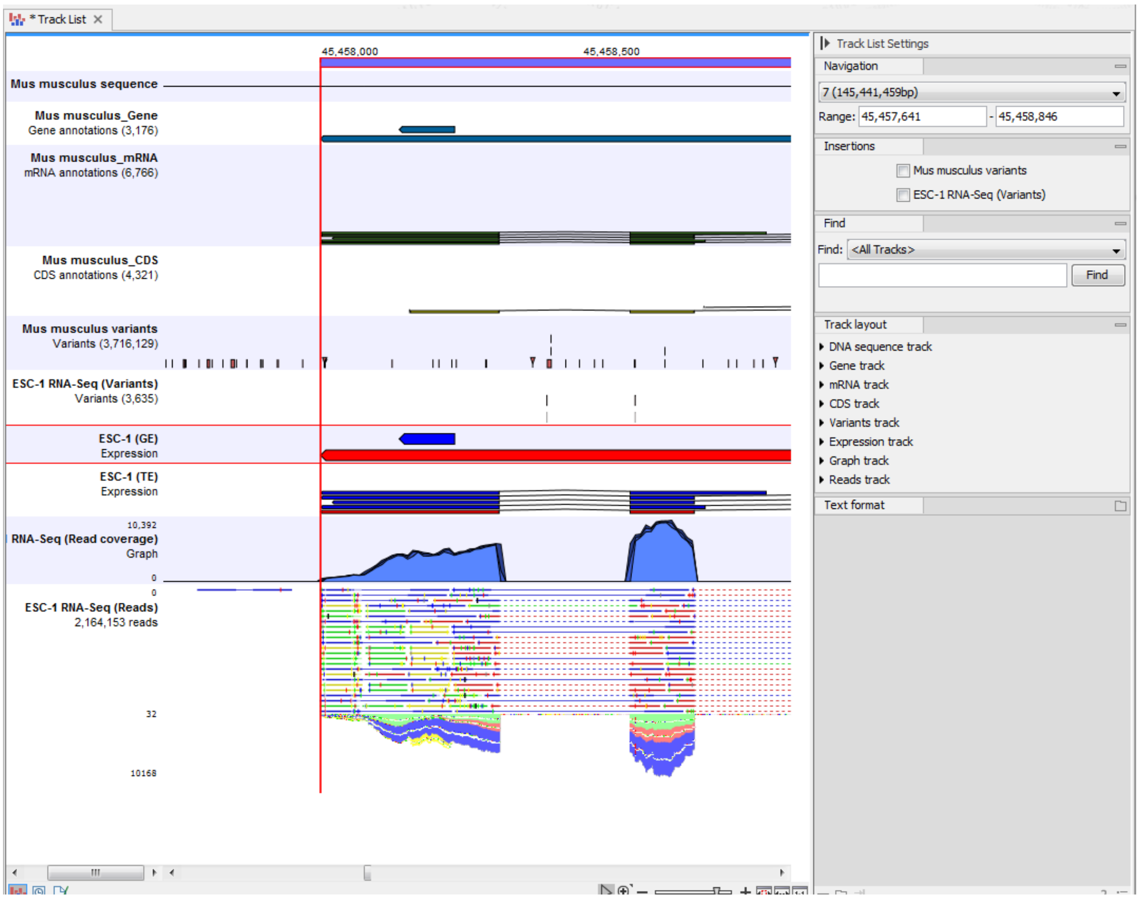
The tool InDels and Structural Variants (see section 20.17) detects structural variants, including insertions, deletions, inversions, translocations and tandem duplications. It will produce a variant track, which will contain some insertions and deletions (the "InDel" track). However, the tool will also detect some insertions for which the "Allele" sequence is not fully, but only partially, known. These insertions do not fulfill the requirements of being a variant and therefore cannot be put in the variant track. Instead they are put in the "SV track", along with the inversions and translocations. The "SV" track is an "annotation" (or "feature") track, which is less strict and more flexible, in the requirements to the types of annotations (or features) that it can contain (see below).

An annotation track (➔) Each annotation track contains a certain type of annotations. Examples are gene or mRNA tracks, which contain gene, respectively mRNA, annotations, UTR tracks, conservation score tracks and target region tracks. They may be obtained either by importing (see section 6.2 or downloading them into the Workbench (e.g from a .bed, .gtf or .gff file or a database, such as ENSEMBL). Also, many of the tools in the *CLC Cancer Research Workbench* will output annotation tracks. Examples are the Indels and Structural Variants tool, which will put the detected structural variants (that do not fulfill the requirements for being of type "variant") in an annotation track, or the ChIP-seq detection tool which will put the detected "peaks" into a "peak" annotation track.

A coverage graph ( ) The coverage graph track is calculated from a reds track and contains a graphical display of the coverage at each position in the reference.

An expression track ( ) The RNA-seq algorithm produces expression tracks; one for genes and one for transcripts. These are tracks that have an annotation for each gene, respectively transcript, *and* an expression value associated to that annotation.

An example of the different types of tracks is given in figure 17.2.



7 (145,441,459bp)

Range: 45,457,641 - 45,458,846

☐ Mus musculus variants

☐ ESC-1 RNA-Seq (Variants)

Find: <All Tracks>

Figure 17.2: *A tracklist containing different types of tracks. From the top: a sequence track, three annotation tracks with gene, mRNA and CDS annotations respectively, two variant tracks, a gene-level (GE) and a transcript level (TE) expression track, a coverage track and a reads track.*

17.1 Create new genome browser view

The tool **Create New Genome Browser View** ( ) can be used to create a list of tracks. Double-click on **Create New Genome Browser View** in the toolbox to run the tool:

Toolbox | **Genome Browser** ( ) | **Create New Genome Browser View** ( )

In the wizard (figure 17.3) you can select all the tracks that you would like to include in your enome browser view. Figure 17.4 shows an example of a genome browser view including a track with the genomic reference sequence at the top followed by the targeted regions, the mapped reads, and in the lower part of the figure a variant detection track.

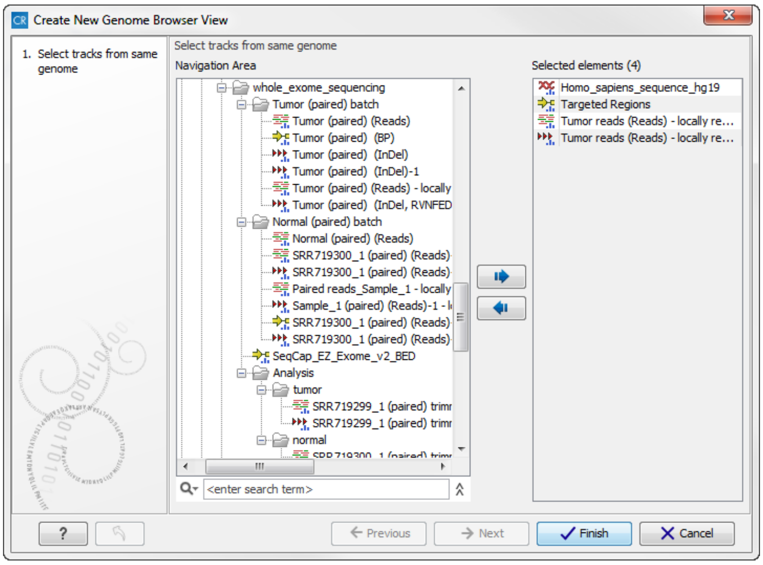

Create New Genome Browser View
1. Select tracks from same genome
Select tracks from same genome
Navigation Area
whole_exome_sequencing
Tumor (paired) batch
Tumor (paired) (Reads)
Tumor (paired) (BP)
Tumor (paired) (InDel)
Tumor (paired) (InDel)-1
Tumor (paired) (Reads) - locally
Tumor (paired) (InDel, RVNFED
Normal (paired) batch
Normal (paired) (Reads)
SRR719300_1 (paired) (Reads)
SRR719300_1 (paired) (Reads)
Paired reads_Sample_1 - locally
Sample_1 (paired) (Reads)-1 - l
SRR719300_1 (paired) (Reads)
SRR719300_1 (paired) (Reads)
SeqCap_EZ_Exome_v2_BED
Analysis
tumor
SRR719299_1 (paired) trim
SRR719299_1 (paired) trim
normal
<enter search term>
Selected elements (4)
Homo_sapiens_sequence_hg19
Targeted Regions
Tumor reads (Reads) - locally re...
Tumor reads (Reads) - locally re...
?
Previous
Next
Finish
Cancel

Figure 17.3: *Three tracks shown in the track list view*

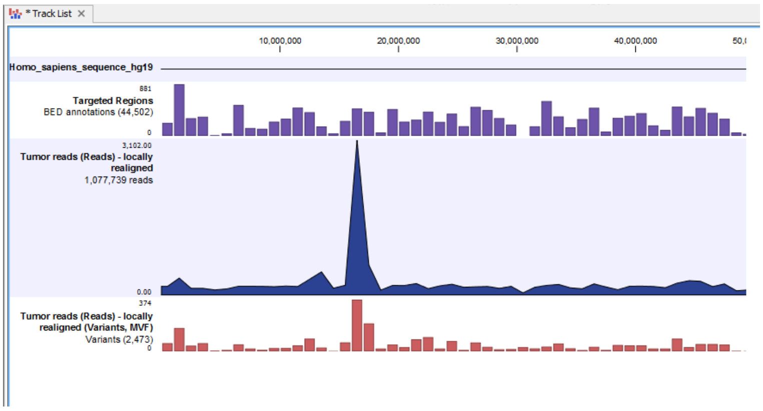


Figure 17.4: *Four tracks shown in the track list view.*

17.2 Genome browser view

For details on how to find and import different tracks see section 6.2. Tracks are saved as files in the **Navigation Area** with specific icons representing each track type, e.g. an annotation track ( ).

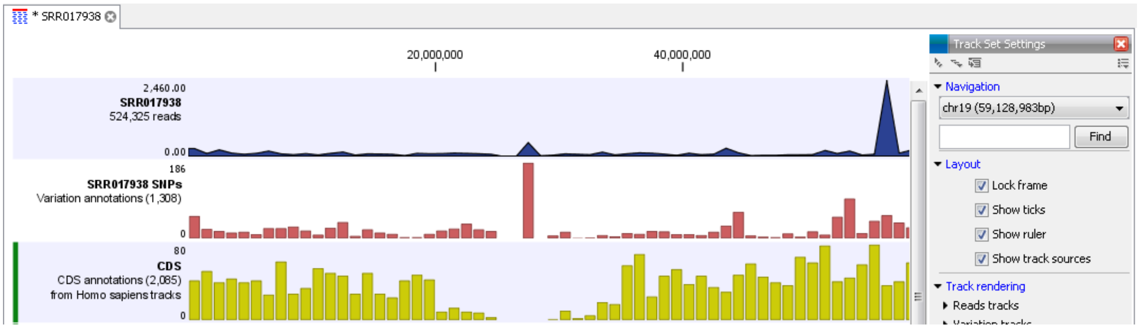
To visualize several tracks together, they can be combined into a **Genome Browser View** ( ). Genome browser views can be created in different ways. One way is via the menu bar:

File | **New** | **Create Genome Browser View ( )**

17.2.1 Zooming and navigating the genome browser views

It is possible to zoom in and out on the view shown in 17.5 with the zoom tools in the lower right-hand corner of the View Area, or by using a mouse scroll wheel while pressing the Ctrl (⌘ on Mac) key.

When zooming in and out you will see that, when zoomed out, the data is visualized in an aggregated format using a density bar plot or a graph. This allows you to navigate the view more smoothly and get an overview of e.g. how many SNPs are located in a certain region.



chr19 (59,128,983bp)

☑ Lock frame
☑ Show ticks
☑ Show ruler
☑ Show track sources

Figure 17.5: *Three tracks shown in the track list view*

In figure 17.6 we have zoomed in on a specific region with a read track at the top showing the individual reads and with CDS and SNP annotations shown below.

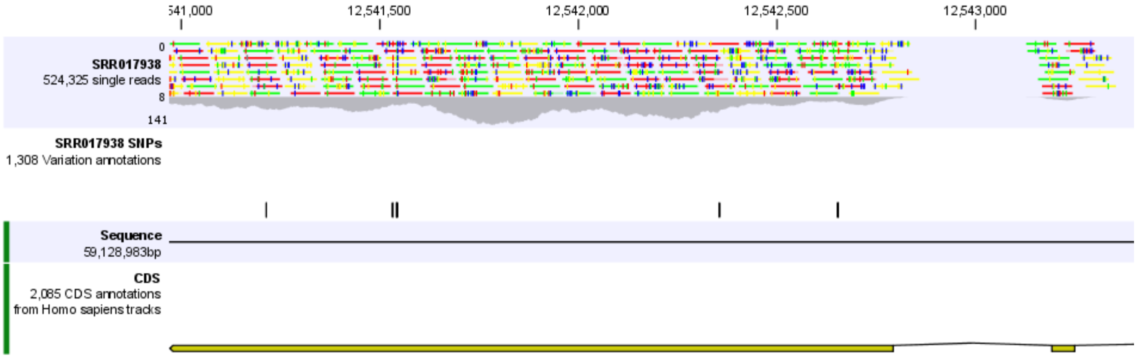


Figure 17.6: *Zooming in on the tracks reveals details*

If you zoom in further the alignment of the reads and the reference sequence can be viewed at single nucleotide level (see figure 17.7).

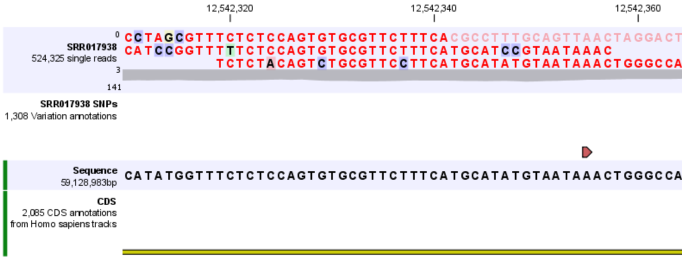

12,542,320
12,542,340
12,542,360
0
SRR017938
524,325 single reads
3
141
CCTAGCGTTTCTCTCCAGTGTGCGTTCTTTCACGCCTTTGCAGTTAACTAGGACT
CATCCGGTTTTTCTCCAGTGTGCGTTCTTTCATGCATCCGTAATAAAC
TCTCTACAGTCTGCGTTCCTTCATGCATATGTAATAAACTGGGCCA
SRR017938 SNPs
1,308 Variation annotations
Sequence
59,128,983bp
CATATGGTTTCTCTCCAGTGTGCGTTCTTTCATGCATATGTAATAAACTGGGCCA
CDS
2,085 CDS annotations
from Homo sapiens tracks

Figure 17.7: *Zoom in to see the bases of the reads and the reference sequence.*

In this case only three reads are visible. In order to see more reads, increase the height of the reads track by dragging down the lower part of the track with the mouse (Figure 17.8).

The options for the **Side Panel** vary depending on which track is shown in the View Area. In figure 17.9 an example is shown for a read mapping:

Navigation. Gives information about which chromosome is currently shown. Below this, you can see the start and end positions of the shown region of the chromosome. The drop-down list can be used to jump to a different chromosome. It is also possible to jump to a new position. This can be done by typing in the start and end positions in the text fields.

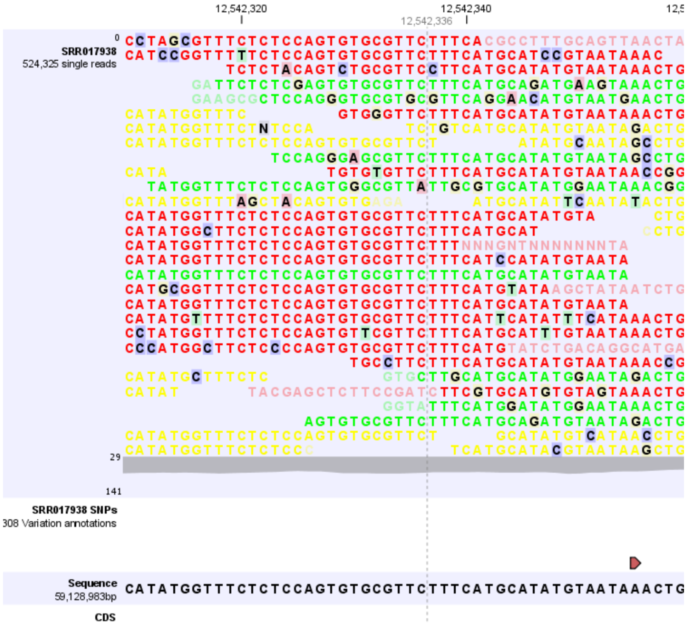


Figure 17.8: *Adjusting the height of the track.*

Thousands separators are supported. The selected region will automatically appear in the viewing area.

Insertions. Only relevant for variant tracks.

Find. Not relevant for reads tracks.

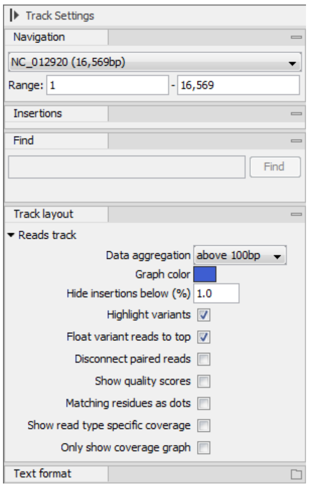
Track layout. The options for the Track layout varies depending on which track type is shown. The options for a read track are:

- **Data aggregation.** Allows you to specify whether the information in the track should be shown in detail or whether you wish to aggregate data. By aggregating data you decrease the detail level but increase the speed of the data display process, which is of particular interest when working with big data sets. The threshold (in bp) for when data should be aggregated can be specified with the drop-down box. The threshold describes the unit (or "bucket") size in base pairs, above which the data will start being aggregated. The bucket size depends on the track length and the zoom level. Hence, a data aggregation threshold with a low value will only show details when zoomed in, whereas a high vfalue means that you can see details even when zoomed out. Please note that when using the high values, it will take longer time to display the data on the screen. Figure 17.9 shows the options for a read track and an annotation track. The data aggregation settings can be adjusted for each displayed track type.
- **Graph color.** Makes it possible to change the graph color.
- **Hide insertions below (%).** Hides insertions where the percentage of reads containing insertions is below this value. To hide all insertions, set this value to 101.
- **Highlight variants.** Variants are highlighted
- **Float variant reads to top.** When checked, reads with variations will appear at the top of the view.
- **Disconnect paired reads.** Disconnects paired end reads.

- **Show quality scores.** Shows the quality score. Ticking this option makes it possible to adjust the colors of the residues based on their quality scores. A quality score of 20 is used as default and will show all residues with a quality score of 20 or below in a blue color. Residues with quality scores above 20 will have colors that correspond to the selected color code. In this case residues with high quality scores will be shown in reddish colors. Clicking once on the color bar makes it possible to adjust the colors. Double clicking on the slider makes it possible to adjust the quality score limits. In cases where no quality scores are available, blue (the color normally used for residues with a low quality score) is used as default color for such residues.
- **Matching residues as dots.** Replaces matching residues with dots, only variants are shown in letters.
- **Show read type specific coverage.** When enabled, the coverage graph that summarizes those reads that could not be explicitly shown is now replaced by one coverage graph for each read type found in the Reads track. This could for instance be used for easy and visual comparison of the strand specific coverage.
- **Only show coverage graph.** When enabled, only the coverage graph is shown and no reads are shown.

When working with other track types such as gene tracks, other options are available:

Labels. Controls where in relation to the annotation features the labels will be shown, i.e. **Flag** places a label at the beginning of the gene, and above the feature graphics as shown in figure 17.10.



NC_012920 (16,569bp)

Range: 1 - 16,569

Data aggregation: above 100bp

Graph color: 🟦

Hide insertions below (%): 1.0

Highlight variants ☑

Float variant reads to top ☑

Disconnect paired reads □

Show quality scores □

Matching residues as dots □

Show read type specific coverage □

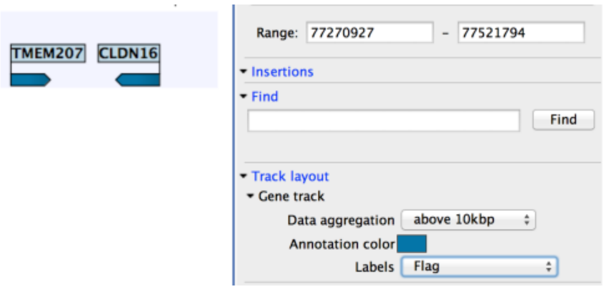
Only show coverage graph □

Figure 17.9: *The Side Panel for reads tracks.*

17.2.2 Adding, removing and reordering tracks

You can organize your tracks by dragging them up and down. Right-clicking on any of the tracks opens up a context menu with several options (Figure 17.11). The options shown in the context menu will vary depending on which tracks you have open in the viewing area. Hence, you may not be presented with all the options described here.

Create Mapping Graph Tracks This will allow you to create a new track from a mapping track (learn more in section 17.3).



Range: 77270927 - 77521794

Data aggregation: above 10kbp

Annotation color

Labels: Flag

Figure 17.10: *The Side Panel for annotation tracks.*

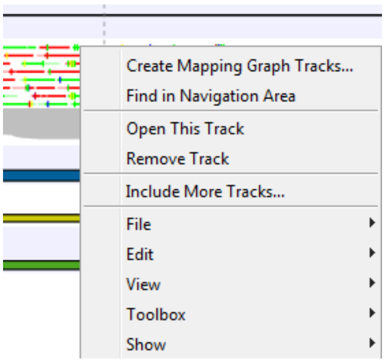


Figure 17.11: *Options to handle and organize tracks.*

Find in Navigation Area This will select the track in the **Navigation Area**.

Open This Track This opens a new view of the track. For annotation and variant tracks, a table view is opened as described in section 17.2.3. This can also be accomplished by double-clicking the track.

Remove Track This will remove the track from the current view. You can add it again by dragging it from the **Navigation Area** into the track list view or by pressing **Undo** (↶).

Include More Tracks This will allow you to add other track sets to your current track set. Please note that the information in the track will still be stored in its original track set. This means that you by including a track in this way at the same time is adding a reference to this track in another track set. An example of this could be the inclusion of a SNP track from another sample to your current analysis.

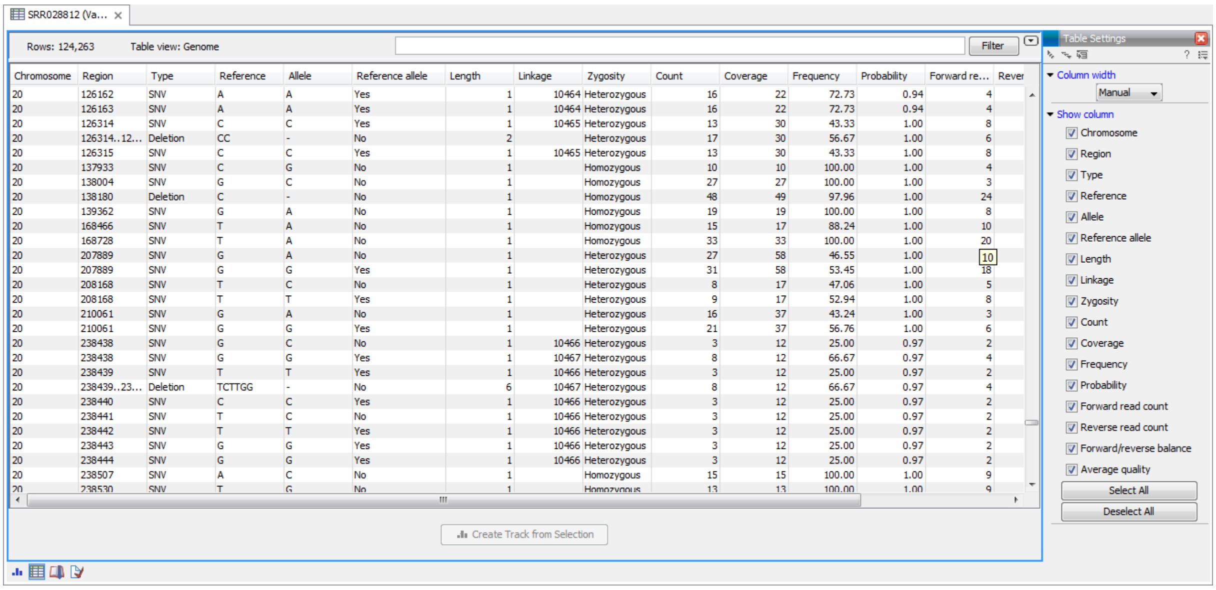
17.2.3 Showing a track in a table

All tracks containing annotations (including variants) can be opened in a table.

From the track list (see section 17.2) this is done either by double-clicking the label of the track or by right-clicking the track and choosing **Open This Track**. Alternatively, you can open the track from the **Navigation Area** and switch to the table view (▦) at the bottom.

The table will have one row for each annotation, and the columns will reflect its information content. Figure 17.12 shows an example of a variant database track that is presented in a table.

You can use the table to sort, filter and select annotations (see Appendix 8.3). Please note that there are two additional options for *filtering on overlaps* in the "Region" column



Rows: 124,263 Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference allele	Length	Linkage	Zygosity	Count	Coverage	Frequency	Probability	Forward re...	Rever
20	126162	SNV	A	A	Yes	1	10464	Heterozygous	16	22	72.73	0.94	4	
20	126163	SNV	A	A	Yes	1	10464	Heterozygous	16	22	72.73	0.94	4	
20	126314	SNV	C	C	Yes	1	10465	Heterozygous	13	30	43.33	1.00	8	
20	126314..12...	Deletion	CC	-	No	2		Heterozygous	17	30	56.67	1.00	6	
20	126315	SNV	C	C	Yes	1	10465	Heterozygous	13	30	43.33	1.00	8	
20	137933	SNV	C	G	No	1		Homozygous	10	10	100.00	1.00	4	
20	138004	SNV	G	C	No	1		Homozygous	27	27	100.00	1.00	3	
20	138180	Deletion	C	-	No	1		Homozygous	48	49	97.96	1.00	24	
20	139362	SNV	G	A	No	1		Homozygous	19	19	100.00	1.00	8	
20	168466	SNV	T	A	No	1		Homozygous	15	17	88.24	1.00	10	
20	168728	SNV	T	A	No	1		Homozygous	33	33	100.00	1.00	20	
20	207889	SNV	G	A	No	1		Heterozygous	27	58	46.55	1.00	10	
20	207889	SNV	G	G	Yes	1		Heterozygous	31	58	53.45	1.00	18	
20	208168	SNV	T	C	No	1		Heterozygous	8	17	47.06	1.00	5	
20	208168	SNV	T	T	Yes	1		Heterozygous	9	17	52.94	1.00	8	
20	210061	SNV	G	A	No	1		Heterozygous	16	37	43.24	1.00	3	
20	210061	SNV	G	G	Yes	1		Heterozygous	21	37	56.76	1.00	6	
20	238438	SNV	G	C	No	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238438	SNV	G	G	Yes	1	10467	Heterozygous	8	12	66.67	0.97	4	
20	238439	SNV	T	T	Yes	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238439..23...	Deletion	TCTTGG	-	No	6	10467	Heterozygous	8	12	66.67	0.97	4	
20	238440	SNV	C	C	Yes	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238441	SNV	T	C	No	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238442	SNV	T	T	Yes	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238443	SNV	G	G	Yes	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238444	SNV	G	G	Yes	1	10466	Heterozygous	3	12	25.00	0.97	2	
20	238507	SNV	A	C	No	1		Homozygous	15	15	100.00	1.00	9	
20	238530	SNV	T	G	No	1		Homozygous	13	13	100.00	1.00	9	

Table Settings

Column width: Manual

Show column
- ☑ Chromosome
- ☑ Region
- ☑ Type
- ☑ Reference
- ☑ Allele
- ☑ Reference allele
- ☑ Length
- ☑ Linkage
- ☑ Zygosity
- ☑ Count
- ☑ Coverage
- ☑ Frequency
- ☑ Probability
- ☑ Forward read count
- ☑ Reverse read count
- ☑ Forward/reverse balance
- ☑ Average quality

Select All
Deselect All

Figure 17.12: *Showing a variant track in a table view.*

When selecting a row in the table the graphical view will jump to this position on the genome. Please note that table filtering only affects the table. The track itself remains unaffected and keeps all annotations. If you also wish to filter tracks in the graphical view, the **Annotate and Filter** tools can be used instead.

At the bottom of the table a button labeled **Create Track from Selection** is available. This function can be used to create tracks showing only a subset of the data and annotations. Select the relevant rows in the table and click the button to create a new track that only includes the selected subset of the annotations. This function is particularly useful when used in combination with the filter.

17.2.4 Open track from a track list in table view

To open a table view of a track that is part of a track list, open the track list by double-clicking on the track name in the **Navigation Area**. The track will open in a graphical view. To open a single track from the track list in table view, either right-click on the track and choose "Open This Track" (see figure 17.13) or double-click on the name of the track you would like to open in table view (in the left side of the track when it is open in the **View Area**. This will automatically open op the specific track in table view.

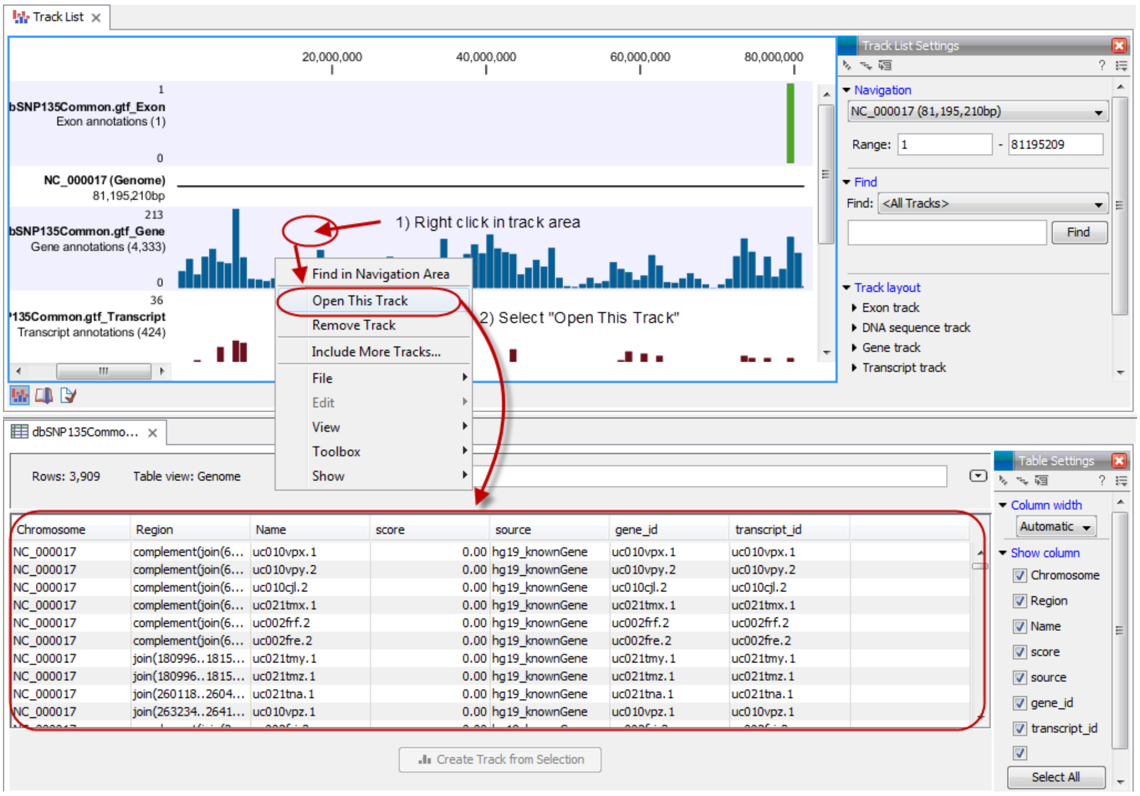
17.2.5 Finding annotations on the genome

In the **Side Panel** under **Find**, a search field allows you to quickly find the annotation that you are looking for. The list of tracks further allows you to restrict the search to a particular track (e.g. a gene track).

In the search field you can enter any kind of text that exists in the annotation track. As an example, consider the gene and tool tip shown in figure 17.14.

If you wish to locate this gene, any of the following entries could be typed in the search field:

BRCA2 This would match the annotation name exactly.



Navigation: NC_000017 (81,195,210bp)

Range: 1 - 81195209

Find: <All Tracks>

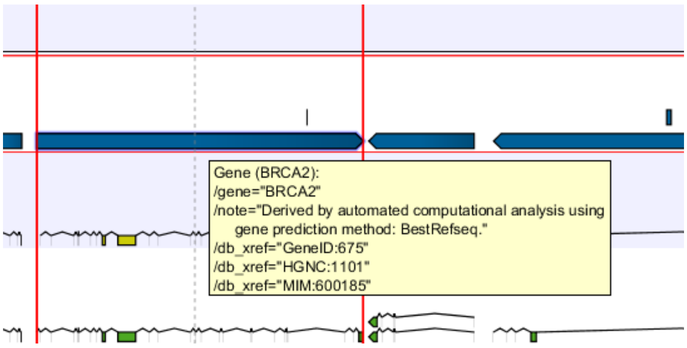
Rows: 3,909 Table view: Genome

Chromosome	Region	Name	score	source	gene_id	transcript_id
NC_000017	complement(join(6...	uc010vpx.1	0.00	hg19_knownGene	uc010vpx.1	uc010vpx.1
NC_000017	complement(join(6...	uc010vpy.2	0.00	hg19_knownGene	uc010vpy.2	uc010vpy.2
NC_000017	complement(join(6...	uc010cjl.2	0.00	hg19_knownGene	uc010cjl.2	uc010cjl.2
NC_000017	complement(join(6...	uc021tmx.1	0.00	hg19_knownGene	uc021tmx.1	uc021tmx.1
NC_000017	complement(join(6...	uc002frf.2	0.00	hg19_knownGene	uc002frf.2	uc002frf.2
NC_000017	complement(join(6...	uc002fre.2	0.00	hg19_knownGene	uc002fre.2	uc002fre.2
NC_000017	join(180996..1815...	uc021tmy.1	0.00	hg19_knownGene	uc021tmy.1	uc021tmy.1
NC_000017	join(180996..1815...	uc021tmz.1	0.00	hg19_knownGene	uc021tmz.1	uc021tmz.1
NC_000017	join(260118..2604...	uc021tna.1	0.00	hg19_knownGene	uc021tna.1	uc021tna.1
NC_000017	join(263234..2641...	uc010vpz.1	0.00	hg19_knownGene	uc010vpz.1	uc010vpz.1

Column width: Automatic

Show column: ☑ Chromosome ☑ Region ☑ Name ☑ score ☑ source ☑ gene_id ☑ transcript_id ☑

Figure 17.13: *One way to open a table view of a track that is part of a track list is to right click on the track of interest and select "Open This Table".*



Gene (BRCA2):
/gene="BRCA2"
/note="Derived by automated computational analysis using
gene prediction method: BestRefseq."
/db_xref="GeneID:675"
/db_xref="HGNC:1101"
/db_xref="MIM:600185"

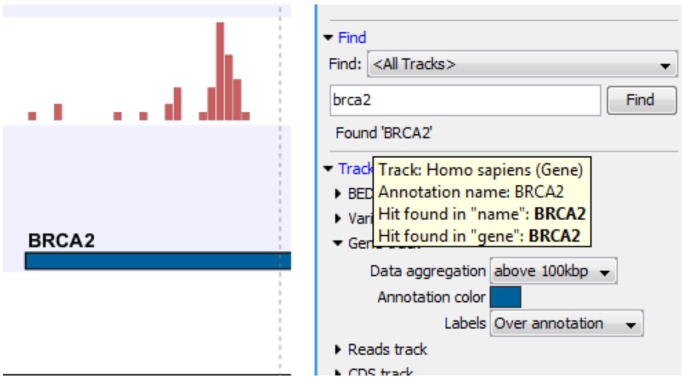
Figure 17.14: *The BRCA2 gene.*

BRCA* This would match the annotation name as well as other genes with a text starting with BRCA (e.g. the BRCA1 gene).

***RCA2** This would match the annotation name as well as other genes with a text ending with RCA2 (e.g. the SMARCA2 gene).

600185 This would match the `db_xref` qualifier for the OMIM database. All the text shown for the annotation in figure 17.14 can be searched this way, both as exact matches and with the * before or after the search term.

Just below the search field in the **Side Panel**, a status label informs about the progress of the search and the hit that has been found. Placing the mouse on top of the label will display a tool tip with more info (see 17.15).



Find: <All Tracks>

brca2

Found 'BRCA2'

Track: Homo sapiens (Gene)
Annotation name: BRCA2
Hit found in "name": **BRCA2**
Hit found in "gene": **BRCA2**

Data aggregation: above 100kbp
Annotation color
Labels: Over annotation

Figure 17.15: *The BRCA2 gene found.*

The search will be performed throughout the entire genome beginning with the chromosome currently shown and stopping when it finds the first hit. Press **Find** again to find the next hit. Once the whole genome has been traversed, the status will inform you that you have searched the whole genome. Click the **Find** button to start the search again.

Please note that you can also use the table view of an annotation track to perform more advanced queries of the data (see section 17.2.3).

17.2.6 Extract sequences from tracks

It is possible to extract sequences from tracks. The sequence of interest can be selected by dragging the mouse over the region of interest followed by a right click on the reads and a click on **Extract sequences** (figure 17.16).

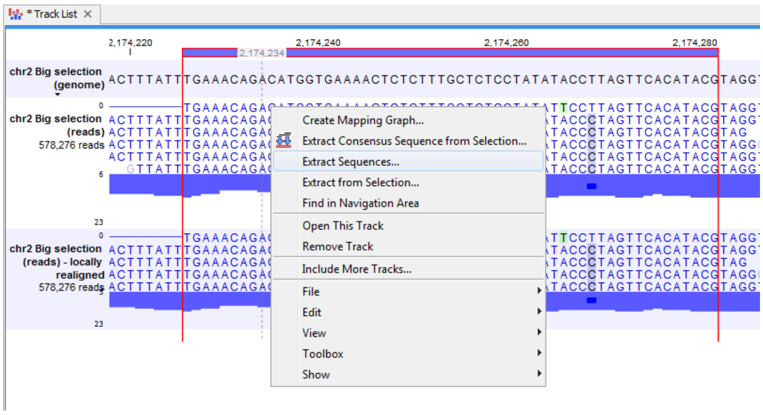
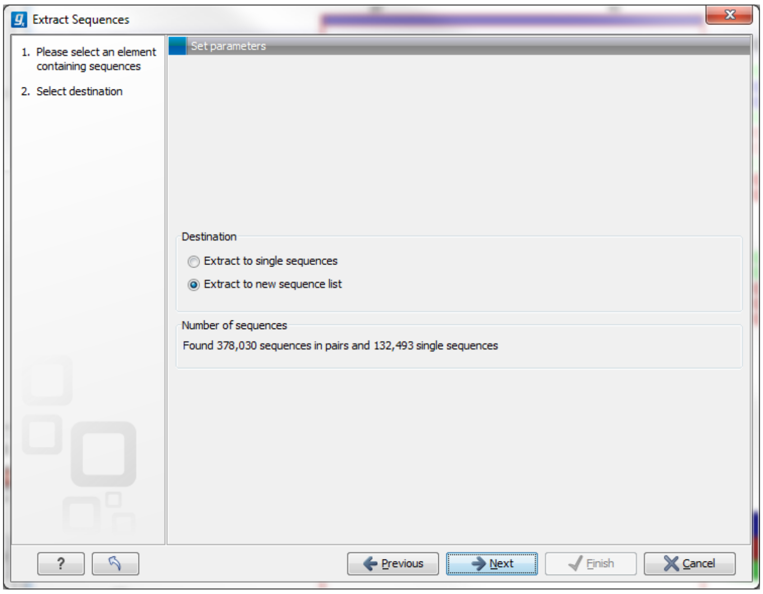

* Track List
2,174,220
2,174,234
2,174,240
2,174,260
2,174,280
chr2 Big selection (genome)
ACTTTATTTGAAACAGACATGGTGAAAACTCTCTTTGCTCCTATATACCTTAGTTCACATACGTAGG
chr2 Big selection (reads)
578,276 reads
chr2 Big selection (reads) - locally realigned
578,276 reads
0
6
23
0
23
Create Mapping Graph...
Extract Consensus Sequence from Selection...
Extract Sequences...
Extract from Selection...
Find in Navigation Area
Open This Track
Remove Track
Include More Tracks...
File
Edit
View
Toolbox
Show

Figure 17.16: *Extract sequences from tracks.*

This opens up the dialog shown in figure 17.17 that allows specification of whether the selected sequences should be extracted as single sequences or as a list of sequences.

Right clicking on the reads also enable the option **Extract from selection**, a function that corresponds to the **Extract from selection** described in section 29.7.5 although with small differences. Common for both versions of the **Extract from selection** function is that when extracting reads in an interval, only reads that are completely covered by the selection will be part of the extracted sequence, which in turn means that the tool can be used to extract only a subset of reads.

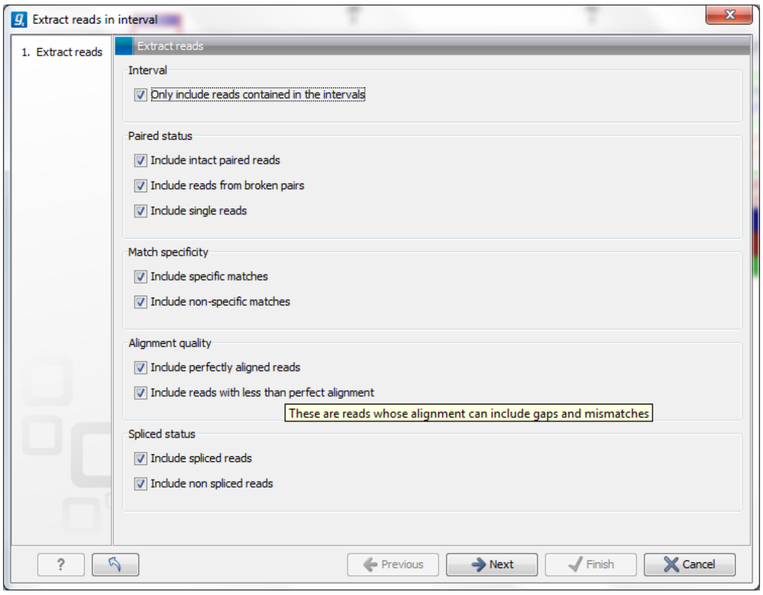
Clicking **Extract from selection** opens up the dialog shown in figure 17.18.


Extract Sequences
1. Please select an element containing sequences
2. Select destination
Set parameters
? Previous Next Finish Cancel

Destination
○ Extract to single sequences
● Extract to new sequence list

Number of sequences
Found 378,030 sequences in pairs and 132,493 single sequences

Figure 17.17: *Select destination for extracted sequences.*



Interval
☑ Only include reads contained in the intervals

Paired status
☑ Include intact paired reads
☑ Include reads from broken pairs
☑ Include single reads

Match specificity
☑ Include specific matches
☑ Include non-specific matches

Alignment quality
☑ Include perfectly aligned reads
☑ Include reads with less than perfect alignment
These are reads whose alignment can include gaps and mismatches

Spliced status
☑ Include spliced reads
☑ Include non spliced reads

Figure 17.18: *Select the reads to include.*

The purpose of this dialog is to let you specify which kinds of reads you wish to include. Per default all reads are included.

The options are:

Interval

Only include reads contained within the intervals Only reads that are included within the selection will be extracted. Reads that continue outside the selected area are not included.

Paired status

Include intact paired reads When paired reads are placed within the paired distance specified, they will fall into this category. Per default, these reads are colored in blue.

Include paired reads from broken pairs When a pair is broken, either because only one read in the pair matches, or because the distance or relative orientation is wrong, the reads are placed and colored as single reads, but you can still extract them by checking this box.

Include single reads This will include reads that are marked as single reads (as opposed to paired reads). Note that paired reads that have been broken during assembly are not included in this category. Single reads that come from trimming paired sequence lists are included in this category.

Match specificity

Include specific matches Reads that only are mapped to one position.

Include non-specific matches Reads that have multiple equally good alignments to the reference. These reads are colored yellow per default.

Alignment quality

Include perfectly aligned reads Reads where *the full read* is perfectly aligned to the reference sequence (or consensus sequence for de novo assemblies). Note that at the end of the contig, reads may extend beyond the contig (this is not visible unless you make a selection on the read and observe the position numbering in the status bar). Such reads are not considered perfectly aligned reads because they don't align in their entire length.

Include reads with less than perfect alignment Reads with mismatches, insertions or deletions, or with unaligned nucleotides at the ends (the faded part of a read).

Spliced status

Include spliced reads Reads that are across an intron.

Include non spliced reads Reads that are not across an intron.

17.2.7 Creating track lists in workflows

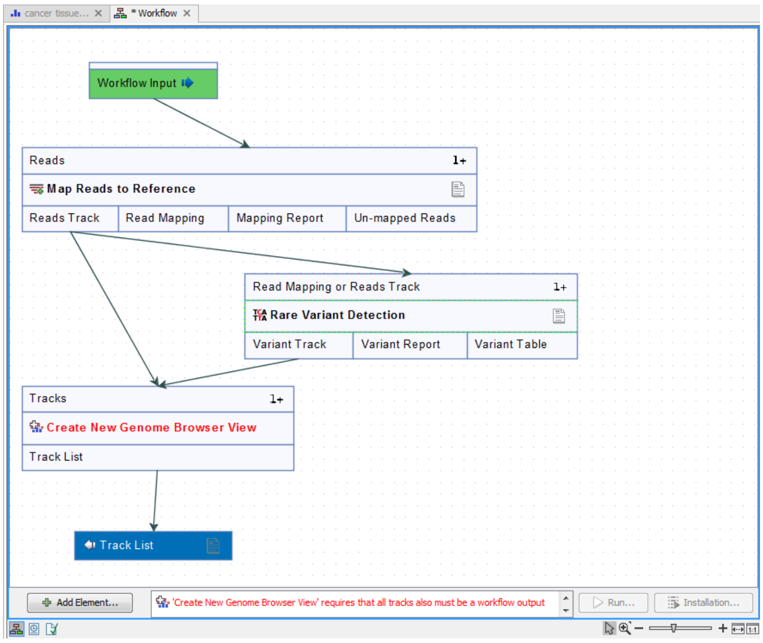
Track lists can be created as part of workflows. Track lists are different from all other workflow outputs in the sense that the tracks inside the track lists have to be saved separately, even if they are included in a track list.

Figure 17.19 shows an example where two tracks are fed into the **Create New Genome Browser View** element.

In this example, there is a warning at the bottom of the editor pointing at the fact that these two tracks need to be selected as output in order for the workflow to be validated. In figure 17.20, this has been corrected by selecting the tracks as output, and the workflow can now be executed.

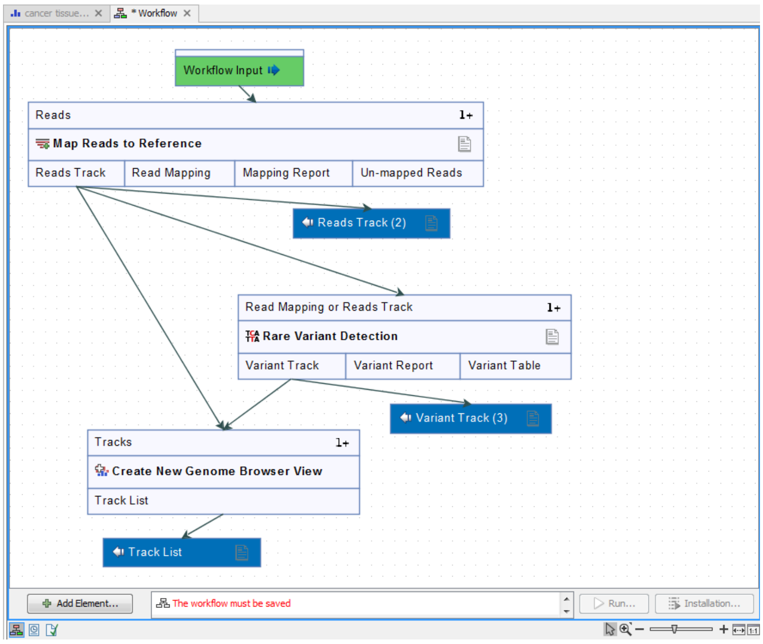
17.3 Creating graph tracks

Graphs can be a good way to quickly get an overview of certain types of information. This is the case for e.g. the GC content in a sequence or the read coverage. The *CLC Cancer Research*



'Create New Genome Browser View' requires that all tracks also must be a workflow output

Figure 17.19: *This workflow does not work because the two tracks need to be marked as output.*



The workflow must be saved

Figure 17.20: *The tracks have been selected as output, and the workflow can now be executed.*

Workbench offers two different tools that can create graph tracks from either a sequence or a read mapping. The two available tools are **Create GC Content Graph** and **Create Mapping Graph**. Both tools are found in the toolbox:

Toolbox | **Genome Browser** ( ) | **Graphs**

Graph tracks can also be created directly from the track view or track list view by right-clicking the track you wish to use as input, which will give access to the toolbox.

To understand what graph tracks are, we will look at an example. We will use the **Create GC Content Graph** tool to go into detail with one type of graph tracks.

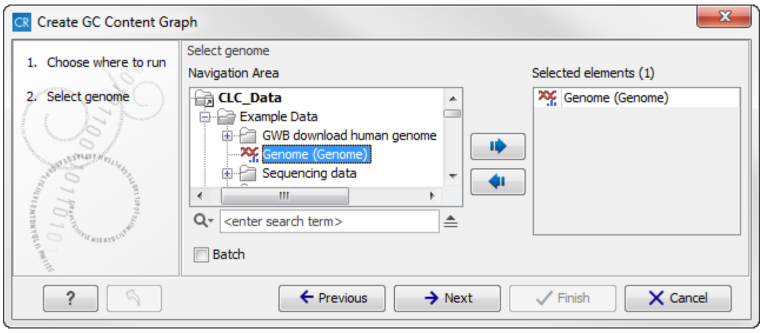
Create GC Content Graph

The **Create GC Content Graph** tool needs a sequence track as input and will create a graph track with the GC contents of that sequence.

To run the tool go to the toolbox:

Toolbox | **Genome Browser** ( ) | **Graphs** | **Create GC Content Graph**

Select the sequence track that should be used as input (see figure 17.21).


Create GC Content Graph

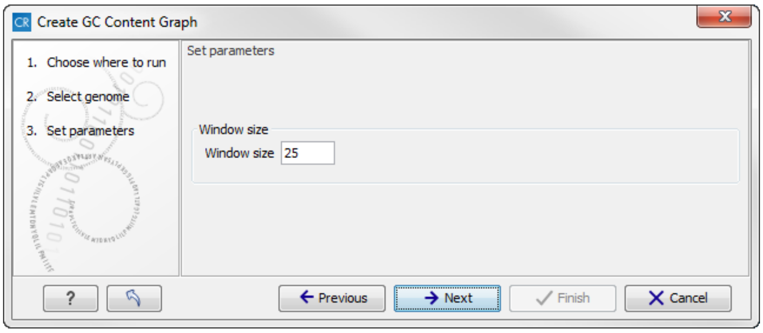
1. Choose where to run
2. Select genome

Select genome
Navigation Area
CLC_Data
Example Data
GWB download human genome
Genome (Genome)
Sequencing data
<enter search term>
□ Batch

Selected elements (1)
Genome (Genome)

Figure 17.21: *Select the sequence track that should be used as input.*

Click on the button labeled **Next** to go to the next wizard step (see figure 17.22). In this wizard step you can specify the window size, which is the size of the window around the central base in the region that is used to calculate the GC content. This number must be odd as you need a central base and an equal number of bases to each side of the central base. E.g. with a window size of 25, the GC content for the central base will be calculated based on the nucleotide composition in the central base and the 12 bases upstream and 12 bases downstream of the central base.


Create GC Content Graph

1. Choose where to run
2. Select genome
3. Set parameters

Set parameters
Window size
Window size 25

Figure 17.22: *Specify the window size. The window size is the region around each individual base that should be used to calculate the GC content in a given region.*

Click on the button labeled **Next**, choose to save your results, and click on the button labeled **Finish**. The output can be seen in figure 17.23. The output from "Create GC Content Graph" is a graph track. The graph track shows one value for each base with one graph being available for each chromosome. When zoomed out as shown in this figure, three different graphs with three different colors can be seen. The top graph with the darkest blue color represents the maximum observed GC content values in the specific region, the graph in the middle with the intermediate blue color shows the mean observed GC content values in the specific region, and the graph at the bottom with the light blue color shows the minimum observed GC content values in the specific region.

When zooming all the way in to single nucleotide level only one graph can now be seen as you ar

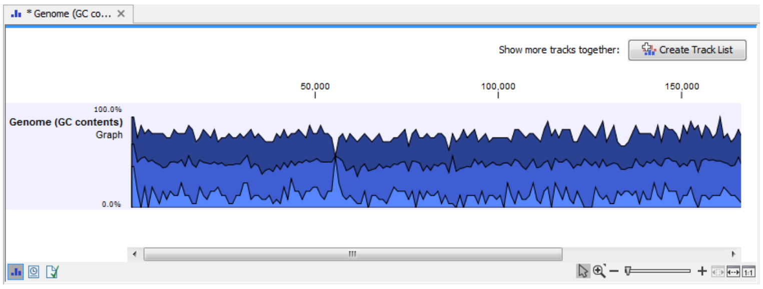


Figure 17.23: *The output from "Create GC Content Graph" is a graph track. The graph track shows one value for each base with one graph being available for each chromosome.*

now no longer looking at large genomic regions. In stead, you can now use the tooltip by mousing over each individual base to look at the GC content for that particular base and the number of bases that you specified as the window size to be used. This is shown in figure 17.24 where the top part of the figure shows the graph track when zoomed all the way out and the bottom part of the figure shows a genome browser view with the sequence track that was used as input together with the output graph track. The input and the output tracks were combined in one view as a track list (see section 17.2) by clicking on the button labeled **Create Genome Browser View** found in the upper right corner of the graph track in the top part of the figure (see the red arrow).

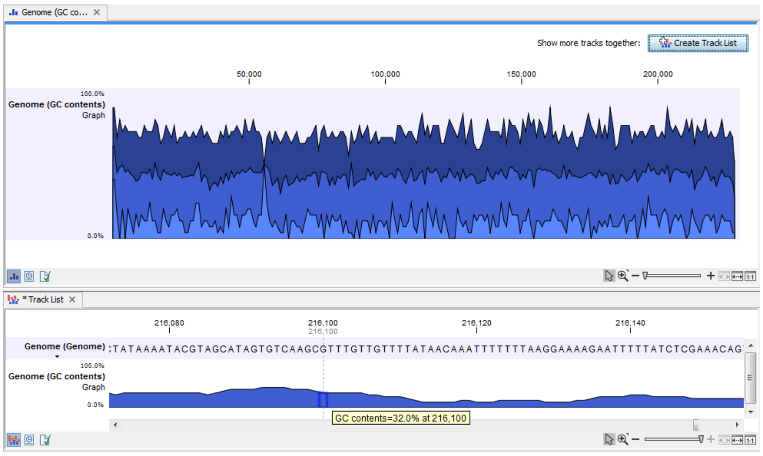


Figure 17.24: *The top part of the figure shows the graph track when zoomed all the way out. The bottom part of the window shows a graph track together with the input genomic sequence at single nucleotide resolution. By mousing over one nucleotide, you can see the GC content for this position. In our example we chose a window size of 25 nucleotides and the GC content that is shown for one nucleotide is the GC content for the central nucleotide and the 12 bases upstream and downstream of this nucleotide.*

This track can then be displayed together with the sequence and other tracks in a genome browser view.

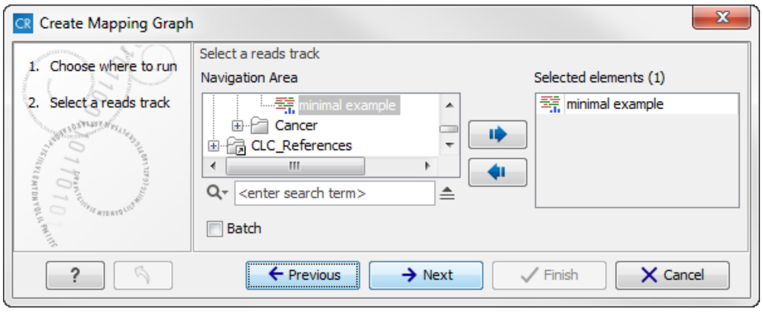
Create Mapping Graph

The **Create Mapping Graph** tool can create a range of different graphs from a read mapping track. To run the tool go to the toolbox:

Toolbox | **Genome Browser** ( ) | **Graphs** | **Create Mapping Graph**

Select the read mapping as shown in figure 17.25 and click on the button labeled **Next**.

Select the graph tracks that you would like to create. The following options exist:



Create Mapping Graph

1. Choose where to run
2. Select a reads track

Select a reads track

Navigation Area

- minimal example
- Cancer
- CLC_References

<enter search term>

□ Batch

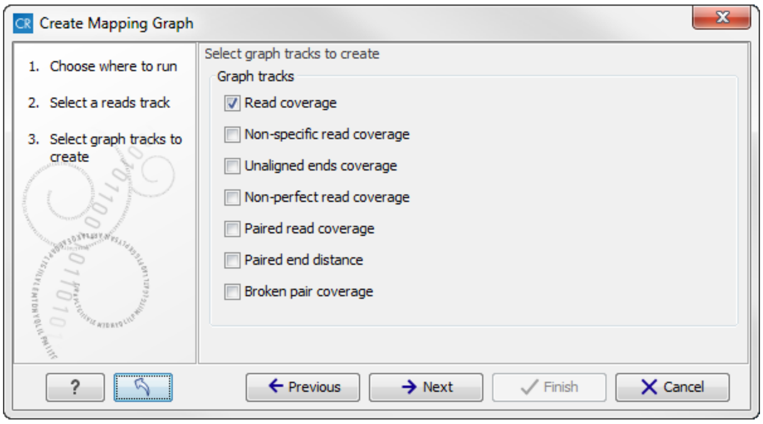
Selected elements (1)

- minimal example

Figure 17.25: *Creating graph track from mappings.*

- Read coverage. For each position this graph shows the number of reads contributing to the alignment (see a more elaborate definition in section 18.2.2).
- Non-specific read coverage. Non-specific reads are reads that would fit equally well other places in the reference genome.
- Unaligned ends coverage. Un-aligned ends arise when a read has been locally aligned to a reference sequence, and then end of the read is left unaligned because there are mismatches or gaps relative to the reference sequence. This part of the read does not contribute to the read coverage above. The unaligned ends coverage graph shows how many reads that have unaligned ends at each position.
- Non-perfect read coverage. Non-perfect reads are reads with one or more mismatches or gaps relative to the reference sequence.
- Paired read coverage. This lists the coverage of intact pairs. If there are no single reads and no pairs are broken, it will be the same as the standard read coverage above.
- Broken pair coverage. A pair is broken either because only one read in the pair matches, or because the distance or relative orientation between the reads is wrong.
- Paired end distance. Displays the average distance between the forward and the reverse read in a pair. A pair contributes to this graph from the beginning of the first read to the end of the second read.

One graph track output will be created for each of the graph tracks you have chosen by checking the boxes shown in figure 17.26.



Create Mapping Graph

1. Choose where to run
2. Select a reads track
3. Select graph tracks to create

Select graph tracks to create

Graph tracks

☑ Read coverage
□ Non-specific read coverage
□ Unaligned ends coverage
□ Non-perfect read coverage
□ Paired read coverage
□ Paired end distance
□ Broken pair coverage

Figure 17.26: *Choose the types of graph tracks you would like to generate.*

Click on the button labeled **Next**, choose where to save the generated output(s) and click on the button labeled **Finish**.

An example of three different outputs is shown in figure 17.27. Two of the views have been dragged and dropped to other areas of the **View Area** to be able to see them in the same window. If you would like to learn more about how to do this, please refer to section 2.1.6.

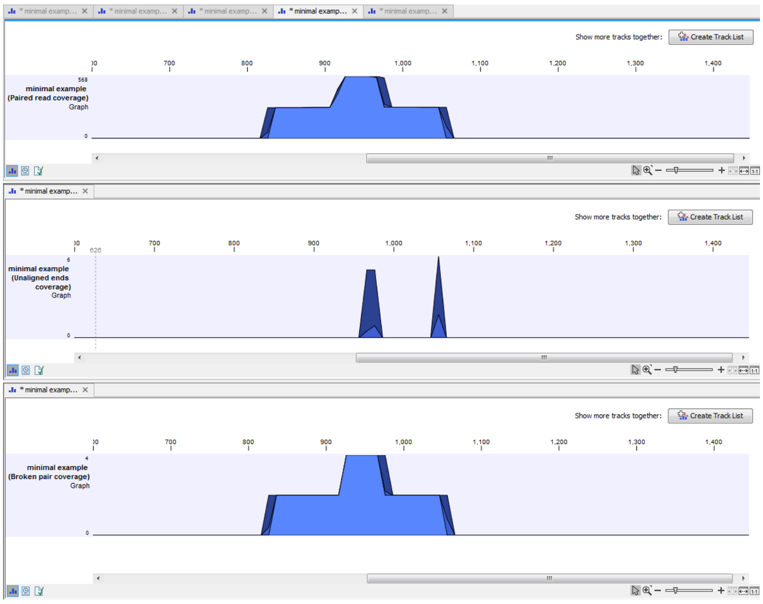


Figure 17.27: *Three types of graph tracks are shown.*

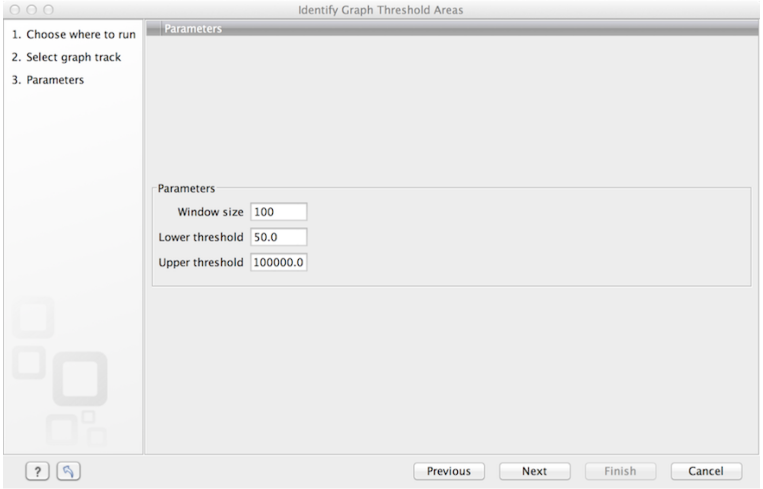
In this example we generated all possible outputs and chose to open them without saving. You can see that the names of the tabs are marked with an asterisk, which indicates that the graph shown in the view area has not been saved or that changes have been made that must be saved if you want to keep them. Three of the generated outputs have been opened. If you would like to see the outputs in the same view, you can do this by creating a genome browser view. Click on the button labeled **Create genome Browser View** in the upper right corner of each of the graph tracks shown in the **View Area**. Combining graph tracks in a genome browser view links the individual tracks together, which makes it much easier to compare the different graph tracks.

Another thing that is worth noting is the options found in the **Side Panel**. In particular the option "Fix graph bounds" found under **Track layout** is useful to know if you would like to manually adjust the numbers on the y-axis.

Identify Graph Threshold Areas

The **Identify Graph Threshold Areas** tool uses graph tracks as input to identify graph regions that fall within certain limits (thresholds). Both a lower and an upper threshold can be specified to create an annotation track for those regions of a graph track where the values are in the given range (see figure 17.28). Consequently, in order to identify only those parts of the track that exceed a certain minimum, one would choose the minimum threshold and set the upper limit to a value well above the maximum occurring in the track (and vice versa for finding ranges that are below a maximum threshold). Obviously, the range chosen for the lower and upper thresholds will depend on the data (coverage, quality etc.).

The "window-size" parameter specifies the width of the window around every position that is used to calculate an average value for that position and hence "smoothes" the graph track beforehand. A window size of 1 will simply use the value present at every individual position and determine if it is within the upper and lower threshold, hence resulting in the same "non-smoothing" behavior as previous versions of the workbench without this parameter. In contrast, a window size of 100 checks if the average value derived from the surrounding 100 positions falls between the minimum and maximum threshold. Such larger windows help to prevent "jumps" in the graph track from fragmenting the output intervals or help to detect over-represented regions in the track that are only visible when looked at in the context of larger intervals and lower resolution. An example output is shown in figure 17.29 where the coverage graph has a couple of local minima near zero. However, by using the averaging window, the tool is able to produce a single unbroken annotation covering the entire region. Of course larger window sizes result in regions that are broader and hence their boundaries are less likely to exactly coincide with the borders of visually recognizable borders of regions in the track.



Parameters

Window size 100

Lower threshold 50.0

Upper threshold 100000.0

Figure 17.28: *Specification of lower and upper thresholds.*

When zoomed out, the graph tracks are composed of three curves showing the maximum, mean, and minimum value observed in a given region (see figure 17.29). When zoomed in all the way down to base resolution only one curve will be shown reflecting the exact observation at each individual position.

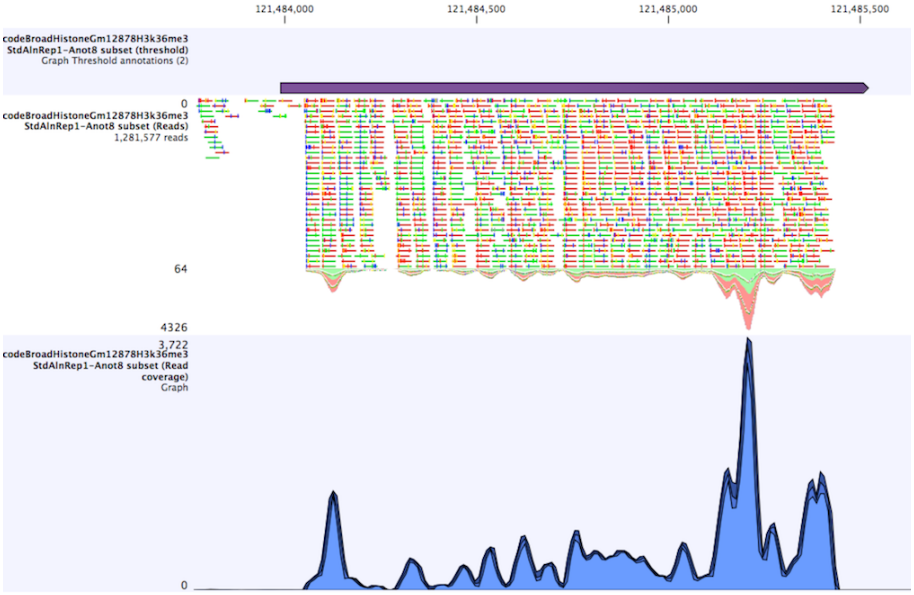

121,484,000
121,484,500
121,485,000
121,485,500
codeBroadHistoneGm12878H3k36me3
StdAlnRep1-Anot8 subset (threshold)
Graph Threshold annotations (2)
0
codeBroadHistoneGm12878H3k36me3
StdAlnRep1-Anot8 subset (Reads)
1,281,577 reads
64
4326
3,722
codeBroadHistoneGm12878H3k36me3
StdAlnRep1-Anot8 subset (Read coverage)
Graph
0

Figure 17.29: *Track list including a region identified by the parameters set above on a dataset of H3K36 methylation from ENCODE. The top track shows the resulting region. Below is the track containing the reads. The graph track at the bottom shows the coverage with the minimum, mean, and maximum observed values.*

Part V

Initial data handling

Chapter 18

Quality control tools

Contents

18.1 QC for Target Sequencing

This tool is designed to report the performance (enrichment and specificity) of a targeted re-sequencing experiment. Targeted re-sequencing is due to its low costs, very popular and several companies provide platforms and protocols (learn more at http://en.wikipedia.org/wiki/Exome_sequencing#Target-enrichment_strategies). Array-based approaches are offered by e.g. Agilent (SureSelect) and Roche Nimblegen. Furthermore, amplicon sequencing with PCR primers is offered by RainDance, Fluidigm and others.

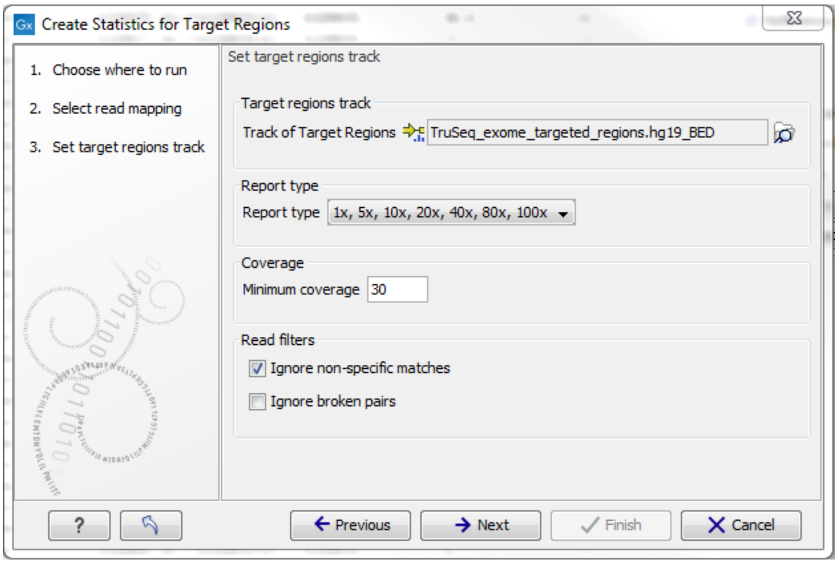
Given an annotation track with the target regions (e.g. imported from a bed file), this tool will investigate a read mapping to determine whether the targeted regions have been appropriately covered by sequencing reads as well as information about how specific the reads map to the targeted regions. The results are provided both as a summary report and as track or table with detailed information about each targeted region.

18.1.1 Running the "QC for Target Sequencing" tool

The tool is found in the Toolbox:

Toolbox | **Quality Control** ( ) | **QC for Target Sequencing** ( )

This opens a wizard where you can select mapping results ( )/ ( )/ ( ). Clicking **Next** will take you to the wizard shown in figure 18.1.



Create Statistics for Target Regions

1. Choose where to run
2. Select read mapping
3. Set target regions track

Set target regions track

Target regions track

Track of Target Regions: TruSeq_exome_targeted_regions.hg19_BED

Report type

Report type: 1x, 5x, 10x, 20x, 40x, 80x, 100x

Coverage

Minimum coverage: 30

Read filters

☑ Ignore non-specific matches

☐ Ignore broken pairs

Figure 18.1: *Specifying the track of target regions.*

Click the **Browse** ( ) icon to select an annotation track that defines the targeted regions of your reference genome.

The **Report type** allows you to select different sets of predefined coverage thresholds to use for reporting (see below). Furthermore, you will be asked to provide a **Minimum coverage** threshold. This will be used to provide the length of each target region that has at least this coverage.

Finally, you are asked to specify whether you want to **Ignore non-specific matches** and **Ignore broken pairs**. When these are applied reads that are non-specifically mapped or belong to broken pairs will be ignored.

Click **Next** to specify the type of output you want (see figure 18.2).

There are three options:

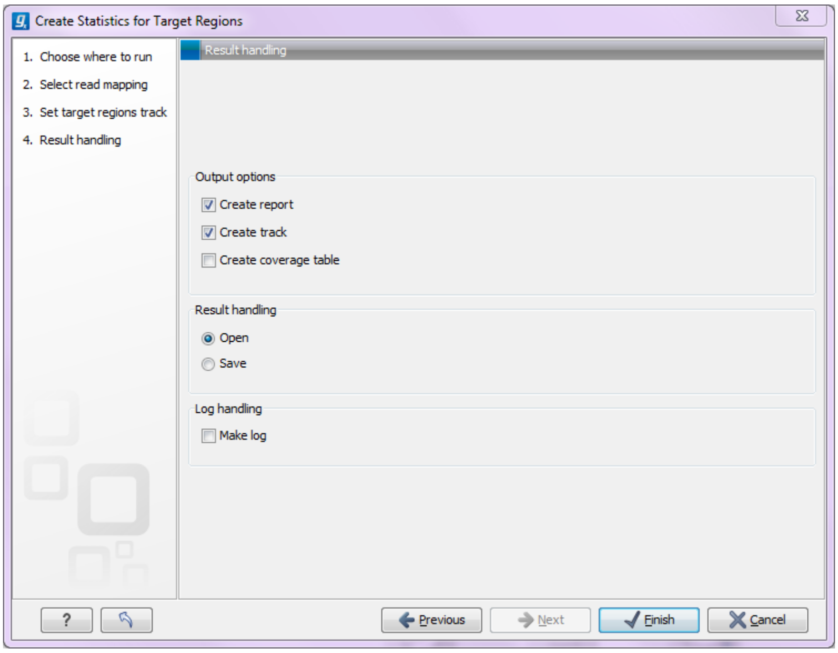
- The report gives an overview of the whole data set as explained in section 18.1.2.
- The track gives information on coverage for each target region as described in section 18.1.3.
- The coverage table outputs coverage for each position in all the targets as described in section 18.1.4.

Click **Finish** to create the reports.

18.1.2 Coverage summary report

An example of a coverage report is shown in figure 18.3).

This figure shows only the top of the report. The full content is explained below:



Create Statistics for Target Regions

1. Choose where to run
2. Select read mapping
3. Set target regions track
4. Result handling

Result handling

Output options

☑ Create report

☑ Create track

☐ Create coverage table

Result handling

● Open

○ Save

Log handling

☐ Make log

Figure 18.2: *Specifying how the result should be reported.*

Coverage summary This table shows overall coverage information.

Number target regions The number of targeted regions.

Total length of target regions The sum of the size of all the targeted regions (this means it is calculated from the annotations alone and is not influenced by the reads).

Average coverage For each position in each target region the coverage is calculated, and stored (you can see the individual coverages in the **Coverage table** output, figure 18.8). The 'average coverage' is calculated by taking the mean of all the calculated coverages in all the positions in all target regions. Note that if the user has chosen **Ignore non-specific matches** or **Ignore broken pairs** these reads will not contribute to the coverage. Note also that bases in over-lapping paired reads will only be counted as 1.

Number of target regions with low coverage The number of target regions which have positions with a coverage that is below the user-specified **Minimum coverage** threshold.

Total length of target regions with low coverage The total length of these regions.

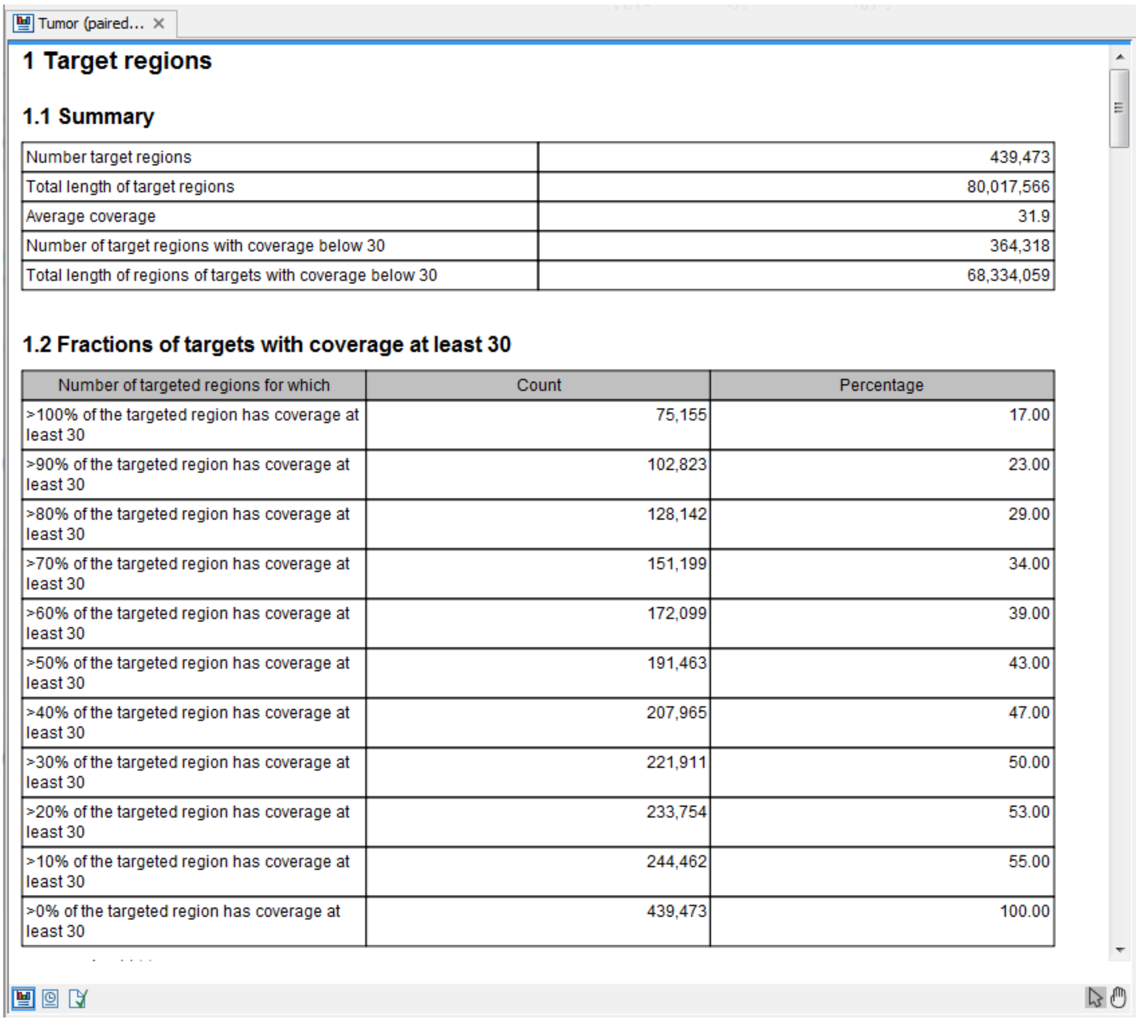
Fractions of targets with coverage at least... This table shows how many target regions have a certain percentage of the region above the user-specified **Minimum coverage** threshold.

Fractions of targets with coverage at least... A histogram presentation of the table above in **Fractions of targets with coverage at least...**.

Coverage of target regions positions This plot shows the coverage level on the x axis, and the number of positions in the target regions with that coverage level.

Coverage of target regions positions A version of the histogram above zoomed in to the values that lie +- 3SDs from the median.

Minimum coverage of target regions This shows the percentage of the targeted regions that are covered by this many bases. The intervals can be specified in the dialog when running the



1 Target regions

1.1 Summary

Number target regions	439,473
Total length of target regions	80,017,566
Average coverage	31.9
Number of target regions with coverage below 30	364,318
Total length of regions of targets with coverage below 30	68,334,059

1.2 Fractions of targets with coverage at least 30

Number of targeted regions for which	Count	Percentage
>100% of the targeted region has coverage at least 30	75,155	17.00
>90% of the targeted region has coverage at least 30	102,823	23.00
>80% of the targeted region has coverage at least 30	128,142	29.00
>70% of the targeted region has coverage at least 30	151,199	34.00
>60% of the targeted region has coverage at least 30	172,099	39.00
>50% of the targeted region has coverage at least 30	191,463	43.00
>40% of the targeted region has coverage at least 30	207,965	47.00
>30% of the targeted region has coverage at least 30	221,911	50.00
>20% of the targeted region has coverage at least 30	233,754	53.00
>10% of the targeted region has coverage at least 30	244,462	55.00
>0% of the targeted region has coverage at least 30	439,473	100.00

Figure 18.3: *The report with overviews of mapped reads.*

analysis. Default is 1, 5, 10, 20, 40, 80, 100 times. In figure 18.4 this means that 26.58 % of the positions on the target are covered by at least 40 bases.

Targeted regions overview This section contains two tables: one that summarizes, for each reference sequence, information relating to the *reads* mapped, and one that summarizes, for each reference, information relating to the *bases* mapped (figures 18.4 and 18.5). Note that, for the table that is concerned with *reads*, reads in over-lapping pairs are counted individually. Also note that, for the table that is concerned with *bases*, bases in overlapping paired reads are counted only as one (Examples are given in figures 18.6 and figure 18.7).

Reference The name of the reference sequence.

Total mapped reads The total number of mapped reads on the reference, including reads mapped outside the target regions.

Mapped reads in targeted region Total number of reads in the targeted regions. Note that if there are overlapping regions, reads covered by two regions will be counted twice. If a read is only partially inside a targeted region, it will still count as a full read.

Specificity The percentage of the total mapped reads that are in the targeted regions.

Total mapped reads excl ingored The total number of mapped reads on the reference, including reads mapped outside the target regions, excluding the non-specific matches or broken pairs, if the user has switched on the option to ignore those.

Mapped reads in targeted region excl ingored Total number of reads in the targeted regions, excluding the non-specific matches or broken pairs, if the user has switched on the option to ignore those.

Specificity excl ingored The percentage of the total mapped reads that are in the targeted regions.

Reference The name of the reference sequence.

Total mapped bases The total number of mapped bases on the reference, including bases mapped outside the target regions.

Mapped bases in targeted region Total number of bases mapped within in the targeted regions. Note that if there are overlapping regions, bases included in two regions will be counted twice.

Specificity The percentage of the total mapped bases that are in the targeted regions.

Total mapped bases excl ingored The total number of mapped bases on the reference, including bases mapped outside the target regions, excluding the bases in non-specific matches or broken pairs, if the user has switched on the option to ignore those.

Mapped bases in targeted region excl ingored Total number of bases in the targeted regions, excluding the bases in non-specific matches or broken pairs, if the user has switched on the option to ignore those.

Specificity excl ingored The percentage of the total mapped bases that are in the targeted regions.

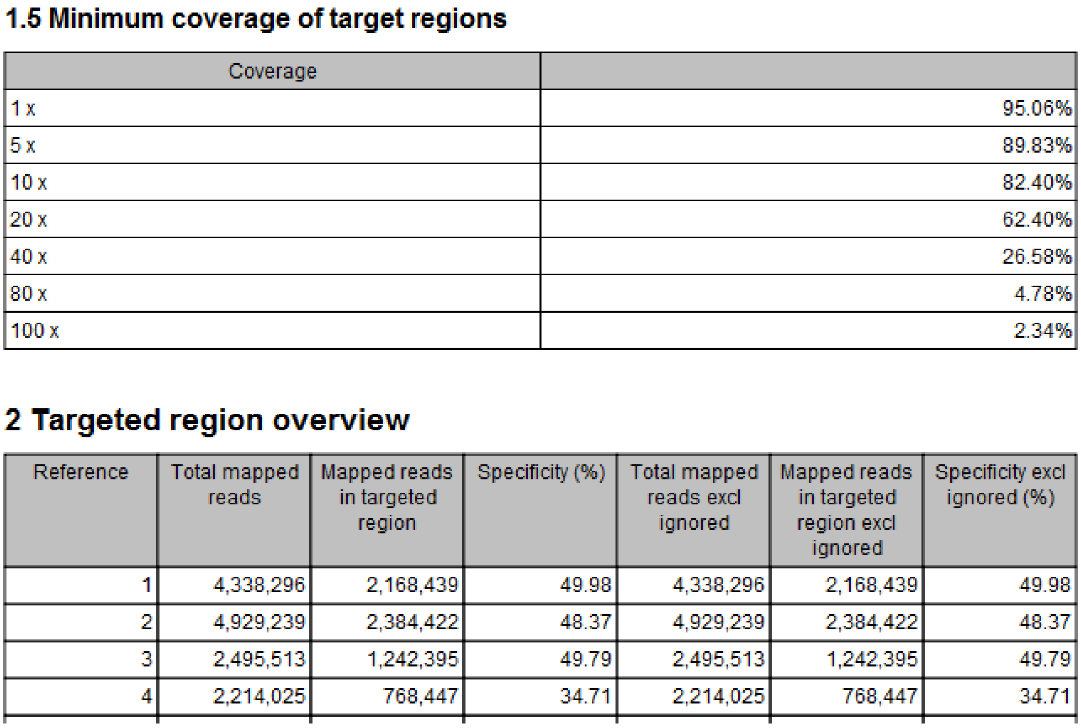
Distribution of target region length A plot of the length of the target regions, and a version of the plot where only the target region lengths that lie within +3SDs of the median target length are shown.

Base coverage The percentage of base positions in the target regions that are covered by respectively 0.1, 0.2, 0.3, 0.4, 0.5 and 1.0 times the mean coverage, where the mean coverage is the **average coverage** given in table 1.1. Because this is based on mean coverage, the numbers can be used for cross-sample comparison of the quality of the experiment.

Base coverage plot A plot showing the relationship between fold mean coverage and the number of positions. This is a graphical representation of the **Base coverage** table above.

Mean coverage per target position Three plots listing the mean coverage for each position of the targeted regions. The first plot shows coverage across the whole target, using a percentage of the target length on the x axis (to make it possible to have targets with different lengths in the same plot). This is reported for reverse and forward reads as well. In addition, there are two plots showing the same but with base positions on the x axis counting from the start and end of the target regions, respectively. These plots can be used to evaluate whether there is a general tendency towards lower coverage at the end of the targeted region, and whether there is a bias in terms of forward and reverse reads coverage.

Read count per %GC The plot shows the GC content of the reference sequence on the X-axis and the number of mapped reads on the Y-axis. This plot will show if there is a basis caused by higher GC-content in the sequence.



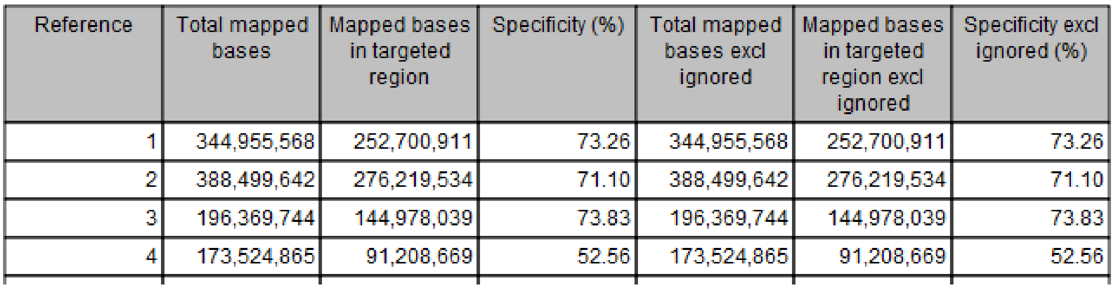
1.5 Minimum coverage of target regions

Coverage	
1 x	95.06%
5 x	89.83%
10 x	82.40%
20 x	62.40%
40 x	26.58%
80 x	4.78%
100 x	2.34%

2 Targeted region overview

Reference	Total mapped reads	Mapped reads in targeted region	Specificity (%)	Total mapped reads excl ignored	Mapped reads in targeted region excl ignored	Specificity excl ignored (%)
1	4,338,296	2,168,439	49.98	4,338,296	2,168,439	49.98
2	4,929,239	2,384,422	48.37	4,929,239	2,384,422	48.37
3	2,495,513	1,242,395	49.79	2,495,513	1,242,395	49.79
4	2,214,025	768,447	34.71	2,214,025	768,447	34.71

Figure 18.4: *The report: mapped reads.*



Reference	Total mapped bases	Mapped bases in targeted region	Specificity (%)	Total mapped bases excl ignored	Mapped bases in targeted region excl ignored	Specificity excl ignored (%)
1	344,955,568	252,700,911	73.26	344,955,568	252,700,911	73.26
2	388,499,642	276,219,534	71.10	388,499,642	276,219,534	71.10
3	196,369,744	144,978,039	73.83	196,369,744	144,978,039	73.83
4	173,524,865	91,208,669	52.56	173,524,865	91,208,669	52.56

Figure 18.5: *The report: mapped bases.*

18.1.3 Per-region statistics

In addition to the summary report, you can see coverage statistics for each targeted region. This is reported as a track, and you can see the numbers by going to the table (▦) view of the track. An example is shown in figure 18.6:

Chromosome The name is taken from the reference sequence used for mapping.

Region The region of the

Name The annotation name derived from the annotation (if there is additional information on the annotation, this is retained in this table as well).

Target region length The length of the region.

Target region length with coverage above... The length of the region that is covered by at least the **Minimum coverage** level provided in figure 18.1.

Percentage with coverage above... The percentage of the positions in the region with coverage above the **Minimum coverage** level provided in figure 18.1.

Read count Number of reads that cover this region. Note that reads that only cover the region partially are also included. Note that reads in over-lapping pairs are counted individually (see figures 18.6 and figure 18.7).

Base count The number of bases in the reads that are covering the target region. Note that bases in overlapping pairs are counted only once (see figures 18.6 and figure 18.7).

%GC The GC content of the region.

Min coverage The lowest coverage in the region.

Max coverage The highest coverage in the region.

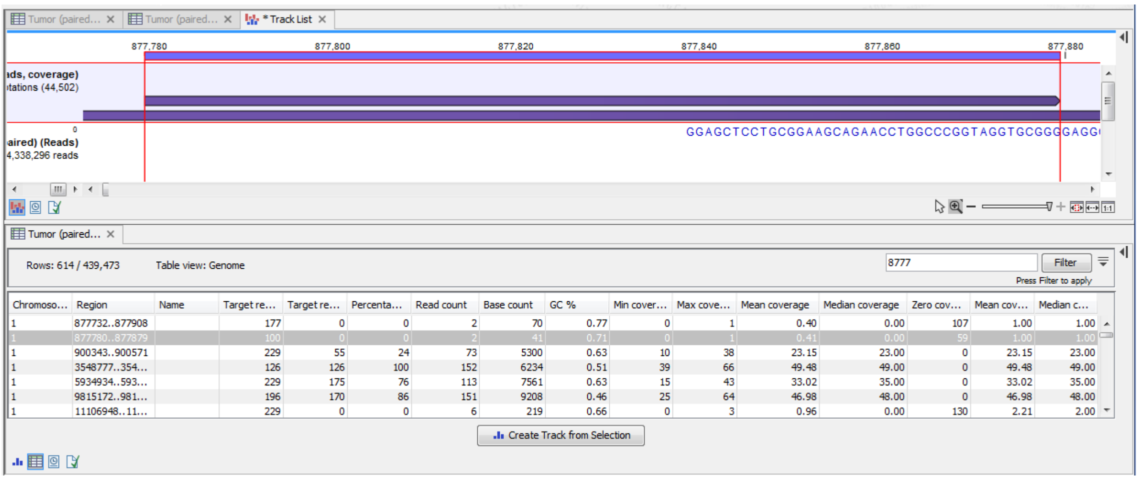
Mean coverage The average coverage in the region.

Median coverage The median coverage in the region.

Zero coverage bases The number of positions with no coverage.

Mean coverage (excluding zero coverage) The average coverage in the region, excluding any zero-coverage parts.

Median coverage (excluding zero coverage) The median coverage in the region, excluding any zero-coverage parts.

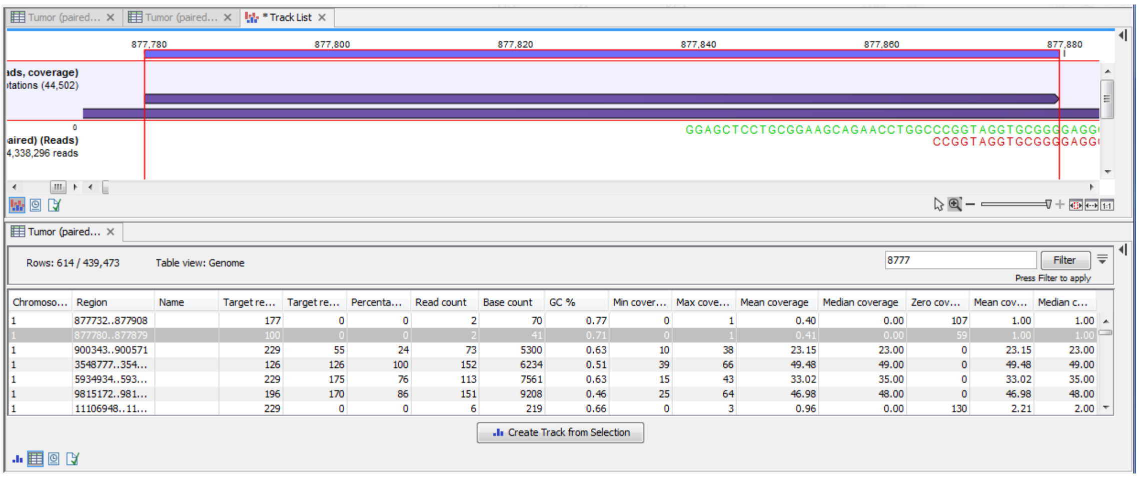


Rows: 614 / 439,473 Table view: Genome

8777

Chromoso...	Region	Name	Target re...	Target re...	Percenta...	Read count	Base count	GC %	Min cover...	Max cove...	Mean coverage	Median coverage	Zero cov...	Mean cov...	Median c...
1	877732..877908		177	0	0	2	70	0.77	0	1	0.40	0.00	107	1.00	1.00
1	877780..877879		100	0	0	2	41	0.71	0	1	0.41	0.00	59	1.00	1.00
1	900343..900571		229	55	24	73	5300	0.63	10	38	23.15	23.00	0	23.15	23.00
1	3548777..354...		126	126	100	152	6234	0.51	39	66	49.48	49.00	0	49.48	49.00
1	5934934..593...		229	175	76	113	7561	0.63	15	43	33.02	35.00	0	33.02	35.00
1	9815172..981...		196	170	86	151	9208	0.46	25	64	46.98	48.00	0	46.98	48.00
1	11106948..11...		229	0	0	6	219	0.66	0	3	0.96	0.00	130	2.21	2.00

Figure 18.6: *A track list containing the target region coverage track and reads track. The target region coverage track has been opened from the track list and is shown in table view. Detailed information on each region is displayed. Only one paired read maps to the region selected.*



Rows: 614 / 439,473 Table view: Genome

Chromoso...	Region	Name	Target re...	Target re...	Percenta...	Read count	Base count	GC %	Min cover...	Max cove...	Mean coverage	Median coverage	Zero cov...	Mean cov...	Median c...
1	877732..877908		177	0	0	2	70	0.77	0	1	0.40	0.00	107	1.00	1.00
1	877780..877879		100	0	0	2	41	0.71	0	1	0.41	0.00	59	1.00	1.00
1	900343..900571		229	55	24	73	5300	0.63	10	38	23.15	23.00	0	23.15	23.00
1	3548777..354...		126	126	100	152	6234	0.51	39	66	49.48	49.00	0	49.48	49.00
1	5934934..593...		229	175	76	113	7561	0.63	15	43	33.02	35.00	0	33.02	35.00
1	9815172..981...		196	170	86	151	9208	0.46	25	64	46.98	48.00	0	46.98	48.00
1	11106948..11...		229	0	0	6	219	0.66	0	3	0.96	0.00	130	2.21	2.00

Figure 18.7: *The same data as shown in figure 18.6, but now the Disconnect paired reads option in the side-panel of the reads track has been ticked, so that the two reads in the paired read are shown disconnected.*

18.1.4 Coverage table

Besides standard information such as position etc, the coverage table (figure 18.8) lists the following information for each position in the whole target:

Name The name of the target region.

Target region position The name of the target region.

Reference base The base in the reference sequence.

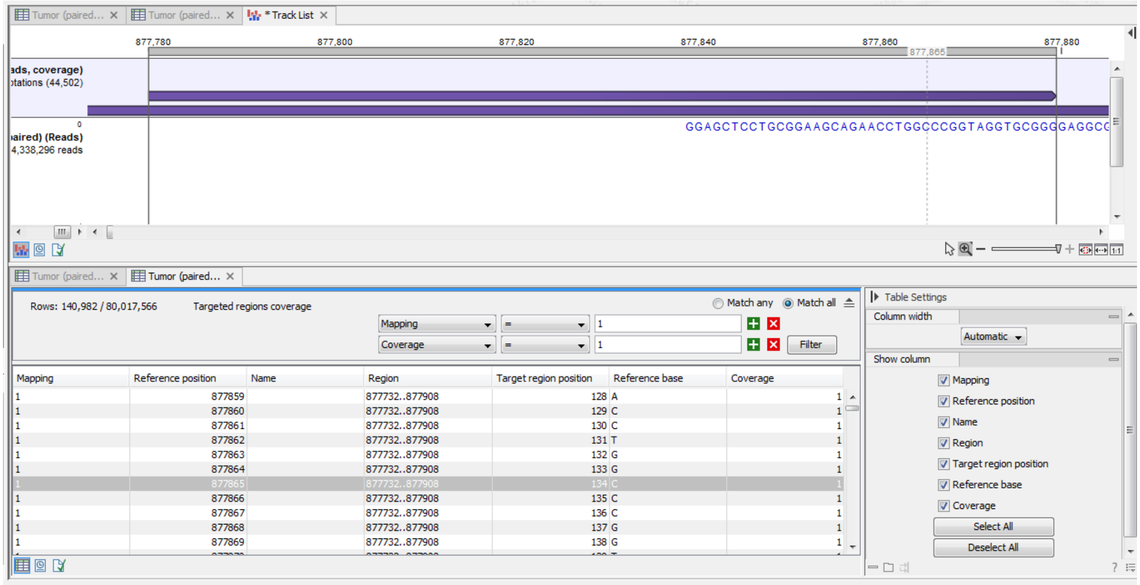
Coverage The number of bases mapped to this position. Note that bases in over-lapping pairs are counted only once. Also note that if the user has chosen the **Ignore non-specific matches** or **Ignore broken pairs** options, these reads will be ignored. (see discussion on coverage in section 18.2.2).

18.2 QC for Sequencing Reads

Quality assurance as well as concern regarding sample authenticity in biotechnology and bioengineering have always been serious topics in both production and research. While next generation sequencing techniques greatly enhance in-depth analyses of DNA-samples, they, however, introduce additional error-sources. Resulting error-signatures can neither be easily removed from resulting sequencing data nor even recognized, which is mainly due to the massive amount of data. Altogether biologists and sequencing facility technicians face not only issues of minor relevance, e.g. suboptimal library preparation, but also serious incidents, including sample-contamination or even mix-up, ultimately threatening the accuracy of biological conclusions.

Unfortunately, most of the problems and evolving questions raised above can't be solved and answered entirely. However, the sequencing data quality control tool of the *CLC Cancer Research Workbench* provides various generic tools to assist in the quality control process of the samples by assessing and visualizing statistics on:

- Sequence-read lengths and base-coverages



Rows: 140,982 / 80,017,566 Targeted regions coverage

○ Match any ● Match all

Mapping = 1

Coverage = 1 Filter

Mapping	Reference position	Name	Region	Target region position	Reference base	Coverage
1	877859		877732..877908	128	A	1
1	877860		877732..877908	129	C	1
1	877861		877732..877908	130	C	1
1	877862		877732..877908	131	T	1
1	877863		877732..877908	132	G	1
1	877864		877732..877908	133	G	1
1	877865		877732..877908	134	C	1
1	877866		877732..877908	135	C	1
1	877867		877732..877908	136	C	1
1	877868		877732..877908	137	G	1
1	877869		877732..877908	138	G	1

☑ Mapping
☑ Reference position
☑ Name
☑ Region
☑ Target region position
☑ Reference base
☑ Coverage

Figure 18.8: *The targeted region coverage table for the same region as shown in same as shown in figures 18.6 and figure 18.7.*

- Nucleotide-contributions and base-ambiguities
- Quality scores as emitted by the base-caller
- Over-represented sequences and hints suggesting contamination events

This tool aims at assessing above quality-indicators and investigates proper and improper result presentation. The inspiration comes from the FastQC-project (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

18.2.1 Report contents

The sections below describe the contents of the report. Note that the two terms "per-sequence" and "per-base" are used frequently in the following descriptions. The generated report is divided into per-sequence and per-base sections. In per-sequence assessments some characteristic (a single value) is assessed for each sequence and then contributes to the overall assessment. In per-base assessments each base position is examined and counted independently.

The report comes in two different flavors: a supplementary report consisting of tables representing all the values that are calculated, and a main summary reports where the tables are visualized in plots (see an example in figure 18.9). Both reports can be exported as pdf files or Excel spread sheets.

Basic analysis

The basic analysis section assesses the most simple characteristics that are supported by all sequencing technologies.

The Summary table provides information regarding the creation date, the author, the software used, the number of data sets the report is based upon, as well as data name and content in terms of read number and total number of nucleotides.

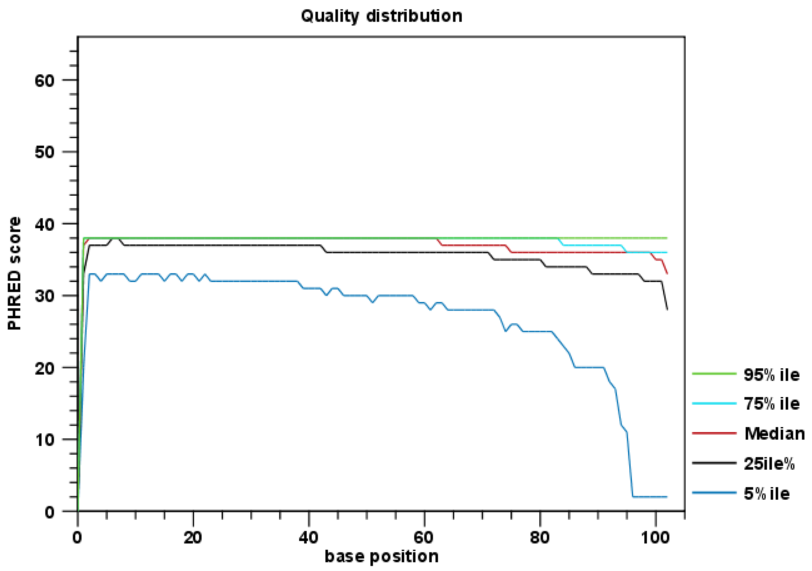


Figure 18.9: *An example of a plot from the graphical report, showing the quality values per base position.*

Sequence length distribution Counts the number of sequences that have been observed for individual sequence lengths. The resulting table correlates sequence-lengths in base-pairs with numbers of sequences observed with that number of base-pairs.

Base coverage distribution Calculates absolute coverages for individual base positions. The resulting table correlates base-positions with the number of sequences that supported (covered) that position.

Sequence-wise %GC-content distribution Counts the number of sequences that feature individual %GC-contents in 101 bins ranging from 0 to 100%.The %GC-content of a sequence is calculated by dividing the absolute number of G/C-nulceotides by the length of that sequence.

Sequence-wise %N-content distribution Counts the number of sequences that feature individual %N-contents in 101 bins ranging from 0 to 100%, where N refers to all ambiguous base-codes as specified by IUPAC.The %N-content of a sequence is calculated by dividing the absolute number of ambiguous nucleotides through the length of that sequence.

Base-wise nucleotide distributions Calculates absolute coverages for the four DNA nucleotides (A, C, G or T) for each base position in the sequences.

Base-wise GC-distribution Calculates absolute coverages of C's + G's for each base position in the sequences.

Base-wise N-distribution Calculates absolute coverages of N's, for each base position in the sequences, where N refers to all ambiguous base-codes as specified by IUPAC.

Quality analysis

The quality analysis examines quality scores reported from technology-dependent base callers. Please note that the NGS import tools of the *CLC Genomics Workbench* and *CLC Genomics Server* convert quality scores to PHRED-scale, regardless of the data source. The following quality distributions are reported:

per-sequence quality distribution Calculates amounts of sequences that feature individual PHRED-scores in 64 bins from 0 to 63. The quality score of a sequence as calculated as arithmetic mean of its base qualities.

per-base quality distribution Calculates amounts of bases that feature individual PHRED-scores in 64 bins from 0 to 63. This results in a three-dimensional table, where dimension 1 refers to the base-position, dimension 2 refers to the quality-score and dimension 3 to amounts of bases observed at that position with that quality score.

Over-representation analysis

The 5-mer analysis examines the enrichment of penta-nucleotides. The enrichment of 5-mers is calculated as the ratio of observed and expected 5-mer frequencies. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5-mer. (Example: given the 5-mer = CCCCC and cytosines have been observed to 20% in the examined sequences, the 5-mer expectation is 0.2^5). Note that 5-mers that contain ambiguous bases (anything different from A/T/C/G) are ignored.

Individual 5-mer distribution Calculates the absolute coverage and enrichment for each 5-mer (observed/expected based on background distribution of nucleotides) for each base position, and plots position vs enrichment data for the top five enriched 5-mers (or fewer if less than five enriched 5-mers are present). This analysis will reveal if there is a pattern of bias at different positions over the read length. Such a bias might origin from non-trimmed adapter sequences, poly-A tails or other sources.

Duplicated sequences analysis

The duplicated sequences analysis identifies sequences that have been sequenced multiple times. In order to achieve reasonable performance, not all input sequences are analyzed. Instead a sequence-dictionary is used, whose entries are sampled evenly from input sequences. Please note that if you select multiple sequence lists as an input, they will all be considered one data set for this analysis (batching can be used to generate separate reports for an individual sequence list). As soon as a sequence makes it into the dictionary (which is a random process), it is tracked for duplicates until all sequences have been examined. The dictionary size is 250 000 sequences.

Because all current sequencing techniques tend to report fading quality scores for the 3' ends of sequences, there is a risk that duplicates are NOT detected, just because of sequencing errors near their 3' ends. Therefore, the identity of two sequences is calculated using only the first 50nt from the 5' end.

Sequence duplication levels This results in a table correlating duplication counts with the number of sequences that featured that duplicate-count. For example, if the dictionary contains 10 sequences and each sequence was seen exactly once, then the table will contain only one row displaying: duplication-count=1 and sequence-count=10. Note: due to space restrictions the corresponding bar-plot shows only bars for duplication-counts of x=[0-100].

Bar-heights of duplication-counts >100 are accumulated at x=100, such that a significantly elevated bar-height at x=100 is a normal observation. Please refer to the table-report for a full list of individual duplication-counts.

Duplicated sequences This results in a list of actual sequences most prevalently observed. The list contains a maximum of 25 (most frequently observed) sequences and is only present in the supplementary report.

18.2.2 Adapters

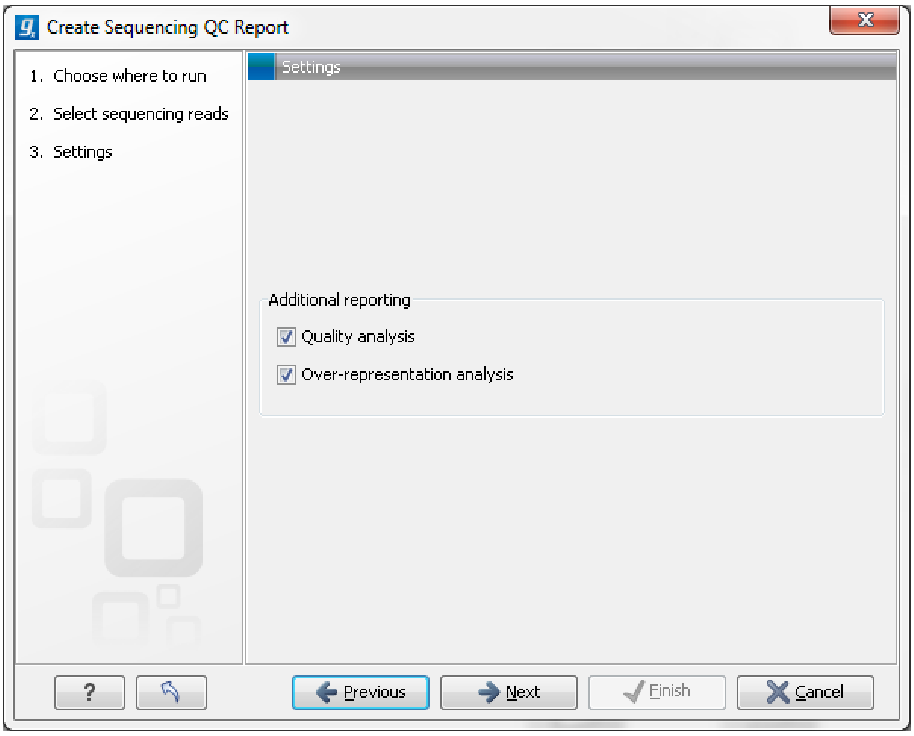
Currently, adapter contamination, i.e. adapter sequences in the reads, cannot be detected in a reliable way with this tool. In some cases, adapter contamination will show up as enriched 5-mers near the end of sequences but only if the contamination is severe.

18.2.3 Running the "QC for Sequencing Reads" tool

The tool is found in the Toolbox:

Toolbox | **Quality Control** ( ) | **QC for Sequencing Reads** ( )

Select one or more sequence lists with sequencing reads as input. If sequence lists in the Navigation Area were already selected, these will be shown in the Selected Elements window. When multiple lists are selected as an input, they are all analyzed in one pool. If you need separate reports for each data set, you can run it in a batch. Clicking **Next** allows you to set parameters as displayed in figure 20.34.



Create Sequencing QC Report

1. Choose where to run
2. Select sequencing reads
3. Settings

Settings

Additional reporting

☑ Quality analysis

☑ Over-representation analysis

Figure 18.10: *Setting parameters for quality control.*

The following parameters can be set:

- **Quality analysis** as described in section 18.2.1.
- **Over-representation analysis** as described in section 18.2.1.

Click **Next** to adjust output options which allow you to select the graphical and supplementary report.

18.3 QC for Read Mapping

You can create two kinds of reports regarding read mappings: *First*, you can choose to generate a summary report about the mapping process itself (see section 20.4). *Second*, you can generate a detailed statistics report after the mapping has finished. This report is useful if you want to generate statistics across results made in different processes, and it generates more detailed statistics than the summary mapping report. Both reports are described below.

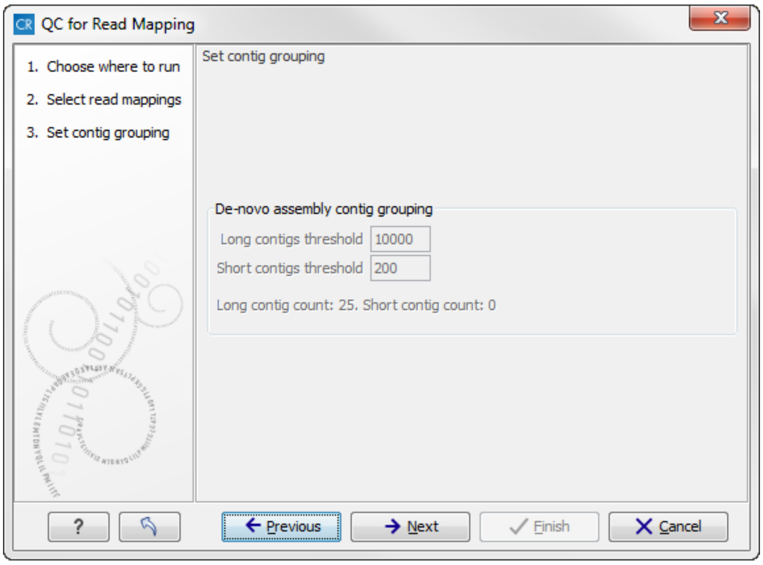
18.3.1 Running the "QC for Read Mapping" tool

The tool is found in the Toolbox:

Toolbox | **Quality Control** ( ) | **QC for Read Mapping** ( )

This opens a dialog where you can select mapping results ( )/ ( )/ ( ) or RNA-Seq analysis results ( ).

Clicking **Next** will display the dialog shown in figure 18.11



QC for Read Mapping

1. Choose where to run
2. Select read mappings
3. Set contig grouping

Set contig grouping

De-novo assembly contig grouping

Long contigs threshold 10000

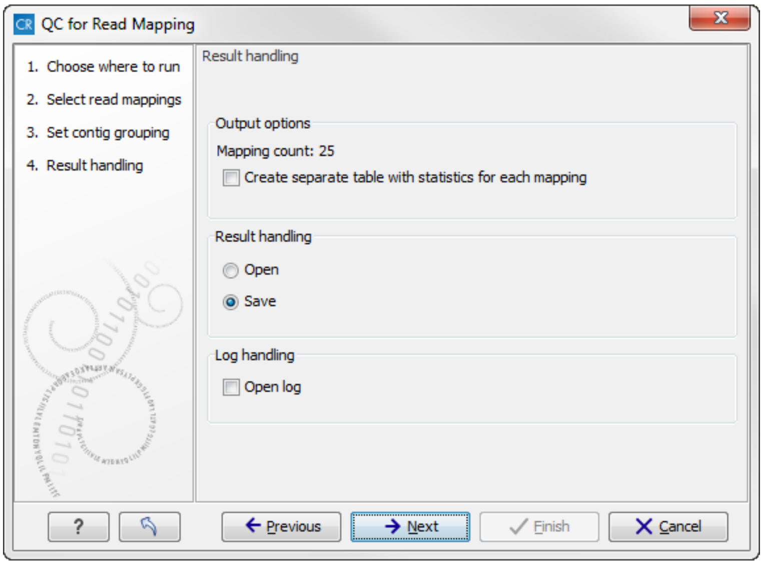
Short contigs threshold 200

Long contig count: 25. Short contig count: 0

Figure 18.11: *Parameters for mapping reports.*

The next wizard step shows the used thresholds for the mapping report. These parameters cannot be modified by the user (as thresholds can only be specified for de novo assemblies that does not have a consensus sequence). Whenever a consensus sequence is present the "De novo assembly contig grouping" options are disabled.

Click **Next** to select output options as shown in figure 18.12



QC for Read Mapping

1. Choose where to run
2. Select read mappings
3. Set contig grouping
4. Result handling

Result handling

Output options

Mapping count: 25

□ Create separate table with statistics for each mapping

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 18.12: *Optionally create a table with detailed statistics per reference.*

Per default, an overall report will be created as described below. In addition, by checking **Create table with statistics for each mapping**, you can create a table showing detailed statistics for each reference sequence. The following sections describe the information produced.

Reference sequence statistics

For reports on results of read mapping, section two concerns the reference sequences. The reference identity part includes the following information:

Reference name The name of the reference sequence.

Reference Latin name The reference sequence's Latin name.

Reference description Description of the reference.

If you want to inspect and edit this information, right-click the reference sequence in the contig and choose **Open Sequence** and switch to the **Element info** ( ) tab (learn more in section 9.4). Note that you need to create a new report if you want the information in the report to be updated. If you update the information for the reference sequence within the contig, you should know that it doesn't affect the original reference sequence saved in the **Navigation Area**.

The next part of the report reports coverage statistics including GC content of the reference sequence. Note that coverage is reported on two levels: including and excluding zero coverage regions. In some cases, you do not expect the whole reference to be covered, and only the coverage levels of the covered parts of the reference sequence are interesting. On the other hand, if you have sequenced the full genome that you use as reference, the overall coverage is probably the most relevant number (i.e. including zero coverage regions).

A position on the reference is counted as "covered" when at least one read is aligned to it. Note that unaligned ends (faded nucleotides at the ends) that are produced when mapping using local alignment do not contribute to the coverage. In the example shown in figure 18.13, there is a region of zero coverage in the middle and one time coverage on each side. Note that the gaps to the very right are within the same read which means that these two positions on the reference sequence are still counted as "covered".

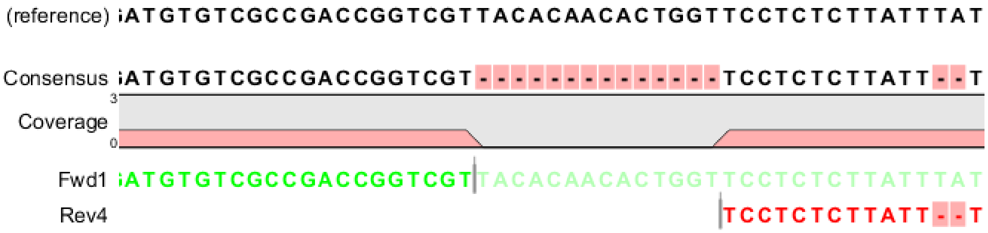

(reference) ;ATGTGTCGCCGACCGGTCGTTACACAACACTGGTTCCTCTCTTATTTAT
Consensus ;ATGTGTCGCCGACCGGTCGT--------------TCCTCTCTTATT--T
Coverage
3
0
Fwd1 ;ATGTGTCGCCGACCGGTCGTTACACAACACTGGTTCCTCTCTTATTTAT
Rev4 TCCTCTCTTATT--T

Figure 18.13: *A region of zero coverage in the middle and one time coverage on each side. Note that the gaps to the very right are within the same read which means that these two positions on the reference sequence are still counted as "covered".*

The identity section is followed by some statistics on the zero-coverage regions; the number, minimum and maximum length, mean length, standard deviation, total length and a list of the regions. If there are too many regions, they will not all be listed in the report (if there are more than 20, only the first 10 are reported).

Next follow two bar plots showing the distribution of coverage with coverage level on the x-axis and number of positions with that coverage on the y-axis. An example is shown in figure 18.16.

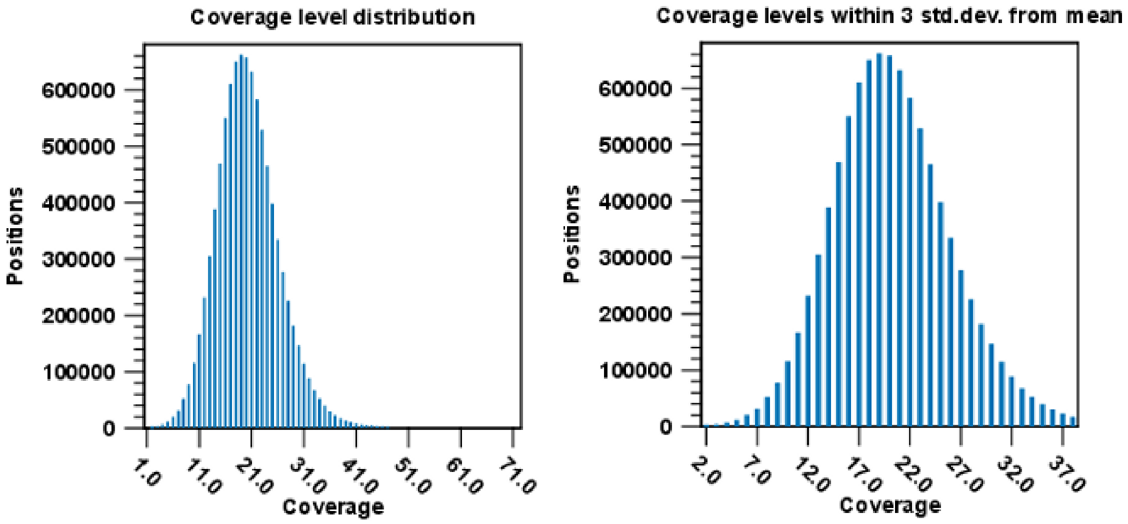


Figure 18.14: *Distribution of coverage - to the left for all the coverage levels, and to the right for coverage levels within 3 standard deviations from the mean.*

The graph to the left shows all the coverage levels, whereas the graph to the right shows coverage levels within 3 standard deviations from the mean. The reason for this is that for complex genomes, you will often have a few regions with extremely high coverage which will affect the resolution of the graph, making it impossible to see the coverage distribution for the majority of the references. These coverage outliers are excluded when only showing coverage within 3 standard deviations from the mean. Note that zero-coverage regions are not shown in the graph but reported in text below (this information is also in the zero-coverage section). Below the second coverage graph there are some statistics on the data that is outside the 3 standard deviations.

One of the biases seen in sequencing data concerns GC content. Often there is a correlation between GC content and coverage. In order to investigate this correlation, the report includes a graph plotting coverage against GC content (see figure 18.15). Note that you can see the GC content for each reference sequence in the table above.

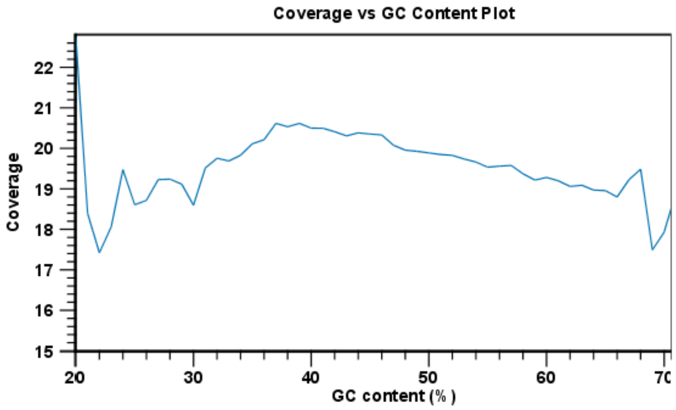


Figure 18.15: *The plot displays, for each GC content level (0-100 %), the mean read coverage of 100bp reference segments with that GC content.*

The plot displays, for each GC content level (0-100 %), the mean read coverage of 100bp reference segments with that GC content.

At the end follows statistics about the reads which are the same for both reference and de novo assembly (see section 18.2.2 below).

Next follow two bar plots showing the distribution of coverage with coverage level on the x-axis and number of positions with that coverage on the y-axis. An example is shown in figure 18.16.

The graph to the left shows all the coverage levels, whereas the graph to the right shows coverage levels within 3 standard deviations from the mean. The reason for this is that for complex genomes, you will often have a few regions with extremely high coverage which will affect the resolution of the graph, making it impossible to see the coverage distribution for the majority of the references. These coverage outliers are excluded when only showing coverage within 3 standard deviations from the mean. Below the second coverage graph there are some statistics on the data that is outside the 3 standard deviations. At the end follows statistics about the reads which are the same for both reference and de novo assembly (see section 18.2.2 below).

Read statistics

This section contains simple statistics for all mapped reads, non-specific matches (reads that match more than place during the assembly), non-perfect matches and paired reads. **Note!** Paired reads are counted as two, even though they form one pair. The section on paired reads also includes information about paired distance and counts the number of pairs that were broken due to:

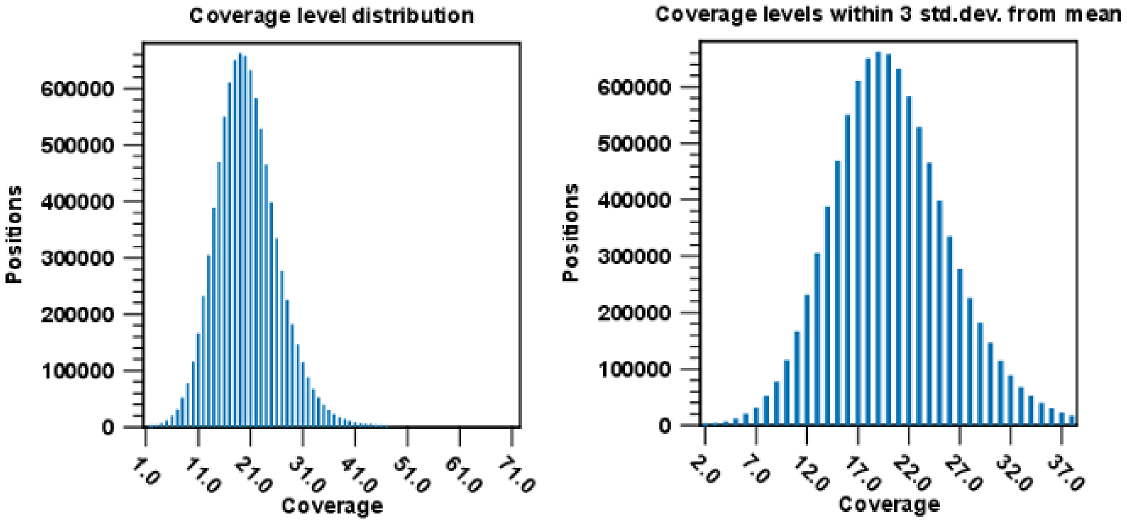


Figure 18.16: *Distribution of coverage - to the left for all the coverage levels, and to the right for coverage levels within 3 standard deviations from the mean.*

Wrong distance When starting the mapping, a distance interval is specified. If the reads during the mapping are placed outside this interval, they will be counted here.

Mate inverted If one of the reads has been matched as reverse complement, the pair will be broken (note that the pairwise orientation of the reads is determined during import).

Mate on other contig If the reads are placed on different contigs, the pair will also be broken.

Mate not matched If only one of the reads match, the pair will be broken as well.

Below these tables follow two graphs showing distribution of paired distances (see figure 18.17) and distribution of read lengths. Note that the distance includes both the read sequence and the insert between them as explained in section 6.3.8.

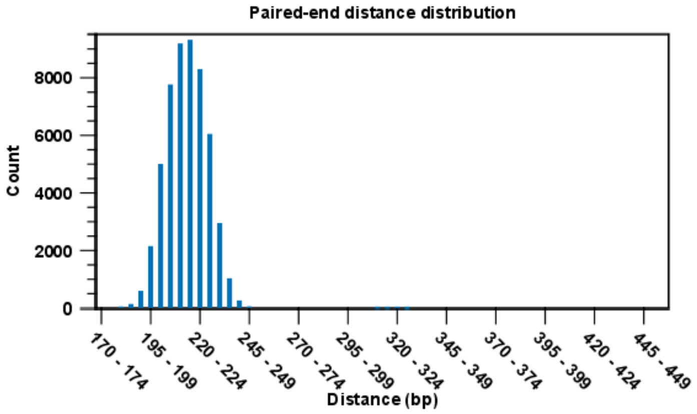


Figure 18.17: *A bar plot showing the distribution of distances between intact pairs.*

Two plots of the distribution of insertion and deletion lengths can bee seen in figure 18.18 and figure 18.19.

Quality and mismatches

Next follows a detailed description of which bases in the reference are substituted to which bases in the reads. This information is plotted in different ways with an example shown here in figure 18.18.

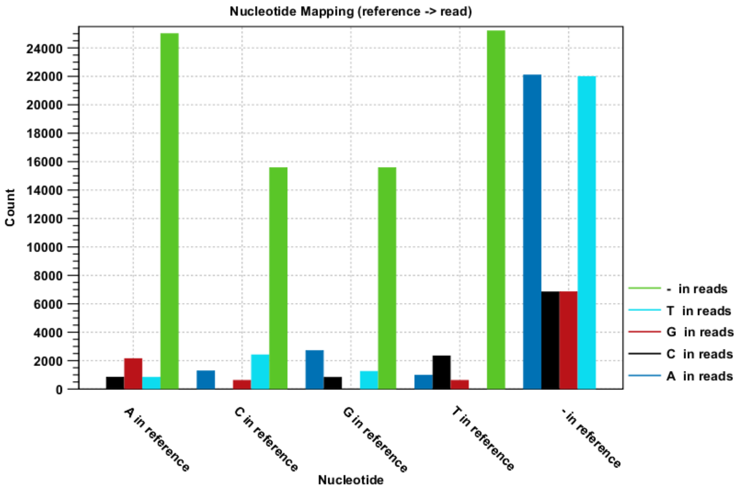


Figure 18.18: *The As and Ts are more often substituted with a gap in the sequencing reads than C and G.*

This example shows for each type of base in the reference sequence, which base (or gap) is found most often. Please note that only mismatches are plotted - the matches are not included. For example, an A in the reference is more often replaced by a G than any other base.

Below these plots, there are two plots of the quality values for matches and mismatches, respectively. Next, there is a plot of the mismatch fraction for each read position. Typically with quality dropping towards the end of a read, there will be more mismatches towards the end as the example in figure 18.19 shows.

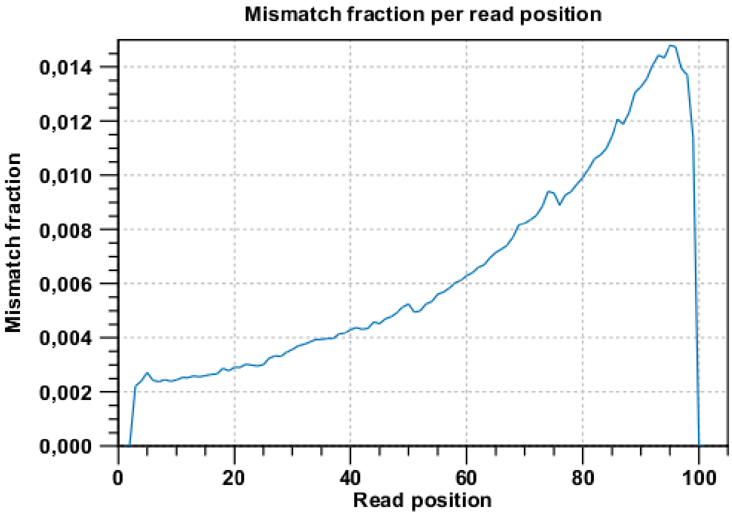
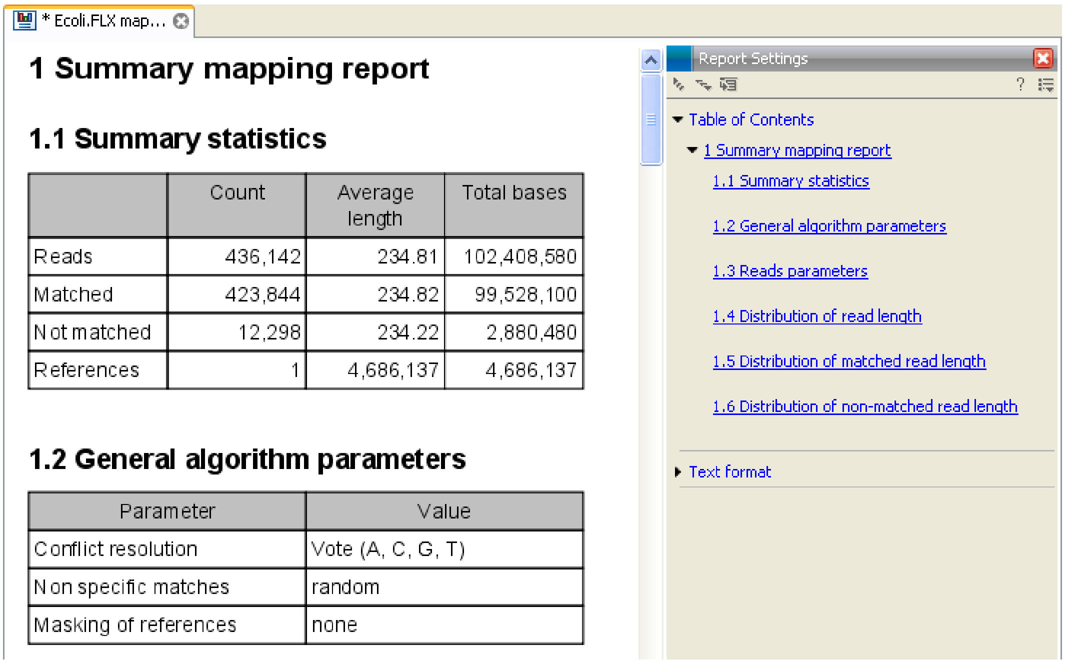


Figure 18.19: *There are mismatches towards the end of the reads.*

The last plots shows the unaligned read lengths.

18.3.2 Summary mapping report

If you choose to create a report as part of the read mapping (see section 20.4), this report will summarize the results of the mapping process. An example of a report is shown in figure 18.20


* Ecoli.FLX map...
Report Settings
?
Table of Contents
1 Summary mapping report
1.1 Summary statistics
1.2 General algorithm parameters
1.3 Reads parameters
1.4 Distribution of read length
1.5 Distribution of matched read length
1.6 Distribution of non-matched read length
Text format

1 Summary mapping report

1.1 Summary statistics

	Count	Average length	Total bases
Reads	436,142	234.81	102,408,580
Matched	423,844	234.82	99,528,100
Not matched	12,298	234.22	2,880,480
References	1	4,686,137	4,686,137

1.2 General algorithm parameters

Parameter	Value
Conflict resolution	Vote (A, C, G, T)
Non specific matches	random
Masking of references	none

Figure 18.20: *The summary mapping report.*

The information included in the report is:

- **Summary statistics**. A summary of the mapping statistics:
  - **Reads**. The number of reads and the average length.
  - **Mapped**. The number of reads that are mapped and their average length.
  - **Not mapped**. The number of reads that do not map and their average length.
  - **References**. Number of reference sequences.
- **Parameters**. The settings used are reported for the process as a whole and for each sequence list used as input.
- **Distribution of read length**. For each sequence length, you can see the number of reads and the distribution in percent. This is mainly useful if you don't have too much variance in the lengths as you have in e.g. Sanger sequencing data.
- **Distribution of matched reads lengths**. Equivalent to the above, except that this includes only the reads that have been matched to a contig.
- **Distribution of non-matched reads lengths**. Show the distribution of lengths of the rest of the sequences.

You can copy the information from the report by selecting in the report and click **Copy** (⧉). You can also export the report in Excel format.

Chapter 19

Preparing raw data tools

Contents

19.1 Merge overlapping pairs

Some paired end library preparation methods using relatively short fragment size will generate data with overlapping pairs. This type of data can be handled as standard paired-end data in the *CLC Cancer Research Workbench*, and it will work perfectly fine (see details for variant detection in section 20.19).

However, in some situations it can be useful to merge the overlapping pair into one sequence read instead. The benefit is that you get longer reads, and that the quality improves (normally the quality drops towards the end of a read, and by overlapping the ends of two reads, the consensus read now reflects two read ends instead of just one).

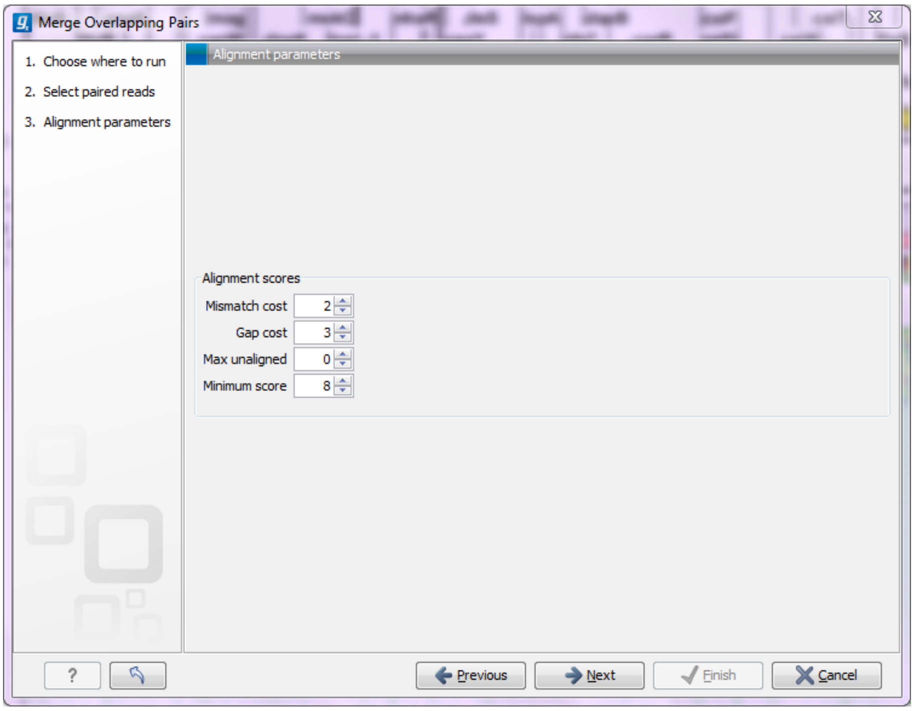
In the *CLC Cancer Research Workbench*, there is a tool for merging overlapping reads, which are in forward-reverse orientation:

Toolbox | **NGS Core Tools** ( ) | **Merge Overlapping Pairs** ( )

Select one or more sequence lists with paired end sequencing reads as input.

Please note that read pairs have to be in forward-reverse orientation. Please also note that after merging the merged reads will always be in the forward orientation.

Clicking **Next** allows you to set parameters as displayed in figure 19.1.



Alignment scores

Mismatch cost 2

Gap cost 3

Max unaligned 0

Minimum score 8

Figure 19.1: *Setting parameters for merging overlapping pairs.*

In order to understand how these parameters should be set, an explanation of the merging algorithm is needed: Because the fragment size is not an exact number of base pairs and is different from fragment to fragment, an alignment of the two reads has to be performed. If the alignment is *good and long enough*, the reads will be merged. *Good enough* in this context means that the alignment has to satisfy some user-specified score criteria (details below). Because of sequencing errors that typically are more abundant towards the end of the read, the alignment is not expected always to be perfect, and the user can decide how many errors are acceptable. *Long enough* in this context means that the overlap between the reads has to be non-coincidental. Merging two reads that do not really overlap, leads to errors in the downstream analysis, thus it is very important to make sure that the overlap is big enough. If only a few bases overlap was required, some read pairs will match by chance, so this has to be avoided.

The following parameters are used to define what is *good enough* and *long enough*

- **Mismatch cost** The alignment awards one point for a match, and the mismatch cost is set by this parameter. The default value is 2.
- **Gap cost** This is the cost for introducing an insertion or deletion in the alignment. The default value is 3.
- **Max unaligned end mismatches** The alignment is local, which means that a number of bases can be left unaligned. If the quality of the reads is dropping to be very poor towards the end of the read, and the expected overlap is long enough, it makes sense to allow some unaligned bases at the end. However, this should be used with great care which is why the default value is 0. As explained above, a wrong decision to merge the reads leads to errors in the downstream analysis, so it is better to be conservative and accept fewer merged reads in the result.
- **Minimum score** This is the minimum score of an alignment to be accepted for merging. The

default value is 10. As an example: with default settings, this means that an overlap of 13 bases with one mismatch will be accepted (12 matches minus 2 for a mismatch).

Please note that even with the alignment scores above the minimum score specified in the tool setup, the paired reads also need to have the number of end mismatches below the "Maximum unaligned end mismatches" value specified in the tool setup to be qualified for merging.

After clicking **Next** you can select whether a report should be generated as part of the output. The main result will be two sequence lists for each list in the input: one containing the merged reads (marked as single end reads), and one containing the reads that could not be merged (still marked as paired data). Since the *CLC Cancer Research Workbench* handles a mix of paired and unpaired data, both of these sequence lists can be used in the further analysis. However, please note that low quality can be one of the reasons why a pair cannot be merged. Hence, the list of reads that could not be paired is more likely to contain more reads with errors than the one with the merged reads.

19.1.1 Using quality scores when merging

Quality scores come into play in two different ways when merging overlapping pairs.

First, if there is a conflict between the reads in a pair (i.e. a mismatch or gap in the alignment), quality scores are used to determine which base the merged read should have at a given position. The base with the highest quality score will be the one used. In the case of gaps, the average of the quality scores of the two surrounding bases will be used. In the case that two conflicting bases have the same quality or both reads have no quality scores, an [IUPAC ambiguity code](see the appendix section D) representing these bases will be inserted.

Second, the quality scores of the merged read reflect the quality scores of the input reads.

We assume independence of errors in calculating the new quality score for a merged position and carry out the following calculations:

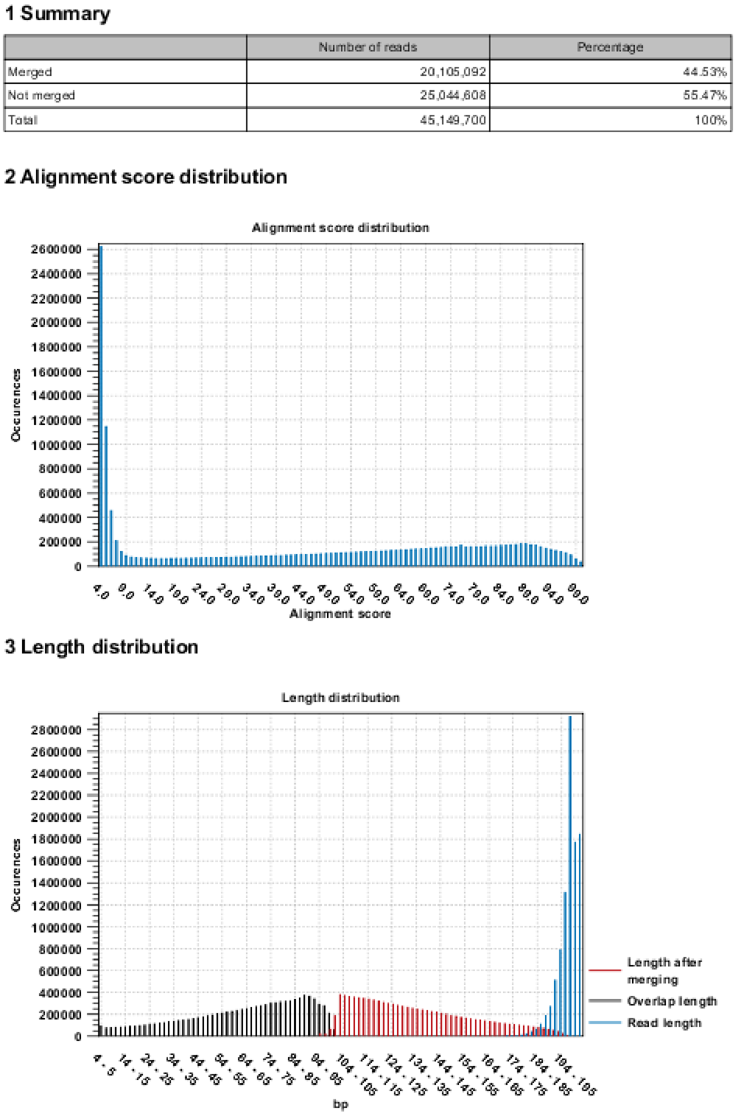
- When the two reads agree at a position, the two quality scores are summed to form the quality score of the base in the new read. The score is capped at the maximum value on the quality score scale which is 64. Phred scores are log scores, so their sums represent the multiplication of the original error probabilities.
- If the two bases disagree at a position, the quality score of the base in the new read is determined by subtracting the lowest score from the highest score of the input reads. Similar to the addition of scores when bases are the same, this adjusts the error probability to reflect a decreased certainty that the base reported at that position is correct.

Thus, if two bases at a given position of an overlapping region are different, and each of those bases was originally given a high phred score, the score assigned to the merged base will be very low. This reflects the fact that the base at this position is unreliable.

If a base at a given position in one read of an overlapping region has a very low quality score and the base at that position in the other read has a high score, it is likely that the base with the high quality score is correct. The adjusted quality score for this position in the merged read would reflect that there is less certainty in the base at that position than before. However, such a position would still be assigned quite a high quality, as the base call is still likely to be correct.

19.1.2 Report of merged pairs

Figure 19.2 shows an example of the report generated when merging overlapping pairs.



1 Summary

	Number of reads	Percentage
Merged	20,105,092	44.53%
Not merged	25,044,608	55.47%
Total	45,149,700	100%

2 Alignment score distribution

3 Length distribution

Figure 19.2: *Report of overlapping pairs.*

It contains three sections:

- A summary showing the numbers and percentages of reads that have been merged.
- A plot of the alignment scores. This can be used to guide the choice of minimum alignment score as explained in section 19.1.
- A plot of read lengths. This shows the distribution of read lengths for the pairs that have been merged:
  - The length of the overlap.
  - The length of the merged read.
  - The combined length of the two reads in the pair before merging.

19.2 Trim Sequences

CLC Cancer Research Workbench offers a number of ways to trim your sequence reads prior to assembly and mapping, including adapter trimming, quality trimming and length trimming. For each original read, the regions of the sequence to be removed for each type of trimming operation are determined independently according to choices made in the trim dialogs. The types of trim operations that can be performed are:

1. Quality trimming based on quality scores
2. Ambiguity trimming to trim off e.g. stretches of Ns
3. Adapter trimming
4. Base trim to remove a specified number of bases at either 3' or 5' end of the reads
5. Length trimming to remove reads shorter or longer than a specified threshold

The trim operation that removes the largest region of the original read from either end is performed while other trim operations are ignored as they would just remove part of the same region.

Note that this may occasionally expose an internal region in a read that has now become subject to trimming. In such cases, trimming may have to be done more than once.

The result of the trim is a list of sequences that have passed the trim (referred to as the trimmed list below) and optionally a list of the sequences that have been discarded and a summary report (list of discarded sequences). The original data will not be changed.

To start trimming:

Toolbox | **NGS Core Tools** () | **Trim Sequences** ()

This opens a dialog where you can add sequences or sequence lists. If you add several sequence lists, each list will be processed separately and you will get a a list of trimmed sequences for each input sequence list.

When the sequences are selected, click **Next**.

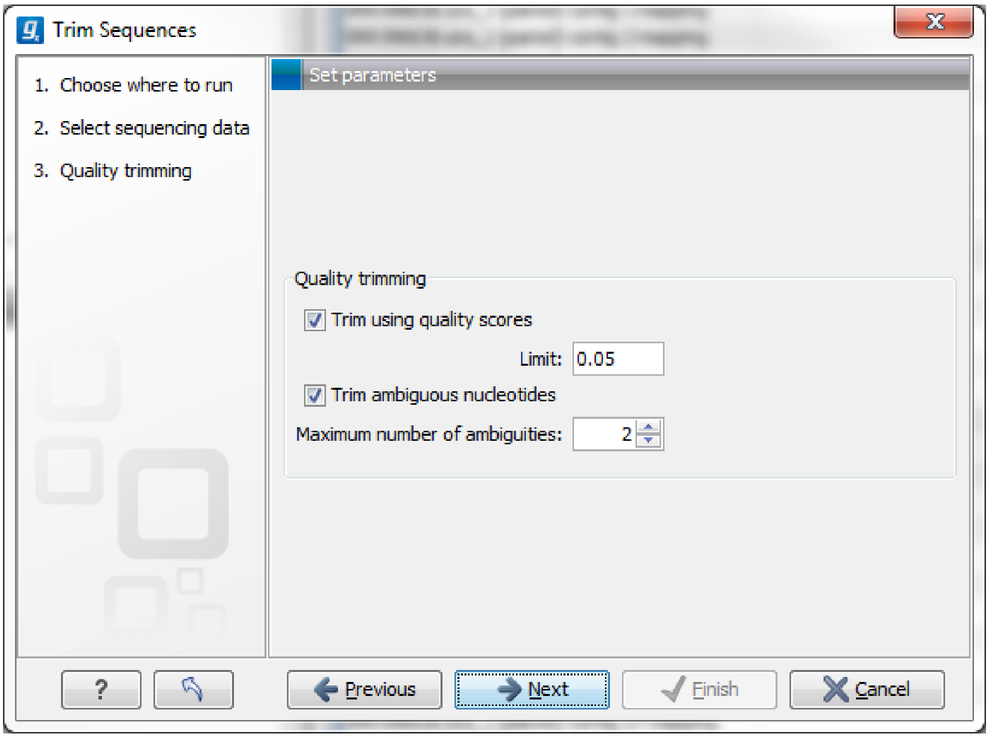
19.2.1 Quality trimming

This opens the dialog displayed in figure 19.3 where you can specify parameters for quality trimming.

The following parameters can be adjusted in the dialog:

- **Trim using quality scores.** If the sequence files contain quality scores from a base-caller algorithm this information can be used for trimming sequence ends. The program uses the modified-Mott trimming algorithm for this purpose (Richard Mott, personal communication):

  Quality scores in the Workbench are on a Phred scale, and formats using other scales will be converted during import. The Phred quality scores (Q), defined as: $Q = -10log10(P)$, where P is the base-calling error probability, can then be used to calculate the error probabilities, which in turn can be used to set the limit for, which bases should be trimmed.



1. Choose where to run
2. Select sequencing data
3. Quality trimming

Quality trimming

☑ Trim using quality scores

Limit: 0.05

☑ Trim ambiguous nucleotides

Maximum number of ambiguities: 2

Figure 19.3: *Specifying quality trimming.*

Hence, the first step in the trim process is to convert the quality score (*Q*) to an error probability: $p_{error} = 10^{\frac{Q}{-10}}$. (This now means that low values are high quality bases.)

Next, for every base a new value is calculated: $Limit - p_{error}$. This value will be negative for low quality bases, where the error probability is high.

For every base, the Workbench calculates the running sum of this value. If the sum drops below zero, it is set to zero. The part of the sequence not trimmed will be the region ending at the highest value of the running sum and starting at the last zero value before this highest score. Everything before and after this region will be trimmed. A read will be completely removed if the score never makes it above zero.

At http://www.clcbio.com/files/usermanuals/trim.zip you find an example sequence and an Excel sheet showing the calculations done for this particular sequence to illustrate the procedure described above.

- **Trim ambiguous nucleotides.** This option trims the sequence ends based on the presence of ambiguous nucleotides (typically N). Note that the automated sequencer generating the data must be set to output ambiguous nucleotides in order for this option to apply. The algorithm takes as input the *maximal number of ambiguous nucleotides allowed in the sequence after trimming*. If this maximum is set to e.g. 3, the algorithm finds the maximum length region containing 3 or fewer ambiguities and then trims away the ends not included in this region.

19.2.2 Adapter trimming

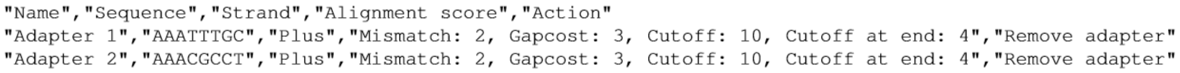
Clicking **Next** will allow you to specify adapter trimming.

In order to trim for adapters, you have to create an adapter list first that must be supplied to the trim tool. A new adapter trim list can be created here:

File | **New** | **Trim Adapter List**

This will create a new empty trim adapter list. Add the adapter(s) that you would like to use for trimming by clicking on the button **Add Row** (✚) found at the bottom of the View Area. Provide the name and sequence of the adapter that should be trimmed away and adjust the parameters if relevant. Click on the button labeled **Finish** to create the adapter trim list. You must now save the generated adapter trim list in the **Navigation Area**. You can do this by clicking on the tab and dragging and dropping the adapter trim list to the desired destination, or you can go to **File** in the menu bar and the choose **Save as**.

You can also create an adapter list by importing a comma separated value (.csv) file of your Adapters. This import can be performed with the standard import using either the Automatic Import option or Force Import as Type: Trim Adapter List. To import a csv file, the names of all adapters must be unique - the Workbench is unable to accept files with multiple rows containing the same adapter name. Additionally, the text between each comma that designates a new column should be quoted. The expected import format for Adapter Lists appears as shown in figure 19.4:



```
"Name","Sequence","Strand","Alignment score","Action"
"Adapter 1","AAATTTGC","Plus","Mismatch: 2, Gapcost: 3, Cutoff: 10, Cutoff at end: 4","Remove adapter"
"Adapter 2","AAACGCCT","Plus","Mismatch: 2, Gapcost: 3, Cutoff: 10, Cutoff at end: 4","Remove adapter"
```

Figure 19.4: *The expected import format for Adapter Lists.*

You can also create an Excel file (.xlsx or .xls) format. In this case, you include the same information per column as indicated above, but do not include the quotes within Excel.

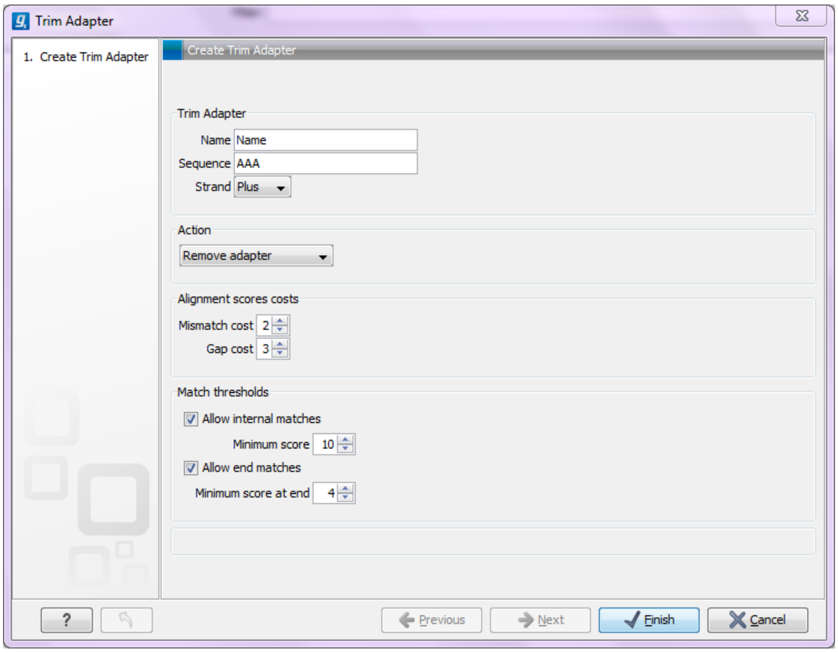
At the bottom of the view, you have the following options:

- **Add Rows.** Add a new adapter. This will bring up a dialog as shown in figure 19.5.
- **Delete Row**. Delete the selected adapter.
- **Edit Row**. Edit the selected adapter. This can also be achieved by double-clicking the row in the table.

The information to be added for each adapter is explained in the following sections, going into detail with the adapter trim. Once the adapters have been added to the list, it should be saved (⇐), and you can select it as shown in figure 19.11.

Action to perform when a match is found For each read sequence in the input to trim, the Workbench performs a Smith-Waterman alignment [Smith and Waterman, 1981] with the adapter sequence to see if there is a match (details described below). When a match is found, the user can specify three kinds of actions:

- **Remove adapter.** This will remove the adapter *and all the nucleotides 5' of the match*. All the nucleotides 3' of the adapter match will be preserved in the read that will be retained in the trimmed reads list. If there are no nucleotides 3' of the adapter match, the read is added to the **List of discarded sequences** (see section 19.2.4).
- **Discard when not found**. If a match is found, the adapter sequence is removed (including all nucleotides 5' of the match as described above) and the rest of the sequence is retained



Trim Adapter

Name: Name

Sequence: AAA

Strand: Plus

Action

Remove adapter

Alignment scores costs

Mismatch cost: 2

Gap cost: 3

Match thresholds

☑ Allow internal matches

Minimum score: 10

☑ Allow end matches

Minimum score at end: 4

Figure 19.5: *Adding a new adapter for adapter trimming.*

in the list of trimmed reads. If no match is found, the whole sequence is discarded and put in the list of discarded sequences. This kind of adapter trimming is useful for small RNA sequencing where the remnants of the adapter is an indication that this is indeed a small RNA.

- **Discard when found**. If a match is found, the read is discarded. If no match is found, the read is retained in the list of trimmed reads. This can be used for quality checking the data for linker contaminations etc.

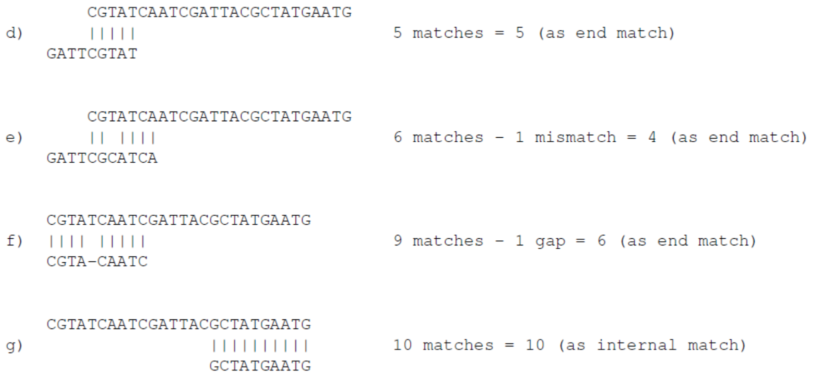
When is there a match? To determine whether there is a match there is a set of scoring thresholds that can be adjusted for each adapter as shown in figure 19.5.

First, you can choose the costs for mismatch and gaps. A match is rewarded one point (this cannot be changed), and per default a mismatch costs 2 and a gap (insertion or deletion) costs 3. A few examples of adapter matches and corresponding scores are shown in figure 19.6.

```
       CGTATCAATCGATTACGCTATGAATG
a)         ||||||| ||||               11 matches - 2 mismatches = 7
          TTCAATCGGTTAC


       CGTATCAATCGATTACGCTATGAATG
          ||||||||| ||||              14 matches - 1 gap = 11
b)        ATCAATCGAT-CGCT


       CGTATCAATCGATTACGCTATGAATG
c)         |||||||                    7 matches - 3 mismatches = 1
          TTCAATCGGG
```

Figure 19.6: *Three examples showing a sequencing read (top) and an adapter (bottom). The examples are artificial, using default setting with mismatch costs = 2 and gap cost = 3.*

In the panel below, you can set the **Minimum score** for a match to be accepted. Note that there is a difference between an **internal match** and an **end match**. The examples above are all internal matches where the alignment of the adapter falls within the read. Figure 19.6 shows a few examples with an adapter match at the end:



```
        CGTATCAATCGATTACGCTATGAATG
d)      |||||                            5 matches = 5 (as end match)
    GATTCGTAT


        CGTATCAATCGATTACGCTATGAATG
e)      || ||||                          6 matches - 1 mismatch = 4 (as end match)
    GATTCGCATCA


    CGTATCAATCGATTACGCTATGAATG
f)  |||| |||||                           9 matches - 1 gap = 6 (as end match)
    CGTA-CAATC


    CGTATCAATCGATTACGCTATGAATG
g)                  ||||||||||           10 matches = 10 (as internal match)
                    GCTATGAATG
```

Figure 19.7: *Four examples showing a sequencing read (top) and an adapter (bottom). The examples are artificial.*

In the first two examples, the adapter sequence extends beyond the end of the read. This is what typically happens when sequencing e.g. small RNAs where you sequence part of the adapter. The third example shows an example which could be interpreted both as an end match and an internal match. However, the Workbench will interpret this as an end match, because it starts at beginning (5' end) of the read. Thus, the definition of an end match is that the alignment of the adapter starts at the read's 5' end. The last example could also be interpreted as an end match, but because it is a the 3' end of the read, it counts as an internal match (this is because you would not typically expect partial adapters at the 3' end of a read). Also note, that if **Remove adapter** is chosen for the last example, the full read will be discarded because everything 5' of the adapter is removed.

Below, the same examples are re-iterated showing the results when applying different scoring schemes. In the first round, the settings are:

- Allowing internal matches with a minimum score of 6
- Not allowing end matches
- Action: Remove adapter

The result (shown in figure 19.8) would be the following (the retained parts are green):

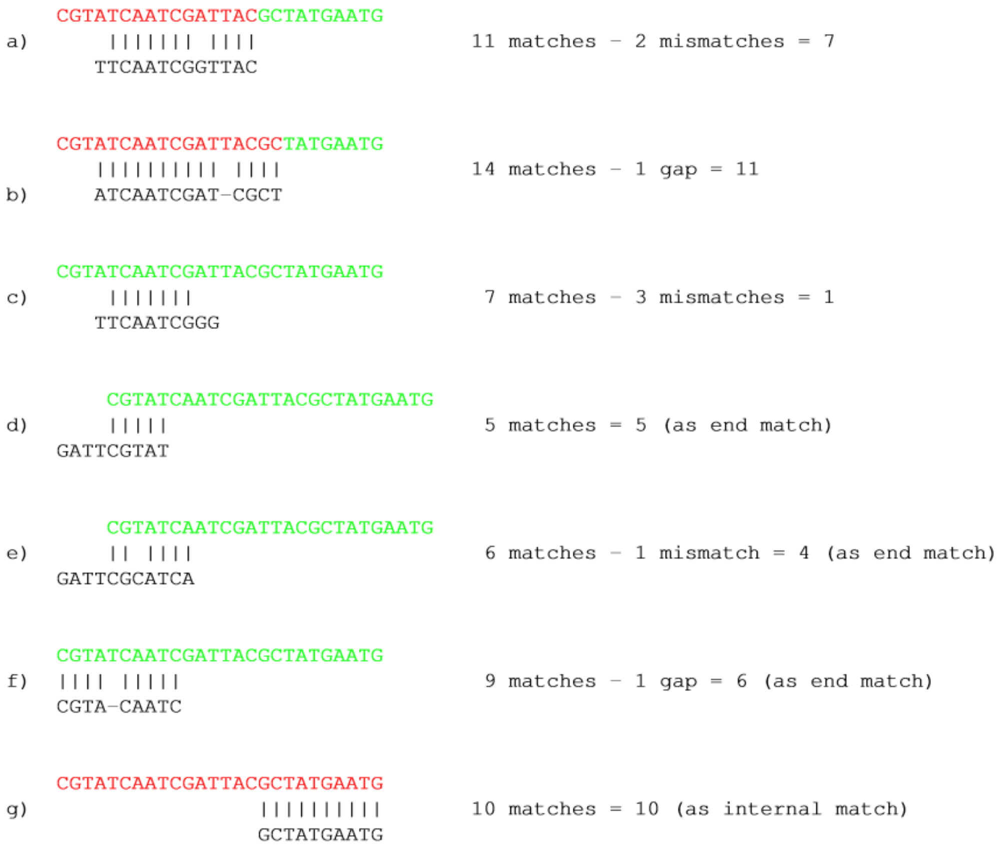
A different set of adapter settings could be:

- Allowing internal matches with a minimum score of 11
- Allowing end match with a minimum score of 4
- Action: Remove adapter

The result would be (shown in figure 19.9):

Strand settings

Each adapter is defined as either **Plus** or **Minus**. Note that all the definitions above regarding 3' end and 5' end also apply to the minus strand (i.e. selecting the Minus strand is equivalent to



```
    CGTATCAATCGATTACGCTATGAATG
a)      ||||||| ||||                 11 matches - 2 mismatches = 7
       TTCAATCGGTTAC

    CGTATCAATCGATTACGCTATGAATG
       |||||||||| ||||               14 matches - 1 gap = 11
b)     ATCAATCGAT-CGCT

    CGTATCAATCGATTACGCTATGAATG
c)      |||||||                       7 matches - 3 mismatches = 1
       TTCAATCGGG

        CGTATCAATCGATTACGCTATGAATG
d)      |||||                         5 matches = 5 (as end match)
    GATTCGTAT

        CGTATCAATCGATTACGCTATGAATG
e)      || ||||                       6 matches - 1 mismatch = 4 (as end match)
    GATTCGCATCA

    CGTATCAATCGATTACGCTATGAATG
f)  |||| |||||                        9 matches - 1 gap = 6 (as end match)
    CGTA-CAATC

    CGTATCAATCGATTACGCTATGAATG
g)                  ||||||||||       10 matches = 10 (as internal match)
                    GCTATGAATG
```

Figure 19.8: *The results of trimming with internal matches only. Red is the part that is removed and green is the retained part. Note that the read at the bottom is completely discarded.*

reverse complementing all the reads). The adapter in this case should be defined as you would see it on the plus strand of the reverse complemented read. The example below (figure 19.10) shows a few examples of an adapter defined on the minus strand. It shows hits for an adapter sequence defined as `CTGCTGTACGGCCAAGGCG`, searching on the minus strand.

You can see that if you reverse complemented the adapter you would find the hit on the plus strand, but then you would have trimmed the wrong end of the read. So it is important to define the adapter as it is, without reverse complementing.

Trimming of 3' ends of the reads

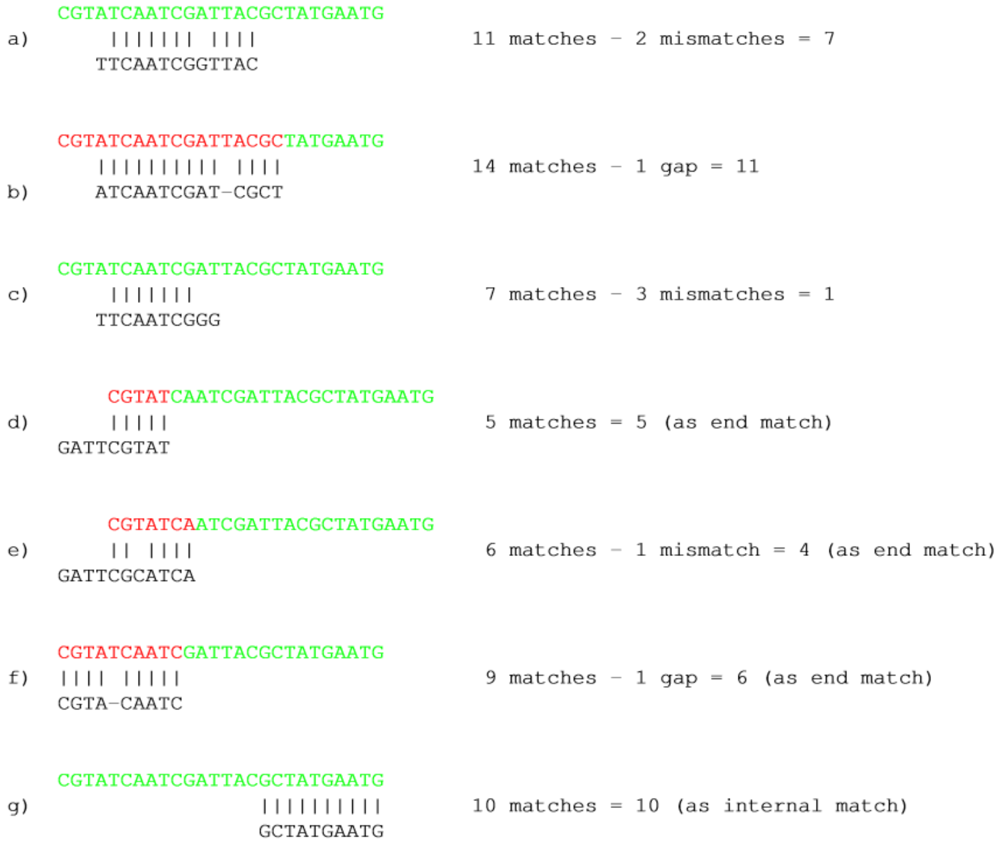
To trim an adapter and everything to the 3' end of the adapter you will need to search for the reverse complement of the adapter on the negative strand. This is achieved by creating a new Trim Adapter List from the reverse complement of your adapter sequence, choosing the minus strand of your reads and run adapter trimming with the new Trim Adapter List as input.

Other adapter trimming options

When you run the trim, you specify the adapter settings as shown in figure 19.11.

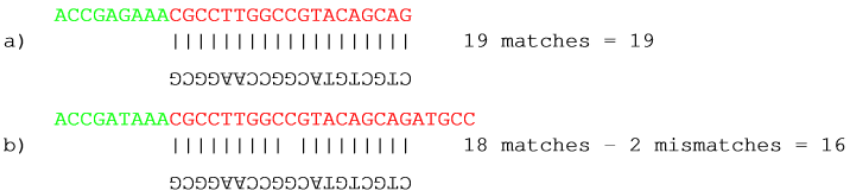
Select an trim adapter list (see section 19.2.2 on how to create an adapter list) that defines the adapters to use.

You can specify if the adapter trimming should be performed in **Color space**. Note that this



```
        CGTATCAATCGATTACGCTATGAATG
a)          ||||||| ||||                 11 matches - 2 mismatches = 7
           TTCAATCGGTTAC

        CGTATCAATCGATTACGCTATGAATG
           |||||||||| ||||               14 matches - 1 gap = 11
b)         ATCAATCGAT-CGCT

        CGTATCAATCGATTACGCTATGAATG
c)          |||||||                       7 matches - 3 mismatches = 1
           TTCAATCGGG

            CGTATCAATCGATTACGCTATGAATG
d)          |||||                         5 matches = 5 (as end match)
        GATTCGTAT

            CGTATCAATCGATTACGCTATGAATG
e)          || ||||                       6 matches - 1 mismatch = 4 (as end match)
        GATTCGCATCA

        CGTATCAATCGATTACGCTATGAATG
f)      |||| |||||                        9 matches - 1 gap = 6 (as end match)
        CGTA-CAATC

        CGTATCAATCGATTACGCTATGAATG
g)                      ||||||||||       10 matches = 10 (as internal match)
                        GCTATGAATG
```

Figure 19.9: *The results of trimming with both internal and end matches. Red is the part that is removed and green is the part that is retained part.*

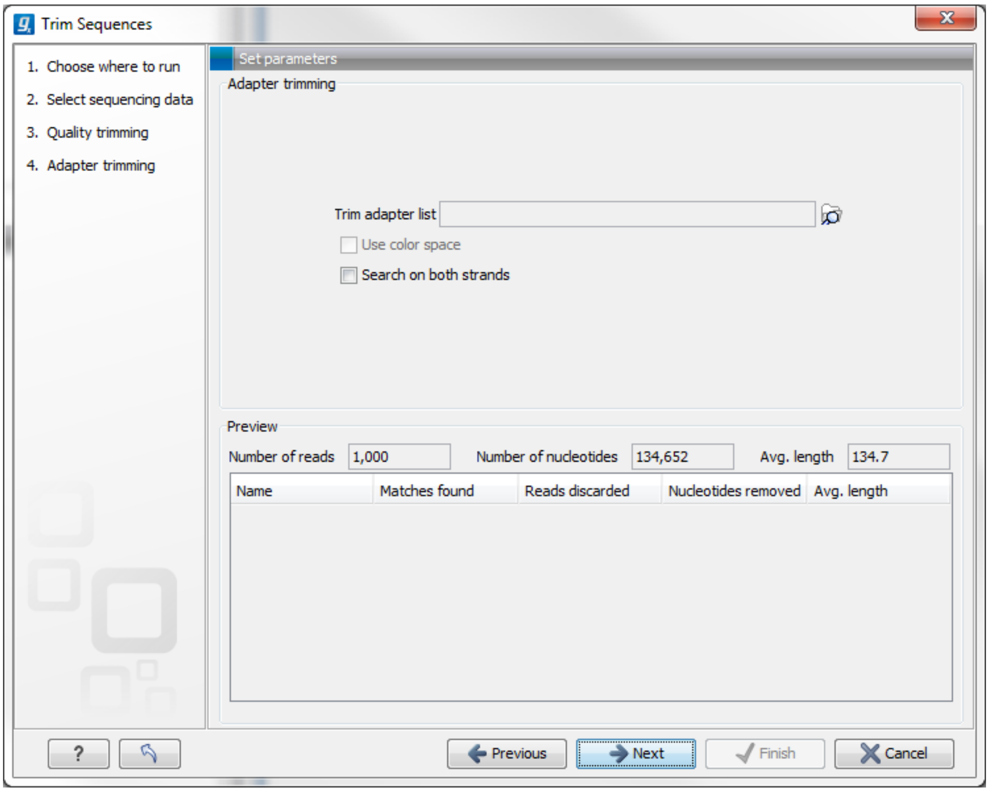


```
    ACCGAGAAACGCCTTGGCCGTACAGCAG
a)           |||||||||||||||||||     19 matches = 19

             CTGCTGTACGGCCAAGGCG

    ACCGATAAACGCCTTGGCCGTACAGCAGATGCC
b)           ||||||||| |||||||||     18 matches - 2 mismatches = 16

             CTGCTGTACGGCCAAGGCG
```

Figure 19.10: *An adapter defined as CTGCTGTACGGCCAAGGCG searching on the minus strand. Red is the part that is removed and green is the retained part. The retained part is 3' of the match on the minus strand, just like matches on the plus strand.*

option is only available for sequencing data imported using the SOLiD import (see section 6.3.3). When doing the trimming in color space, the Smith-Waterman alignment is simply done using colors rather than bases. The adapter sequence is still input in base space, and the Workbench then infers the color codes. Note that the scoring thresholds apply to the color space alignment (this means that a perfect match of 10 bases would get a score of 9 because 10 bases are represented by 9 color residues). Learn more about color space in section 20.5.

Checking the **Search on both strands** checkbox will search both the minus and plus strand for the adapter sequence. **Note!** If a match is found on the reverse strand the Trim action will reverse complement the read before trimming and output the trimmed reverse complement. Its intended use is for removal of multiplexing barcodes and primers.



Trim adapter list

□ Use color space

□ Search on both strands

Number of reads 1,000 Number of nucleotides 134,652 Avg. length 134.7

Name	Matches found	Reads discarded	Nucleotides removed	Avg. length

Figure 19.11: *Trimming your sequencing data for adapter sequences.*

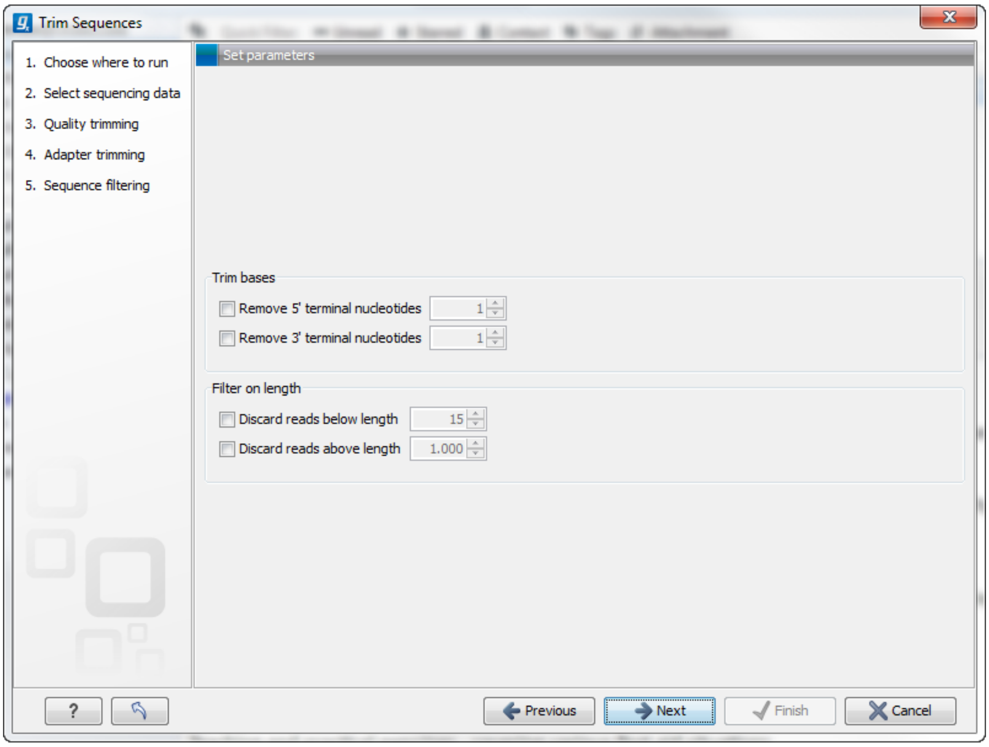
Below you find a preview listing the results of trimming with the current settings on 1000 reads in the input file (reads 1001-2000 when the read file is long enough). This is useful for a quick feedback on how changes in the parameters affect the trimming (rather than having to run the full analysis several times to identify a good parameter set). The following information is shown:

- **Name**. The name of the adapter.
- **Matches found**. Number of matches found based on the strand and alignment score settings.
- **Reads discarded**. This is the number of reads that will be completely discarded. This can either be because they are completely trimmed (when the **Action** is set to Remove adapter and the match is found at the 3' end of the read), or when the **Action** is set to Discard when found or Discard when not found.
- **Nucleotides removed**. The number of nucleotides that are trimmed include both the ones coming from the reads that are discarded and the ones coming from the parts of the reads that are trimmed off.
- **Avg. length** This is the average length of the reads that are retained (excluding the ones that are discarded).

Note that the preview panel is only showing how the adapter trim affects the results. If other kinds of trimming (quality or length trimming) is applied, this will not be reflected in the preview but still influence the results.

19.2.3 Length trimming

Clicking **Next** will allow you to specify length trimming as shown in figure 19.12.



Trim bases

□ Remove 5' terminal nucleotides 1

□ Remove 3' terminal nucleotides 1

Filter on length

□ Discard reads below length 15

□ Discard reads above length 1.000

Figure 19.12: *Trimming on length.*

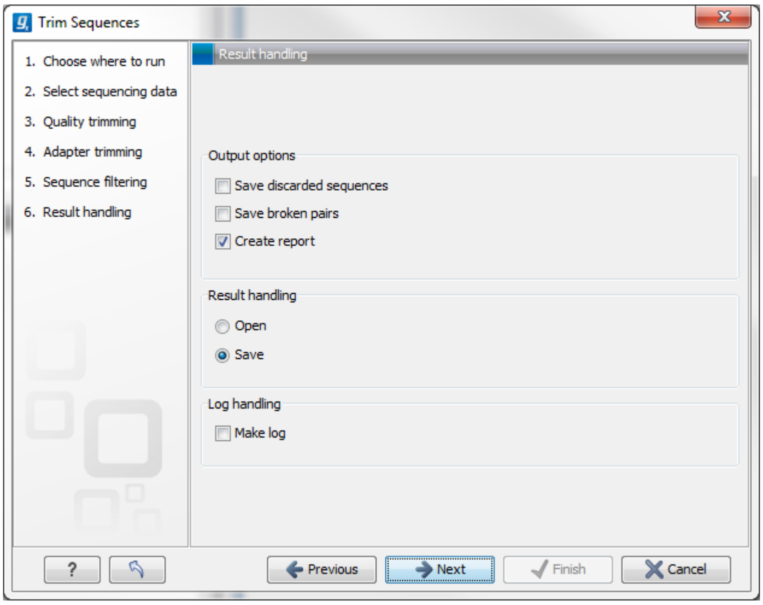
At the top you can choose to **Trim bases** by specifying a number of bases to be removed from either the 3' or the 5' end of the reads. Below you can choose to **Discard reads below length**. This can be used if you wish to simply discard reads because they are too short. Similarly, you can discard reads above a certain length. This will typically be useful when investigating e.g. small RNAs (note that this is an integral part of the small RNA analysis together with adapter trimming).

19.2.4 Trim output

Clicking **Next** will allow you to specify the output of the trimming as shown in figure 19.13.

No matter what is chosen here, the list of trimmed reads will always be produced. In addition the following can be output as well:

- **Create list of discarded sequences**. This will produce a list of reads that have been discarded during trimming. Sections trimmed from reads that are not themselves discarded will not appear in this list.
- **Create report**. An example of a trim report is shown in figure 19.14. The report includes the following:
  - **Trim summary.**
    - **Name.** The name of the sequence list used as input.



Output options

☐ Save discarded sequences

☐ Save broken pairs

☑ Create report

Result handling

○ Open

● Save

Log handling

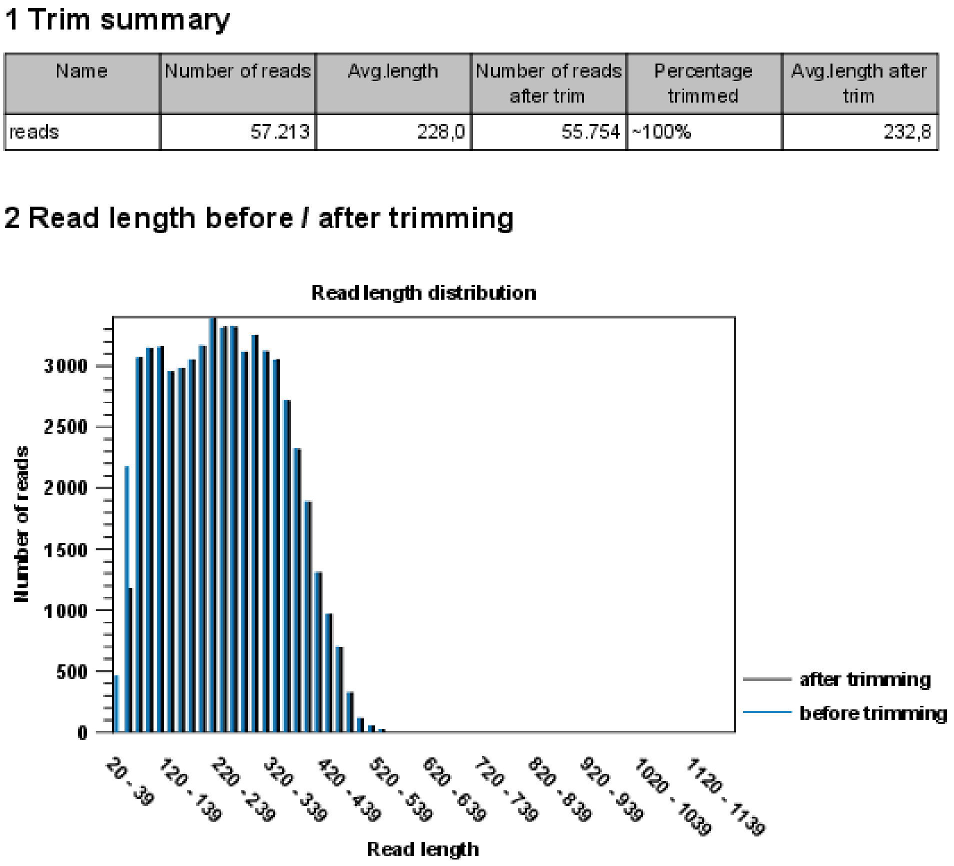
☐ Make log

Figure 19.13: *Specifying the trim output. No matter what is chosen here, the list of trimmed reads will always be produced.*

    - **Number of reads.** Number of reads in the input file.
    - **Avg. length.** Average length of the reads in the input file.
    - **Number of reads after trim.** The number of reads retained after trimming.
    - **Percentage trimmed.** The percentage of the input reads that are retained.
    - **Avg. length after trim.** The average length of the retained sequences.
  - **Read length before / after trimming**. This is a graph showing the number of reads of various lengths. The numbers before and after are overlayed so that you can easily see how the trimming has affected the read lengths (right-click the graph to open it in a new view).
  - **Trim settings** A summary of the settings used for trimming.
  - **Detailed trim results**. A table with one row for each type of trimming:
    - **Input reads.** The number of reads used as input. Since the trimming is done sequentially, the number of retained reads from the first type of trim is also the number of input reads for the next type of trimming.
    - **No trim.** The number of reads that have been retained, unaffected by the trimming.
    - **Trimmed.** The number of reads that have been partly trimmed. This number plus the number from **No trim** is the total number of retained reads.
    - **Nothing left or discarded.** The number of reads that have been discarded either because the full read was trimmed off or because they did not pass the length trim (e.g. too short) or adapter trim (e.g. if **Discard when not found** was chosen for the adapter trimming).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. This will start the trimming process.

If you trim paired data, the result will be a bit special. In the case where one part of a paired read has been trimmed off completely, you no longer have a valid paired read in your sequence list.



1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
reads	57.213	228,0	55.754	~100%	232,8

2 Read length before / after trimming

Figure 19.14: *A report with statistics on the trim results.*

In order to use paired information when doing assembly and mapping, the Workbench therefore creates two separate sequence lists: one for the pairs that are intact, and one for the single reads where one part of the pair has been deleted. When running assembly and mapping, simply select both of these sequence lists as input, and the Workbench will automatically recognize that one has paired reads and the other has single reads.

19.3 Demultiplex reads

Multiplexing techniques are often used When sequencing of different samples in one sequencing run. One method used is to *tag* the sequences with a unique identifier during the preparation of the sample for sequencing [Meyer et al., 2007].

With this technique, each sequence read will have a sample-specific tag, which is a specific sequence of nucleotides before and after the sequence of interest. This principle is shown in figure 19.15 (please refer to [Meyer et al., 2007] for more detailed information).

The sample-specific tag - also called the barcode - can then be used to distinguish between the different samples when analyzing the sequence data.

The post-processing of the sequencing data to separate the reads into their corresponding samples based on their barcodes, can be done using the demultiplexing functionality of the *CLC Cancer Research Workbench*. Using this tool, sequences are associated with a particular samples when they contain an exact match to a particular barcode. Sequences that do not contain an exct

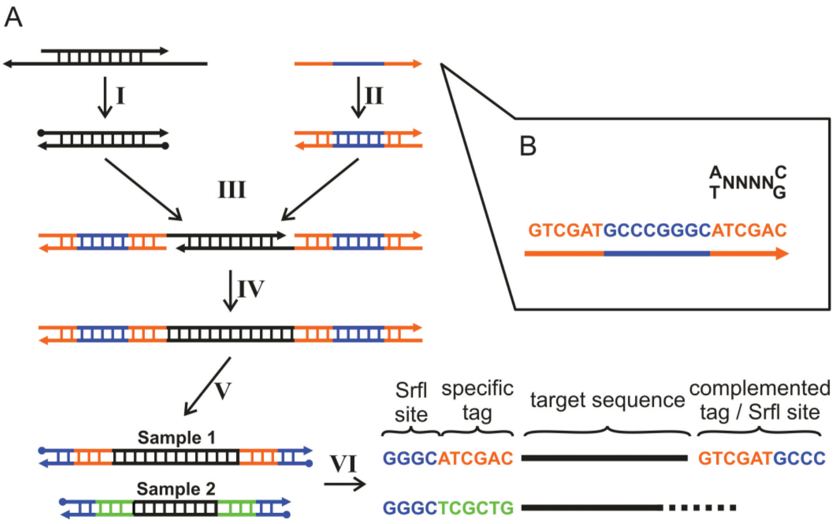


Figure 19.15: *Tagging the target sequence. Figure from [Meyer et al., 2007].*

match to any of the barcode sequences provided are classfied as not grouped and are put into a sequence list with the name "Not grouped".

Note that there is also an example using Illumina data at the end of this section.

Before processing the data, you need to import it as described in section 6.3.

Please note that demultiplexing is often carried out on the sequencing machine so that the sequencing reads are already separated according to sample. This is often the best option, if it is available to you. Of course, in such cases, the data will not need to be demuliplexed again after import into the *CLC Cancer Research Workbench*.

To de-multiplex your data, please go to:

Toolbox | **Preparing Raw Data** ( ) | **Demultiplex Reads** ( )

This opens a dialog where you can specify the sequences to process.

When you click on the button labeled **Next**, you can then specify the details of how the demultiplexing should be performed. At the bottom of the dialog, there are three buttons, which are used to **Add**, **Edit** and **Delete** the elements that describe how the barcode is embedded in the sequences.

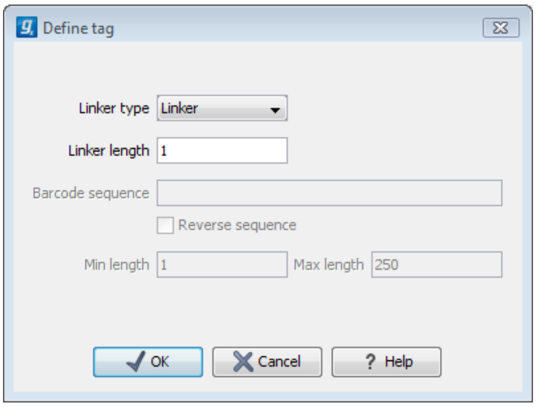
First, click **Add** to define the first element. This will bring up the dialog shown in 19.16.

At the top of the dialog, you can choose which kind of element you wish to define:

- **Linker**. This is a sequence which should just be ignored - it is neither the barcode nor the sequence of interest. Following the example in figure 19.15, it would be the four nucleotides of the *SrfI* site. For this element, you simply define its length - nothing else.

- **Barcode**. The barcode is the stretch of nucleotides used to group the sequences. In this dialog, you simply need to specify the length of the barcode. The valid sequences for your barcodes need to be provided at a later stage in setting up this job.

- **Sequence**. This element defines the sequence of interest. You can define a length interval for how long you expect this sequence to be. The sequence part is the only part of the read



Define tag

Linker type: Linker

Linker length: 1

Barcode sequence:

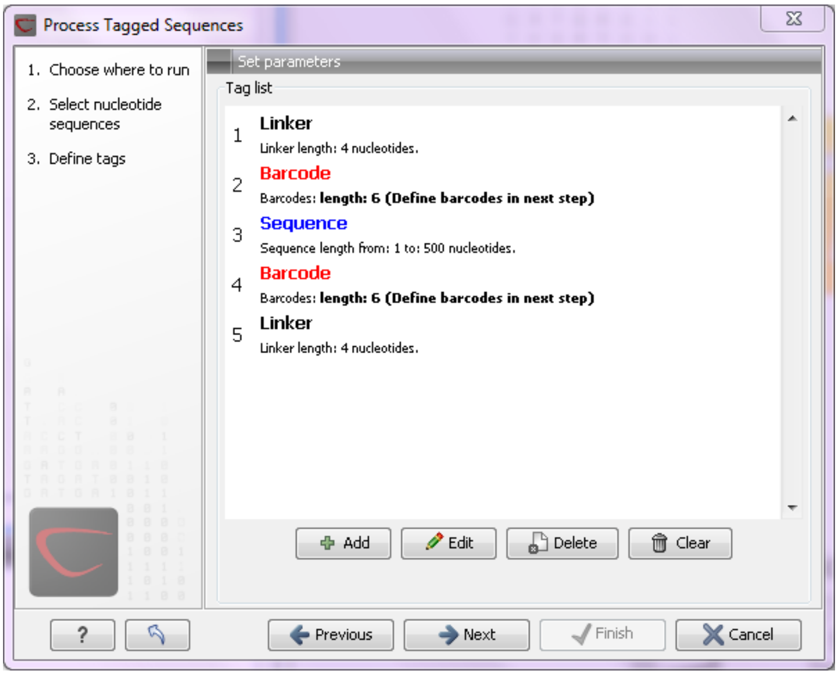
□ Reverse sequence

Min length: 1 Max length: 250

Figure 19.16: *Defining an element of the barcode system.*

that is retained in the output. Both barcodes and linkers are removed.

The concept when adding elements is that you add e.g. a linker, a barcode and a sequence in the desired sequential order to describe the structure of each sequencing read. You can of course edit and delete elements by selecting them and clicking the buttons below. For the example from figure 19.15, the dialog should include a linker for the *SrfI* site, a barcode, a sequence, a barcode (now reversed) and finally a linker again as shown in figure 19.17.



Process Tagged Sequences

1. Choose where to run
2. Select nucleotide sequences
3. Define tags

Set parameters

Tag list

1 Linker
Linker length: 4 nucleotides.

2 Barcode
Barcodes: **length: 6 (Define barcodes in next step)**

3 Sequence
Sequence length from: 1 to: 500 nucleotides.

4 Barcode
Barcodes: **length: 6 (Define barcodes in next step)**

5 Linker
Linker length: 4 nucleotides.

Figure 19.17: *Processing the tags as shown in the example of figure 19.15.*

If you have paired data, the dialog shown in figure 19.17 will be displayed twice - one for each part of the pair.

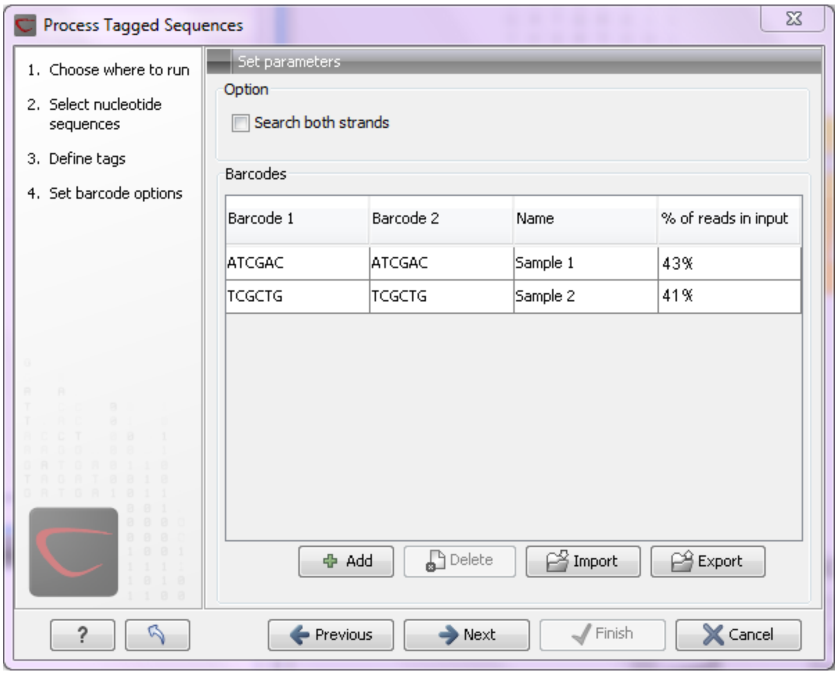
In case, where paired reads are expected to be barcoded in the same way (see example below), you would set the parameters for read1 (wizard step 3) and read2 (wizard step 4) to be the same.

Read1 : –Linker–Barcode1–Sequence

Read2 : –Linker–Barcode1–Sequence

However, if read2 of the pair is not expected to be the same as read1 in the pair, it is necessary to adjust these settings accordingly. For example, it is possible that read2 does not contain any barcode sequence at all. In this case, you would simply set the sequence parameter for the mate and exclude the barcode and linker parameters.

Clicking **Next** will display a dialog as shown in figure 19.18.



Option

□ Search both strands

Barcodes

Barcode 1	Barcode 2	Name	% of reads in input
ATCGAC	ATCGAC	Sample 1	43%
TCGCTG	TCGCTG	Sample 2	41%

Figure 19.18: *Specifying the barcodes as shown in the example of figure 19.15.*

Barcodes can be entered manually or imported from a properly formatted CSV or Excel file:

Manually The barcodes can be entered manually by clicking the **Add** () button. You can edit the barcodes and the names by clicking the cells in the table. The barcode name is used when naming the results.

Import from CSV or Excel To import a file of barcodes, click on the **Import** () button. The input format consists of two columns: the first contains the barcode sequence, the second contains the name of the barcode. An acceptable csv format file would contain columns of information that looks like:

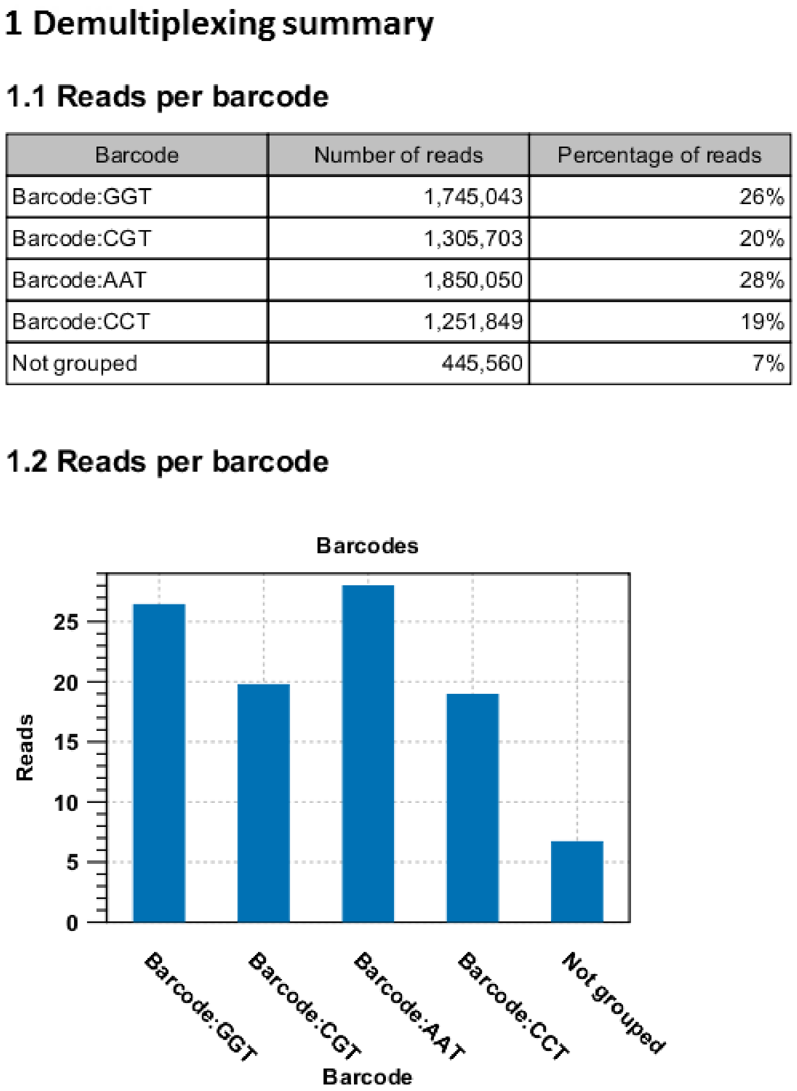
```
 "AAAAAA","Sample1"
"GGGGGG","Sample2"
 "CCCCCC","Sample3"
```

The **Preview** column will show a preview of the results by running through the first 10,000 reads.

At the top, you can choose to search on both strands for the barcodes (this is needed for some 454 protocols where the MID is located at either end of the read).

Click **Next** to specify the output options. First, you can choose to create a list of the reads that could not be grouped. Second, you can create a summary report showing how many reads were found for each barcode (see figure 19.19).

There is also an option to create subfolders for each sequence list. This can be handy when the



1 Demultiplexing summary

1.1 Reads per barcode

Barcode	Number of reads	Percentage of reads
Barcode:GGT	1,745,043	26%
Barcode:CGT	1,305,703	20%
Barcode:AAT	1,850,050	28%
Barcode:CCT	1,251,849	19%
Not grouped	445,560	7%

1.2 Reads per barcode

Barcodes
Reads
25
20
15
10
5
0
Barcode:GGT
Barcode:CGT
Barcode:AAT
Barcode:CCT
Not grouped
Barcode

Figure 19.19: *An example of a report showing the number of reads in each group.*

results need to be processed in batch mode (see section 8.1).

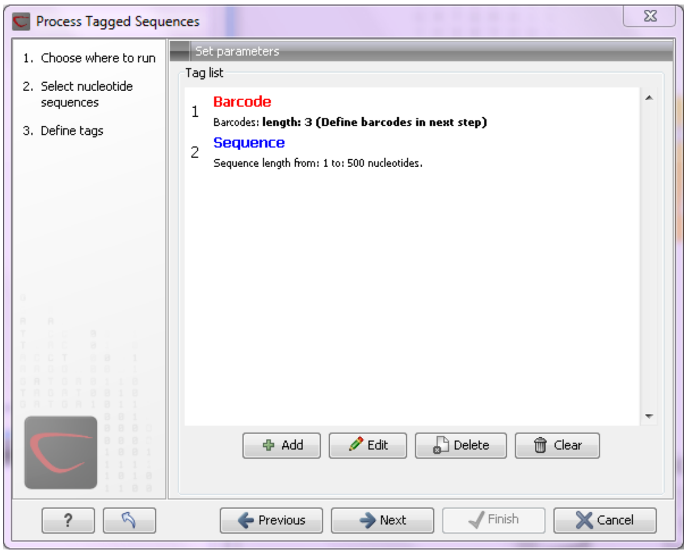
A new sequence list will be generated for each barcode containing all the sequences where this barcode is identified. Both the linker and barcode sequences are removed from each of the sequences in the list, so that only the target sequence remains. This means that you can continue the analysis by doing trimming or mapping. Note that you have to perform separate mappings for each sequence list.

An example using Illumina barcoded sequences

The data set in this example can be found at the Short Read Archive at NCBI: http://www.ncbi.nlm.nih.gov/sra/SRX014012. It can be downloaded directly in fastq format via the URL http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=dload&run_list=SRR030730&format=fastq. The file you download can be imported directly into the Workbench.

The barcoding was done using the following tags at the beginning of each read: CCT, AAT, GGT, CGT (see supplementary material of [Cronn et al., 2008] at http://nar.oxfordjournals.org/cgi/data/gkn502/DC1/1).

The settings in the dialog should thus be as shown in figure 19.20.

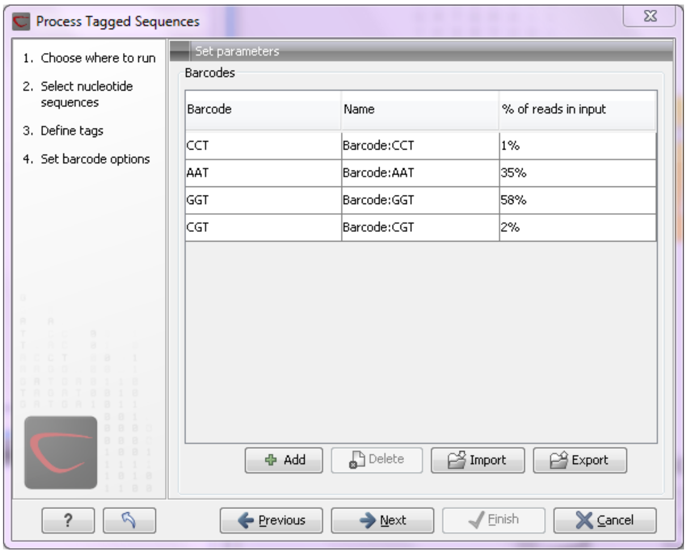


1 Barcode
Barcodes: length: 3 (Define barcodes in next step)

2 Sequence
Sequence length from: 1 to: 500 nucleotides.

Figure 19.20: *Setting the barcode length at three*

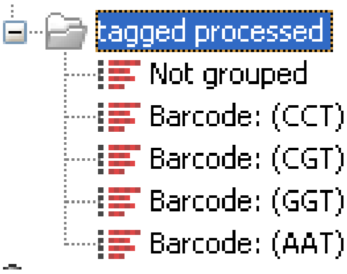
Click **Next** to specify the bar codes as shown in figure 19.21 (use the **Add** button).



Barcode	Name	% of reads in input
CCT	Barcode:CCT	1%
AAT	Barcode:AAT	35%
GGT	Barcode:GGT	58%
CGT	Barcode:CGT	2%

Figure 19.21: *A preview of the result*

With this data set we got the four groups as expected (shown in figure 19.22). The **Not grouped** list contains 445,560 reads that will have to be discarded since they do not have any of the barcodes.



tagged processed
- Not grouped
- Barcode: (CCT)
- Barcode: (CGT)
- Barcode: (GGT)
- Barcode: (AAT)

Figure 19.22: *The result is one sequence list per barcode and a list with the remainders*

Part VI

Resequencing analysis

Chapter 20

Resequencing analysis tools

Contents

20.1 Identify Known Mutations from Sample Mappings

The **Identify Known Mutations from Sample Mappings** tool can be used to look up known genomic variants in read mappings. This can be done in one or more samples by comparing a track of known variants with the read mappings of interest in order to test for the presence or absence of clinical (or other relevant) variants in e.g. patient samples.

The **Identify Known Mutations from Sample Mappings** tool does not perform any kind of variant calling, which means that this tool cannot be used to find de novo variants. Rather, the tool is intended for identification of variants that have already been reported and described regarding potential clinical relevance.

20.1.1 Input and Parameters

You need two types of input for the **Identify Known Mutations from Sample Mappings** tool:

1. A variant track that hold the specific variants that you wish to test for.
2. The read mapping(s) that you wish to check for the presence (or absence) of specific variants.

20.1.2 Output from the "Identify Known Mutations from Sample Mappings" tool

The **Identify Known Mutations from Sample Mappings** tool has two kinds of outputs:

- An overview track with information about:
  - whether the variant could be detected or not
  - whether the coverage was sufficient at the given position
  - the frequency of the variant in each sample.
- Individual output tracks for each sample that show the observed frequency, average base quality, forward/reverse read balance, zygosity and observed allele count.

20.1.3 How to run the "Identify Known Mutations from Sample Mappings" tool

To run the "Identify Known Mutations from Sample Mappings" tool go to the toolbox:

Toolbox | **Resequencing Analysis** () | **Identify Known Mutations from Sample Mappings** ()

This opens the wizard shown in figure 20.1.

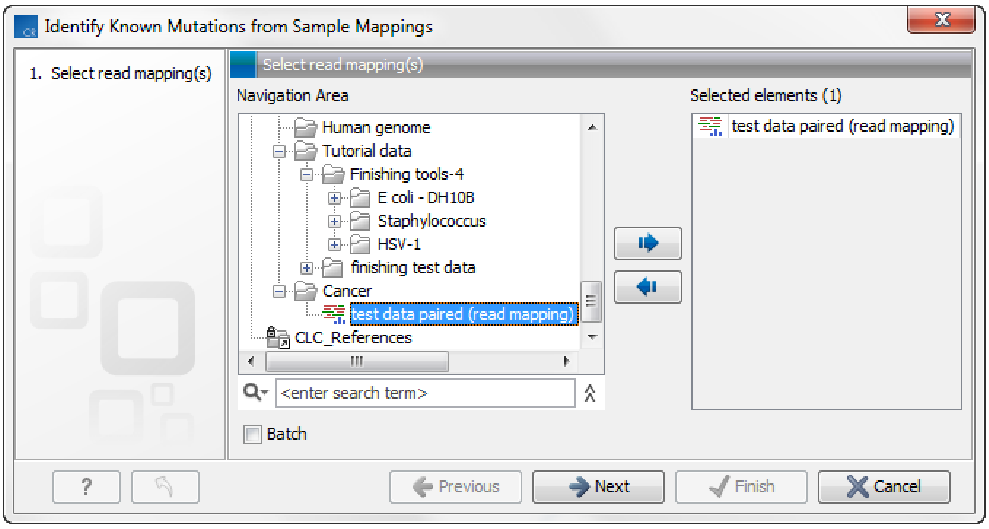


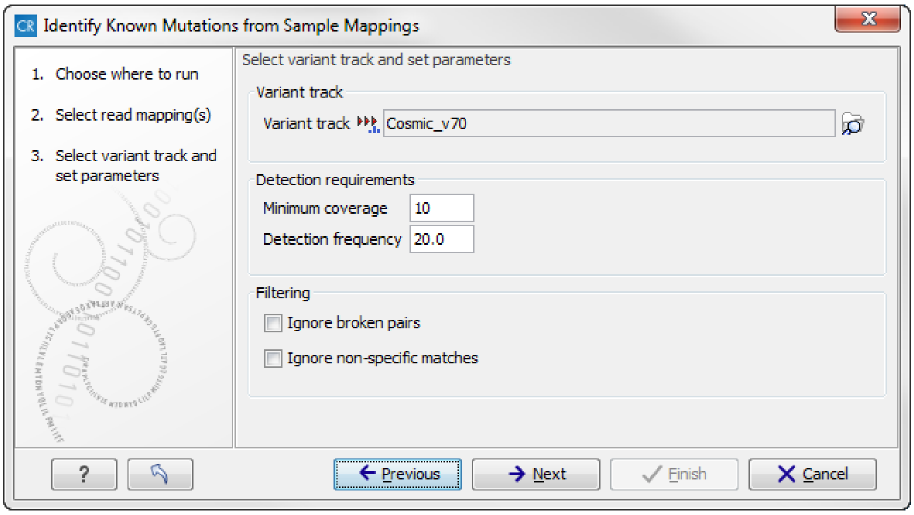
Figure 20.1: *Select the read mapping(s) to analyze.*

Select the read mapping to analyze and click on the button labeled **Next**.

In the next wizard that appears, you get the following options:

Variant track

- **Variant track** Select the variant track that contains the specific variants that you wish to test for in your read mapping (figure 20.2). **Note!** You can only select one variant track at the time. If you wish to compare with more than one variant track, you must run the analysis with each individual variant track at the time.



1. Choose where to run
2. Select read mapping(s)
3. Select variant track and set parameters

Select variant track and set parameters

Variant track

Variant track: Cosmic_v70

Detection requirements

Minimum coverage: 10

Detection frequency: 20.0

Filtering

□ Ignore broken pairs

□ Ignore non-specific matches

Figure 20.2: *Select the variant track with the variants that you wish to use for variant testing. In this example we will use "COSMIC".*

Detection requirements

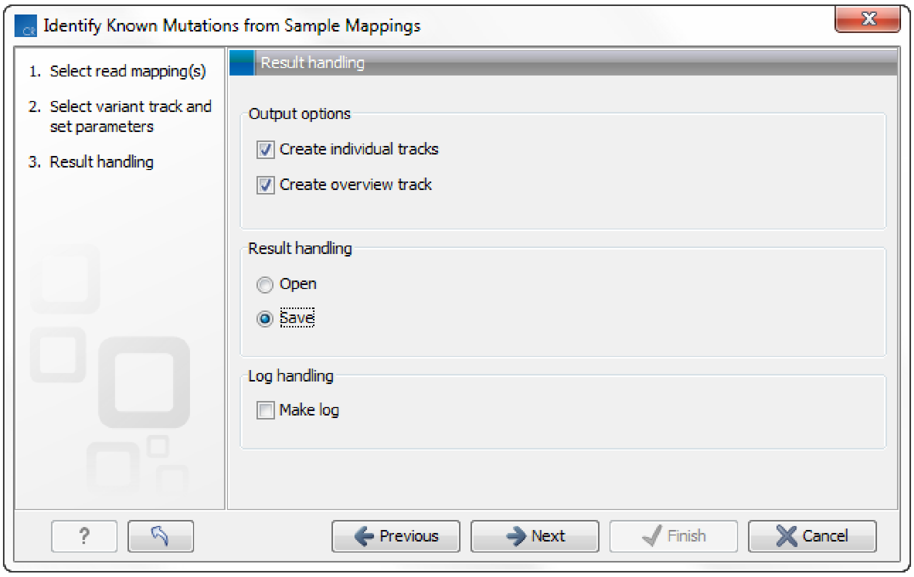
- **Minimum coverage** The minimum number of reads that covers the position of the variant, which is required to set "Sufficient Coverage" to YES.
- **Detection frequency** The minimum allele frequency that is required to annotate a variant as being present in the sample. The same threshold will also be used to determine if a variant is homozygous or heterozygous. In case the most frequent alternative allele at the position of the considered variant has a frequency of less than this value, the zygosity of the considered variant will be reported as being homozygous.

Filtering

- **Ignore broken pairs** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected.
- **Ignore non-specific matches** Reads that have an equally good match elsewhere on the reference genome (these reads are colored yellow in the mapping view) can be ignored in the analysis. Whether you include these reads or not will be a tradeoff between sensitivity and specificity. Including them may lead to the prediction of transcripts

that are not correct, whereas excluding them may mean that you will loose some true transcripts.

Click on the button labeled **Next** to go to the next wizard step (figure 20.3). At this step the output options can be adjusted.



Identify Known Mutations from Sample Mappings

Result handling

Output options

☑ Create individual tracks

☑ Create overview track

Result handling

○ Open

● Save

Log handling

□ Make log

1. Select read mapping(s)
2. Select variant track and set parameters
3. Result handling
?
Previous
Next
Finish
Cancel

Figure 20.3: *Select the desired output format(s). If using the default settings, two types of output will be generated; individual tracks and overview tracks.*

The output options are:

- **Create individual track** For each read mapping an individual track is created with the observed frequency, average base quality, forward/reverse read balance, zygosity and observed allele count.
- **Create overview track** The overview track is a summary for all samples with information about whether the coverage is sufficient at a given variant position and if the variant has been detected; the frequency of the variant.

Specify where to save the results and click on the button labeled **Finish**.

The individual sample track

For each mapping track from a sample, one individual sample output track will be created (figure 20.4). The track provides more detailed information about each variant tested in this specific sample.

The following information is annotated to the variant in the overview track:

- **("Sample name") Coverage** Either Yes or No, depending on whether the coverage at the position of the variant was higher or lower than the user given threshold for minimum coverage.
- **("Sample name") detection** Either Yes or No, depending on the minimum frequency settings chosen by the user.

- **("Sample name") frequency** The variant frequency observed in this sample.
- **("Sample name") zygosity** The zygosity observed in the sample. This setting is based on the minimum frequency setting made by the user. If this variant has been detected and the most frequent alternative allele at this position is also over the cutoff, the value is heterozygote.

The following information is annotated to the variant in the individual track:

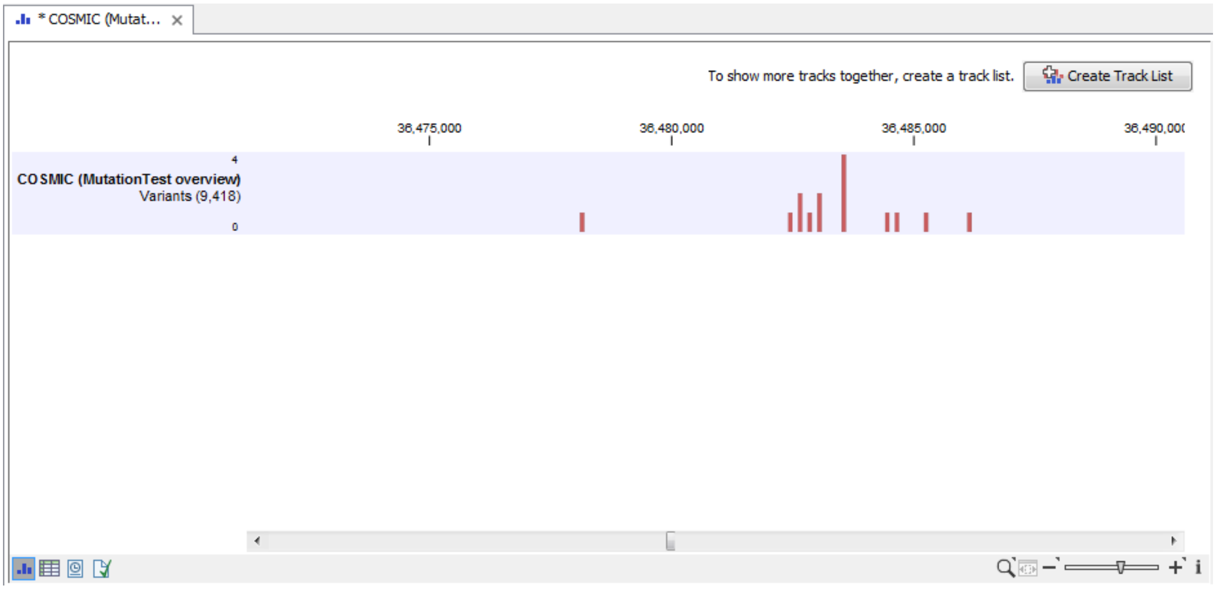
- **Zygosity** Homozygous or Heterozygous (based on the parameter "Detection frequency" setting)
- **Count** Number of reads supporting the variant
- **Frequency** Frequency of the reads supporting the variant
- **Average Quality** Average quality of all bases supporting the variant
- **Forward/Reverse balance** Minimum ratio of forward and reverse reads supporting the variant
- **MFAA count** Count of reads supporting the most frequent alternative allele at the position of the variant
- **MFAA frequency** Frequency of reads supporting the most frequent alternative allele at the position of the variant
- **MFAA forward read count** forward reads supporting the most frequent alternative allele at the position of the variant
- **MFAA reverse read count** reverse reads supporting the most frequent alternative allele at the position of the variant
- **MFAA forward/reverse balance** forward/reverse balance of the most frequent alternative allele at the position of the variant
- **MFAA average quality** average quality of the most frequent alternative allele at the position of the variant

At the bottom of the window it is possible to switch to a table view that lists all the mutations from the variant track that were found in your sample mapping. An example of the "Mutation Test overview" table can be seen in figure 20.5.

20.2 Trim primers of mapped reads

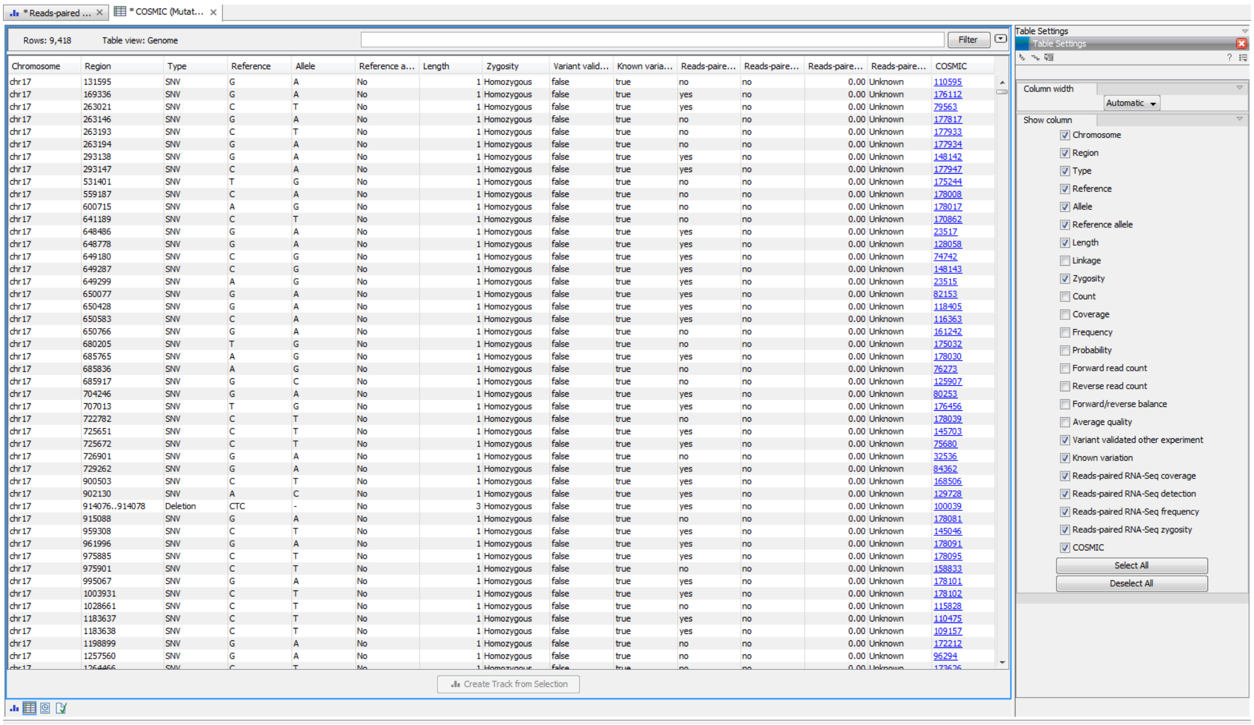
The **Trim primers of mapped reads** tool has been developed to targeted amplicon sequencing experiments with many targets (and as a consequence many primers). The tool makes use of the primer pairs in the trimming process, which greatly speeds up the trimming process time.

Removal of primers from the mapped reads ensures that no bias is introduced in the variant calling as would be the case if the primers were considered to be part of the sequencing reads.



To show more tracks together, create a track list.

Figure 20.4: *Summary output of the variant tester tool.*



Rows: 9,418   Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference a...	Length	Zygosity	Variant valid...	Known varia...	Reads-paire...	Reads-paire...	Reads-paire...	Reads-paire...	COSMIC
chr17	131595	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	110595
chr17	169336	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	176112
chr17	263021	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	79563
chr17	263146	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	177817
chr17	263193	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	177933
chr17	263194	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	177934
chr17	293138	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	148142
chr17	293147	SNV	C	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	177947
chr17	531401	SNV	T	G	No	1	Homozygous	false	true	no	no	0.00	Unknown	175244
chr17	559187	SNV	C	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	178008
chr17	600715	SNV	A	G	No	1	Homozygous	false	true	no	no	0.00	Unknown	178017
chr17	641189	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	170862
chr17	648486	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	23517
chr17	648778	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	128058
chr17	649180	SNV	C	G	No	1	Homozygous	false	true	yes	no	0.00	Unknown	74742
chr17	649287	SNV	C	G	No	1	Homozygous	false	true	yes	no	0.00	Unknown	148143
chr17	649299	SNV	A	G	No	1	Homozygous	false	true	yes	no	0.00	Unknown	23515
chr17	650077	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	82153
chr17	650428	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	118405
chr17	650583	SNV	C	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	116363
chr17	650766	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	161242
chr17	680205	SNV	T	G	No	1	Homozygous	false	true	no	no	0.00	Unknown	175032
chr17	685765	SNV	A	G	No	1	Homozygous	false	true	yes	no	0.00	Unknown	178030
chr17	685836	SNV	A	G	No	1	Homozygous	false	true	no	no	0.00	Unknown	76273
chr17	685917	SNV	G	C	No	1	Homozygous	false	true	no	no	0.00	Unknown	125907
chr17	704246	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	80253
chr17	707013	SNV	T	G	No	1	Homozygous	false	true	yes	no	0.00	Unknown	176456
chr17	722782	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	178039
chr17	725651	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	145703
chr17	725672	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	75680
chr17	726901	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	32536
chr17	729262	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	84362
chr17	900503	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	168506
chr17	902130	SNV	A	C	No	1	Homozygous	false	true	yes	no	0.00	Unknown	129728
chr17	914076..914078	Deletion	CTC	-	No	3	Homozygous	false	true	yes	no	0.00	Unknown	100039
chr17	915088	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	178081
chr17	959308	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	145046
chr17	961996	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	178091
chr17	975885	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	178095
chr17	975901	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	158833
chr17	995067	SNV	G	A	No	1	Homozygous	false	true	yes	no	0.00	Unknown	178101
chr17	1003931	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	178102
chr17	1028661	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	115828
chr17	1183637	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	110475
chr17	1183638	SNV	C	T	No	1	Homozygous	false	true	yes	no	0.00	Unknown	109157
chr17	1198899	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	172212
chr17	1257560	SNV	G	A	No	1	Homozygous	false	true	no	no	0.00	Unknown	96294
chr17	1264466	SNV	C	T	No	1	Homozygous	false	true	no	no	0.00	Unknown	173626

Table Settings

Column width: Automatic

Show column

- ☑ Chromosome
- ☑ Region
- ☑ Type
- ☑ Reference
- ☑ Allele
- ☑ Reference allele
- ☑ Length
- □ Linkage
- ☑ Zygosity
- □ Count
- □ Coverage
- □ Frequency
- □ Probability
- □ Forward read count
- □ Reverse read count
- □ Forward/reverse balance
- □ Average quality
- ☑ Variant validated other experiment
- ☑ Known variation
- ☑ Reads-paired RNA-Seq coverage
- ☑ Reads-paired RNA-Seq detection
- ☑ Reads-paired RNA-Seq frequency
- ☑ Reads-paired RNA-Seq zygosity
- ☑ COSMIC

Select All

Deselect All

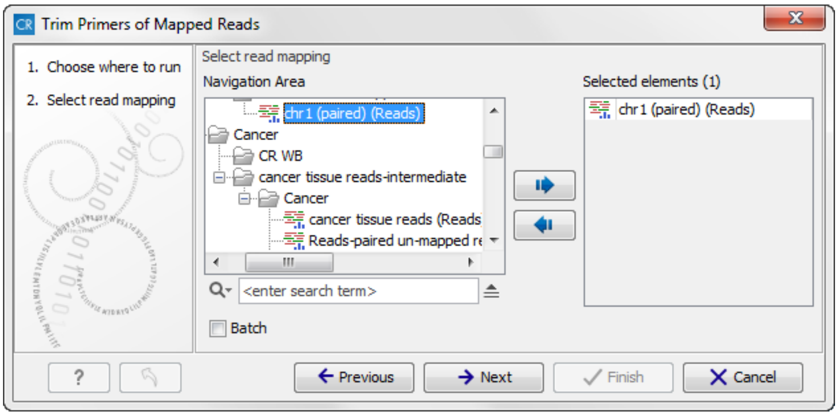
Figure 20.5: *Overview output of the "Identify Known Mutations from Sample Mappings" tool.*

To be able to trim off the primers used in your sequencing experiment you must know the primer sequences as you will need to specify which target primer sequence file to use.

The **Trim Primers of Mapped Reads** can be found in the toolbox:

Toolbox | **Resequencing Analysis** ( ) | **Trim Primers of Mapped Reads** ( )

This will open the wizard shown in figure 20.6. In the first wizard step you are asked to select the read mapping. If you would like to analyze more than one read mapping, you can choose to run the analysis in batch mode by ticking the "Batch" box in the lower left corner of the wizard and then selecting the folder that hold the read mappings you want to analyze.



Trim Primers of Mapped Reads

1. Choose where to run
2. Select read mapping

Select read mapping

Navigation Area

- chr1 (paired) (Reads)
- Cancer
  - CR WB
  - cancer tissue reads-intermediate
    - Cancer
      - cancer tissue reads (Reads
      - Reads-paired un-mapped re

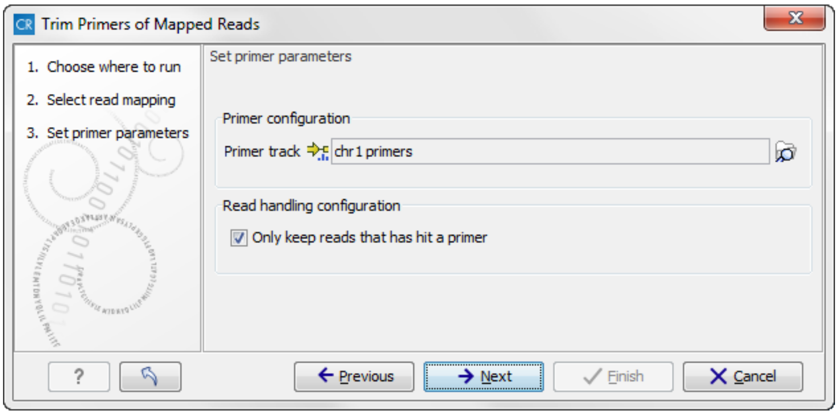
Selected elements (1)

- chr1 (paired) (Reads)

□ Batch

Figure 20.6: *Select files to import.*

Click on the button labeled **Next** to go to the next wizard step (see figure 20.7).



Trim Primers of Mapped Reads

1. Choose where to run
2. Select read mapping
3. Set primer parameters

Set primer parameters

Primer configuration

Primer track: chr1 primers

Read handling configuration

☑ Only keep reads that has hit a primer

Figure 20.7: *Select your primer location file and choose whether you want to keep or discard reads with no matching primers.*

- **Primer track** Click on the folder icon in the right side to select your primer location file.
- **Read handling configuration** If you tick "Only keep reads that have hit a primer", reads with no matching primers will be discarded.

Click on the button labeled **Next** to go to the next wizard step, choose to save the result of the primer trimming and click on the button labeled **Finish**. The output corresponds to the input with the only difference that the primers have been trimmed off and that the output file has "trimmed reads" appended to the name.

20.3 Map Reads to Reference

Read mapping is a very fundamental step in most applications of high-throughput sequencing data. The *CLC Cancer Research Workbench* includes read mapping in several other tools (e.g. in the RNA-Seq Analysis), but this chapter will focus on the core read mapping algorithm. At the end of the chapter you can find descriptions of the read mapping reports and a tool to merge read mappings.

There are two different versions of the core mapper: one for color space data, and one for base space data. At http://www.clcbio.com/white-paper you can find white papers with detailed benchmarks and descriptions of both algorithms.

The following description focuses on the parameters that can be directly influenced by the user.

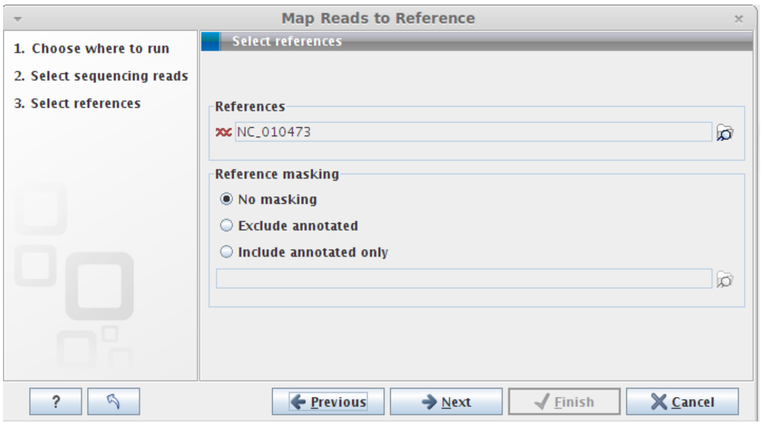
20.3.1 Selecting reads and reference

To start the read mapping:

> **Toolbox | Resequencing Analysis ( ) | Map Reads to Reference ( )**

In this dialog, select the sequences or sequence lists containing the sequencing data. Note that the reference sequences should be selected in the next step.

When the sequences are selected, click **Next**, and you will see the dialog shown in figure 20.8.



Map Reads to Reference

1. Choose where to run
2. Select sequencing reads
3. Select references

Select references

References

NC_010473

Reference masking

● No masking

○ Exclude annotated

○ Include annotated only

Figure 20.8: *Specifying the reference sequences and masking.*

At the top you select one or more reference sequences by clicking the **Browse and select element** ( ) button. You can select either single sequences, a list of sequences or a sequence track as reference.

20.3.2 Including or excluding regions (masking)

The next part of the dialog shown in figure 20.8 lets you *mask* the reference sequences. Masking refers to a mechanism where parts of the reference sequence are not considered in the mapping. This can be useful for example when mapping data is captured from specific regions (e.g. for amplicon resequencing). The read mapping will still base its output on the full reference - it is only the core read mapping that ignores regions.

Masking is performed by discarding the masked out nucleotides. As a result the reference is split into separate sequences, which are positioned according to the original unmasked reference sequence.

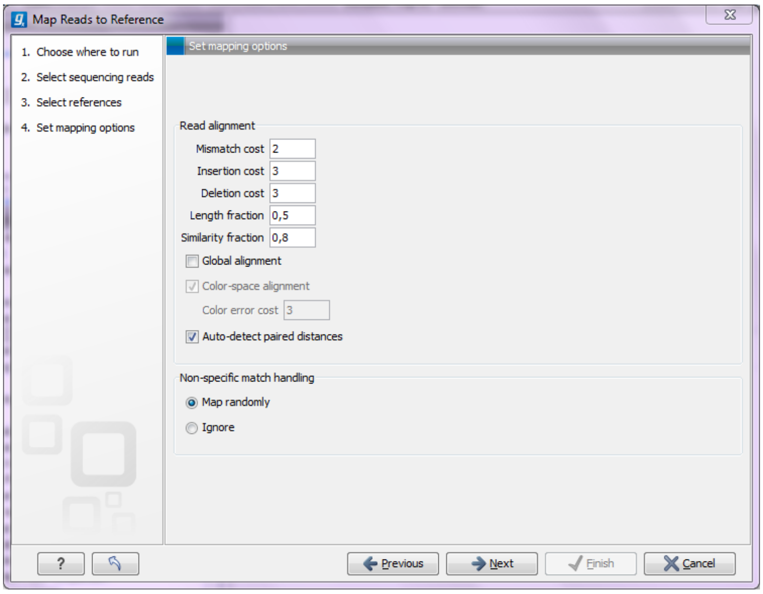
Note that you should be careful that your data is indeed only sequenced from the target regions. If not, some of the reads that would have matched a masked-out region perfectly may be placed wrongly at another position with a less-perfect match and lead to wrong results for subsequent variant calling. For resequencing purposes, we recommend testing whether masking is appropriate by running the same data set through two rounds of read mapping and variant calling: one with masking and one without. At the end, comparing the results will reveal if any off-target sequences cause problems in the variant calling.

Masking out repeats or using other masks with many regions is not recommended. Repeats are handled well and does not cause any slowdown. On the contrary, masking repeats is likely to cause a dramatic slowdown in speed, increase memory requirements and lead to incorrect read placement.

To mask a reference sequence, first click the **Include** or **Exclude** options, and second click the **Browse** ( ) button to select a track to use for masking.

20.3.3 Mapping parameters

Clicking **Next** leads to the parameters for the read mapping (see figure 20.9).



Map Reads to Reference

1. Choose where to run
2. Select sequencing reads
3. Select references
4. Set mapping options

Set mapping options

Read alignment

Mismatch cost 2

Insertion cost 3

Deletion cost 3

Length fraction 0,5

Similarity fraction 0,8

□ Global alignment

☑ Color-space alignment

Color error cost 3

☑ Auto-detect paired distances

Non-specific match handling

● Map randomly

○ Ignore

Figure 20.9: *Setting parameters for the mapping.*

The first three parameters allow mismatch and gap costs to be adjusted:

Mismatch cost The cost of a mismatch between the read and the reference sequence.

Insertion cost The cost of an insertion in the read (a gap in the reference sequence).

Deletion cost The cost of having a gap in the read (an insertion in the reference sequence).

Ambiguous nucleotides such as "N", "R" or "Y" in read or reference sequences are treated as a mismatches.

The score for a match between the read and the reference is always 1. Adjusting the three cost parameters above can improve the mapping quality, e.g. when the read error rate is high or the reference is expected to differ significantly from the sequenced organism. For example, if the reads contain many insertions and/or deletions, it can be a good idea to lower the insertion and deletion costs to allow more of such errors in the reads. When adjusting these settings one should consider the possible drawbacks. For example, reducing the insertion and deletion cost increases the risk of mapping reads to the wrong positions in the reference.

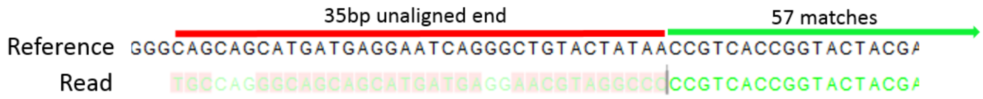

35bp unaligned end
57 matches
Reference GGGCAGCAGCATGATGAGGAATCAGGGCTGTACTATAACCGTCACCGGTACTACGA
Read TGCCAGGGCAGCAGCATGATGAGGAACGTAGGCCCCCGTCACCGGTACTACGA

Figure 20.10: *An alignment of a read where a region of 35bp at the start of the read is unaligned while the remaining 57 nucleotides matches the reference.*

Figure 20.10 shows an example where the read mapper is unable to map a region in a read due to insertions in the read and mismatches between the read and the reference. The aligned region of the read has a total of 57 matching nucleotides which result in an alignment score of 57 which is optimal when using the default cost for insertions and mismatches (2 and 3 respectively). If the mapper had aligned the remaining 35bp of the read as shown in Figure 20.11 using the default scoring scheme, the score would become:

$$(26 + 1 + 3 + 57) * 1 - 5 * 2 - 8 * 3 = 53 \tag{20.1}$$

In this case the alignment shown in Figure 20.10 is optimal since it has the highest score. However, if either the cost of deletions or mismatches were reduced by one, the score of the alignment shown in Figure 20.11 would become 61 and 58 respectively and thus make it optimal.

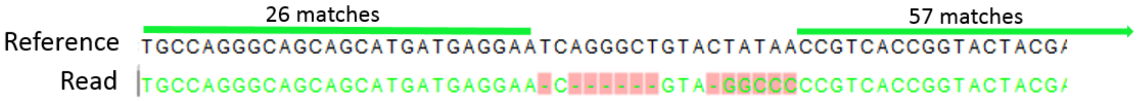

26 matches
57 matches
Reference TGCCAGGGCAGCAGCATGATGAGGAATCAGGGCTGTACTATAACCGTCACCGGTACTACGA
Read TGCCAGGGCAGCAGCATGATGAGGAA-C------GTA-GGCCCCCGTCACCGGTACTACGA

Figure 20.11: *An alignment of a read containing a region with several mismatches and deletions. By reducing the default cost of either mismatches or deletions the read mapper can make an alignment that spans the full length of the read.*

Once the optimal alignment of the read is found, based on the costs parameters described above, a filtering process determines whether this match is good enough for the read to be included in the output. The filtering threshold is determined by two factors:

Length fraction The minimum percentage of the total alignment length that must match the reference sequence at the selected similarity fraction. A fraction of 0.5 means that at least half of alignment must match the reference sequence before the read is included in the mapping (if the similarity fraction is set to 1). Note, that the minimal seed (word) size for read mapping is 15 bp, so reads shorter than this will not be mapped.

Similarity fraction The minimum percentage identity between the aligned region of the read and the reference sequence. For example, if the identity should be at least 80% for the read to be included in the mapping, set this value to 0.8. Note that the similarity fraction relates to the length fraction, i.e. when the length fraction is set to 50% then *at least 50% of the alignment must have at least 80% identity* (see figure 20.12).

By default, mapping is done with **local alignment** of the reads to the reference. The advantage of performing local alignment instead of global alignment is that the ends are automatically left unaligned if there are many differences from the reference at the ends. For many sequencing platforms, the quality of the bases drop along the read, and a local alignment approach is desirable. Note that the aligned region has to be greater than the length threshold set. If **global alignment** is preferred, it can be enabled with a checkbox as shown in figure 20.9.

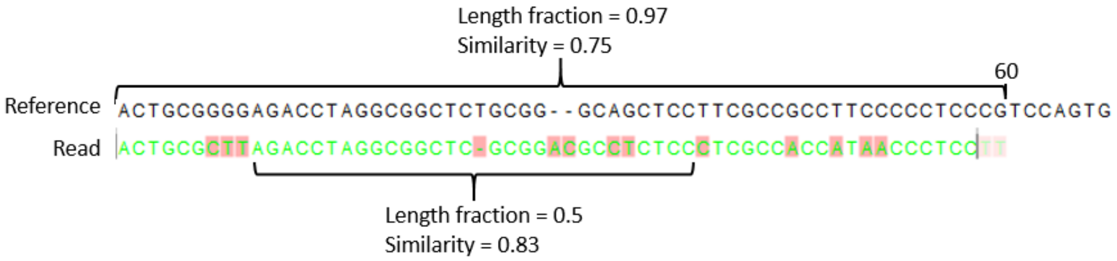

Length fraction = 0.97
Similarity = 0.75
60
Reference
ACTGCGGGGAGACCTAGGCGGCTCTGCGG--GCAGCTCCTTCGCCGCCTTCCCCCTCCCGTCCAGTG
Read
ACTGCGCTTAGACCTAGGCGGCTC-GCGGACGCCTCTCCCTCGCCACCATAACCCTCCTT
Length fraction = 0.5
Similarity = 0.83

Figure 20.12: *A read containing 59 nucleotides where the total alignment length is 60. The part of the alignment that gave rise to the optimal score has length 58 which excludes 2 bases at the left end of the read. The length fraction of the matching region in this example is therefore 58/60 = 0.97. Given a minimum length fraction of 0.5, the similarity fraction of the alignment is computed as the maximum similarity fraction of any part of the alignment which constitute at least 50% of the total alignment. In this example the marked region in the alignment with length 30 (50% of the alignment length) has a similarity fraction of 0.83 which will satisfy the default minimum similarity fraction requirement of 0.8.*

When mapping data in color space (data from SOLiD systems), the **color space** checkbox is enabled, and a corresponding cost for color errors can be set. If you do not have color space data, these will be disabled and are not relevant. For more details about this, please see section 20.5 which explains how color space mapping is performed in greater detail.

Mapping paired reads

At the bottom of the dialog shown in figure 20.9 you can specify how **Paired reads** should be handled. You can read more about how paired data is imported and handled in section 6.3.8. If the sequence list used as input contains paired reads, this option will automatically be enabled - if it contains single reads, this option will not be applicable.

The *CLC Cancer Research Workbench* offers as the default choice to automatically calculate the distance between the pairs. If this is selected, the distance is estimated in the following way:

1. A sample of 100,000 reads is extracted randomly from the full data set and mapped against the reference using a very wide distance interval.
2. The distribution of distances between the paired reads is analyzed, and an appropriate distance interval is selected:
    - If less than 10,000 reads map, a simple calculation is used where the minimum distance is one standard deviation below the average distance, and the maximum distance is one standard deviation above the average distance.
    - If more than 10,000 reads map, a more sophisticated method is used which investigates the shape of the distribution and finds the boundaries of the peak.
3. The full sample is mapped using this distance interval.
4. The **history** (🕓) of the result records the distance interval used.

The above procedure will be run for each sequence list used as input, assuming that they do not necessarily share the same library preparation and could have different distributions of paired

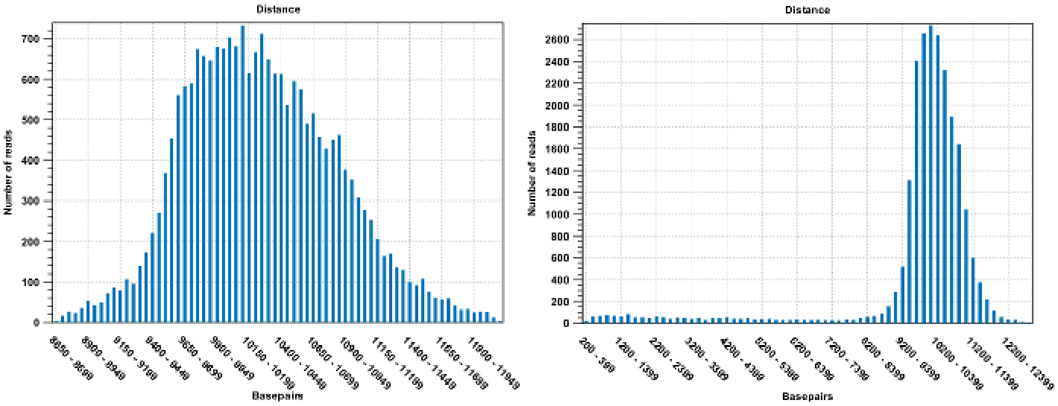


Figure 20.13: *To the left: mapping with a narrower distance interval estimated by the workbench. To the right: mapping with a large paired distance interval (note the large right tail of the distribution).*

distances. Figure 20.13 shows an example of the distribution of intervals with and without automatic pair distance interval estimation.

Sometimes the automatic estimation of the distance between the pairs may return a warning "multiple intervals detected". This may happen if e.g. the reads derive from multiple libraries or from certain types of amplicon sequencing protocols. In this case, the estimates may still be correct, but, if in doubt, the user may want to disable the option to automatically estimate paired distances and instead manually specify minimum and maximum distances between pairs on the input sequence list.

If the automatic detection of paired distances is not checked, the mapper will use the information about minimum and maximum distance recorded on the input sequence lists (see section 6.3.8).

The 'automatic detection of paired distance' option when mapping should be used with caution. It is possible that the estimated distance setting is too narrow and consequently many read pairs will be flagged broken. Sometimes, a second peak in the Paired Distance distribution graph is not picked up on by the estimation tool.

If a large portion of pairs are flagged 'Broken' we recommend the following:

1. Inspect the detailed mapping report (see section **??**) to deduce a distance setting interval - and compare this to the estimated distance used by the mapper (found in the mapping history).

2. Open the paired reads list and set a broad paired distance in the Elements tab. Then run a new mapping with the 'auto-detect...' OFF. Make sure to have a report produced. Open this report and look at the Paired Distance Distribution graph. This will tell you the distances that your pairs did map with. Use this information to narrow down the distance setting and perhaps run a third mapping using this.

3. Another cause of excessive amounts of broken pairs is misspecification of the read pair orientation. This can be changed in the Elements tab of the paired reads list prior to running a mapping.

See (section 18.3) for further information about the mapping reports.

When a paired distance interval is set, the following approach is used for determining the placement of read pairs:

- First, all the optimal placements for the two individual reads are found.
- Then, the allowed placements according to the paired distance interval are found.
- If both reads can be placed independently but no pairs satisfies the paired criteria, the reads are treated as independent and marked as a **broken pair**.
- If only one pair of placements satisfy the criteria, the reads are placed accordingly and marked as uniquely placed even if either read may have multiple optimal placements.
- If several placements satisfy the paired criteria, the pair is treated as a non-specific match (see section 20.3.3 for more information.)
- If one read is uniquely mapped but the other read has several placements that are valid given the distance interval, the mapper chooses the location that is closest to the first read.

Non-specific matches

At the bottom of the dialog, you can specify how **Non-specific matches** should be treated. The concept of Non-specific matches refers to a situation where a read aligns at *more than one position with an equally good score*. In this case you have two options:

- **Random**. This will place the read in one of the positions randomly.
- **Ignore**. This will not include the read in the final mapping.

Note that a read is only considered non-specific when the read matches equally well at several alignment positions. If there are e.g. two possible alignment positions and one of them is a perfect match and the other involves a mismatch, the read is placed at the position with the perfect match and it is not marked as a non-specific match.

For paired data, reads are only considered non-specific matches if the entire pair could be mapped elsewhere with equal scores for both reads, or if the pair is broken in which case a read can be categorized as non-specific in the same way as single reads (see section 20.3.3).

When looking at the mapping, the default color for non-specific matches is yellow.

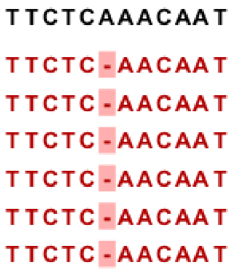
20.3.4 Gap placement

In the case of insertions or deletions in homopolymeric or repetitive regions, the precise placement of the insertion or deletion cannot be determined from the data. An example is shown in figure 20.14.

In this example, three A's in the reference (top) have been replaced by two A's in the reads (shown in red). The gap is placed towards the 5' end (left side), but could have been placed towards the 3' end with an equally good mapping score for the read as shown in figure 20.15.

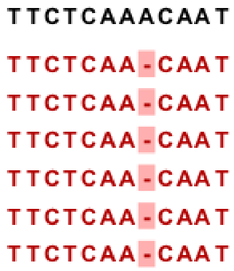
Since either way of placing the gap is arbitrary, the goal of the mapper is to place the gaps consistently at the same side for all reads.

Many insertions and deletions in homopolymeric or repetitive regions reported in the public databases dbSNP and 1000Genomes have been identified based on mappings done with tools



```
TTCTCAAACAAT
TTCTC-AACAAT
TTCTC-AACAAT
TTCTC-AACAAT
TTCTC-AACAAT
TTCTC-AACAAT
TTCTC-AACAAT
```

Figure 20.14: *Three A's in the reference (top) have been replaced by two A's in the reads (shown in red). The gap is placed towards the 5' end, but could have been placed towards the 3' end with an equally good mapping score for the read.*



```
TTCTCAAACAAT
TTCTCAA-CAAT
TTCTCAA-CAAT
TTCTCAA-CAAT
TTCTCAA-CAAT
TTCTCAA-CAAT
TTCTCAA-CAAT
```

Figure 20.15: *Three A's in the reference (top) have been replaced by two A's in the reads (shown in red). The gap is placed towards the 3' end, but could have been placed towards the 5' end with an equally good mapping score for the read.*

like BWA and Bowtie, that place insertions or deletions at the left side of a homopolymeric tract. Thus, to help facilitate the comparison of variant results with such public resources, the CLC bio **Map Reads to Reference** tool, as of version 6.5 of the *CLC Cancer Research Workbench*, will place insertions or deletions in homopolymeric tracts at the left hand side.

This is a change to earlier versions of the *CLC Cancer Research Workbench* (version 6.0.5 and earlier) where the CLC bio read mapper placed insertions and deletions in homopolymeric tracts at the right hand side of the homopolymer, as viewed in the Workbench.

This has the implication that insertion and deletion variants called in homopolymeric regions will be in different positions relative to the reference when based on mappings run in version 6.0.5 and earlier, compared to variant calls based on mappings run in version 6.5 and later. Thus, if comparisons between sample variant tracks will be done in the *CLC Cancer Research Workbench*, we recommend re-running mappings so all samples are mapped using the mapping tool in version 6.5 of the *CLC Cancer Research Workbench* or higher, or all samples to be compared have been mapped using version 6.0.5 and lower.

For users of the COSMIC database or other clinical databases following the recommendations from the Human Genome Variation Society (HGVS)

The Human Genome Variation Society (HGVS) recommendations, which pertain to variants within genes, state that for insertions and deletions in homopolymeric or repetitive regions, the most 3' position (corresponding to the strand of the gene) possible should be arbitrarily assigned as the site of change (see http://www.hgvs.org/mutnomen/recs-DNA.html#del). Resources such as COSMIC adhere to these recommendations. In this case, placement to the farthest possible left hand position, as viewed in the *CLC Cancer Research Workbench*, of insertions or deletions in repetitive or homopolymeric tracts, has a different effect, depending on whether the gene involved is on the positive or negative strand of the reference. Such variants located within genes on the negative strand can be compared with the COSMIC database, while those within genes lying on the positive strand cannot be, as the positions relative to the reference will be different in this case. The opposite situation is true when variant calls are based on mappings run in version 6.0.5 of the *CLC Cancer Research Workbench* or earlier. That is, if comparing to a resource following HGVS recommendations, like COSMIC, insertions and deletions in homopolymeric or repetitive regions called within genes that lie on the positive strand will be comparable based on position relative to the reference, while those within genes on the negative strand will not be.

20.3.5 Computational requirements

The memory requirements of **Map Reads to Reference** depends on four factors. The size of the reference, the length of the reads, the read error rate and the number of CPU cores available. The limiting factor is often the size of the reference while the contribution of the other three factors to the total memory consumption is usually small (see below).

A good estimate for the memory required by the base space read mapper to represent a reference is one MB for each Mbp in the reference. For example the human reference genome requires $3200 * 1MB = 3.2GB$ of memory. The color space mapper is able to scale down its memory consumption, such that even large references can be represented using small amounts of memory. However, when the memory consumption is scaled down it causes the read mapping to become slower.

When mapping short high quality reads, such as Illumina reads, the added memory consumption per CPU core is small. However, when mapping long reads with a high error rate, such as PacBio reads, each CPU core can add several hundred MB to the total memory consumption. Consequently, mapping long reads with high error rate on a machine with many CPU cores, can cause a large increase in the memory requirements for all CLC read mappers. An additional 4GB of memory should be reserved for the *CLC Cancer Research Workbench*, and thus the recommended minimum amount of memory for mapping short high quality reads (e.g. Illumina reads) to the human genome is 8GB.

20.3.6 Reference Caching

In some cases repeated mappings against the same reference will result in a dramatically reduced runtime because the internal data structure used for mapping the reads, which is reference specific, can be reused. This has been enabled by storing files in the system tmp folder as a caching mechanism. Only a certain amount of disk space will be used and once reaching the limit, the oldest files are cleaned up. Consequently, the reference data structure files will automatically have to be recreated if the cache was filled or the tmp folder was cleaned up.

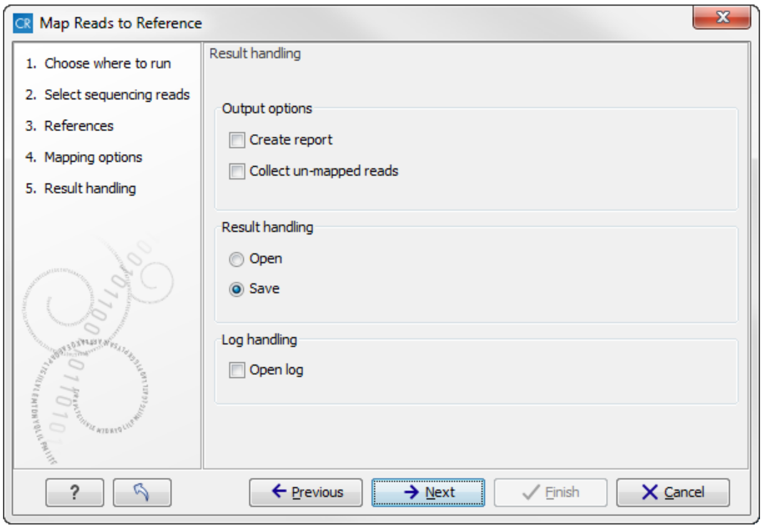
The default space limit is 8 GB which can be changed by going to

Edit | **Preferences** | **Advanced** | **Read Mapper**

On the server and for webstart the cache size can be controlled by creating a settings file "readmapper.properties" with an entry going like this "referencecachesize = 8589934592" where the size is in bytes. On grid setups, the "readmapper.properties" file will have to be added manually to each grid worker directory.

20.4 Mapping output options

Click **Next** lets you choose how the output of the mapping should be reported (see figure 20.16).



Result handling

Output options

□ Create report

□ Collect un-mapped reads

Result handling

○ Open

● Save

Log handling

□ Open log

Figure 20.16: *Mapping output options.*

There are two independent output options available that can be (de-)activated in both cases:

- **Create report**. This will generate a summary report as described in section 18.3.2.
- **Collect un-mapped reads**. This will collect all the reads that could not be mapped to the reference into a sequence list (there will be one list of unmapped reads per sample, and for paired reads, there will be one list for intact pairs and one for single reads where the mate could be mapped).

However, the main output is a reads track:

Reads track A reads track is very "lean" (i.e. with respect to memory requirements) since it only contains the reads themselves. Additional information about the reference, consensus sequence or annotations can be added and viewed alongside in the context of a Track List later (by adding, for example, a reference and/or annotation track, respectively). This kind of output is useful when working with tracks in general and especially for resequencing purposes this is recommended. Details about wiewing and editing of reads-tracks are described in chapter 17.

Finally, you can choose to save or open the results, and if you wish to see a log of the process (see section 8.2).

Clicking **Finish** will start the mapping.

20.5 Color space

20.5.1 Sequencing

The SOLiD sequencing technology from Applied Biosystems is different from other sequencing technologies since it does not sequence one base at a time. Instead, two bases are sequenced at a time in an overlapping pattern. There are 16 different dinucleotides, but in the SOLiD technology, the dinucleotides are grouped in four carefully chosen sets, each containing four dinucleotides. The colors are as follows:

Base 1	**Base 2**			
	A	C	G	T
A	🔵	🟢	🟡	🔴
C	🟢	🔵	🔴	🟡
G	🟡	🔴	🔵	🟢
T	🔴	🟡	🟢	🔵

Notice how a base and a color uniquely defines the following base. This approach can be used to deduce a whole sequence from the initial nucleotide and a series of colors. Here is a sequence and the corresponding colors.

Sequence T A C T C C A T G C A
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔴 🟢 🔴 🟢

The colors do not uniquely define the sequence. Here is another sequence with the same list of colors:

Sequence A T G A G G T A C G T
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔴 🟢 🔴 🟢

But if the first nucleotide is known, the colors do uniquely define the remaining sequence. This is exactly the strategy used in SOLiD sequencing: The first nucleotide is known from the primer used, and the remaining nucleotides are deduced from the colors.

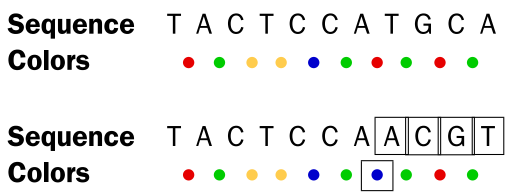
20.5.2 Error modes

As with other sequencing technologies, errors do occur with the SOLiD technology. If a single nucleotide is changed, two colors are affected since a single nucleotide is contained in two overlapping dinucleotides:

Sequence T A C T C C A T G C A
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔴 🟢 🔴 🟢

Sequence T A C T C C A $\boxed{\text{A}}$ G C A
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔵 🟡 🔴 🟢

Sometimes, a wrong color is determined at a given position. Due to the dependence between dinucleotides and colors, this affects the remaining sequence from the point of the error:

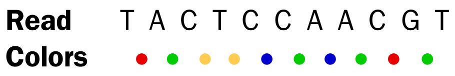

Sequence T A C T C C A T G C A
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔴 🟢 🔴 🟢
Sequence T A C T C C A A C G T
Colors 🔴 🟢 🟡 🟡 🔵 🟢 🔵 🟢 🔴 🟢

Thus, when the instrument makes an error while determining a color, the error mode is very different from when a single nucleotide is changed. This ability to differentiate different types of errors and differences is a very powerful aspect of SOLiD sequencing. With other technologies sequencing errors always appear as nucleotide differences.

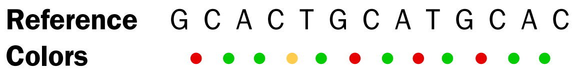
20.5.3 Mapping in color space

Reads from a SOLiD sequencing run may exhibit all the same differences to a reference sequence as reads from other technologies: mismatches, insertions and deletions. On top if this, SOLiD reads may exhibit color errors, where a color is read wrongly and the rest of the read is affected. If such an error is detected, it can be corrected and the rest of the read can be converted to what it would have been without the error.

Consider this SOLiD read:

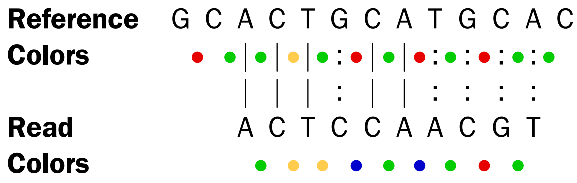


The first nucleotide (T) is from the primer, so this is ignored in the following analysis. Now, assume that a reference sequence is this:


Reference G C A C T G C A T G C A C
Colors 🔴 🟢 🟢 🟡 🟢 🔴 🟢 🔴 🟢 🔴 🟢 🟢

Here, the colors are just inferred since they are not the result of a sequencing experiment.

Looking at the colors, a possible alignment presents itself:



In the beginning of the read, the nucleotides match (ACT), then there is a mismatch (G in reference and C in read), then two more matches (CA), and finally the rest of the read does not match. But, the colors match at the end of the read. So a possible interpretation of the alignment is that there is a nucleotide change in position four of the read and a color space error between positions six and seven in the read. Such an interpretation can be represented as:

```
Reference   G C A C T G C A T G C A C
                | | | : | | | | | |
Read            A C T C C A*T G C A
```

Here, the * represents a color error. The remaining part of the displayed read sequence has been adjusted according to the inferred error. So this alignment scores nine times the match score minus the mismatch cost and a color error cost. This color error cost is a new parameter that is introduced when performing read mapping in color space.

Note that a color error may be inferred before the first nucleotide of a read. This is the very first color after the known primer nucleotide that is wrong, changing the whole read.

Here is an example from a set of real SOLiD data that was reference assembled by taking color space into account using ungapped global alignments.

```
444_1840_767_F3 has 1 match with a score of 35:

    1046535 GATACTCAATGCCGCCAAAGATGGAAGCCGGGCCA 1046569    reference
            |||||||||||||||||||||||||||||||||||
            GATACTCAATGCCGCCAAAGATGGAAGCCGGGCCA            reverse read

444_1840_803_F3 has 0 matches

444_1840_980_F3 has 1 match with a score of 29:

    2620828 GCACGAAAACGCCGCGTGGCTGGATGGT*CAAC*GTC 2620862    reference
            ||||||||||||||||||||||||||||*||||*|||
            GCACGAAAACGCCGCGTGGCTGGATGGT*CAAC*GTC            read

444_1840_1046_F3 has 1 match with a score of 32:

    3673206 TT*GGTCAGGGTCTGGGCTTAGGCGGTGAATGGGGC 3673240    reference
            ||*|||||||||||||||||||||||||||||||||
            TT*GGTCAGGGTCTGGGCTTAGGCGGTGAATGGGGC            reverse read

444_1841_22_F3 has 0 matches

444_1841_213_F3 has 1 match with a score of 29:

    1593797 CTTTG*AGCGCATTGGTCAGCGTGTAATCTCCTGCA 1593831    reference
            |||||*|||||||| ||||||||||||||||||||
            CTTTG*AGCGCATTAGTCAGCGTGTAATCTCCTGCA            reverse read
```

The first alignment is a perfect match and scores 35 since the reads are all of length 35. The next alignment has two inferred color errors that each count is -3 (marked by * between residues), so the score is *35 - 2 x 3 = 29*. Notice that the read is reported as the inferred sequence taking the color errors into account. The last alignment has one color error and one mismatch giving a score of *34 - 3 - 2 = 29*, since the mismatch cost is 2.

Running the same reference assembly without allowing for color errors, the result is:

```
444_1840_767_F3 has 1 match with a score of 35:

    1046535 GATACTCAATGCCGCCAAAGATGGAAGCCGGGCCA 1046569    reference
            |||||||||||||||||||||||||||||||||||
            GATACTCAATGCCGCCAAAGATGGAAGCCGGGCCA            reverse read
```

```
444_1840_803_F3 has 0 matches

444_1840_980_F3 has 0 matches

444_1840_1046_F3 has 1 match with a score of 29:

    3673206 TTGGTCAGGGTCTGGGCTTAGGCGGTGAATGGGGC 3673240    reference
              |||||||||||||||||||||||||||||||||
            AAGGTCAGGGTCTGGGCTTAGGCGGTGAATGGGGC            reverse read

444_1841_22_F3 has 0 matches

444_1841_213_F3 has 0 matches
```

The first alignment is still a perfect match, whereas two of the other alignment now do not match since they have more than two errors. The last alignment now only scores 29 instead of 32, because two mismatches replaced the one color error above. This shows the power of including the possibility of color errors when aligning: many more matches are found.

The reference assembly program in *CLC Cancer Research Workbench* does not directly support alignment in color space only, but if such an alignment was carried out, sequence 444_1841_213_F3 would have three errors, since a nucleotide mismatch leads to two color space differences. The alignment would look like this:

```
444_1841_213_F3 has 1 match with a score of 26:

    1593797 CTTTG*AGCGCATT*G*GTCAGCGTGTAATCTCCTGCA 1593831    reference
            |||||*||||||||*|*|||||||||||||||||||||
            CTTTG*AGCGCATT*G*GTCAGCGTGTAATCTCCTGCA            reverse read
```

So, the optimal solution is to both allow nucleotide mismatches and color errors in the same program when dealing with color space data. This is the approach taken by the assembly program in *CLC Cancer Research Workbench*.

Note! If you set the color error cost as low as 1 while keeping the mismatch cost at 2 or above, a mismatch will instead be represented as two adjacent color errors.

20.5.4 Viewing color space information

Importing data from SOLiD systems (see section 6.3.3) will from *CLC Cancer Research Workbench* version 3.1 be imported as color space. This means that if you open the imported data, it will look like figure 20.17

In the **Side Panel** under **Nucleotide info**, you find the **Color space encoding** group which lets you define a few settings for how the colors should appear. These settings are also found in the side panel of mapping results and single sequences.

Infer encoding This is used if you want to display the colors for non-color space sequence (e.g. a reference sequence). The colors are then simply inferred from the sequence.

Show corrections This is only relevant for mapping results - it will show where the mapping process has detected color errors. An example of a color error is shown in figure 20.18.

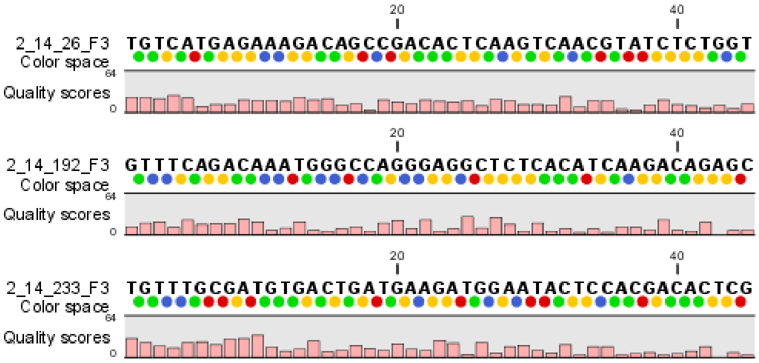
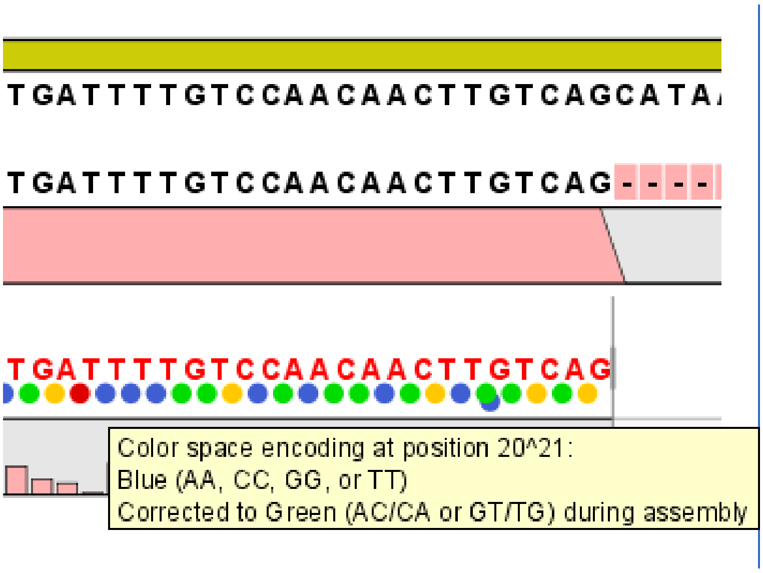

20
40
2_14_26_F3
Color space
Quality scores
64
0
TGTCATGAGAAAGACAGCCGACACTCAAGTCAACGTATCTCTGGT
2_14_192_F3
GTTTCAGACAAATGGGCCAGGGAGGCTCTCACATCAAGACAGAGC
2_14_233_F3
TGTTTGCGATGTGACTGATGAAGATGGAATACTCCACGACACTCG

Figure 20.17: *Color space sequence list.*

Hide unaligned ends This option determines whether color for the unaligned ends of reads should be displayed. It also controls whether colors should be shown for gaps. The idea behind this is that these color dots will interfere with the color alignment, so it is possible to turn them off.



Color space encoding at position 20^21:
Blue (AA, CC, GG, or TT)
Corrected to Green (AC/CA or GT/TG) during assembly

Figure 20.18: *One of the dots have both a blue and a green color. This is because this color has been corrected during mapping. Putting the mouse on the dot displays the small explanatory message.*

20.6 Mapping result

Reads can be mapped to linear and circular chromosomes. Read mappings to circular genomes are visualized linearly as shown in figure 20.19.

Reads that map across the starting point of the sequence are shown both at the start and end of the reference sequence. Such reads are marked with >> at the end of the read to indicate that the alignment continues at the other end of the reference sequence.

The result of the read mapping is a read mapping track ().

Note: If your reads track shows the message 'Too much data for rendering' on a grey background,

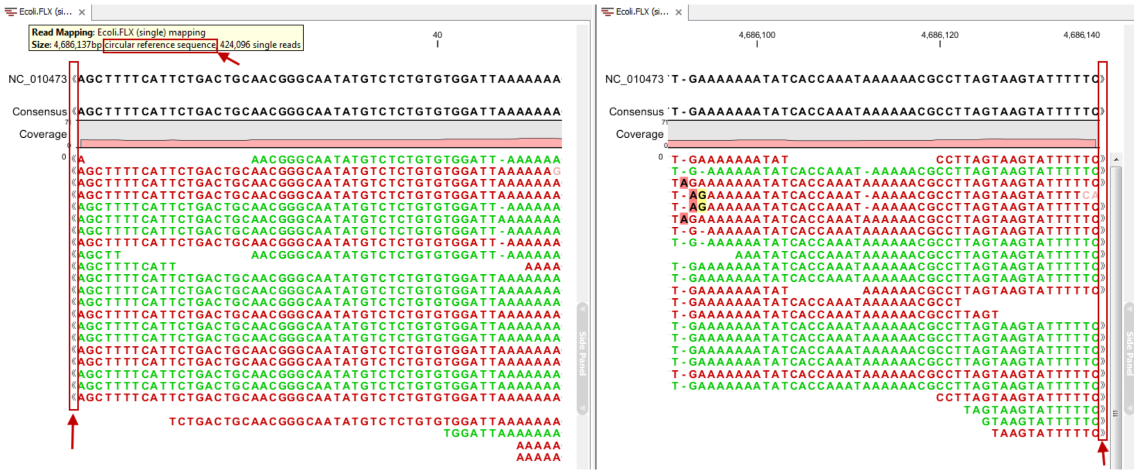

Ecoli.FLX (si...
Read Mapping: Ecoli.FLX (single) mapping
Size: 4,686,137bp circular reference sequence 424,096 single reads
40
NC_010473 AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAA
Consensus AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAA
Coverage
Ecoli.FLX (si...
4,686,100
4,686,120
4,686,140
NC_010473 T-GAAAAAAATATCACCAAATAAAAAACGCCTTAGTAAGTATTTTTC
Consensus T-GAAAAAAATATCACCAAATAAAAAACGCCTTAGTAAGTATTTTTC
Coverage
Side Panel

Figure 20.19: *Mapping reads to a circular chromosome. Reads that are marked with double arrows at the ends are reads that map across the starting point of the sequence. The arrows indicate that the alignment continues at the other end of the reference sequence.*

simply zoom in to see your reads in more detail. This occurs when there are too many reads (more than 500,000) to be displayed clearly.

20.6.1 View settings in the Side Panel

When you open a single mapping, the following settings are available in the **Side Panel** for customizing the layout.

- **Read layout.** This section appears at the top of the **Side Panel** when viewing a stand alone read mapping:
  - **Compactness.** The compactness setting options let you control the level of detail to be displayed. This setting affects many of the other settings in the **Side Panel** as well as the general behavior of the view. For example: if the compactness is set to **Compact**, you will not be able to see quality scores or annotations on the reads, even if these are turned on via the Nucleotide info section of the Side Panel. You can change the Compactness setting in the Side Panel directly, or you can use the shortcut: press and hold the Alt key while you scroll with the mouse wheel or touchpad.
    - **Not compact.** This allows the mapping to be viewed full detail, including quality scores and trace data for the reads, where this is relevant. To view such information, additional viewing options under the **Nucleotide info** view settings must also selected. For further details on these, please see section 29.1.2 and section 9.1.
    - **Low.** Hides trace data, quality scores and puts the reads' annotations on the sequence.
    - **Medium.** The labels of the reads and their annotations are hidden, and the residues of the reads cannot be seen.
    - **Compact.** Even less space between the reads.
    - **Packed.** All the other compactness settings will stack the reads on top of each other, but the packed setting will use all space available for displaying the reads. When zoomed in to 100%, you can see the residues but when zoomed

out the reads will be represented as lines just as with the Compact setting. The packed mode is very useful when viewing large amounts of data. However certain functionality possible with other views are not available in packed view. For example, no editing of the read mapping or selections of it can be done and color coding changes are not possible. An example of the packed setting is shown in figure 29.14.

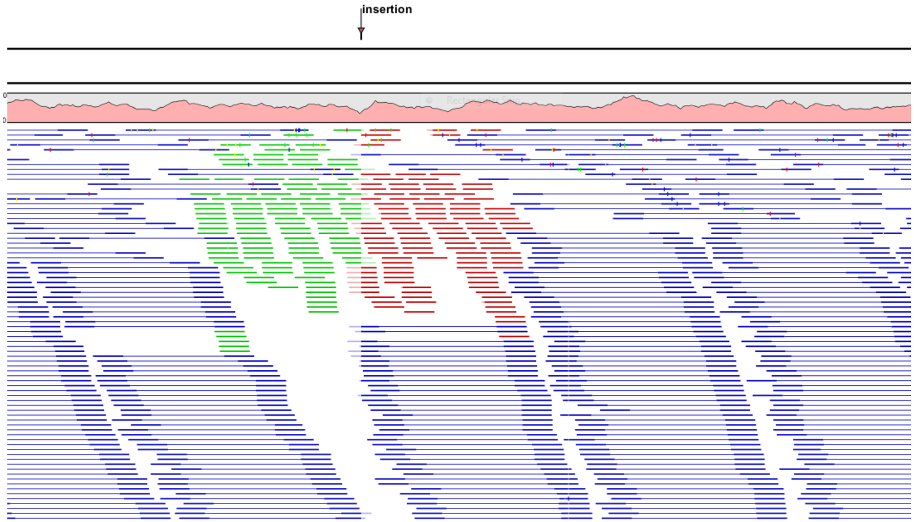


Figure 20.20: *An example of the packed compactness setting.*

- **Gather sequences at top.** Enabling this option affects the view that is shown when scrolling horizontally. If selected, the sequence reads which did not contribute to the visible part of the mapping will be omitted whereas the contributing sequence reads will automatically be placed right below the reference. This setting is not relevant when the compactness is packed.
- **Show sequence ends.** Regions that have been trimmed are shown with faded traces and residues. This illustrates that these regions have been ignored during the assembly.
- **Show mismatches.** When the compactness is packed, you can highlight mismatches which will get a color according to the Rasmol color scheme. A mismatch is whenever the base is different from the reference sequence at this position. This setting also causes the reads that have mismatches to be floated at the top of the view.
- **Disconnect paired reads.** This option will break up the paired reads in the display (they are still marked as pairs - this just affects the visualization). The reads are marked with colors for the direction (default red and green) instead of the color for pairs (default blue). This is particularly useful when investigating overlapping pairs in packed view and when the strand / read orientation is important.
- **Packed read height.** When the compactness is set to "packed", you can choose the height of the visible reads. When there are more reads than the height specified, an overflow graph will be displayed below the reads. The overflow graph is shown in the same colors as the sequences, and mismatches in reads are shown as narrow horizontal lines in. The colors of the small lines represent the mismatching residue. The color codes for the horizontal lines correspond to the color used for highlighting mismatches in the sequences (red = A, blue = C, yellow = G, and green = T). E.g.

    a red line with half the height of the blue part of the overflow graph will represent a mismatching "A" in half of the paired reads at this particular position.

  - **Find Conflict.** Clicking this button selects the next position where there is an conflict between the sequence reads. Residues that are different from the reference are colored (as default), providing an overview of the conflicts. Since the next conflict is automatically selected it is easy to make changes. You can also use the Space key to find the next conflict.
  - **Low coverage threshold.** All regions with coverage up to and including this value are considered low coverage. When clicking the 'Find low coverage' button the next region in the read mapping with low coverage will be selected.

- **Alignment info.** There is one additional parameter:
  - **Coverage**: Shows how many sequence reads that are contributing information to a given position in the mapping. The level of coverage is relative to the overall number of sequence reads.
    - **Foreground color.** Colors the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage.
    - **Background color.** Colors the background of the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage
    - **Graph.** The coverage is displayed as a graph (Learn how to export the data behind the graph in section 6.7).
      - **Height.** Specifies the height of the graph.
      - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
      - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. If a Color bar is chosen, the color box is replaced by a gradient color box as described under Foreground color.

- **Residue coloring.** There is one additional parameter:
  - **Sequence colors.** This option lets you use different colors for the reads.
    - **Main**. The color of the consensus and reference sequence. Black per default.
    - **Forward**. The color of forward reads (single reads). Green per default.
    - **Reverse**. The color of reverse reads (single reads). Red per default.
    - **Paired**. The color of paired reads. Blue per default. Note that reads from **broken pairs** are colored according to their Forward/Reverse orientation or as a Non-specific match, but with a darker nuance than ordinary single reads.
    - **Non-specific matches**. When a read would have matched equally well another place in the mapping, it is considered a non-specific match. This color will "overrule" the other colors. Note that if you are mapping with several reference sequences, a read is considered a double match when it matches more than once *across all the contigs/references*. A non-specific match is yellow per default.

- **Sequence layout.** At the top of the **Side Panel**:
  - **Matching residues as dots** Matching residues will be presented as dots. Only the top sequence will be preserved in its original format.

There are many other viewing options available, both general and aimed as specifice elements of a mapping, which can be adjusted in the View settings. Those covered here were the key ones relevant standard review of mapping results.

20.7 Local realignment

The goal of the local realignment tool is to improve on the alignments of the reads in an existing read mapping. The local realignment algorithm works by exploiting the information available in the alignments of *other* reads when it is attempting to re-align any given read. Most mappers do not use cross-read information as it would be computationally prohibitive to do within the mapping algorithm. However, once the reads have been mapped, local realignment procedures can exploit this information.

Realignment will typically occur in areas around insertions and deletions in the sample reads relative to the reference. In such regions we wish to see our reads mapped with one end of the read on one side of the indel and the rest mapped on the other side. However, the mapper that originally mapped the reads to the reference does not have information about the existence of an indel to use when mapping a given read. Thus, reads that are mapped to such regions, but that only have a short part of the read representing the region on one side of the indel, will typically not be mapped properly across the indel, but instead be mapped with this end unaligned, or into the indel region with many mismatches. The Local Realignment tool can use information from the other reads mapping to a region containing an indel, including reads that are located more centered across the indel and thus have been mapped with ends on either side of the indel. As a result an alternative mapping, as good as or better than the original, can be generated.

Local realignment will typically have an effect on any read mapping, whether the reads were mapped using a local or global alignment algorithm (i.e. with the Global alignment option of the mapping tool unchecked (the default) or checked, respectively). An example of the effect of using the Local Realignment tool on a read mapping made using the the local alignment algorithm is shown in figure 20.21. An example in the case of a mapping made using the global alignment algorithm is shown in figure 20.22.

20.7.1 Method

The local realignment algorithm uses a variant of the approach described by Homer et al. [Homer N, 2010]. In the first step, alignment information of all input reads are collected in an efficient graph-based data structure, which is essentially similar to a de-Brujn graph. This realignment graph represents how reads are aligned to the reference sequence and how reads overlap each other. In the second step, metadata are derived from the graph structure that indicate at which alignment positions realignment could potentially improve the read mapping, and also provides hypotheses as to how reads should be realigned to yield the most concise multiple alignment. In the third step the realignment graph and its metadata are used to actually perform the local realignment of each individual read. Figure 20.23 depicts a partial realignment graph for the read mapping shown in figure 20.21.

20.7.2 Realignment of unaligned ends

A typical error in read alignments is the occurrence of unaligned ends (also known as soft-clipped read ends). These unaligned ends are introduced by the read mapper as a consequence of an

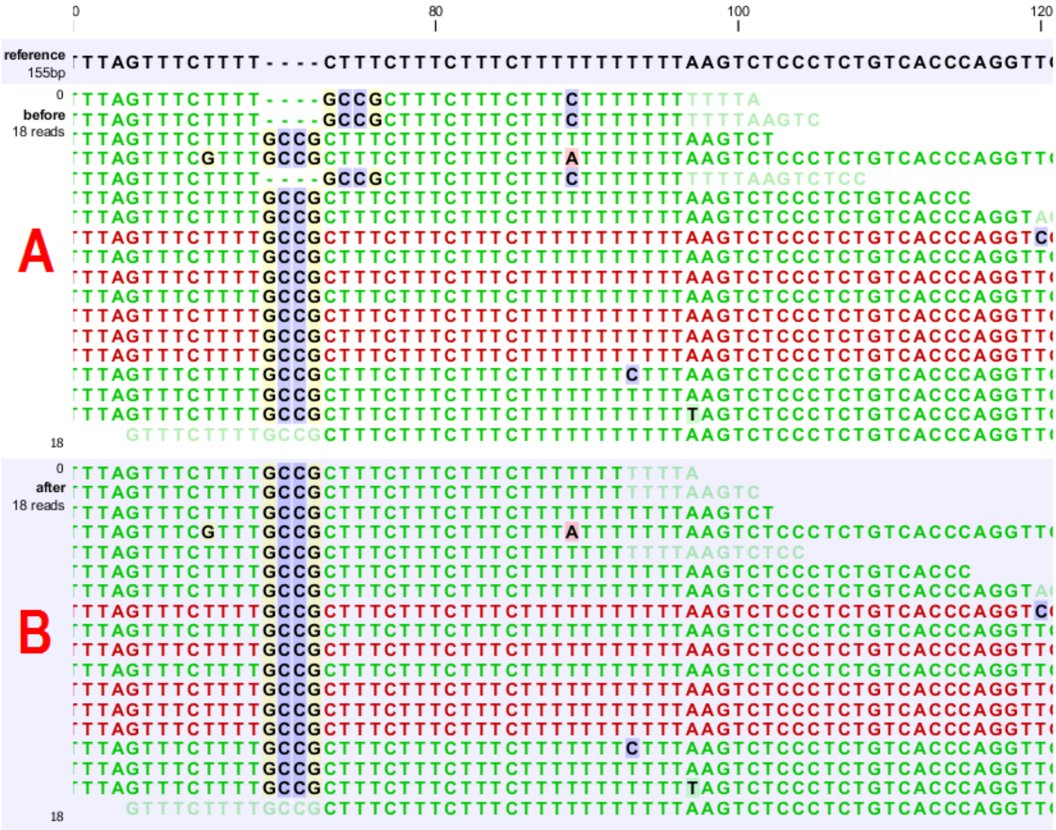


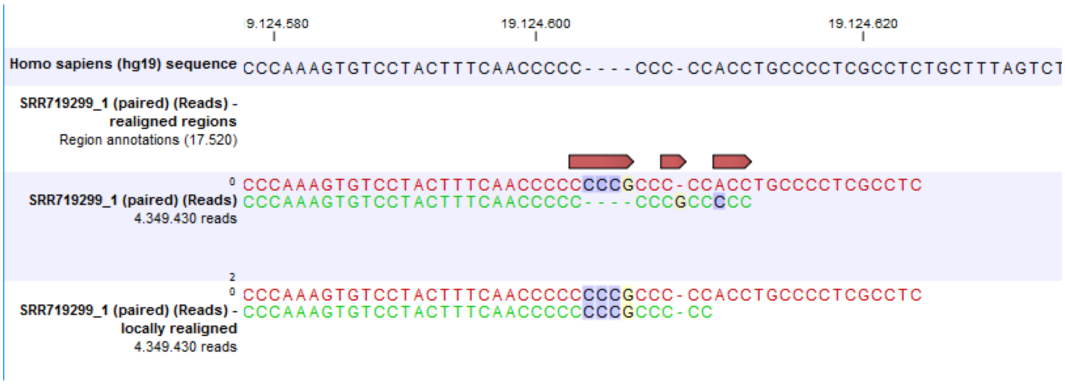
Figure 20.21: *Local realignment of a read mapping produced with the 'local' option. [A] The alignments of the first, second, and fifth read in this read mapping do not support the four-nucleotide insertion supported by the remaining reads. A variant caller might be tempted to call a heterozygous insertion of four nucleotides in one allele and heterozygous replacement of four nucleotides in a second allele. [B] After applying local realignment, the first, second, and fifth read consistently support the four-nucleotide insertion.*

unresolved indel towards the end of a read. Those unaligned ends can be realigned in many cases, after the read itself has been locally realigned according to the indel that prevented the read mapper from aligning the read ends correctly. Figure 20.24 depicts such an example.

20.7.3 Guided Realignment

One limitation of the local realignment algorithm employed is that at least one read must be aligned correctly according to the true indel present in the data. If none of the reads is aligned correctly, local realignment cannot improve the alignment, since it lacks information about how to do so. To overcome this limitation, local realignment can be guided in two ways:

1. **Guidance variants:** By supplying the Local realignment tool with a track of guidance variants. There are two modes for using the guidance variant track: either the 'un-forced' guidance mode (if the 'Force realignment to guidance-variants' is left un-ticked) or the 'forced' guidance mode (if the 'Force realignment to guidance-variants' is ticked). In the 'unforced' mode, 'pseudo-reads' are given to the local realignment algorithm representing the guidance variants, allowing the local realignment algorithm to explore the paths in the graph corresponding to these alignments. In the 'forced' mode, 'pseudo-references'



```
                                9.124.580           19.124.600                    19.124.620
Homo sapiens (hg19) sequence    CCCAAAGTGTCCTACTTTCAACCCCC----CCC-CCACCTGCCCCTCGCCTCTGCTTTAGTCT
SRR719299_1 (paired) (Reads) -
            realigned regions
      Region annotations (17.520)
                              0 CCCAAAGTGTCCTACTTTCAACCCCCCCCGCCC-CCACCTGCCCCTCGCCTC
SRR719299_1 (paired) (Reads)    CCCAAAGTGTCCTACTTTCAACCCCC----CCCGCCCCC
                4.349.430 reads
                              2
                              0 CCCAAAGTGTCCTACTTTCAACCCCCCCCGCCC-CCACCTGCCCCTCGCCTC
SRR719299_1 (paired) (Reads) -  CCCAAAGTGTCCTACTTTCAACCCCCCCCGCCC-CC
              locally realigned
                4.349.430 reads
```

Figure 20.22: *Local realignment of a read mapoping produced with the 'global' option. Before realignment the green read was mapped with two mismatches. After realignment it is mapped with the inserted 'CCCG' sequence (seen in the alignment of the red read) and no mismatches.*

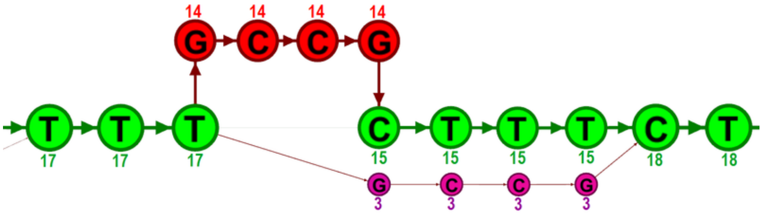


Figure 20.23: *The green nodes represent nucleotides of the reference sequence. The four red nodes represent the four-nucleotide insertion observed in fourteen mapped reads. The four violet nodes represent the four mismatches to the reference sequence observed in three mapped reads. During realignment of the original reads, two possible paths through the graph are discovered. One path leads through the four red nodes, the other through the four violet nodes. Since red nodes have been observed in fourteen of the original reads, whereas the violet nodes have only been seen in three original reads, the path through the four red nodes is preferred over the path through the violet nodes.*

are given to the local realignment algorithm representing the guidance variants, allowing the reads to be aligned to allele sequences of these in addition to the original reference sequence - with matches being awarded and encouraged equally much. The 'unforced' mode can be used with any guidance variant track as input. The 'force' mode should *only* be used with guidance variants for which there is prior evidence that they exist in the data (e.g., the 'InDel' track from the Structural Variants' tool (see Section 20.17) produced on the read mapping that is being aligned).

2. **Concurrent local realignment of multiple samples:** Multiple input read mappings increase the chance to encounter at least one read mapped correctly. This guiding mechanism has been particularly designed for scenarios, where samples are known to be related, such as in family trials.

Figure 20.25 and figure 20.26 show examples that can be improved by guiding the local realignment algorithm.

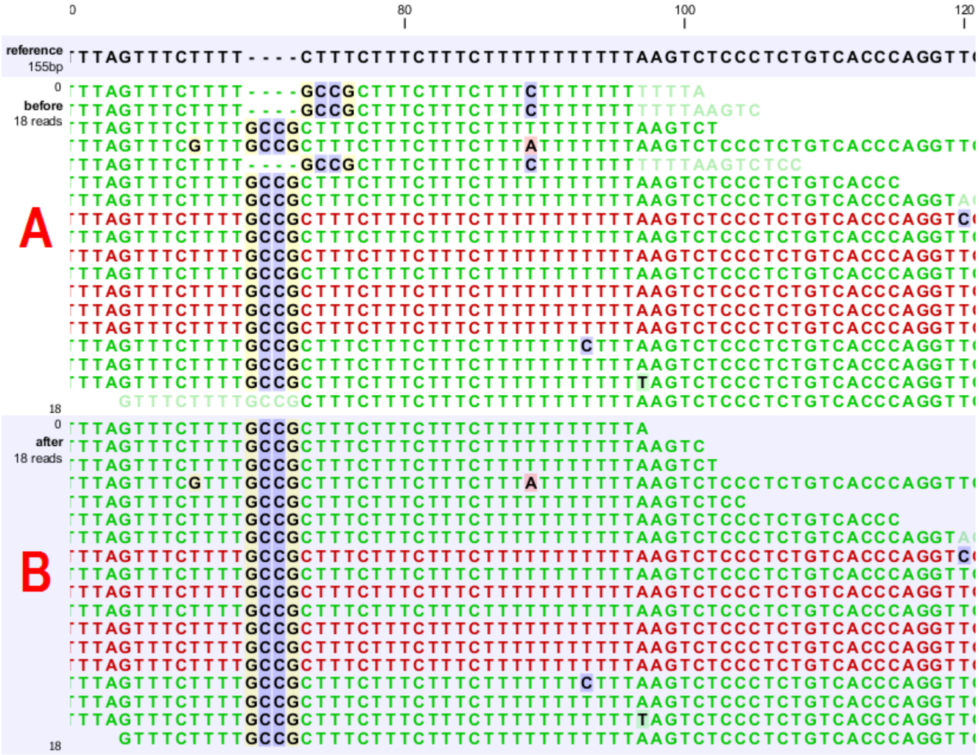

0
80
100
120
reference
155bp
TTAGTTTCTTTT----CTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
0
before
18 reads
A
18
0
after
18 reads
B
18

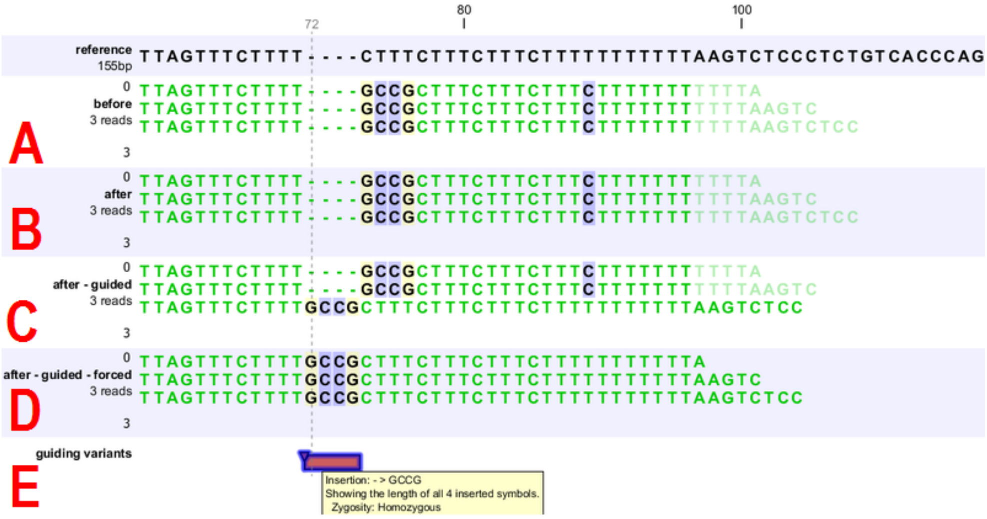
Figure 20.24: *[A] The alignments of the first, second, and fifth read in this read mapping do not support the four-nucleotide insertion supported by the remaining reads. Additionally, the first, second, fifth and the last reads have unaligned ends. [B] After applying local realignment the first, second and fifth read consistently support the four-nucleotide insertion. Additionally, all previously unaligned ends have been realigned, because they perfectly match the reference sequence now (see also figure 20.21).*

20.7.4 Multi-pass local realignment

As described in section 20.7.1 the algorithm initially builds the realignment graph using the input read mapping. After the graph has been built the algorithm realigns individual reads based on information inferred from the realignment graph structure and its associated metadata. In some cases repetitive realignment iterations yield even more improvements, because with each realignment iteration the structure of the realignment graph changes slightly, potentially permitting further improvements. Local realignment therefore supports to perform multiple iterations implicitly. This is not only considered a convenience feature, but also saves a great deal of runtime by avoiding repeated transfers of large input data sets. For most samples local realignment will quickly saturate in the number of improvements. Generally, two realignment passes are strongly recommended. More than three passes rarely yield further improvements.

20.7.5 Known Limitations

The major limitation of the local realignment algorithm is the necessity of at least one read being mapped correctly according to an indel present in the data. Insufficient alignment data results in suboptimal realignments or no realignments at all. As a work-around, local realignment can be guided by supplying a track of variants that enable the algorithm to determine improvements. Further guidance can be achieved by increasing the amount of alignment information and thereby



```
      reference
          155bp  TTAGTTTCTTTT----CTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAG
              0  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTA
         before  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTAAGTC
        3 reads  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTAAGTCTCC
              3
              0  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTA
          after  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTAAGTC
        3 reads  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTAAGTCTCC
              3
              0  TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTA
  after - guided TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTTAAGTC
        3 reads  TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCC
              3
              0  TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTA
after - guided - forced TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTC
        3 reads  TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCC
              3
guiding variants
```

Insertion: - > GCCG
Showing the length of all 4 inserted symbols.
Zygosity: Homozygous

Figure 20.25: *[A] Three reads are misaligned in the presence of a four nucleotide insertion relative to the reference. [B] When applying local realignment without guidance the alignment is not improved. [C] Here local realignment is performed in the presence of the guiding variant track seen in (E). This enables the algorithm to consider alternative alignments, which are accepted whenever they have significant improvements over the original (as in read three that has a comparatively long unaligned-end). [D] If the alignment is performed with the option "Force realignment to guidance-variants" enabled, the realignment will be forced to realign according to the guiding variants track shown in (E), and this will result in realignment of all three reads. [E] The guiding variants track contains, amongst others, the four nucleotide insertion.*

increasing the chance to observe at least one read mapped correctly.

Reads are ignored, but retained in outputs, if:

- Lengths are longer than 50,000 base pairs.
- The alignment is longer than 50,000 base pairs.
- Crossing the boundaries of circular chromosomes.

Guiding variants are ignored, if:

- They are of type "Replacement".
- They are longer than 100 bp.
- If they are inter-chromosomal structural variations.
- If they contain ambiguous nucleotides.

20.7.6 Computational Requirements

The realignment graph is produced using a sliding-window approach with a window size of 250,000 bp. If local realignment is run with multiple passes, then each pass has its own realignment graph. While memory consumption is typically below two gigabytes for single-pass, processor

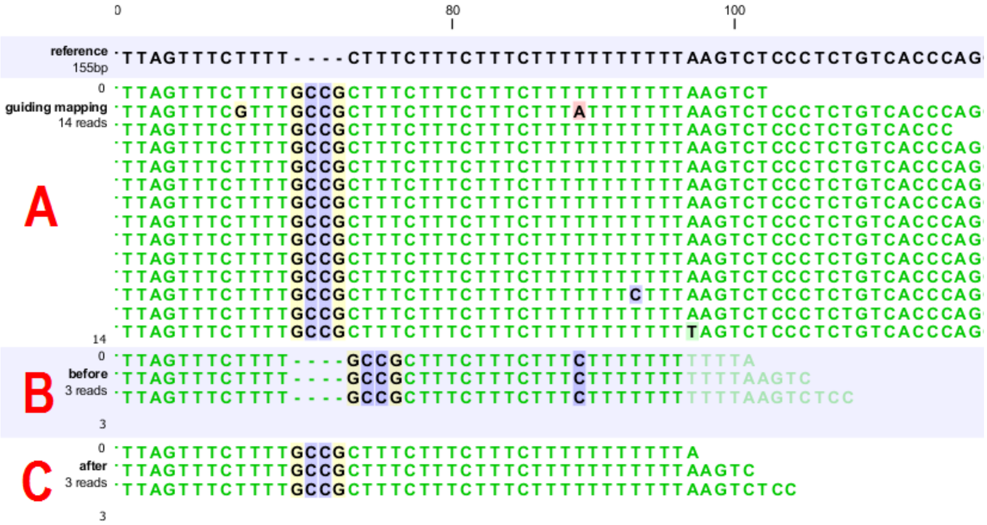


Figure 20.26: *[B] Three reads are misaligned in the presence of a four nucleotide insertion into the reference. Applying local realignment without guiding information would not yield any improvements (not shown). [C] Performing local realignment on both samples (A) and (B) enables the algorithm to improve the alignments of sample (B).*

loads are substantial. Realigning a human sample of approximately 50x coverage will take around 24 hours on a typical desktop machine with four physical cores. Building the realignment graph and realignment of reads are parallelized actions, such that the algorithm scales very well with the number of physical cores. Server machines exploiting 12 or more physical cores typically run three times faster than the desktop with only four cores.

20.7.7 How to run the Local Realignment tool

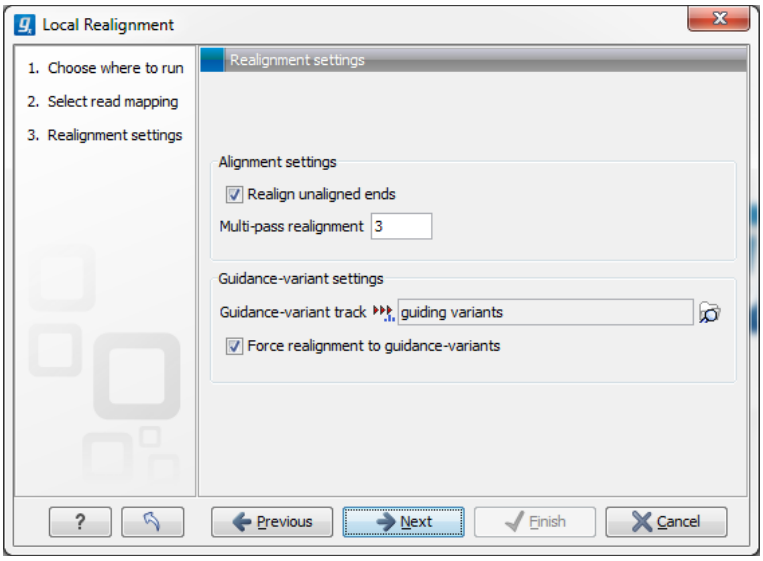
The tool is found in the Toolbox:

Toolbox | NGS Core Tools ( ) | Local Realignment ( )

Select one or multiple read mappings as input. If one read mapping is selected, local realignment will attempt to realign all contained reads, if appropriate. If multiple read mappings are selected, their reference genome must exactly match. Local realignment will realign all reads from all input read mappings as if they came from the same input. However, local realignment will create one output read mapping for each input read mapping, thereby preserving the affiliation of each read to its sample. Clicking Next allows you to set parameters as displayed in figure 20.27.

Alignment settings

- **Realign unaligned ends** This option, if enabled, will trigger the realignment algorithm to attempt to realign unaligned ends as described in section "Realignment of unaligned ends (soft clipped reads)". This option should be enabled by default unless unaligned ends arise from known artifacts (such as adapter remainders in amplicon sequencing setups) and are thus not expected to be realignable anyway. Ignoring unaligned ends will yield a significant run time improvement in those cases. Realigning unaligned ends under normal conditions (where unaligned ends are expected to be realignable), however, does not contribute a lot of processing time.



Alignment settings

☑ Realign unaligned ends

Multi-pass realignment 3

Guidance-variant settings

Guidance-variant track guiding variants

☑ Force realignment to guidance-variants

Figure 20.27: *Set the realignment options.*

- **Multi-pass realignment** This option is used to specify, how many realignment passes shall be performed by the algorithm. More passes improve accuracy at the cost of longer run time (approx. 25% per pass). Two passes are recommended; more than three passes barely yield further improvements.

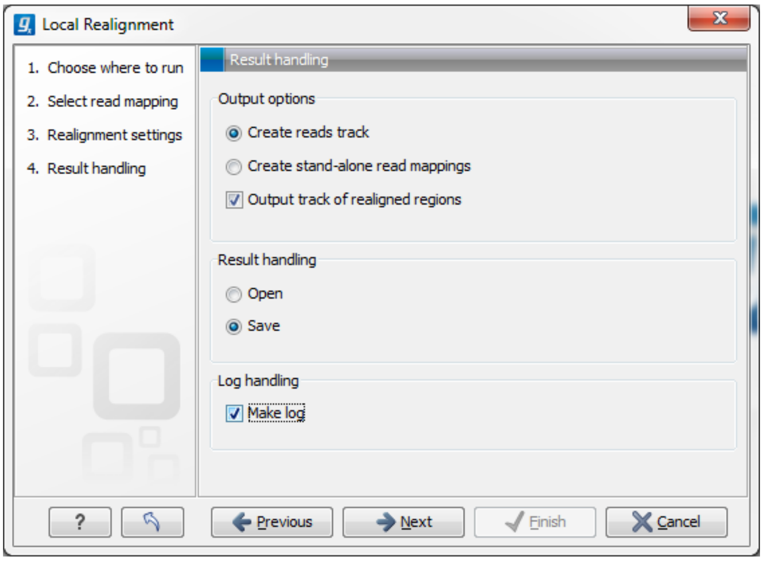
Guidance-variant settings

- **Guidance-variant track** A track of variants to guide realignment of reads. Guiding can be used in at least two scenarios: (1) if reads are short or expected variants are long and (2) if cross sample comparisons are performed and some samples are already well genotyped. A track of variants can be produced by either of the variant callers, The Structural Variant tool or by importing variants from external data sources, such as COSMIC, dbSNP, etc.

  There are two modes for using the guidance track:

  - **Un-forced** If the 'Force realignment to guidance-variants' is un-ticked the guidance variants are used as 'weak' prior evidence: each guidance variant will be represented by a pseudo-read, allowing the local realignment to explore the alignments that the guidance variants suggest. Any variant track may be used to guide the realignment when the un-forced mode is chosen.
  - **Force realignment to guidance-variants** If the 'Force realignment to guidance-variants' is ticked the guidance variants are used as 'strong' prior evidence: a 'pseudo' reference will be generated for each guidance variant, and the alignment of nucleotides to their sequences will be awarded and encouraged *as much as* the alignment to the original reference sequence. Thus, the 'Force realignment to guidance-variants' options should *only* be used when there is prior information that the variants in the guidance variant track are infact present in the sample. This would e.g. be the case for an 'InDel' track produced by the Structural Variant tool (see Section 20.17), in an analysis of the same sample as the realignment is carried out on. Using 'forced' realignment to a general variant data base track is generally *strongly* discouraged.

The next dialog allows specification of the result handling. Under "Output options" it is possible to specify whether the results should be presented as a reads track or a stand-alone read mapping (figure 20.28).



Output options

● Create reads track

○ Create stand-alone read mappings

☑ Output track of realigned regions

Result handling

○ Open

● Save

Log handling

☑ Make log

Figure 20.28: *An output track of realigned regions can be created.*

If enabled, the option **Output track of realigned regions** will cause the algorithm to output a track of regions that help pinpoint regions that have been improved by local realignment. This track has purely informative intention and cannot be used for anything else.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

20.8 Remove duplicate mapped reads

The purpose of this tool is to efficiently remove duplicate reads from a mapping, when duplicate reads have arisen due to the use of PCR amplification (or other enrichment) during sample preparation. The tool may be used on mappings of single end reads, paired end reads or both.

A read duplication can be easily distinguished when mapping reads to a reference sequence as shown in the example in figure 20.29.

When sequencing library preparation involves a PCR amplification step, it is common to observe multiple reads where identical nucleotide sequences are disproportionably represented in the final results. Thus, to facilitate processing of mappings based on this kind of data, it may be necessary to perform a duplicate read removal step, which flags identical reads and subsequently removes them from the data set. However, this step is complicated by the low, but consistent, presence of sequencing errors that may cause otherwise identical sequences to differ slightly. Thus, it is important that the duplicate read removal includes some tolerance for nearly identical sequences, which could still be reads from the same PCR artifact.

In samples that have been mapped to a reference genome, duplicate reads from PCR amplification typically result in areas of disproportionally high coverage and are often the cause of significant skew in allelic ratios, particularly when replication errors are made by the enzymes (e.g. polymerases) used during amplification. Sequencing errors incorporated post-amplification can affect both sequence- and coverage-based analysis methods, such as variant calling, where introduced errors can create false positive SNPs, and ChIP-Seq, where artificially inflated

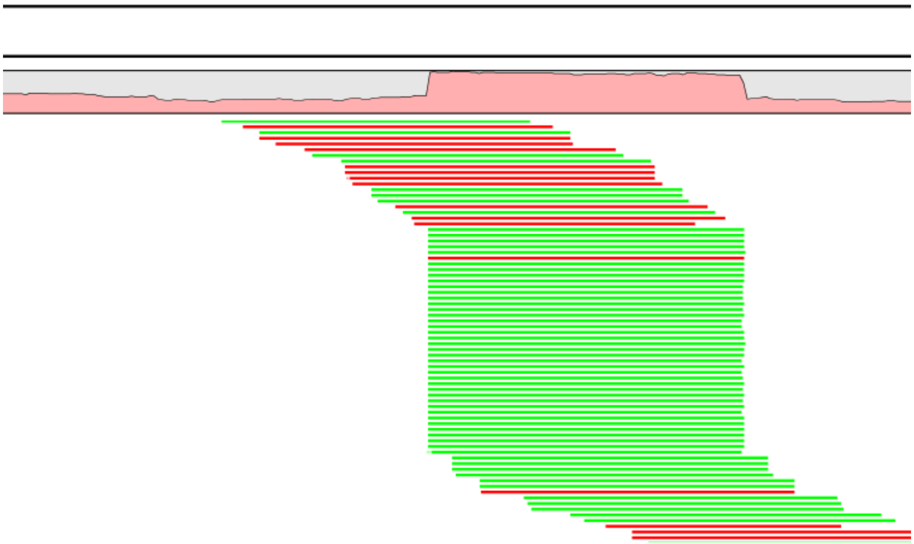


Figure 20.29: *Mapped reads with a set of duplicate reads, the colors denote the strand (green is forward and red is reverse).*

coverage can skew the significance of certain locations. By utilizing the mapping information, it is possible to perform the duplicate removal process rapidly and efficiently.

Note! We only recommend using the duplicate read removal if there are amplification steps involved in the library preparation.

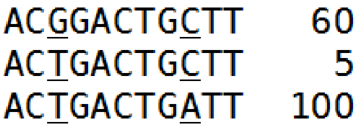
The method used by the duplicate read removal is to identify reads that share common coordinates (e.g. the same start and end coordinate), sequencing direction (or mapped strand) and the same sequence, these being the unifying characteristics behind sequencing reads that originate from the same amplified fragments of nuclear material. However, due to the frequent occurrence of sequencing errors, the tool utilizes simple heuristics to prune sequences with small variations from the consensus, as would be expected from errors observed in data from next-generation sequencing platforms. Base mismatch errors that were incorporated during amplification or prior to amplification will be indistinguishable from SNPs and may not be filtered out by this tool.

20.8.1 Algorithm details and parameters

The algorithm operates with the following assumption: Mapped reads from duplicated DNA fragments will share a mapping orientation (e.g. will map to the same strand), and depending on their orientation, will share either a start coordinate (forward reads), an end coordinate (reverse reads) or both (paired end reads).

Based on this assumption, a group of reads that share identical start and end coordinates (or start coordinate and length for single end reads) and also share identical sequences can be considered as potential duplications of the same DNA fragment. These reads are then investigated to find reads to be removed and reads to be kept. In order to explain how this works, we will use a small example shown in figure 20.30.

The example shows 165 reads that share the same start position and orientation and are considered for duplicate read removal. 60 reads share the sequence shown at the top, 5 reads share the middle sequence, and 100 reads share the sequence at the bottom. The differences



```
ACGGACTGCTT   60
ACTGACTGCTT    5
ACTGACTGATT  100
```

Figure 20.30: *An alignment of three different sequence. The numbers are read counts, e.g. the read at the top occurs 60 times.*

are at position 3 and 9 (underlined).

The tool will now create a tree structure out of these reads as illustrated in figure 20.31.

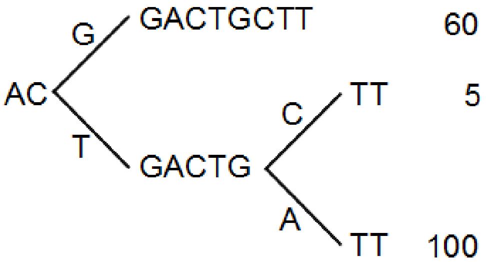


Figure 20.31: *The reads from figure 20.30 represented as a Patricia tree [Morrison, 1968].*

The first branch point in the tree is at the third position, where sequence number one has a G and the other sequences have a T. The other two sequence disagree at position nine, where one has a C and another has an A.

The next step is to iteratively merge the branches, starting from the end of the tree. The first branch point to consider is at position nine. Since only 5 reads have a C and 100 reads have an A, the C branch is collapsed. This is shown in figure 20.32.

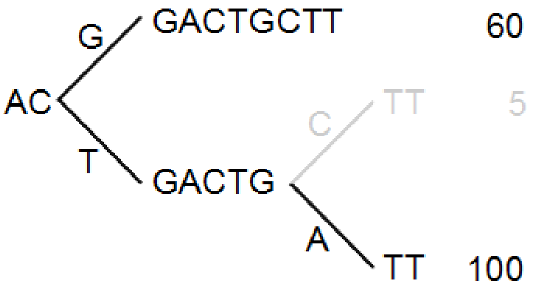


Figure 20.32: *Merging the sequences.*

As a user, you can specify the **threshold** for when the reads should be merged. The default is 20 %: when the minority branch has less than 20 % of the read count of the both branches, it is collapsed.

The next branch to consider is at the third position, where there are now 105 reads that have a T and 60 reads that have a G (see figure 20.33).

With the default setting at 20%, these two branches will not be collapsed, because there are too

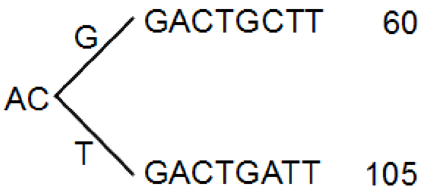


Figure 20.33: *Merging the sequences.*

many reads on the minority branch (60 reads versus 105 reads). Since this process is aimed at collapsing reads that are only distinguished apart by sequencing errors, you would not expect this situation to be caused by sequencing errors, but rather true biological variation (or PCR errors in the early cycles that are indistinguishable from true variation).

If we raised the threshold to 60%, the two branches above would be merged into one if it was not for the second rule governing the merging of branches: The sequences have to be identical except for the difference at the branch point. Looking at the sequences in figure 20.33, there is a difference at position 9 which means that these two branches would never be merged, regardless the threshold and the read counts.

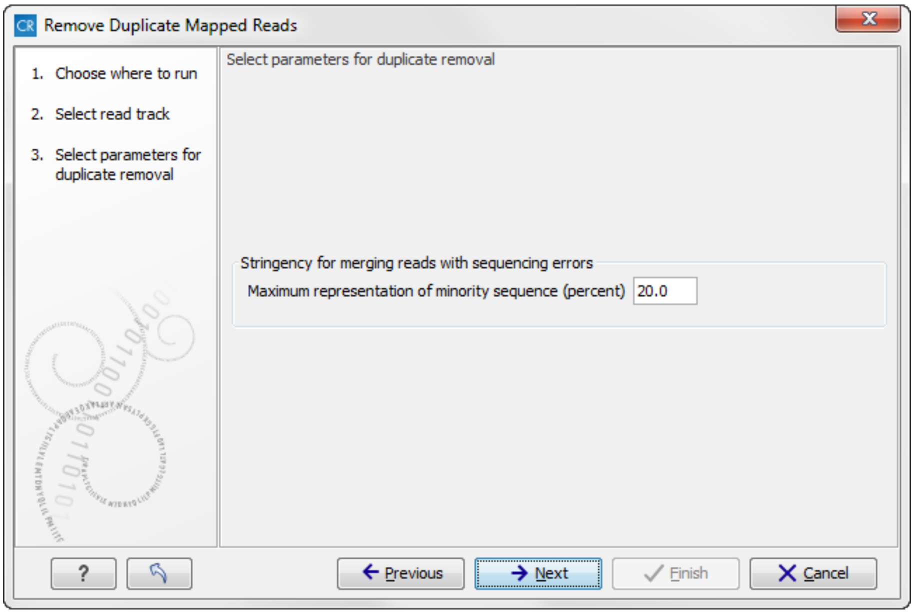
The result of the duplicate reads removal in this example would be that the 165 reads are reduced to two in the result.

20.8.2 Running the duplicate reads removal

The tool is found in the Toolbox:

Toolbox | **Resequencing Analysis** ( ) | **Remove Duplicate Mapped Reads** ( )

This opens a dialog where you can select mapping results in read tracks ( ) format. Clicking **Next** allows you to set the threshold parameters as displayed in figure 20.34.


Remove Duplicate Mapped Reads
1. Choose where to run
2. Select read track
3. Select parameters for duplicate removal
?
Previous
Next
Finish
Cancel

Select parameters for duplicate removal

Stringency for merging reads with sequencing errors

Maximum representation of minority sequence (percent) 20.0

Figure 20.34: *Setting the stringency for merging similar reads.*

The parameter is explained in detail in section 20.8.1.

20.9 Coverage analysis

The coverage analysis tool is designed to identify regions in read mappings with unexpectedly low or high coverage. Such regions may e.g. be indicative of a deletion or an amplification in the sample relative to the reference. The algorithm fits a Poisson distribution to the observed coverages in the positions of the mapping. This distribution is used as the basis for identifying the regions of 'Low coverage' or 'High coverage'. The user chooses two parameter values in the wizard: (1) a 'Minimum length' and (2) a 'P-value threshold' value. The algorithm inspects the coverages in each of the positions in the read mapping and marks the ones with coverage in the lower or upper tails of the estimated Poisson distribution, using the provided p-value as cut-off. Regions with consecutive positions marked consistently as having low (respectively high) coverage, longer than the user specified 'Minimum length' value are called as 'Low coverage' (respectively 'High coverage') regions.

The coverage analysis tool may produce either an annotation track or a table, depending on the users choice, and, optionally, a report. The annotation track (or table) contains a row for each detected low or high coverage region, with information describing the location, the type and the p-value of the detected region. The p-value of a region is defined as the average of the p-values calculated for each of the positions in the region.

20.9.1 Running the Coverage analysis tool

To run the Coverage analysis tool:

Toolbox | **Resequencing Analysis** ( ) | **Coverage Analysis** ( )

This opens the dialog shown in figure 20.35.

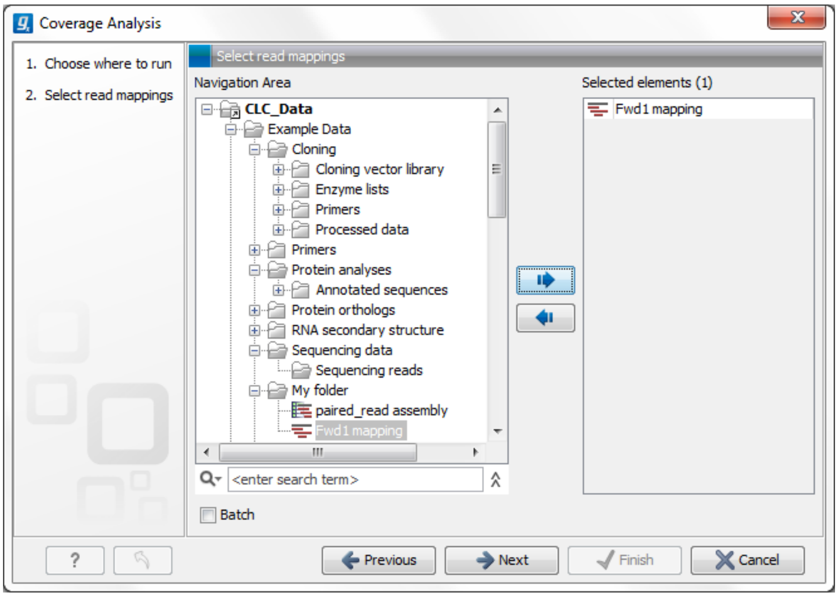
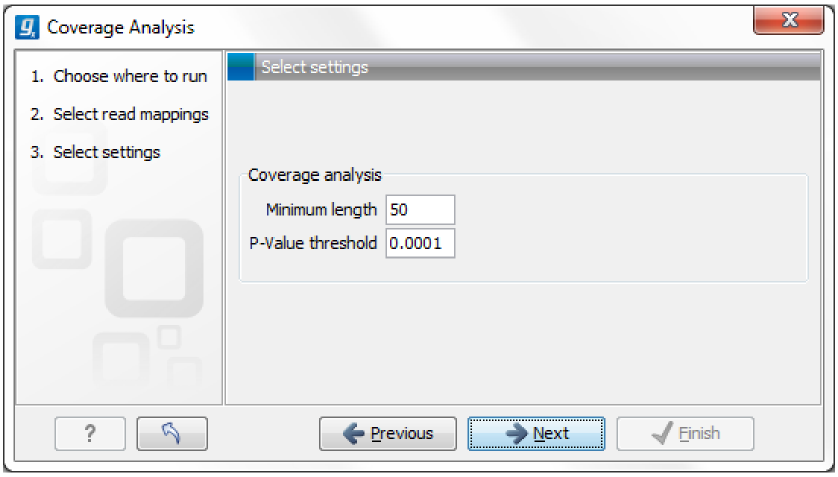


Figure 20.35: *Select read mapping results.*

Select a reads track or read mapping and click **Next**. This opens the dialog shown in figure 20.36.



Coverage analysis

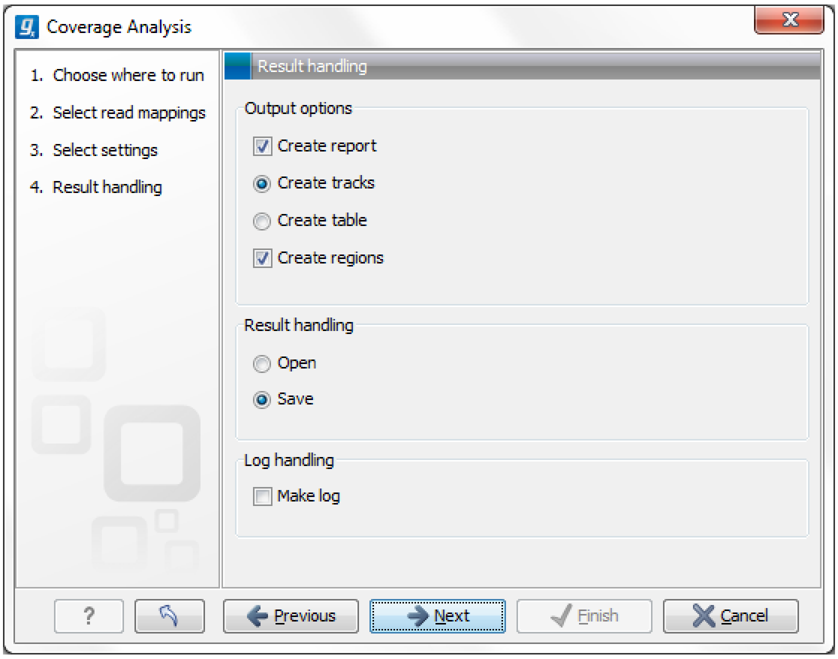
Minimum length: 50

P-Value threshold: 0.0001

Figure 20.36: *Specify the p-value cutoff.*

Set the p-value and minimum length cutoff.

Click **Next** and specify the result handling (figure 20.37).



Output options

☑ Create report

● Create tracks

○ Create table

☑ Create regions

Result handling

○ Open

● Save

Log handling

□ Make log

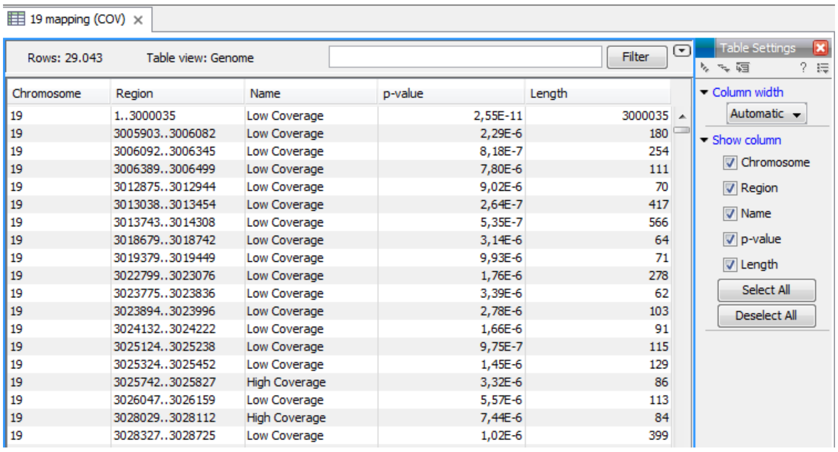
Figure 20.37: *Specify the output.*

Open or save and click **Finish**.

An example of a track output of the Coverage analysis tool is shown in figure 20.38.

In the *CLC Cancer Research Workbench resequencing* is the overall category for applications comparing genetic variation of a sample to a reference sequence. This can be targeted resequencing of a single locus or whole genome sequencing. The overall workflow will typically involve read mapping, some sort of variant detection and interpretation of the variants.

This chapter describes the tools relevant for the resequencing workflows downstream from the



Rows: 29.043 Table view: Genome

Chromosome	Region	Name	p-value	Length
19	1..3000035	Low Coverage	2,55E-11	3000035
19	3005903..3006082	Low Coverage	2,29E-6	180
19	3006092..3006345	Low Coverage	8,18E-7	254
19	3006389..3006499	Low Coverage	7,80E-6	111
19	3012875..3012944	Low Coverage	9,02E-6	70
19	3013038..3013454	Low Coverage	2,64E-7	417
19	3013743..3014308	Low Coverage	5,35E-7	566
19	3018679..3018742	Low Coverage	3,14E-6	64
19	3019379..3019449	Low Coverage	9,93E-6	71
19	3022799..3023076	Low Coverage	1,76E-6	278
19	3023775..3023836	Low Coverage	3,39E-6	62
19	3023894..3023996	Low Coverage	2,78E-6	103
19	3024132..3024222	Low Coverage	1,66E-6	91
19	3025124..3025238	Low Coverage	9,75E-7	115
19	3025324..3025452	Low Coverage	1,45E-6	129
19	3025742..3025827	High Coverage	3,32E-6	86
19	3026047..3026159	Low Coverage	5,57E-6	113
19	3028029..3028112	High Coverage	7,44E-6	84
19	3028327..3028725	Low Coverage	1,02E-6	399

Figure 20.38: *An example of a track output of the Coverage analysis tool.*

actual read mapping.

First comes a description of a tool to perform quality check of targeted resequencing approaches, next we describe the three variant callers that come with the *CLC Cancer Research Workbench* for finding variants, followed by a section describing a coverage analysis tool used to identify fluctuations in coverage. Next, the format of the variants are described, and finally we go through the various tools for filtering, comparing and annotating variants.

20.10 Variant Detectors - overview

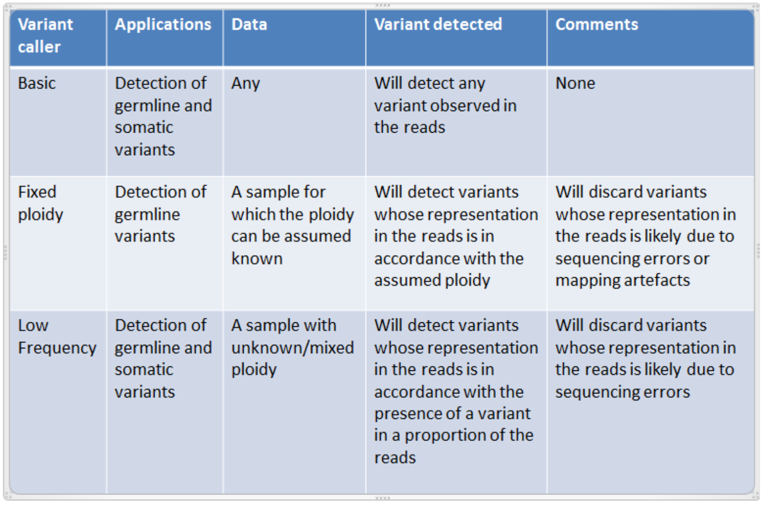
The Variant Detectors tools consist of three different tools for variant detection: The 'Basic Variant Detection' tool, the 'Fixed Ploidy Variant Detection' tool, and the 'Low Frequency Variant Detection' tool. The tools are designed for analysis of different types of samples. They differ in their underlying assumptions about the data, and hence differ in their assessments of when there is enough information in the data for a variant to be called. They share a number of options regarding which reads in the data should be considered, and options related to the filtering of called variants.

The Cancer and Genomics Workbenches offer three tools for detecting variants:

- The 'Basic Variant Detection' tool
- The 'Fixed Ploidy Variant Detection' tool and
- The 'Low Frequency Variant Detection' tool

The 'Basic Variant Detection' and the 'Fixed Ploidy Variant Detection' tools are new versions of the 'Quality-based Variant Detection' and the 'Probabilistic Variant Detection' tools, in which the filtering options have been unified and extended. The 'Quality-based Variant Detection' and the 'Probabilistic Variant Detection' are on their way to be retired and have been moved to the Legacy Tools.

The Basic Variant Detection, Fixed Ploidy Variant Detection and the Low Frequency Variant Detection tools *differ in their underlying assumptions about the data*, and are suitable for different types of applications. An overview is given in figure 20.39.



Variant caller	Applications	Data	Variant detected	Comments
Basic	Detection of germline and somatic variants	Any	Will detect any variant observed in the reads	None
Fixed ploidy	Detection of germline variants	A sample for which the ploidy can be assumed known	Will detect variants whose representation in the reads is in accordance with the assumed ploidy	Will discard variants whose representation in the reads is likely due to sequencing errors or mapping artefacts
Low Frequency	Detection of germline and somatic variants	A sample with unknown/mixed ploidy	Will detect variants whose representation in the reads is in accordance with the presence of a variant in a proportion of the reads	Will discard variants whose representation in the reads is likely due to sequencing errors

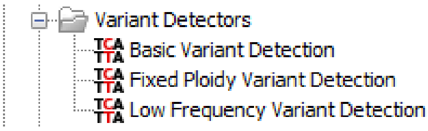
Figure 20.39: *An overview of the applicabilities of the variant detection tools.*

Below we first describe the tools. Each of the tools have a set of parameters that are specific to that tool. Next, we describe the filtering and output options that are shared among the tools (section 20.15).

To run the Variant Detection tools, go to:

Toolbox | **Resequencing Analysis** ( ) | **Variant Detectors**

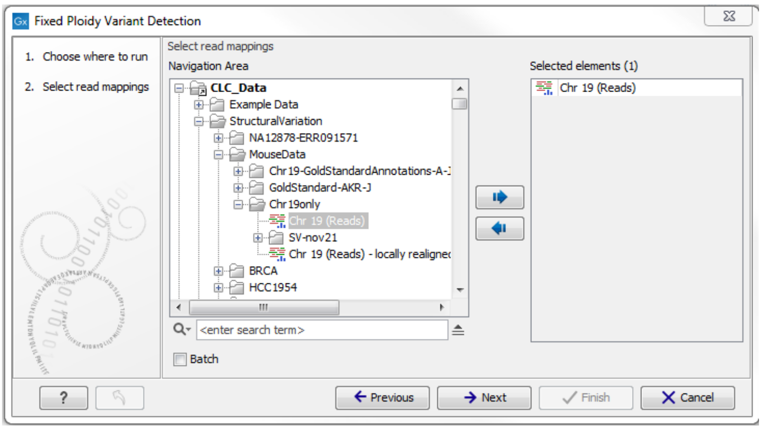
Here you are presented with the three tools (see figure 20.40).



Variant Detectors
- Basic Variant Detection
- Fixed Ploidy Variant Detection
- Low Frequency Variant Detection

Figure 20.40: *The Variant Detection tools.*

When double-clicking one of the tools, a dialog is opened where you select the **reads track** or read mapping you want to analyze.



Fixed Ploidy Variant Detection

1. Choose where to run
2. Select read mappings

Select read mappings

Navigation Area

- CLC_Data
  - Example Data
  - StructuralVariation
    - NA12878-ERR091571
    - MouseData
      - Chr19-GoldStandardAnnotations-A-J
      - GoldStandard-AKR-J
      - Chr19only
        - Chr 19 (Reads)
        - SV-nov21
        - Chr 19 (Reads) - locally realigned
    - BRCA
    - HCC1954

<enter search term>

□ Batch

Selected elements (1)

Chr 19 (Reads)

Figure 20.41: *Select the read mapping that you want to analyze.*

Click **Next** when the reads track is listed in the right-hand side of the dialog.

The user is next asked to set the parameters that are specific for the variant detection tool. The three tools, their assumptions, and the tool-specific parameters are described in the sections

20.11, 20.12, 20.13. First we will here give some examples of different variants called by the three variant detection tools and the overall mode by which the variant detection is carried out.

20.10.1 Differences among the variants called by the three variant callers

The Variant Detection tools will call SNVs, MNVs (which are neighboring SNVs for which there is evidence in the data that they occur together), small to medium-sized insertions and deletions (the size of the insertions and deletions that the variant detection tools are able to call is restricted by the fact that they need to be represented within a single read), and replacements (which are neighboring SNVs and indels).

As the tools *differ in their underlying assumptions about the data*, they differ in their assessments of when there is enough information in the data for a variant to be called, and hence *will call different variants*. However, when run with the *same* filter settings (section 20.15 for a description of the filters), you will generally have that:

- The Basic Variant Caller will call the highest number of variants. It will also do this relatively quickly, as it does not do any error-model estimation.
- The Low Frequency Variant Caller will call *a subset of the variants called by the Basic Variant caller*. The variants called by the Basic Variant Caller that the Low Frequency Variant Caller will NOT call, are those that, according the error model that the Low Frequency Variant Caller estimates from the data, are likely to have been caused by sequencing errors. The Low Frequency Variant Caller will be the slowest of the three variant callers as it (1) estimates an error-model and (2) calls Low Frequency variants (and not just those that are in accordance with a specified ploidy model).
- The Fixed Ploidy Variant Caller will call *a subset of the variants called by the Low Frequency Variant caller*. The variants called by the Low Frequency Variant Caller that the Fixed Ploidy Variant Caller will NOT call, are those that, according to the assumed ploidy of the sample analyzed and the error model that the Fixed Ploidy Variant Caller estimates from the data, are likely to have been caused by either mapping errors or by sequencing errors.

Figure 20.42 shows variant calls produced by the three variant callers when run with the same filter settings, more precisely those that are default for the Low Frequency Variant Caller. The numbers of called variants are shown in the left part of the figure, under the variant track names 'basicV2', 'LowFreq' and 'FixedV2'. The Basic Variant Caller calls most variants and the Fixed Ploidy the least. The Fixed Ploidy Variant Caller calls a subset of those called by the Low Frequency Variant caller, which in turn calls a subset of those called by the Basic Variant caller – in spite of the fact that there are 9 variants in the Low Frequency variant track that are not in the Basic Variant track. Although those 9 variants are in fact not in the Basic Variant track, they are 'sub-variants' of variants in that track. The highlighted variants in the figure is an example of this: The Basic variant caller has called a heterozygous 2bp MNV. The Low Frequency variant caller has judged that one on the SNVs constituting this 2bp MNV is likely to be the result of sequencing errors, and has only called one of the SNVs.

In figure 20.43 a variant is highlighted that is detected by the Basic Variant Caller but not by the Low Frequency or the Fixed Ploidy Variant Caller. The variant is present at a low frequency in a high coverage position. The Low Frequency Variant Caller compares this evidence to the error model, and has decided that the three reads carrying the variant are likely to be the result of sequencing

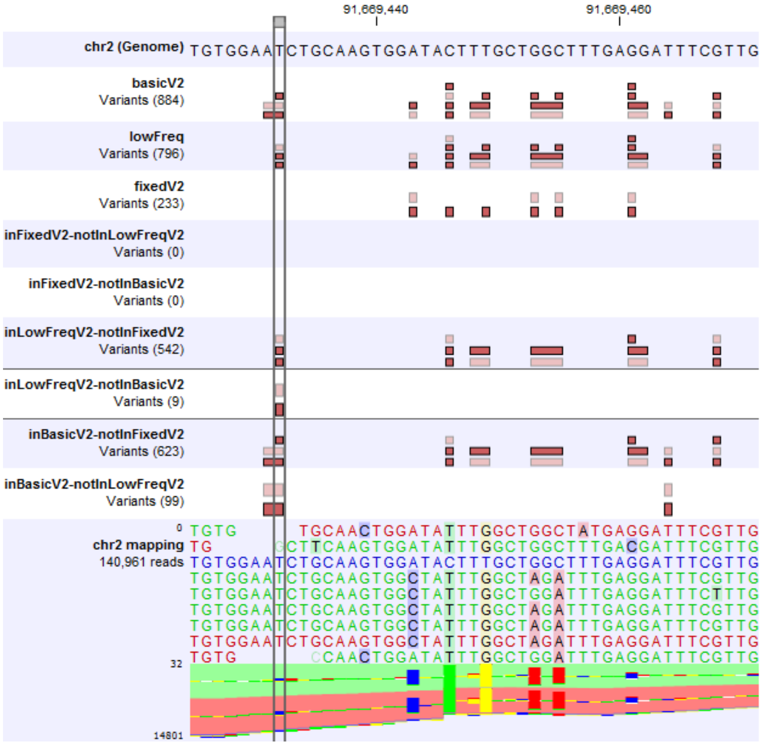
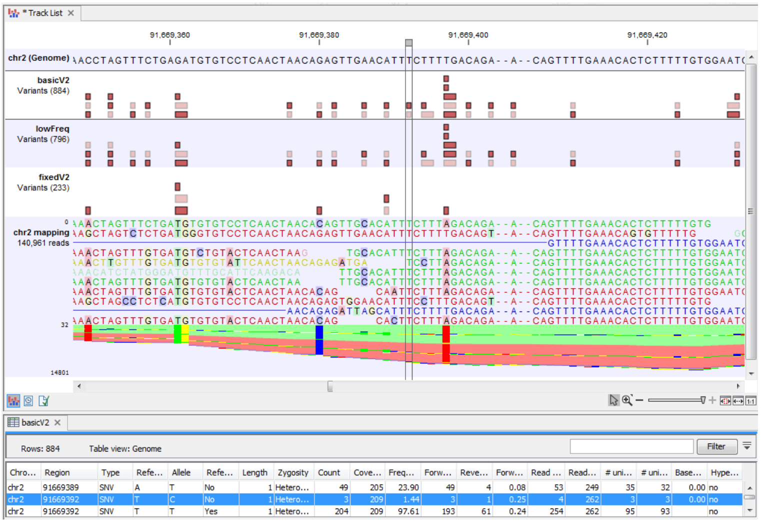


Figure 20.42: *The differences in variants called by the three variant callers. The variant callers have all been run with same the filter settings (those that are the defaults for the Low Frequency Variant Caller).*

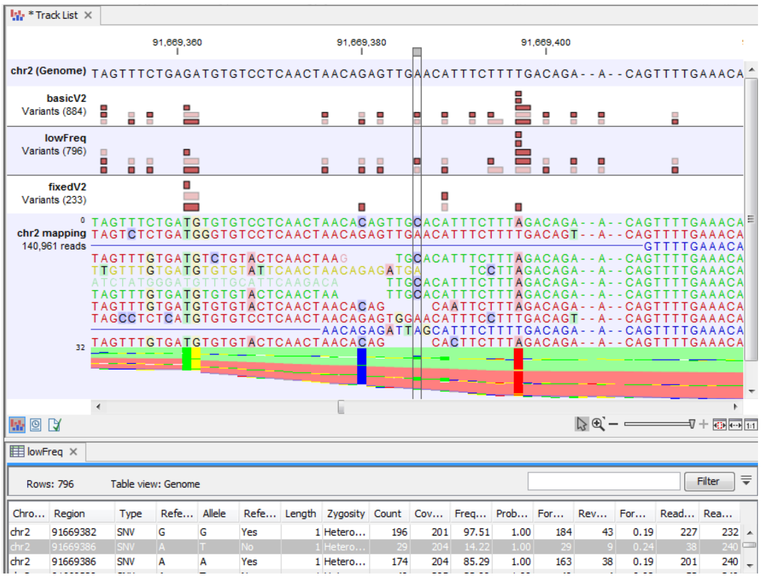


Rows: 884 Table view: Genome

Chro...	Region	Type	Refe...	Allele	Refe...	Length	Zygosity	Count	Cove...	Freq...	Forw...	Reve...	Forw...	Read ...	Read...	# uni...	# uni...	Base...	Hype...
chr2	91669389	SNV	A	T	No	1	Hetero...	49	205	23.90	49	4	0.08	53	249	35	32	0.00	no
chr2	91669392	SNV	T	C	No	1	Hetero...	3	209	1.44	3	1	0.25	4	262	3	3	0.00	no
chr2	91669392	SNV	T	T	Yes	1	Hetero...	204	209	97.61	193	61	0.24	254	262	95	93		no

Figure 20.43: *A variant is highlighted that is detected by the Basic Variant Caller but not by the Low Frequency or the Fixed Ploidy Variant Caller. The variant track for the Basic variant Caller variants is opened in the table-view at the bottom of the figure. The variant is present at a low frequency in a high coverage position, and is likely to have been caused by sequencing error.*

errors, rather than the result of a true variant. Figure 20.44 highlights a variant that is detected by both the Basic and the Low Frequency Variant Caller, but not the Fixed Ploidy. The variant is present at a higher frequency (14.22%) in a high coverage region (coverage 204). Observing the variant in 29 out of 204 reads is not likely to be due to sequencing errors. However, observing 29 reads from one allele and the remaining from the other in a diploid sample is highly unlikely, and the Fixed Ploidy Variant Caller judges that this variant is most likely caused by mapping errors (that is, a subset of the reads in the region being mapped there spuriously) and filters out this variant.



Rows: 796 Table view: Genome

Chro...	Region	Type	Refe...	Allele	Refe...	Length	Zygosity	Count	Cov...	Freq...	Prob...	For...	Rev...	For...	Read...	Rea...
chr2	91669382	SNV	G	G	Yes	1	Hetero...	196	201	97.51	1.00	184	43	0.19	227	232
chr2	91669386	SNV	A	T	No	1	Hetero...	29	204	14.22	1.00	29	9	0.24	38	240
chr2	91669386	SNV	A	A	Yes	1	Hetero...	174	204	85.29	1.00	163	38	0.19	201	240

Figure 20.44: *A variant is highlighted that is detected by the Basic and the Low Frequency but not by the Fixed Ploidy Variant Caller. The variant track for the Low Frequency Variant Caller variants is opened in the table-view at the bottom of the figure. The variant is present at a moderate frequency in a high coverage position, and is,* under the assumed ploidy, *most likely to have been caused by mapping error.*

20.10.2 How the variant detectors work

The Variant detectors share a set of filters. They relate to (i) which areas and positions of the read mappings that should be inspected for variants, (ii) which reads in the data should be considered when this assessment is done, (iii) requirements to the coverage, frequency and absolute counts of variant carrying reads and (iv) the quality and neighborhood composition of the area surrounding the variant. The filters are described in detail in section 20.15.

The variant detectors operate in the following step-wise fashion:

1. Estimate an error model (Fixed Ploidy and Low Frequency Variant Detectors only).
2. Examine each single nucleotide position for the presence of a potential variant while:
   (a) Ignoring the positions, regions and reads specified by the 'Reference masking' and 'Read filter' parts of the 'General filters' (section 20.15).
   (b) Applying half the cut-offs specified by the user in the 'Count and coverage filters' part of the 'General filters' (section 20.15).
   (c) Applying the cut-offs specified in the Noise filters (section 20.15).

   Discard the single nucleotide variant if it's presence in the reads is not 'significantly' better explained by being variant than by being due to sequencing and/or mapping errors (Fixed Ploidy and Low Frequency Variant Detectors only. For details see section 20.12 and 20.13.

3. For the single position variants that survived the initial screening in 2., examine if neighboring variants are present in the same reads. If so, 'join' them into MNVs, longer insertions or deletions, or into replacements.

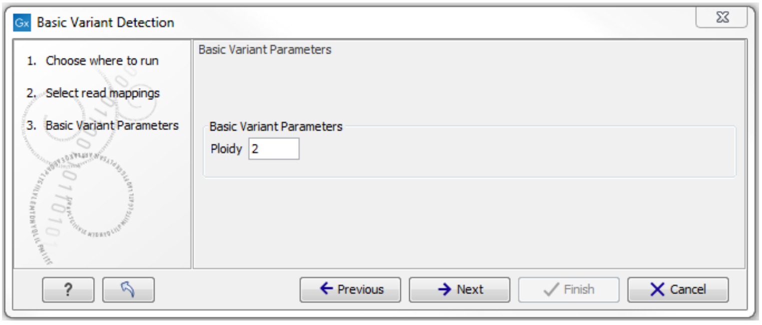
4. Apply the full cut-offs of the 'Count and coverage filters' part of the 'General filters' to all the variants (single and multiple positions) that were obtained after 3.

The reason for first examining the read mapping for variants in single positions is that it is at the single position level that we can estimate an error-model, and hence it is at the single position level that we can distinguish true variants from likely sequencing errors. The reason for joining the single position variants, *when they are present in the same reads*, is that this gives us evidence that they occur together (e.g., it is unsatisfactory to call three single base neighboring deletions if we can see that they occur in the same reads). The reason for applying only half of the cut-off of the 'Count and coverage filters' part of the 'General filters' on the initial single position examination (in step 2.), and wait with the full cut-off till after the variants have been joined (in step 4), is that it will sometimes happen that coverage drops within a longer variant. This can result in single position variants within a longer variant not being called, if the full filters were initially applied. By applying only half the cut-off in the initial examination, this risk is decreased.

Having described the overall mode of the variant detection tools, we will now describe individual characteristics and specific assumptions of the three variant detection tools. The filtering and output options that are shared among the tools are described in the (section 20.15) and section 20.16.

20.11 Basic Variant Detection

The Basic Variant Detection tool does not rely on any assumptions on the data, and does not estimate any error model. It can be used on any type of sample. It will call a variant if it satisfies the requirements that you specify when you set the filters (see section 20.15). The tool has a single parameter (figure 20.45) that is specific to this tool: the user is asked to specify the 'ploidy' of the sample that is being analyzed. The value of this parameter does not have an impact on which variants are called - it will merely determine the contents of the 'hyper-allelic' column that is added to the variant track table: variants that occur in positions with more variants than expected given the specified ploidy, will have 'Yes' in this column, other variants will have 'No' (see section 20.16 for a description of the outputs).



Ploidy 2

Figure 20.45: *The Basic Variant Detection parameters.*

20.12 Fixed Ploidy Variant Detection

The Fixed Ploidy Variant Detection tool relies on two models:

1. A model for the possible 'site-types' and
2. A model for the sequencing errors.

For (i), the set of possible 'site-types' depend on the user-specified ploidy parameter: For a diploid organism there are two alleles and thus the site types are A/A, A/C, A/G, A/T, A/-, C/C, and so on until -/-. The error model, (ii), specifies the probabilities of having a certain base in the read, but calling a different base. The error model is estimated from the data prior to calling the variants (see section 20.14). The Fixed Ploidy algorithm will, given the estimated error model and the data observed in the site, calculate the probabilities of each of the site types. One of those site types is the site that is homozygous for the reference - that is, it stipulates that whatever differences are observed from the reference nucleotide in the reads is due to sequencing errors. The remaining site-types are those which stipulate that at least one of the alleles in the sample is different from the reference. The sum of the probabilities for these latter site types is the posterior probability that the sample contains at least one allele that differs from the reference at this site. We refer to this posterior probability as the 'variant probability'.

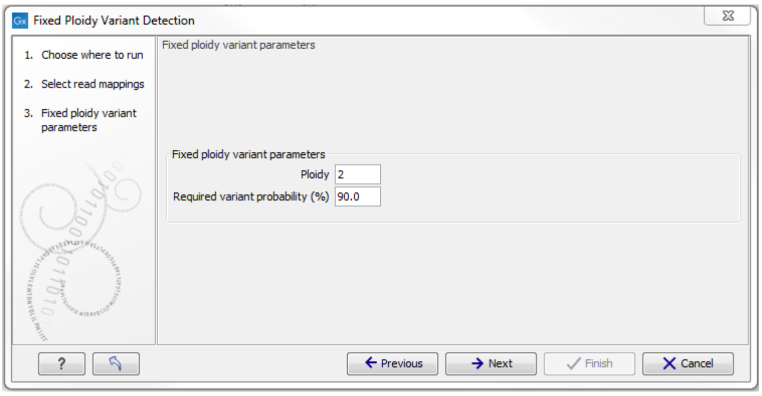
The Fixed Ploidy Variant Detection tool has two parameters: the 'Ploidy' and the 'Variant probability' parameters (figure 20.46):

- The 'ploidy' is the ploidy of the analyzed sample. The value that the user sets for this parameter determines the site types that are considered in the model. For more information about ploidy please see section 20.12.1.
- The 'Required variant probability' is the minimum value of the variant probability required for the variant to be called. That is, only variants with a probability higher than the specified value will be called. That means that the higher the value you set, the fewer variants are called.

As the Fixed Ploidy Variant Detection tool strongly depends on the model assumed for the ploidy, the user should carefully consider the validity of the ploidy assumption that he makes for his sample. The tool allows ploidy values up to and including 4 (tetraploids). For higher ploidy values the number of possible site types is too large for estimation and computation to be feasible, and the user should use the Low Frequency or Basic Variant Detection Tool instead.

20.12.1 Ploidy and sensitivity

Core to how this tool works is the ploidy level you set. This defines the statistical model that will be used during the variant detection analysis and thereby also defines what will be reported. When you set up a Fixed Ploidy Variant Detection, you provide a ploidy and you also provide a variant probability. The variant probability parameter defines how good the evidence has to be at a particular site for the tool to report a variant at that location. If the site passes this threshold, (note that it's 'the site' and not 'the variant' here), then the variant with the highest probability at that site will be reported. That is, at a given location, you get one variant reported. The number of alleles that variant may have depends on the value that has been chosen for the ploidy parameter. For example, if you chose a ploidy of 2, then the variant at a site could be a


Fixed Ploidy Variant Detection
1. Choose where to run
2. Select read mappings
3. Fixed ploidy variant parameters
Fixed ploidy variant parameters
Previous
Next
Finish
Cancel

Fixed ploidy variant parameters

Ploidy 2

Required variant probability (%) 90.0

Figure 20.46: *The Fixed ploidy Variant Detection parameters.*

homozygote (two alleles the same in the sample, but different to the reference), or a heterozygote (two alleles different than each other in the sample, with at least one of them different from the reference). If you had chosen a ploidy of three, then the variant at a site could be a homozygote (three alleles the same in the sample, but different to the reference), or a heterozygote (three alleles different than each other in the sample, with at least one of them different from the reference). So how sensitive this tool is, in terms of what it detects and reports to you, will be down to:

- The statistical model. Here the ploidy you set is crucial to the model and what will be detected.
- The variant probability setting. This affects which locations will be reported.

So, to increase sensitivity, you could decrease the variant probability setting or increase the ploidy. The increased sensitivity due to decreasing the variant probability setting is due to more sites being reported on. This could decrease false negatives, but it could also increase your reporting of false positives.

The increased sensitivity you would get by increasing the ploidy level is down to the extra allele types that will be included in the model and thus reported to you in the results. For example, let's say you set a ploidy of 2 and at a particular location, where the evidence that the position was different than the reference was high enough, the reference was a T. The variant with the highest probability at this location will be reported. Let's say that it was a homozygote with C at that position. So that's C|C. However, let's say that in fact, there are some Gs reported at that site in some reads. In the diploid model, all the possibilities will have been tested (e.g. A|A, A|C....C|C, C|G. C|T....and so on). However, in this example, C|C had the highest probability, so it is reported. It doesn't really matter how much deeper the mapping is at that location if the relative prevalence of Gs is low compared to Cs. (That is, the probability of C|C will stay higher than C|G if there were relatively few Gs at this site, so C|C would be reported as long as this stayed the case.) Let's say that you chose a ploidy of 3 instead. Now the model will test all the triploid possibilities (e.g. A|A|A, A|A|C, A|A|G.....C|C|A, C|C|C, C|C|G.... and so on). Now, for the same site, say that the evidence in the reads resulted in the variant C|C|G having a higher probability than C|C|C, then it would be the variant reported. So in this way, you have increased the sensitivity. i.e you have reported a variant that represents the evidence of the reads with G at that position as well as the ones reporting a C at that position.

20.13 Low Frequency Variant Detection

As the Fixed Ploidy Variant Detection tool, the Low frequency variant Detection tool relies on

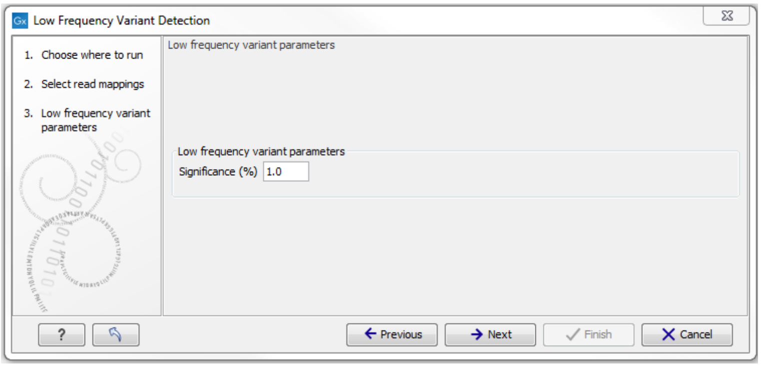
1. A statistical model for the analyzed sample and
2. A model for the sequencing errors.

The method employed in the Low Frequency Variant Detection tool for estimating the sequencing error rates is similar to that of the Fixed Ploidy Variant Detection tool (see section 20.14), but the statistical model for the sample is different. It does not make any assumptions about the ploidy of the sample. Instead a statistical test is performed at each site to determine if the nucleotides observed in the reads at that site could be due simply to sequencing errors, or if they are significantly better explained by there being one (or more) alleles than the reference present in the sample at some unknown frequency. If the latter is the case, a variant corresponding to the significant allele will be called, with estimated frequency.

The Low Frequency Variant Detection tool has one parameter (figure 20.47):

- 'Required Significance': this parameter determines the cut-off value for the statistical test for the variant not being due to sequencing errors. Only variants that are at least this significant will be called. That means that, the lower you set this cut-off, the fewer variants will be called.

The Low Frequency Variant Detection tool is suitable for analysis of samples of mixed tissue types (such as cancer samples) in which low frequent variants are likely to be present, as well as for samples for which the ploidy is unknown or not well defined. The tool also calls more abundant variants, and can be used for analysis of samples with ploidy larger than four. Note that, as the tool looks for all variants, abundant as well as low frequency ones, analysis will generally be slower than those of the other variant detection tools. In particular, it will be very slow, and possibly prohibitively so, for samples with extremely high coverage, or a very large number of variants (as in cases where the sample differs considerably from the reference).


Low Frequency Variant Detection
1. Choose where to run
2. Select read mappings
3. Low frequency variant parameters
Low frequency variant parameters
?
Previous
Next
Finish
Cancel

Low frequency variant parameters

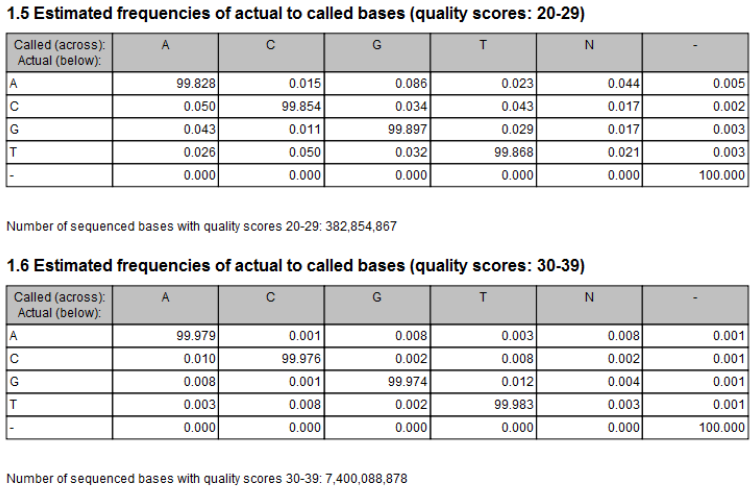
Significance (%) 1.0

Figure 20.47: *The Low Frequency Variant Detection parameters.*

20.14 Variant Detectors - error model estimation

The Fixed Ploidy and Low Frequency Variant Detection tools both rely on statistical models for the sequencing error rates. An error model is assumed and estimated for each quality score.

Typically low quality read nucleotides will have a higher error rate than high quality nucleotides. In the error models, different types of errors have their own parameter, so if A's for example more often tend to result in erroneous G's than other nucleotides, that is also recognized by the error models. The parameters are all estimated from the data set being analyzed, so will adapt to the sequencing technology used and the characteristics of the particular sequencing runs. Information on the estimated error rates can be found in the Reports (see section 20.16).



1.5 Estimated frequencies of actual to called bases (quality scores: 20-29)

Called (across):
Actual (below):
	A	C	G	T	N	-

A	99.828	0.015	0.086	0.023	0.044	0.005
C	0.050	99.854	0.034	0.043	0.017	0.002
G	0.043	0.011	99.897	0.029	0.017	0.003
T	0.026	0.050	0.032	99.868	0.021	0.003
-	0.000	0.000	0.000	0.000	0.000	100.000

Number of sequenced bases with quality scores 20-29: 382,854,867

1.6 Estimated frequencies of actual to called bases (quality scores: 30-39)

Called (across):
Actual (below):
	A	C	G	T	N	-

A	99.979	0.001	0.008	0.003	0.008	0.001
C	0.010	99.976	0.002	0.008	0.002	0.001
G	0.008	0.001	99.974	0.012	0.004	0.001
T	0.003	0.008	0.002	99.983	0.003	0.001
-	0.000	0.000	0.000	0.000	0.000	100.000

Number of sequenced bases with quality scores 30-39: 7,400,088,878

Figure 20.48: *Example of estimated error rates. The figure shows average estimated error rates across bases in the given quality score intervals (20-29 and 30-39, respectively). Higher error rates are estimated for bases with lower quality scores.*

An example of error rates estimated from a whole exome sequencing Illumina data set is shown in figure 20.48. As expected, the estimated error rates (that is, the off-diagonal elements in the matrices in the figure) are higher for the lower quality nucleotides than for higher. Note that, although the matrices in the figure show error rates of bases within *ranges of* quality scores, a separate matrix is estimated for each quality score in the error model estimation.

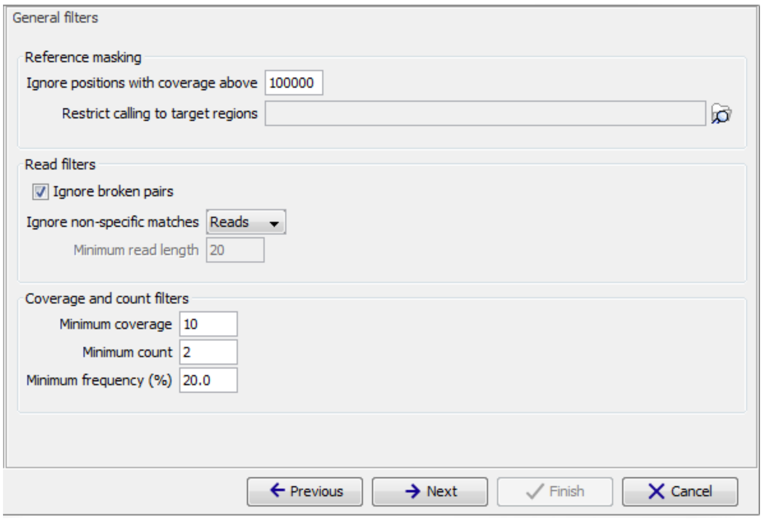
20.15 Variant Detectors - filters

The variant detectors offer a number of filters. These relate both to which reads should be used, and how much evidence should be required for a variant to be called. The user is asked to set the values of these filters in two wizard steps: the 'General filters' step (figure 20.49) and the 'Noise filters' step (figure 20.50). Note that, for most of the filter, the values that they calculate and base their filtering on are also added as annotations to the variants (see section 20.16). This means that they are available in the variant track and that the user can choose to perform the filtering in a post-processing step from the variant track table, rather than choosing to apply the filtering within the variant calling detection it self. The filters are described below.

20.15.1 General filters

The 'General' filters relate to the regions and reads in the read mappings that should be considered, and the amount of evidence the user wants to require for a variant to be called:

Note on the use of the rare variant caller with WGS data: The default settings for the rare variant caller are optimized for targeted resequencing protocols, not whole genome sequencing (e.g. cancer gene panels). In order to run the tool on WGS data the the parameter 'Ignore



General filters

Reference masking

Ignore positions with coverage above 100000

Restrict calling to target regions

Read filters

☑ Ignore broken pairs

Ignore non-specific matches Reads

Minimum read length 20

Coverage and count filters

Minimum coverage 10

Minimum count 2

Minimum frequency (%) 20.0

Figure 20.49: *General filters. The values shown are those that are default for Fixed Ploidy Variant detection.*

positions with coverage above' should be adjusted (to e.g. 1000) as it is not uncommon to have modest coverage for most part of the mapping, and abnormal areas (typically repeats around the centromeres) with very high coverage. Looking for low frequency variants in high coverage areas will exhaust the machine memory simply because there will be many low frequency variants (due to some reads originating from near identical repeat sequences or simple sequencing errors). As these 'high coverage' regions typically are not of interest, disregard these areas by using a lower 'Ignore positions with coverage above' filtering setting is recommended for WGS data.

Reference masking

The 'Reference masking' filters allow the user to only perform variant calling (incl. error model estimation) in specific regions. There are two parameters to specify this:

- **Ignore positions with coverage above:** All positions with coverage above this value will be ignored when inspecting the read mapping for variants.
- **Restrict calling to target regions:** Only positions in the regions specified will be inspected for variants.

Note that when the option **Ignore positions with coverage above** is chosen, any position with coverage above the specified value will be ignored by the algorithm, and no variants will be called at these positions. The option is highly useful in cases where you have a read mapping which has areas of extremely high coverage, that you are not really interested in. An example is areas around centromeres in whole genome sequencing applications. The computational requirements may be prohibitive if you consider the full read mapping, but manageable if you exclude the (un-interesting) high-coverage areas.

Also note that the **Restrict calling to target regions** parameter is optional. When not specified, the full read mapping will be examined.

Read filters

The 'Read filters' determine which reads (or regions) should be considered when calling the variants.

- **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected.
- **Non-specific match filter:** Non-specific matches are likely to come from some type of repeat region, and the exact mapping location of them is uncertain. In general, variants based on non-specific matches are likely to be less reliable. However as there are regions in the genome that are entirely perfect repeats, ignoring non-specific matches may have the effect that true variants go undetected in these regions.

  There are three options for specifying to which 'extent' the non-specific matches should be ignored:

  - 'No': when this option is chosen they are not ignored.
  - 'Reads': when this option is chosen they are ignored.
  - 'Region': when this option is chosen no variants are called in regions covered by at least one non-specific match.

  When ignoring regions containing a non-specific match (the last of the options mentioned above), the minimum length of reads that are allowed to trigger this effect has to be stated. The reason is that we want to avoid really short reads triggering the effect, as really short reads will usually be non-specific even if they do not stem from repeat regions.

Coverage and count filters

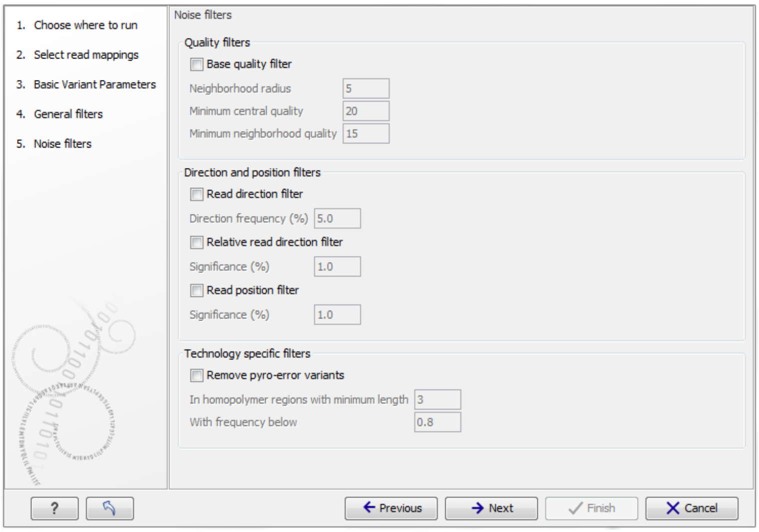
These filters specify absolute requirements for the variants to be called. Note that suitable values for these filters are highly dependent on the coverage in the sample being analyzed:

- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at, at least, this frequency (calculated as 'count'/'coverage') are called.

These values are calculated for each of the detected candidate variants. If the candidate variant meets the specified requirements, it is called. Note that when the values are calculated, only the 'countable reads' are considered. The 'countable reads' are those that the user has not chosen to ignore. This means that, if the user, in the read filter, has specified that reads from broken pairs should be ignored, broken pair reads will not be countable. Similarly goes for the non-specific reads, and for reads with bases at the variant position that does not fulfill the base quality requirements specified by the 'Base Quality Filter' (see the section on 'Noise filters' below). Also note that overlapping paired reads only count as one read (since they only represent one fragment).

20.15.2 Noise filters

The 'Noise filters' examine each candidate variant at a more detailed level. They are intended as a means of filtering out variants that are likely to be the result of various types of systematic errors and/or biases, e.g. induced by the amplification or sequencing protocol, that may occur in samples. They should be used with care, as there is always the risk that a real variant has the characteristics of systematically induced variant.



1. Choose where to run
2. Select read mappings
3. Basic Variant Parameters
4. General filters
5. Noise filters

Noise filters

Quality filters

□ Base quality filter

Neighborhood radius 5

Minimum central quality 20

Minimum neighborhood quality 15

Direction and position filters

□ Read direction filter

Direction frequency (%) 5.0

□ Relative read direction filter

Significance (%) 1.0

□ Read position filter

Significance (%) 1.0

Technology specific filters

□ Remove pyro-error variants

In homopolymer regions with minimum length 3

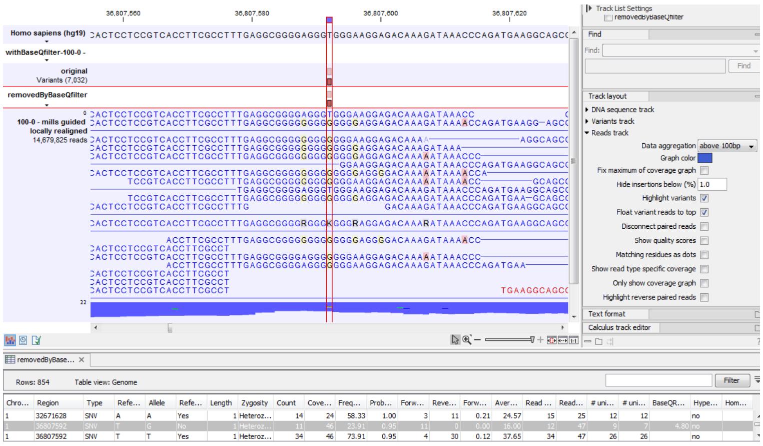
With frequency below 0.8

Figure 20.50: *Noise filters.*

Quality filters

- **Base quality filter:** The base quality filter can be used to ignore the reads whose nucleotide at the potential variant position is of dubious quality. This is assessed by considering the quality of the nucleotides in the read in the region around the nucleotide position. There are three parameters to determine the base quality filter:
  - **Neighborhood radius**: This parameter determines the region size: when a neighborhood radius of five is used, each nucleotide in a read is evaluated based on the nucleotides in the read 5 positions upstream and 5 positions downstream of the examined site - a total of 11 nucleotides. (Note that, near the end of the reads, eleven nucleotides are still considered, by changing the region offset relative to the nucleotide in question).
  - **Minimum central quality:** Reads whose central base has a quality below this value are ignored. This parameter does not apply to deletions, since there is no 'central base' in these cases.
  - **Minimum neighborhood quality:** Read for which the minimum quality of the bases *within the specified neighborhood radius* is below this value, are ignored.

Figure 20.51 gives an example of a variant that is called when the base quality filter is NOT applied, and not called when it is. To understand why it is not called when the base quality filter is applied look at the data in figure 20.52. This figure shows the same data as in figure 20.51, however, now with the 'Show quality scores' option in the side panel of the reads track switched on. This reveals that the reads that carry the potential 'G' variant tend to have poor quality. As all reads that have a base with quality less than 20 in this potential variant position are ignored when the 'Base quality filter' is turned on, no variant is called, most likely because it now does not meet the requirements of either the 'Minimum coverage', 'Minimum count' or 'Minimum frequency' filters. Note that the error in the example shown is a 'typical' Illumina error: the reference has a 'T' that is surrounded by stretches of 'G'. The 'G' signals 'drown' the signal of the 'T'.



Track layout

- DNA sequence track
- Variants track
- Reads track

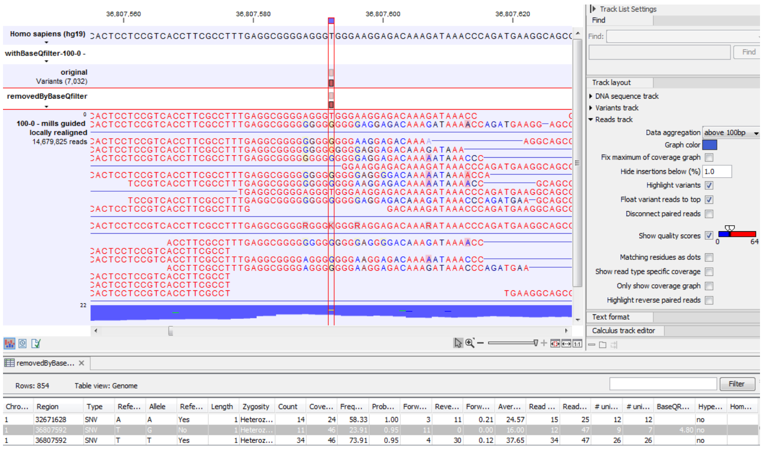
Data aggregation: above 100bp
Graph color: 🟦
Fix maximum of coverage graph ☐
Hide insertions below (%): 1.0
Highlight variants ☑
Float variant reads to top ☑
Disconnect paired reads ☐
Show quality scores ☐
Matching residues as dots ☐
Show read type specific coverage ☐
Only show coverage graph ☐
Highlight reverse paired reads ☐

removedByBase...

Rows: 854 Table view: Genome

Chro...	Region	Type	Refe...	Allele	Refe...	Length	Zygosity	Count	Cove...	Freq...	Prob...	Forw...	Reve...	Forw...	Aver...	Read ...	Read...	# uni...	# uni...	BaseQR...	Hype...	Hom...
1	32671628	SNV	A	A	Yes	1	Heteroz...	14	24	58.33	1.00	3	11	0.21	24.57	15	25	12	12		no	
1	36807592	SNV	T	G	No	1	Heteroz...	11	46	23.91	0.95	11	0	0.00	16.00	12	47	9	7	4.80	no	
1	36807592	SNV	T	T	Yes	1	Heteroz...	34	46	73.91	0.95	4	30	0.12	37.65	34	47	26	26		no	

Figure 20.51: *An example of a variant that is removed by the base quality filter.*



Track layout

- DNA sequence track
- Variants track
- Reads track

Data aggregation: above 100bp
Graph color: 🟦
Fix maximum of coverage graph ☐
Hide insertions below (%): 1.0
Highlight variants ☑
Float variant reads to top ☑
Disconnect paired reads ☐
Show quality scores ☑
Matching residues as dots ☐
Show read type specific coverage ☐
Only show coverage graph ☐
Highlight reverse paired reads ☐

removedByBase...

Rows: 854 Table view: Genome

Chro...	Region	Type	Refe...	Allele	Refe...	Length	Zygosity	Count	Cove...	Freq...	Prob...	Forw...	Reve...	Forw...	Aver...	Read ...	Read...	# uni...	# uni...	BaseQR...	Hype...	Hom...
1	32671628	SNV	A	A	Yes	1	Heteroz...	14	24	58.33	1.00	3	11	0.21	24.57	15	25	12	12		no	
1	36807592	SNV	T	G	No	1	Heteroz...	11	46	23.91	0.95	11	0	0.00	16.00	12	47	9	7	4.80	no	
1	36807592	SNV	T	T	Yes	1	Heteroz...	34	46	73.91	0.95	4	30	0.12	37.65	34	47	26	26		no	

Figure 20.52: *The same data as in figure 20.51, now with the 'Show quality scores' option in the reads track switched on.*

Direction and position filters

Many sequencing protocols are prone to various types of amplification induced biases and errors. The 'Read direction' and 'Read position' filters are aimed at providing means for weeding out variants that are likely to originate from such biases.

- **Read direction filter:** The read direction filter removes variants that are almost exclusively present in either forward or reverse reads. For many sequencing protocols such variants are most likely to be the result of amplification induced errors. Note, however, that the filter is **NOT suitable for amplicon data**, as for this you will not expect coverage of both forward and reverse reads. The filter has a single parameter:
  - **Direction frequency:** Variants that are not supported by at least this frequency of reads from each direction are removed.

- **Relative read direction filter:** The relative read direction filter attempts to do the same thing as the 'Read direction filter', but does this in a statistical, rather than absolute, sense: it tests whether the distribution among forward and reverse reads of the variant carrying reads is different from that of the total set of reads covering the site. The statistical, rather than absolute, approach makes the filter less stringent. The filter has one parameter:
  - **Significance:** Variants whose read direction distribution is significantly different from the expected with a test at this level, are removed. The lower you set the significance cut-off, the fewer variants will be filtered out.
- **Read position filter:** The read position filter is a filter that attempts to remove systematic errors in a similar fashion as the 'Read direction filter', *but* that is also **suitable for hybridization-based data**. It removes variants that are located differently in the reads carrying it, than would be expected given the general location of the reads covering the variant site. This is done by categorizing each sequenced nucleotide (or gap) according to the mapping direction of the read and also where in the read the nucleotide is found; each read is divided in five parts along its length and the part number of the nucleotide is recorded. This gives a total of ten categories for each sequenced nucleotide and a given site will have a distribution between these ten categories for the reads covering the site. If a variant is present in the site, you would expect the variant nucleotides to follow the same distribution. The read position filter carries out a test for whether the read position distribution of the variant carrying reads is different from that of the total set of reads covering the site. The filter has one parameter:
  - **Significance:** Variants whose read position distribution is significantly different from the expected with a test at this level, are removed. The lower you set the significance cut-off, the fewer variants will be filtered out.

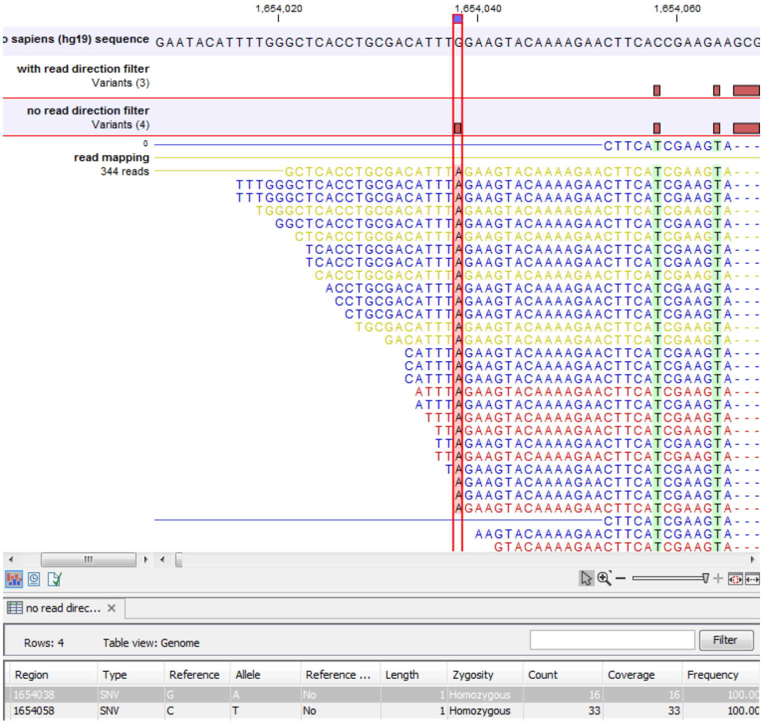
Figure 20.53 shows an example of a variant that is removed by the 'Read direction' filter.

Note that variant calling was done ignoring non-specific matches and broken pair reads, so only the 16 intact paired reads (the blue reads) are considered. To see the direction of the reads, you must adjust the viewer settings in the 'Reads track' side panel, to 'Disconnect paired reads'. This has been done in figure 20.54. Now it becomes apparent that the variant is found in the forward reads (that is, the green reads) of the 16 intact paired reads, and in no reverse reads (except the three that come from broken pairs, and which were ignored), and therefore removed by the read direction data.

Figure 20.55 shows an example of a variant that is removed by the read position filter, but not by the read direction filter. The variant is only present in a portion of the reads that cover the variant, and the portion or the reads that carry the variant have the variant occurring in read positions that are systematically different from what you would expect, given the general placement of reads covering the variant (e.g., none of the reads that start after position 186,641,600 carry the variant).

Technology specific filters

- **Remove pyro-error variants:** This filter can be used to remove insertions and deletions *in the reads* that are likely to be due to pyro-like errors in homopolymer regions. There are two types of such errors: They may occur either at (1) the immediate ends of homopolymer



Rows: 4 Table view: Genome

Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency
1654038	SNV	G	A	No	1	Homozygous	16	16	100.00
1654058	SNV	C	T	No	1	Homozygous	33	33	100.00

Figure 20.53: *An example of a variant that is filtered out by the Read Direction filter.*

regions or (2) as an 'overspill' a few nucleotides downstream of a homopolymer region. In case (1) the exact numbers of the same number of nucleotide is uncertain and a sequence like "AAAAAAAA" is sometimes reported as "AAAAAAAAA". In case (2) a sequence like "CGAAAAAGTCG" may sometimes get an 'overspill' insertion of an A between the T and C so that the reported sequence is C "CGAAAAAGT**A**CG". Note that the removal is done in the reads as a very first step, before calling the initial 1 bp variants.

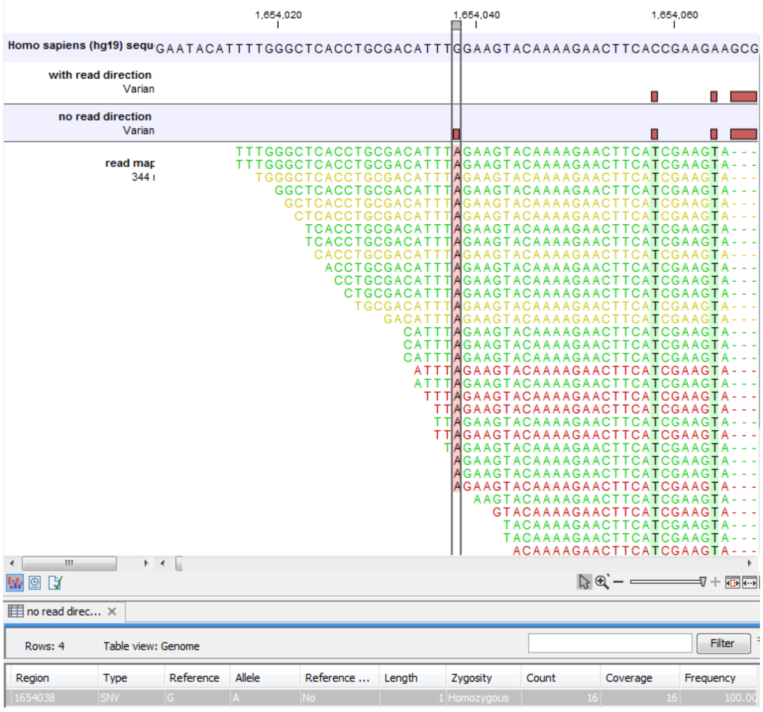
There are two parameters that must be specified for this filter:

- **In homopolymer regions with minimum length**: Only insertion or deletion variants in homopolymer regions of at least this length will be removed.
- **With frequency below:** Only insertion or deletion variants whose frequency (ignoring all non-reference and non-homopolymer variant reads) is lower than this threshold will be removed.

Note that the higher you set the **With maximum frequency** parameter, the more variants will be removed. Figure 20.56 shows an example of a variant that is called when the pyro-error filter with minimum length setting 3 and frequency setting 0.5 is used, but that is filtered when the frequency setting is increased to 0.8. The variant has a frequency of 55.71.

20.16 Variant Detectors - the outputs

The Variant Detection Tools have the following outputs: a variant track, an annotated variant table and a report (figure 20.57). The report contains information on the estimated error model and, as only the Fixed ploidy and the Low Frequency variant callers uses an error model, the report is only available for those, and not for the Basic Variant caller. The outputs are described below.



Rows: 4 Table view: Genome

Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency
1654038	SNV	G	A	No	1	Homozygous	16	16	100.00

Figure 20.54: *The same data as shown in figure 20.53, but now with 'Disconnect paired reads' option switched on in the 'reads track' side panel.*

20.16.1 The variant track output

The variant track contains information on each of the variants called. When opened in the table view there is a number of columns for each of the variants (see figure 20.58).

The contents of these are:

Chromosome The name of the reference sequence on which the variant is located.

Region The region on the reference sequence at which the variant is located. The region may be either a 'single position', a 'region' or a 'between position region'.

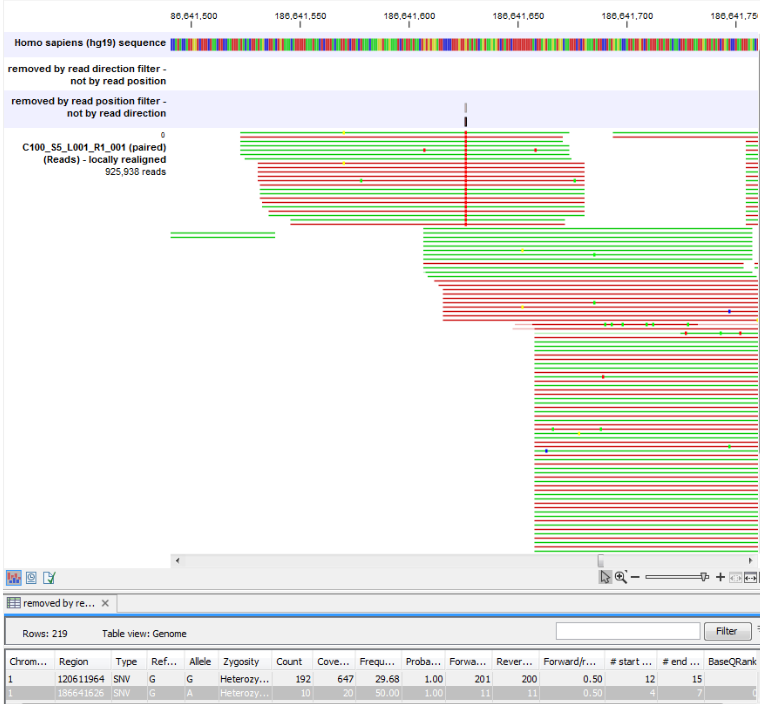
Type The type of variant. This can either be SNV (single-nucleotide variant), MNV (multi-nucleotide variant), insertion, deletion, or replacement.

Reference The reference sequence at the position of the variant.

Allele The allele sequence of the variant.

Reference allele Describes whether the variant is identical to the reference. This will be the case for one of the alleles for most, but not all, detected heterozygous variants (e.g. the variant caller might detect two variants, A and G, at a given position in which the reference is 'A'. In this case the variant corresponding to allele 'A' will have 'Yes' in the 'reference allele' column entry, and the variant corresponding to allele 'G' would have 'No'. Had the variant caller called the two variants 'C' and 'G' at the position, both would have had 'No' in the 'Reference allele' column).

Length The length of the variant. The length is 1 for SNVs, and for MNVs it is the number of allele or reference bases (which will always be the same). For deletions, it is the length


86,641,500
186,641,550
186,641,600
186,641,650
186,641,700
186,641,75
Homo sapiens (hg19) sequence
removed by read direction filter - not by read position
removed by read position filter - not by read direction
0
C100_S5_L001_R1_001 (paired) (Reads) - locally realigned
925,938 reads
removed by re...
Filter

Rows: 219 Table view: Genome

Chrom...	Region	Type	Ref...	Allele	Zygosity	Count	Cove...	Frequ...	Proba...	Forwa...	Rever...	Forward/r...	# start ...	# end ...	BaseQRank
1	120611964	SNV	G	G	Heterozy...	192	647	29.68	1.00	201	200	0.50	12	15	
1	186641626	SNV	G	A	Heterozy...	10	20	50.00	1.00	11	11	0.50	4	7	0

Figure 20.55: *A variant that is filtered out by the Read position filter but not by the Read direction filter.*

of the deleted sequence, and for insertions it is the length of the inserted sequence. For replacements, both the length of the replaced reference sequence and the length of the inserted sequence are considered, and the longest of those two is reported.

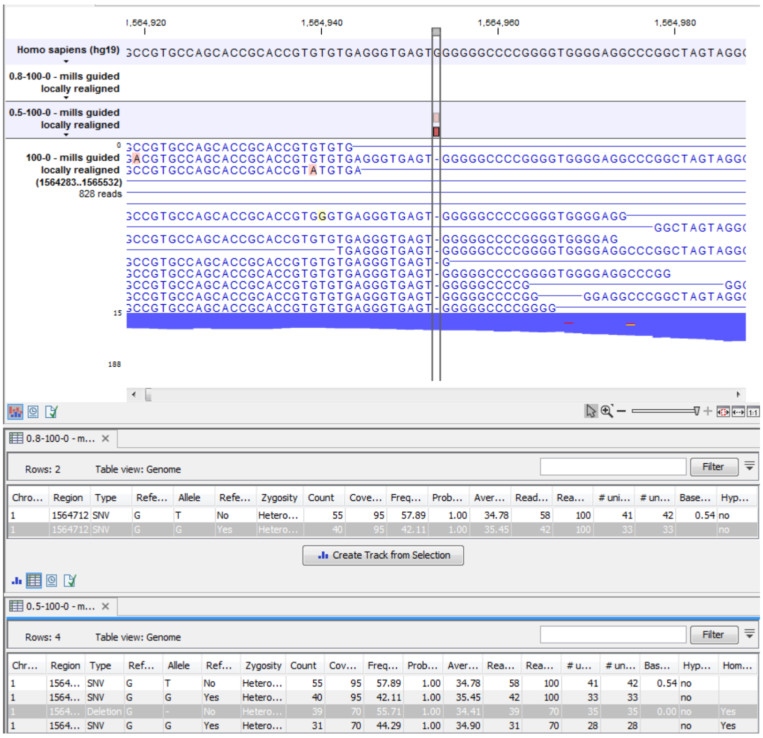
Zygosity The zygosity of the variant called, as determined by the variant caller. This will be either 'Homozygous', where there is only one variant called at that position or 'Heterozygous' where more than one variant was called at that position.

Count The number of 'countable' fragments supporting the allele. The 'countable' fragments are those that are used by the variant caller when calling the variant. Which fragments are 'countable' depends on the user settings when the variant calling is performed - if e.g. the user has chosen 'Ignore broken pairs', reads belonging to broken pairs are not 'countable'. Note that, although overlapping paired reads have two reads in their overlap region, they only represent one fragment, and are counted only as one. (Please see the column 'Read count' below for a column that reports the value for 'reads' rather than for 'fragments').

Coverage The fragment coverage at this position. Only 'countable' fragments are considered (see under 'Count' above for an explanation of 'countable' fragments). Note that, although overlapping paired reads have two reads in their overlap region, they only represent one fragment, and overlapping paired reads contribute only 1 to the coverage. (Please see the column 'Read coverage' below for a column that reports the value for 'reads' rather than for 'fragments').

Frequency 'Count' divided by 'Coverage'.

Probability The contents of the Probability column (for Low frequency and Fixed Ploidy variant callers only) depend on the variant caller that produced and the type of variant:


1,564,920
1,564,940
1,564,960
1,564,980
Homo sapiens (hg19)
0.8-100-0 - mills guided locally realigned
0.5-100-0 - mills guided locally realigned
100-0 - mills guided locally realigned (1564283..1565532)
828 reads
0
15
188

0.8-100-0 - m...

Rows: 2 Table view: Genome

Chro...	Region	Type	Refe...	Allele	Refe...	Zygosity	Count	Cove...	Freq...	Prob...	Aver...	Read...	Rea...	# uni...	# un...	Base...	Hyp...
1	1564712	SNV	G	T	No	Hetero...	55	95	57.89	1.00	34.78	58	100	41	42	0.54	no
1	1564712	SNV	G	G	Yes	Hetero...	40	95	42.11	1.00	35.45	42	100	33	33		no

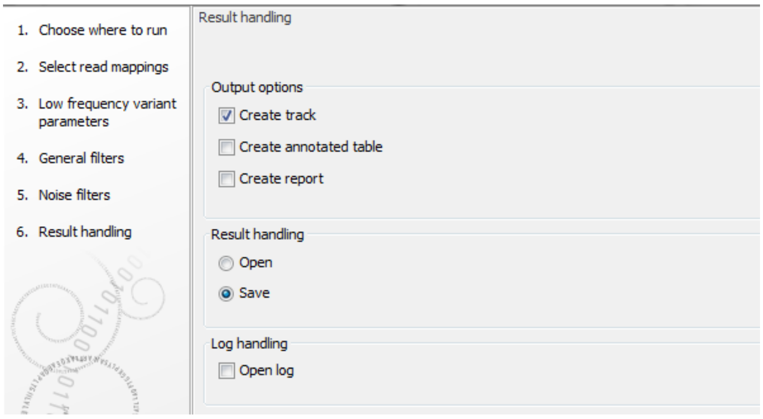
Create Track from Selection

0.5-100-0 - m...

Rows: 4 Table view: Genome

Chr...	Region	Type	Ref...	Allele	Ref...	Zygosity	Count	Cov...	Freq...	Prob...	Aver...	Rea...	Rea...	# u...	# un...	Bas...	Hyp...	Hom...
1	1564...	SNV	G	T	No	Hetero...	55	95	57.89	1.00	34.78	58	100	41	42	0.54	no	
1	1564...	SNV	G	G	Yes	Hetero...	40	95	42.11	1.00	35.45	42	100	33	33		no	
1	1564...	Deletion	G	-	No	Hetero...	39	70	55.71	1.00	34.41	39	70	35	35	0.00	no	Yes
1	1564...	SNV	G	G	Yes	Hetero...	31	70	44.29	1.00	34.90	31	70	28	28		no	Yes

Figure 20.56: *An example of a variant that is filtered out when the pyro-error filter is applied with settings 3 and 0.8, but not with settings 3 and 0.5.*


1. Choose where to run
2. Select read mappings
3. Low frequency variant parameters
4. General filters
5. Noise filters
6. Result handling

Result handling

Output options

☑ Create track

☐ Create annotated table

☐ Create report

Result handling

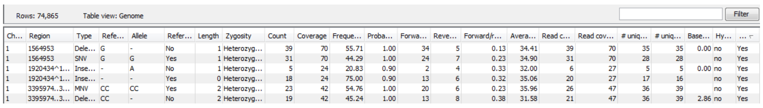
○ Open

● Save

Log handling

☐ Open log

Figure 20.57: *Output options.*


Rows: 74,865 Table view: Genome

Ch...	Region	Type	Refe...	Allele	Refer...	Length	Zygosity	Count	Coverage	Freque...	Proba...	Forwa...	Reve...	Forward/r...	Avera...	Read c...	Read cov...	# uniq...	# uniq...	Base...	Hy...	...
1	1564953	Dele...	G	-	No	1	Heterozyg...	39	70	55.71	1.00	34	5	0.13	34.41	39	70	35	35	0.00	no	Yes
1	1564953	SNV	G	G	Yes	1	Heterozyg...	31	70	44.29	1.00	24	7	0.23	34.90	31	70	28	28		no	Yes
1	1920434^1...	Inse...	-	A	No	1	Heterozyg...	5	24	20.83	0.90	2	4	0.33	32.00	6	27	5	5	0.00	no	Yes
1	1920434^1...	Inse...	-	-	Yes	0	Heterozyg...	18	24	75.00	0.90	13	6	0.32	35.06	20	27	17	16		no	Yes
1	3395974..3...	MNV	CC	CC	Yes	2	Heterozyg...	23	42	54.76	1.00	20	6	0.23	35.96	26	47	36	39		no	Yes
1	3395974..3...	Dele...	CC	-	No	2	Heterozyg...	19	42	45.24	1.00	13	8	0.38	31.58	21	47	36	39	2.86	no	Yes

Figure 20.58: *A variant track shown in the table view.*

- In the Fixed Ploidy Variant Detection Tool, the probability in the resulting variant track's 'Probability' column is NOT the probability referred to in the wizard. The probability referred to in the wizard, is the required minimum (posterior) probability that the site is NOT homozygous for the reference. The probability in the variant track 'Probability' column is the posterior probability of the particular site-type called. The fixed ploidy tool calculates the probability of the different possible configurations at each site. So using this tool, for single site variants the probability column just contains this quantity (for variants that span multiple positions see below).

- The rare variant tool makes statistical tests for the various possible explanations for each site. This means that the probability for the called variant must be estimated separately since it is not part of the actual variant calling. This is done by assigning prior probabilities to the various explanations for a site in a way that makes the probability for two explanations equal in exactly the situation where the statistical test shifts from preferring one explanation to the other. For a given single site variant, the probability is then calculated as the sum of probabilities for all the explanations containing that variant. So if a G variant is called, the reported probability is the sum of probabilities for these configurations: G, A/G, C/G, G/T, A/C/G, A/G/T, C/G/T, and A/C/G/T (and also all the configurations containing deletions together with G).

For multi position variants, an estimate is made of the probability of observing the same read data if the variant did not exist and all observations of the variant were due to sequencing errors. This is possible since a sequencing error model is found for both the fixed ploidy and rare variant tools. The probability column contains one minus this estimated probability. If this value is less than 50%, the variant might as well just be the result of sequencing errors and it is not reported at all.

Forward read count The number of 'countable' forward reads supporting the allele (see under 'Count' above for an explanation of 'countable' reads).

Reverse read count The number of 'countable' reverse reads supporting the allele (see under 'Count' above for an explanation of 'countable' reads).

Forward/reverse balance The minimum of the fraction of 'countable' forward reads and 'countable' reverse reads carrying the variant among all 'countable' reads carrying the variant (see under 'Count' above for an explanation of 'countable' reads).

Average quality The average base quality score of the bases supporting a variant. In the case of a deletion, the quality score is taken from the average quality of the two bases neighbouring the deleted one, and the lowest is reported. Similarly for insertions, the quality in reads where the insertion is absent, is taken from the minimum average of the two bases either side of the position. It can be possible in rare cases, that the quality score reported in this column for a deletion or insertion is below the threshold set for 'Minimum central quality', because this parameter is not applied to any quality value calculated from positions *outside* of the central variant. To remove low quality variants from the output, use the **Filter Marginal Variant Calls** tool (see section **??**).

Read count The number of 'countable' reads supporting the allele. Only 'countable' reads are considered (see under 'Count' above for an explanation of 'countable' reads). Note that each read in an overlapping pair contribute 1. To view the reads in pairs in a reads track as single reads, check the 'Disconnect paired reads' option in the side-panel of the reads track. (Please see the column 'Count' above for a column that reports the value for 'fragments' rather than for 'reads').

Read coverage The read coverage at this position. Only 'countable' reads are considered (see under 'Count' above for an explanation of 'countable' reads). Note that each read in an overlapping pair contribute 1. To view the reads in pairs in a reads track as single reads, check the 'Disconnect paired reads' option in the side-panel of the reads track. (Please see the column 'Coverage' above for a column that reports the value for 'fragments' rather than for 'reads').

# Unique start positions The number of unique start positions for 'countable' fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same start position, you could suspect that it is a result of an amplification error.

# Unique end positions The number of unique end positions for 'countable' fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same end position, you could suspect that it is a result of an amplification error.

BaseQRankSum The BaseQRankSum column contains an evaluation of the quality scores in the reads that has a called variant compared with the quality scores of the reference allele. Variants for which no corresponding reference allele is called does not have a BaseQRankSum value. Likewise, no values are calculated for reference alleles. The score is a Z score, so a value of 2.0 means that the observed qualities for the variant two standard deviations below the qualities for the reference allele. The scoring is performed using a Mann-Whitney U for comparing the two sets of quality scores from the reference allele and the variant.

Read position test probability The test probability for the test of whether the distribution of the read positions variant in the variant carrying reads is different from that of all the reads covering the variant position.

Read direction test probability The test probability for the test of whether the distribution among forward and reverse reads of the variant carrying reads is different from that of all the reads covering the variant position.

Hyper-allelic Basic and Fixed Ploidy Variant detectors only: Contains "yes", if the site contains more variants than the user-specified ploidy predicts, "no" if not.

Genotype Fixed Ploidy only: Contains the most probable genotype for the site.

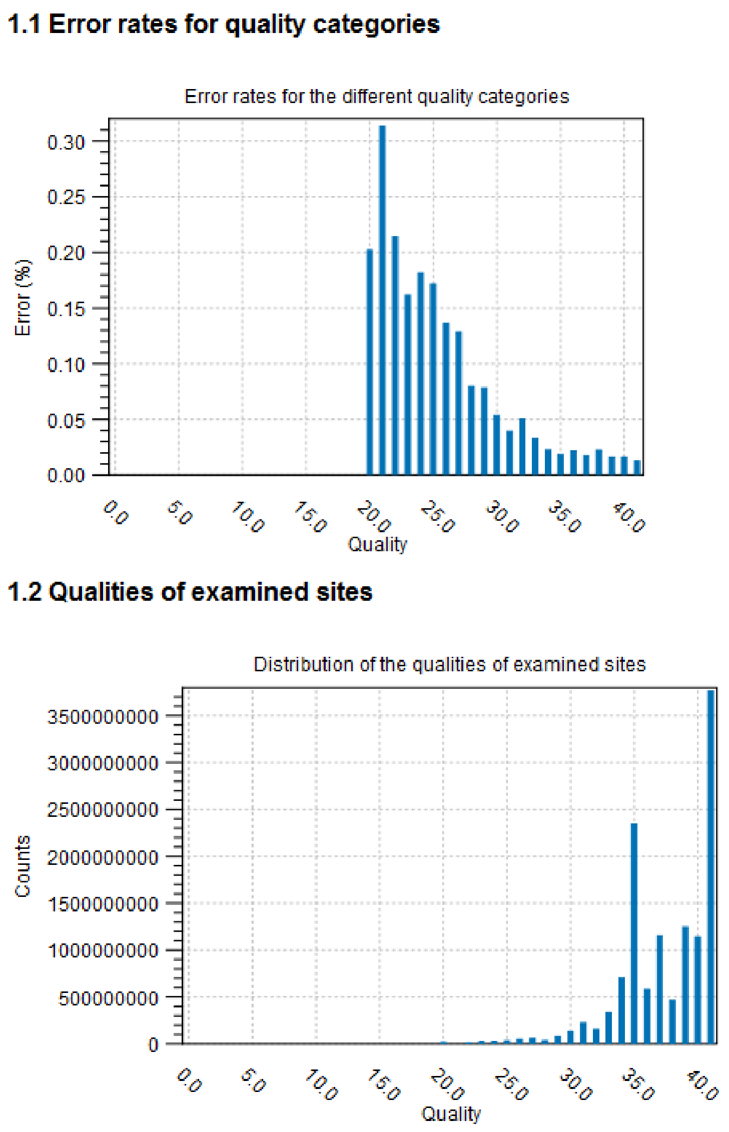
Homopolymer The column contains "Yes" if the variant is likely to be a homopolymer error and "No" if not. This is assessed by inspecting all variants in homopolymeric regions longer than 2. A variant will get the mark "yes" if it is a homopolymeric length variation of the reference allele, or a length variation of another variant that is a homopolymeric variation of the reference allele.

20.16.2 The annotated table output

The 'Annotated table' output contains an 'old' style variant format output.

20.16.3 The report

In addition to the estimated error rates of the different types of errors shown in figure 20.48, the report contains information on the total error rates for each quality score as well as a distribution of the qualities of the individual bases in the reads in the read mapping, at the sites that were examined for variants (see figure 20.59).



1.1 Error rates for quality categories

1.2 Qualities of examined sites

Figure 20.59: *Part of the contents of the report on the variant calling.*

20.17 InDels and Structural Variants

The InDels and Structural Variants tool is designed to identify structural variants such as insertions, deletions, inversions, translocations and tandem duplications in read mappings. The tool relies *exclusively* on information derived from unaligned ends (also called 'soft clippings') of the reads in the mappings. This means that:

- The tool will detect NO structural variants if there are NO reads with unaligned ends in the read mapping.
- Read mappings made with the CLC 'Map reads to reference' tool with the *'global'* option switched on will have NO unaligned ends and the Structural Variation tool will thus find NO structural variants on these. (The 'global' option means that reads are aligned in their entirety - irrespectively of whether that introduces mismatches towards the ends of the reads. In the 'local' option such reads will be mapped with unaligned ends).

- Read mappings based on really short reads (say, below 35 bp) are not likely to produce many reads with unaligned ends of any useful length, and the tool is thus not likely to produce many structural variant predictions for these read mappings.
- Read mappings generated with the Large Gap Read Mapper are NOT optimal for the detection of structural variants with this tool. This is due to the fact that, the Large Gap Read Mapper will map some reads with (large) gaps, that would be mapped with unaligned ends with standard read mappers, and thus will leave a weaker unaligned end signal in the mappings for the Structural Variation tool to work with.

In it's current version the InDels and Structural Variants tool has the following known limitations:

- It will only detect intra-chromosomal structural variants.

20.17.1 How to run the InDels and Structural Variants tool

To start the structural variant detection:

Toolbox | **Resequencing** ( ) | **InDels and Structural Variants tool** ( )

This will open up a dialog. Select the read mapping of interest as shown in figure 20.60 and click on the button labeled **Next**.

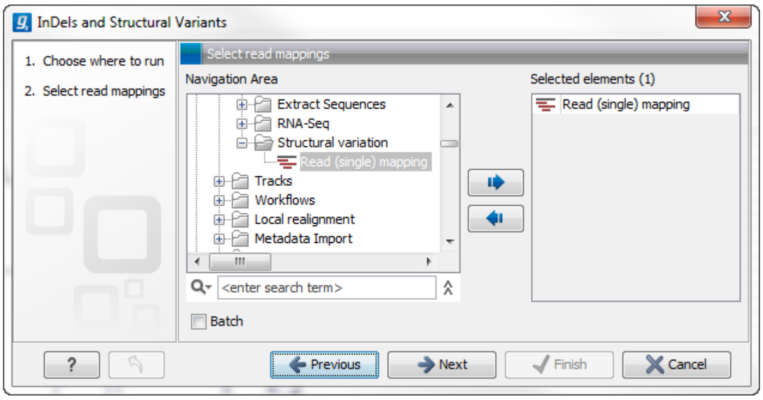
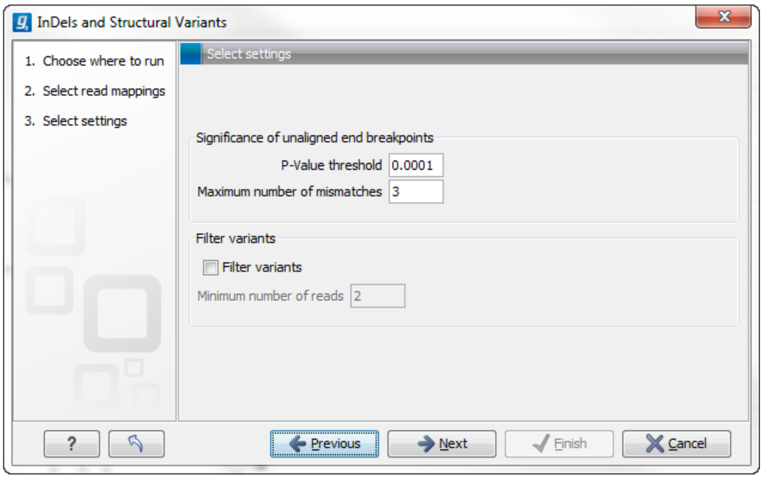


Figure 20.60: *Select the read mapping of interest.*

The next wizard step (Figure 20.61) is concerned with specifying parameters related to the algorithm used for calling structural variants. The algorithm first identifies positions in the mapping(s) with an *excess* of reads with left (or right) unaligned ends. Once these positions and the consensus sequences of the unaligned ends are determined, the algorithm maps the determined consensus sequences to the reference sequence around other positions with unaligned ends. If mappings are found that are in accordance with a 'signature' of a structural variant, a structural variant is called. For further details about the algorithm see Section 20.17.3.

The **'Significance of unaligned end breakpoints'** parameters are concerned with when a position with unaligned ends should be considered by the algorithm, and when it should be ignored:

- **P-value threshold**: Only positions in which the fraction of reads with unaligned ends is sufficiently high will be considered. The 'P-value threshold' determines the cut-off value in a Binomial Distribution for this fraction. The higher the P-value threshold is set, the more unaligned breakpoints will be identified.



InDels and Structural Variants

Significance of unaligned end breakpoints

P-Value threshold 0.0001

Maximum number of mismatches 3

Filter variants

□ Filter variants

Minimum number of reads 2

Figure 20.61: *Select the relevant settings.*

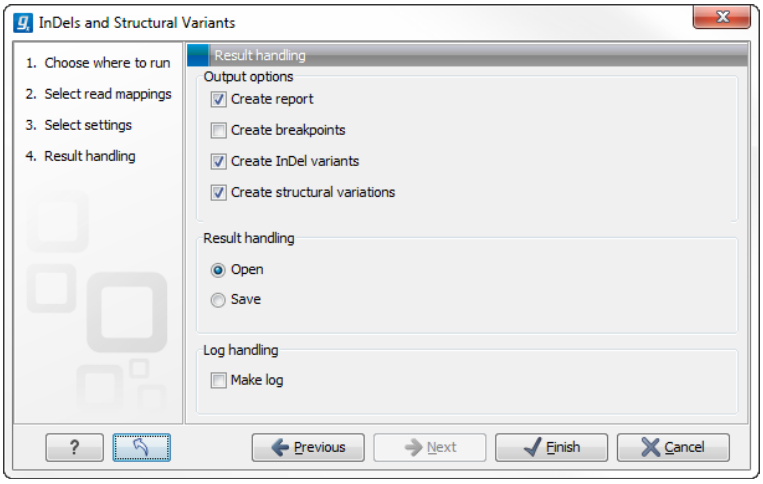
- **Maximum number of mismatches**: The 'Maximum number of mismatches' parameter determines which reads should be considered when inferring unaligned end breakpoints. Poorly map reads tend to have many mis-matches and unaligned ends, and it may be preferable to let the algorithm ignore reads with too many mis-matches in order to avoid false positives and reduce computational time. On the other hand, if the allowed number of mis-matches is set too low, unaligned end breakpoints in proximities of other variants (e.g. SNVs) may be lost. Again, the higher the number of mis-matches allowed, the more unaligned breakpoints will be identified.

The **'Filter variants'** parameters are concerned with the amount of evidence for each structural variant required for it to be called:

- **Filter variants**: When the **Filter variants** box is checked, only variants that are inferred by breakpoints that *together* are supported by at least the specified **Minimum number of reads** will be called.

Specify these settings and click **Next**. The "Results handling" dialog (Figure 20.62) will be opened. The Indels and Structural variants tool has the following output options:

- **Create report** When ticked, a report that summarizes information about the inferred breakpoints and variants is created.
- **Create breakpoints** When ticked, a track containing the detected breakpoints is created.
- **Create InDel variants** When ticked, a variant track containing the detected InDels that fulfill the requirements for being 'variants' is created. These include the detected insertions for which the allele sequence is inferred, but not those for which it is not, or only partly, known. Also, only deletions of six up to 200 bp are included in the variant track. See section 20.18.1 for a definition of the requirements for 'variants'. Note that insertions and deletions that are not included in the InDel track, will be present in the 'Structural variants track' (described below).
- **Create structural variations** When ticked, a track containing the detected structural variants is created.



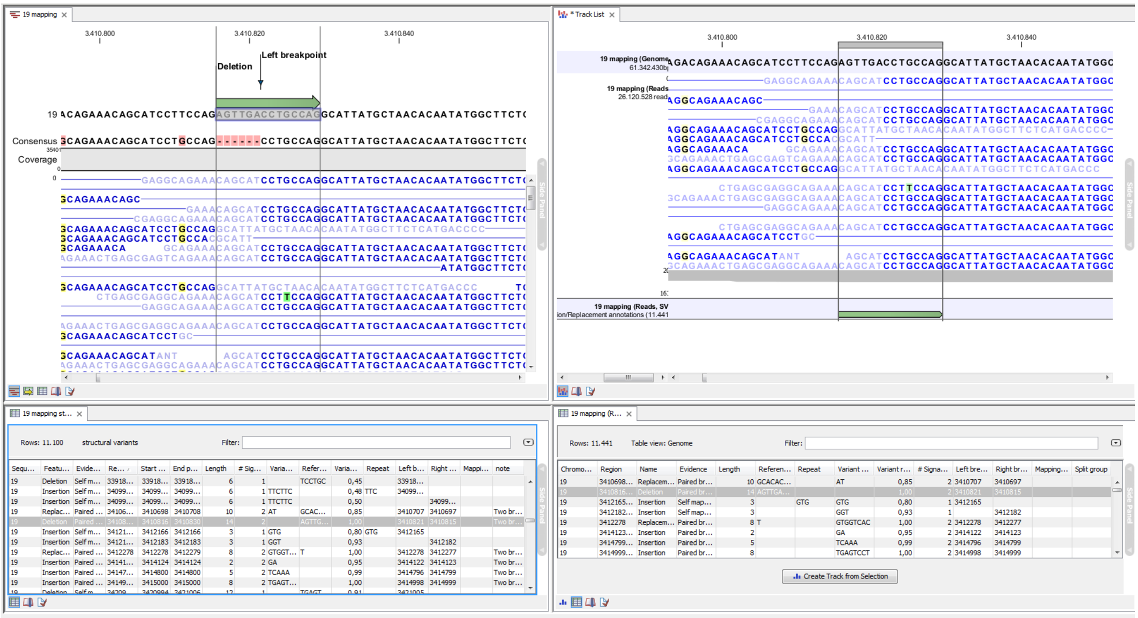
Output options
☑ Create report
□ Create breakpoints
☑ Create InDel variants
☑ Create structural variations

Result handling
● Open
○ Save

Log handling
□ Make log

Figure 20.62: *Select output formats.*

An example of the output from the InDel and Structural Variant tool is shown in Figure 20.63. The output is described in detail in the next section (Section 20.17.2).



19 mapping st...

Rows: 11.100 structural variants Filter:

Sequ...	Featu...	Evide...	Re...	Start ...	End p...	Length	# Sig...	Varia...	Refer...	Varia...	Repeat	Left b...	Right ...	Mappi...	note
19	Deletion	Self m...	33918...	33918...	33918...	6	1		TCCTGC	0,45		33918...			
19	Insertion	Self m...	34099...	34099...	34099...	6	1	TTCTTC		0,48	TTC	34099...			
19	Insertion	Self m...	34099...	34099...	34099...	6	1	TTCTTC		0,50			34099...		
19	Replac...	Paired ...	34106...	3410698	3410708	10	2	AT	GCAC...	0,85		3410707	3410697		Two br...
19	Deletion	Paired ...	34108...	3410816	3410830	14	2		AGTTG...	1,00		3410821	3410815		Two br...
19	Insertion	Self m...	34121...	3412166	3412166	3	1	GTG		0,80	GTG	3412165			
19	Insertion	Self m...	34121...	3412183	3412183	3	1	GGT		0,93			3412182		
19	Replac...	Paired ...	3412278	3412278	3412279	8	2	GTGGT...	T	1,00		3412278	3412277		Two br...
19	Insertion	Paired ...	34141...	3414124	3414124	2	2	GA		0,95		3414122	3414123		Two br...
19	Insertion	Paired ...	34147...	3414800	3414800	5	2	TCAAA		0,99		3414796	3414799		Two br...
19	Insertion	Paired ...	34149...	3415000	3415000	8	2	TGAGT...		1,00		3414998	3414999		Two br...
19	Deletion	Self m...	34209...	3420994	3421006	12	1		TGAGT...	0,91		3421005			

19 mapping (R...

Rows: 11.441 Table view: Genome Filter:

Chromo...	Region	Name	Evidence	Length	Referen...	Repeat	Variant ...	Variant r...	# Signa...	Left bre...	Right br...	Mapping...	Split group
19	3410698...	Replacem...	Paired br...	10	GCACAC...		AT	0,85	2	3410707	3410697		
19	3410816...	Deletion	Paired br...	14	AGTTGA...			1,00	2	3410821	3410815		
19	3412165...	Insertion	Self map...	3		GTG	GTG	0,80	1	3412165			
19	3412182...	Insertion	Self map...	3			GGT	0,93	1		3412182		
19	3412278	Replacem...	Paired br...	8	T		GTGGTCAC	1,00	2	3412278	3412277		
19	3414123...	Insertion	Paired br...	2			GA	0,95	2	3414122	3414123		
19	3414799...	Insertion	Paired br...	5			TCAAA	0,99	2	3414796	3414799		
19	3414999...	Insertion	Paired br...	8			TGAGTCCT	1,00	2	3414998	3414999		

Figure 20.63: *Example of the result of an analysis on a standalone read mapping (to the left) and on a reads track (to the right).*

20.17.2 The Structural Variants and InDels output

The report

The report gives an overview of the numbers and types of structural variants found in the sample. It contains

- A table with a row for each reference sequence, and information on the number of breakpoint signatures and structural variants found.
- A table giving the total number of left and right unaligned end breakpoint signatures found, and the total number of reads supporting them.

- A distribution of the logarithm of the sequence complexity of the unaligned ends of the left and right breakpoint signatures (see Section 20.17.7 for how the complexity is calculated).
- A distribution of the length of the unaligned ends of the left and right breakpoint signatures.
- A table giving the total number of the different types of structural variants found.
- Plots depicting the distribution of the lengths of structural variants identified.

The Breakpoints track (BP):

The breakpoints track contains a row for each called breakpoint with the following information:

- 'Chromosome': The chromosome on which the breakpoint is located.
- 'Region': The location on the chromosome of the breakpoint.
- 'Name': The type of the breakpoint ('left breakpoint' or 'right breakpoint').
- 'p-value': The p-value (in the Binomial distribution) of the unaligned end breakpoint.
- 'Unaligned': The consensus sequence of the unaligned ends at the breakpoint.
- 'Unaligned length': The length of the consensus sequence of the unaligned ends at the breakpoint.
- 'Mapped to self': If the unaligned end sequence at the breakpoint was found to map back to the reference in the vicinity of the breakpoint itself, a 'Deletion' or 'Insertion' based on 'self-mapping' evidence is called. This column will contain 'Deletion' or 'Insertion' if that is the case, or be empty if the unaligned end did not map back to the reference in the vicinity of the breakpoint itself.
- 'Perfect mapped': The number of 'perfect mapped' reads. This number is intended as a proxy for the number of reads that fit with the reference sequence. When calculating this number we consider all reads that extend across the breakpoint. We ignore reads that are non-specifically mapped, in a broken pair, or has more than the **maximum number of mismatches**. A read is perfectly mapped if (1) it has no insertions or deletions (mismatches are allowed) and (2) it has no unaligned end.
- 'Not perfect mapped': The number of 'not perfect mapped' reads. This number is intended as a proxy for the number of reads that fit with the predicted InDel. When calculating this number we consider all reads that extend across the breakpoint or that has an unaligned end starting at the breakpoint. We ignore reads that are non-specifically mapped, in a broken pair, or has more than the **maximum number of mismatches**. A read is not perfect mapped if (1) it has an insertion or deletion or (2) it has an unaligned end.
- 'Fraction non-perfectly mapped': the 'Non perfect mapped' divided by the 'Non perfect mapped' + 'Perfect mapped'.
- 'Sequence complexity': The sequence complexity of the unaligned end of the breakpoint (see Section 20.17.7 for how the sequence complexity is calculated).
- 'Reads': The number of reads supporting the breakpoint.

Note that typically, breakpoints will be found for which it is not possible to infer a structural variant. There may be a number of reasons for that: (1) the unaligned ends from which the breakpoint signature was derived might not be caused by an underlying structural variant, but merely be due to read mapping issues or noise, or (2) the breakpoint(s) which the detected breakpoint should have been matched to was/were not detected, and therefore no matching breakpoint(s) were found. Breakpoints may go un-detected either because of lack of coverage in the breakpoint region or because they are located within regions with exclusively non-uniquely mapped reads (only unaligned ends of uniquely mapping reads are used).

The InDel variants track (InDel):

The Indel variants track contains a row for each of the called InDels that *fulfills the requirements for being of a 'variant' type* (see Section 20.18 for a description of the 'variant type'). These are the small to medium sized insertions and deletions (up to 200 bp in length) for which the algorithm was able to identify the allele sequence (that is, the exact inserted sequence, or the exact deleted sequence). For insertions, the full allele sequence is found from the unaligned ends of mapped reads. For some insertions the length and allele sequence cannot be determined and as these do not fulfill the requirements of a 'variant', they do not qualify for representation in the 'InDel variant' track but instead appear in the Structural Variants track (see below). The information provided for each of the InDels in the InDel variant track is the 'Chromosome', 'Region', 'Type', 'Reference', 'Allele', 'Reference Allele', 'Length' and 'Zygosity' columns that are provided for all variants (see Section 20.18.1). In addition the following information, which is primarily intended to allow the user to assess the degree of evidence supporting each predicted InDel, is provided:

- 'Evidence': The mapping evidence on which the call of the InDel was based. This may be either 'Self mapped', 'Paired breakpoint', Cross mapped breakpoint' or 'Tandem duplication' depending of the mapping signature of the unaligned ends of the breakpoint(s) from which the InDel was inferred.

- 'Repeat': The algorithm attempts to identify if the variant sequence contains perfect repeats. This is done by searching the region around the structural variant for perfect repeat sequences. The region searched is 3 times the length of variant around the insertion/deletion point. The maximum repeat length searched for is 10. If a repeat sequence is found, the repeated sequence is given in this column. If not, the column is empty.

- 'Variant ratio': This column contains the sum of the 'Non perfect mapped' reads for the breakpoints used to infer the InDel, divided by the sum of the 'Non perfect mapped' and 'Perfect mapped' reads for the breakpoints used to infer the InDel (see Section the description above of the breakpoints track). This fraction is intended to give a hint towards the zygosity of the InDel. The closer the value to 1, the higher the likelihood that the variant is homozygous.

- '# Reads': The total number of reads supporting the breakpoints from which the InDel was constructed.

- 'Sequence complexity': The sequence complexity of the unaligned end of the breakpoint (see Section 20.17.7). InDels with higher complexity are typically more reliable than those with low complexity.

The 'Zygosity' field is set to 'Homozygous' if the 'Variant ratio' is 0.80 or above, and 'Heterozygous' otherwise.

The Structural variants track (SV):

The Structural variants track contains a row for each of the called Structural variants that is not already reported in the InDel track. It contains the following information:

- 'Chromosome': The chromosome on which the structural variant is located.
- 'Region': The location on the chromosome of the structural variant.
- 'Name': The type of the structural variant ('deletion', 'insertion', 'inversion', 'replacement', 'translocation' or 'complex').
- 'Evidence': The breakpoint mapping evidence ('that is, the 'unaligned end 'signature') on which the call of the structural variant was based. This may be either 'Self mapped', 'Paired breakpoint', 'Cross mapped breakpoints', 'Cross mapped breakpoints (invalid orientation)', 'Close breakpoints', 'Multiple breakpoints' or 'Tandem duplication', depending on which type of signature that was found.
- 'Length': the length of the allele sequence of the structural variant. Note that the length of variants for which the allele sequence could not be determined is reported as 0 (e.g insertions inferred from 'Close breakpoints').
- 'Reference sequence': The sequence of the reference in the region of the structural variant.
- 'Variant sequence': The allele sequence of the structural variant if it is known. If not, the column will be empty.
- 'Repeat': The same as in the InDel track.
- 'Variant ratio': The same as in the InDel track.
- 'Signatures': The number of unaligned breakpoints involved in the signature of the structural variant. In most cases these will be pairs of breakpoints, and the value is 2, however some structural variants that have signatures involving more than two breakpoint (See Section 20.17.6). Typically structural variants of type 'complex' will be inferred from more than 2 breakpoint signatures.
- 'Left breakpoints': The positions of the 'Left breakpoints' involved in the signature of the structural variant.
- 'Right breakpoints': The positions of the 'Right breakpoints' involved in the signature of the structural variant.
- 'Mapping scores fraction': The mapping scores of the unaligned ends for each of the breakpoints. These are the similarity values between the unaligned end and the region of the reference to which it was mapped. The values lie between 0 and 1. The closer the value is to 1, the better the match, suggesting better reliability of the inferred variant.
- 'Reads': The total number of reads supporting the breakpoints from which the InDels was constructed.

- ’Sequence complexity’: The sequence complexity of the unaligned end of the breakpoint (see Section 20.17.7).
- ’Split group’: Some structural variants extend over a very large a region. For these visualization is challenging, and instead of reporting them in a single row we split them in multiple rows - one for each ’end’ of the variant. To allow the user to see which of these ’split features’ belong together, we give features that belong to the same structural variant a common ’split group’ identifier. If the column is empty the structural variant is not split, but contained within a single row.

20.17.3 The InDels and Structural Variants detection algorithm

The Indels and Structural Variants detection algorithm has two steps:

1. Identify ’breakpoint signatures’: First, the algorithm identifies positions in the mapping(s) with an *excess* of reads with left (or right) unaligned ends. For each of these, it creates a Left breakpoint (LB) or Right breakpoint (RB) signature.
2. Identify ’structural variant signatures’: Secondly, the algorithm creates structural variant signatures from the identified breakpoint signatures. This is done by mapping the consensus unaligned ends of the identified LB and RB signatures to selected areas of the references as well as to each other. The mapping patterns of the consensus unaligned ends are examined and structural variant annotations consistent with the mapping patterns are created.

The two steps of the algorithm are described in detail in sections 20.17.4 and 20.17.5.

20.17.4 The InDels and Structural Variants detection algorithm - Step 1: Creating Left- and Right breakpoint signatures

In the first step of the InDels and Structural Variants detection algorithm points in the read mapping are identified which have a significant proportion of reads mapped with unaligned ends. There are typically numerous reads with unaligned ends in read mappings — some are due to structural variants in the sample relative to the reference, others are due to poorly mapped, or poor quality reads. An example is given in figure 20.64. In order to make reliable predictions, attempts must be made to distinguish the unaligned ends caused by noisy read(mappings) from those caused by structural variants, so that the signal from the structural variants comes through as clearly as possible — both in terms of where the ’significant’ unaligned ends are and in terms of what they look like.

To identify positions with a ’significant’ portion of ’consistent’ unaligned end reads we first estimate ’null-distributions’ of the fractions of left and right unaligned end reads at each position in the read mapping, and subsequently use these distributions to identify positions with an ’excess’ of unaligned end reads. In these positions we create a Left (LB) or Right (RB) breakpoint signature. To estimate the null-distributions we:

1. Calculate the coverage, c_i, in each position, i of all uniquely mapped reads (Non-specifically mapped reads are ignored. Furthermore, for paired read data sets, only intact paired reads pairs are considered — broken paired reads are ignored).

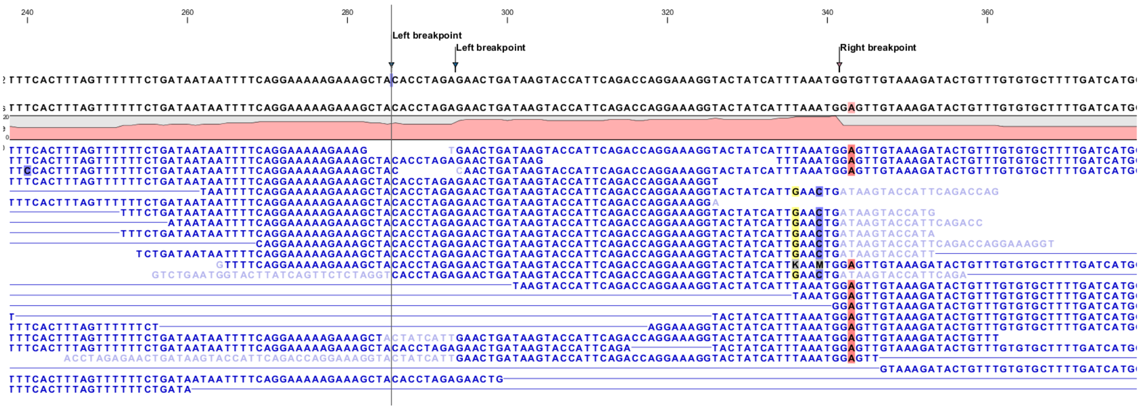


Figure 20.64: *Example of a read mapping containing unaligned ends with three unaligned end signatures.*

2. Calculate the coverage in each position of 'valid' reads with a starting left unaligned end, l_i (of minimum consensus length 3bp).

3. Calculate the coverage in each position of 'valid' reads with a starting right unaligned end, r_i (of minimum consensus length 3bp).

We then use the observed fractions of 'Left unaligned ends' ($\sum_i l_i / \sum_i c_i$) and 'Right unaligned ends' ($\sum_i r_i / \sum_i c_i$) as frequencies in binomial distributions of 'Left unaligned end' and 'Right unaligned end' read fractions. We go through each position in the read mapping and examine it for an excess of left (or right) unaligned end reads: if the probability of obtaining the observed number of left (or right) unaligned ends in a position with the observed coverage, is 'small', a Left breakpoint signature (LB), respectively Right breakpoint signature (RB), is created.

The two user-specified settings 'The P-value threshold' and the 'Maximum number of mismatches' determine which breakpoint signatures the algorithm will detect (see Section 20.17.1 and Figure 20.61). The p-value is used as a cutoff in the binomial distributions estimated above: if the probability of obtaining the observed number of left (or right) unaligned ends in a position with the observed coverage, is smaller than the user-specified cut-off, a Left breakpoint signature (LB), respectively Right breakpoint signature (RB), is created. The 'Maximum number of mis-matches' parameter is used to determine which reads are considered 'valid' unaligned end reads. Only reads that have at most this number of mis-matches in their aligned parts are counted. The higher these two values are set, the more breakpoints will be called. The more breakpoints are called, the larger the search space for the Structural variation detection algorithm, and thus the longer the computation time.

In figure 20.64, three unaligned end signatures are shown. The left-most LB signature is called only when the p-value cut-off is chosen high (0.01 as opposed to 0.0001).

20.17.5 The InDels and Structural Variants detection algorithm - Step 2: Creating Structural variant signatures

In the second step of the InDels and Structural Variants detection algorithm the unaligned end 'breakpoint signatures' (identified in step 1) are used to derive 'structural variant signatures'. This is done by:

1. Generating a consensus sequence of the reads with unaligned ends at each identified

breakpoint.

2. Mapping the generated consensus sequences against the reference sequence in the regions around *other* identified breakpoints ('cross-mapping').
3. Mapping the generated consensus sequences of breakpoints that are *near each other against each other* ('aligning').
4. Mapping the generated consensus sequences against the reference sequence in the *region around the breakpoint itself* ('self-mapping').
5. Considering the breakpoints whose unaligned end consensus sequences are found to cross map against each other together, and compare their mapping patterns to the set of theoretically expected 'structural variants signatures' (See Section 20.17.6).
6. Creating a 'structural variant signature' for each of the groups of breakpoints whose mapping patterns were in accordance with one of the expected 'structural variants signatures'.

A structural variant is called for each of the created 'structural variant signatures'. For each of the groups of breakpoints whose mapping patterns were NOT in accordance with one of the expected 'structural variants signatures', we call a structural variant of type 'complex'.

The steps above require a number of decisions to be made regarding (1) When is the consensus sequence reliable enough to work with?, and (2) When does an unaligned end map well enough that we will call it a match? The algorithm uses a number of hard-coded values when making those decisions. The values are described below.

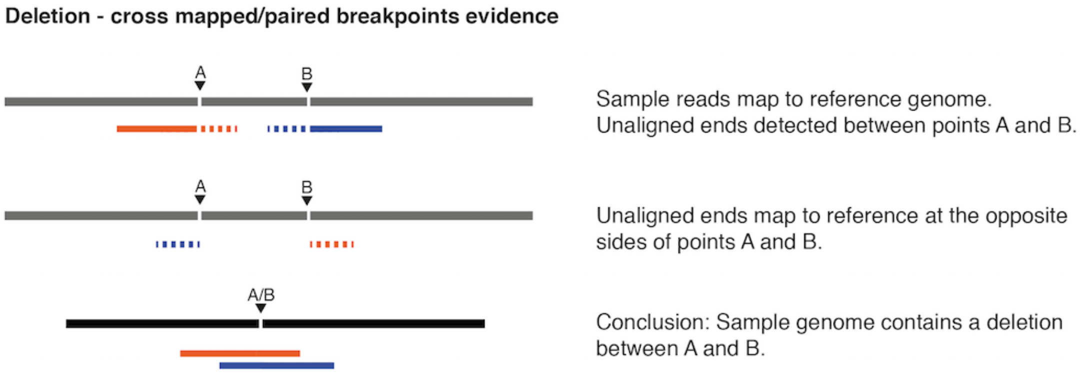
Algorithmic details

- **Generating a consensus:** The consensus of the unaligned ends is calculated by simple alignment without gaps. Having created the consensus, we exclude the unaligned ends which differ by more than 20% from the consensus, and recalculate the consensus. This prevents 'spuriously' unaligned ends that extend longer than other unaligned ends from impacting the tail of the consensus unaligned end.
- **Mapping of the consensus:**
  - **'Cross mapping':** When mapping the consensus sequences against the reference sequence around other breakpoints we require that:
    - The consensus is at least 16 bp long.
    - The score of the alignment is at least 70% of the maximal possible score of the alignment.
  - **'Aligning':** When aligning the consensus sequences two closely located breakpoints against each other we require that:
    - The breakpoints are within a 100 bp distance of each other.
    - The overlap in the alignment of the consensus sequences is least 4 nucleotides long.
  - **'Self-mapping':** When mapping the consensus sequences of breakpoints against the reference sequence in a region around the breakpoint itself we require that:
    - The consensus is at least 9 bp long.

- ∗ A match is found within 400 bp window of the breakpoint.
- ∗ The score of the alignment is at least 90% of the maximal possible score of the alignment of the part of the consensus sequence that does not include the variant allele part.

20.17.6 Theoretically expected structural variant signatures

Different types of structural variants will leave different 'signatures' in terms of the mapping patterns of the unaligned ends. The 'structural variant signatures' of the set of structural variants that are considered by the Indel and Structural variant tool are drawn in Figures 20.65, 20.66, 20.67, 20.68, 20.69, 20.70, 20.71, 20.72 and 20.73.

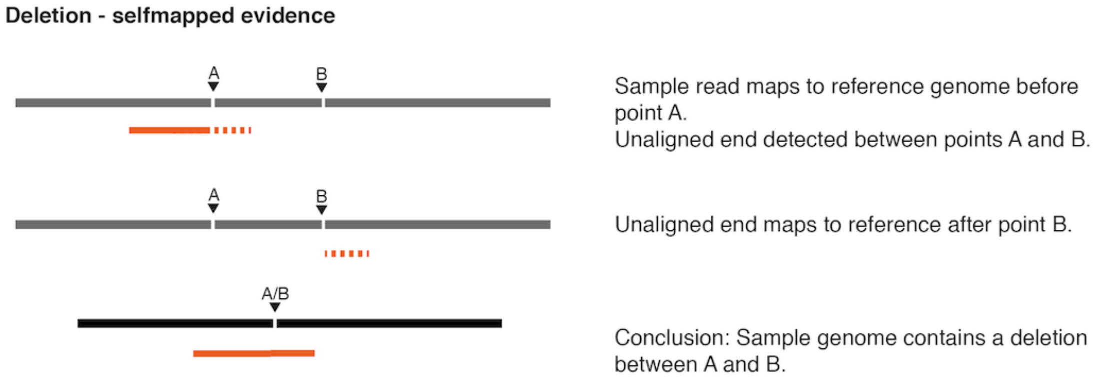

Deletion - cross mapped/paired breakpoints evidence

Sample reads map to reference genome.
Unaligned ends detected between points A and B.

Unaligned ends map to reference at the opposite sides of points A and B.

Conclusion: Sample genome contains a deletion between A and B.

Figure 20.65: *A deletion with cross-mapping breakpoint evidence.*


Deletion - selfmapped evidence

Sample read maps to reference genome before point A.
Unaligned end detected between points A and B.

Unaligned end maps to reference after point B.

Conclusion: Sample genome contains a deletion between A and B.

Figure 20.66: *A deletion with selfmapping breakpoint evidence.*

20.17.7 How sequence complexity is calculated

The sequence complexity of an unaligned end is calculated as the product of 'the observed vocabulary-usages' divided by 'the maximal possible vocabulary-usages', for word sizes from one to seven. When multiple breakpoints are used to construct a structural variant, the complexity is calculated as the product of the individual sequence complexities of the breakpoints constituting the structural variant.

The observed vocabulary usage for word size, k, for a given sequence is the number of different "words" of size k that exist in that sequence. The maximal possible vocabulary usage for word

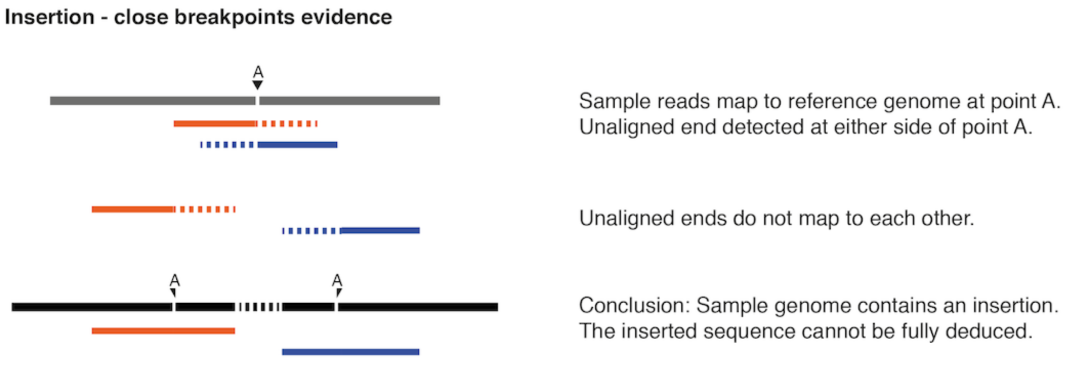

Insertion - close breakpoints evidence
A
Sample reads map to reference genome at point A.
Unaligned end detected at either side of point A.
Unaligned ends do not map to each other.
A
A
Conclusion: Sample genome contains an insertion.
The inserted sequence cannot be fully deduced.

Figure 20.67: *An insertion with close breakpoint evidence.*

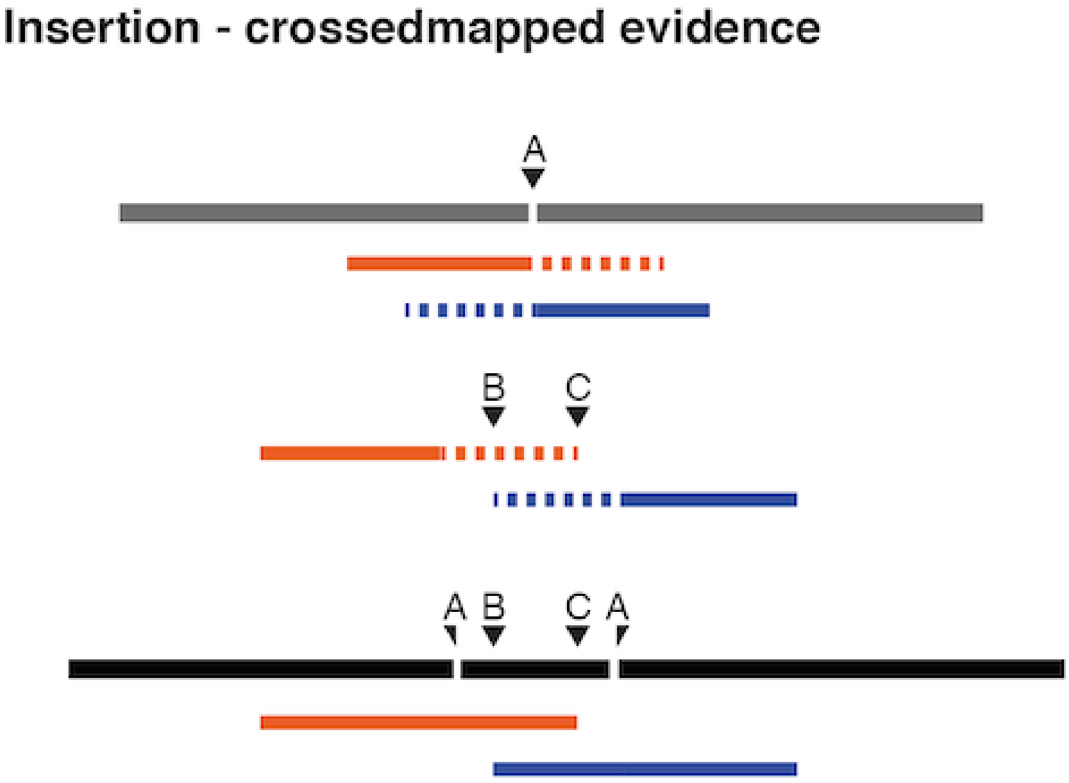


Figure 20.68: *An insertion with cross-mapped breakpoints evidence.*

size k for a given sequence is the maximal number of different words of size k that can possibly be observed in a sequence of a given length. For DNA sequences, the set of all possible letters in such words is four, that is, there are four letters that represent the possible nucleotides: A, C, G and T. The calculation is most easily described using an example.

Consider the sequence CAGTACAG. In this sequence we observe:

- 4 different words of size 1 ('A,', 'C', 'G' and 'T').
- 5 different words of size 2 ('CA', 'AG', 'GT', 'TA' and 'AC') Note that 'CA' and 'AG' are found

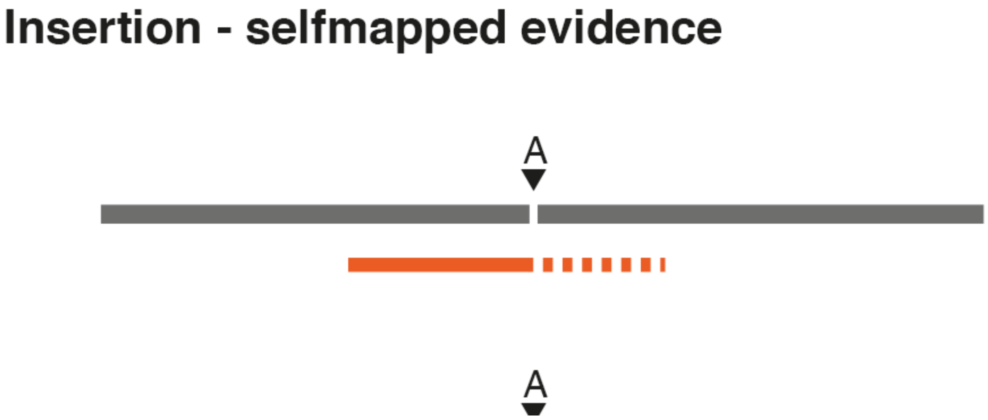
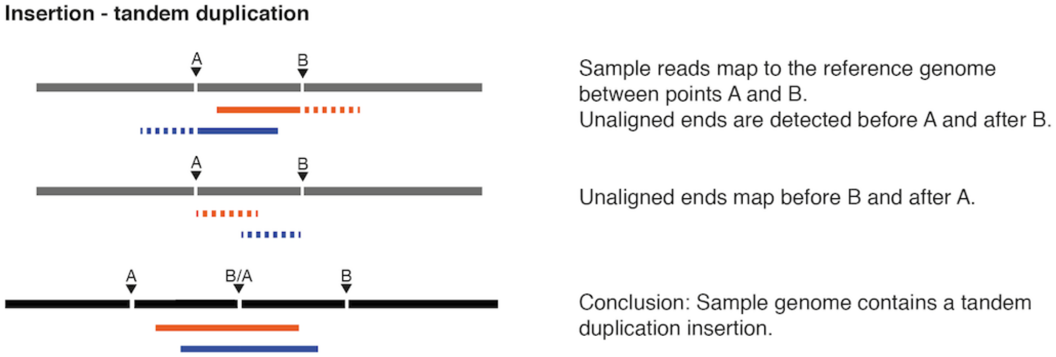


Figure 20.69: *An insertion with selfmapped breakpoint evidence.*



Sample reads map to the reference genome between points A and B.
Unaligned ends are detected before A and after B.

Unaligned ends map before B and after A.

Conclusion: Sample genome contains a tandem duplication insertion.

Figure 20.70: *An insertion with breakpoint mapping evidence corresponding to a 'Tandem duplication'.*

twice in this sequence.

- 5 different words of size 3 ('CAG', 'AGT', 'GTA', 'TAC' and 'ACA') Note that 'CAG' is found twice in this sequence.
- 5 different words of size 4 ('CAGT', 'AGTA', 'GTAC', 'TACA' and 'ACAG')
- 4 different words of size 5 ('CAGTA', 'AGTAC' , 'GTACA' and 'TACAG')
- 3 different words of size 6 ('CAGTAC', 'AGTACA' and 'GTACAG')
- 2 different words of of size 7 ('CAGTACA' and 'AGTACAG')

Note that we only do the calculations for word sizes up to 7, even when the unaligned end is longer than this.

Now we consider the maximal possible number of words we could observe in a DNA sequence of this length, again restricting our considerations to word lengths of 7.

- Word size of 1: The maximum number of different letters possible here is 4, the single characters, A, G, C and T. There are 8 positions in our example sequence, but there are only 4 possible unique nucleotides.

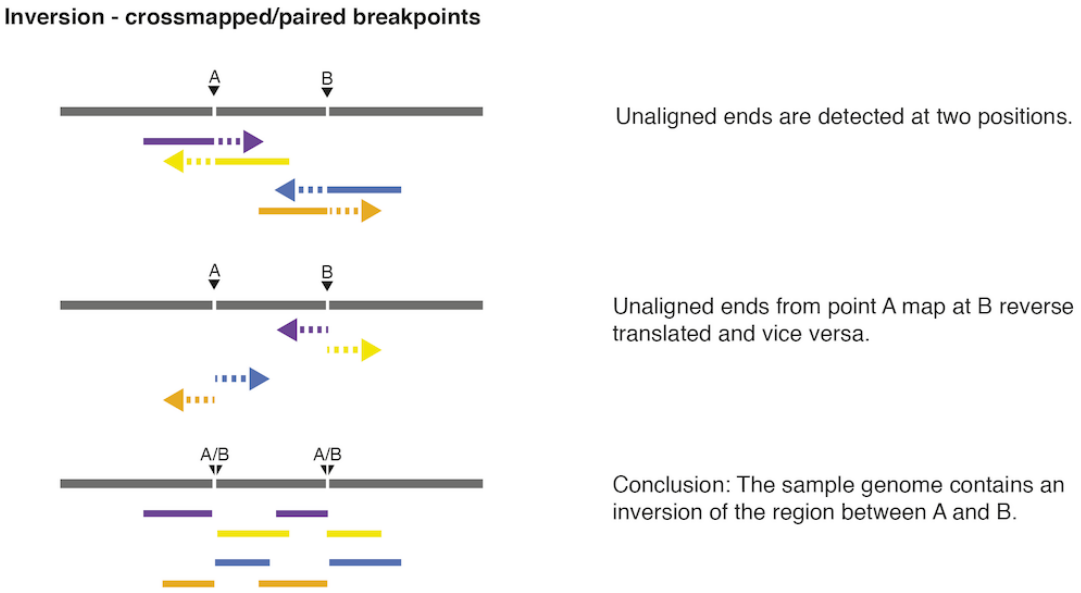

Inversion - crossmapped/paired breakpoints
A
B
Unaligned ends are detected at two positions.
A
B
Unaligned ends from point A map at B reverse translated and vice versa.
A/B
A/B
Conclusion: The sample genome contains an inversion of the region between A and B.

Figure 20.71: *The unaligned end mapping pattern of an inversion.*

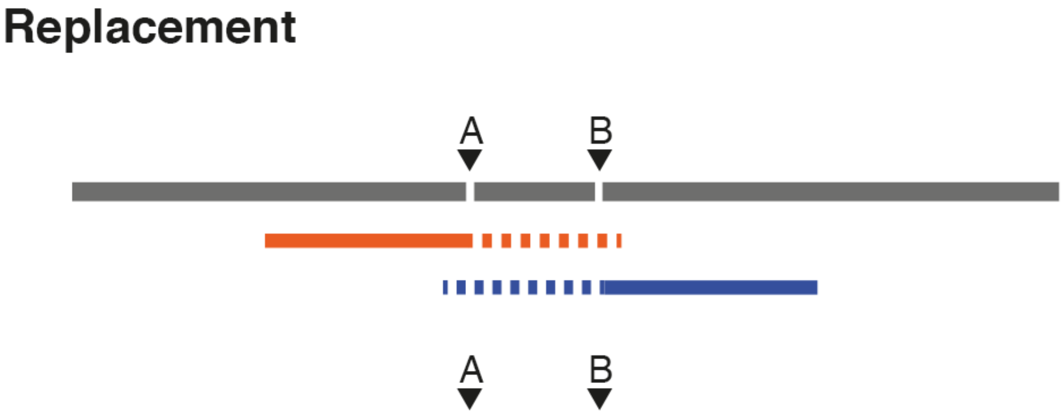


Figure 20.72: *The unaligned end mapping pattern of a replacement.*

- Word size of 2: The maximum number of different words possible here is 7. For DNA generally, there is a total of 16 different dinucleotides (4*4). For a sequence of length 8, we can have a total of 7 dinucleotides, so with 16 possibilities, the dinucleotides at each of our 7 positions could be unique.
- Word size of 3: The maximum number of different words possible here is 6. For DNA generally, there is a total of 64 different dinucleotides (4*4*4). For a sequence of length 8, we can have a total of 6 trinucleotides, so with 64 possibilities, the trinucleotides at each of our 6 positions could be unique.
- Word size of 4: The maximum number of different words possible here is 5. For DNA generally, there is a total of 256 different dinucleotides (4*4*4*4). For a sequence of length 8, we can have a total of 5 quatronucleotides, so with 256 possibilities, the quatronucleotides at each of our 5 positions could be unique.

We then continue, using the logic above, to calculate a maximum possible number of words for

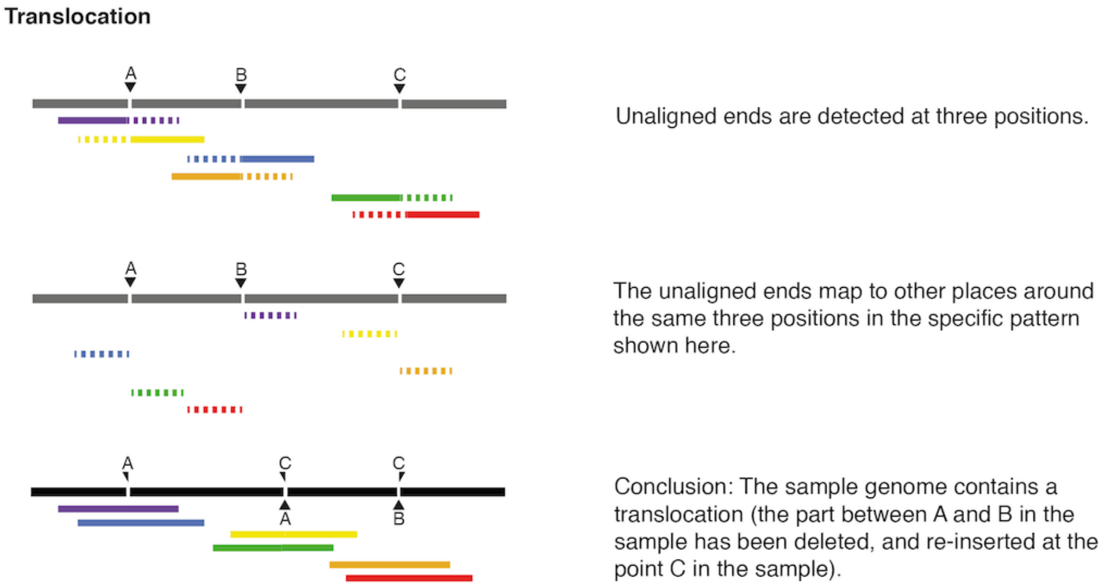

Translocation
A
B
C
Unaligned ends are detected at three positions.
A
B
C
The unaligned ends map to other places around the same three positions in the specific pattern shown here.
A
C
C
A
B
Conclusion: The sample genome contains a translocation (the part between A and B in the sample has been deleted, and re-inserted at the point C in the sample).

Figure 20.73: *The unaligned end mapping pattern of a translocation.*

a word size of 5 being 4, a maximum possible number of words for a word size of 6 being 3, and a maximum possible number of words for a word size of 7 being 2.

Now we can compute the complexity for this 7 nucleotide sequence by taking the number of different words we observe for each word length from 1 to 7 nucleotides and dividing them by the maximum possible number of words for each word length from 1 to 7. Here that gives us:

(4/4)(5/7)(5/6)(5/5)(4/4)(3/3)(2/2) = 0.595

As an extreme example of a sequence of low complexity, consider the 7 base sequence AAAAAAA. Here, we would get the complexity:

(1/4)(1/6)(1/5)(1/4)(1/3)(1/2)(1/1) = 0.000347

20.18 Variant data

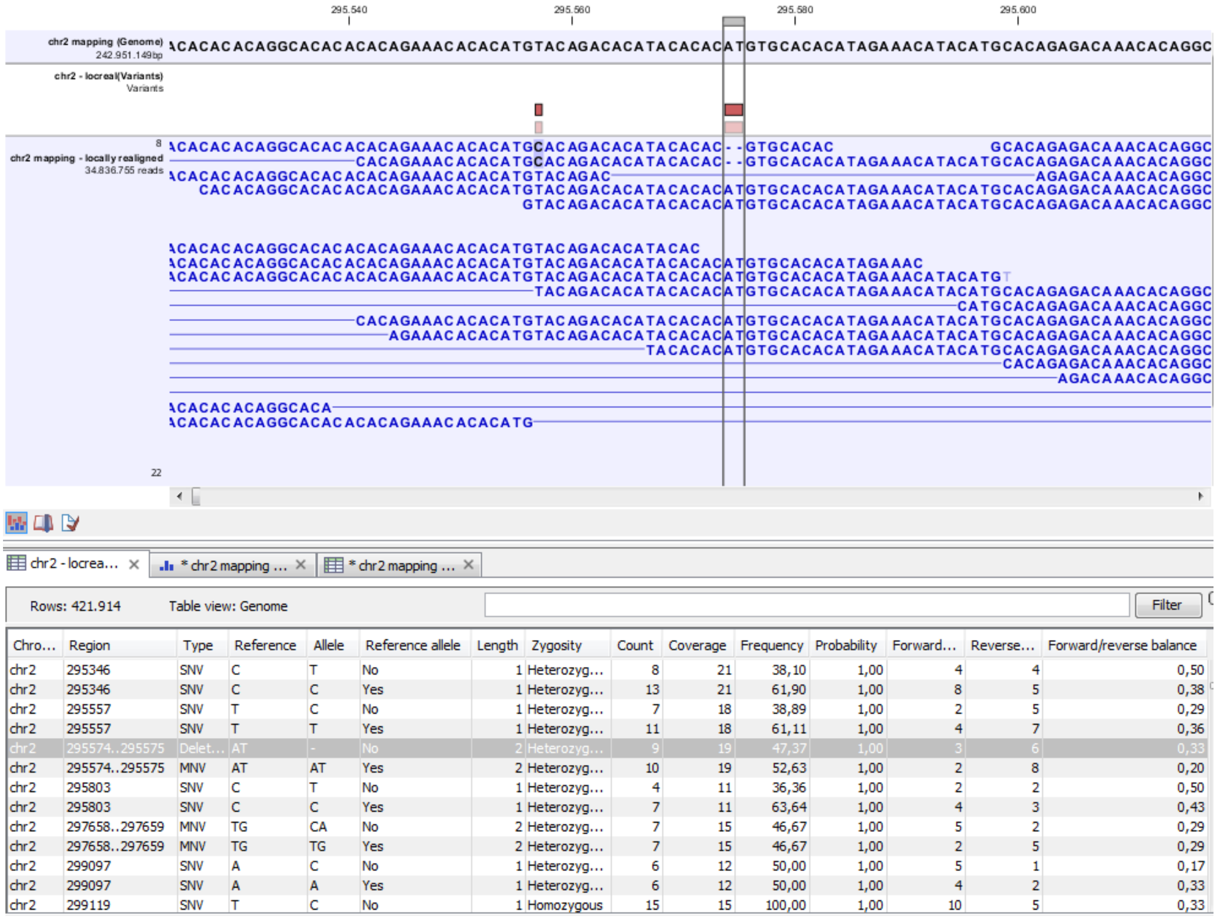
Variant data may be obtained either by importing variants from files (e.g. gvf or vcf files - as described in section 6.2), by downloading variants from external databases (e.g. dbSNP, HapMap, 1000genomes or COSMIC - (described in section 6.2)) or by calling variants on read tracks or read mappings using the CLC Basic Variant Detection (section 20.11), Fixed Ploidy Variant Detection (section 20.12), or the Low Frequency Variant Detection (section 20.13) tools.

Variant types include SNVs, MNVs, insertions, deletions or replacements. They may be presented either in a variant track (see figure 20.74) or in an annotated variant table (see figure 20.77).

20.18.1 Variant tracks

A variant track (figure 20.74), created with the *CLC Cancer Research Workbench* variant callers (see section 20.10), has the following information for each variant:

Chromosome The name of the reference sequence on which the variant is located.



chr2 - locrea... | * chr2 mapping ... | * chr2 mapping ...

Rows: 421.914 Table view: Genome Filter

Chro...	Region	Type	Reference	Allele	Reference allele	Length	Zygosity	Count	Coverage	Frequency	Probability	Forward...	Reverse...	Forward/reverse balance
chr2	295346	SNV	C	T	No	1	Heterozyg...	8	21	38,10	1,00	4	4	0,50
chr2	295346	SNV	C	C	Yes	1	Heterozyg...	13	21	61,90	1,00	8	5	0,38
chr2	295557	SNV	T	C	No	1	Heterozyg...	7	18	38,89	1,00	2	5	0,29
chr2	295557	SNV	T	T	Yes	1	Heterozyg...	11	18	61,11	1,00	4	7	0,36
chr2	295574..295575	Delet...	AT	-	No	2	Heterozyg...	9	19	47,37	1,00	3	6	0,33
chr2	295574..295575	MNV	AT	AT	Yes	2	Heterozyg...	10	19	52,63	1,00	2	8	0,20
chr2	295803	SNV	C	T	No	1	Heterozyg...	4	11	36,36	1,00	2	2	0,50
chr2	295803	SNV	C	C	Yes	1	Heterozyg...	7	11	63,64	1,00	4	3	0,43
chr2	297658..297659	MNV	TG	CA	No	2	Heterozyg...	7	15	46,67	1,00	5	2	0,29
chr2	297658..297659	MNV	TG	TG	Yes	2	Heterozyg...	7	15	46,67	1,00	2	5	0,29
chr2	299097	SNV	A	C	No	1	Heterozyg...	6	12	50,00	1,00	5	1	0,17
chr2	299097	SNV	A	A	Yes	1	Heterozyg...	6	12	50,00	1,00	4	2	0,33
chr2	299119	SNV	T	C	No	1	Homozygous	15	15	100,00	1,00	10	5	0,33

Figure 20.74: *Variant track. The figure shows a track list (top), consisting of a reference sequence track, a variant track and a read mapping. The variant track was produced by running the Probabilistic Variant Caller on the read track. The variant track has been opened in a separate table view by double-clicking on it in the track list. By selecting a row in the variant track table, the track list view is centered on the corresponding variant.*

Region The region on the reference sequence at which the variant is located. The region may be either a 'single position', a 'region' or a 'between position region'. Examples are given in figure 20.75. An extract of a gvf-file giving rise to these three variants after import is shown in figure 20.76.

Variant type The type of variant. This can either be SNV (single-nucleotide variant), MNV (multi-nucleotide variant), insertion, deletion, or replacement. Learn more in section 20.18.3.

Reference The reference sequence at the position of the variant.

Allele The allele sequence of the variant.

Reference allele Describes whether the variant is identical to the reference. This will be the case one of the alleles for most, but not all, detected heterozygous variants (e.g. the variant caller might detect two variants, A and G, at a given position in which the reference is 'A'. In this case the variant corresponding to allele 'A' will have 'Yes' in the 'reference allele' column entry, and the variant corresponding to allele 'G' would have 'No'. Had the variant caller called the two variants 'C' and 'G' at the position, both would have had 'No' in the

'Reference allele' column).

Length The length of the variant. The length is 1 for SNVs, and for MNVs it is the number of allele or reference bases (which will always be the same). For deletions, it is the length of the deleted sequence, and for insertions it is the length of the inserted sequence. For replacements, both the length of the replaced reference sequence and the length of the inserted sequence are considered, and the longest of those two is reported.

Zygosity The zygosity of the variant called, as determined by the variant caller. This will be either 'Homozygous', where there is only one variant called at that position or 'Heterozygous' where more than one variant was called at that position.

Count The number of 'countable' reads supporting the allele. The 'countable' reads are those that are used by the variant caller when calling the variant. Which reads are 'countable' depends on the user settings when the variant calling is performed - if e.g. the user has chosen 'Ignore broken pairs', reads belonging to broken pairs are not 'countable'.

Coverage The read coverage at this position. Only 'countable' reads are considered (see under 'Count' above for an explanation of 'countable' reads. Also see overlapping pairs in section 20.19 for how overlapping paired reads are treated.)

Frequency The number of 'countable' reads supporting the allele divided by the number of 'countable' reads covering the position of the variant ('see under 'Count' above for an explanation of 'countable' reads).

Probability The probability that this particular variant exists in the sample. (For further information please refer to the White paper on Probabilistic Variant Caller: `http://www.clcbio.com/files/whitepapers/whitepaper-probabilistic-variant-caller-1.pdf`).

Forward read count The number of 'countable' forward reads supporting the allele (see under 'Count' above for an explanation of 'countable' reads). Also see more information about overlapping pairs in section 20.19.

Reverse read count The number of 'countable' reverse reads supporting the allele (see under 'Count' above for an explanation of 'countable' reads). Also see more information about overlapping pairs in section 20.19.

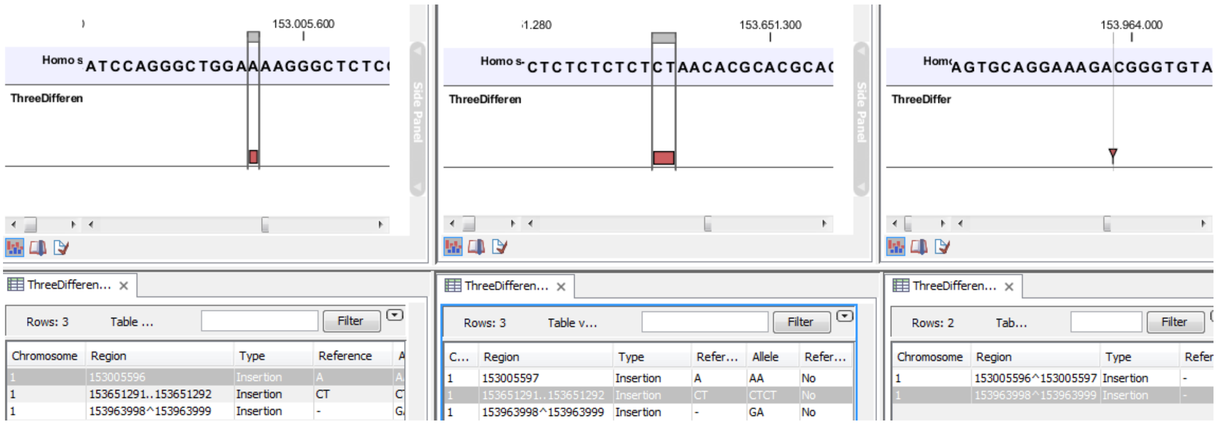
Forward/reverse balance The minimum of the fraction of 'countable' forward reads and 'countable' reverse reads carrying the variant among all 'countable' reads carrying the variant (see under 'Count' above for an explanation of 'countable' reads).[1]

Average quality The average read quality score of the bases supporting a variant.If there are no values in this column, it is probably because the sequencing data was imported without quality scores (learn more about importing quality scores from different sequencing platforms in section 6.3). For deletions, the quality scores of the two surrounding bases are taken into account, and the lowest value of these two is reported.

[1] Some systematic sequencing errors can be triggered by a certain combination of bases. This means that sequencing one strand may lead to sequencing errors that are not seen when sequencing the other strand (see [Nguyen et al., 2011] for a recent study with Illumina data). In order to evaluate whether the distribution of forward and reverse reads is approximately random, this value is calculated as the minimum of the number of forward reads divided by the total number of reads and the number of reverse reads divided by the total number of reads supporting the variant. An equal distribution of forward and reverse reads for a given allele would give a value of 0.5. (See also more information about overlapping pairs in section 20.19.)

Hyper-allelic Relevant for "Quality-based Variant Detection". Reports hyper-allelic status of variants based on the specified threshold "Maximum expected allele" in the "Set genome information" wizard under "Ploidy". The output in the table is "Yes" or "No" with respect to whether the threshold has been exceeded.

Variant tracks that have been created with Genomics Workbench 6.0 will have an additional column with the header 'Linkage'.



Chromosome	Region	Type	Reference	A
1	153005596	Insertion	A	A
1	153651291..153651292	Insertion	CT	C
1	153963998^153963999	Insertion	-	G

C...	Region	Type	Refer...	Allele	Refer...
1	153005597	Insertion	A	AA	No
1	153651291..153651292	Insertion	CT	CTCT	No
1	153963998^153963999	Insertion	-	GA	No

Chromosome	Region	Type	Refer
1	153005596^153005597	Insertion	-
1	153963998^153963999	Insertion	-

Figure 20.75: *Examples of variants with different types of 'Region' column contents. The left-most variant has a 'single position' region, the middle variant has a 'region' region and the right-most has a 'between positions' region.*

Please note that the variants in the variant track can be enriched with information using the annotation tools in section 21.

A variant track can be imported and exported in VCF or GVF formats. An example of the gvf-file giving rise to the variants shown in figure 20.75 is given in figure 20.76.

```
##gff-version 3
##gvf-version 1.06
##file-date 2013-09-23
#file-encoding windows-1252
1   CLC insertion   153005596   153005596   0   .   .   ID=CLC_1;Variant_seq=AA;Reference_seq=A;
1   CLC insertion   153651291   153651292   0   .   .   ID=CLC_2;Variant_seq=CTCT;Reference_seq=CT;
1   CLC insertion   153963999   153963998   0   .   .   ID=CLC_3;Variant_seq=GA;Reference_seq=-;
```

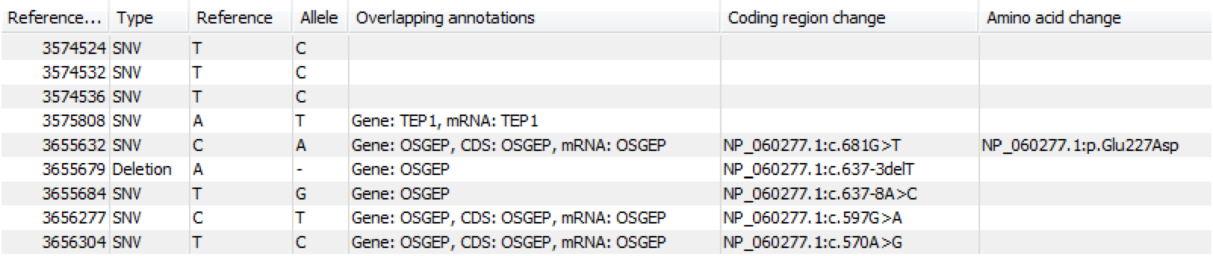
Figure 20.76: *A gvf file giving rise to the variants in the figure above.*

20.18.2 The annotated variant table

The annotated variant table (see figure 20.77) contains a subset of the columns of the variant track table and additionally the three columns below.

When the variant calling is performed on a read mapping in which gene and cds annotations are present on the reference sequence, the three columns will contain the following information:

Overlapping annotation This shows if the variant is covered by an annotation. The annotation's type and name will displayed. For annotated reference sequences, this information can be used to tell if the variant is found in e.g. a coding or non-coding region of the genome. **Note** that annotations of type `Variation` and `Source` are not reported.



Reference...	Type	Reference	Allele	Overlapping annotations	Coding region change	Amino acid change
3574524	SNV	T	C			
3574532	SNV	T	C			
3574536	SNV	T	C			
3575808	SNV	A	T	Gene: TEP1, mRNA: TEP1		
3655632	SNV	C	A	Gene: OSGEP, CDS: OSGEP, mRNA: OSGEP	NP_060277.1:c.681G>T	NP_060277.1:p.Glu227Asp
3655679	Deletion	A	-	Gene: OSGEP	NP_060277.1:c.637-3delT	
3655684	SNV	T	G	Gene: OSGEP	NP_060277.1:c.637-8A>C	
3656277	SNV	C	T	Gene: OSGEP, CDS: OSGEP, mRNA: OSGEP	NP_060277.1:c.597G>A	
3656304	SNV	T	C	Gene: OSGEP, CDS: OSGEP, mRNA: OSGEP	NP_060277.1:c.570A>G	

Figure 20.77: *An example of an annotated variant table.*

Coding region change For variants that fall within a coding region of a gene, the change is reported according to the standard conventions as outlined in http://www.hgvs.org/mutnomen/.

Amino acid change If the reference sequence of the mapping is annotated with ORF or CDS annotations, the variant caller will also report whether the variant is synonymous or non-synonymous. If the variant changes the amino acid in the protein translation, the new amino acid will be reported. The nomenclature used for reporting is taken from http://www.hgvs.org/mutnomen/.

If the reference sequence has no gene and cds annotations these columns will have the entry 'NA'.

The table can be **Exported** (📤) as a csv file (comma-separated values) and imported into e.g. Excel. Note that the CSV export includes all the information in the table, regardless of filtering and what has been chosen in the **Side Panel**. If you only want to use a subset of the information, simply select and **Copy** (⧉) the information.

Note that if you make a split view of the table and the mapping (see section 2.1.6), you will be able to browse through the variants by clicking in the table. This will cause the view to jump to the position of the variant.

This table view is not well-suited for downstream analysis, in which case we recommend working with tracks instead (see section 20.18.1).

20.18.3 Variant types

Variants are classified into five different types:

SNV A single nucleotide variant. This means that one base is replaced by one other base. This is also often referred to as a SNP. *SNV* is preferred over *SNP* because the latter includes an extra layer of interpretation about variants in a population. This means that an SNV could potentially be a SNP but this cannot be determined at the point where the variant is detected in a single sample.

MNV This type represents two or more SNVs in succession.

Insertion This refers to the event where one or more bases are inserted in the experimental data compared to the reference.

Deletion This refers to the event where one or more bases are deleted from the experimental data compared to the reference.

Replacement This is a more complex event where one or more bases have been replaced by one or more bases, where the identified allele has a length different from the reference (i.e. involving an insertion or deletion). Basically, this type represents variants that cannot be represented in the other four categories. An example could be `AAA->CC`. This cannot be resolved into a SNV or an MNV because the number of bases is different between the experimental data and the reference, it is not an insertion because something is also deleted from the reference, and it is not a deletion because something is also inserted.

20.19 Detailed information about overlapping paired reads

Paired reads that overlap introduce additional complexity for variant detection. This section describes how this is handled by *CLC Cancer Research Workbench*.

When it comes to **coverage** in the overlapping region, each pair is contributing once to the coverage. Even if there are indeed two reads in this region, they do not both contribute to coverage. The reason is that the two reads represent the same fragment, so they are essentially treated as one.

When it comes to counting the number of **forward and reverse reads**, including the forward/reverse reads balance, each read contribute. This is because this information is intended to account for systematic sequencing errors in one direction, and the fact that the two reads are from the same fragment is less important than the fact that they are sequenced on different strands.

If the two overlapping reads do not agree about the variant base, they are both ignored. Please note that there can be a special situation with the quality-based variant detection: If the two reads disagree, and one read does not pass the quality filter, the other read will contribute to the variant just as if there had been only that read and no overlapping pair.

Part VII

Working with variants

Chapter 21

Add information to variants tools

Contents

21.1 Add information from variant databases

It is also possible to annotate with information from variant databases. To run the Add information from variant databases tool, go to:

> **Toolbox | Add Information to Variants** ( ) **| Add Information from Variant Databases** ( )

This tool will create a new track with all the experimental variants including added information about overlapping variants found in the track of known variants. The annotations are marked in three different ways:

Exact match This means that the variant position and allele both have to be identical in the input and the known variants track (however, note the extra option for joining adjacent SNVs and MNVs described below).

Partial MNV match This applies to MNVs which can be annotated with partial matches if an SNV or a shorter MNV in the database has an allele sequence that is contained in the allele sequence of the annotated MNV.

Overlap This will report if the known variant track has an overlapping variant.

For exact matches, all the information about the variant from the known variants track is transferred to the annotated variant. For partial matches and overlaps, the information from the known variants are not transferred.

21.2 Add conservation scores

The possible functional consequence of a variant can be interrogated by comparing to a conservation score that tells how conserved this particular position is among a set of different species. The underlying line of thought is that conserved bases are functionally important - otherwise they would have been mutated during evolution. If a variant is found at a position that is otherwise well conserved, it is an indication that the variant is functionally important. Of course this is only a prediction, as non-conserved regions could have functional roles too.

Conservation scores can be computed by several tools e.g. PhyloP and PhastCons and can be downloaded as pre-computed scores from an whole genome alignment of different species from different sources. See how to find and import tracks with conservation scores in section 6.2.

Toolbox | **Add Information to Variants** ( ) | **Add Conservation Scores** ( )

Select the variant track as input and when you click **Next** you will need to provide the track with conservation scores (see figure 21.1).

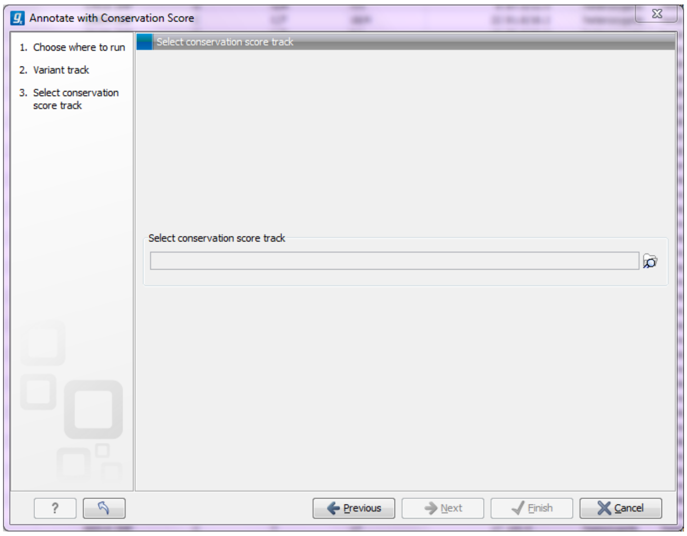
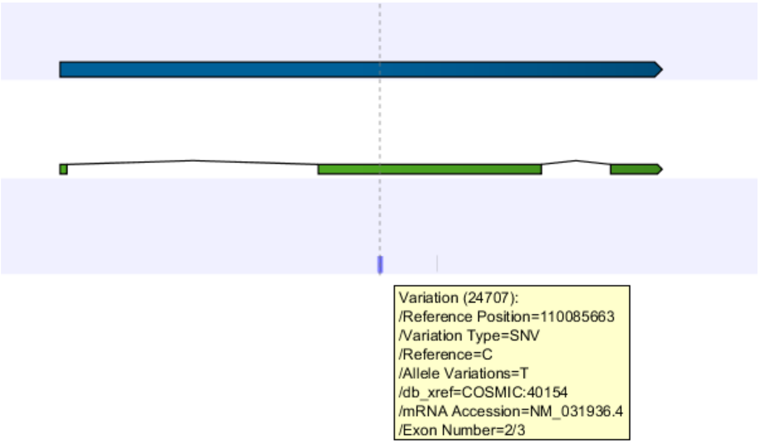

Annotate with Conservation Score
1. Choose where to run
2. Variant track
3. Select conservation score track
Select conservation score track
Select conservation score track
?
Previous
Next
Finish
Cancel

Figure 21.1: *The conservation score track.*

In the resulting track, all the variants will have quality scores annotated, and this can be used for sorting and filtering the track (see section 17.2.3).

21.3 Add exon number

Given a track with mRNA annotations, a new track will be created in which variants are annotated with the numbering of the corresponding exon with numbered exons based on the transcript annotations in the input track (see an example of a result in figure 21.2).



Variation (24707):
/Reference Position=110085663
/Variation Type=SNV
/Reference=C
/Allele Variations=T
/db_xref=COSMIC:40154
/mRNA Accession=NM_031936.4
/Exon Number=2/3

Figure 21.2: *A variant found in the second exon out of three in total.*

When there are multiple isoforms, a comma-separated list of the exon numbers is given.

21.4 Add flanking sequence

In some situations, it is useful to see a variant in the context of the bases of the reference sequence. This information can be added using the **Annotate with Flanking Sequence** tool:

Toolbox | **Add Information to Variants** ( ) | **Add Flanking Sequence** ( )

This opens a dialog where you can select a variant track ( ) to be annotated.

Clicking **Next** will display the dialog shown in figure 21.3

Select a sequence track that should be used for adding the flanking sequence, and specify how large the flanking region should be.

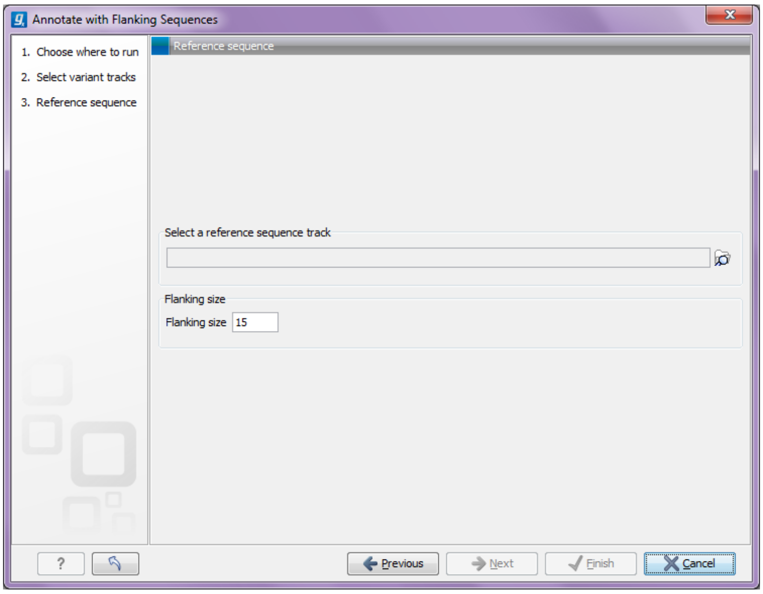
The result will be a new track with an additional column for the flanking sequence formatted like this: `CGGCT[T]AGTCC` with the base in square brackets being the variant allele.

21.5 Add fold changes

With this tool you can add the expression fold changes to your variants. You will create a copy of the input variant track and add the gene name and expression fold changes to this track. When you have added the expression fold changes to the variant track, they can be seen in the tooltip when you zoom all the way in on the individual variants or in the table view.

You can create a fold change track with the tool **Create Fold Change Track** that is described in section 27.1.7.

To add fold changes, go to the toolbox:



Select a reference sequence track

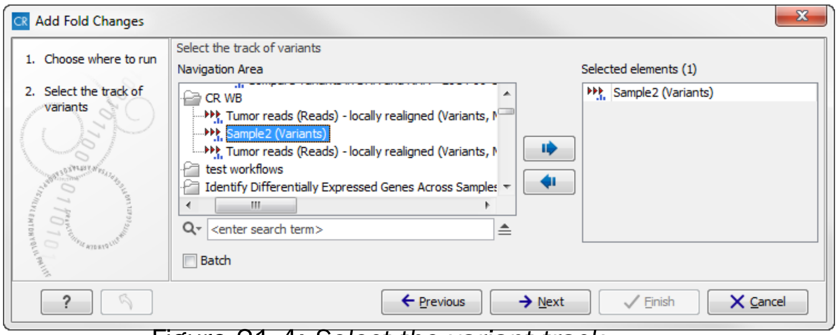
Flanking size

Flanking size 15

Figure 21.3: *Specifying a reference sequence and the amount of flanking bases to include.*

Toolbox | **Add Information to Variants** ( ) | **Add Fold Changes** ( )

If you are connected to a server, you will first be asked where you would like to run the analysis. Next, you will be asked to select the variant track that you would like to add fold changes to (figure 21.4). To select the variant track, double-click on the file name or click once on the file and then on the arrow pointing to the right side in the middle of the wizard. Click on the button labeled **Next**.


Add Fold Changes
1. Choose where to run
2. Select the track of variants
Navigation Area
CR WB
Tumor reads (Reads) - locally realigned (Variants, N
Sample2 (Variants)
Tumor reads (Reads) - locally realigned (Variants, N
test workflows
Identify Differentially Expressed Genes Across Samples
<enter search term>
Previous
Next
Finish
Cancel

Select the track of variants

Selected elements (1)

Sample2 (Variants)

□ Batch

Figure 21.4: *Select the variant track.*

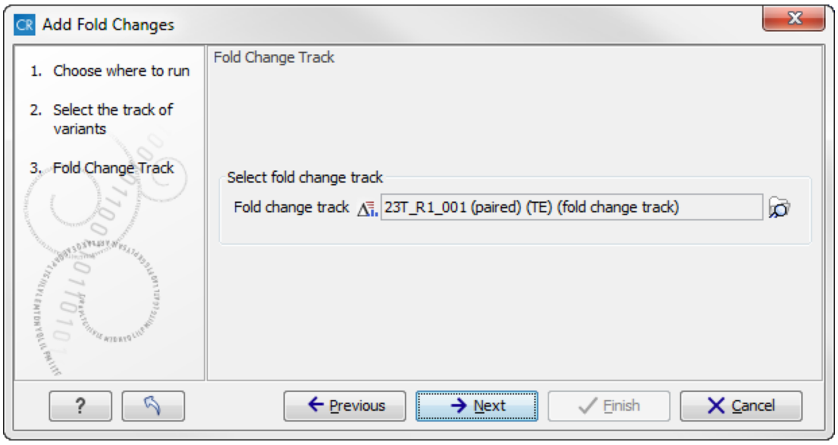
In the next step you can choose the fold change track (see figure 21.5).

Click on the button labeled **Next**, choose to save the results and click **Finish**.

The generated output is a variant track.If you open the variant track in table view by clicking on the table icon ( ) in the lower left corner of the **View Area**, you can see in the **Side Panel** under Table Settings that "Fold change" and "Gene" have been added to the list. If you would like to look into the numbers behind the fold changes, you can see the expression values in the original fold change file that was used as input in the "Add Fold Changes" analysis.

21.6 Add information about amino acid changes

To add information about amino acid changes to a variant track:



Add Fold Changes

1. Choose where to run
2. Select the track of variants
3. Fold Change Track

Fold Change Track

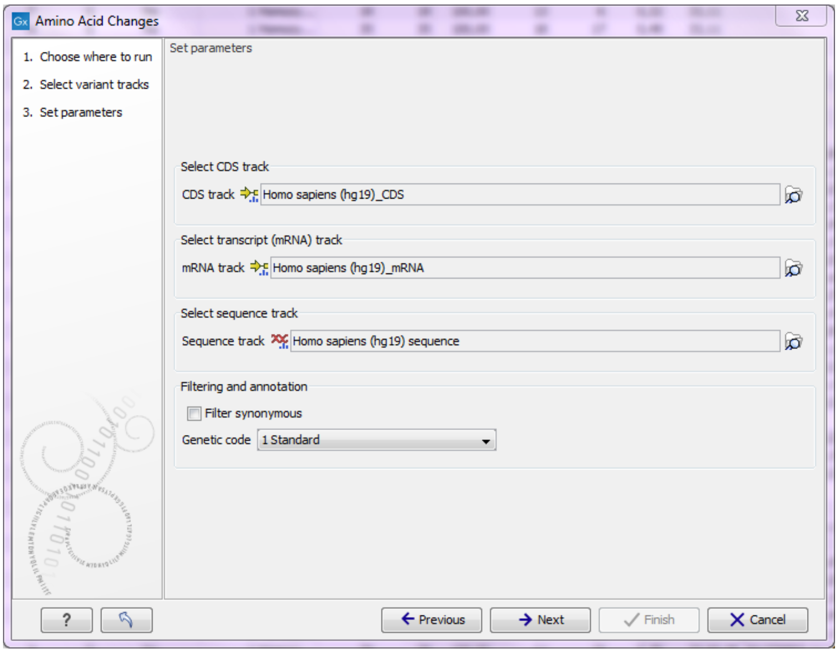
Select fold change track

Fold change track: 23T_R1_001 (paired) (TE) (fold change track)

Figure 21.5: *Select the fold change track.*

Toolbox | Add Information to Variants ( ) | Add Information about Amino Acid Changes ( )

This tool annotates variants with amino acid changes given a track with coding regions and a reference sequence (see figure 21.6).



Amino Acid Changes

1. Choose where to run
2. Select variant tracks
3. Set parameters

Set parameters

Select CDS track

CDS track: Homo sapiens (hg19)_CDS

Select transcript (mRNA) track

mRNA track: Homo sapiens (hg19)_mRNA

Select sequence track

Sequence track: Homo sapiens (hg19) sequence

Filtering and annotation

□ Filter synonymous

Genetic code: 1 Standard

Figure 21.6: *The amino acid changes annotation tool.*

The CDS track is used to determine the reading frame to be used for translation. The mRNA track is used to determine whether the variant is inside or outside the region covered by the transcript.

For each variant in the input track, the following information is added:

- **Coding region change**. This will annotate the relative position on the coding DNA level, using the nomenclature proposed at http://www.hgvs.org/mutnomen/. Variants inside exons and in the untranslated regions of the transcript will also be annotated with the distance to the nearest exon. E.g. "c.-4A>C" describes a SNV four bases upstream of the start codon, while "c.*4A>C" describes a SNV four bases downstream of the stop codon.

- **Amino acid change**. This will annotate the change on the protein level. For example, single amino-acid changes caused by SNVs are listed as "p.[Gly261Cys]", denoting that in the protein sequence (hence the "p.") the Glycine at position 261 is changed into Cysteine. Frame-shifts caused by indels are listed with the extension *fs*, for example p.[Pro244fs] denoting a frameshift at position 244 coding for Proline. For further details of the nomenclature see the "Recommendations for the description of protein sequence variants (v2.0)" at http://www.hgvs.org/mutnomen/.

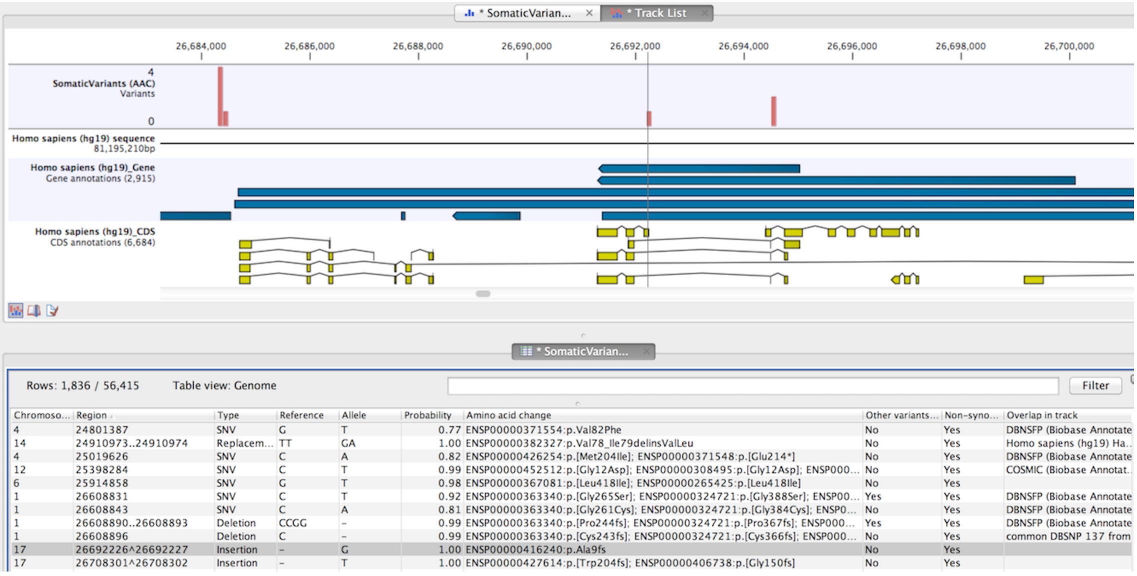
- **Coding region change in longest transcript**. When there are many transcript variants for a gene, the coding region change for all transcripts are listed in the "Coding region change" column. For quick reference, the longest transcript is often used, and there is a special column only listing the coding region change for the longest transcript.

- **Amino acid change in longest transcript**. This is similar to the above, just on the protein level.

- **Other variants within codon**. If there are other variants within the same codon, this column will have a "Yes". In this case, it should be manually investigated whether the two variants are linked by reads and the amino acid change annotated by the amino acid changes may not be correct in this case.

- **Non-synonymous**. Will have a "Yes" if the variant is non-synonymous.

By filtering in the table view of the result track on the column "Non-synonymous" for "Yes", only variants that change the protein product will be retained in the result track.



Rows: 1,836 / 56,415   Table view: Genome

Chromoso...	Region	Type	Reference	Allele	Probability	Amino acid change	Other variants...	Non-syno...	Overlap in track
4	24801387	SNV	G	T	0.77	ENSP00000371554:p.Val82Phe	No	Yes	DBNSFP (Biobase Annotate
14	24910973..24910974	Replacem...	TT	GA	1.00	ENSP00000382327:p.Val78_Ile79delinsValLeu	No	Yes	Homo sapiens (hg19) Ha..
4	25019626	SNV	C	A	0.82	ENSP00000426254:p.[Met204Ile]; ENSP00000371548:p.[Glu214*]	No	Yes	DBNSFP (Biobase Annotate
12	25398284	SNV	C	T	0.99	ENSP00000452512:p.[Gly12Asp]; ENSP00000308495:p.[Gly12Asp]; ENSP000...	No	Yes	COSMIC (Biobase Annotat..
6	25914858	SNV	G	T	0.98	ENSP00000367081:p.[Leu418Ile]; ENSP00000265425:p.[Leu418Ile]	No	Yes	
1	26608831	SNV	C	T	0.92	ENSP00000363340:p.[Gly265Ser]; ENSP00000324721:p.[Gly388Ser]; ENSP00...	Yes	Yes	DBNSFP (Biobase Annotate
1	26608843	SNV	C	A	0.81	ENSP00000363340:p.[Gly261Cys]; ENSP00000324721:p.[Gly384Cys]; ENSP0...	No	Yes	DBNSFP (Biobase Annotate
1	26608890..26608893	Deletion	CCGG	–	0.99	ENSP00000363340:p.[Pro244fs]; ENSP00000324721:p.[Pro367fs]; ENSP000...	Yes	Yes	DBNSFP (Biobase Annotate
1	26608896	Deletion	C	–	0.99	ENSP00000363340:p.[Cys243fs]; ENSP00000324721:p.[Cys366fs]; ENSP000...	No	Yes	common DBSNP 137 from
17	26692226^26692227	Insertion	–	G	1.00	ENSP00000416240:p.Ala9fs	No	Yes	
17	26708301^26708302	Insertion	–	T	1.00	ENSP00000427614:p.[Trp204fs]; ENSP00000406738:p.[Gly150fs]	No	Yes	

Figure 21.7: *The resulting amino acid changes in track and table views.*

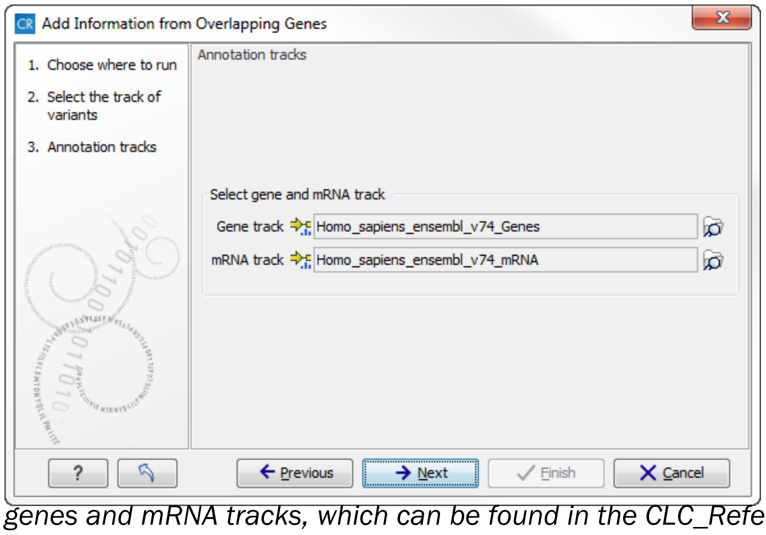
An example of the output is given in Figure 21.7. The top track view displays the variant track, sequence track, gene annotation and CDS track. The lower table view is filtered for non-synonymous variants.

21.7 Add information from overlapping genes

This will create a copy of the track used as input and add information from overlapping genes and mRNA tracks. To run the Add information from overlapping genes tool, go to the toolbox:

Toolbox | **Add Information to Variants** ( ) | **Add Information from Overlapping Genes** ( )

First, select the track you wish to annotate and click **Next**. You can choose any kind of variant or annotation track as input. Next, select the a gene track and a mRNA track for overlap comparison (figure 21.8).


Add Information from Overlapping Genes

1. Choose where to run
2. Select the track of variants
3. Annotation tracks

Annotation tracks

Select gene and mRNA track

Gene track: Homo_sapiens_ensembl_v74_Genes

mRNA track: Homo_sapiens_ensembl_v74_mRNA

Figure 21.8: *Select the genes and mRNA tracks, which can be found in the CLC_References folder.*

You can find the gene and mRNA tracks in the CLC_References folder (figure 21.9).

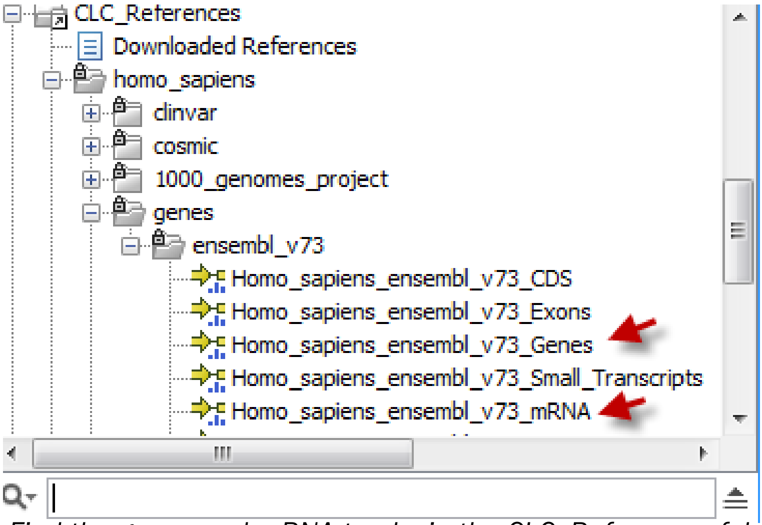

CLC_References
Downloaded References
homo_sapiens
clinvar
cosmic
1000_genomes_project
genes
ensembl_v73
Homo_sapiens_ensembl_v73_CDS
Homo_sapiens_ensembl_v73_Exons
Homo_sapiens_ensembl_v73_Genes
Homo_sapiens_ensembl_v73_Small_Transcripts
Homo_sapiens_ensembl_v73_mRNA

Figure 21.9: *Find the genes and mRNA tracks in the CLC_References folder.*

The result of this tool is a new track showing all the variants that now have been annotated with the information about genes and mRNA that overlap with the identified variants. The requirement for being registered as an overlap is that parts of the annotations are overlapping, regardless of the strandedness of the annotations (note that this makes it unsuitable for comparing e.g. two gene tracks but great for annotating variants with overlapping genes or regulatory regions). The added information can be visualized in two ways; 1) In the track tooltips when mousing over the individual variants or 2) in the table view where you can see that new columns describing the added gene and mRNA tracks have been added to the table. The table view can be accessed by clicking on the table icon (▦) in the lower part of the **View Area**.

21.8 Add information from genomic regions

This will create a copy of the track used as input and add information from overlapping annotations or regions. To run the Add information from genomic regions tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** ( ) | **Add Information from genomic regions** ( )

First, select the track you wish to annotate and click **Next**. You can choose any kind of variant or annotation track as input. Next, select the track holding the overlapping region of interest (e.g. regulatory regions from ENCODE or if you have imported other databases containing regions that you would like to use for overlap comparison).

The result of this tool is a new track showing all the variants that now have been annotated with the information about the regions that overlap with the identified variants. The requirement for being registered as an overlap is that parts of the annotations are overlapping, regardless of the strandedness of the annotations. The added information can be visualized in two ways; 1) In the track tooltips when mousing over the individual variants or 2) in the table view where you can see that new columns describing the added overlap information have been added to the table. The table view can be accessed by clicking on the table icon (▦) in the lower part of the **View Area**.

21.9 From databases

21.9.1 Add information from 1000 Genomes Project

To run the Add information from 1000 Genomes Project tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** ( ) | **From Databases** ( )| **Add Information from 1000 Genomes Project** ( )

This tools adds information from variants identified by the 1000 Genomes Project to your variants. All you have to do when running this tool is to select the variant track that you want to annotate with information from the 1000 Genomes Project and specify where you would like to save the output.

21.9.2 Add information from COSMIC

To run the Add information from COSMIC tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** ( ) | **From Databases** ( )| **Add Information from COSMIC** ( )

This tools adds information from known cancer variants in the COSMIC database to your variants. All you have to do when running this tool is to select the variant track that you want to annotate with information from the COSMIC database and specify where you would like to save the output.

21.9.3 Add information from ClinVar

To run the Add information from ClinVar tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** () | **From Databases** ()| **Add Information from 1000 Genomes Project** ()

This tools adds information from known clinically relevant variants in the ClinVar database to your variants.

21.9.4 Add information from common dbSNP

To run the Add information from common dbSNP tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** () | **From Databases** ()| **Add Information from 1000 Genomes Project** ()

This tools adds information from common variants in the common dbSNP database to your variants.

21.9.5 Add information from HapMap

To run the Add information from HapMap tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** () | **From Databases** ()| **Add Information from 1000 Genomes Project** ()

This tools adds information from common variants in the HapMap database to your variants.

21.9.6 Add information from dbSNP

To run the Add information from dbSNP tool, go to the toolbox:

> **Toolbox** | **Add Information to Variants** () | **From Databases** ()| **Add Information from 1000 Genomes Project** ()

This tools adds information from variants in the dbSNP database to your variants.

Chapter 22

Remove variants tools

Contents

Comparison with known variants from variant databases is a key concept when working with resequencing data. The *CLC Cancer Research Workbench* provides two types of tools for facilitating this task: one for *adding information to your experimental variants* with information from known variants (e.g. adding information about phenotypes like cancer associated with a certain variant allele), and one for *removing your experimental variants* based on this information (e.g. for removing common variants).

22.1 Remove variants found in external database

Any variant track can be used as "external database". It may either be produced by the *CLC Cancer Research Workbench*, imported or downloaded from variant database resources like dbSNP, 1000 genomes, HapMap etc. (see section 6.2). Hence, this tool has overlapping function with the three "From Databases" tools.

To run the Remove variants found in external database, go to the toolbox:

Toolbox | **Remove Variants** ( ) | **Remove Variants Found in External Database** ( )

(figure 22.1)

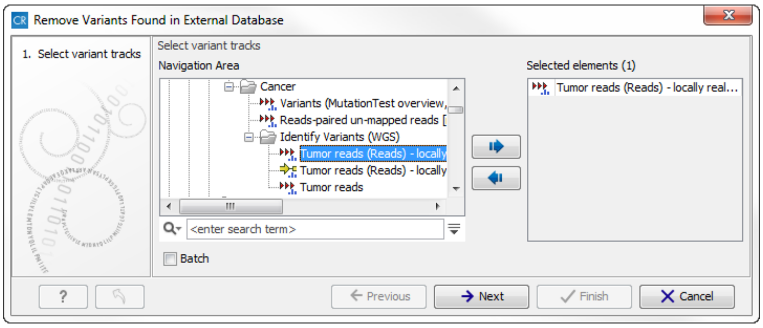

Remove Variants Found in External Database
1. Select variant tracks
Select variant tracks
Navigation Area
Cancer
Variants (MutationTest overview,
Reads-paired un-mapped reads [
Identify Variants (WGS)
Tumor reads (Reads) - locally
Tumor reads (Reads) - locally
Tumor reads
<enter search term>
Selected elements (1)
Tumor reads (Reads) - locally real...
□ Batch
?
Previous
Next
Finish
Cancel

Figure 22.1: *Wizard Step 1.*

22.2 Remove variants not found in external database

To run the Remove variants not found in external database tool, go to the toolbox:

> **Toolbox** | **Remove Variants** ( ) | **Remove Variants not Found in External Database** ( )

This tool removes variants that are found in your data set, but not in the specific external database, which you have imported into the Cancer Research Workbench as a track. The track with the external variants has to be specified as a parameter to the tool.

22.3 Remove false positives

To run the Remove false positive, go to the toolbox:

> **Toolbox** | **Remove Variants** ( ) | **Remove false positives** ( )

This tool will remove false positives by removing variants with low frequency, low average quality, and a bad forward/reverse balance.

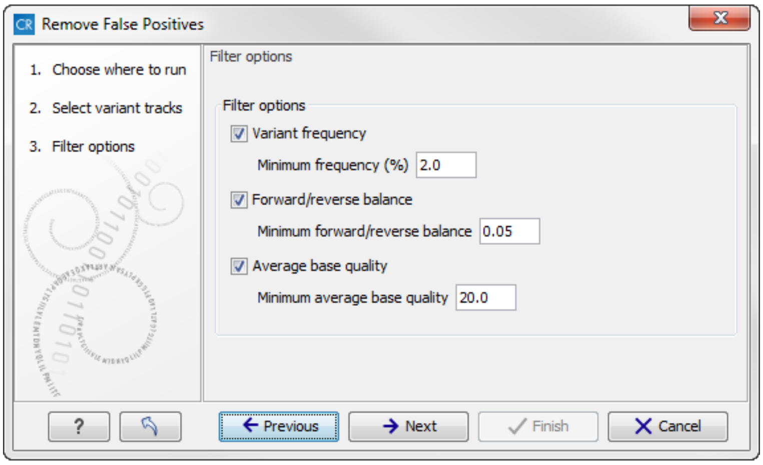
After you have selected the variant track that you would like to remove false positives from, you can adjust the filter parameters to specify how many reads should be supporting a variant to pass the filter (figure 22.2).

The **Filter options** are:

- **Variant frequency** Checking this box allows to specify the minimum frequency %. Variants that are present below this frequency (calculated as 'count'/'coverage') will be removed.
- **Forward/reverse balance** Checking this box allows to specify the forward/reverse balance. Variants with a threshold below the specified threshold will be removed. E.g. if you check the forward/reverse balance setting and adjust this parameter to be a value greater than 0 and less than 0.5, it will then be necessary for at least two reads to be supporting the variant in order to pass this filter.
- **Average base quality** Checking this box allows to specify the minimum average base quality. Variants with an average base quality below the specified threshold will be removed.

22.4 Remove Germline Variants

Running the variant caller on a case and control sample separately and filtering away variants



Filter options

☑ Variant frequency

Minimum frequency (%) 2.0

☑ Forward/reverse balance

Minimum forward/reverse balance 0.05

☑ Average base quality

Minimum average base quality 20.0

Figure 22.2: *This wizard step allows you to adjust the parameters for filtering.*

found in the control data set does not always give a satisfactory result as many variants in the control sample have not been called. This is often due to lack of read coverage in the corresponding regions or too stringent parameter settings. Therefore, instead of calling variants in the control sample, the Remove Germline Variants tool can be used to remove variants found in both samples from the set of candidate variants identified in the case sample.

Toolbox | **Remove Variants** ( ) | **Remove Germline Variants**

The variant track from the case sample must be used as input. When clicking **Next**, you are asked to supply the number of reads in the control data set that should support the variant allele in order to include it as a match (see figure 22.3). All the variants where at least this number of control reads show the particular allele will be filtered away in the result track.

Please note that variants, which have no coverage in the mapped control reads will be reported too. You can identify them by looking for a 0 value in the column 'Control coverage'.

The following annotations will be added to each variant not found in the control data set:

Control count For each allele the number of reads supporting the allele.

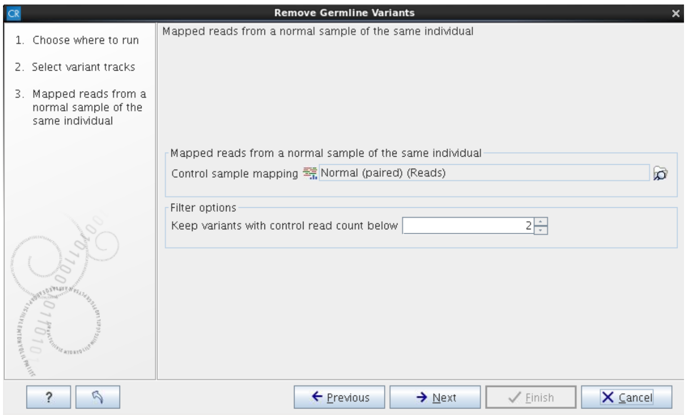
Control coverage Read coverage in the control dataset for the position in which the allele has been identified in the case dataset.

Control frequency Percentage of reads supporting the allele in the control sample.

The filter option can be used to set a threshold for which variants should be kept. In the dialog shown in figure 22.3 the threshold is set at two. This means that if a variant is found in only two or less of the control reads, it will be filtered away.

22.5 Remove reference variants

The variant tracks produced by the variant detection tools of *CLC Cancer Research Workbench* include reference alleles complementing a non-reference allele (i.e. a heterozygous variant where only one allele is different from the reference). In some situations this information is not necessary and these reference allele variants can be filtered away. To run the Remove reference variants, go to the toolbox:



Remove Germline Variants

1. Choose where to run
2. Select variant tracks
3. Mapped reads from a normal sample of the same individual

Mapped reads from a normal sample of the same individual

Mapped reads from a normal sample of the same individual

Control sample mapping: Normal (paired) (Reads)

Filter options

Keep variants with control read count below: 2

Figure 22.3: *Specify here the read mapping of the control sample and the minimum number of reads, which should include the variant before it will be removed as germline variant.*

Toolbox | **Remove Variants** () | **Remove Reference Variants** ()

This opens a wizard where you can select a variant track () that should be filtered.

Click on the button labeled **Next** and **Finish** to create a new track without the reference variants.

22.6 Remove variants inside genome regions

To run the Remove variants inside genome regions, go to the toolbox:

Toolbox | **Remove Variants** () | **Remove variants inside genome regions** ()

This tool will remove variants that are present in specific genome regions. Genomic regions have to be available as a track and have to be specified as a parameter.

22.7 Remove variants outside genome regions

To run the Remove variants outside genome regions, go to the toolbox:

Toolbox | **Remove Variants** () | **Remove variants outside genome regions** ()

This tool will remove variants that are present outside specific genome regions. Genomic regions have to be available as a track and have to be specified as a parameter.

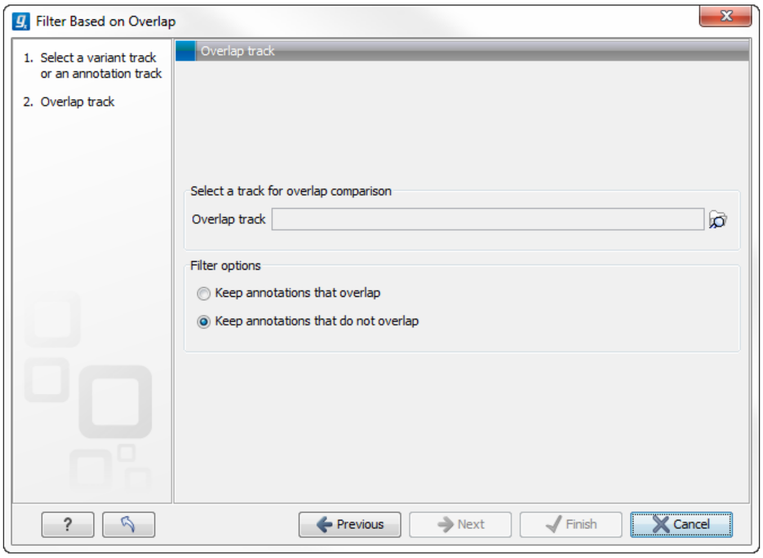
22.8 Remove variants outside targeted regions

The overlap filter will be used to filter an annotation track based on an overlap with another annotation track. This can be used to e.g. only show variants that fall within genes or regulatory regions or for restricting variant results to only cover a subset of genes as explained in section 26.4. Please note that for comparing variant tracks, more specific filters should be used (see chapter 21).

To run the Remove variants outside targeted regions, go to the toolbox:

Toolbox | **Remove Variants** () | **Remove variants outside targeted regions** ()

Select the track you wish to filter and click on the button labeled **Next** to specify the track of overlapping annotations (see figure 26.5).



Select a track for overlap comparison

Overlap track

Filter options

○ Keep annotations that overlap

● Keep annotations that do not overlap

Figure 22.4: *Select overlapping annotations track.*

Next, select the track that should be used for comparison and tick whether you wish to keep annotations that overlap, or whether to keep annotations that do not overlap with the selected track. An overlap has a simple definition – if the annotation used as input has at least one shared position with the other track, there is an overlap. The boundaries of the annotations do not need to match.

22.9 From databases

22.9.1 Remove variants found in 1000 genomes project

To run the Remove variants found in 1000 genomes project, go to the toolbox:

> **Toolbox** | **Remove Variants** ( ) | **From Databases** ( )| **Remove Variants Found in 1000 Genomes Project** ( )

This tool will remove variants present in 1000 Genomes Data.

22.9.2 Remove variants found in common dbSNP

To run the Remove variants found in common dbSNP, go to the toolbox:

> **Toolbox** | **Remove Variants** ( ) | **From Databases** ( )| **Remove Variants Found in common dbSNP** ( )

This tool will remove variants present in the common dbSNP database with common variants in a specific population (population frequency > 1%).

22.9.3 Remove variants found in HapMap

To run the Remove variants found in HapMap, go to the toolbox:

Toolbox | **Remove Variants** ( ) | **From Databases** ( )| **Remove variants found in HapMap** ( )

This tool will remove variants present in the HapMap database with common variants in a specific population.

Chapter 23

Add information to genes tool

Contents

23.1 Add information from overlapping variants

This will create a copy of the track used as input and add information from overlapping annotations or variants:

> **Toolbox** | **Add Information to Genes** ( ) | **Add Information from Overlapping Variants** ( )

First, select the track you wish to annotate and click **Next**. You can choose any kind of variant or annotation track as input. Next, select the track for overlap comparison, again you can choose any variant or annotation track.

The result of this tool is a new track with all the annotations from the input track and with additional information from the annotations that overlap from the other track. The requirement for being registered as an overlap is that parts of the annotations are overlapping, regardless of the strandedness of the annotations (note that this makes it unsuitable for comparing e.g. two gene tracks but great for annotating variants with overlapping genes or regulatory regions).

When running the "Annotate with overlap information" tool with a gene track as input and a variant track as parameter track, a new column describing the specific variant is added to the Track Table. The variant description also appears in the track tooltips when mousing over the individual variants.

Chapter 24

Compare samples tools

Contents

24.1 Compare shared variants within a group of samples

This tool should be used if you are interested in finding common (frequent) variants in a group of samples. For example one use case could be that you have 50 unrelated patients with the same disease and would like to identify variants that are present in at least 70% of all patients. It can also be used to do an overall comparison between samples (a frequency threshold of 0% will report all alleles).

> **Toolbox | Compare Samples ( ) | Compare Shared Variants within a Group of Samples ( )**

This opens a dialog where you can select the variant tracks ( ) from the samples in the group.

Clicking **Next** will display the dialog shown in figure 24.1.

The **Frequency threshold** is the percentage of samples that have this variant. Setting it to 70% means that at least 70% of the samples selected as input have to contain a given variant for it to be reported in the output.

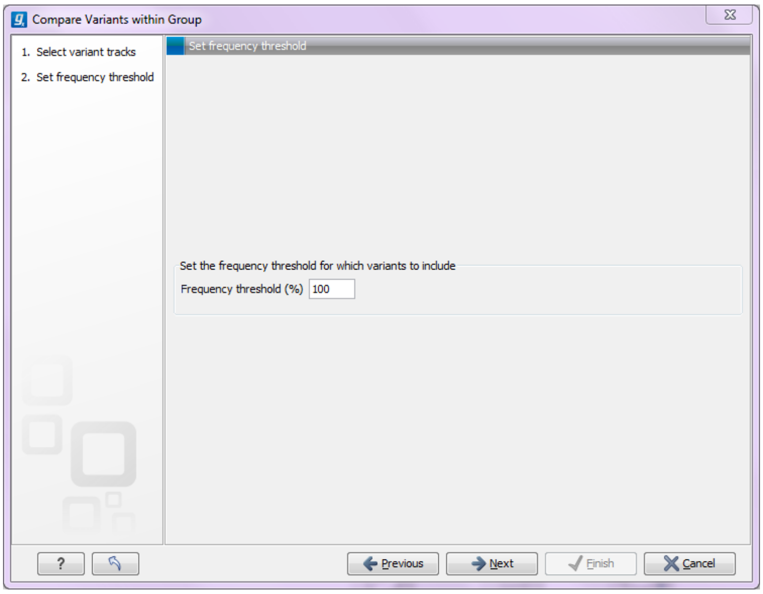
The output of the analysis is a track with all the variants that passed the frequency thresholds and with additional reporting of:

Sample count The number of samples that have the variant

Total number of samples The total number of samples (this will be identical for all variants).

Sample frequency This is the same frequency that is also used as a threshold (see figure 24.1).

Origin tracks A comma-separated list of the name of the tracks that contain the variant.



Compare Variants within Group

1. Select variant tracks
2. Set frequency threshold

Set frequency threshold

Set the frequency threshold for which variants to include

Frequency threshold (%) 100

Figure 24.1: *Frequency treshold.*

Note that this tool can be used for merging all variants from a number of variant tracks into one track by setting the frequency threshold to 0.

24.2 Identify Enriched Variants in Case vs Control Group

This tool should be used if you have a case-control study. This could be patients with a disease (case) and healthy individuals (control). The idea is to identify variants which are more common in the case samples than in the control samples.

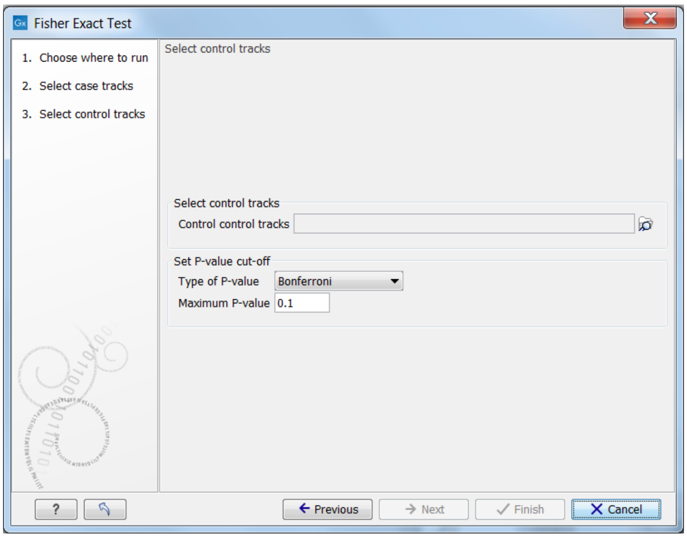
The Fisher exact test is applied on the number of occurrences of each allele of each variant in the case and the control data set. The alleles from each variant are considered separately, i.e. for an SNV with two alleles; a Fisher Exact test will be applied to each of the two. The test will also check whether an SNV in the case group is part of an MNV in the control group. Those with a low p-value are potential candidates for variants playing a role in the disease/phenotype. Please note that a low p-value can only be reached if the number of samples in the data set is high.

The tool is found in the Toolbox:

> **Toolbox** | **Compare Samples** ( ) | **Identify Enriched Variants in Case vs Control Group** ( )

In the first step of the dialog, you select the case variant tracks. Clicking **Next** shows the dialog in figure 24.2.

At the top, select the variant tracks from the control group. Furthermore, you must set a threshold for the p-value (default is 0.05); only variants having a p-value below this threshold will be reported. You can choose whether the threshold p-value refers to a corrected value for multiple tests (either Bonferroni Correction, or False Discovery Rate (FDR)), or an uncorrected p-value. A variant table is created as output (see figure 24.3), reporting only those variants with p-values lower than the threshold. All corrected and uncorrected p-values are shown here, so alternatively, variants with non-significant p-values can also be filtered out or more stringent thresholds can be applied at this stage, using the manual filtering options.



Fisher Exact Test

1. Choose where to run
2. Select case tracks
3. Select control tracks

Select control tracks

Select control tracks

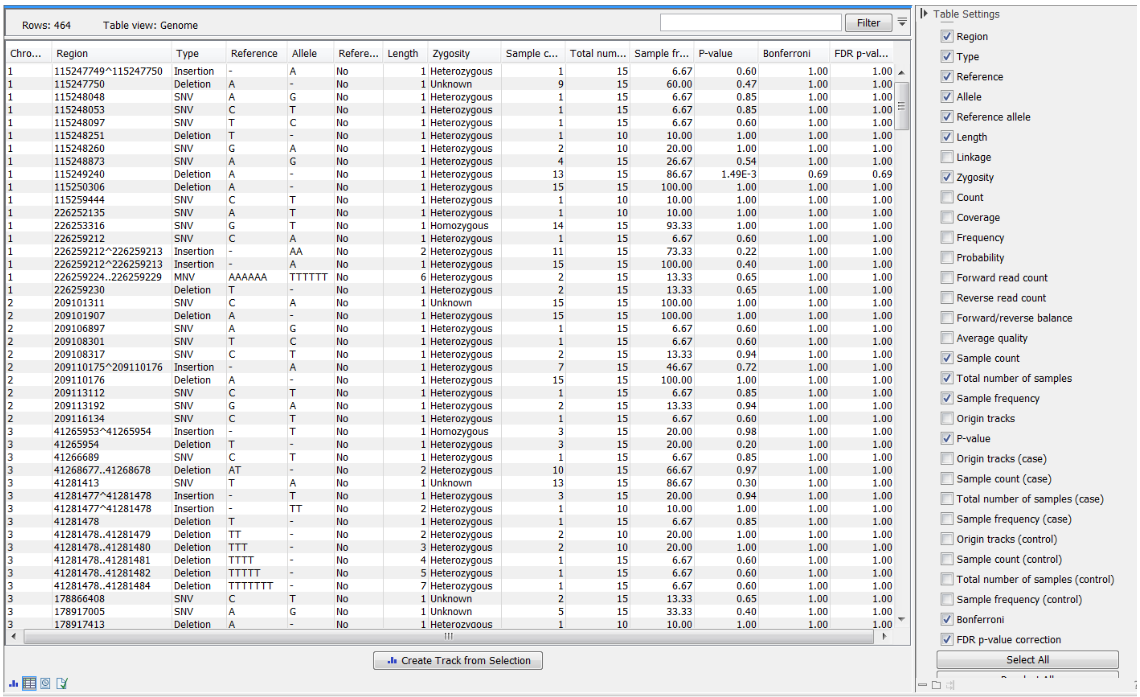
Control control tracks

Set P-value cut-off

Type of P-value: Bonferroni

Maximum P-value: 0.1

Figure 24.2: *In this dialog you can select the control tracks, a p-value correction method, and specify the p-value threshold for the fisher exact test.*



Rows: 464 Table view: Genome

Chro...	Region	Type	Reference	Allele	Refere...	Length	Zygosity	Sample c...	Total num...	Sample fr...	P-value	Bonferroni	FDR p-val...
1	115247749^115247750	Insertion	-	A	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
1	115247750	Deletion	A	-	No	1	Unknown	9	15	60.00	0.47	1.00	1.00
1	115248048	SNV	A	G	No	1	Heterozygous	1	15	6.67	0.85	1.00	1.00
1	115248053	SNV	C	T	No	1	Heterozygous	1	15	6.67	0.85	1.00	1.00
1	115248097	SNV	T	C	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
1	115248251	Deletion	T	-	No	1	Heterozygous	1	10	10.00	1.00	1.00	1.00
1	115248260	SNV	G	A	No	1	Heterozygous	2	10	20.00	1.00	1.00	1.00
1	115248873	SNV	A	G	No	1	Heterozygous	4	15	26.67	0.54	1.00	1.00
1	115249240	Deletion	A	-	No	1	Heterozygous	13	15	86.67	1.49E-3	0.69	0.69
1	115250306	Deletion	A	-	No	1	Heterozygous	15	15	100.00	1.00	1.00	1.00
1	115259444	SNV	C	T	No	1	Heterozygous	1	10	10.00	1.00	1.00	1.00
1	226252135	SNV	A	T	No	1	Heterozygous	1	10	10.00	1.00	1.00	1.00
1	226253316	SNV	G	T	No	1	Homozygous	14	15	93.33	1.00	1.00	1.00
1	226259212	SNV	C	A	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
1	226259212^226259213	Insertion	-	AA	No	2	Heterozygous	11	15	73.33	0.22	1.00	1.00
1	226259212^226259213	Insertion	-	A	No	1	Heterozygous	15	15	100.00	0.40	1.00	1.00
1	226259224..226259229	MNV	AAAAAA	TTTTTT	No	6	Heterozygous	2	15	13.33	0.65	1.00	1.00
1	226259230	Deletion	T	-	No	1	Heterozygous	2	15	13.33	0.65	1.00	1.00
2	209101311	SNV	C	A	No	1	Unknown	15	15	100.00	1.00	1.00	1.00
2	209101907	Deletion	A	-	No	1	Heterozygous	15	15	100.00	1.00	1.00	1.00
2	209106897	SNV	A	G	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
2	209108301	SNV	T	C	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
2	209108317	SNV	C	T	No	1	Heterozygous	2	15	13.33	0.94	1.00	1.00
2	209110175^209110176	Insertion	-	A	No	1	Heterozygous	7	15	46.67	0.72	1.00	1.00
2	209110176	Deletion	A	-	No	1	Heterozygous	15	15	100.00	1.00	1.00	1.00
2	209113112	SNV	C	T	No	1	Heterozygous	1	15	6.67	0.85	1.00	1.00
2	209113192	SNV	G	A	No	1	Heterozygous	2	15	13.33	0.94	1.00	1.00
2	209116134	SNV	C	T	No	1	Heterozygous	1	15	6.67	0.60	1.00	1.00
3	41265953^41265954	Insertion	-	T	No	1	Homozygous	3	15	20.00	0.98	1.00	1.00
3	41265954	Deletion	T	-	No	1	Heterozygous	3	15	20.00	0.20	1.00	1.00
3	41266689	SNV	C	T	No	1	Heterozygous	1	15	6.67	0.85	1.00	1.00
3	41268677..41268678	Deletion	AT	-	No	2	Heterozygous	10	15	66.67	0.97	1.00	1.00
3	41281413	SNV	T	A	No	1	Unknown	13	15	86.67	0.30	1.00	1.00
3	41281477^41281478	Insertion	-	T	No	1	Heterozygous	3	15	20.00	0.94	1.00	1.00
3	41281477^41281478	Insertion	-	TT	No	2	Heterozygous	1	10	10.00	1.00	1.00	1.00
3	41281478	Deletion	T	-	No	1	Heterozygous	1	15	6.67	0.85	1.00	1.00
3	41281478..41281479	Deletion	TT	-	No	2	Heterozygous	2	10	20.00	1.00	1.00	1.00
3	41281478..41281480	Deletion	TTT	-	No	3	Heterozygous	2	10	20.00	1.00	1.00	1.00
3	41281478..41281481	Deletion	TTTT	-	No	4	Heterozygous	1	15	6.67	0.60	1.00	1.00
3	41281478..41281482	Deletion	TTTTT	-	No	5	Heterozygous	1	15	6.67	0.60	1.00	1.00
3	41281478..41281484	Deletion	TTTTTTT	-	No	7	Heterozygous	1	15	6.67	0.60	1.00	1.00
3	178866408	SNV	C	T	No	1	Unknown	2	15	13.33	0.65	1.00	1.00
3	178917005	SNV	A	G	No	1	Unknown	5	15	33.33	0.40	1.00	1.00
3	178917413	Deletion	A	-	No	1	Heterozygous	1	10	10.00	1.00	1.00	1.00

Table Settings

- ☑ Region
- ☑ Type
- ☑ Reference
- ☑ Allele
- ☑ Reference allele
- ☑ Length
- □ Linkage
- ☑ Zygosity
- □ Count
- □ Coverage
- □ Frequency
- □ Probability
- □ Forward read count
- □ Reverse read count
- □ Forward/reverse balance
- □ Average quality
- ☑ Sample count
- ☑ Total number of samples
- ☑ Sample frequency
- □ Origin tracks
- ☑ P-value
- □ Origin tracks (case)
- □ Sample count (case)
- □ Total number of samples (case)
- □ Sample frequency (case)
- □ Origin tracks (control)
- □ Sample count (control)
- □ Total number of samples (control)
- □ Sample frequency (control)
- ☑ Bonferroni
- ☑ FDR p-value correction

Figure 24.3: *In the output table, you can view information about all significant variants, select which columns to view, and filter manually on certain criteria.*

There are many other columns displaying information about the variants in the output table, such as the type, sequence, and length of the variant, its frequency and read count in case and control samples, and its overall zygosity. The zygosity information refers to **all** of the case samples; a label of 'homozygous' means the variant is homozygous in all case samples, a label of 'heterozygous' means the variant is heterozygous in all case samples, whereas a label of 'unknown' means it is heterozygous in some, and homozygous in others.

Overlapping variants: If two different types of variants occur in the same location, these are reported separately in the output table. This is particularly important, where SNPs occur in the same position as an MNV. Usually, multiple SNVs occurring alongside each other would simply be reported as one MNV, but if one SNV of the MNV is found in additional case samples by itself, it will be reported separately. For example, if an MNV of AAT -> GCA at position 1 occurs in five of the case samples, and the SNV at position 1 of A -> G, occurs in an *additional* 3 samples (so 8 samples in total), the output table will list the MNV and SNV information separately (however, the SNV will be shown as being present in only 3 samples, as this is the number in which it appears 'alone').

The test will also check whether an SNV in the case group is part of an MNV in the control group.

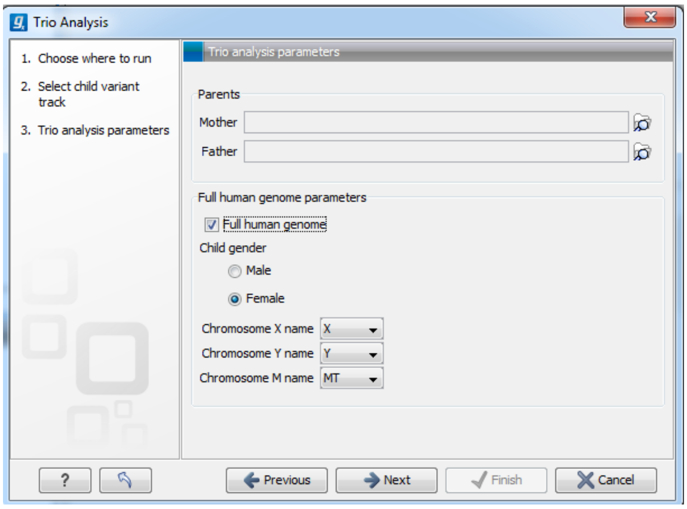
24.3 Trio analysis

This tool should be used if you have a trio study with one child and its parents. It should be mainly used for investigating differences in the child in comparison to its parents.

To start the Trio analysis:

Toolbox | **Compare Samples** () | **Trio Analysis** ()

In the first step of the dialog, select the variant track of the child. Clicking **Next** shows the dialog in figure 24.4.



Parents

Mother

Father

Full human genome parameters

☑ Full human genome

Child gender

○ Male

● Female

Chromosome X name X

Chromosome Y name Y

Chromosome M name MT

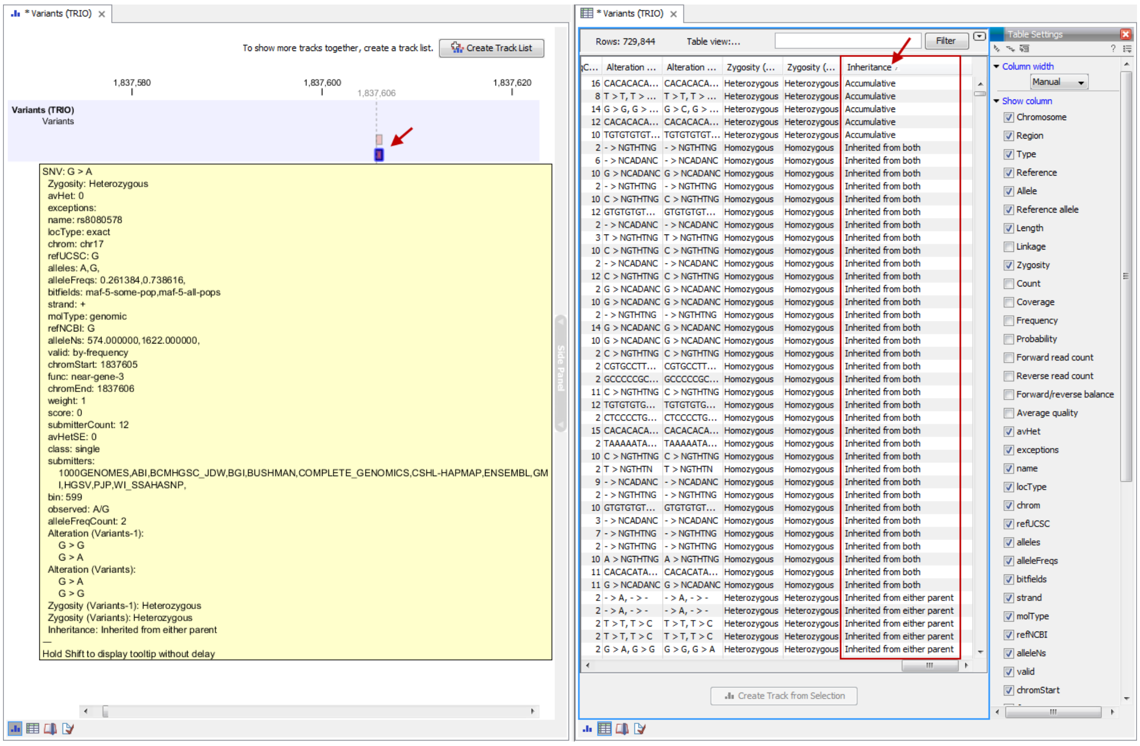
Figure 24.4: *Selecting variant tracks of the parents.*

Click on the folder () to select the two variant tracks for the mother and the father. In case you have a human TRIO, please specify if the child is male or female and how the X, Y chromosomes as well as the mitochondrion are named in the genome track. These parameters are important in order to apply specific inheritance rules to these chromosomes.

Click **Next** and **Finish**.

The output is a variant track showing all variants detected in the child. For each variant in the child, it is reported whether the variant is inherited from the father, mother, both, either or is a de novo mutation. This information can be found in the tooltip for each variant or by switching to the table view (see the column labeled "Inheritance") (figure 24.5).

In cases where both parents are heterozygous with respect to a variant allele, and the child has the same phenotype as the parents, it is unclear which allele was inherited from which parent.



SNV: G > A
Zygosity: Heterozygous
avHet: 0
exceptions:
name: rs8080578
locType: exact
chrom: chr17
refUCSC: G
alleles: A,G,
alleleFreqs: 0.261384,0.738616,
bitfields: maf-5-some-pop,maf-5-all-pops
strand: +
molType: genomic
refNCBI: G
alleleNs: 574.000000,1622.000000,
valid: by-frequency
chromStart: 1837605
func: near-gene-3
chromEnd: 1837606
weight: 1
score: 0
submitterCount: 12
avHetSE: 0
class: single
submitters:
1000GENOMES,ABI,BCMHGSC_JDW,BGI,BUSHMAN,COMPLETE_GENOMICS,CSHL-HAPMAP,ENSEMBL,GMI,HGSV,PJP,WI_SSAHASNP,
bin: 599
observed: A/G
alleleFreqCount: 2
Alteration (Variants-1):
G > G
G > A
Alteration (Variants):
G > A
G > G
Zygosity (Variants-1): Heterozygous
Zygosity (Variants): Heterozygous
Inheritance: Inherited from either parent
—
Hold Shift to display tooltip without delay

Rows: 729,844

qC...	Alteration ...	Alteration ...	Zygosity (...	Zygosity (...	Inheritance
16	CACACACA...	CACACACA...	Heterozygous	Heterozygous	Accumulative
8	T > T, T > ...	T > T, T > ...	Heterozygous	Heterozygous	Accumulative
14	G > G, G > ...	G > C, G > ...	Heterozygous	Heterozygous	Accumulative
12	CACACACA...	CACACACA...	Heterozygous	Heterozygous	Accumulative
10	TGTGTGTGT...	TGTGTGTGT...	Heterozygous	Heterozygous	Accumulative
2	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
6	- > NCADANC	- > NCADANC	Homozygous	Homozygous	Inherited from both
10	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
2	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
10	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
12	GTGTGTGT...	GTGTGTGT...	Homozygous	Homozygous	Inherited from both
2	- > NCADANC	- > NCADANC	Homozygous	Homozygous	Inherited from both
3	T > NGTHTNG	T > NGTHTNG	Homozygous	Homozygous	Inherited from both
10	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
2	- > NCADANC	- > NCADANC	Homozygous	Homozygous	Inherited from both
12	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
2	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
10	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
2	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
14	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
10	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
2	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
2	CGTGCCTT...	CGTGCCTT...	Homozygous	Homozygous	Inherited from both
2	GCCCCCGC...	GCCCCCGC...	Homozygous	Homozygous	Inherited from both
11	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
12	TGTGTGTG...	TGTGTGTG...	Homozygous	Homozygous	Inherited from both
2	CTCCCCTG...	CTCCCCTG...	Homozygous	Homozygous	Inherited from both
15	CACACACA...	CACACACA...	Homozygous	Homozygous	Inherited from both
2	TAAAAATA...	TAAAAATA...	Homozygous	Homozygous	Inherited from both
10	C > NGTHTNG	C > NGTHTNG	Homozygous	Homozygous	Inherited from both
2	T > NGTHTN	T > NGTHTN	Homozygous	Homozygous	Inherited from both
9	- > NCADANC	- > NCADANC	Homozygous	Homozygous	Inherited from both
2	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
10	GTGTGTGT...	GTGTGTGT...	Homozygous	Homozygous	Inherited from both
3	- > NCADANC	- > NCADANC	Homozygous	Homozygous	Inherited from both
7	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
2	- > NGTHTNG	- > NGTHTNG	Homozygous	Homozygous	Inherited from both
10	A > NGTHTNG	A > NGTHTNG	Homozygous	Homozygous	Inherited from both
11	CACACATA...	CACACATA...	Homozygous	Homozygous	Inherited from both
11	G > NCADANC	G > NCADANC	Homozygous	Homozygous	Inherited from both
2	- > A, - > -	- > A, - > -	Heterozygous	Heterozygous	Inherited from either parent
2	- > A, - > -	- > A, - > -	Heterozygous	Heterozygous	Inherited from either parent
2	T > T, T > C	T > T, T > C	Heterozygous	Heterozygous	Inherited from either parent
2	T > T, T > C	T > T, T > C	Heterozygous	Heterozygous	Inherited from either parent
2	G > A, G > G	G > G, G > A	Heterozygous	Heterozygous	Inherited from either parent

Table Settings

Column width: Manual

Show column
☑ Chromosome
☑ Region
☑ Type
☑ Reference
☑ Allele
☑ Reference allele
☑ Length
☐ Linkage
☑ Zygosity
☐ Count
☐ Coverage
☐ Frequency
☐ Probability
☐ Forward read count
☐ Reverse read count
☐ Forward/reverse balance
☐ Average quality
☑ avHet
☑ exceptions
☑ name
☑ locType
☑ chrom
☑ refUCSC
☑ alleles
☑ alleleFreqs
☑ bitfields
☑ strand
☑ molType
☑ refNCBI
☑ alleleNs
☑ valid
☑ chromStart

Figure 24.5: *Output from Trio Analysis showing the variants found in the child in track and table format.*

Such mutations are described as 'Inherited from either parent'.

In cases where both parents are homozygous with respect to a variant allele, and the child has the same phenotype as the parents, it is also unclear which allele was inherited from which parent. Such mutations are described as 'Inherited from both parents'.

In cases where both parents are heterozygous and the child homozygous for the variant, the child has inherited a variant from both parents. In such cases the tool will also check for a potential 'accumulative' mutation. Accumulative mutations are present in a heterozygous state in each of the parents, but are homozygous in the child. To investigate potential disease relevant variants, 'accumulative' variants and de novo variants are the most interesting (in case the parents are not affected). The tool will also add information about the genotype (homozygote or heterozygote) in all samples.

For humans, special rules apply for chromosome X (in male children) and chromosome Y, as well as the mitochondrion, as these are haploid and always inherited from the same parent. Heterozygous variants in the child that do not follow mendelian inheritance patterns will be marked in the result.

Let's look at an example where these special rules apply - in this case the trio analysis is performed with a boy:

The boy has a position on the Y chromosome that is heterozygous for C/T. The heterozygous C is not present in neither the mother or father, but the T is present in the father. In this case the inheritance result for the T variant will be: 'Inherited from the father', and for the C variant 'de novo'. However, both variants will also be marked with 'Yes' in the column 'Mendelian inheritance problem' because of this aberrant situation. In case the child is female, all variants on the Y chromosome will be marked in the same way.

The following annotations will be added to the resulting child track:

Zygosity Zygosity in the child as reported from the variant caller. Can be either homozygote or heterozygote.

Zygosity (Name of parent track 1) Zygosity in the corresponding parent (e.g. father) as reported from the variant caller. Can be either homozygote or heterozygote.

Allele variant (Name of parent track 1) Alleles called in the corresponding parent (e.g. father).

Zygosity (Name of parent track 2) Zygosity in the corresponding parent (e.g. mother) as reported from the variant caller. Can be either homozygote or heterozygote.

Allele variant (Name of parent track 2) Alleles called in the corresponding parent (e.g. mother).

Inheritance Inheritance status. Can be one of the following values: 'De novo', 'Accumulative', 'Inherited from both', 'Inherited from either', 'Inherited from (Name of parent track)'.

Mendelian inheritance problem Variants not following the mendelian inheritance pattern are marked here with 'Yes'.

Note! If the variant at this position cannot be found in either of the parents, the zygosity status of the parent where the variant has not been found is unknown, and the allele variant column will be left empty.

Chapter 25

Identify candidate variants tools

Contents

25.1 Create Filter Criteria

To run the Create Filter Criteria, go to the toolbox:

Toolbox | **Identify Candidate Variants** ( ) | **Create Filter Criteria** ( )

After the variants have been identified and post-filtered (e.g. for somatic variants), the next task is to identify e.g. driver mutations or at to identify those variants, which should be validated first.

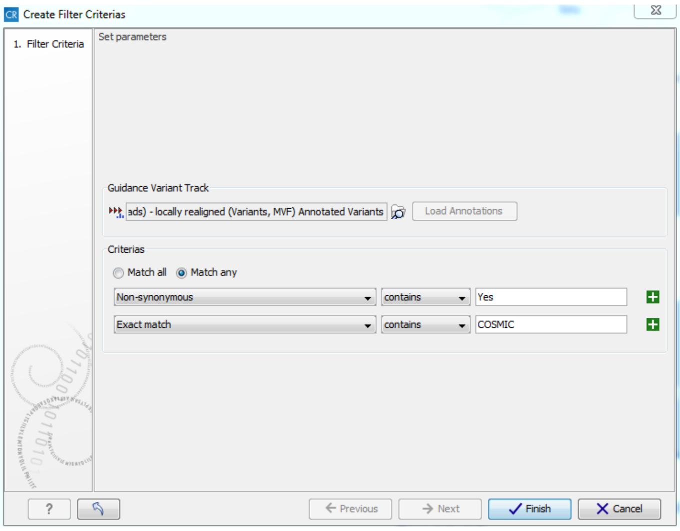
The Create Filter Criteria tool can help setting up advanced filter criteria that can be used in the **Identify Candidate Variants** tool (see section 25.2) to identify candidate variants. The section 25.2) can be used in analysis workflows later on.

An example of a filter criterion is shown in figure 25.1 where you will filter out everything that is not non-synonymous variants or variants that are present in the COSMIC database. As a result you will only keep the variants that are already known in cancer (present in the COSMIC database) or that are non-synonymous.

Please select and load one of your annotated variant tracks in the dialog box by clicking on the folder icon. This track is used as a guidance variant track in order to obtain the annotation categories (annotation column headers) to be used for selection of which annotations the filter should be applied to. Click on the "Load Annotations" button to activate the filter categories found under "Criteria".

The created filter criteria can be modified by opening the generated output from the **Create Filter Criteria** by clicking on the name of the filter criteria file where you saved it the **Navigation Area**.

Caution! Please note, that when you create filter criteria, you should use a guiding variant track that contains the same annotations as the annotations that are present in the variant tracks on



Guidance Variant Track
ads) - locally realigned (Variants, MVF) Annotated Variants

Criterias
○ Match all ● Match any

Non-synonymous | contains | Yes
Exact match | contains | COSMIC

Figure 25.1: *A filter criterion to extract those variants, which are known in cancer (present in the COSMIC database) or are non-synonymous*

which the filter criteria should be applied to in a workflow.

25.2 Identify candidate variants

To run the Identify candidate variants, go to the toolbox:

> **Toolbox** | **Identify Candidate Variants** ( ) | **Identify candidate variants** ( )

The **Identify Candidate Variants** tool can be used to identify only the variants that fulfill certain criteria. This means that the **Identify candidate variants** tool is to be used together with a filter criteria file created with the **Create Filter Criteria** tool (see section 25.1). Hence, the **Identify Candidate Variants** tool takes tracks with annotated or non-annotated variants as input as well as the relevant filter criteria file that was generated with the **Create Filter Criteria** tool.

The output from the **Identify Candidate Variants** tool is a variant track (and table) that contains only the variants that fulfill the specified filter criteria.

25.3 Remove information from variants

When you use information from various databases to annotate your variants, you may end up with many duplicated annotations or even annotations that you are not really interested in. This tool can help you remove annotations that have been added to variants, so that the results track/results table only includes the information that is of relevance to you.

To run the Remove information from variants, go to the toolbox:

> **Toolbox** | **Identify Candidate Variants** ( ) | **Remove information from variants** ( )

The input for the tool is an annotated variant track (please make sure that you select a variant track that contains annotations e.g. Amino Acid Change, Exact Match, Conservation Score etc.). If you in the wizard (shown in figure 25.3) click on the button labeled "Load Annotations" , the annotations that have been added to the variants in the input track are preloaded in the window below. You can choose, which annotations should be kept or removed. Please use the Ctrl or

Shift keys on your keyboard to select the annotations.

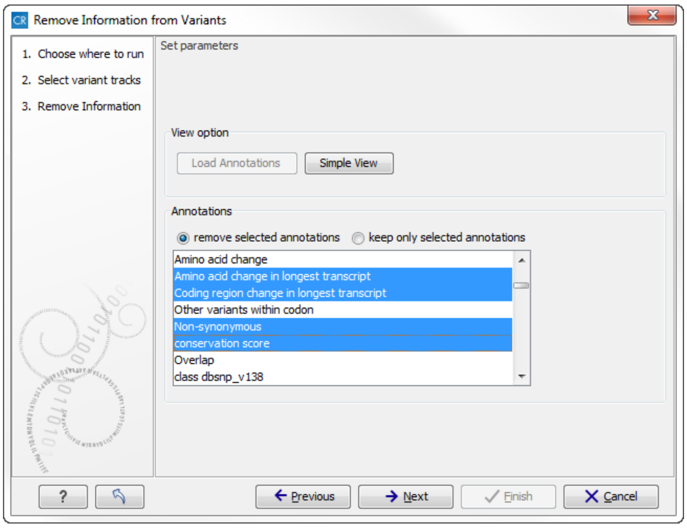


Figure 25.2: *The selected annotation columns preload from the input variant track will be removed in the output variant track.*

25.4 Identify variants with effect on splicing

This tool will analyze a variant track to determine whether the variants fall within potential splice sites. A transcript track has to be selected as shown in figure 25.3.

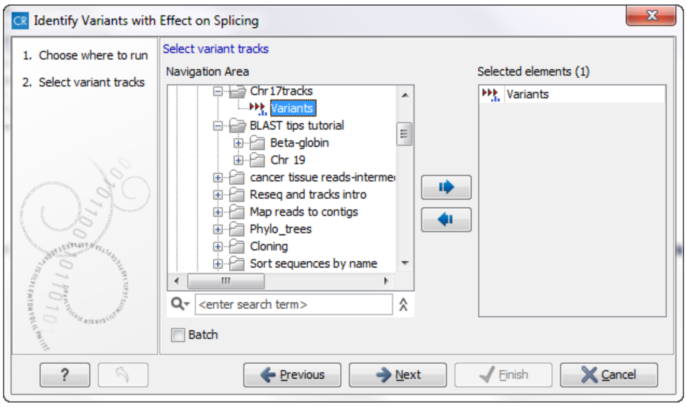


Figure 25.3: *Select the variant track that you wish to check for potential influence on splicing.*

If a variant falls within two base pairs of an intron-exon boundary, it will annotated as a possible splice site disruption. As part of the dialog box you can choose to exclude all variants that do not fall within a splice site.

Chapter 26

Identify candidate genes tools

Contents

26.1 Identify differentially expressed gene groups and pathways

This tool can be used to investigate candidate differentially expressed genes for a common functional role. For example if you would like to compare different cancer patients to check whether e.g. the same pathways are affected in different individuals, you can use this tool.

For this, you need a GO association file, which includes gene names and associated Gene Ontology terms. A GO association file with the top-level GO terms annotated (GO slim) is provided with the CLC Cancer Research Workbench and can be downloaded using the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).

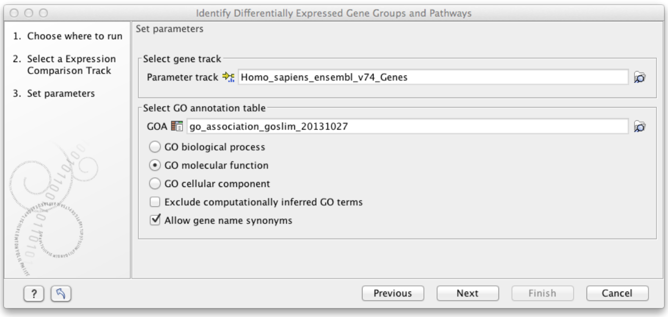
To run the analysis go to the toolbox:

> **Toolbox** | **Identify Candidate Genes** ( ) | **Identify Differentially Expressed Gene Groups and Pathways** ( )

When you run the Identify Differentially Expressed Gene Groups and Pathways analysis, you first have to select the expression comparison track ( ) you wish to annotate with the GO term enrichment analysis. Expression comparison tracks can be created e.g. by the create fold change track tool (see section 27.1.7).

After clicking **Next**, you have to specify the annotation association file, a gene track, and finally which ontology (cellular component, biological process or molecular function) you would like to test for (see figure 26.1).

Next, the Workbench tries to match gene names from the expression comparison track with the gene names in the GO association file. Please be aware that the same gene name definition should be used in both files.



Identify Differentially Expressed Gene Groups and Pathways

1. Choose where to run
2. Select a Expression Comparison Track
3. Set parameters

Set parameters

Select gene track

Parameter track: Homo_sapiens_ensembl_v74_Genes

Select GO annotation table

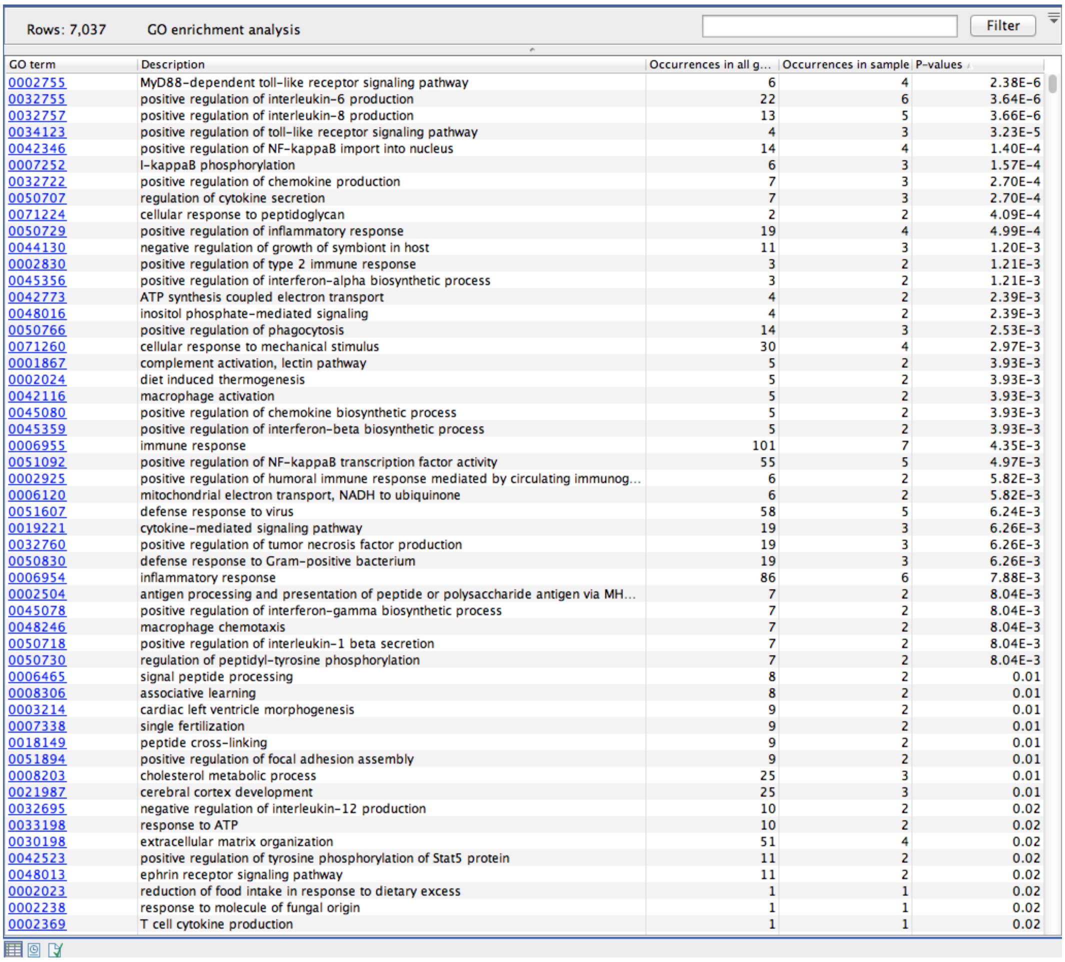
GOA: go_association_goslim_20131027

○ GO biological process
● GO molecular function
○ GO cellular component
☐ Exclude computationally inferred GO terms
☑ Allow gene name synonyms

Figure 26.1: *Select gene track, GO annotation table, and ontology.*

Based on this, the Workbench finds GO terms that are over-represented in the list. A hypergeometric test is used to identify over-represented GO terms by testing whether some of the GO terms are over-represented in a given gene set, compared to a randomly selected set of genes.

The result is a table with GO terms and the calculated p-value for the differentially expressed genes, and a new expression comparison track with annotated GO terms and the corresponding p-value (see figure 26.2). The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, or in other words how significant (trustworthy) a result is. In case of a small p-value the probability of achieving the same result by chance with the same test statistic is very small.



Rows: 7,037 GO enrichment analysis

GO term	Description	Occurrences in all g...	Occurrences in sample	P-values
0002755	MyD88-dependent toll-like receptor signaling pathway	6	4	2.38E-6
0032755	positive regulation of interleukin-6 production	22	6	3.64E-6
0032757	positive regulation of interleukin-8 production	13	5	3.66E-6
0034123	positive regulation of toll-like receptor signaling pathway	4	3	3.23E-5
0042346	positive regulation of NF-kappaB import into nucleus	14	4	1.40E-4
0007252	I-kappaB phosphorylation	6	3	1.57E-4
0032722	positive regulation of chemokine production	7	3	2.70E-4
0050707	regulation of cytokine secretion	7	3	2.70E-4
0071224	cellular response to peptidoglycan	2	2	4.09E-4
0050729	positive regulation of inflammatory response	19	4	4.99E-4
0044130	negative regulation of growth of symbiont in host	11	3	1.20E-3
0002830	positive regulation of type 2 immune response	3	2	1.21E-3
0045356	positive regulation of interferon-alpha biosynthetic process	3	2	1.21E-3
0042773	ATP synthesis coupled electron transport	4	2	2.39E-3
0048016	inositol phosphate-mediated signaling	4	2	2.39E-3
0050766	positive regulation of phagocytosis	14	3	2.53E-3
0071260	cellular response to mechanical stimulus	30	4	2.97E-3
0001867	complement activation, lectin pathway	5	2	3.93E-3
0002024	diet induced thermogenesis	5	2	3.93E-3
0042116	macrophage activation	5	2	3.93E-3
0045080	positive regulation of chemokine biosynthetic process	5	2	3.93E-3
0045359	positive regulation of interferon-beta biosynthetic process	5	2	3.93E-3
0006955	immune response	101	7	4.35E-3
0051092	positive regulation of NF-kappaB transcription factor activity	55	5	4.97E-3
0002925	positive regulation of humoral immune response mediated by circulating immunog...	6	2	5.82E-3
0006120	mitochondrial electron transport, NADH to ubiquinone	6	2	5.82E-3
0051607	defense response to virus	58	5	6.24E-3
0019221	cytokine-mediated signaling pathway	19	3	6.26E-3
0032760	positive regulation of tumor necrosis factor production	19	3	6.26E-3
0050830	defense response to Gram-positive bacterium	19	3	6.26E-3
0006954	inflammatory response	86	6	7.88E-3
0002504	antigen processing and presentation of peptide or polysaccharide antigen via MH...	7	2	8.04E-3
0045078	positive regulation of interferon-gamma biosynthetic process	7	2	8.04E-3
0048246	macrophage chemotaxis	7	2	8.04E-3
0050718	positive regulation of interleukin-1 beta secretion	7	2	8.04E-3
0050730	regulation of peptidyl-tyrosine phosphorylation	7	2	8.04E-3
0006465	signal peptide processing	8	2	0.01
0008306	associative learning	8	2	0.01
0003214	cardiac left ventricle morphogenesis	9	2	0.01
0007338	single fertilization	9	2	0.01
0018149	peptide cross-linking	9	2	0.01
0051894	positive regulation of focal adhesion assembly	9	2	0.01
0008203	cholesterol metabolic process	25	3	0.01
0021987	cerebral cortex development	25	3	0.01
0032695	negative regulation of interleukin-12 production	10	2	0.02
0033198	response to ATP	10	2	0.02
0030198	extracellular matrix organization	51	4	0.02
0042523	positive regulation of tyrosine phosphorylation of Stat5 protein	11	2	0.02
0048013	ephrin receptor signaling pathway	11	2	0.02
0002023	reduction of food intake in response to dietary excess	1	1	0.02
0002238	response to molecule of fungal origin	1	1	0.02
0002369	T cell cytokine production	1	1	0.02

Figure 26.2: *The results of the analysis.*

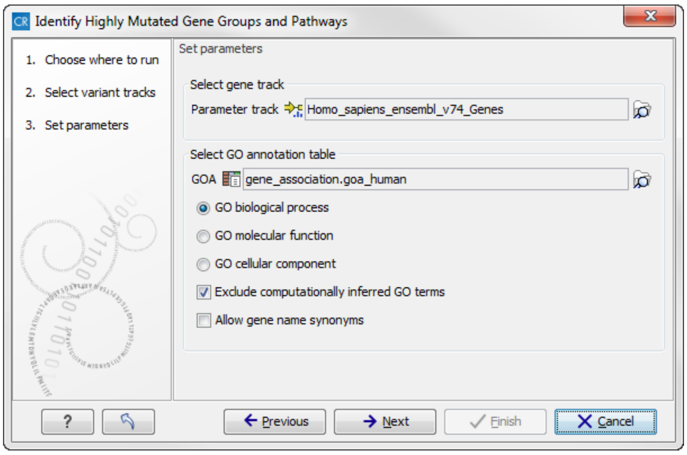
26.2 Identify highly mutated gene groups and pathways

This tool can be used to investigate candidate variants or better their corresponding altered genes for a common functional role. For example if you would like to compare different cancer patients to check whether e.g. the same pathways are affected in different individuals, you can use this tool. For this, you need a GO association file, which includes gene names and associated Gene Ontology terms. A GO association file with the top-level GO terms annotated (GO slim) is provided with the CLC Cancer Research Workbench and can be downloaded using the **Data Management** ( ) function found in the top right corner of the Workbench (see section 10.1.4).

To run the analysis go to the toolbox:

> **Toolbox** | **Identify Candidate Genes** ( ) | **Identify highly mutated gene groups and pathways** ( )

When you run the Identify highly mutated gene groups and pathways analysis, you have to specify both the annotation association file, a gene track, and finally which ontology (cellular component, biological process or molecular function) you would like to test for (see figure 26.3).


Identify Highly Mutated Gene Groups and Pathways

1. Choose where to run
2. Select variant tracks
3. Set parameters

Set parameters

Select gene track

Parameter track: Homo_sapiens_ensembl_v74_Genes

Select GO annotation table

GOA: gene_association.goa_human

● GO biological process
○ GO molecular function
○ GO cellular component
☑ Exclude computationally inferred GO terms
□ Allow gene name synonyms

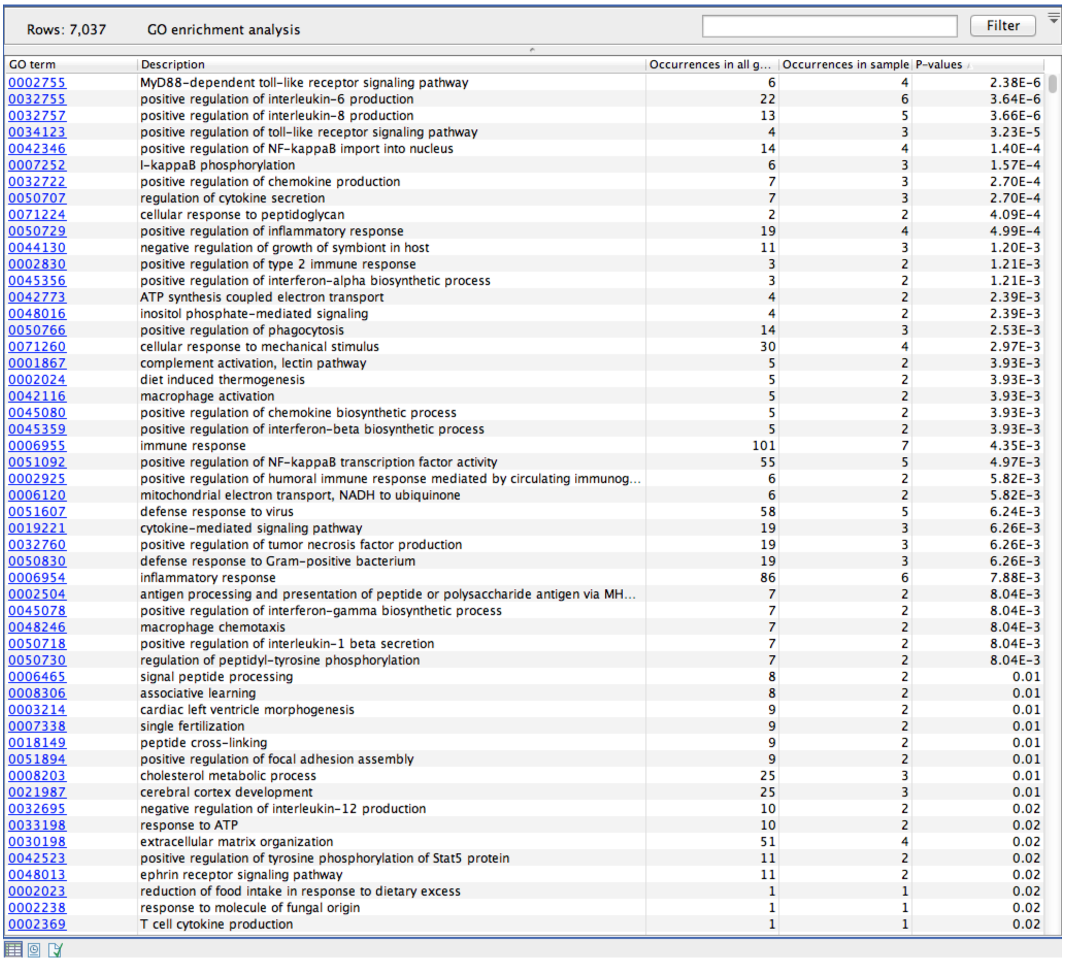
Figure 26.3: *Select gene track, GO annotation table, and ontology.*

The analysis starts by associating all of the variants from the input variant file with genes in the gene track, based on overlap with the gene annotations. A variant track can be created with the *CLC Cancer Research Workbench* variant callers (section 32.1, 20.17 and 32.2).

Next, the Workbench tries to match gene names from the gene (annotation) track with the gene names in the GO association file. Please be aware that the same gene name definition should be used in both files.

Based on this, the Workbench finds GO terms that are over-represented in the list. A hypergeometric test is used to identify over-represented GO terms by testing whether some of the GO terms are over-represented in a given gene set, compared to a randomly selected set of genes.

The result is a table with GO terms and the calculated p-value for the candidate variants, and a new variant file with annotated GO terms and the corresponding p-value (see figure 26.4). The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, or in other words how significant (trustworthy) a result is. In case of a small p-value the probability of achieving the same result by chance with the same test statistic is very small.



Rows: 7,037 GO enrichment analysis

GO term	Description	Occurrences in all g...	Occurrences in sample	P-values
0002755	MyD88-dependent toll-like receptor signaling pathway	6	4	2.38E-6
0032755	positive regulation of interleukin-6 production	22	6	3.64E-6
0032757	positive regulation of interleukin-8 production	13	5	3.66E-6
0034123	positive regulation of toll-like receptor signaling pathway	4	3	3.23E-5
0042346	positive regulation of NF-kappaB import into nucleus	14	4	1.40E-4
0007252	I-kappaB phosphorylation	6	3	1.57E-4
0032722	positive regulation of chemokine production	7	3	2.70E-4
0050707	regulation of cytokine secretion	7	3	2.70E-4
0071224	cellular response to peptidoglycan	2	2	4.09E-4
0050729	positive regulation of inflammatory response	19	4	4.99E-4
0044130	negative regulation of growth of symbiont in host	11	3	1.20E-3
0002830	positive regulation of type 2 immune response	3	2	1.21E-3
0045356	positive regulation of interferon-alpha biosynthetic process	3	2	1.21E-3
0042773	ATP synthesis coupled electron transport	4	2	2.39E-3
0048016	inositol phosphate-mediated signaling	4	2	2.39E-3
0050766	positive regulation of phagocytosis	14	3	2.53E-3
0071260	cellular response to mechanical stimulus	30	4	2.97E-3
0001867	complement activation, lectin pathway	5	2	3.93E-3
0002024	diet induced thermogenesis	5	2	3.93E-3
0042116	macrophage activation	5	2	3.93E-3
0045080	positive regulation of chemokine biosynthetic process	5	2	3.93E-3
0045359	positive regulation of interferon-beta biosynthetic process	5	2	3.93E-3
0006955	immune response	101	7	4.35E-3
0051092	positive regulation of NF-kappaB transcription factor activity	55	5	4.97E-3
0002925	positive regulation of humoral immune response mediated by circulating immunog...	6	2	5.82E-3
0006120	mitochondrial electron transport, NADH to ubiquinone	6	2	5.82E-3
0051607	defense response to virus	58	5	6.24E-3
0019221	cytokine-mediated signaling pathway	19	3	6.26E-3
0032760	positive regulation of tumor necrosis factor production	19	3	6.26E-3
0050830	defense response to Gram-positive bacterium	19	3	6.26E-3
0006954	inflammatory response	86	6	7.88E-3
0002504	antigen processing and presentation of peptide or polysaccharide antigen via MH...	7	2	8.04E-3
0045078	positive regulation of interferon-gamma biosynthetic process	7	2	8.04E-3
0048246	macrophage chemotaxis	7	2	8.04E-3
0050718	positive regulation of interleukin-1 beta secretion	7	2	8.04E-3
0050730	regulation of peptidyl-tyrosine phosphorylation	7	2	8.04E-3
0006465	signal peptide processing	8	2	0.01
0008306	associative learning	8	2	0.01
0003214	cardiac left ventricle morphogenesis	9	2	0.01
0007338	single fertilization	9	2	0.01
0018149	peptide cross-linking	9	2	0.01
0051894	positive regulation of focal adhesion assembly	9	2	0.01
0008203	cholesterol metabolic process	25	3	0.01
0021987	cerebral cortex development	25	3	0.01
0032695	negative regulation of interleukin-12 production	10	2	0.02
0033198	response to ATP	10	2	0.02
0030198	extracellular matrix organization	51	4	0.02
0042523	positive regulation of tyrosine phosphorylation of Stat5 protein	11	2	0.02
0048013	ephrin receptor signaling pathway	11	2	0.02
0002023	reduction of food intake in response to dietary excess	1	1	0.02
0002238	response to molecule of fungal origin	1	1	0.02
0002369	T cell cytokine production	1	1	0.02

Figure 26.4: *The results of the analysis.*

26.3 Identify mutated genes

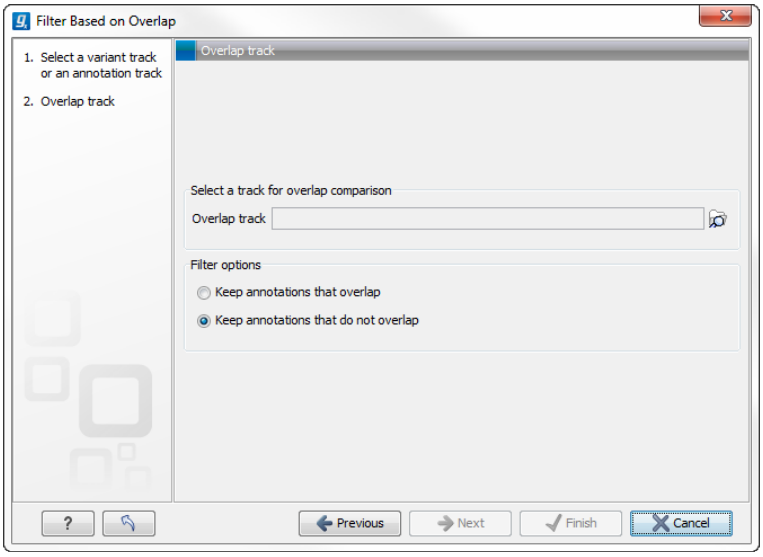
The overlap filter will be used for filtering an annotation track based on an overlap with another annotation track. This can be used to e.g. only show variants that fall within genes or regulatory regions or for restricting variants results to only cover a subset of genes as explained in section 26.4. Please note that for comparing variant tracks, more specific filters should be used (see section 21).

If you are just interested in finding out whether one particular position overlaps any of the annotations, you can use the advanced table filter and filter on the region column (track tables are described in section 17.2.3).

Toolbox | **Identify Candidate Genes** () | **Identify Mutated Genes** ()

Select the track you wish to filter and click **Next** to specify the track of overlapping annotations (see figure 26.5).

Next, select the track that should be used for comparison and tick whether you wish to keep annotations that overlap, or whether to keep annotations that do not overlap with the track selected. An overlap has a simple definition – if the annotation used as input has at least one shared position with the other track, there is an overlap. The boundaries of the annotations do not need to match.



Select a track for overlap comparison

Overlap track

Filter options

○ Keep annotations that overlap

● Keep annotations that do not overlap

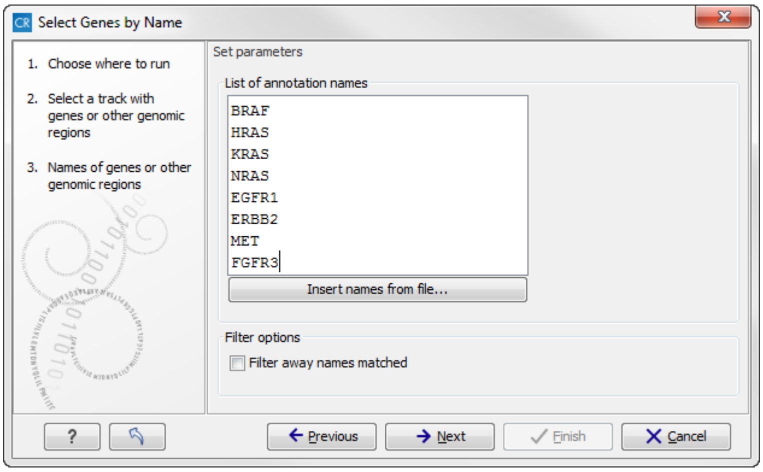
Figure 26.5: *Select overlapping annotations track.*

26.4 Select genes by name

The name filter allows you to use a list of names as input to create a new track only with these names. This is useful if you wish to filter your variants so that only those within certain genes are reported.

Toolbox | **Identify Candidate Genes** ( ) | **Select Genes By Name** ( )

Select the track you wish to filter and click **Next**.


Select Genes by Name
1. Choose where to run
2. Select a track with genes or other genomic regions
3. Names of genes or other genomic regions
Set parameters
Insert names from file...
Previous
Next
Finish
Cancel

List of annotation names

```
BRAF
HRAS
KRAS
NRAS
EGFR1
ERBB2
MET
FGFR3
```

Filter options

□ Filter away names matched

Figure 26.6: *Specify names for filtering.*

As shown in figure 26.6, you can specify a list of annotation names. Each name should be on a separate line.

In the bottom part of the wizard you can choose whether you wish to keep the annotations that are found, or whether you wish to exclude them. In the use case described above a track was created with only those annotations being kept that matched the specified names. Sometimes the other option may be useful, for example if you wish to screen certain categories of genes from the analysis (for example excluding all cancer genes to reduce the risk of coincidental findings when analyzing patient samples).

Part VIII

Transcriptomic analysis

Chapter 27

Transcriptomics tools

Contents

27.1 RNA-Seq analysis

Two tools are available for RNA-seq analysis, the tool **RNA-Seq Analysis** and the tool **Create Fold Change Track** Based on an annotated reference genome, the *CLC Cancer Research Workbench* supports RNA-Seq analysis by mapping next-generation sequencing reads and counting and distributing the reads across genes and transcripts. Subsequently, the results can be used for expression analysis using the tools in the **Transcriptomics Analysis** toolbox.

The tool for RNA-Seq analysis can be found here:

Toolbox | Transcriptomics Analysis () | **RNA-Seq Analysis ()**

The approach taken by the *CLC Cancer Research Workbench* is based on [Mortazavi et al., 2008].

The following describes the overall process of the RNA-Seq analysis when using an annotated eukaryote genome. See section 27.1.1 for more information on other types of reference data.

The RNA-Seq analysis is done in several steps: First, all genes are extracted from the reference genome (using a *gene* track). Next, all annotated transcripts are extracted (using an *mRNA* track). If there are several annotated splice variants, they are all extracted.

An example is shown in figure 27.1.

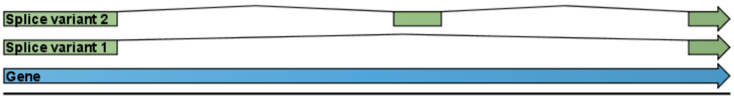


Figure 27.1: *A simple gene with three exons and two splice variants.*

This is a simple gene with three exons and two splice variants. The transcripts are extracted as shown in figure 27.2.

Next, the reads are mapped against all the transcripts plus the entire gene (see figure 27.3) and optionally to the whole genome.

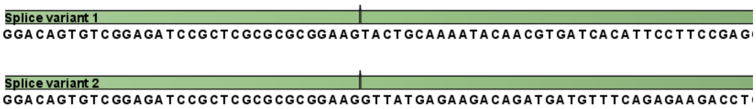

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT

Figure 27.2: *All the exon-exon junctions are joined in the extracted transcript.*

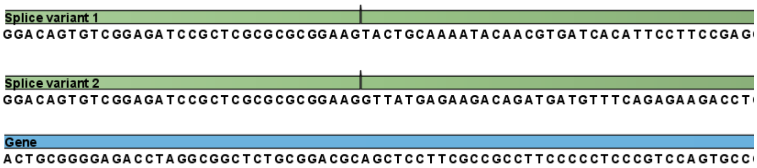

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT
Gene
ACTGCGGGGAGACCTAGGCGGCTCTGCGGACGCAGCTCCTTCGCCGCCTTCCCCCTCCCGTCCAGTGCC

Figure 27.3: *The reference for mapping: all the exon-exon junctions and the gene.*

From this mapping, the reads are categorized and assigned to the genes (elaborated later in this section), and expression values for each gene and each transcript are calculated.

Details on the process are elaborated in the following sections, which describe how to run RNA-seq analyses.

27.1.1 Specifying reads and reference

To start the RNA-Seq analysis, go to:

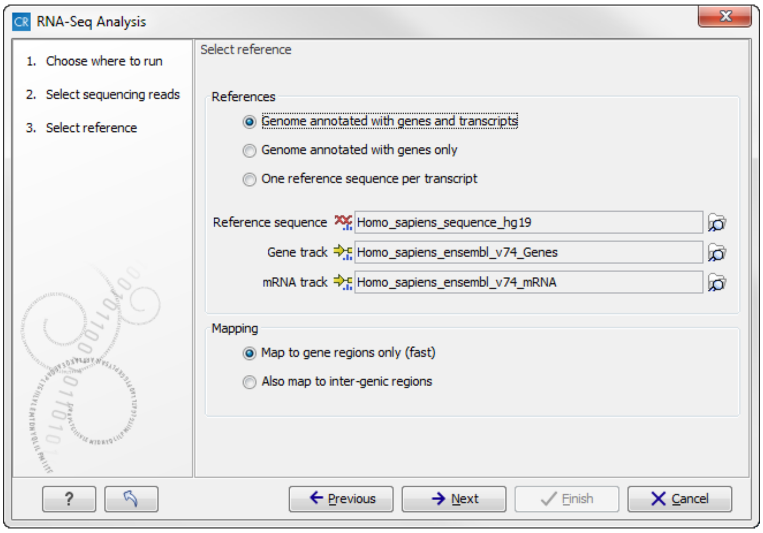
Toolbox | Transcriptomics Analysis ( ) | RNA-Seq Analysis ( )

This opens a dialog where you select the **sequencing reads**. Note that you need to import the sequencing data into the Workbench before it can be used for analysis. Importing read data is described in section 6.3.

If you have several samples that you wish to analyze independently and compare afterwards, you can run the analysis in batch mode (see section 8.1).

Click **Next** when the sequencing data are listed in the right-hand side of the dialog.

You are now presented with the dialog shown in figure 27.4.


RNA-Seq Analysis

1. Choose where to run
2. Select sequencing reads
3. Select reference

Select reference

References

● Genome annotated with genes and transcripts
○ Genome annotated with genes only
○ One reference sequence per transcript

Reference sequence: Homo_sapiens_sequence_hg19
Gene track: Homo_sapiens_ensembl_v74_Genes
mRNA track: Homo_sapiens_ensembl_v74_mRNA

Mapping

● Map to gene regions only (fast)
○ Also map to inter-genic regions

Figure 27.4: *Defining a reference genome for RNA-Seq.*

At the top, there are three options concerning how the reference sequences are annotated.

- **Genome annotated with genes and transcripts**. This option is the only option where splicing is taken into account. When this option is selected, both a *Gene* and an *mRNA* track should be provided in the boxes below. The mRNA annotations are used to define how the transcripts are spliced (as shown in figure 27.1). The reference sequence, gene, and mRNA tracks are provided with the CLC Cancer Research Workbench and can be downloaded using the **Data Management** () function found in the top right corner of the Workbench (see section 10.1.4).

- **Genome annotated with genes only**. This option should be used for in situations where you are not interested in transcript level expression. When this option is selected, a *Gene* track should be provided in the box below.

- **One reference sequence per transcript**. This option is suitable for situations where the reference is a list of sequences. Each sequence in the list will be treated as a "transcript" and expression values are calculated for each sequence. This option is most often used if the reference is a product of a *de novo* assembly of RNA-Seq data. When this option is selected, only the reference sequence should be provided, either as a sequence track or a sequence list.

At the bottom of the dialog you can choose the reference content to map to. Note that this is only relevant when using an annotated reference:

- **Map to gene regions only (fast)**. This option will ignore all inter-genic regions in the reference. Since only genes are considered, this options is also significantly faster than the alternative option. The effect of restricting the mapping to genes only is that any reads coming from genes or transcripts that are not part of the annotations will either be unmapped or map to another transcript with a similar sequence (e.g. a pseudo-gene). For poorly annotated references, it is possible to improve the annotations using the **Transcript Discovery** plugin which is freely available for download in the **Plugin Manager** (see section 1.7.1).

- **Also map to inter-genic regions**. This option will include the inter-genic regions as well. Please note that reads that map outside genes are counted as intergenic hits only and thus do not contribute to the expression values[1]. If a read maps equally well to a gene and to an inter-genic region, the read will be placed in the gene.

27.1.2 Tightly packed genes and genes in operons

For annotated references containing genes located very close to each other, as commonly found in bacteria, including operon structures, we recommend that the option "Map genes to gene regions only (fast)" is chosen. In this case, reads that map over a gene boundary will be counted towards the expression value for the gene they best map to.

With the mapping option "Also map to inter-genic regions" selected, only reads mapping completely within a genes boundaries will be counted towards the expression value for that gene. If any part of a read maps outside a given gene's boundaries then it will be considered intergenic and will not be counted towards the expression value. For tightly packed genes, especially in

[1] The reads will indirectly impact the RPKM expression values as they will be counted in the total number of mapped reads which is used to calculate RPKM (section 27.1.6)

cases where non-coding 5' regions are not included in the gene annotation, this can be too conservative.

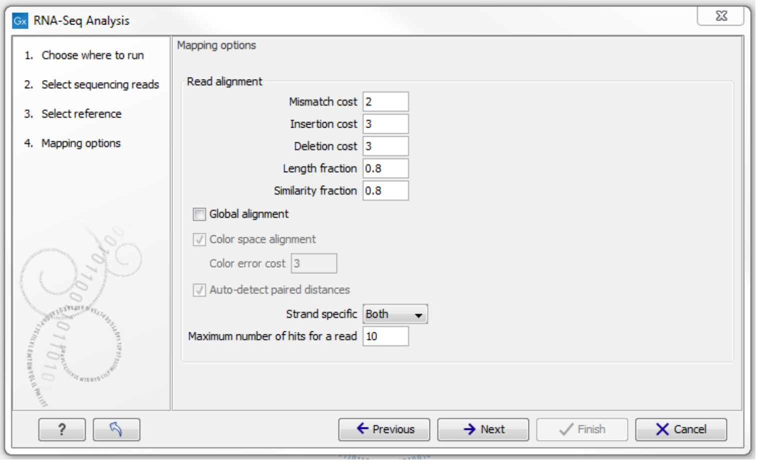
In both cases, reads are only mapped to, and thus counted towards the expression value of, one gene.

If differential expression is the main interest, then whether the intergenic mapping option is chosen or not may make little difference to the final results. However, if there are short genes, where the read length exceeds the gene length in some cases, then some granularity may be lost. That is, reads mapping to short genes might not be counted at all when the "Also map to inter-genic regions" is chosen.

Please also refer to the section entitled "Genes in Operons", which is found in section 27.1.4.

27.1.3 Defining mapping options for RNA-Seq

When the reference has been defined, click **Next** and you are presented with the dialog shown in figure 27.5.



RNA-Seq Analysis

1. Choose where to run
2. Select sequencing reads
3. Select reference
4. Mapping options

Mapping options

Read alignment

Mismatch cost 2
Insertion cost 3
Deletion cost 3
Length fraction 0.8
Similarity fraction 0.8

☐ Global alignment
☑ Color space alignment
Color error cost 3
☑ Auto-detect paired distances

Strand specific Both
Maximum number of hits for a read 10

Figure 27.5: *Defining mapping parameters for RNA-Seq.*

The mapping parameters are identical to those applying to **Map Reads to Reference**, as the underlying mapping is performed in the same way. For a description of the parameters, please see section 20.3.3.

For the estimation of paired reads distances, RNA-Seq uses the transcript level reference sequence information. This means that intronic regions are not included during this step - reflecting the true nature of the transcript on based on which the paired reads were produced.

In addition to the generic mapping parameters, two RNA-Seq specific parameters can be set:

- **Maximum number of hits for a read**. A read that matches equally well to more *distinct* places in the references than the 'Maximum number of hits for a read' specified will not be mapped (the notion of *distinct* places is elaborated below). If a read matches to multiple distinct places, but less than the specified maximum number, it will be randomly assigned to one of these places. The random distribution is done proportionally to the number of unique matches that the genes to which it matches have, normalized by the exon length (to ensure that genes with no unique matches have a chance of having multi-matches assigned to them, 1 will be used instead of 0, for their count of unique matches). This means that if there are 10 reads that match two different genes with equal exon length, the 10 reads will be distributed according to the number of unique matches for these two genes. The gene that has the highest number of unique matches will thus get a greater proportion of the 10 reads.

The definition of a *distinct* place in the references is complicated because each annotated transcript is extracted and used as reference for the read mapping (if the "Genome annotated with genes and transcripts" is selected in figure 27.4). To exemplify, consider a gene with 10 transcripts and 11 exons, where all transcripts have exon 1, and each of the 10 transcripts have only one of the exons 2 to 11. Exon 1 will be represented 11 times in the references (once for the gene region and once for each of the 10 transcripts). Reads that match to exon 1 will thus match to 11 of the extracted references. However, when the mappings are considered in the coordinates of the the main reference genome, it becomes evident that the 11 match places are not distinct but in fact identical. In this case this will just count as one distinct placement of the read, and it will *not* be discarded for exceeding the maximum number of hits limit. Similarly, when a multi-match read is randomly assigned to one of it's match places, each distinct place is considered only once.

The limit for how many non-specific matches a read is allowed to have, is applied first to the set of gene matches (if any), and then second to the intergenic matches. As an example using the default value of 10, if a read matches equally well 8 places within genes and 50 places in intergenic regions, it is still considered a valid match. It will only be discarded if the number of matches within genes is above the limit, or if there are no gene matches at all and the number of intergenic matches exceeds the limit.

Note that, although a read is mapped *distinctly* at the gene level, it does not necessarily map *uniquely* to a particular transcript of the gene. The above example with a gene with 10 transcripts and 11 exons, where all transcripts have exon 1, and each of the 10 transcripts have only one of the exons 2 to 11, is a good an easy to understand example of this: all reads that are mapped to exon 1 are *uniquely* mapped at the gene level but are *non-specific matches* at the transcript level. A more complicated example is that you may have a gene with transcript annotations where one transcript has a longer version of an exon than the other. In this case you may have reads that may either be mapped entirely within the long version of the exon, or across the exon-exon boundary of one of the transcripts with the short version of the exon. Such an example is provided by the gene 'Ftl1' in the RNA-seq Mouse Chromosome 7 Tutorial data. The gene and mRNA annotations for that gene are shown in Figure 27.6, along with the reads mapping to the gene.

When you zoom in on the regions at the end of the second exons and the beginning of the third exons (Figure 27.7) you see that the reference sequence is identical in the start of the part of the second exons that is only present in the long version, and in the start of the third exons (they share the sequence 'CTGCACA'). So a read that is e.g. '...TCATCTTGAGATGGCTTCTGCACA' may be either mapped entirely within the long version of the second exons, or across the exon-exon boundary of the short version of the second exon and the third exon. When it comes to reporting expression levels at the *transcript* level, reads are randomly assigned among the transcripts to which they map. Note that this introduces some randomness in the numbers of total exon reads for the *transcripts* (but not for the *gene*), even when you require that only specific matches are used. Also, as there is the chance that a read may sometimes be assigned to a transcript for which it is an exon-exon read, and sometimes to a transcript for which it is mapped entirely within an

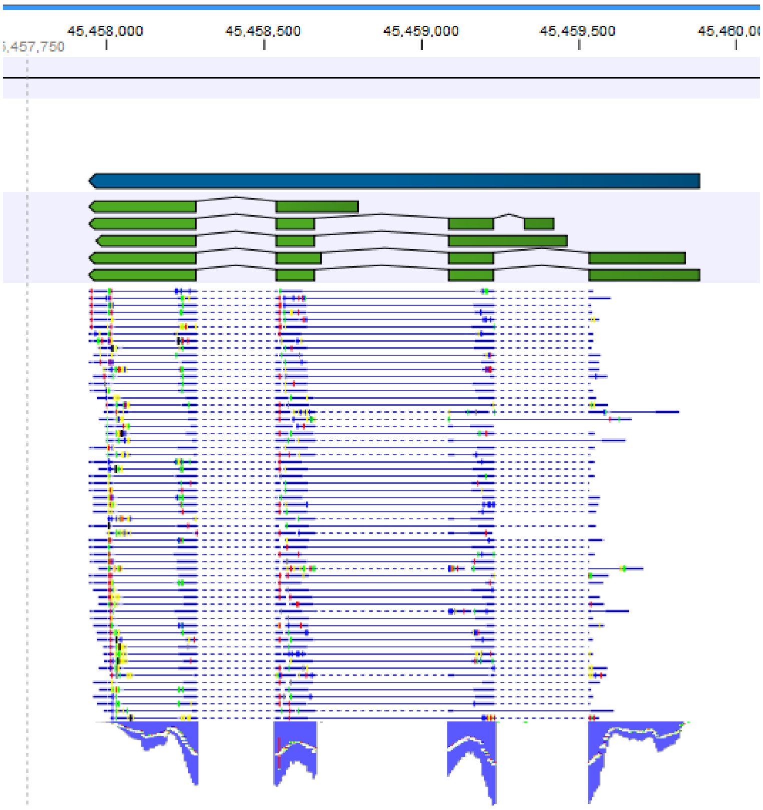


Figure 27.6: *The gene 'Ftl1' from the mouse chromosome 7 tutorial data.*

exon, *even* for a run with the **maximum number of hits for a read** parameter set to 1, there is a random component to the number of exon-exon reads reported, but not to the total number of exon reads.

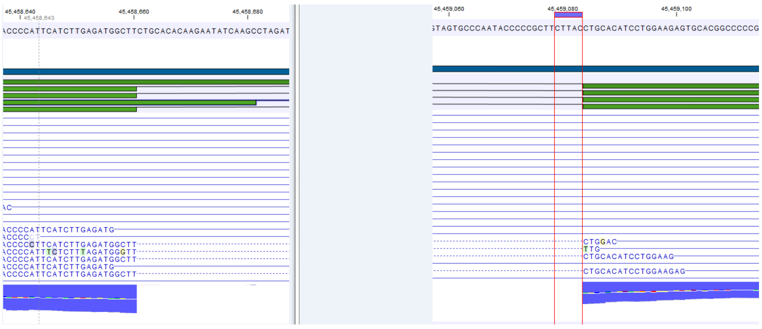


Figure 27.7: *The regions at the end of the second exons and the beginning of the third exons of the mRNA transcripts for the gene 'Ftl1'.*

- **Strand-specific alignment**. When this option is checked, the user can specify whether the reads should be mapped only in their forward (or reverse) orientation. This will typically be appropriate when a strand specific protocol for read generation has been used. It allows assignment of the reads to the right gene in cases where overlapping genes are located on different strands. Without the strand-specific protocol, this would not be possible (see [Parkhomchuk et al., 2009]). Also, applying the 'strand specific' 'reverse' option in an RNA-seq run could allow the user to assess the degree of antisense transcription.

27.1.4 Calculating expression values from RNA-Seq

When the reference has been defined, click **Next** and you are presented with the dialog shown in figure 27.8.

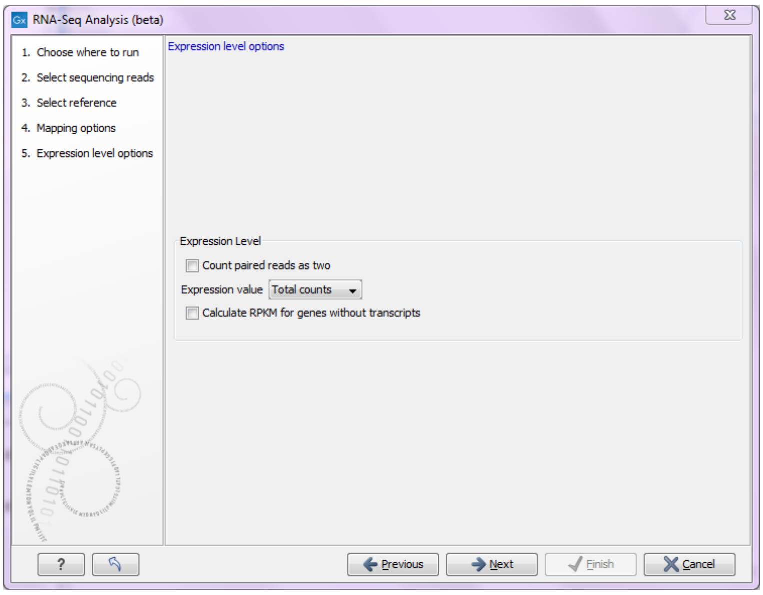

RNA-Seq Analysis (beta)
1. Choose where to run
2. Select sequencing reads
3. Select reference
4. Mapping options
5. Expression level options
Expression level options
Expression Level
☐ Count paired reads as two
Expression value Total counts
☐ Calculate RPKM for genes without transcripts
?
Previous
Next
Finish
Cancel

Figure 27.8: *Defining how expression values should be calculated.*

These parameters determine the way expression values are counted. Some background information on how paired reads are handled is useful before describing the parameters.

Paired reads in RNA-Seq
The *CLC Cancer Research Workbench* supports the direct use of paired data for RNA-Seq. A combination of single reads and paired reads can also be used. There are three major advantages of using paired data:

- Since the mapped reads span a larger portion of the reference, there will be fewer non-specifically mapped reads. This means that generally there is a greater accuracy in the expression values.
- This in turn means that there is a greater chance of accurately measuring the expression of transcript splice variants. As single reads (especially from the short reads platforms) typically only span one or two exons, many cases will occur where expression of splice variants sharing the same exons cannot be determined accurately. With paired reads, more combinations of exons will be identified as being unique for a particular splice variant.[2]
- It is possible to detect **Gene fusions** when one read in a pair maps in one gene and the other part maps in another gene. Several reads exhibiting the same pattern is supporting the presence of a fusion gene.

You can read more about how paired data are imported and handled in section 6.3.8.

When counting the mapped reads to generate expression values, the *CLC Cancer Research Workbench* needs to be told how to handle the counting of paired reads. The default behavior of

[2] Note that the *CLC Cancer Research Workbench* only calculates the expression of the transcripts already annotated on the reference.

the *CLC Cancer Research Workbench* is to count fragments rather than individual reads when two reads map as an intact pair. That is, an intact pair is given a count of one. Reads from a pair are considered part of a broken pair when the reads map outside the estimated pair distance, either map in wrong orientation or only one of the reads of the pair map. Neither member of a broken pair are counted when the default counting scheme is used. The reasoning is that when reads map as a broken pair, it is in indication that something is not right. For example, perhaps the transcripts are not represented correctly on the reference or there are errors in the data. In general, more confidence can be placed on an intact pair representing transcription within the sample. If a combination of paired and single reads are input into the analysis, then single reads that map are given a count of one. This is different from reads input into the analysis as part of a pair, but where their partner did not map.

In some situations it may be too strict to disregard broken pairs as is done using the default counting scheme. This could be the case where there is a high degree of variation in the sample compared to the reference or where the reference lacks comprehensive transcript annotations. By checking the **Count paired reads as two** option, you choose to count mapped 'reads' rather than mapped 'fragments. That means that, the two reads in an intact pair are each counted as one mapped read (so an intact pair contributes with a total count of two), and mapped members of broken pairs will each get given a count of one. Single mapped reads are also given a count of one. Note that this approach does not represent the abundance of fragments being sequenced correctly, since the two reads of a pair derive from the same fragment, whereas a fragment sequenced with single reads only give rise to one read.

When looking at the mappings, reads from broken pairs have a darker color than reads that are intact pairs or originally single reads.

Expression value
The expression values are created on two levels as two separate result files: one for genes and one for transcripts (if the "Genome annotated with genes and transcripts" is selected in figure 27.4). The content of the result files is described in section 27.1.5.

The **Expression value** parameter describes how expression per gene or transcript can be defined in different ways on both levels:

- **Total counts**. When the reference is annotated with genes only, this value is the total number of reads mapped to the gene. For un-annotated references, this value is the total number of reads mapped to the reference sequence. For references annotated with transcripts and genes, the value reported for each gene is the number of reads that map to the exons of that gene. The value reported per transcript is the total number of reads mapped to the transcript.

- **Unique counts**. This is similar to the above, except only reads that are non-specifically mapped are counted (read more about the distribution of non-specific matches in section 27.1.3).

- **RPKM**. This is a normalized form of the "Total counts" option (see more in section 27.1.6).

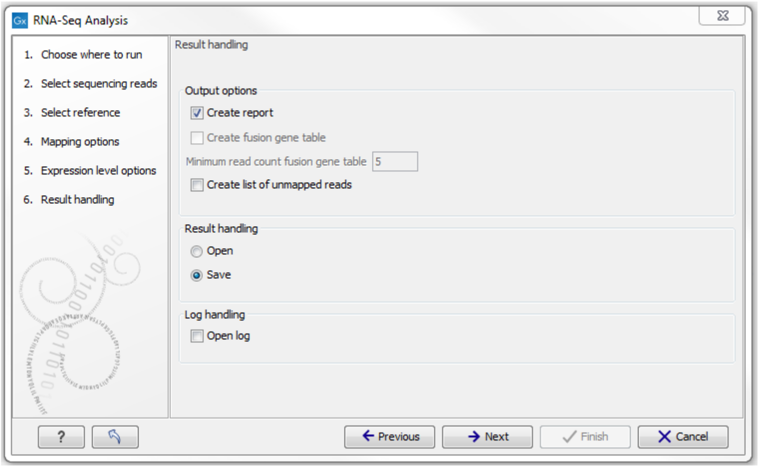
Please note that all values are present in the output. The **Expression value** in this dialog is solely used to inform the Workbench about which expression value should be applied when using the result in downstream analysis.

For genes without annotated transcripts, the RPKM cannot be calculated since the total length of all exons is needed. By checking the **Calculate RPKM for genes without transcripts**, the length of the gene will be used in place of an "exon length". If the option is not checked, there will be no RPKM value reported for those genes.

Genes in Operons
In the case of operons, where several genes are transcribed in the same mRNA transcript and are therefore located directly alongside each other, it is likely that some RNA-seq reads will map across the boundary of two different genes. If the mapping option "Also map to inter-genic regions" has been turned on, then reads like this will not be counted towards expression of any gene. If the option "Map to gene regions only (fast)" has been chosen, then reads will be counted towards only one gene within the operon: the one it mapped to best, which will usually mean the gene which the longest segment of the read mapped to. This means that the value for the expression of a particular operon in the RNA-seq results will be divided across the component genes of that operon.

27.1.5 Specifying RNA-Seq outputs

Clicking **Next** will allow you to specify the output options as shown in figure 27.9.



Result handling

Output options
☑ Create report
□ Create fusion gene table
Minimum read count fusion gene table 5
□ Create list of unmapped reads

Result handling
○ Open
● Save

Log handling
□ Open log

Figure 27.9: *Selecting the output of the RNA-Seq analysis.*

The main results of the RNA-Seq analysis are two expression tracks (one for gene-level and one for transcript-level expression) and a mapping track. In addition, the following optional results can be selected:

- **Create list of unmapped reads**. Creates a list of the reads that either did not map to the reference at all or that were non-specific matches with more placements than specified (see section 27.1.3).
- **Create report**. Creates a report of the results. See **RNA-Seq report** below for a description of the information contained in the report.
- **Create fusion gene table**. An option that is enabled when using paired data. Creates a table that lists potential fusion genes. This, along with the **Minimum read count**, is described further below in section **Gene fusion reporting**.

27.1.6 Interpreting the RNA-Seq analysis result

The main results of RNA-Seq analysis are:

- **Expression Tracks** One track summarizing expression at the gene level is produced. The track name ends in **(GE)**. If the "Genome annotated with genes and transcripts" option was selected, as shown in figure 27.4, then a second track summarizing expression at the transcript level is also produced. This track has a name ending with **(TE)**.
- **Reads track** This track contains the mapping of the reads to the references. This track has a name ending with **(Reads)**.

Other outputs, if these have been chosen when setting up the RNA-seq analysis are:

- **RNA-seq report**
- **Fusion gene table**
- **A sequence list of unmapped reads**

The rest of this page describes these output types in more detail.

Expression tracks

Both tracks can be shown in a **Table** (▦) and a **Graphical** (📊) view. By creating a **Track list**, the graphical view can be shown together with the read mapping track and tracks from other samples:

File | **New** | **Track List** (📊)

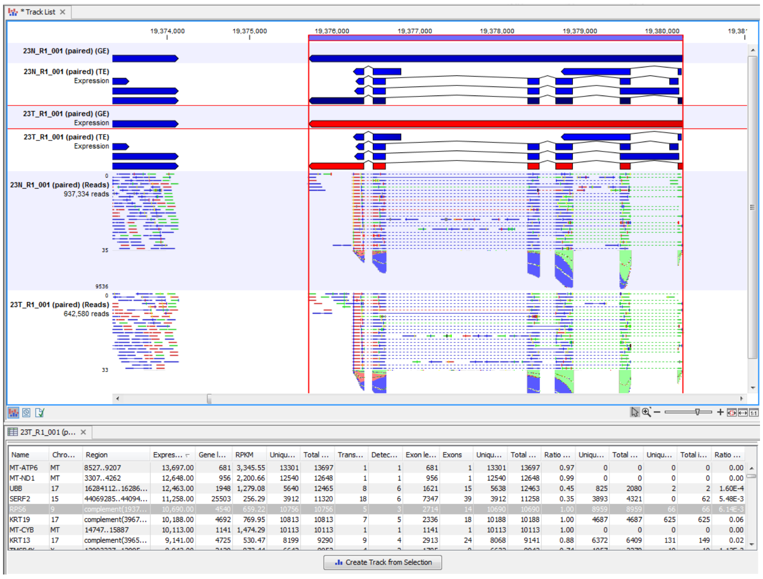
Select the mapping and expression tracks of the samples you wish to visualize together and select the annotation tracks used as reference for the RNA-Seq and click **Finish**.

Once the track list is shown, double-click the label of the expression track to show it in a table view. Clicking a row in the table makes the track list view jump to that location, allowing for quick inspection of interesting parts of the RNA-Seq read mapping (see an example in figure 27.10).

Reads spanning two exons are shown with a dashed line between each end as shown in figure 27.10, and the thin solid line represents the connection between two reads in a pair.

When doing comparative analysis and opening an experiment (see section 27.3) and a track list, clicking a row in the experiment will cause the track list to jump to the corresponding position, allowing for quick inspection of the reads underlying the counts in the experiment. Please note that at least one of the expression tracks used in the experiment have to be included in the track list in order for the link between the two to work.

Expression tracks can also be used to annotate variants using the **Annotate with Overlap Information** tool. Select the variant track as input and annotate with the expression track. For variants inside genes or transcripts, information will be added about expression (counts, expression value etc) from the gene or transcript in the expression track. Read more about the annotation tool in section 23.1.

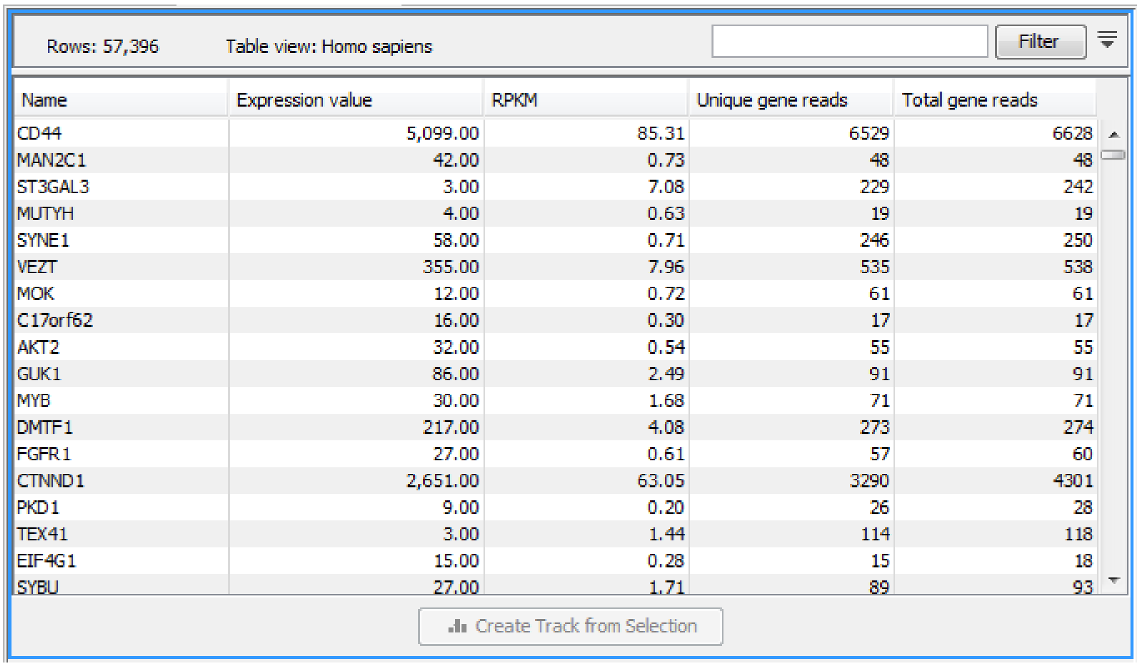


Name	Chro...	Region	Expres...	Gene l...	RPKM	Uniqu...	Total ...	Trans...	Detec...	Exon le...	Exons	Uniqu...	Total ...	Ratio ...	Uniqu...	Total ...	Uniqu...	Total i...	Ratio ...
MT-ATP6	MT	8527..9207	13,697.00	681	3,345.55	13301	13697	1	1	681	1	13301	13697	0.97	0	0	0	0	0.00
MT-ND1	MT	3307..4262	12,648.00	956	2,200.66	12540	12648	1	1	956	1	12540	12648	0.99	0	0	0	0	0.00
UBB	17	16284112..16286...	12,463.00	1948	1,279.08	5640	12465	8	6	1621	15	5638	12463	0.45	825	2080	2	2	1.60E-4
SERF2	15	44069285..44094...	11,258.00	25503	256.29	3912	11320	18	6	7347	39	3912	11258	0.35	3893	4321	0	62	5.48E-3
RPS6	9	complement(1937...	10,690.00	4540	659.22	10756	10756	5	3	2714	14	10690	10690	1.00	8959	8959	66	66	6.14E-3
KRT19	17	complement(3967...	10,188.00	4692	769.95	10813	10813	7	5	2336	18	10188	10188	1.00	4687	4687	625	625	0.06
MT-CYB	MT	14747..15887	10,113.00	1141	1,474.29	10113	10113	1	1	1141	1	10113	10113	1.00	0	0	0	0	0.00
KRT13	17	complement(3965...	9,141.00	4725	530.47	8199	9290	9	4	2913	24	8068	9141	0.88	6372	6409	131	149	0.02

Figure 27.10: *RNA-Seq results shown in a split view with an expression track at the bottom and a track list with read mappings of two samples at the top.*

Gene-level expression

The gene-level expression track holds information about counts and expression values for each gene. It can be opened in a **Table view** (▦) allowing sorting and filtering on all the information in the track (see figure 27.11 for an example subset of an expression track).



Name	Expression value	RPKM	Unique gene reads	Total gene reads
CD44	5,099.00	85.31	6529	6628
MAN2C1	42.00	0.73	48	48
ST3GAL3	3.00	7.08	229	242
MUTYH	4.00	0.63	19	19
SYNE1	58.00	0.71	246	250
VEZT	355.00	7.96	535	538
MOK	12.00	0.72	61	61
C17orf62	16.00	0.30	17	17
AKT2	32.00	0.54	55	55
GUK1	86.00	2.49	91	91
MYB	30.00	1.68	71	71
DMTF1	217.00	4.08	273	274
FGFR1	27.00	0.61	57	60
CTNND1	2,651.00	63.05	3290	4301
PKD1	9.00	0.20	26	28
TEX41	3.00	1.44	114	118
EIF4G1	15.00	0.28	15	18
SYBU	27.00	1.71	89	93

Figure 27.11: *A subset of a result of an RNA-Seq analysis on the gene level. Not all columns are shown in this figure*

Each row in the table corresponds to a gene (or reference sequence, if the **One reference sequence per transcript** option was used). The corresponding counts and other information is shown for each gene:

- **Name**. This is the name of the gene or the reference sequence, if the **One reference sequence per transcript** is used.

- **Chromosome and region**. The position of the gene on the genome.

- **Expression value**. This is based on the expression measure chosen as described in section 27.1.4.

- **Gene length** The length of the gene as annotated.

- **RPKM**. This is the expression value measured in RPKM [Mortazavi et al., 2008]: $RPKM = \frac{\textit{total exon reads}}{\textit{mapped reads(millions)} \times \textit{exon length (KB)}}$. See exact definition at end of subsection.

- **Unique gene reads**. This is the number of reads that match uniquely to the gene or its transcripts.

- **Total gene reads**. This is all the reads that are mapped to this gene — both reads that map uniquely to the gene or its transcripts and reads that matched to more positions in the reference (but fewer than the 'Maximum number of hits for a read' parameter) which were assigned to this gene.

- **Transcripts annotated**. The number of transcripts based on the mRNA annotations on the reference. Note that this is not based on the sequencing data - only on the annotations already on the reference sequence(s).

- **Detected transcripts**. The number of annotated transcripts to which reads have been assigned (see the description of transcript-level expression below).

- **Exon length**. The total length of all exons (not all transcripts).

- **Exons**. The total number of exons across all transcripts.

- **Unique exon reads**. The number of reads that match uniquely to the exons (including across exon-exon junctions).

- **Total exon reads**. Number of reads mapped to this gene that fall entirely within an exon or in exon-exon or exon-intron junctions. As for the 'Total gene reads' this includes both uniquely mapped reads and reads with multiple matches that were assigned to an exon of this gene.

- **Ratio of unique to total (exon reads)**. The ratio of the unique reads to the total number of reads in the exons. This can be convenient for filtering the results to exclude the ones where you have low confidence because of a relatively high number of non-unique exon reads.

- **Unique exon-exon reads**. Reads that uniquely match across an exon-exon junction of the gene (as specified in figure 27.10). The read is only counted once even though it covers several exons.

- **Total exon-exon reads**. Reads that match across an exon-exon junction of the gene (as specified in figure 27.10). As for the 'Total gene reads' this includes both uniquely mapped reads and reads with multiple matches that were assigned to an exon-exon junction of this gene.

- **Unique intron-exon reads**. Reads that uniquely map across an intron-exon boundary of the gene.
- **Total intron-exon reads**. Reads that maps across an intron-exon boundary of the gene.
- **Ratio of intron to total gene reads**. This can be convenient to identify genes with poor or lacking transcript annotations. If one or more exons are missing from the annotations, there will be a relatively high number of reads mapping in the intron.

Transcript-level expression
If the "Genome annotated with genes and transcripts" option is selected in figure 27.4, a transcript-level expression track is also generated.

The track can be opened in a **Table view** (▦) allowing sorting and filtering on all the information in the track. Each row in the table corresponds to an mRNA annotation in the mRNA track used as reference.

- **Name**. This is the name of the transcript, if the **One reference sequence per transcript** is used.
- **Chromosome and region**. The position of the gene on the genome.
- **Expression value**. This is based on the expression measure chosen as described in section 27.1.4.
- **RPKM**. This is the expression value measured in RPKM [Mortazavi et al., 2008]: $RPKM = \frac{\textit{total exon reads}}{\textit{mapped reads(millions)} \times \textit{exon length (KB)}}$. See exact definition at end of subsection.
- **Relative RPKM**. The RPKM for the transcript divided by the maximum of the RPKM values among all transcripts of the same gene. This value describes the relative expression of alternative transcripts for the gene.
- **Gene name**. The name of the corresponding gene.
- **Transcript length**. This is the length of the transcript.
- **Exons**. The total number of exons in the transcript.
- **Transcript ID**. The transcript ID is taken from the transcript_id note in the mRNA track annotations and can be used to differentiate between different transcripts of the same gene.
- **Transcripts annotated**. The number of transcripts based on the mRNA annotations on the reference. Note that this is not based on the sequencing data - only on the annotations already on the reference sequence(s).
- **Detected transcripts**. The number of annotated transcripts to which reads have been assigned.
- **Unique transcript reads**. This is the number of reads in the mapping for the gene that are uniquely assignable to the transcript. This number is calculated after the reads have been mapped and both single and multi-hit reads from the read mapping may be unique transcript reads.

- **Total transcript reads**. Once the 'Unique transcript read's have been identified and their counts calculated for each transcript, the remaining (non-unique) transcript reads are assigned randomly to one of the transcripts to which they match. The 'Total transcript reads' counts are the total number of reads that are assigned to the transcript once this random assignment has been done. As for the random assignment of reads among genes, the random assignment of reads within a gene but among transcripts, is done proportionally to the 'unique transcript counts' normalized by transcript length, that is, using the RPKM. Unique transcript counts of 0 are not replaced by 1 for this proportional assignment of non-unique reads among transcripts.
- **Ratio of unique to total (transcript reads)**. The ratio of the unique reads to the total number of reads in the transcripts. This can be convenient for filtering the results to exclude the ones where you have low confidence because of a relatively high number of non-unique transcript reads.

Reads track

A track containing the mapped reads is generated. Details about viewing and editing of reads-tracks are described in section **??**, and the resequencing is detailed in section **??**.

If you have chosen the strand specific option when setting up your analysis, it may be helpful to note that the colors of mapped single reads represent the orientation of the read relative to the reference provided. When a gene track is provided along with the reference genome, the reads will be mapped using the strand you specified, but the coloring of the read will be relative to the reference gene. If the reads matches the orientation of the gene it is coloured green, and if it is opposite to the orientation of the gene it is coloured red. A summary list of the colors to expect with different combinations of gene orientation and strand specific mapping options is:

- Strand specific, forward orientation chosen + gene on plus strand of reference = single reads colored green.
- Strand specific, forward orientation chosen + gene on minus strand of reference = single reads colored red.
- Strand specific, reverse orientation chosen + gene on plus strand of reference = single reads colored red.
- Strand specific, reverse orientation chosen + gene on minus strand of reference = single reads colored green.

See Figure 27.12 for an example of forward and reverse reads mapped to a gene on the plus strand.

Note: Reads mapping to intergenic regions (if the **Also map to inter-genic regions** box is selected) will not be mapped in a strand specific way.

Although paired reads are coloured blue, they can be viewed as red and green 'single' reads by selecting the **Disconnect paired reads** box, within the Read Mapping Settings bar on the right-hand side of the track.

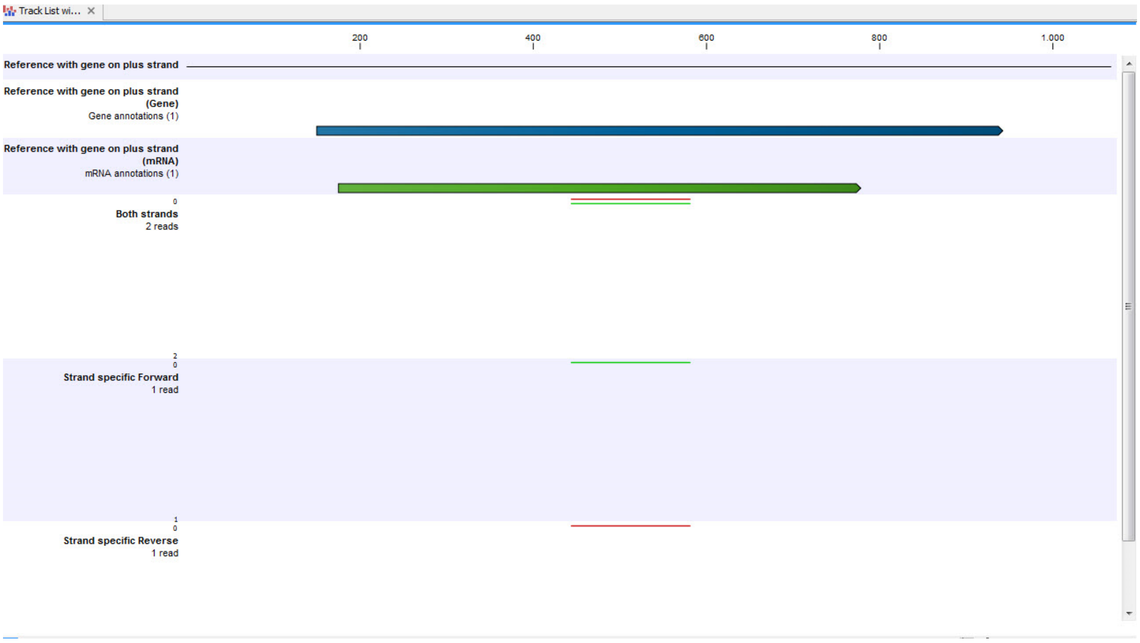


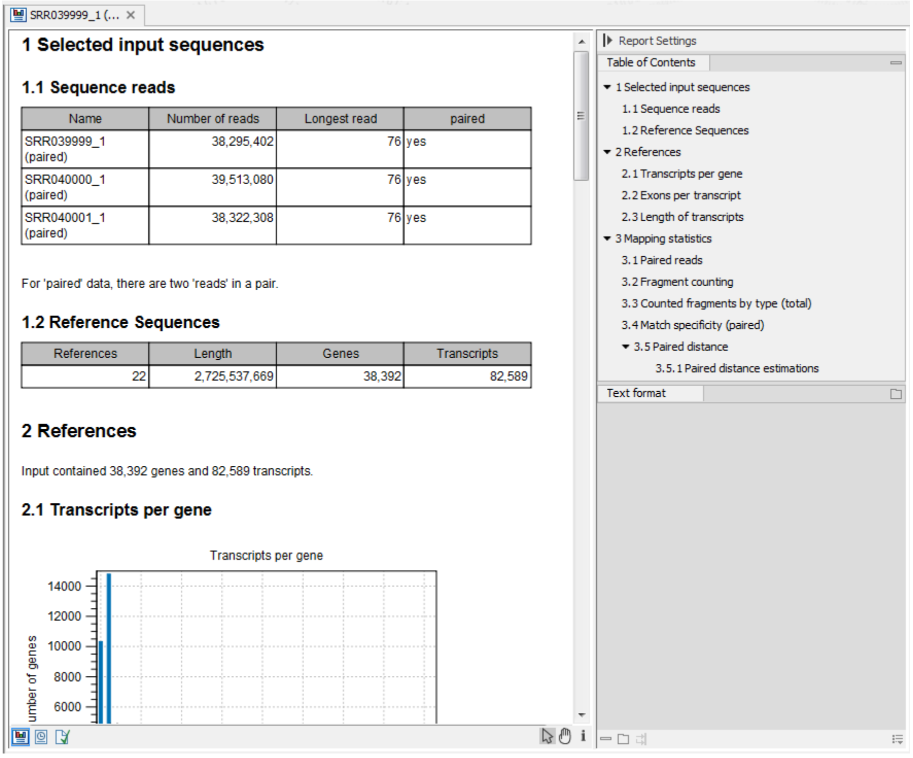
Figure 27.12: *A track list showing a gene and transcript on the plus strand, and various mapping results. The first reads track shows a mapping of two reads (one 'forward' and one 'reverse') using strand specific 'both' option. Both reads map successfully; the forward read coloured green (because it matches the direction of the gene), and the reverse read coloured red. The second reads track shows a mapping of the same reads using strand specific ('forward') option. The reverse read does not map because it is not in the correct direction, therefore only the green forward read is shown. The final reads track shows a mapping of the same reads again but using strand specific 'reverse' option. This time, the green forward read does not map because it is in the wrong direction, and only the red reverse read is shown.*

RNA-Seq report

An example of an RNA-seq report that is created if you choose the **Create report** option is shown in figure 27.13.

The report contains the following information:

- **Sequence reads**. Information about the number of reads.
- **Reference sequences**. Information about the reference sequences used and their lengths.
- **Reference**. Information about the total number of genes and transcripts (for eukaryotes only) found in the reference.
- **Transcripts per gene**. A graph showing the number of transcripts per gene. For eukaryotes, this will be equivalent to the number of mRNA annotations per gene annotation.
- **Exons per transcript**. A graph showing the number of exons per transcript.
- **Length of transcripts**. A graph showing the distribution of transcript lengths.
- **Mapping statistics**. Shows statistics on:



1 Selected input sequences

1.1 Sequence reads

Name	Number of reads	Longest read	paired
SRR039999_1 (paired)	38,295,402	76	yes
SRR040000_1 (paired)	39,513,080	76	yes
SRR040001_1 (paired)	38,322,308	76	yes

For 'paired' data, there are two 'reads' in a pair.

1.2 Reference Sequences

References	Length	Genes	Transcripts
22	2,725,537,669	38,392	82,589

2 References

Input contained 38,392 genes and 82,589 transcripts.

2.1 Transcripts per gene

Figure 27.13: *Report of an RNA-Seq run.*

- **Paired reads**. (Only included if paired reads are used). Shows the number of reads mapped in pairs, the number of reads in broken pairs and the number of unmapped reads.
- **Fragment counting**. Lists the total number of fragments used for calculating expression, divided into uniquely and non-specifically mapped reads (see the point below on match specificity for details).
- **Counted fragments by type**. Divides the fragments that are counted into different types
  - **Exon**. Reads that map completely within an exon
  - **Exon-exon reads**. Reads that map across an exon junction as specified in figure 27.10.
  - **Total exon reads**. Number of reads that fall entirely within an exon or in an exon-exon junction.
  - **Intron**. Reads that fall partly or entirely within an intron.
  - **Total gene reads**. All reads that map to the gene.
  - **Intergenic**. All reads that map partly or entirely between genes (will only be shown if the **Also map to inter-genic regions** option is used).

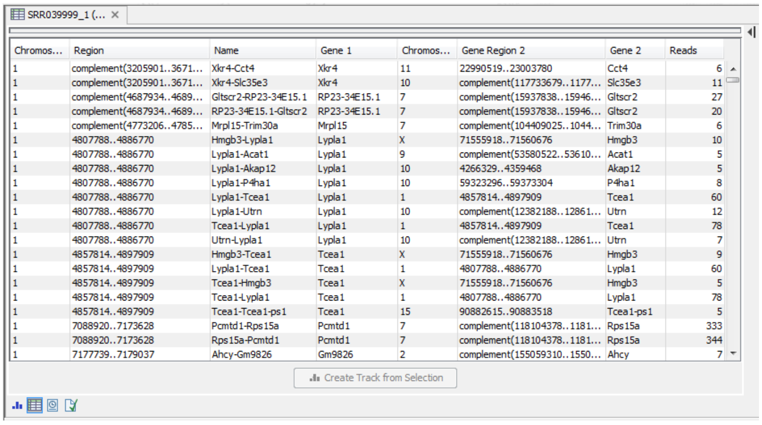
- **Match specificity**. Shows a graph of the number of match positions for the reads. Most reads will be mapped 0 or 1 time, but there will also be reads matching more than once in the reference. The maximum number of match positions is limited in the **Maximum number of hits for a read** setting in figure 27.4. Note that the number of reads that are mapped 0 times includes both the number of reads that cannot be mapped at all and the number of reads that matches to more than the **Maximum number of hits for a read** parameter.

- **Paired distance**. (Only included if paired reads are used). Shows a graph of the distance between mapped reads in pairs.

Note that the report can be exported in pdf or Excel format.

Gene fusion reporting

When using paired data, there is also an option to create an annotation track summarizing the evidence for gene fusions. An example is shown in figure 27.14.



SRR039999_1 (...

Chromos...	Region	Name	Gene 1	Chromos...	Gene Region 2	Gene 2	Reads
1	complement(3205901..3671...	Xkr4-Cct4	Xkr4	11	22990519..23003780	Cct4	6
1	complement(3205901..3671...	Xkr4-Slc35e3	Xkr4	10	complement(117733679..1177...	Slc35e3	11
1	complement(4687934..4689...	Gltscr2-RP23-34E15.1	RP23-34E15.1	7	complement(15937838..15946...	Gltscr2	27
1	complement(4687934..4689...	RP23-34E15.1-Gltscr2	RP23-34E15.1	7	complement(15937838..15946...	Gltscr2	20
1	complement(4773206..4785...	Mrpl15-Trim30a	Mrpl15	7	complement(104409025..1044...	Trim30a	6
1	4807788..4886770	Hmgb3-Lypla1	Lypla1	X	71555918..71560676	Hmgb3	10
1	4807788..4886770	Lypla1-Acat1	Lypla1	9	complement(53580522..53610...	Acat1	5
1	4807788..4886770	Lypla1-Akap12	Lypla1	10	4266329..4359468	Akap12	5
1	4807788..4886770	Lypla1-P4ha1	Lypla1	10	59323296..59373304	P4ha1	8
1	4807788..4886770	Lypla1-Tcea1	Lypla1	1	4857814..4897909	Tcea1	60
1	4807788..4886770	Lypla1-Utrn	Lypla1	10	complement(12382188..12861...	Utrn	12
1	4807788..4886770	Tcea1-Lypla1	Lypla1	1	4857814..4897909	Tcea1	78
1	4807788..4886770	Utrn-Lypla1	Lypla1	10	complement(12382188..12861...	Utrn	7
1	4857814..4897909	Hmgb3-Tcea1	Tcea1	X	71555918..71560676	Hmgb3	9
1	4857814..4897909	Lypla1-Tcea1	Tcea1	1	4807788..4886770	Lypla1	60
1	4857814..4897909	Tcea1-Hmgb3	Tcea1	X	71555918..71560676	Hmgb3	5
1	4857814..4897909	Tcea1-Lypla1	Tcea1	1	4807788..4886770	Lypla1	78
1	4857814..4897909	Tcea1-Tcea1-ps1	Tcea1	15	90882615..90883518	Tcea1-ps1	5
1	7088920..7173628	Pcmtd1-Rps15a	Pcmtd1	7	complement(118104378..1181...	Rps15a	333
1	7088920..7173628	Rps15a-Pcmtd1	Pcmtd1	7	complement(118104378..1181...	Rps15a	344
1	7177739..7179037	Ahcy-Gm9826	Gm9826	2	complement(155059310..1550...	Ahcy	7

Figure 27.14: *An example of a gene fusion table.*

Each row represents one gene where read pairs suggest it could be fused with another gene. This means that each fusion is represented by two rows.

The **Minimum read count** option in figure 27.9 is used to make sure that only combinations of genes supported by at least this number of read pairs are included. The default value is 5, which means that at least 5 pairs need to connect two genes in order to report it in the result.

The result table shows the following information for each row:

- **Name**. The name of the fusion (the two gene names combined).
- **Information per gene**. Gene name, chromosome and position are included for both genes.
- **Reads**. How many reads that are mapped across the two genes.

Note that the reporting of gene fusions is very simple and should be analyzed in much greater detail before any evidence of gene fusions can be verified. The table should be considered more of a pointer to genes to explore rather than evidence of gene fusions. Please note that you can include the fusion genes track in a track list together with the reads tracks to investigate the mapping patterns in greater detail:

File | New | Track List (📊)

Sequence list of unmapped reads

If you chose the option to **Create a list of unmapped reads**, then a sequence list of these will be produced. If you started with paired reads then more than one list of unmapped reads may be produced: paired reads are put in one list, with a name that ends in (paired), singe reads, including members of broken pairs, are put in a read list with a name than ends in (single).

Definition of RPKM
RPKM, Reads Per Kilobase of exon model per Million mapped reads, is defined in this way [Mortazavi et al., 2008]:

$$RPKM = \frac{\textit{total exon reads}}{\textit{mapped reads(millions)} \times \textit{exon length (KB)}}.$$

For prokaryotic genes and other non-exon based regions, the calculation is performed in this way:

$$RPKM = \frac{\textit{total gene reads}}{\textit{mapped reads(millions)} \times \textit{gene length (KB)}}.$$

Total exon reads This value can be found in the column with header **Total exon reads** in the expression track. This is the number of reads that have been mapped to exons (either within an exon or at the exon junction). When the reference genome is annotated with gene and transcript annotations, the mRNA track defines the exons, and the total exon reads are the reads mapped to all transcripts for that gene. When only genes are used, each gene in the gene track is considered an exon. When an un-annotated sequence list is used, each sequence is considered an exon.

Exon length This is the number in the column with the header **Exon length** in the expression track, divided by 1000. This is calculated as the sum of the lengths of all exons (see definition of exon above). Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region.

Mapped reads The sum of all mapped reads as listed in the RNA-Seq analysis report. If paired reads were used in the mapping, mapped fragments are counted here instead of reads, unless the **Count paired reads as two** option was selected. For more information on how expression is calculated in this case, see section 27.1.4. Please note that the option to **Map to gene regions only** will affect the number of mapped reads, since all intergenic reads will not be mapped if this option is selected. This means that comparison of RPKM values between samples should only be carried out if this parameter was set in the same way for all samples.

27.1.7 Create fold change track

The **Create Fold Change Track** tool can be used to compare RNA expression levels in two samples (e.g. matched tumor and normal samples).

After RNA-seq analysis, the resulting expression tracks can be compared using the **Create Fold Change Track** tool. For each gene or transcript, this tool will compute the ratio of the expression value in the case track to the expression value of the same gene in the control track. The tool enables filtering on the basis of fold-changes as well as expression values. This can give a quick first indication of possible differentially expressed genes or transcripts between two RNA-seq samples. For a more detailed and statistically robust analysis, we recommend running a Statistical Analysis tool in the *CLC Cancer Research Workbench*.

The tool for creating a fold-change track can be found here:

Toolbox | Transcriptomics Analysis ( ) | RNA-Seq Analysis ( ) | Create Fold Change Track ( )

After starting the Create Fold Change Track tool, select the expression track corresponding to the case sample in the input dialog, as shown in figure 27.15, and click on the button labeled **Next**.

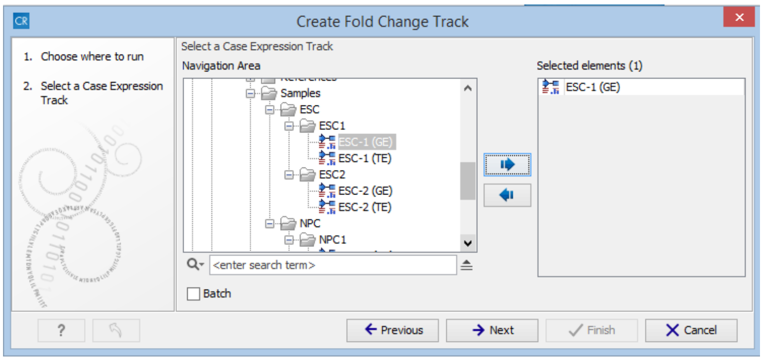
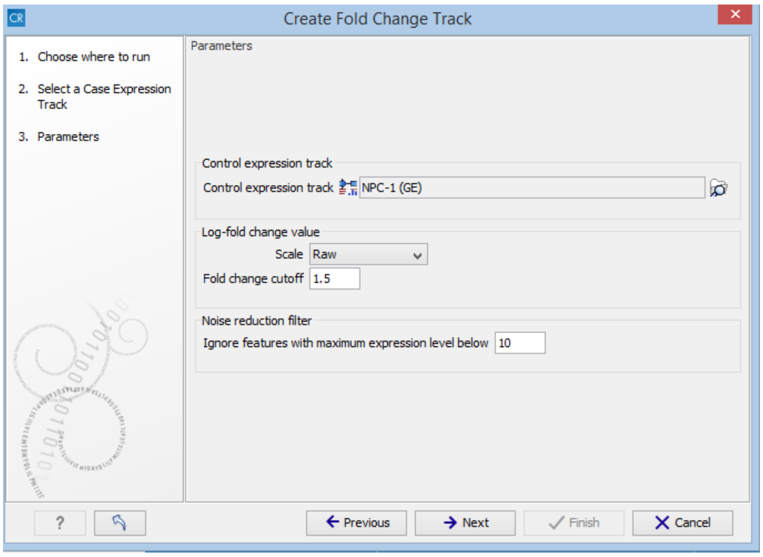


Figure 27.15: *Choosing the case expression track for the Create Fold Change Track tool.*

Setting the parameters

You are now presented with the wizard step shown in figure 27.16 where you can set the following input parameters:



Control expression track

Control expression track: NPC-1 (GE)

Log-fold change value

Scale: Raw

Fold change cutoff: 1.5

Noise reduction filter

Ignore features with maximum expression level below: 10

Figure 27.16: *Setting the input parameters for the Create Fold Change Track tool.*

- **Control expression track** At the top of the wizard, a control expression track must be chosen. The control expression track will be used as the reference in creating the fold-change track.
- **Log fold change value** In the middle of the wizard, the parameters relating to the reported fold-change values can be specified.
  - *Scale* Determines on what scale the fold-changes should be reported and interpreted:
    - **Raw** The fold-changes will be reported as a direct ratio: $\text{case expression} = \frac{\text{case expression}}{\text{control expression}}$

- **Log-2** The fold-changes will be reported as log-2 values: case expression $= \log_2\left(\frac{\text{case expression}}{\text{control expression}}\right)$
- **Log-10** The fold-changes will be reported as log-10 values: case expression $= \log_{10}\left(\frac{\text{case expression}}{\text{control expression}}\right)$
- **Natural logarithm** The fold-changes will be reported as log-e values: case expression $= \log\left(\frac{\text{case expression}}{\text{control expression}}\right)$

- *Fold-change cutoff* Enables filtering in the results based on fold-changes. Note that the value entered into the fold-change cutoff field will be interpreted according to the scale specified in the *scale* parameter. For example, if the expression level of a gene is 100 in the case sample and 200 in a control sample, then a cutoff of 1.5 with the scale parameter set to **Raw** will not result in the gene being filtered on the basis of fold-changes, and the gene will appear in the final results. However, if the scale parameter is set to **Log10**, the gene will be filtered and will not appear in the final results. To include all genes or transcripts in the output without filtering on the basis of fold-changes, enter 0 in this field (this will work for any value specified for the *scale* parameter).

- **Noise reduction filter** A second level of filtering is possible based on absolute expression levels. The rationale behind this filtering is that seemingly very large fold-changes can occur by chance if the expression levels are very low in both samples, creating a false-positive noise in the resulting fold-change track. Therefore, it is desirable to require a certain level of expression for a gene in at least one of the samples. This can be specified using the *Ignore features with maximum expression level below* parameter.

  - *Ignore features with maximum expression level below* The value entered in this field corresponds to the minimum expression level required in *either* the case *or* the control track, in order for the gene or transcript to appear in the results. If the expression level does not exceed this value in either the case or the control sample, the gene or transcript will be filtered out from the final results. Entering a value of 0 for this parameter will not filter any genes or transcripts based on absolute expression levels.

When finished with setting the parameters, click **Next**.

Interpreting the results

The Create Fold Change Track tool will produce an annotation track, including the genes or transcripts that were not filtered out based on the filtering criteria. Each feature in this track is annotated with the following information:

- **Fold-change** Represents the fold-change according to the specified scale. A positive value indicates that the expression level was higher in the case. A negative value indicates that the expression level was higher in the control. If **Raw** was specified for the *scale* parameter, a value of 0 represents no difference. If a logarithm-based value was specified for the *scale* parameter, a value of 1 represents no difference.

- **Difference** Represents the difference in expressions. A positive value indicates that the expression level was higher in the case. A negative value indicates the expression level was higher in the control.

- **Maximum expression** Represents the larger of the expression values observed in the case and the control samples.
- **Expression (case)** Gives the expression value in the case sample
- **Expression (control)** Gives the expression value in the control sample
- **Fold-change** Represents the fold-change according to the specified scale. A positive value indicates that the expression level was higher in the case. A negative value indicates that the expression level was higher in the control. If **Raw** was specified for the *scale* parameter, a value of 0 represents no difference. If a logarithm-based value was specified for the *scale* parameter, a value of 1 represents no difference.
- **Difference** Represents the difference in expressions. A positive value indicates that the expression level was higher in the case. A negative value indicates the expression level was higher in the control.
- **Maximum expression** Represents the larger of the expression values observed in the case and the control samples.
- **Expression (case)** Gives the expression value in the case sample
- **Expression (control)** Gives the expression value in the control sample

As is the case with all annotation tracks in the *CLC Cancer Research Workbench*, it is possible to sort and filter the results based on any of the above values, and to create track lists for further analysis of the results.

27.2 Small RNA analysis

The small RNA analysis tools in *CLC Cancer Research Workbench* are designed to facilitate trimming of sequencing reads, counting and annotating of the resulting tags using miRBase or other annotation sources and performing expression analysis of the results. The tools are general and flexible enough to accommodate a variety of data sets and applications within small RNA profiling, including the counting and annotation of both microRNAs and other non-coding RNAs from any organism. Illumina, 454 and SOLiD sequencing platforms are supported. For SOLiD, adapter trimming and annotation is done in color space.

The annotation part is designed to make special use of the information in miRBase but more general references can be used as well.

There are generally two approaches to the analysis of microRNAs or other smallRNAs: (1) count the different types of small RNAs in the data and compare them to databases of microRNAs or other smallRNAs, or (2) map the small RNAs to an annotated reference genome and count the numbers of reads mapped to regions which have smallRNAs annotated. The approach taken by *CLC Cancer Research Workbench* is (1). This approach has the advantage that it does not require an annotated genome for mapping — you can use the sequences in miRBase or any other sequence list of smallRNAs of interest to annotate the small RNAs. In addition, small RNAs that would not have mapped to the genome (e.g. when lacking a high-quality reference genome or if the RNAs have not been transcribed from the host genome) can still be measured and their expression be compared. The methods and tools developed for *CLC Cancer Research Workbench* are inspired by the findings and methods described in [Creighton et al., 2009], [Wyman et al., 2009], [Morin et al., 2008] and [Stark et al., 2010].

In the following, the tools for working with small RNAs are described in detail. Look at the tutorials on http://www.clcbio.com/tutorials to see examples of analyzing specific data sets.

27.2.1 Extract and count

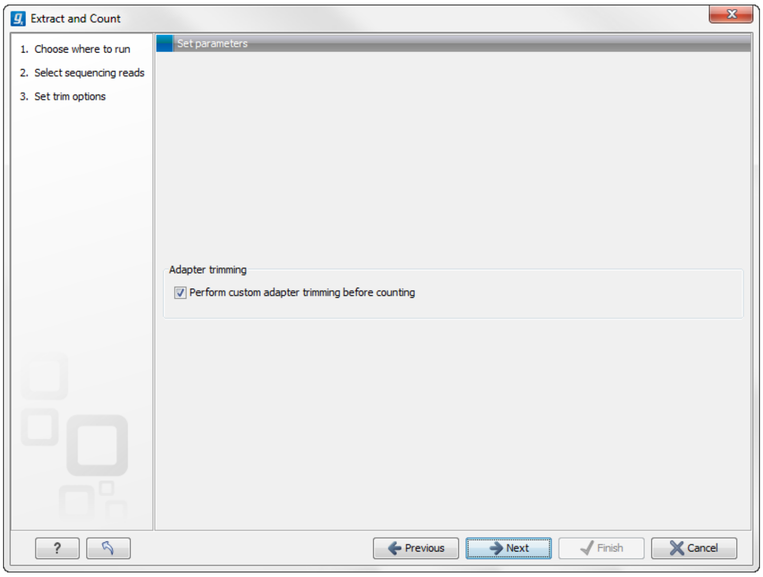
First step in the analysis is to import the data (see section 6.3).

The next step is to extract and count the small RNAs to create a *small RNA sample* that can be used for further analysis (either annotating or analyzing using the expression analysis tools):

> **Toolbox** | **Transcriptomics Analysis** ( ) | **Small RNA Analysis** ( ) | **Extract and Count** ( )

This will open a dialog where you select the sequencing reads that you have imported. Click **Next** when the sequencing data is listed in the right-hand side of the dialog. Note that if you have several samples, they should be processed separately.

This dialog (see figure 27.17) is where you specify whether the reads should be trimmed for adapter sequences prior to counting. It is often necessary to trim off remainders of adapter sequences from the reads before counting.



Adapter trimming

☑ Perform custom adapter trimming before counting

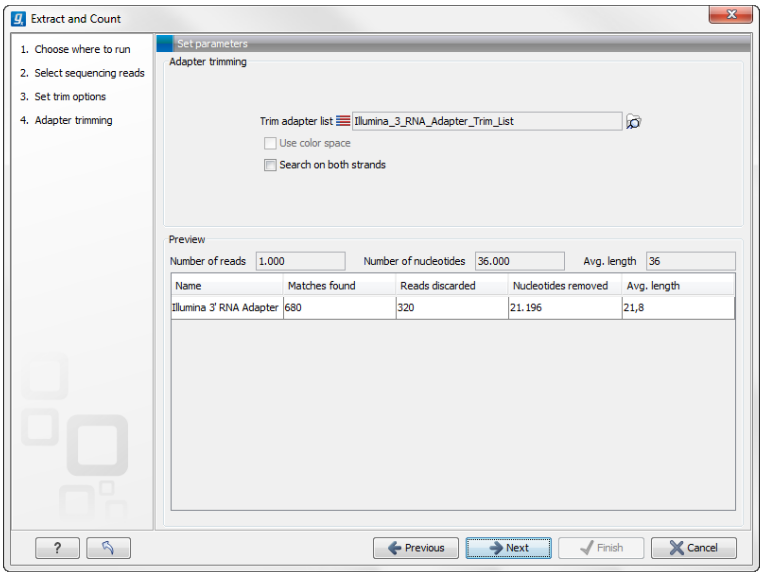
Figure 27.17: *Specifying whether adapter trimming is needed.*

When you click **Next**, you will be able to specify how the trim should be performed as shown in figure 27.18.

If you have chosen not to trim the reads for adapter sequence, you will see figure 27.19 instead.

The trim options shown in figure 27.18 are the same as described under adapter trim in section 19.2.2. Please refer to this section for more information.

It should be noted that if you expect to see part of adapters in your reads, you would typically choose **Discard when not found** as the action. By doing this, only reads containing the adapter sequence will be counted as small RNAs in the further analysis. If you have a data set where the



Extract and Count

1. Choose where to run
2. Select sequencing reads
3. Set trim options
4. Adapter trimming

Set parameters

Adapter trimming

Trim adapter list: Illumina_3_RNA_Adapter_Trim_List

□ Use color space

□ Search on both strands

Preview

Number of reads: 1.000 Number of nucleotides: 36.000 Avg. length: 36

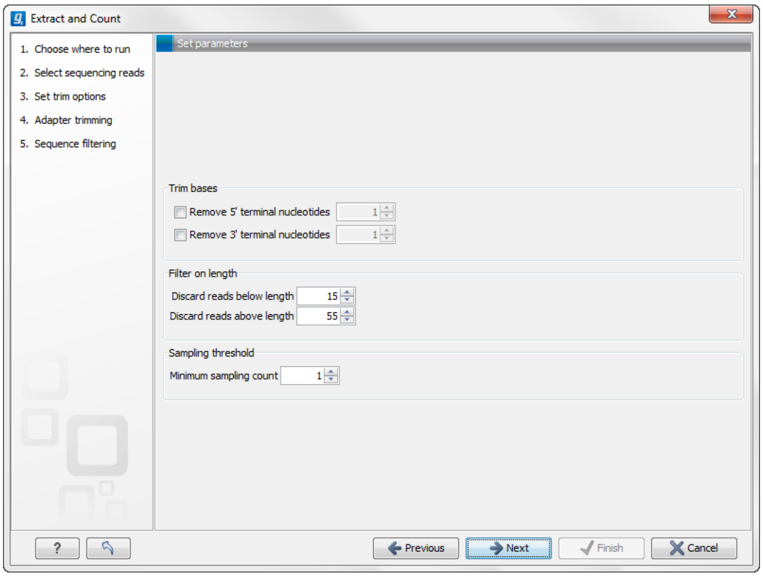
Name	Matches found	Reads discarded	Nucleotides removed	Avg. length
Illumina 3' RNA Adapter	680	320	21.196	21,8

Figure 27.18: *Setting parameters for adapter trim.*

adapter may be there or not you would choose **Remove adapter**.

Note that all reads will be trimmed for ambiguity symbols such as N before the adapter trim.

Clicking **Next** allows you to specify additional options regarding trimming and counting as shown in figure 27.19.



Extract and Count

1. Choose where to run
2. Select sequencing reads
3. Set trim options
4. Adapter trimming
5. Sequence filtering

Set parameters

Trim bases

□ Remove 5' terminal nucleotides: 1

□ Remove 3' terminal nucleotides: 1

Filter on length

Discard reads below length: 15

Discard reads above length: 55

Sampling threshold

Minimum sampling count: 1

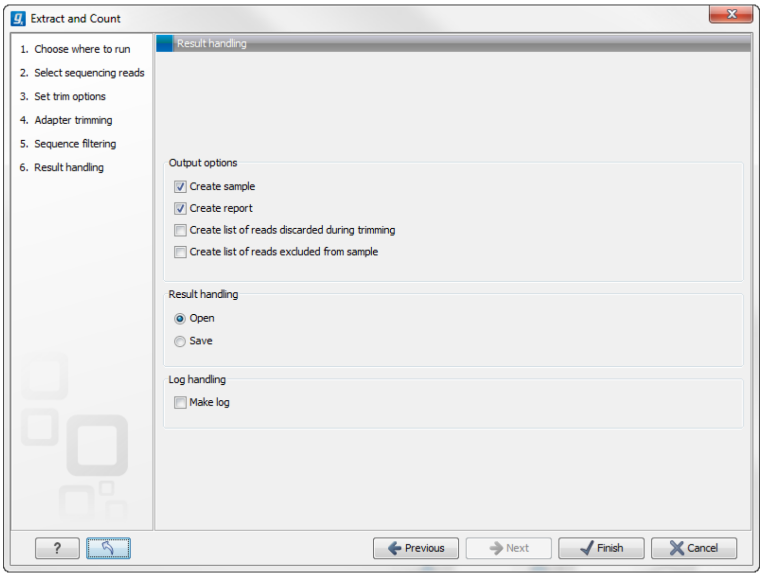
Figure 27.19: *Defining length interval and sampling threshold.*

At the top you can choose to **Trim bases** by specifying a number of bases to be removed from either the 3' or the 5' end of the reads. Below, you can specify the minimum and maximum lengths of the small RNAs to be counted (this is the length after trimming). The minimum length that can be set is 15 and the maximum is 55.

At the bottom, you can specify the **Minimum sampling count**. This is the number of copies of the small RNAs (tags) that are needed in order to include it in the resulting count table (the small RNA sample). The actual counting is very simple and relies on **perfect match** between the reads to be counted together[3]. This also means that a count threshold of 1 will include a lot of unique tags as a result of sequencing errors. In order to set the threshold right, the following should be considered:

- If the sample is going to *be annotated*, annotations may be found for the tags resulting from sequencing errors. This means that there is no negative effect of including tags with a low count in the output.
- When using *un-annotated sequences* for discovery of novel small RNAs, it may be useful to apply a higher threshold to eliminate the noise from sequencing errors. However, this can be done at a later stage by filtering the sample and creating a sub-set.
- When multiple samples are compared, it is interesting to know if one tag which is abundant in one sample is also found in another, even at a very low number. In this case, it is useful to include the tags with very low counts, since they may become more trustworthy in combination with information from other samples.
- Setting the count threshold higher will reduce the size of the sample produced which will reduce the memory and disk usage when working with the results.

Clicking **Next** allows you to specify the output of the analysis as shown in 27.20.



Output options

☑ Create sample

☑ Create report

□ Create list of reads discarded during trimming

□ Create list of reads excluded from sample

Result handling

● Open

○ Save

Log handling

□ Make log

Figure 27.20: *Output options.*

The options are:

Create sample This is the primary result showing all the tags and respective counts (an example is shown in figure 27.21). Each row represents a tag with the actual sequence as the feature ID and a column with **Length** and **Count**. The actual count is based on 100 % similarity[4]. The sample can be used in further analysis by the tools of the **Transcriptomics Analysis** toolbox in the "raw" form, or you can annotate it (see below). The tools for working with the data in the sample are described in section 27.2.4.

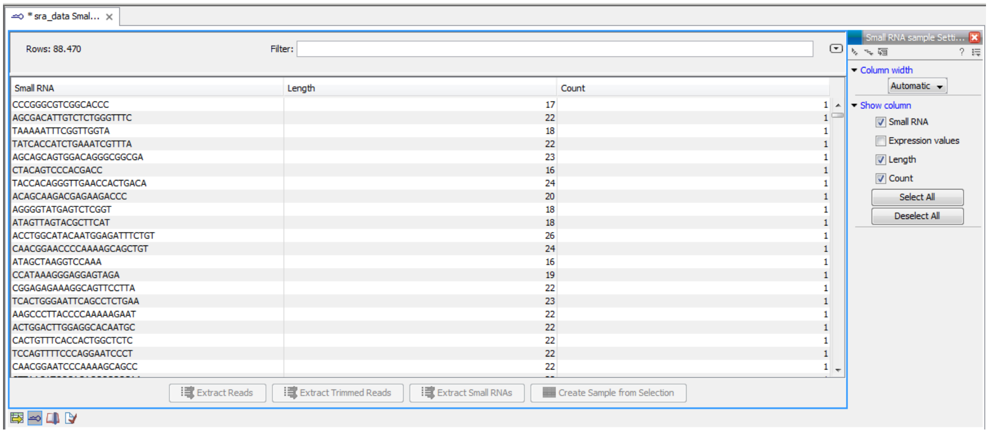
[3] Note that you can identify variants of the same miRNA when annotating the sample (see below).

[4] Note that you can identify variants of the same miRNA when annotating the sample (see below).

Create report This will create a summary report as described below.

Create list of reads discarded during trimming This list contains the reads where no adapter was found (when choosing **Discard when not found** as the action).

Create list of reads excluded from sample This list contains the reads that passed the trimming but failed to meet the sampling thresholds regarding minimum/maximum length and number of copies.



Rows: 88.470

Filter:

Small RNA	Length	Count
CCCGGGCGTCGGCACCC	17	1
AGCGACATTGTCTCTGGGTTTC	22	1
TAAAAATTTCGGTTGGTA	18	1
TATCACCATCTGAAATCGTTTA	22	1
AGCAGCAGTGGACAGGGCGGCGA	23	1
CTACAGTCCCACGACC	16	1
TACCACAGGGTTGAACCACTGACA	24	1
ACAGCAAGACGAGAAGACCC	20	1
AGGGGTATGAGTCTCGGT	18	1
ATAGTTAGTACGCTTCAT	18	1
ACCTGGCATACAATGGAGATTTCTGT	26	1
CAACGGAACCCCAAAAGCAGCTGT	24	1
ATAGCTAAGGTCCAAA	16	1
CCATAAAGGGAGGAGTAGA	19	1
CGGAGAGAAAGGCAGTTCCTTA	22	1
TCACTGGGAATTCAGCCTCTGAA	23	1
AAGCCCTTACCCCAAAAAGAAT	22	1
ACTGGACTTGGAGGCACAATGC	22	1
CACTGTTTCACCACTGGCTCTC	22	1
TCCAGTTTTCCCAGGAATCCCT	22	1
CAACGGAATCCCAAAAGCAGCC	22	1

Column width
Automatic

Show column
☑ Small RNA
☐ Expression values
☑ Length
☑ Count

Figure 27.21: *The tags have been extracted and counted.*

The summary report includes the following information (an example is shown in figure 27.22):

Trim summary Shows the following information for each input file:

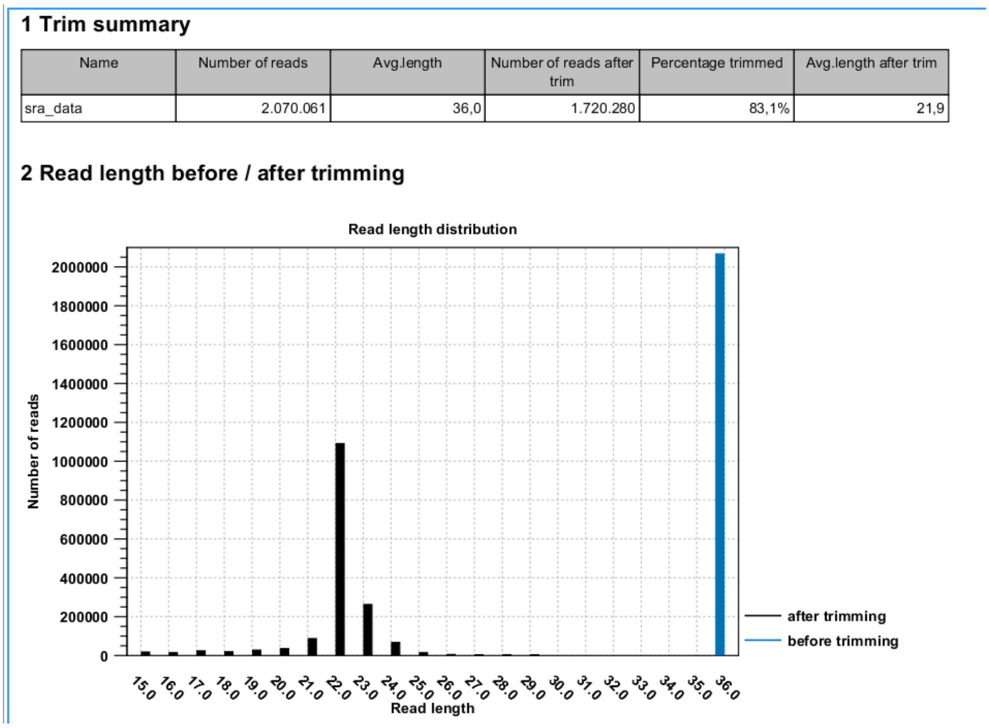
- Number of reads in the input.
- Average length of the reads in the input.
- Number of reads after trim. The difference between the number of reads in the input and this number will be the number of reads that are discarded by the trim.
- Percentage of the reads that pass the trim.
- Average length after trim. When analyzing miRNAs, you would expect this number to be around 22. If the number is significantly lower or higher, it could indicate that the trim settings are not right. In this case, check that the trim sequence is correct, that the strand is right, and adjust the alignment scores. Sometimes it is preferable to increase the minimum scores to get rid of low-quality reads. The average length after trim could also be somewhat larger than 22 if your sequenced data contains a mixture of miRNA and other (longer) small RNAs.

Read length before/after trimming Shows the distribution of read lengths before and after trim. The graph shown in figure 27.22 is typical for miRNA sequencing where the read lengths after trim peaks at 22 bp.

Trim settings The trim settings summarized. Note that ambiguity characters will automatically be trimmed.

Detailed trim results This is described under adapter trim in section 19.2.2.

Tag counts The number of tags and two plots showing on the x-axis the counts of tags and on the y-axis the number of tags for which this particular count is observed. The plot is in a zoomed version where only the lower part of the y-axis is shown to make it possible to see the numbers of tags higher counts.



1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
sra_data	2.070.061	36,0	1.720.280	83,1%	21,9

2 Read length before / after trimming

Figure 27.22: *A summary report of the counting.*

27.2.2 Downloading miRBase

In order to make use of the additional information about mature regions on the precursor miRNAs in miRBase, you need to use the integrated tool to download miRBase rather than downloading it from http://www.mirbase.org/:

Toolbox | **Transcriptomics Analysis** () | **Small RNA Analysis** () | **Download miRBase** ()

This will download a sequence list with all the precursor miRNAs including annotations for mature regions. The list can then be selected when annotating the samples with miRBase (see section 27.2.3).

The downloaded version will always be the latest version (it is downloaded from ftp://mirbase.org/pub/mirbase/CURRENT/miRNA.dat.gz). Information on the version number of miRBase is also available in the **History** () of the downloaded sequence list, and when using this for annotation, the annotated samples will also include this information in their **History** ().

Importing the miRBase data file

You can also import the miRBase data file directly into the Workbench. The file can be downloaded from ftp://mirbase.org/pub/mirbase/CURRENT/miRNA.dat.gz.

In order for the file to be recognized as a miRBase file, you have to select `miRBase dat` in the **Force import as type** menu of the import dialog.

Creating your own miRBase file

If you wish to construct a file yourself to be used as a miRBase file for annotation, this is also possible if you format the file in the same way as the miRBase data file. In particular, the following needs to be in place:

- The sequences needs "miRNA" annotation on the precursor sequences. In the Workbench, you can add a miRNA annotation by selecting a region and right click to **Add Annotation**. You should have a max 2 miRNA annotations per precursor sequences. Matches to first miRNA annotation are counting in `5′` column. Matches to second miRNA annotation are counted as `3′` matches.
- If you have sequence list containing sequences form multiple species, the **Latin name** of the sequences should be set. This is used in the annotation dialog (see section 27.2.3) where you can select the species. If the Latin name is not set, the dialog will show "N/A".

Once you have created the file, it has to be imported as described above.

27.2.3 Annotating and merging small RNA samples

The small RNA sample produced when counting the tags (see section 27.2.1) can be enriched by *CLC Cancer Research Workbench* by comparing the tag sequences with annotation resources such as miRBase and other small RNA annotation sources. Note that the annotation can also be performed on an experiment, set up from small RNA samples (see section 27.3.1).

Besides adding annotations to known small RNAs in the sample, it is also possible to merge variants of the same small RNA to get a cumulative count. When initially counting the tags, the Workbench requires that the trimmed reads are identical for them to be counted as the same tag. However, you will often see different variants of the same miRNA in a sample, and it is useful to be able to count these together. This is also possible using the tool to annotate and merge samples.

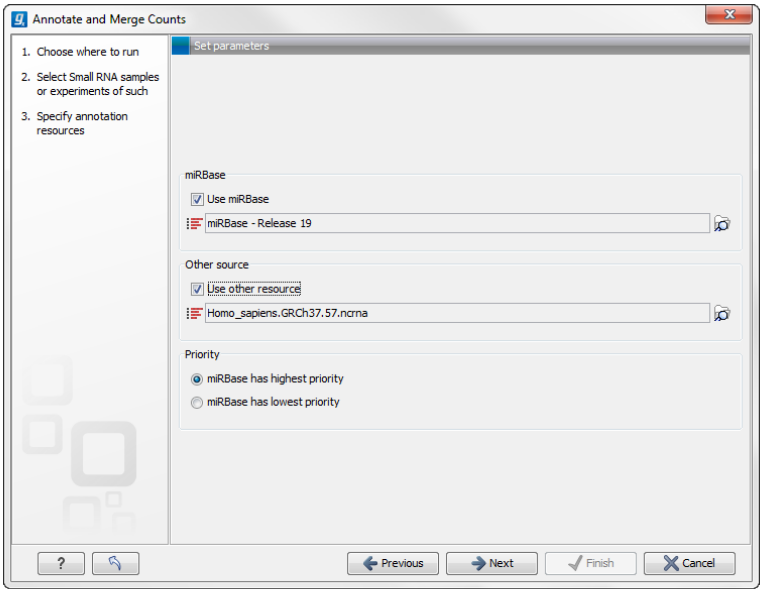
> **Toolbox | Transcriptomics Analysis ( ) | Small RNA Analysis ( ) | Annotate and Merge Counts ( )**

This will open a dialog where you select the small RNA samples ( ) to be annotated. Note that if you have included several samples, they will be processed separately but summarized in one report providing a good overview of all samples. You can also input **Experiments** ( ) (see section 27.3.1) created from small RNA samples. Click **Next** when the data is listed in the right-hand side of the dialog.

This dialog (figure 27.23) is where you define the annotation resources to be used.

There are two ways of providing annotation sources:

- Downloading miRBase using the integrated download tool (explained in section 27.2.2).
- Importing a list of sequences, e.g. from a fasta file. This could be from Ensembl, e.g. `ftp://ftp.ensembl.org/pub/release-57/fasta/homo_sapiens/ncrna/`



Annotate and Merge Counts

1. Choose where to run
2. Select Small RNA samples or experiments of such
3. Specify annotation resources

Set parameters

miRBase

☑ Use miRBase

miRBase - Release 19

Other source

☑ Use other resource

Homo_sapiens.GRCh37.57.ncrna

Priority

● miRBase has highest priority

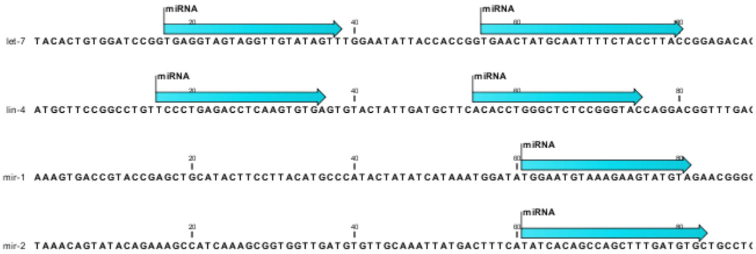
○ miRBase has lowest priority

Figure 27.23: *Defining annotation resources.*

Homo_sapiens.GRCh37.57.ncrna.fa.gz or from ncRNA.org: http://www.ncrna.org/frnadb/files/ncrna.zip.

Note: We recommend using the integrated download tool to import miRBase. Although it is possible to import it as a fasta file, the same options with regards to species will not be available if you import from a file.

The downloaded miRBase file contains all precursor sequences from the latest version of miRBase http://www.mirbase.org/ including annotations defining the mature regions (see an example in figure 27.24).



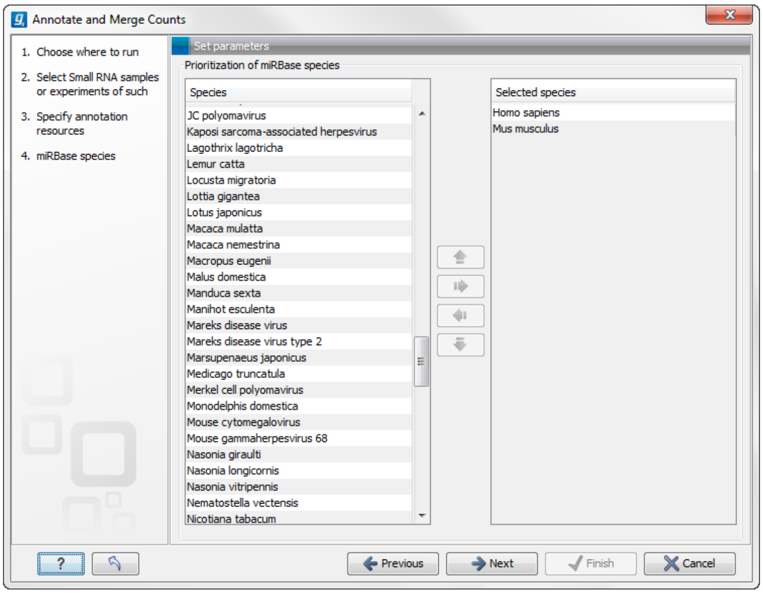
miRNA
miRNA
let-7 TACACTGTGGATCCGGTGAGGTAGTAGGTTGTATAGTTTGGAATATTACCACCGGTGAACTATGCAATTTTCTACCTTACCGGAGACA
miRNA
miRNA
lin-4 ATGCTTCCGGCCTGTTCCCTGAGACCTCAAGTGTGAGTGTACTATTGATGCTTCACACCTGGGCTCTCCGGGTACCAGGACGGTTTGA
miRNA
mir-1 AAAGTGACCGTACCGAGCTGCATACTTCCTTACATGCCCATACTATATCATAAATGGATATGGAATGTAAAGAAGTATGTAGAACGGG
miRNA
mir-2 TAAACAGTATACAGAAAGCCATCAAAGCGGTGGTTGATGTGTTGCAAATTATGACTTTCATATCACAGCCAGCTTTGATGTGCTGCCT
20
40
60
80

Figure 27.24: *Some of the precursor miRNAs from miRBase have both 3' and 5' mature regions (previously referred to as mature and mature*) annotated (as the two first in this list).*

This means that it is possible to have a more fine-grained classification of the tags using miRBase compared to a simple fasta file resource containing the full precursor sequence. This is the reason why the miRBase annotation source is specified separately in figure 27.23.

At the bottom of the dialog, you can specify whether miRBase should be prioritized over the additional annotation resource. The prioritization is explained in detail later in this section. To prioritize one over the other can be useful when there is redundant information (e.g. if you have an additional source that also contains all the miRNAs from miRBase and you prefer the miRBase annotations when possible).

When you click **Next**, you will be able to choose which species from miRBase should be used and in which order (see figure 27.25). Note that if you have not selected a miRBase annotation source, you will go directly to the next step shown in figure 27.26.



Annotate and Merge Counts

1. Choose where to run
2. Select Small RNA samples or experiments of such
3. Specify annotation resources
4. miRBase species

Set parameters

Prioritization of miRBase species

Species

JC polyomavirus
Kaposi sarcoma-associated herpesvirus
Lagothrix lagotricha
Lemur catta
Locusta migratoria
Lottia gigantea
Lotus japonicus
Macaca mulatta
Macaca nemestrina
Macropus eugenii
Malus domestica
Manduca sexta
Manihot esculenta
Mareks disease virus
Mareks disease virus type 2
Marsupenaeus japonicus
Medicago truncatula
Merkel cell polyomavirus
Monodelphis domestica
Mouse cytomegalovirus
Mouse gammaherpesvirus 68
Nasonia giraulti
Nasonia longicornis
Nasonia vitripennis
Nematostella vectensis
Nicotiana tabacum

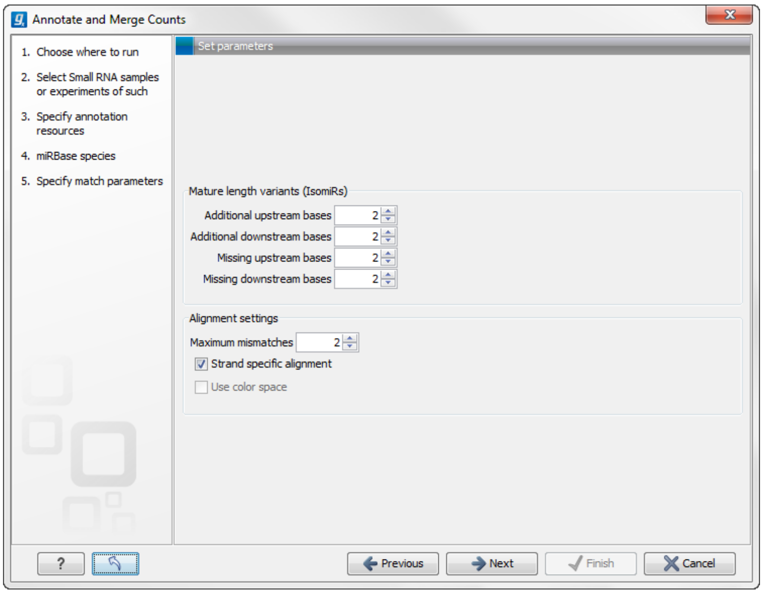
Selected species

Homo sapiens
Mus musculus

Figure 27.25: *Defining and prioritizing species in miRBase.*

To the left, you see the list of species in miRBase. This list is dynamically created based on the information in the miRBase file. Using the arrow button (➡) you can add species to the right-hand panel. The order of the species is important since the tags are annotated iteratively based on the order specified here. This means that in the example in figure 27.25, a human miRNA will be preferred over mouse, even if they are identical in sequence (the prioritization is elaborated below). The up and down arrows (⬆)/ (⬇) can be used to change the order of species.

When you click **Next**, you will be able to specify how the alignment of the tags against the annotation sources should be performed (see figure 27.26).



Annotate and Merge Counts

1. Choose where to run
2. Select Small RNA samples or experiments of such
3. Specify annotation resources
4. miRBase species
5. Specify match parameters

Set parameters

Mature length variants (IsomiRs)

Additional upstream bases 2
Additional downstream bases 2
Missing upstream bases 2
Missing downstream bases 2

Alignment settings

Maximum mismatches 2
☑ Strand specific alignment
☐ Use color space

Figure 27.26: *Setting parameters for aligning.*

The panel at the top is active only if you have chosen to annotate with miRBase. It is used to define the requirements to the alignment of a read for it to be counted as a 3' or 5' mature region (previously referred to as mature and mature*) tag:

Additional upstream bases This defines how many bases the tag is allowed to extend the annotated mature region at the 5' end and still be categorized as mature.

Additional downstream bases This defines how many bases the tag is allowed to extend the annotated mature region at the 3' end and still be categorized as mature.

Missing upstream bases This defines how many bases the tag is allowed to miss at the 5' end compared to the annotated mature region and still be categorized as mature.

Missing downstream bases This defines how many bases the tag is allowed to miss at the 3' end compared to the annotated mature region and still be categorized as mature.

At the bottom of the dialog you can specify the **Maximum mismatches** (default value is 2). Furthermore, you can specify if the alignment and annotation should be performed in **color space** which is available when your small RNA sample is based on SOLiD data. [5] Finally, you can choose whether the tags should be aligned against both strands of the reference or only the positive strand. Usually it is only necessary to align against the positive strand.

At this point, a more elaborate explanation of the annotation algorithm is needed. The short read mapping algorithm in the *CLC Cancer Research Workbench* is used to map all the tags to the reference sequences which comprise the full precursor sequences from miRBase and the sequence lists chosen as additional resources. The mapping is done in several rounds: the first round is done requiring a perfect match, the second allowing one mismatch, the third allowing two mismatches etc. No gaps are allowed. The number of rounds depend on the number of mismatches allowed[6] (default is two which means three rounds of read mapping, see figure 27.26).

After each round of mapping, the tags that are mapped will be removed from the list of tags that continue to the next round. This means that a tag mapping with perfect match in the first round will not be considered for the subsequent one-mismatch round of mapping.

Following the mapping, the tags are classified into the following categories according to where they match.

- Mature 5' exact
- Mature 5' super
- Mature 5' sub
- Mature 5' sub/super
- Mature 3' exact

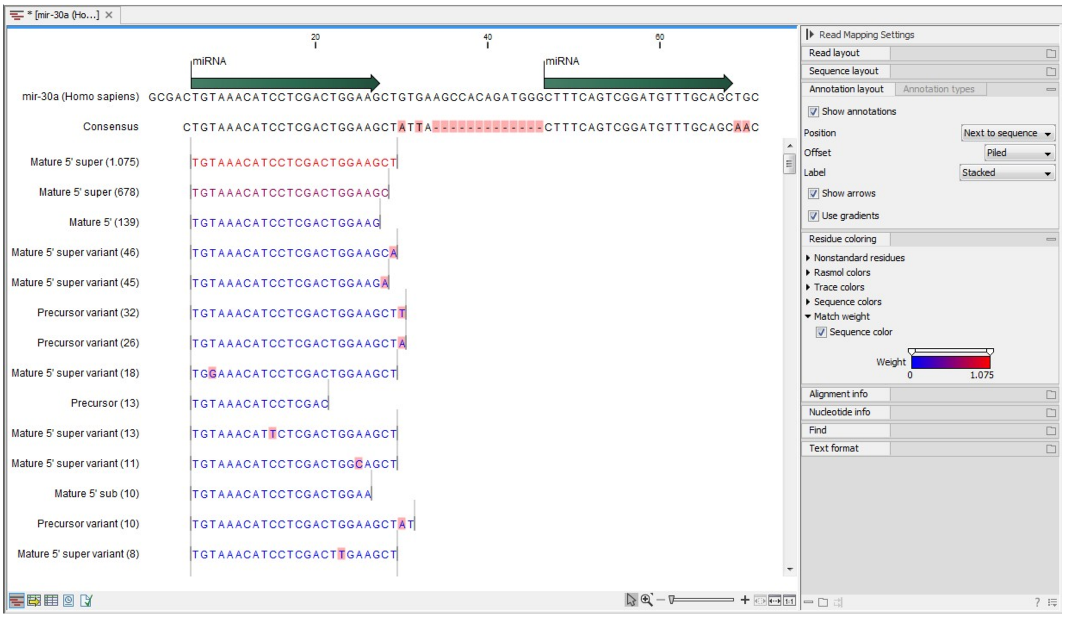
[5] Note that this option is only going to make a difference for tags with low counts. Since the actual tag counting in the first place is done based on perfect matches, the highly abundant tags are not likely to have sequencing errors, and aligning in color space does not add extra benefit for these.

[6] For color space, the maximum number of mismatches is 2.

- Mature 3' super
- Mature 3' sub
- Mature 3' sub/super
- Precursor
- Other

All these categories except *Other* refer to hits in miRBase. For hits on mirBase sequences we distinguish between where on the sequences the tags match. The mirBase sequences may have up to two mature micro RNAs annotated. We refer to a mature miRNA that is located closer (or equally close) to the 5' end than to the 3' end as 'Mature 5''. A mature miRNA that is located closer to the 3' end is referred to as 'Mature 3''. *Exact* means that the tag matches exactly to the annotated mature 5' or 3' region; *Sub* means that the observed tag is shorter than the annotated mature 5' or mature 3'; *super* means that the observed tag is longer than the annotated mature 5' or mature 3'. The combination *sub/super* means that the observed tag extends the annotation in one end and is shorter at the other end. Precursor means that the tag matches on a mirBase sequence, but not within the extended annotated mature region(s). These are defined by the "mature length variants (IsomiRs)" parameters in the "specify match parameters" wizard step (by default these parameters are set to 2 which means that reads that are at most 2 bases too long or too short relative to the annotated mature region are all considered mature hits). The Other category is for hits in the other resources (the information about resource is also shown in the output).

An example of an alignment is shown in figure 27.27 using the same alignment settings as in figure 27.26.



```
mir-30a (Homo sapiens)       GCGACTGTAAACATCCTCGACTGGAAGCTGTGAAGCCACAGATGGGCTTTCAGTCGGATGTTTGCAGCTGC
Consensus                        CTGTAAACATCCTCGACTGGAAGCTATTA-------------CTTTCAGTCGGATGTTTGCAGCAAC
Mature 5' super (1.075)            TGTAAACATCCTCGACTGGAAGCT
Mature 5' super (678)              TGTAAACATCCTCGACTGGAAGC
Mature 5' (139)                    TGTAAACATCCTCGACTGGAAG
Mature 5' super variant (46)       TGTAAACATCCTCGACTGGAAGCA
Mature 5' super variant (45)       TGTAAACATCCTCGACTGGAAGA
Precursor variant (32)             TGTAAACATCCTCGACTGGAAGCTT
Precursor variant (26)             TGTAAACATCCTCGACTGGAAGCTA
Mature 5' super variant (18)       TGGAAACATCCTCGACTGGAAGCT
Precursor (13)                     TGTAAACATCCTCGAC
Mature 5' super variant (13)       TGTAAACATTCTCGACTGGAAGCT
Mature 5' super variant (11)       TGTAAACATCCTCGACTGGCAGCT
Mature 5' sub (10)                 TGTAAACATCCTCGACTGGAA
Precursor variant (10)             TGTAAACATCCTCGACTGGAAGCTAT
Mature 5' super variant (8)        TGTAAACATCCTCGACTTGAAGCT
```

Read Mapping Settings
Read layout
Sequence layout
Annotation layout | Annotation types
☑ Show annotations
Position: Next to sequence
Offset: Piled
Label: Stacked
☑ Show arrows
☑ Use gradients
Residue coloring
- Nonstandard residues
- Rasmol colors
- Trace colors
- Sequence colors
- Match weight
  - ☑ Sequence color
  - Weight 0 1.075

Alignment info
Nucleotide info
Find
Text format

Figure 27.27: *Alignment of length variants of mir-30a.*

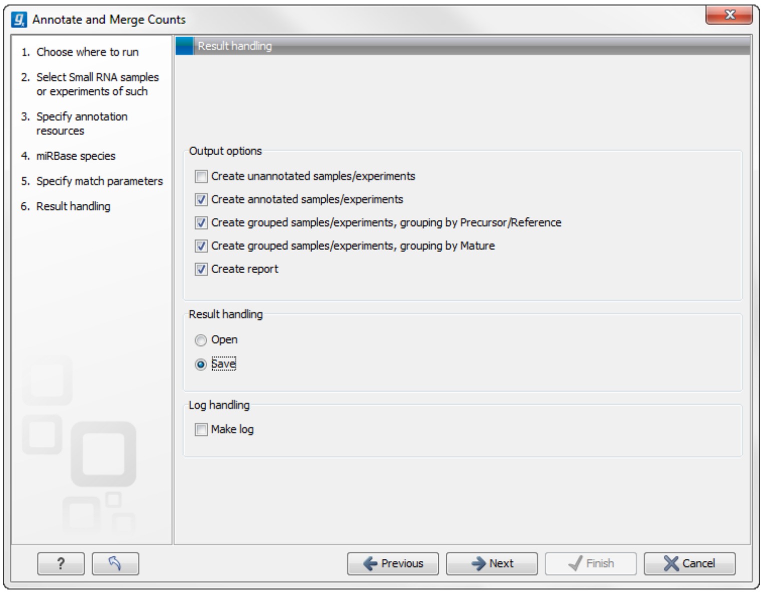
The two tags at the top are both classified as *mature 5' super* because they cover and extend beyond the annotated mature 5' RNA. The third tag is identical to the annotated mature 5'.

If a tag has several hits, the list above is used for prioritization. This means that e.g. a *Mature 5' sub* is preferred over a *Mature 3' exact*. Note that if miRBase was chosen as lowest priority (figure 27.23), the *Other* category will be at the top of the list. All tags mapping to a miRBase reference without qualifying to any of the mature 5' and mature 3' types will be typed as *Other*. Also note that if a tag has several hits to references *with the same priority* (e.g. the tag matches the mature regions of two different miRBase sequences) it will be annotated with all these sequences. In the report we refer to these tags as 'ambiguously annotated'.

In case you have selected more than one species for miRBase annotation (e.g. Homo Sapiens and Mus Musculus) the following rules for adding annotations apply:

1. If a tag has hits with the same priority for both species, the annotation for the top-prioritized species will be added.
2. Read category priority is stronger than species category priority: If a read is a higher priority match for a mouse miRBase sequence than it is for a human miRBase sequence the annotation for the mouse will be used

Clicking **Next** allows you to specify the output of the analysis as shown in 27.28.


Annotate and Merge Counts
1. Choose where to run
2. Select Small RNA samples or experiments of such
3. Specify annotation resources
4. miRBase species
5. Specify match parameters
6. Result handling
Result handling
?
Previous
Next
Finish
Cancel

Output options
- ☐ Create unannotated samples/experiments
- ☑ Create annotated samples/experiments
- ☑ Create grouped samples/experiments, grouping by Precursor/Reference
- ☑ Create grouped samples/experiments, grouping by Mature
- ☑ Create report

Result handling
- ○ Open
- ● Save

Log handling
- ☐ Make log

Figure 27.28: *Output options.*

The options are:

Create unannotated sample All the tags where no hit was found in the annotation source are included in the unannotated sample. This sample can be used for investigating novel miRNAs, see section 27.2.5. No extra information is added, so this is just a subset of the input sample.

Create annotated sample This will create a sample as described in section 27.2.4. In this sample, the following columns have been added to the counts.

Name This is the name of the annotation sequence in the annotation source. For miRBase, it will be the names of the miRNAs (e.g. *let-7g* or *mir-147*), and for other source, it will be the name of the sequence.

Resource This is the source of the annotation, either miRBase (in which case the species name will be shown) or other sources (e.g. Homo_sapiens.GRCh37.57.ncrna).

Match type The match type can be exact or variant (with mismatches) of the following types:

- Mature 5'
- Mature 5' super
- Mature 5' sub
- Mature 5' sub/super
- Mature 3'
- Mature 3' super
- Mature 3' sub
- Mature 3' sub/super
- Other

Mismatches The number of mismatches.

Note that if a tag has two equally prioritized hits, they will be shown with // between the names. This could be e.g. two precursor sequences sharing the same mature sequence (also see the sample grouped on mature below).

Create grouped sample, grouping by Precursor/Reference This will create a sample as described in section 27.2.4. All variants of the same reference sequence will be merged to create one expression value for all.

Expression values. The expression value can be changed at the bottom of the table. The default is to use the counts in the mature 5' column.

Name. The name of the reference. For miRBase this will then be the name of the precursor.

Resource. The name of the resource that the reference comes from.

Exact mature 5'. The number of exact mature 5' reads.

Mature 5'. The number of all mature 5' reads including sub, super and variants.

Unique exact mature 5'. In cases where one tag has several hits (as denoted by the // in the ungrouped annotated sample as described above), the counts are distributed evenly across the references. The difference between *Exact mature 5'* and *Unique exact mature 5'* is that the latter only includes reads that are unique to this reference.

Unique mature 5'. Same as above but for all mature 5's, including sub, super and variants.

Exact mature 3'. Same as above, but for mature 3'.

Mature 3'. Same as above, but for mature 3'.

Unique exact mature '3. Same as above, but for mature 3'.

Unique mature '3. Same as above, but for mature 3'.

Exact other. Exact matches in miRBase sequences, but outside annotated mature regions.

Other. All matches in miRBase sequences, but outside annotated mature regions, including variants.

Total. The total number of tags mapped and classified to the precursor/reference sequence.

Note that, for non-mirBase sequences, the counts are collected in the 'Mature 5'' columns: 'Exact mature 5'' (number reads that map to the sequence without mismatches), 'Mature 5' (number reads that map to the sequence, including those with mismatches), 'Unique exact mature 5' (number reads that map uniquely to the sequence without mismatches) and 'Unique mature 5' (number reads that map uniquely to the sequence, including those with mismatches).

Create grouped sample, grouping by Mature This will create a sample as described in section 27.2.4. This is also a grouped sample, but in addition to grouping based on the same reference sequence, the tags in this sample are grouped on the same mature 5'. This means that two precursor variants of the same mature 5' miRNA are merged. Note that it is only possible to create this sample when using miRBase as annotation resource (because the Workbench has a special interpretation of the miRBase annotations for mature as described previously). To find identical mature 5' miRNAs, the Workbench compares all the mature 5' sequences and when they are identical, they are merged. The names of the precursor sequences merged are all shown in the table.

Expression values. The expression value can be changed at the bottom of the table. The default is to use the counts in the mature 5' column.

Name. The name of the reference. When several precursor sequences have been merged, all the names will be shown separated by //.

Resource. The species of the reference.

Exact mature 5'. The number of exact mature 5' reads.

Mature 5'. The number of all mature 5' reads including sub, super and variants.

Unique exact mature 5'. In cases where one tag has several hits (as denoted by the // in the ungrouped annotated sample as described above), the counts are distributed evenly across the references. The difference between *Exact mature 5'* and *Unique exact mature 5'* is that the latter only includes reads that are unique to one of the precursor sequences that are represented under this mature 5' sequence.

Unique mature 5'. Same as above but for all mature 5's, including sub, super and variants.

Create report. A summary report described below.

The summary report includes the following information (an example is shown in figure 27.29):

Summary Shows the following information for each input sample:

- Number of small RNAs(tags) in the input.
- Number of annotated tags (number and percentage).
- Number of ambiguously annotated tags (number and percentage).
- Number of reads in the sample (one tag can represent several reads)
- Number of annotated reads (number and percentage).
- Number of ambiguously annotated reads (number and percentage).

Resources Shows how many matches were found in each resource:

- Number of sequences in the resource.

- Number of sequences where a match was found (i.e. this sequence has been observed at least once in the sequencing data).

Reads Shows the number of reads that fall into different categories (there is one table per input sample). On the left hand side are the annotation resources. For each resource, the count and percentage of reads in that category are shown. Note that the percentage are relative to the overall categories (e.g. the miRBase reads are a percentage of all the *annotated* reads, not all reads). This is information is shown for each mismatch level.

Small RNAs Similar numbers as for the reads but this time for each small RNA tag and without mismatch differentiation.

Read count proportions A histogram showing, for each interval of read counts, the proportion of annotated (respectively, unannotated) small RNAs with a read count in that interval. Annotated small RNAs may be expected to be associated with higher counts, since the most abundant small RNAs are likely to be known already.

Annotations (miRBase) Shows an overview table for classifications of the number of reads that fall in the miRBase categories for each species selected.

Annotations (Other) Shows an overview table with read numbers for total, exact match and mutant variants for each of the other annotation resources.

27.2.4 Working with the small RNA sample

Generally speaking, the small RNA sample comes in two variants:

- The *un-grouped* sample, either as it comes directly from the **Extract and Count** ( ) or when it has been annotated. In this sample, there is one row per tag, and the feature ID is the tag sequence.
- The *grouped* sample created using the **Annotate and Merge Counts** ( ) tool. In this sample, each row represents several tags grouped by a common Mature or Precursor miRNA or other reference.

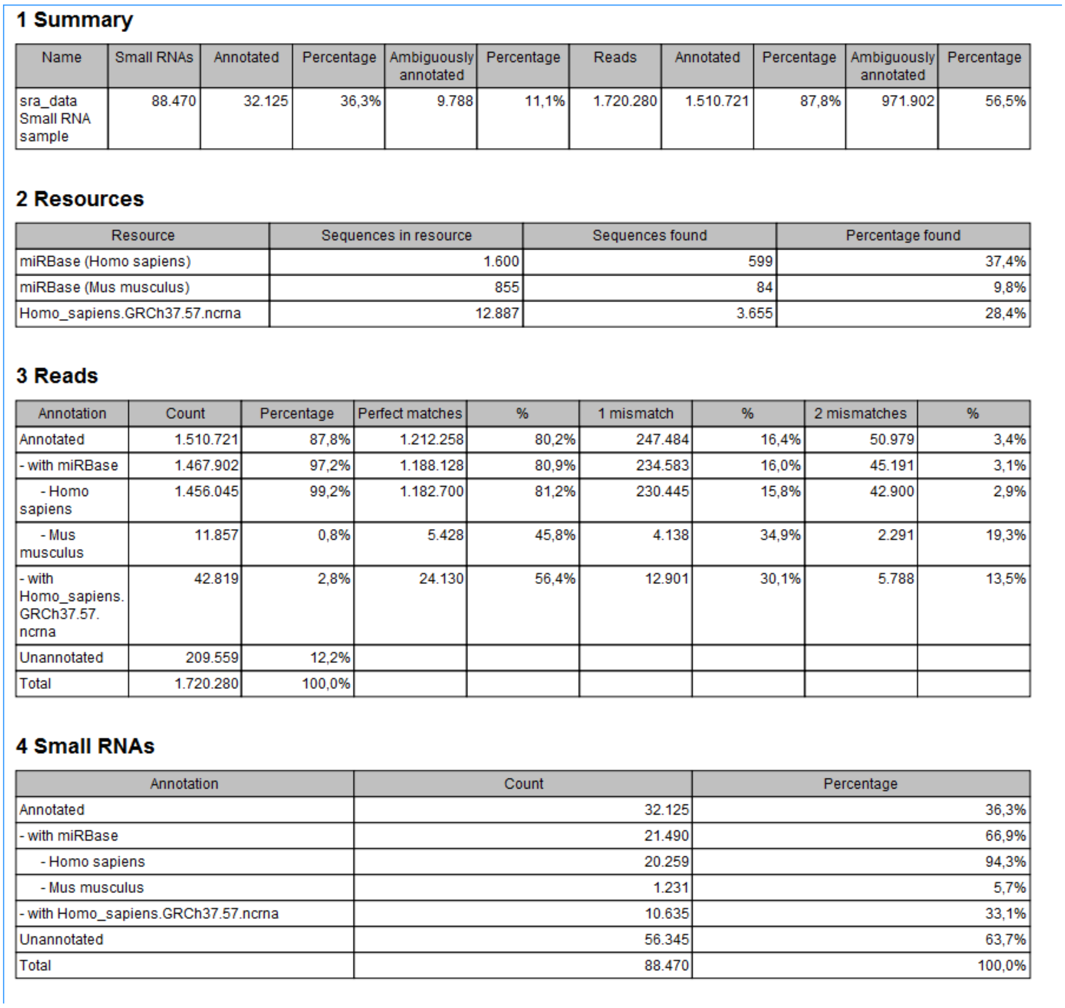
Below, these two kinds of samples are described in further detail. Note that for both samples, filtering and sorting can be applied, see section 8.3.

The un-grouped sample

An example of an un-grouped annotated sample is shown in figure 27.30.

By selecting one or more rows in the table, the buttons at the bottom of the view can be used to extract sequences from the table:

Extract Reads ( ) This will extract the original sequencing reads that contributed to this tag. Figure 27.31 shows an example of such a read. The reads include trim annotations (for use when inspecting and double-checking the results of trimming). Note that if these reads are used for read mapping, the trimmed part of the read will automatically be removed. If all rows in the sample are selected and extracted, the sequence list would be the same as



1 Summary

Name	Small RNAs	Annotated	Percentage	Ambiguously annotated	Percentage	Reads	Annotated	Percentage	Ambiguously annotated	Percentage
sra_data Small RNA sample	88.470	32.125	36,3%	9.788	11,1%	1.720.280	1.510.721	87,8%	971.902	56,5%

2 Resources

Resource	Sequences in resource	Sequences found	Percentage found
miRBase (Homo sapiens)	1.600	599	37,4%
miRBase (Mus musculus)	855	84	9,8%
Homo_sapiens.GRCh37.57.ncrna	12.887	3.655	28,4%

3 Reads

Annotation	Count	Percentage	Perfect matches	%	1 mismatch	%	2 mismatches	%
Annotated	1.510.721	87,8%	1.212.258	80,2%	247.484	16,4%	50.979	3,4%
- with miRBase	1.467.902	97,2%	1.188.128	80,9%	234.583	16,0%	45.191	3,1%
- Homo sapiens	1.456.045	99,2%	1.182.700	81,2%	230.445	15,8%	42.900	2,9%
- Mus musculus	11.857	0,8%	5.428	45,8%	4.138	34,9%	2.291	19,3%
- with Homo_sapiens.GRCh37.57.ncrna	42.819	2,8%	24.130	56,4%	12.901	30,1%	5.788	13,5%
Unannotated	209.559	12,2%						
Total	1.720.280	100,0%						

4 Small RNAs

Annotation	Count	Percentage
Annotated	32.125	36,3%
- with miRBase	21.490	66,9%
- Homo sapiens	20.259	94,3%
- Mus musculus	1.231	5,7%
- with Homo_sapiens.GRCh37.57.ncrna	10.635	33,1%
Unannotated	56.345	63,7%
Total	88.470	100,0%

Figure 27.29: *A summary report of the annotation.*

the input except for the reads that did not meet the adapter trim settings and the sampling thresholds (tag length and number of copies).

Extract Trimmed Reads () The same as above, except that the trimmed part has been removed.

Extract Small RNAs () This will extract only one copy of each tag.

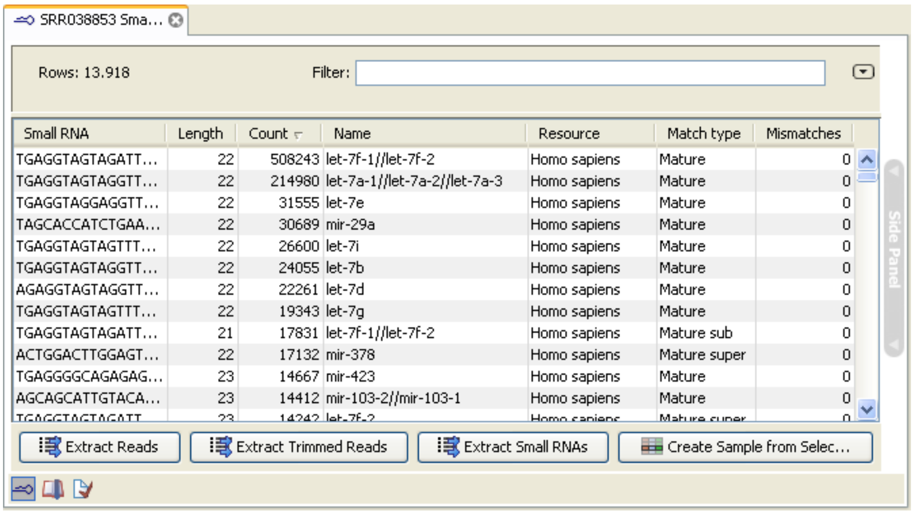
Note that for all these, you will be able to determine whether a list of DNA or RNA sequences should be produced (when working within the *CLC Cancer Research Workbench* environment, this only effects the RNA folding tools).

The button **Create Sample from Selection** () can be used to create a new sample based on the tags that are selected. This can be useful in combination with filtering and sorting.

The grouped sample

An example of a grouped annotated sample is shown in figure 27.32.

The contents of the table are explained in section 27.2.3. In this section, we focus on the tools available for working with the sample.



Rows: 13.918 Filter:

Small RNA	Length	Count	Name	Resource	Match type	Mismatches
TGAGGTAGTAGATT...	22	508243	let-7f-1//let-7f-2	Homo sapiens	Mature	0
TGAGGTAGTAGGTT...	22	214980	let-7a-1//let-7a-2//let-7a-3	Homo sapiens	Mature	0
TGAGGTAGGAGGTT...	22	31555	let-7e	Homo sapiens	Mature	0
TAGCACCATCTGAA...	22	30689	mir-29a	Homo sapiens	Mature	0
TGAGGTAGTAGTTT...	22	26600	let-7i	Homo sapiens	Mature	0
TGAGGTAGTAGGTT...	22	24055	let-7b	Homo sapiens	Mature	0
AGAGGTAGTAGGTT...	22	22261	let-7d	Homo sapiens	Mature	0
TGAGGTAGTAGTTT...	22	19343	let-7g	Homo sapiens	Mature	0
TGAGGTAGTAGATT...	21	17831	let-7f-1//let-7f-2	Homo sapiens	Mature sub	0
ACTGGACTTGGAGT...	22	17132	mir-378	Homo sapiens	Mature super	0
TGAGGGGCAGAGAG...	23	14667	mir-423	Homo sapiens	Mature	0
AGCAGCATTGTACA...	23	14412	mir-103-2//mir-103-1	Homo sapiens	Mature	0
TGAGGTAGTAGATT	23	14242	let-7f-2	Homo sapiens	Mature super	0

Figure 27.30: *An ungrouped annotated sample.*

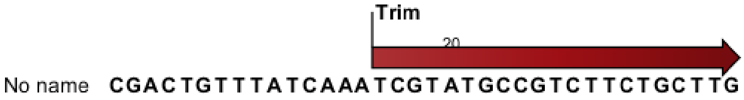
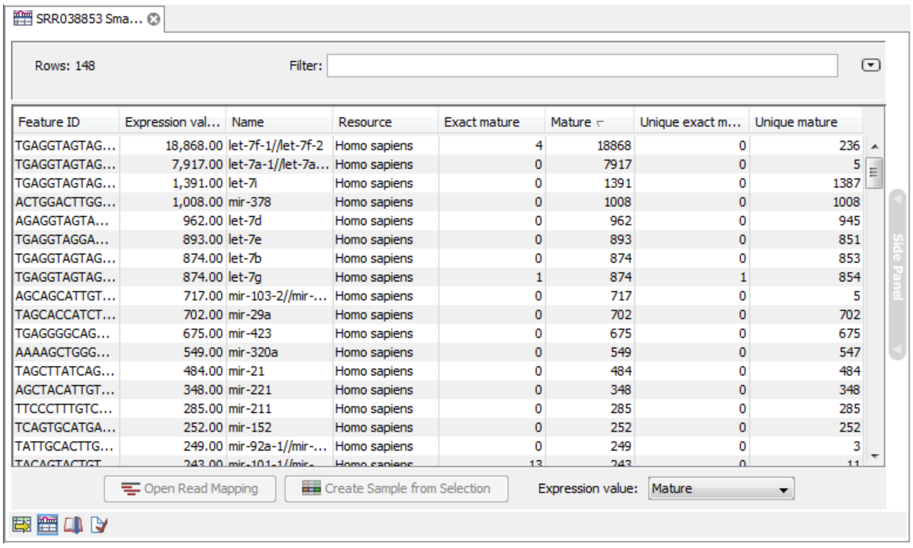


Figure 27.31: *Extracting reads from a sample.*



Rows: 148 Filter:

Feature ID	Expression val...	Name	Resource	Exact mature	Mature	Unique exact m...	Unique mature
TGAGGTAGTAG...	18,868.00	let-7f-1//let-7f-2	Homo sapiens	4	18868	0	236
TGAGGTAGTAG...	7,917.00	let-7a-1//let-7a...	Homo sapiens	0	7917	0	5
TGAGGTAGTAG...	1,391.00	let-7i	Homo sapiens	0	1391	0	1387
ACTGGACTTGG...	1,008.00	mir-378	Homo sapiens	0	1008	0	1008
AGAGGTAGTA...	962.00	let-7d	Homo sapiens	0	962	0	945
TGAGGTAGGA...	893.00	let-7e	Homo sapiens	0	893	0	851
TGAGGTAGTAG...	874.00	let-7b	Homo sapiens	0	874	0	853
TGAGGTAGTAG...	874.00	let-7g	Homo sapiens	1	874	1	854
AGCAGCATTGT...	717.00	mir-103-2//mir-...	Homo sapiens	0	717	0	5
TAGCACCATCT...	702.00	mir-29a	Homo sapiens	0	702	0	702
TGAGGGGCAG...	675.00	mir-423	Homo sapiens	0	675	0	675
AAAAGCTGGG...	549.00	mir-320a	Homo sapiens	0	549	0	547
TAGCTTATCAG...	484.00	mir-21	Homo sapiens	0	484	0	484
AGCTACATTGT...	348.00	mir-221	Homo sapiens	0	348	0	348
TTCCCTTTGTC...	285.00	mir-211	Homo sapiens	0	285	0	285
TCAGTGCATGA...	252.00	mir-152	Homo sapiens	0	252	0	252
TATTGCACTTG...	249.00	mir-92a-1//mir-...	Homo sapiens	0	249	0	3
TACAGTACTGT	243.00	mir-101-1//mir-	Homo sapiens	13	243	0	11

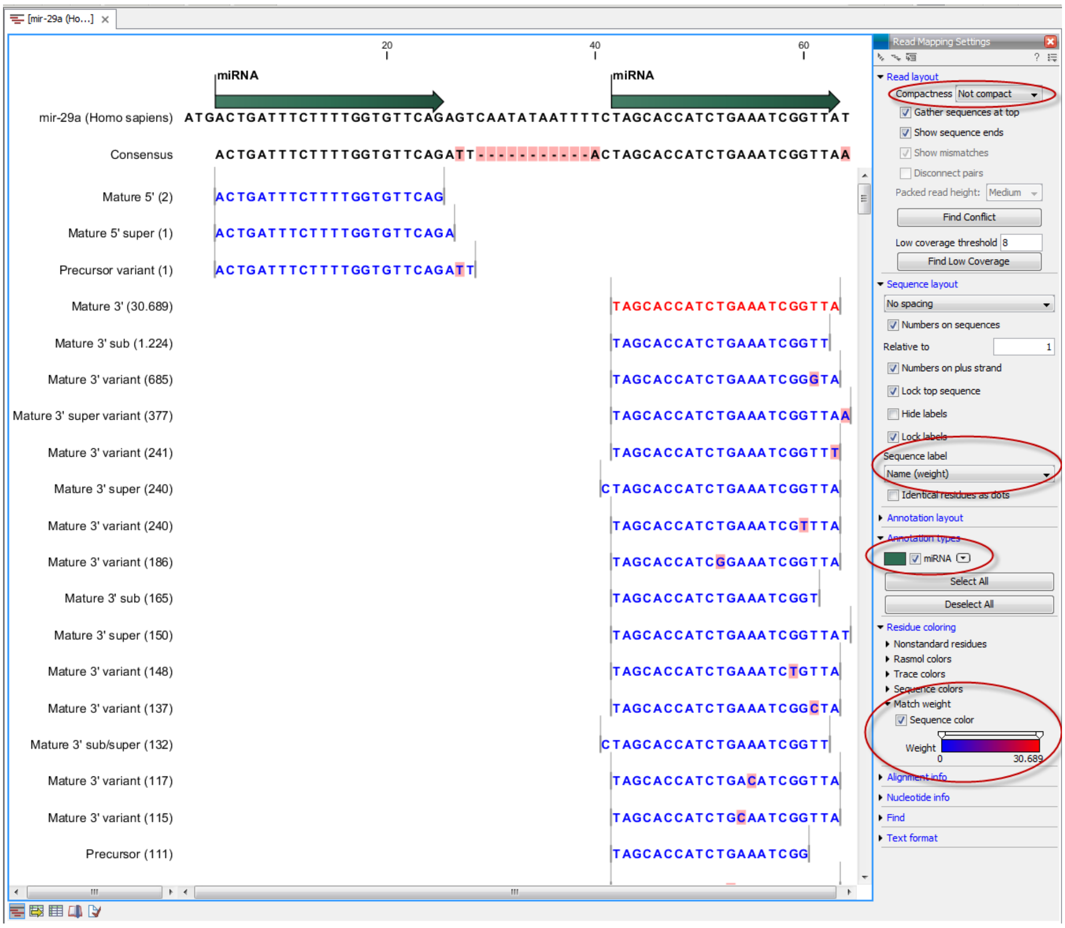
Expression value: Mature

Figure 27.32: *A sample grouped on mature 5' miRNAs.*

By selecting one or more rows in the table, the buttons at the bottom of the view become active:

Open Read Mapping (☰) This will open a view showing the annotation reference sequence at the top and the tags aligned to it as shown in figure 27.33. The names of the tags indicate their status compared with the reference (e.g. Mature 5', Mature super 5', Precursor). This categorization is based on the choices you make when annotating. You can also see the annotations when using miRBase as the annotation source. In this example both the mature 5' and the mature 3' are annotated, and you can see that both are found in the sample. In the **Side Panel** to the right you can see the **Match weight** group under **Residue coloring** which is used to color the tags according to their relative abundance. The weight is also shown next to the name of the tag. The left side color is used for tags with low counts and the right side color is used for tags with high counts, relative to the total counts of this annotation reference. The sliders just above the gradient color box can be dragged to highlight relevant levels of abundance. The colors can be changed by clicking the box. This will show a list of gradients to choose from.

Create Sample from Selection (▦) This is used to create a new sample based on the tags that are selected. This can be useful in combination with filtering and sorting.



Read layout
Compactness: Not compact
☑ Gather sequences at top
☑ Show sequence ends
☑ Show mismatches
□ Disconnect pairs
Packed read height: Medium
Find Conflict
Low coverage threshold: 8
Find Low Coverage

Sequence layout
No spacing
☑ Numbers on sequences
Relative to: 1
☑ Numbers on plus strand
☑ Lock top sequence
□ Hide labels
☑ Lock labels
Sequence label: Name (weight)
□ Identical residues as dots

Annotation layout

Annotation types
🟩 ☑ miRNA
Select All
Deselect All

Residue coloring
Nonstandard residues
Rasmol colors
Trace colors
Sequence colors
Match weight
☑ Sequence color
Weight 0 30.689

Alignment info
Nucleotide info
Find
Text format

Figure 27.33: *Aligning all the variants of this miRNA from miRBase, providing a visual overview of the distribution of tags along the precursor sequence.*

27.2.5 Exploring novel miRNAs

One way of doing this would be to identify interesting tags based on their counts (typically you would be interested in pursuing tags with not too low counts in order to avoid wasting efforts on tags based on reads with sequencing errors), **Extract Small RNAs** (☰) and use this list of tags as input to **Map Reads to Reference** (☰) using the genome as reference. You could then examine where the reads match, and for reads that map in otherwise unannotated regions you could select a region around the match and create a subsequence from this. The subsequence could be folded and examined to see whether the secondary structure was in agreement with the expected hairpin-type structure for miRNAs. The *CLC Cancer Research Workbench* is able to analyze expression data produced on microarray platforms and high-throughput sequencing platforms (also known as Next-Generation Sequencing platforms). The *CLC Cancer Research Workbench* provides tools for performing quality control of the data, transformation and normalization, statistical analysis to measure differential expression and annotation-based tests. A number of visualization tools such as volcano plots, MA plots, scatter plots, box plots, and heat maps are used to aid the interpretation of the results.

27.3 Experimental design

In order to make full use of the various tools for interpreting expression data, you need to know the central concepts behind the way the data is organized in the *CLC Cancer Research Workbench*.

The first piece of data you are faced with is the **sample**. In the Workbench, a sample contains the expression values from either one array or from sequencing data of one sample.

Note that the calculation of expression levels based on the raw sequence data is described in section 27.1 and section 27.2.

See more below on how to get your expression data into the Workbench as samples in section G.

In a sample, there is a number of **features**, usually genes, and their associated expression levels.

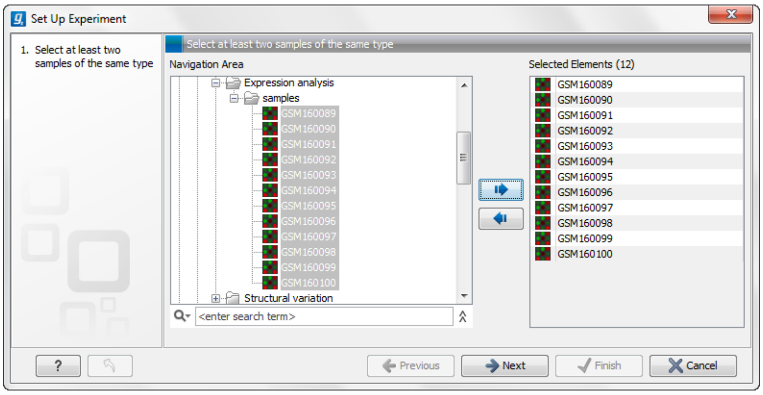
To analyze differential expression, you need to tell the workbench how the samples are related. This is done by setting up an **experiment**. An experiment is essentially a set of samples which are grouped. By creating an experiment defining the relationship between the samples, it becomes possible to do statistical analysis to investigate differential expression between the groups. The **Experiment** is also used to accumulate calculations like t-tests and clustering because this information is closely related to the grouping of the samples.

27.3.1 Setting up an experiment

To set up an experiment:

Toolbox | Transcriptomics Analysis ( )| Set Up Experiment ( )

Select the samples that you wish to use by double-clicking or selecting and pressing the **Add** (➡) button (see figure 27.34).



1. Select at least two samples of the same type

Navigation Area: GSM160089, GSM160090, GSM160091, GSM160092, GSM160093, GSM160094, GSM160095, GSM160096, GSM160097, GSM160098, GSM160099, GSM160100

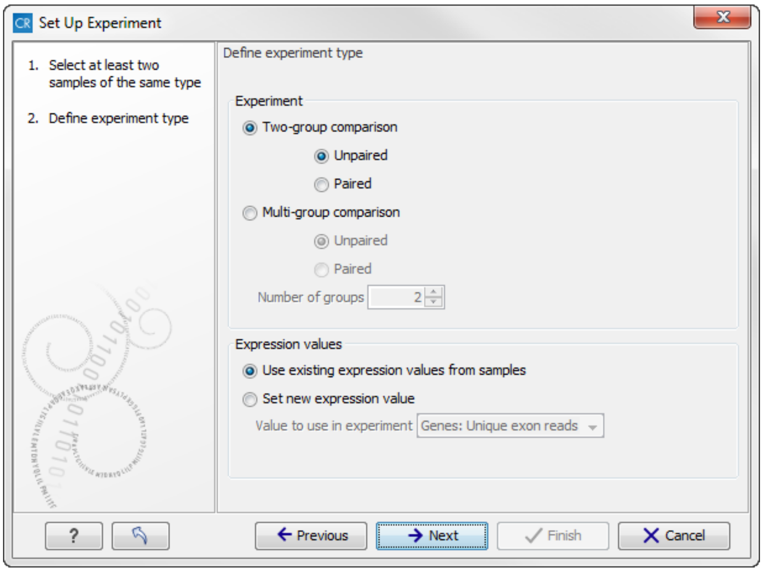
Selected Elements (12)

GSM160089
GSM160090
GSM160091
GSM160092
GSM160093
GSM160094
GSM160095
GSM160096
GSM160097
GSM160098
GSM160099
GSM160100

Figure 27.34: *Select the samples to use for setting up the experiment.*

Note that we use "samples" as the general term for both microarray-based sets of expression values and sequencing-based sets of expression values (e.g. an expression track from RNA-Seq).

Clicking **Next** shows the dialog in figure 27.35.



Set Up Experiment

1. Select at least two samples of the same type
2. Define experiment type

Define experiment type

Experiment

● Two-group comparison

● Unpaired

○ Paired

○ Multi-group comparison

● Unpaired

○ Paired

Number of groups 2

Expression values

● Use existing expression values from samples

○ Set new expression value

Value to use in experiment: Genes: Unique exon reads

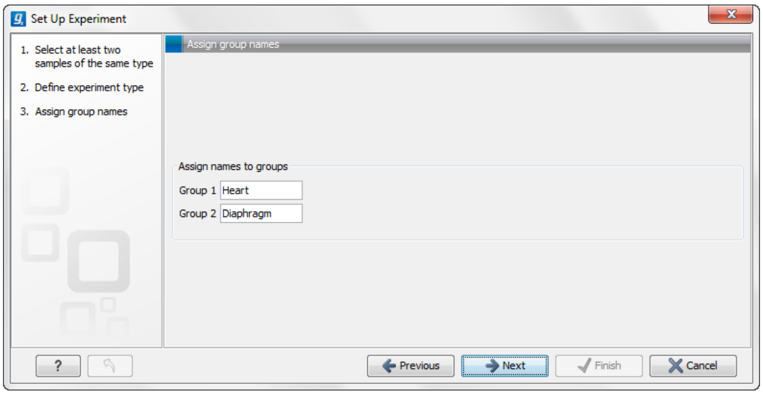
Figure 27.35: *Defining the number of groups.*

Here you define the number of groups in the experiment. At the top you can select a two-group experiment, and below you can select a multi-group experiment and define the number of groups.

Note that you can also specify if the samples are paired. Pairing is relevant if you have samples from the same individual under different conditions, e.g. before and after treatment, or at times 0, 2 and 4 hours after treatment. In this case statistical analysis becomes more efficient if effects of the individuals are taken into account, and comparisons are carried out not simply by considering *raw* group means but by considering these *corrected for* effects of the individual. If the **Paired** is selected, a paired rather than a standard t-test will be carried out for two group comparisons. For multiple group comparisons a repeated measures rather than a standard ANOVA will be used.

For RNA-Seq experiments, you can also choose which expression value to be used when setting up the experiment. This value will then be used for all subsequent analyses.

Clicking **Next** shows the dialog in figure 27.36.



Set Up Experiment

1. Select at least two samples of the same type
2. Define experiment type
3. Assign group names

Assign group names

Assign names to groups

Group 1: Heart

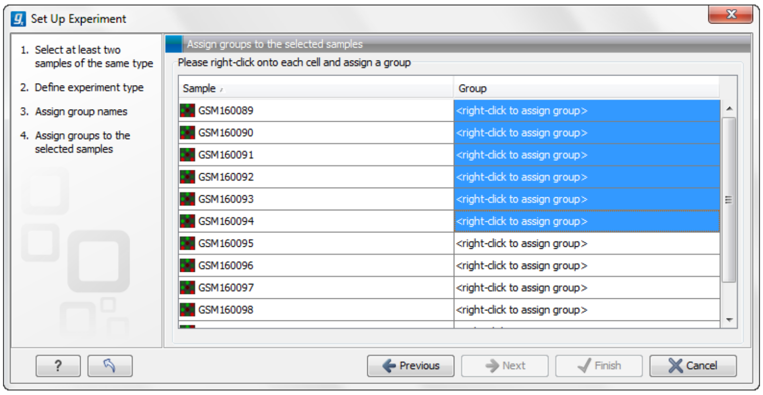
Group 2: Diaphragm

Figure 27.36: *Naming the groups.*

Depending on the number of groups selected in figure 27.35, you will see a list of groups with text fields where you can enter an appropriate name for that group.

For multi-group experiments, if you find out that you have too many groups, click the **Delete** (✖) button. If you need more groups, simply click **Add New Group**.

Click **Next** when you have named the groups, and you will see figure 27.37.



Set Up Experiment

1. Select at least two samples of the same type
2. Define experiment type
3. Assign group names
4. Assign groups to the selected samples

Assign groups to the selected samples

Please right-click onto each cell and assign a group

Sample	Group
GSM160089	<right-click to assign group>
GSM160090	<right-click to assign group>
GSM160091	<right-click to assign group>
GSM160092	<right-click to assign group>
GSM160093	<right-click to assign group>
GSM160094	<right-click to assign group>
GSM160095	<right-click to assign group>
GSM160096	<right-click to assign group>
GSM160097	<right-click to assign group>
GSM160098	<right-click to assign group>

Figure 27.37: *Putting the samples into groups.*

This is where you define which group the individual sample belongs to. Simply select one or more samples (by clicking and dragging the mouse), right-click (Ctrl-click on Mac) and select the appropriate group.

Note that the samples are sorted alphabetically based on their names.

If you have chosen **Paired** in figure 27.35, there will be an extra column where you define which samples belong together. Just as when defining the group membership, you select one or more samples, right-click in the pairing column and select a pair.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

27.3.2 Organization of the experiment table

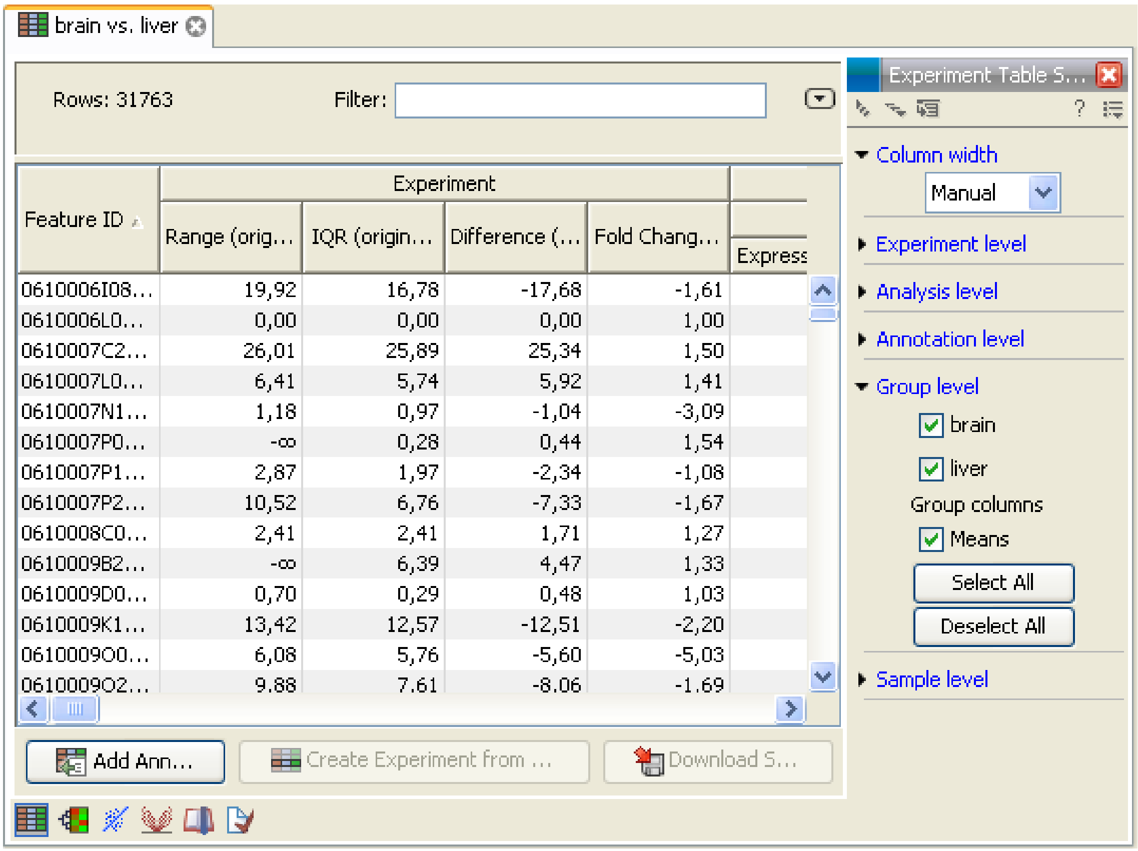
The resulting experiment includes all the expression values and other information from the samples (the values are copied - the original samples are not affected and can thus be deleted with no effect on the experiment). In addition it includes a number of summaries of the values across all, or a subset of, the samples for each feature. Which values are included is described in the sections below.

When you open it, it is shown in the experiment table (see figure 27.38).

For a general introduction to table features like sorting and filtering, see section 8.3.

Unlike other tables in *CLC Cancer Research Workbench*, the experiment table has a hierarchical grouping of the columns. This is done to reflect the structure of the data in the experiment. The **Side Panel** is divided into a number of groups corresponding to the structure of the table. These are described below. Note that you can customize and save the settings of the **Side Panel** (see section 4.5).

Whenever you perform analyses like normalization, transformation, statistical analysis etc, new columns will be added to the experiment. You can at any time **Export** () all the data in the



Rows: 31763 Filter:

Feature ID	Experiment				
	Range (orig...	IQR (origin...	Difference (...	Fold Chang...	Express
0610006I08...	19,92	16,78	-17,68	-1,61	
0610006L0...	0,00	0,00	0,00	1,00	
0610007C2...	26,01	25,89	25,34	1,50	
0610007L0...	6,41	5,74	5,92	1,41	
0610007N1...	1,18	0,97	-1,04	-3,09	
0610007P0...	-∞	0,28	0,44	1,54	
0610007P1...	2,87	1,97	-2,34	-1,08	
0610007P2...	10,52	6,76	-7,33	-1,67	
0610008C0...	2,41	2,41	1,71	1,27	
0610009B2...	-∞	6,39	4,47	1,33	
0610009D0...	0,70	0,29	0,48	1,03	
0610009K1...	13,42	12,57	-12,51	-2,20	
0610009O0...	6,08	5,76	-5,60	-5,03	
0610009O2...	9.88	7.61	-8.06	-1.69	

Figure 27.38: *Opening the experiment.*

experiment in csv or Excel format or **Copy** (◻) the full table or parts of it.

Column width

There are two options to specify the width of the columns and also the entire table:

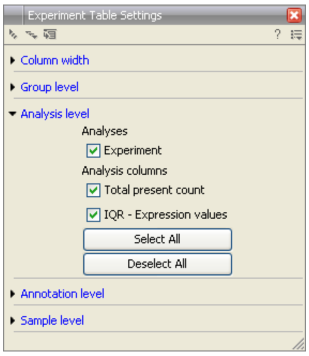
- **Automatic**. This will fit the entire table into the width of the view. This is useful if you only have a few columns.
- **Manual**. This will adjust the width of all columns evenly, and it will make the table as wide as it needs to be to display all the columns. This is useful if you have many columns. In this case there will be a scroll bar at the bottom, and you can manually adjust the width by dragging the column separators.

Experiment level

The rest of the **Side Panel** is devoted to different levels of information on the values in the experiment. The experiment part contains a number of columns that, for each feature ID, provide summaries of the values across all the samples in the experiment (see figure 27.39).

Initially, it has one header for the whole **Experiment**:

- **Range (original values)**. The 'Range' column contains the difference between the highest and the lowest expression value for the feature over all the samples. If a feature has the



Analyses

☑ Experiment

Analysis columns

☑ Total present count

☑ IQR - Expression values

Figure 27.39: *The initial view of the experiment level for a two-group experiment.*

value NaN in one or more of the samples the range value is NaN.

- **IQR (original values).** The 'IQR' column contains the inter-quantile range of the values for a feature across the samples, that is, the difference between the 75 %-ile value and the 25 %-ile value. For the IQR values, only the numeric values are considered when percentiles are calculated (that is, NaN and +Inf or -Inf values are ignored), and if there are fewer than four samples with numeric values for a feature, the IQR is set to be the difference between the highest and lowest of these.

- **Difference (original values).** For a two-group experiment the 'Difference' column contains the difference between the mean of the expression values across the samples assigned to group 2 and the mean of the expression values across the samples assigned to group 1. Thus, if the mean expression level in group 2 is higher than that of group 1 the 'Difference' is positive, and if it is lower the 'Difference' is negative. For experiments with more than two groups the 'Difference' contains the difference between the maximum and minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs before the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).

- **Fold Change (original values).** For a two-group experiment the 'Fold Change' tells you how many times bigger the mean expression value in group 2 is relative to that of group 1. If the mean expression value in group 2 is bigger than that in group 1 this value is the mean expression value in group 2 divided by that in group 1. If the mean expression value in group 2 is smaller than that in group 1 the fold change is the mean expression value in group 1 divided by that in group 2 with a negative sign. Thus, if the mean expression levels in group 1 and group 2 are 10 and 50 respectively, the fold change is 5, and if the and if the mean expression levels in group 1 and group 2 are 50 and 10 respectively, the fold change is -5. For experiments with more than two groups, the 'Fold Change' column contains the ratio of the maximum of the mean expression values of the groups to the minimum of the mean expression values of the groups, multiplied by -1 if the group with the maximum mean expression value occurs before the group with the minimum mean expression value (with the ordering: group 1, group 2, ...).

Thus, the sign of the values in the 'Difference' and 'Fold change' columns give the direction of the trend across the groups, going from group 1 to group 2, etc.

If the samples used are Affymetrix GeneChips samples and have 'Present calls' there will also be a 'Total present count' column containing the number of present calls for all samples.

The columns under the 'Experiment' header are useful for filtering purposes, e.g. you may wish to ignore features that differ too little in expression levels to be confirmed e.g. by qPCR by filtering on the values in the 'Difference', 'IQR' or 'Fold Change' columns or you may wish to ignore features that do not differ at all by filtering on the 'Range' column.

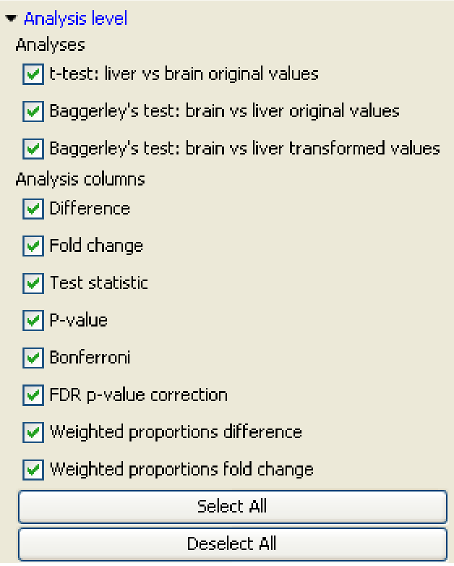
If you have performed normalization or transformation (see sections 27.5.3 and 27.5.2, respectively), the IQR of the normalized and transformed values will also appear. Also, if you later choose to transform or normalize your experiment, columns will be added for the transformed or normalized values.

Note! It is very common to filter features on fold change values in expression analysis and fold change values are also used in volcano plots, see section 27.7.2. There are different definitions of 'Fold Change' in the literature. The definition that is used typically depends on the original scale of the data that is analyzed. For data whose original scale is *not* the log scale the standard definition is the ratio of the group means [Tusher et al., 2001]. This is the value you find in the 'Fold Change' column of the experiment. However, for data whose original *is* the log scale, the difference of the mean expression levels is sometimes referred to as the fold change [Guo et al., 2006], and if you want to filter on fold change for these data you should filter on the values in the 'Difference' column. Your data's original scale will e.g. be the log scale if you have imported Affymetrix expression values which have been created by running the RMA algorithm on the probe-intensities.

Analysis level

If you perform statistical analysis (see section 27.7), there will be a heading for each statistical analysis performed. Under each of these headings you find columns holding relevant values for the analysis (P-value, corrected P-value, test-statistic etc. - see more in section 27.7).

An example of a more elaborate analysis level is shown in figure 27.40.



▾ Analysis level

Analyses

☑ t-test: liver vs brain original values

☑ Baggerley's test: brain vs liver original values

☑ Baggerley's test: brain vs liver transformed values

Analysis columns

☑ Difference

☑ Fold change

☑ Test statistic

☑ P-value

☑ Bonferroni

☑ FDR p-value correction

☑ Weighted proportions difference

☑ Weighted proportions fold change

Figure 27.40: *Transformation, normalization and statistical analysis has been performed.*

Annotation level

If your experiment is annotated (see section 27.3.4), the annotations will be listed in the **Annotation level** group as shown in figure 27.41.

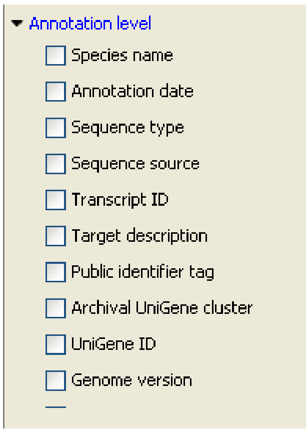


Figure 27.41: *An annotated experiment.*

In order to avoid too much detail and cluttering the table, only a few of the columns are shown per default.

Note that if you wish a different set of annotations to be displayed each time you open an experiment, you need to save the settings of the **Side Panel** (see section 4.5).

Group level

At the group level, you can show/hide entire groups (*Heart* and *Diaphragm* in figure 27.38). This will show/hide everything under the group's header. Furthermore, you can show/hide group-level information like the group means and present count within a group. If you have performed normalization or transformation (see sections 27.5.3 and 27.5.2, respectively), the means of the normalized and transformed values will also appear.

Sample level

In this part of the side panel, you can control which columns to be displayed for each sample. Initially this is the all the columns in the samples.

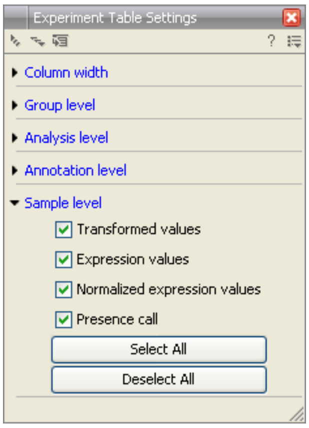
If you have performed normalization or transformation (see sections 27.5.3 and 27.5.2, respectively), the normalized and transformed values will also appear.

An example is shown in figure 27.42.

Creating a sub-experiment from a selection

If you have identified a list of genes that you believe are differentially expressed, you can create a subset of the experiment. (Note that the filtering and sorting may come in handy in this situation, see section 8.3).

To create a sub-experiment, first select the relevant features (rows). If you have applied a filter and wish to select all the visible features, press Ctrl + A (⌘ + A on Mac). Next, press the **Create**

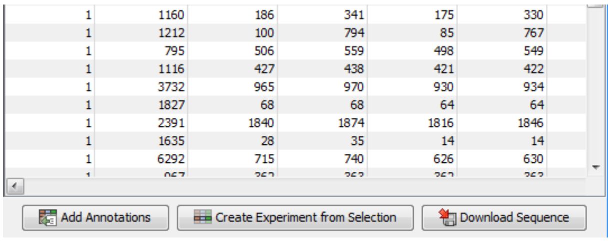


Experiment Table Settings

- ▸ Column width
- ▸ Group level
- ▸ Analysis level
- ▸ Annotation level
- ▾ Sample level
  - ☑ Transformed values
  - ☑ Expression values
  - ☑ Normalized expression values
  - ☑ Presence call

Figure 27.42: *Sample level when transformation and normalization has been performed.*

Experiment from Selection (▦) button at the bottom of the table (see figure 27.43).



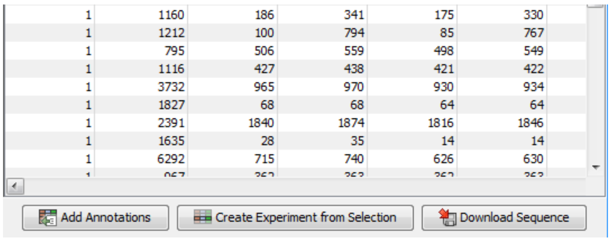
1	1160	186	341	175	330
1	1212	100	794	85	767
1	795	506	559	498	549
1	1116	427	438	421	422
1	3732	965	970	930	934
1	1827	68	68	64	64
1	2391	1840	1874	1816	1846
1	1635	28	35	14	14
1	6292	715	740	626	630

Figure 27.43: *Create a subset of the experiment by clicking the button at the bottom of the experiment table.*

This will create a new experiment that has the same information as the existing one but with less features.

Downloading sequences from the experiment table

If your experiment is annotated, you will be able to download the GenBank sequence for features which have a GenBank accession number in the 'Public identifier tag' annotation column. To do this, select a number of features (rows) in the experiment and then click **Download Sequence** (⬇) (see figure 27.44).



1	1160	186	341	175	330
1	1212	100	794	85	767
1	795	506	559	498	549
1	1116	427	438	421	422
1	3732	965	970	930	934
1	1827	68	68	64	64
1	2391	1840	1874	1816	1846
1	1635	28	35	14	14
1	6292	715	740	626	630

Figure 27.44: *Select sequences and press the download button.*

This will open a dialog where you specify where the sequences should be saved. You can learn more about opening and viewing sequences in chapter 9. You can now use the downloaded sequences for further analysis in the Workbench, e.g. performing BLAST searches and designing primers for QPCR experiments.

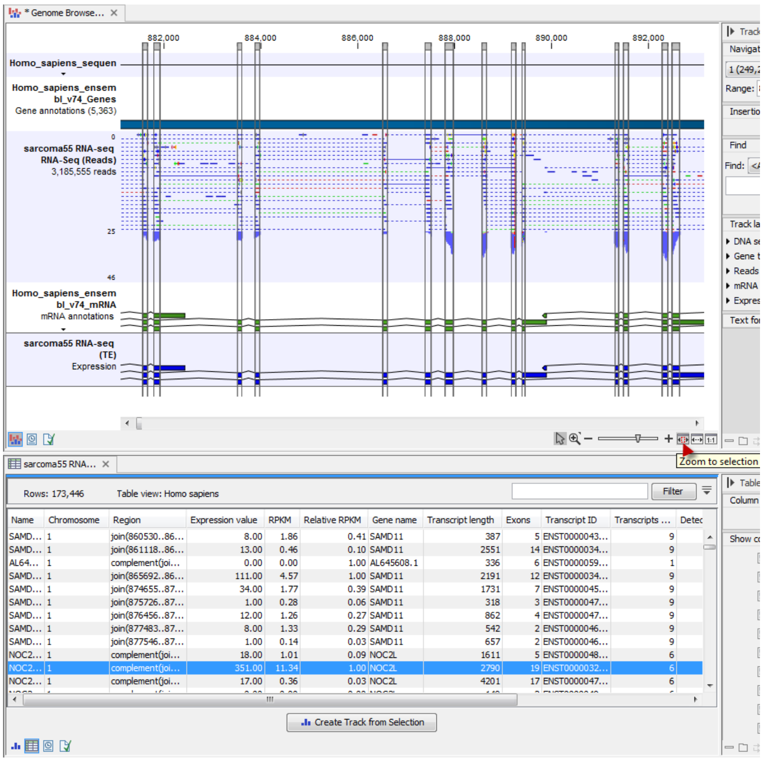
27.3.3 Visualizing RNA-Seq read tracks for the experiment

When working with RNA-Seq data, the experiment can be used to browse the read mappings to investigate how the reads supporting each sample are mapped. This is done by creating a track list:

File | New | Track List (📊**)**

Select the mapping and expression tracks of the samples you wish to visualize together and select any annotation tracks (e.g. gene and mRNA) to be included for visualization **Finish**.

Once the track list is shown, create a split view or drag the tab of the view on to a second screen (if you have two screens). Clicking a row in the table makes the track list view jump to that location, allowing for quick inspection of interesting parts of the RNA-Seq read mapping (see an example in figure 27.45. Note that the **Zoom to selection** (⊡) button can be used to adjust the zoom level to fit the region selection.



Rows: 173,446 Table view: Homo sapiens

Name	Chromosome	Region	Expression value	RPKM	Relative RPKM	Gene name	Transcript length	Exons	Transcript ID	Transcripts ...	Detec
SAMD...	1	join(860530..86...	8.00	1.86	0.41	SAMD11	387	5	ENST0000043...	9	
SAMD...	1	join(861118..86...	13.00	0.46	0.10	SAMD11	2551	14	ENST0000034...	9	
AL64...	1	complement(joi...	0.00	0.00	1.00	AL645608.1	336	6	ENST0000059...	1	
SAMD...	1	join(865692..86...	111.00	4.57	1.00	SAMD11	2191	12	ENST0000034...	9	
SAMD...	1	join(874655..87...	34.00	1.77	0.39	SAMD11	1731	7	ENST0000045...	9	
SAMD...	1	join(875726..87...	1.00	0.28	0.06	SAMD11	318	3	ENST0000047...	9	
SAMD...	1	join(876456..87...	12.00	1.26	0.27	SAMD11	862	4	ENST0000047...	9	
SAMD...	1	join(877483..87...	8.00	1.33	0.29	SAMD11	542	2	ENST0000046...	9	
SAMD...	1	join(877546..87...	1.00	0.14	0.03	SAMD11	657	2	ENST0000046...	9	
NOC2...	1	complement(joi...	18.00	1.01	0.09	NOC2L	1611	5	ENST0000048...	6	
NOC2...	1	complement(joi...	351.00	11.34	1.00	NOC2L	2790	19	ENST0000032...	6	
NOC2...	1	complement(joi...	17.00	0.36	0.03	NOC2L	4201	17	ENST0000047...	6	

Figure 27.45: *RNA-Seq results shown in a split view with an experiment table at the bottom and a track list with read mappings of several samples at the top.*

Please note that at least one of the expression tracks used in the experiment have to be included in the track list in order for the link between the two to work.

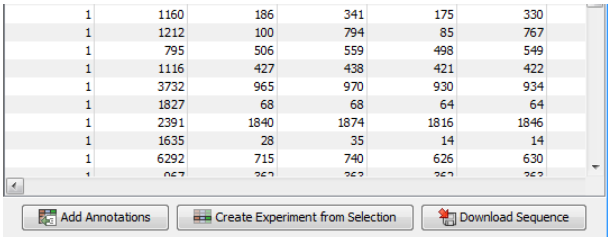
27.3.4 Adding annotations to an experiment

Annotation files provide additional information about each feature. This information could be which GO categories the protein belongs to, which pathways, various transcript and protein identifiers etc. See section G for information about the different annotation file formats that are supported *CLC Cancer Research Workbench*.

The annotation file can be imported into the Workbench and will get a special icon (). See an overview of annotation formats supported by *CLC Cancer Research Workbench*in section G. In order to associate an annotation file with an experiment, either select the annotation file when you set up the experiment (see section 27.3.1), or click:

Toolbox | Transcriptomics Analysis ()| Annotation Test | Add Annotations ()

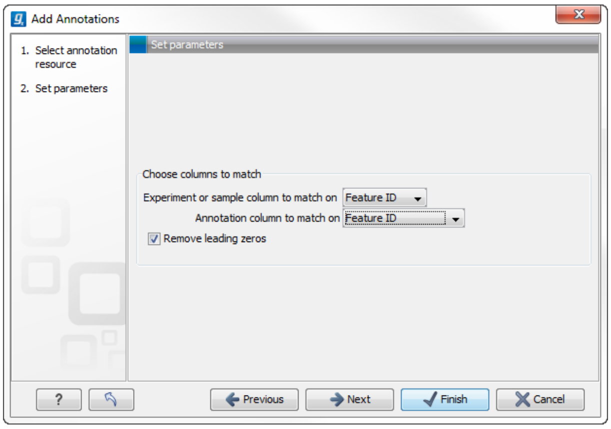
Select the experiment () and the annotation file () and click **Finish**. You will now be able to see the annotations in the experiment as described in section 27.3.2. You can also add annotations by pressing the **Add Annotations** () button at the bottom of the table (see figure 27.46).



1	1160	186	341	175	330
1	1212	100	794	85	767
1	795	506	559	498	549
1	1116	427	438	421	422
1	3732	965	970	930	934
1	1827	68	68	64	64
1	2391	1840	1874	1816	1846
1	1635	28	35	14	14
1	6292	715	740	626	630

Figure 27.46: *Adding annotations by clicking the button at the bottom of the experiment table.*

This will bring up a dialog where you can select the annotation file that you have imported together with the experiment you wish to annotate. Click **Next** to specify settings as shown in figure 27.47).



Choose columns to match

Experiment or sample column to match on: Feature ID

Annotation column to match on: Feature ID

☑ Remove leading zeros

Figure 27.47: *Choosing how to match annotations with samples.*

In this dialog, you can specify how to match the annotations to the features in the sample. The Workbench looks at the columns in the annotation file and lets you choose which column that should be used for matching to the feature IDs in the experimental data (experiment or sample) as well as for the annotations. Usually the default is right, but for some annotation files, you need to select another column.

Some annotation files have leading zeros in the identifier which you can remove by checking the **Remove leading zeros** box.

Note! Existing annotations on the experiment will be overwritten.

27.3.5 Scatter plot view of an experiment

At the bottom of the experiment table, you can switch between different views of the experiment (see figure 27.48).

Figure 27.48: *An experiment can be viewed in several ways.*

One of the views is the **Scatter Plot** (⁘). The scatter plot can be adjusted to show e.g. the group means for two groups (see more about how to adjust this below).

An example of a scatter plot is shown in figure 27.49.

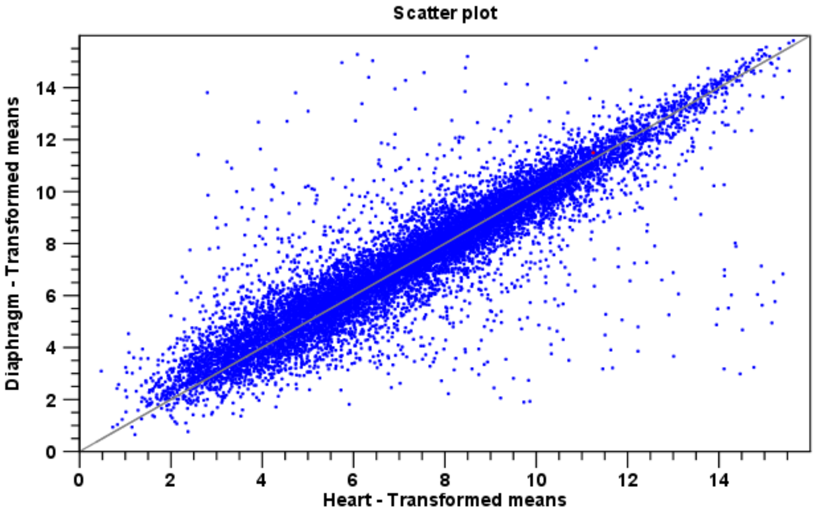


Figure 27.49: *A scatter plot of group means for two groups (transformed expression values).*

In the **Side Panel** to the left, there are a number of options to adjust this view. Under **Graph preferences**, you can adjust the general properties of the scatter plot:

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks

- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Draw x = y axis**. This will draw a diagonal line across the plot. This line is shown per default.
- **Line width**
  - Thin
  - Medium
  - Wide
- **Line type**
  - None
  - Line
  - Long dash
  - Short dash
- **Line color.** Allows you to choose between many different colors. Click the color box to select a color.
- Show Pearson correlation When checked, the Pearson correlation coefficient (r) is displayed on the plot.

Below the general preferences, you find the **Dot properties** preferences, where you can adjust coloring and appearance of the dots:

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.

Finally, the group at the bottom - **Values to plot** - is where you choose the values to be displayed in the graph. The default for a two-group experiment is to plot the group means.

Note that if you wish to use the same settings next time you open a scatter plot, you need to save the settings of the **Side Panel** (see section 4.5).

27.3.6 Cross-view selections

There are a number of different ways of looking at an experiment as shown in figure 27.50).

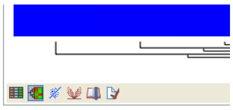


Figure 27.50: *An experiment can be viewed in several ways.*

Beside the **Experiment table** (▦) which is the default view, the views are: **Scatter plot** (⁘), **Volcano plot** (⋎) and the **Heat map** (▨). By pressing and holding the Ctrl (⌘ on Mac) button while you click one of the view buttons in figure 27.50, you can make a split view. This will make it possible to see e.g. the experiment table in one view and the volcano plot in another view.

An example of such a split view is shown in figure 27.51.

Selections are shared between all these different views of an experiment. This means that if you select a number of rows in the table, the corresponding dots in the scatter plot, volcano plot or heatmap will also be selected. The selection can be made in any view, also the heat map, and all other open views will reflect the selection.

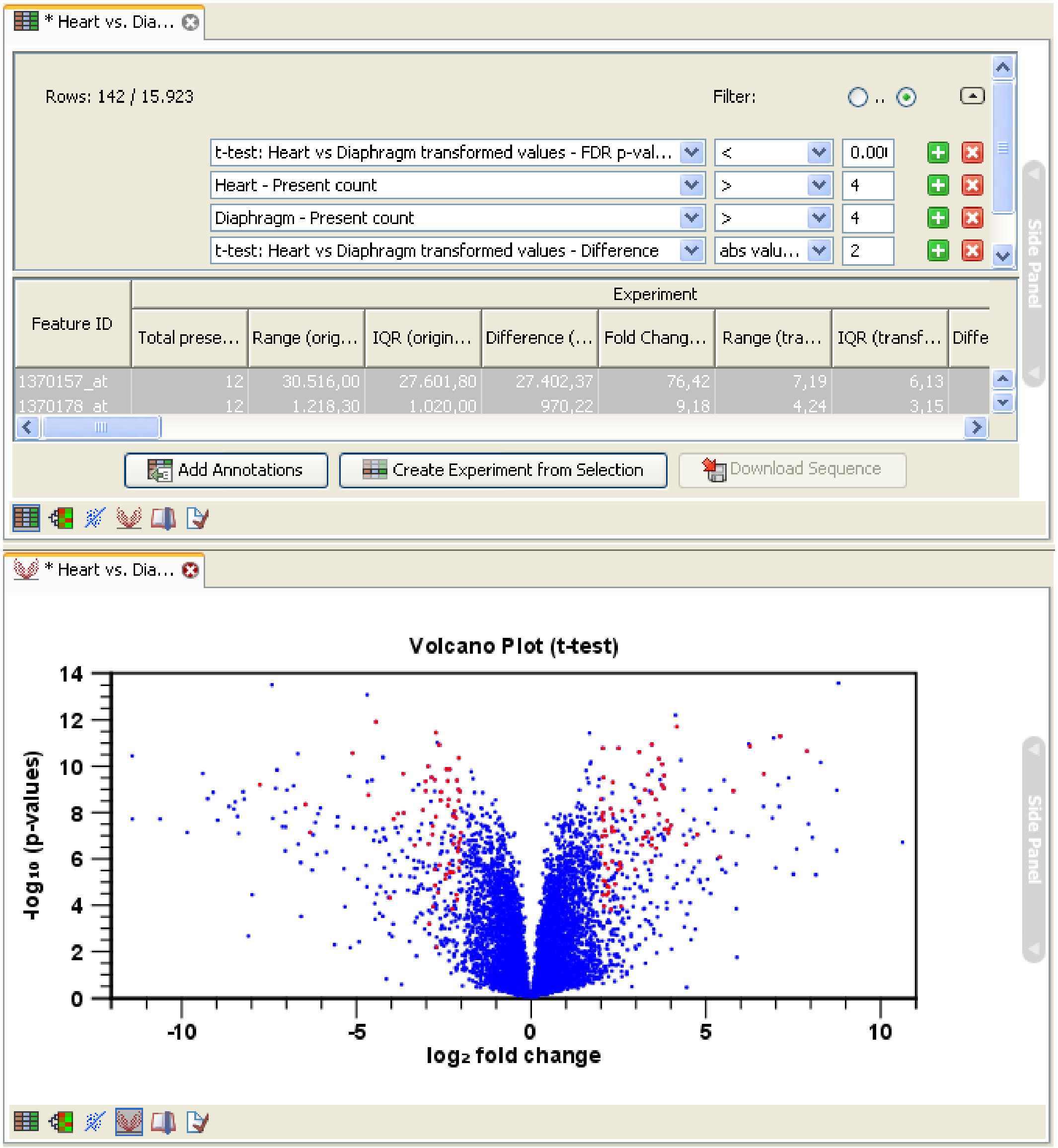
A common use of the split views is where you have an experiment and have performed a statistical analysis. You filter the experiment to identify all genes that have an FDR corrected p-value below 0.05 and a fold change for the test above say, 2. You can select all the rows in the experiment table satisfying these filters by holding down the Cntrl button and clicking 'a'. If you have a split view of the experiment and the volcano plot all points in the volcano plot corresponding to the selected features will be red. Note that the volcano plot allows two sets of values in the columns under the test you are considering to be displayed on the x-axis: the 'Fold change's and the 'Difference's. You control which to plot in the side panel. If you have filtered on 'Fold change' you will typically want to choose 'Fold change' in the side panel. If you have filtered on 'Difference' (e.g. because your original data is on the log scale, see the note on fold change in 27.3.2) you typically want to choose 'Difference'.

27.4 Working with tracks and experiments

The *CLC Cancer Research Workbench* provides several tools for the analysis, organization, and visualization of expression data. In this section, we describe how Tracks and Experiments complement each other, and how they can be used together for the analysis of transcriptomics data using the tools found in the **Transcriptomics** toolbox.

27.4.1 Data structures for transcriptomics

The two main data structures used for transcriptomics data analysis in the *CLC Cancer Research Workbench* are tracks and experiments. Tracks, also known as 'Genome Browser View', (see



Rows: 142 / 15.923

t-test: Heart vs Diaphragm transformed values - FDR p-val...	<	0.001
Heart - Present count	>	4
Diaphragm - Present count	>	4
t-test: Heart vs Diaphragm transformed values - Difference	abs valu...	2

Feature ID	Experiment							
	Total prese...	Range (orig...	IQR (origin...	Difference (...	Fold Chang...	Range (tra...	IQR (transf...	Diffe
1370157_at	12	30.516,00	27.601,80	27.402,37	76,42	7,19	6,13	
1370178_at	12	1.218,30	1.020,00	970,22	9,18	4,24	3,15	

Figure 27.51: *A split view showing an experiment table at the top and a volcano plot at the bottom (note that you need to perform statistical analysis to show a volcano plot, see section 27.7).*

section 17) are the fundamental building blocks for data analysis in the *CLC Cancer Research Workbench*, where all information is tied to genomic positions. A central coordinate-system is provided by a reference genome, which allows that different types of data or results for different samples can be seen and analyzed together.

Experiments (see section 27.3), on the other hand, are used to represent complex relationships between expression samples, and to carry out statistical analysis (see section 27.7) of differential expression.

Tracks and experiments are intimately related, and it is possible in most cases to convert from one type to the other.

From Tracks to Experiments

When carrying out RNA-seq analysis using the **RNA-Seq Analysis** tool, the starting point is a set of reads from a sequencing study. As part of the RNA-Seq Analysis tool, these reads are mapped onto a reference genome. The RNA-Seq tool produces expression tracks, which are compatible with the reference genome, and can be visualized together with the genome in the **Genome Browser View** (see section 17). You can find more information about the RNA-Seq Analysis tool in section 27.1).

Once expression tracks have been obtained from the RNA-Seq Analysis tool, they can be used as sequencing-based sets of expression values in setting up an experiment. This can be done using the **Set up Experiment** tool, also found in the **Transcriptomics** toolbox. This is described in more detail in section 27.3.

An experiment set up in this manner from expression tracks is intimately coupled to the tracks it originated from. To see this coupling in action, perform the following steps:

1. Use the **Set up Experiment** tool on two or more expression tracks to set up an experiment, as described in section 27.3.
2. Save and open the resulting experiment, by double-clicking its name in the **Navigation Area**.
3. Use the **Create New Genome Browser View** tool to create a new Genome Browser View from the expression tracks you used to set up the experiment, as described in section 17.
4. Save and open the resulting Genome Browser View by double-clicking its name in the **Navigation Area**.
5. Drag the experiment tab downwards, until you see the blue shadow indicating the resulting placement (figure 27.52), and drop it in place. You should now have a divided view, with the experiment in the bottom half (figure 27.53).
6. Clicking on any line in the experiment will now automatically jump to the corresponding genomic location in the upper view. Use the **Zoom to Selection** ( ) button to zoom in to the desired genomic region.

From Experiments to Tracks

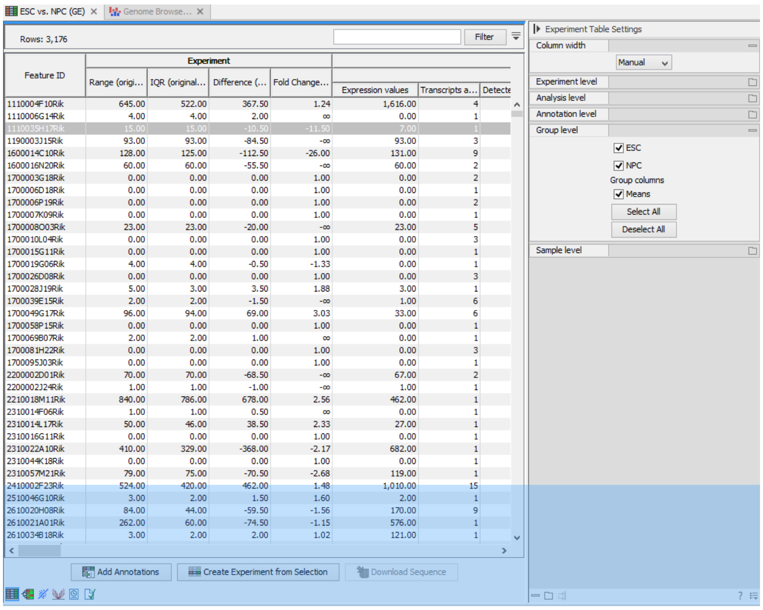
Experiments can be used to carry out statistical analysis on the expression values obtained from RNA-seq analysis as described in section 27.7. The results of the statistical analysis are annotated on the experiment as additional columns.

It can be advantageous to visualize the results of the statistical analysis as tracks. The **Create Track from Experiment** tool in the *CLC Cancer Research Workbench* enables the conversion of experiments to tracks. You can find the **Create Track from Experiment** tool here:

Toolbox | **Transcriptomics Analysis** ( ) | **Create Track from Experiment** ( ) .

27.4.2 Running the Create Track from Experiment tool

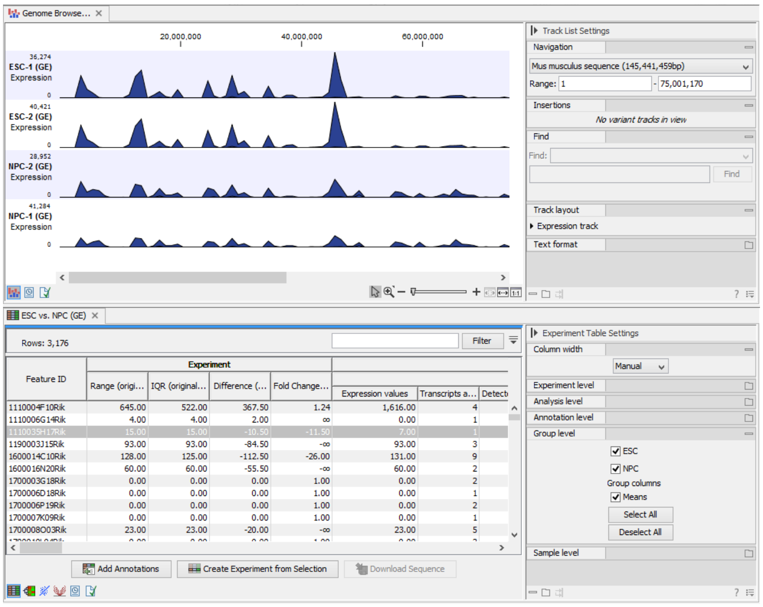
After you start the tool, you are presented with a wizard where you can choose the experiment that you would like to create a track of. The **Create Track from Experiment** tool can be run on two types of experiments:



Rows: 3,176

Feature ID	Experiment						
	Range (origi...	IQR (original...	Difference (...	Fold Change...	Expression values	Transcripts a...	Detecte
1110004F10Rik	645.00	522.00	367.50	1.24	1,616.00	4	
1110006G14Rik	4.00	4.00	2.00	∞	0.00	1	
1110035H17Rik	15.00	15.00	-10.50	-11.50	7.00	1	
1190003J15Rik	93.00	93.00	-84.50	-∞	93.00	3	
1600014C10Rik	128.00	125.00	-112.50	-26.00	131.00	9	
1600016N20Rik	60.00	60.00	-55.50	-∞	60.00	2	
1700003G18Rik	0.00	0.00	0.00	1.00	0.00	2	
1700006D18Rik	0.00	0.00	0.00	1.00	0.00	1	
1700006P19Rik	0.00	0.00	0.00	1.00	0.00	2	
1700007K09Rik	0.00	0.00	0.00	1.00	0.00	1	
1700008O03Rik	23.00	23.00	-20.00	-∞	23.00	5	
1700010L04Rik	0.00	0.00	0.00	1.00	0.00	3	
1700015G11Rik	0.00	0.00	0.00	1.00	0.00	1	
1700019G06Rik	4.00	4.00	-0.50	-1.33	0.00	1	
1700026D08Rik	0.00	0.00	0.00	1.00	0.00	3	
1700028J19Rik	5.00	3.00	3.50	1.88	3.00	1	
1700039E15Rik	2.00	2.00	-1.50	-∞	1.00	6	
1700049G17Rik	96.00	94.00	69.00	3.03	33.00	6	
1700058P15Rik	0.00	0.00	0.00	1.00	0.00	1	
1700069B07Rik	2.00	2.00	1.00	∞	0.00	1	
1700081H22Rik	0.00	0.00	0.00	1.00	0.00	3	
1700095J03Rik	0.00	0.00	0.00	1.00	0.00	1	
2200002D01Rik	70.00	70.00	-68.50	-∞	67.00	2	
2200002J24Rik	1.00	1.00	-1.00	-∞	1.00	1	
2210018M11Rik	840.00	786.00	678.00	2.56	462.00	1	
2310014F06Rik	1.00	1.00	0.50	∞	0.00	1	
2310014L17Rik	50.00	46.00	38.50	2.33	27.00	1	
2310016G11Rik	0.00	0.00	0.00	1.00	0.00	1	
2310022A10Rik	410.00	329.00	-368.00	-2.17	682.00	1	
2310044K18Rik	0.00	0.00	0.00	1.00	0.00	1	
2310057M21Rik	79.00	75.00	-70.50	-2.68	119.00	1	
2410002F23Rik	524.00	420.00	462.00	1.48	1,010.00	15	
2510046G10Rik	3.00	2.00	1.50	1.60	2.00	1	
2610020H08Rik	84.00	44.00	-59.50	-1.56	170.00	9	
2610021A01Rik	262.00	60.00	-74.50	-1.15	576.00	1	
2610034B18Rik	3.00	2.00	2.00	1.02	121.00	1	

Figure 27.52: *Dragging a tab to the lower half of the view area.*



Mus musculus sequence (145,441,459bp)

Range: 1 - 75,001,170

Rows: 3,176

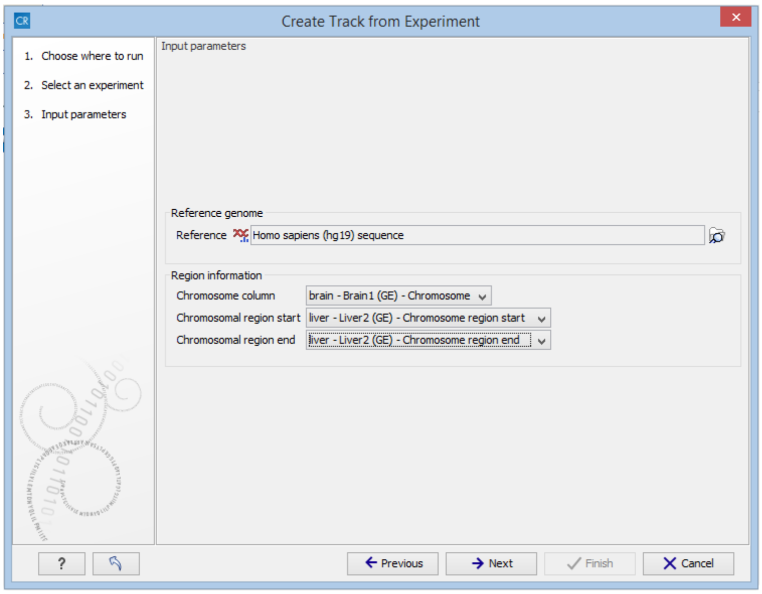
Feature ID	Experiment						
	Range (origi...	IQR (original...	Difference (...	Fold Change...	Expression values	Transcripts a...	Detecte
1110004F10Rik	645.00	522.00	367.50	1.24	1,616.00	4	
1110006G14Rik	4.00	4.00	2.00	∞	0.00	1	
1110035H17Rik	15.00	15.00	-10.50	-11.50	7.00	1	
1190003J15Rik	93.00	93.00	-84.50	-∞	93.00	3	
1600014C10Rik	128.00	125.00	-112.50	-26.00	131.00	9	
1600016N20Rik	60.00	60.00	-55.50	-∞	60.00	2	
1700003G18Rik	0.00	0.00	0.00	1.00	0.00	2	
1700006D18Rik	0.00	0.00	0.00	1.00	0.00	1	
1700006P19Rik	0.00	0.00	0.00	1.00	0.00	2	
1700007K09Rik	0.00	0.00	0.00	1.00	0.00	1	
1700008O03Rik	23.00	23.00	-20.00	-∞	23.00	5	

Figure 27.53: *After dropping a tab to the lower half othe view area.*

1. Experiments with associated genomic information, such as those created using expression tracks from the **RNA-Seq Analysis** tool.
2. Experiments without associated genomic information, such as those created using samples from the legacy RNA-Seq Analysis tool.

In the case where the experiment has associated genomic information, the **Create Track from Experiment** tool will automatically infer these and the wizard will jump directly to the Filtering step, as shown in figure 27.55.

In the case where the experiment does not have associated genomic information, you will first need to specify how the genomic information should be obtained in the Parameters step of the **Create Track from Experiment** tool (figure 27.56).



Reference: Homo sapiens (hg19) sequence

Chromosome column: brain - Brain1 (GE) - Chromosome

Chromosomal region start: liver - Liver2 (GE) - Chromosome region start

Chromosomal region end: liver - Liver2 (GE) - Chromosome region end

Figure 27.54: *The parameter step in the Create Track from Experiment tool.*

In the Input parameters step, you must specify the following parameters:

- **Reference genome.** The chosen genome will be used as the reference genome for the resulting track.
- **Chromosome column.** The column containing the chromosome names must be chosen from the drop-down menu.
- **Chromosomal region start.** The column containing the start of the genomic regions must be chosen from the drop-down menu.
- **Chromosomal region end.** The column containing the end of the genomic regions must be chosen from the drop-down menu.

Note! The drop-down menus will only contain the columns that potentially represent the information required by the given parameter. If the experiment does not contain any columns that potentially represent the required genomic information, the drop-down menus may appear empty. In this case, it is not possible to convert the given experiment to a track.

In the Filtering step (figure 27.55), you have the following options:

- **Filter based on statistical analysis results** This allows to filter which annotations are transferred to the track on the basis of the statistical analysis. To enable filtering, check the **Filter based on statistical analysis results** checkbox. The filtering option is only available if a statistical analysis has previously been carried out on the Experiment, and the drop-down menu will only contain the statistical analyses that are present on the Experiment.
- **Statistical analysis** Allows you to choose statistical analysis from the drop-down list. The selection of available statistical analyses depends on which tests have been used when you set up the experiment that you are about to convert to track format.

- **Type of p-value** This drop-down menu allows you to select between raw and corrected p-values (see section **??**). Only the types of p-values available for the given statistical analysis will be present in the drop-down menu.
- **Maximum p-value** In this input field, you can enter the maximum allowed p-value, as a number between 0 and 1. Any If you do not want any filtering based on p-value, enter 1.
- **Minimum fold-change value** You can also specify the minimum allowed fold-change value as a number greater than zero. If you do not want any filtering based on fold-change, enter 0.

You can then select in the drop-down menu which analysis you want to use for filtering.

The fold change values are stored as different columns in the experiment, depending on the statistical analysis performed. The Create Track from Experiment tool will automatically use the fold-change column appropriate for the different statistical analyses:

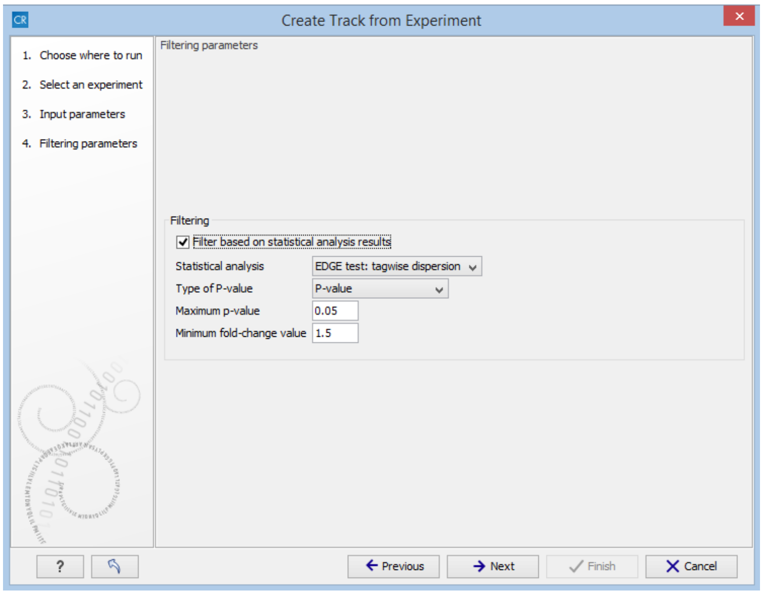
- Kal's Z-test (see section **??**): Proportions fold change.
- Baggerley's test(see section **??**): Weighted proportions fold change.
- T-test (see section 27.7.1): Fold change.
- ANOVA (see section **??**): Max fold change.
- Empirical analysis of DGE (see section 27.7.1): Fold change.

The resulting track will contain only differentially expressed genes whose p-value is lower than the specified threshold and whose fold-enrichment is above the specified threshold.

If the chosen statistical analysis was performed on several pairs of groups, there will be an output track for each tested pair of groups. For example, if the same statistical analysis has been carried out on 'group 1 vs. group 2' and 'group 1 vs. group 3', then the output will contain two tracks, where one is filtered according to the 'group 1 vs. group 2' analysis results and the other one is filtered according to the 'group 1 vs. group 3' analysis results.

When running the **Create Track from Experiment** tool as part of a workflow, there are a few differences in how the parameters are set (see figure 27.57).

- The **Source of genomic information** parameter determines the behavior of the algorithm if the incoming experiment is *not* coupled to a genome. If the value of this parameter is set to **Require genomic information in experiment**, then the algorithm will expect the incoming experiment to be coupled to a genome, and will fail with an error alerting the user in case the experiment does not fulfill this criterion. If the value of the parameter is set to **Automatic: use genomic information if available**, then the algorithm will still use the genomic information in a genome-coupled experiment. But if this information is not available, the algorithm will attempt to use the information specified by the user in the workflow parameters. *Note:* If the incoming experiment *is* coupled to a genome (as will usually be the case), the value of this parameter makes no difference.
- In a workflow setting, the column titles for the chromosome, region end and region start fields can be specified as texts. These fields may be left empty, if the incoming experiment



Create Track from Experiment

1. Choose where to run
2. Select an experiment
3. Input parameters
4. Filtering parameters

Filtering parameters

Filtering

☑ Filter based on statistical analysis results

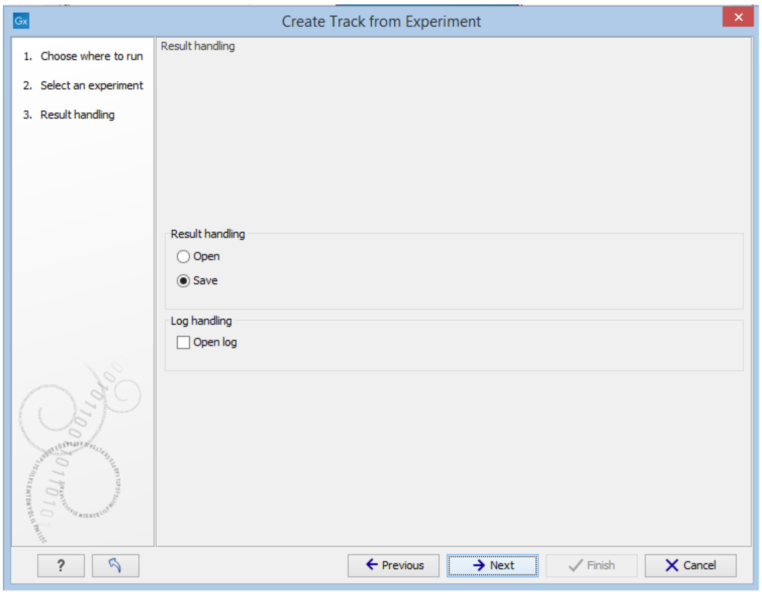
Statistical analysis: EDGE test: tagwise dispersion

Type of P-value: P-value

Maximum p-value: 0.05

Minimum fold-change value: 1.5

Figure 27.55: *The filtering step in the Create Track from Experiment tool.*



Create Track from Experiment

1. Choose where to run
2. Select an experiment
3. Result handling

Result handling

Result handling

○ Open

● Save

Log handling

□ Open log

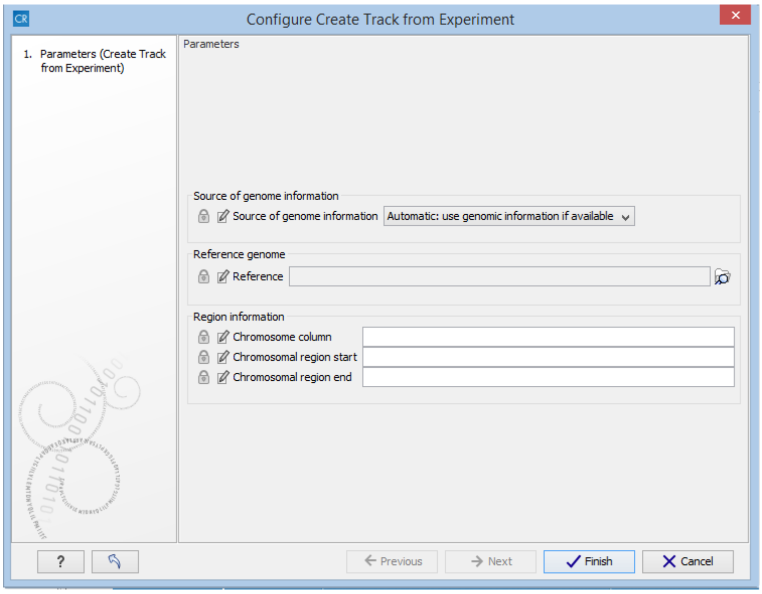
Figure 27.56: *The result handling step in the Create Track from Experiment tool.*

contains the genomic information. If filling out these fields, note that the format for this text is very strict, and must exactly match the text appearing in the drop-down menu when running the tool from the toolbox. For example, if 'Chromosome' is a sample-specific column, for a sample called 'Liver (GE)' in the 'liver' group in the experiment, then the column name text will be: 'liver - Liver (GE) - Chromosome'.

27.4.3 Interpreting the results of the Create Track from Experiment tool

The **Create Track from Experiment** tool will produce a track or several tracks, if filtering based on analysis results was chosen. The track(s) will contain the following annotations:

- All experiment-specific columns from the experiment



Configure Create Track from Experiment

1. Parameters (Create Track from Experiment)

Parameters

Source of genome information

Source of genome information: Automatic: use genomic information if available

Reference genome

Reference:

Region information

Chromosome column:

Chromosomal region start:

Chromosomal region end:

Figure 27.57: *Setting the parameters for the Create Track from Experiment tool in a workflow*

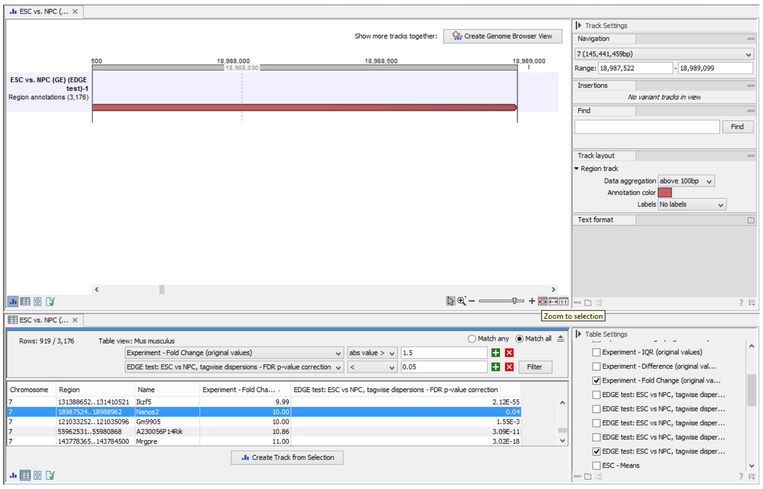
- All user-defined annotations added to the experiment
- All analysis-specific columns from the experiment
- All group-specific columns from the experiment
- Those of the following sample-specific columns when present in the experiment (for each sample): Expression values, Total exon reads, and RPKM.

Two different view options exist: the Genome Browser View and the Table View. When opening the annotated output result, the default view is the Genome Browser View. It is possible to open both views in split view by holding down the Ctrl key while clicking on the table icon in the lower left corner of the View Area. The two different views are linked together. This means that when you click once on an entry in the table, the Genome Browser View will jump the the selected region. With the **Zoom to Selection** ( ) button it is possible to jump to and zoom in on the selected region (figure 27.58).

The results of any statistical test executed on the experiment, including fold-changes and p-values, can be seen in the tooltip when hovering over each region in the annotation track shown in Genome Browser View (figure 27.59).

27.5 Transformation and normalization

The original expression values often need to be transformed and/or normalized in order to ensure that samples are comparable and assumptions on the data for analysis are met [Allison et al., 2006]. These are essential requirements for carrying out a meaningful analysis. The raw expression values often exhibit a strong dependency of the variance on the mean, and it may be preferable to remove this by log-transforming the data. Furthermore, the sets of expression values in the different samples in an experiment may exhibit systematic differences that are likely due to differences in sample preparation and array processing, rather being the result of the underlying biology. These noise effects should be removed before statistical analysis is carried out.


ESC vs. NPC (...
Show more tracks together:
Create Genome Browser View
500
18,988,000
18,988,030
18,988,500
18,989,000
ESC vs. NPC (GE) (EDGE test)-1
Region annotations (3,176)
Zoom to selection
ESC vs. NPC (...
Create Track from Selection
Track Settings
Table Settings
Find
Filter

Track Settings

Navigation: 7 (145,441,459bp)
Range: 18,987,522 - 18,989,099

Insertions: *No variant tracks in view*

Track layout
- Region track
  - Data aggregation: above 100bp
  - Annotation color: 🟥
  - Labels: No labels

Text format

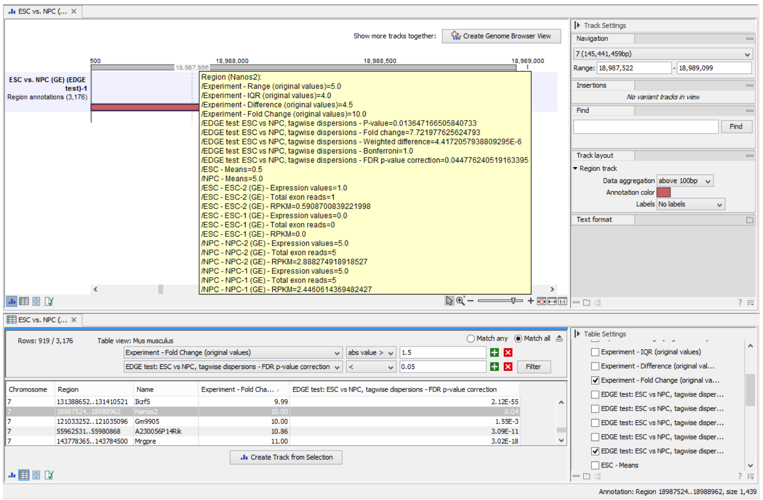
Rows: 919 / 3,176 Table view: Mus musculus ○ Match any ● Match all

Experiment - Fold Change (original values) | abs value > | 1.5
EDGE test: ESC vs NPC, tagwise dispersions - FDR p-value correction | < | 0.05

Chromosome	Region	Name	Experiment - Fold Cha...	EDGE test: ESC vs NPC, tagwise dispersions - FDR p-value correction
7	131388652..131410521	Ikzf5	9.99	2.12E-55
7	18987524..18988962	Nanos2	10.00	0.04
7	121033252..121035096	Gm9905	10.00	1.55E-3
7	55962531..55980868	A230056P14Rik	10.86	3.09E-11
7	143778365..143784500	Mrgpre	11.00	3.02E-18

Table Settings
- ☐ Experiment - IQR (original values)
- ☐ Experiment - Difference (original val...
- ☑ Experiment - Fold Change (original va...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☑ EDGE test: ESC vs NPC, tagwise disper...
- ☐ ESC - Means

Figure 27.58: *Viewing the track produced by the Create Track from Experiment Tool*


ESC vs. NPC (...
Show more tracks together:
Create Genome Browser View
500
18,987,866
18,988,000
18,988,500
18,989,000
ESC vs. NPC (GE) (EDGE test)-1
Region annotations (3,176)
ESC vs. NPC (...
Create Track from Selection
Track Settings
Table Settings
Find
Filter

```
Region (Nanos2):
/Experiment - Range (original values)=5.0
/Experiment - IQR (original values)=4.0
/Experiment - Difference (original values)=4.5
/Experiment - Fold Change (original values)=10.0
/EDGE test: ESC vs NPC, tagwise dispersions - P-value=0.013647166505840733
/EDGE test: ESC vs NPC, tagwise dispersions - Fold change=7.721977625624793
/EDGE test: ESC vs NPC, tagwise dispersions - Weighted difference=4.4172057938809295E-6
/EDGE test: ESC vs NPC, tagwise dispersions - Bonferroni=1.0
/EDGE test: ESC vs NPC, tagwise dispersions - FDR p-value correction=0.044776240519163395
/ESC - Means=0.5
/NPC - Means=5.0
/ESC - ESC-2 (GE) - Expression values=1.0
/ESC - ESC-2 (GE) - Total exon reads=1
/ESC - ESC-2 (GE) - RPKM=0.5908700839221998
/ESC - ESC-1 (GE) - Expression values=0.0
/ESC - ESC-1 (GE) - Total exon reads=0
/ESC - ESC-1 (GE) - RPKM=0.0
/NPC - NPC-2 (GE) - Expression values=5.0
/NPC - NPC-2 (GE) - Total exon reads=5
/NPC - NPC-2 (GE) - RPKM=2.888274918918527
/NPC - NPC-1 (GE) - Expression values=5.0
/NPC - NPC-1 (GE) - Total exon reads=5
/NPC - NPC-1 (GE) - RPKM=2.4460614369482427
```

Track Settings

Navigation: 7 (145,441,459bp)
Range: 18,987,522 - 18,989,099

Insertions: *No variant tracks in view*

Track layout
- Region track
  - Data aggregation: above 100bp
  - Annotation color: 🟥
  - Labels: No labels

Text format

Rows: 919 / 3,176 Table view: Mus musculus ○ Match any ● Match all

Experiment - Fold Change (original values) | abs value > | 1.5
EDGE test: ESC vs NPC, tagwise dispersions - FDR p-value correction | < | 0.05

Chromosome	Region	Name	Experiment - Fold Cha...	EDGE test: ESC vs NPC, tagwise dispersions - FDR p-value correction
7	131388652..131410521	Ikzf5	9.99	2.12E-55
7	18987524..18988962	Nanos2	10.00	0.04
7	121033252..121035096	Gm9905	10.00	1.55E-3
7	55962531..55980868	A230056P14Rik	10.86	3.09E-11
7	143778365..143784500	Mrgpre	11.00	3.02E-18

Table Settings
- ☐ Experiment - IQR (original values)
- ☐ Experiment - Difference (original val...
- ☑ Experiment - Fold Change (original va...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☐ EDGE test: ESC vs NPC, tagwise disper...
- ☑ EDGE test: ESC vs NPC, tagwise disper...
- ☐ ESC - Means

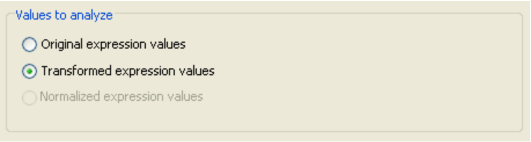
Annotation: Region 18987524..18988962, size 1,439

Figure 27.59: *The annotations on the track produced by the Create Track from Experiment Tool*

When you perform transformation and normalization, the original expression values will be kept, and the new values will be added. If you select an experiment ( ), the new values will be added to the experiment (not the original samples). And likewise if you select a sample (( ) or ( )) - in this case the new values will be added to the sample (the original values are still kept on the sample).

27.5.1 Selecting transformed and normalized values for analysis

A number of the tools in the **Expression Analysis** ( ) folder use expression levels. All of these tools let you choose between *Original*, *Transformed* and *Normalized* expression values as shown in figure 27.60.


Values to analyze

○ Original expression values
● Transformed expression values
○ Normalized expression values

Figure 27.60: *Selecting which version of the expression values to analyze. In this case, the values have not been normalized, so it is not possible to select normalized values.*

In this case, the values have not been normalized, so it is not possible to select normalized values.

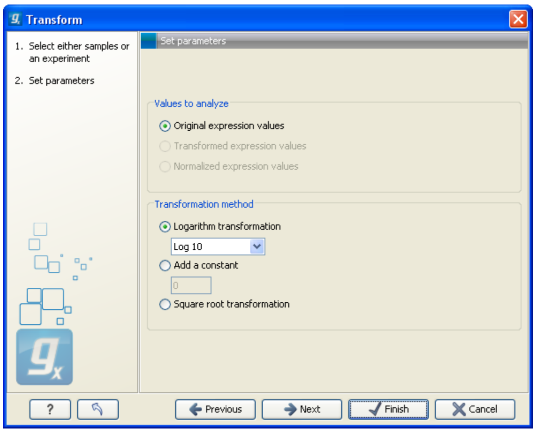
27.5.2 Transformation

The *CLC Cancer Research Workbench* lets you transform expression values based on logarithm and adding a constant:

> **Toolbox** | **Transcriptomics Analysis** ()| **Transformation and Normalization** | **Transform** ()

Select a number of samples (() or ()) or an experiment () and click **Next**.

This will display a dialog as shown in figure 27.61.



Transform

1. Select either samples or an experiment
2. Set parameters

Set parameters

Values to analyze

● Original expression values
○ Transformed expression values
○ Normalized expression values

Transformation method

● Logarithm transformation
Log 10
○ Add a constant
0
○ Square root transformation

Figure 27.61: *Transforming expression values.*

At the top, you can select which values to transform (see section 27.5.1).

Next, you can choose three kinds of transformation:

- **Logarithm transformation**. Transformed expression values will be calculated by taking the logarithm (of the specified type) of the values you have chosen to transform.
  - **10**.
  - **2**.
  - **Natural logarithm**.
- **Adding a constant**. Transformed expression values will be calculated by adding the specified constant to the values you have chosen to transform.
- **Square root transformation**.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

27.5.3 Normalization

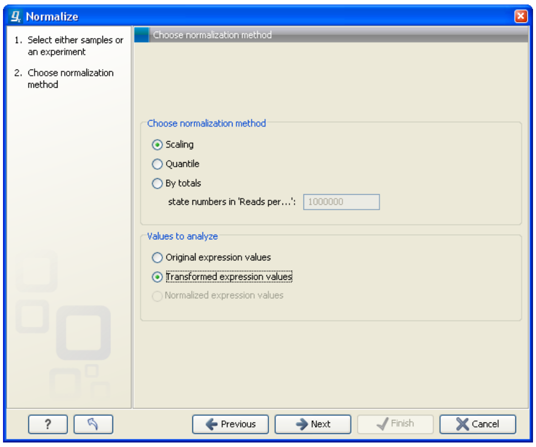
The *CLC Cancer Research Workbench* lets you normalize expression values.

To start the normalization:

Toolbox | Transcriptomics Analysis ()| Transformation and Normalization | Normalize ()

Select a number of samples (() or ()) or an experiment () and click **Next**.

This will display a dialog as shown in figure 27.62.



1. Select either samples or an experiment
2. Choose normalization method

Choose normalization method

Choose normalization method

● Scaling
○ Quantile
○ By totals

state numbers in 'Reads per...': 1000000

Values to analyze

○ Original expression values
● Transformed expression values
○ Normalized expression values

Figure 27.62: *Choosing normalization method.*

At the top, you can choose three kinds of normalization (for mathematical descriptions see [Bolstad et al., 2003]):

- **Scaling**. The sets of the expression values for the samples will be multiplied by a constant so that the sets of normalized values for the samples have the same 'target' value (see description of the **Normalization value** below).
- **Quantile**. The empirical distributions of the sets of expression values for the samples are used to calculate a common target distribution, which is used to calculate normalized sets of expression values for the samples.
- **By totals**. This option is intended to be used with count-based data, i.e. data from RNA-seq, small RNA or expression profiling by tags. A sum is calculated for the expression values in a sample. The transformed value are generated by dividing the input values by the sample sum and multiplying by the factor (e.g. per '1,000,000').

Figures 27.63 and 27.64 show the effect on the distribution of expression values when using scaling or quantile normalization, respectively.

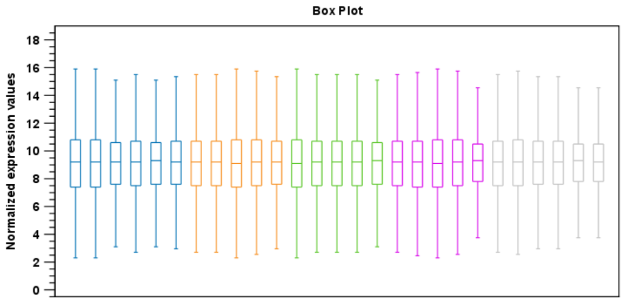


Figure 27.63: *Box plot after scaling normalization.*

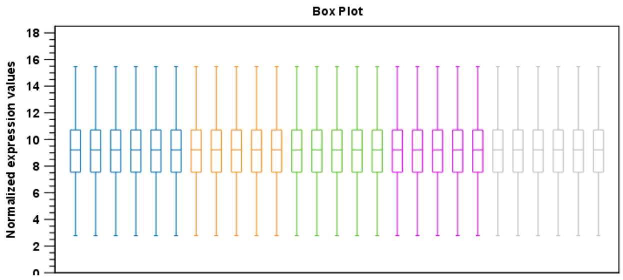
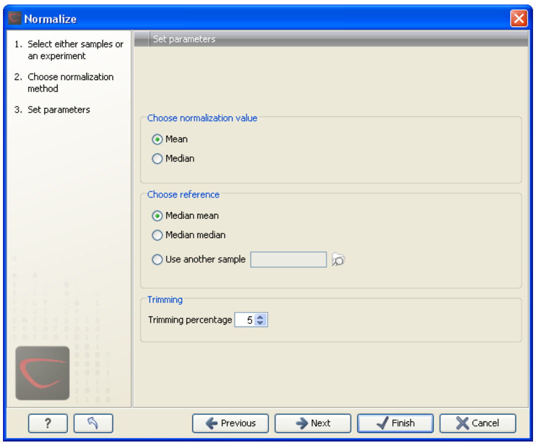


Figure 27.64: *Box plot after quantile normalization.*

At the bottom of the dialog in figure 27.62, you can select which values to normalize (see section 27.5.1).

Clicking **Next** will display a dialog as shown in figure 27.65.


Normalize
1. Select either samples or an experiment
2. Choose normalization method
3. Set parameters
Set parameters
? Previous Next Finish Cancel

Choose normalization value
- ● Mean
- ○ Median

Choose reference
- ● Median mean
- ○ Median median
- ○ Use another sample

Trimming
Trimming percentage 5

Figure 27.65: *Normalization settings.*

The following parameters can be set:

- **Normalization value.** The type of value of the samples which you want to ensure are equal for the normalized expression values
  - **Mean.**
  - **Median.**
- **Reference.** The specific value that you want the normalized value to be after normalization.
  - **Median mean.**
  - **Median median.**
  - **Use another sample.**
- **Trimming percentage.** Expression values that lie below the value of this percentile, or above 100 minus the value of this percentile, in the empirical distribution of the expression values in a sample will be excluded when calculating the normalization and reference values.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

27.6 Quality control

The *CLC Cancer Research Workbench* includes a number of tools for quality control. These allow visual inspection of the overall distributions, variability and similarity of the sets of expression values in samples, and may be used to spot unwanted systematic differences between samples, outlying samples and samples of poor quality, that you may want to exclude.

27.6.1 Creating box plots - analyzing distributions

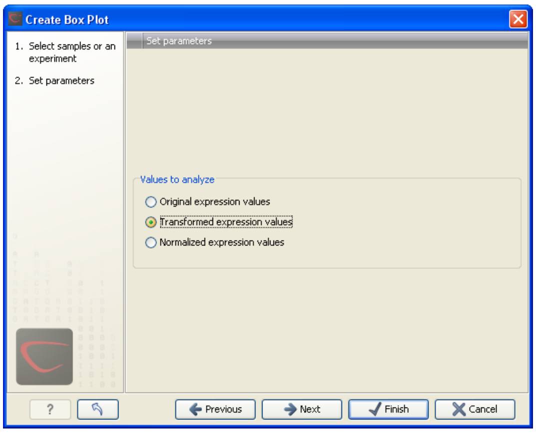
In most cases you expect the majority of genes to behave similarly under the conditions considered, and only a smaller proportion to behave differently. Thus, at an overall level you would expect the distributions of the sets of expression values in samples in a study to be similar. A boxplot provides a visual presentation of the distributions of expression values in samples. For each sample the distribution of it's values is presented by a line representing a center, a box representing the middle part, and whiskers representing the tails of the distribution. Differences in the overall distributions of the samples in a study may indicate that normalization is required before the samples are comparable. An atypical distribution for a single sample (or a few samples), relative to the remaining samples in a study, could be due to imperfections in the preparation and processing of the sample, and may lead you to reconsider using the sample(s).

To create a box plot:

Toolbox | **Transcriptomics Analysis ( )**| **Quality Control** | **Create Box Plot ( )**

Select a number of samples (( ) or ( )) or an experiment ( ) and click **Next**.

This will display a dialog as shown in figure 27.66.



Values to analyze
○ Original expression values
● Transformed expression values
○ Normalized expression values

Figure 27.66: *Choosing values to analyze for the box plot.*

Here you select which values to use in the box plot (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Viewing box plots

An example of a box plot of a two-group experiment with 12 samples is shown in figure 27.67.

Note that the boxes per default are colored according to their group relationship. At the bottom you find the names of the samples, and the y-axis shows the expression values (note that sample

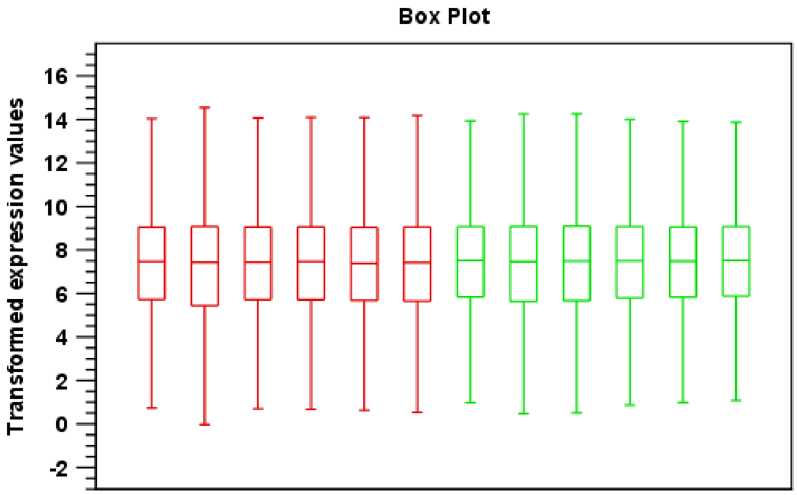
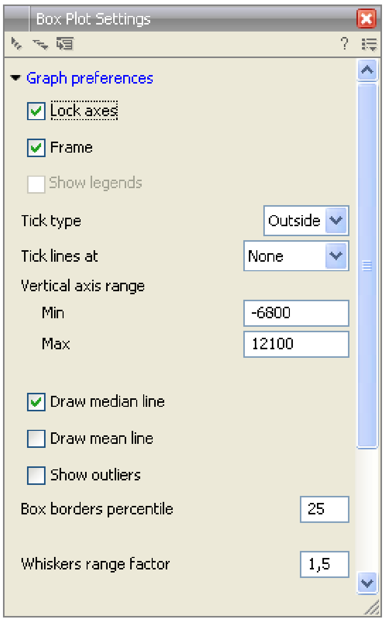


Figure 27.67: *A box plot of 12 samples in a two-group experiment, colored by group.*

names are not shown in figure 27.67).

Per default the box includes the IQR values (from the lower to the upper quartile), the median is displayed as a line in the box, and the whiskers extend 1.5 times the height of the box.

In the **Side Panel** to the left, there is a number of options to adjust this view. Under **Graph preferences**, you can adjust the general properties of the box plot (see figure 27.68).



Box Plot Settings

Graph preferences

☑ Lock axes

☑ Frame

□ Show legends

Tick type: Outside

Tick lines at: None

Vertical axis range

Min: -6800

Max: 12100

☑ Draw median line

□ Draw mean line

□ Show outliers

Box borders percentile: 25

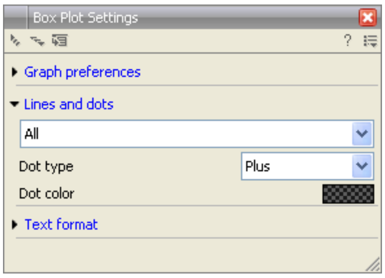
Whiskers range factor: 1,5

Figure 27.68: *Graph preferences for a box plot.*

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.

- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Draw median line**. This is the default - the median is drawn as a line in the box.
- **Draw mean line**. Alternatively, you can also display the mean value as a line.
- **Show outliers**. The values outside the whiskers range are called outliers. Per default they are not shown. Note that the dot type that can be set below only takes effect when outliers are shown. When you select and deselect the **Show outliers**, the vertical axis range is automatically re-calculated to accommodate the new values.

Below the general preferences, you find the **Lines and dots** preferences, where you can adjust coloring and appearance (see figure 27.69).



All

Dot type: Plus

Dot color: ▩

Figure 27.69: *Lines and dot preferences for a box plot.*

- **Select sample or group**. When you wish to adjust the properties below, first select an item in this drop-down menu. That will apply the changes below to this item. If your plot is based on an experiment, the drop-down menu includes both group names and sample names, as well as an entry for selecting "All". If your plot is based on single elements, only sample names will be visible. Note that there are sometimes "mixed states" when you select a group where two of the samples e.g. have different colors. Selecting a new color in this case will erase the differences.
- **Dot type**
  - None
  - Cross
  - Plus
  - Square

  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot

- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.

Note that if you wish to use the same settings next time you open a box plot, you need to save the settings of the **Side Panel** (see section 4.5).

Interpreting the box plot

This section will show how to interpret a box plot through a few examples.

First, if you look at figure 27.70, you can see a box plot for an experiment with 5 groups and 27 samples.

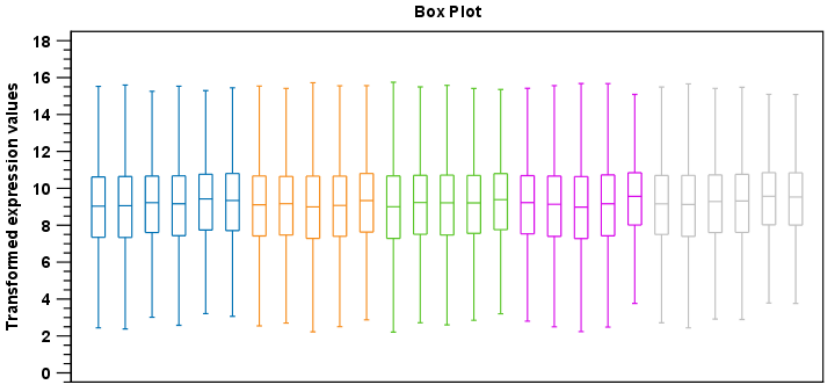


Figure 27.70: *Box plot for an experiment with 5 groups and 27 samples.*

None of the samples stand out as having distributions that are atypical: the boxes and whiskers ranges are about equally sized. The locations of the distributions however, differ some, and indicate that normalization may be required. Figure 27.71 shows a box plot for the same experiment after quantile normalization: the distributions have been brought into par.

In figure 27.72 a box plot for a two group experiment with 5 samples in each group is shown.

The distribution of values in the second sample from the left is quite different from those of other samples, and could indicate that the sample should not be used.

27.6.2 Hierarchical clustering of samples

A hierarchical clustering of samples is a tree representation of their relative similarity. The tree structure is generated by

1. letting each feature be a cluster

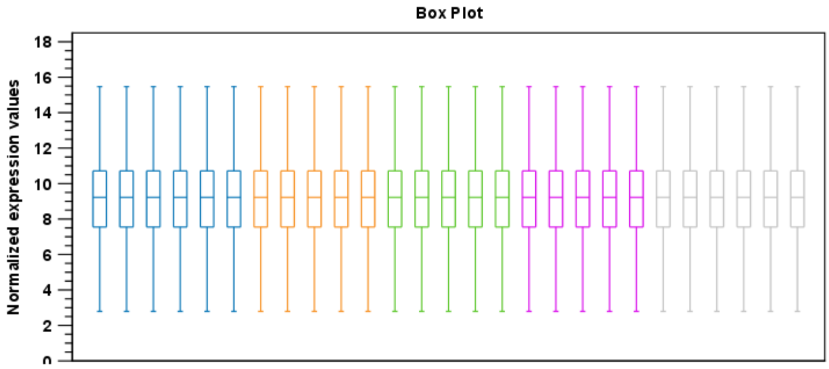


Figure 27.71: *Box plot after quantile normalization.*

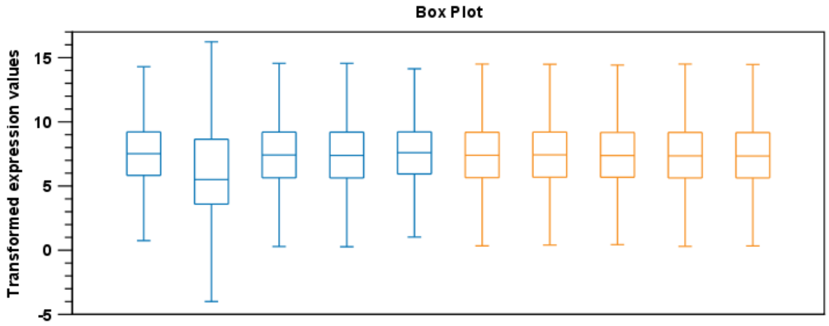


Figure 27.72: *Box plot for a two-group experiment with 5 samples.*

2. calculating pairwise distances between all clusters
3. joining the two closest clusters into one new cluster
4. iterating 2-3 until there is only one cluster left (which will contain all samples).

The tree is drawn so that the distances between clusters are reflected by the lengths of the branches in the tree. Thus, features with expression profiles that closely resemble each other have short distances between them, those that are more different, are placed further apart. (See [Eisen et al., 1998] for a classical example of application of a hierarchical clustering algorithm in microarray analysis. The example is on features rather than samples).

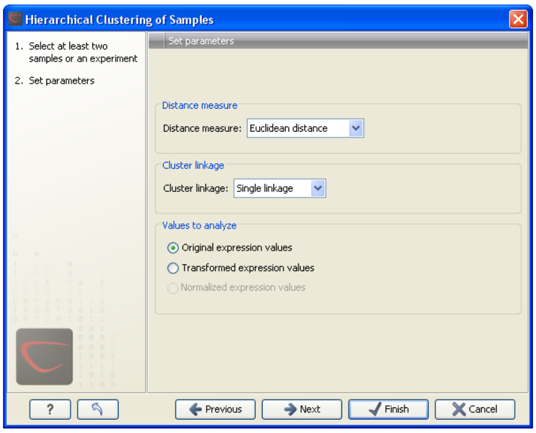
To start the clustering:

> **Toolbox | Transcriptomics Analysis ()| Quality Control | Hierarchical Clustering of Samples ()**

Select a number of samples (() or ()) or an experiment () and click **Next**.

This will display a dialog as shown in figure 27.73. The hierarchical clustering algorithm requires that you specify a distance measure and a cluster linkage. The similarity measure is used to specify how distances between two samples should be calculated. The cluster distance metric specifies how you want the distance between two clusters, each consisting of a number of samples, to be calculated.

At the top, you can choose three kinds of **Distance measures**:



Hierarchical Clustering of Samples

1. Select at least two samples or an experiment
2. Set parameters

Set parameters

Distance measure

Distance measure: Euclidean distance

Cluster linkage

Cluster linkage: Single linkage

Values to analyze

● Original expression values

○ Transformed expression values

○ Normalized expression values

Figure 27.73: *Parameters for hierarchical clustering of samples.*

- **Euclidean distance**. The ordinary distance between two points - the length of the segment connecting them. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Euclidean distance between u and v is

$$|u - v| = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}.$$

- **1 - Pearson correlation**. The Pearson correlation coefficient between two elements $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is defined as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \overline{x}}{s_x}) * (\frac{y_i - \overline{y}}{s_y})$$

  where $\overline{x}/\overline{y}$ is the average of values in x/y and s_x/s_y is the sample standard deviation of these values. It takes a value $\in [-1, 1]$. Highly correlated elements have a high absolute value of the Pearson correlation, and elements whose values are un-informative about each other have Pearson correlation 0. Using $1 - |Pearsoncorrelation|$ as distance measure means that elements that are highly correlated will have a short distance between them, and elements that have low correlation will be more distant from each other.

- **Manhattan distance**. The Manhattan distance between two points is the distance measured along axes at right angles. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Manhattan distance between u and v is

$$|u - v| = \sum_{i=1}^{n}|u_i - v_i|.$$

Next, you can select the cluster linkage to be used:

- **Single linkage**. The distance between two clusters is computed as the distance between the two closest elements in the two clusters.

- **Average linkage**. The distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster. The averaging is performed over all pairs (x, y), where x is an object from the first cluster and y is an object from the second cluster.

- **Complete linkage**. The distance between two clusters is computed as the maximal object-to-object distance $d(x_i, y_j)$, where x_i comes from the first cluster, and y_j comes from the second cluster. In other words, the distance between two clusters is computed as the distance between the two farthest objects in the two clusters.

At the bottom, you can select which values to cluster (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Result of hierarchical clustering of samples

The result of a sample clustering is shown in figure 27.74.

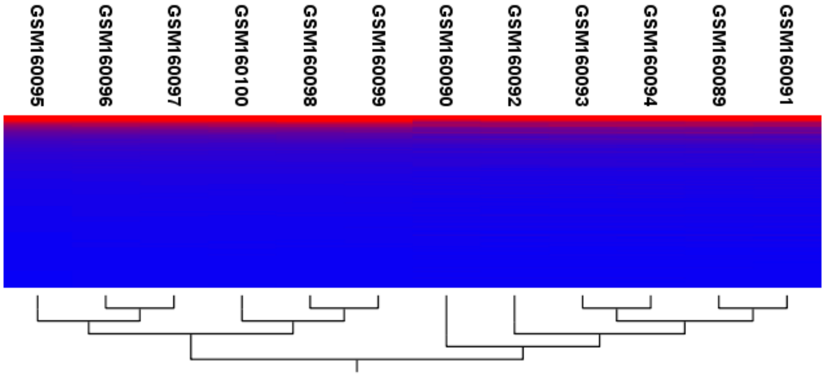


Figure 27.74: *Sample clustering.*

If you have used an **experiment** (▦) as input, the clustering is added to the experiment and will be saved when you save the experiment. It can be viewed by clicking the **Show Heat Map** (▦) button at the bottom of the view (see figure 27.75).

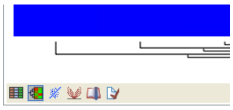


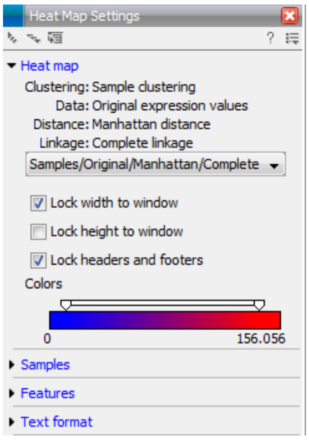
Figure 27.75: *Showing the hierarchical clustering of an experiment.*

If you have selected a number of **samples** ((▦) or (☰)) as input, a new element will be created that has to be saved separately.

Regardless of the input, the view of the clustering is the same. As you can see in figure 27.74, there is a tree at the bottom of the view to visualize the clustering. The names of the samples are listed at the top. The features are represented as horizontal lines, colored according to the expression level. If you place the mouse on one of the lines, you will see the names of the feature to the left. The features are sorted by their expression level in the first sample (in order to cluster the features, see section 27.8.1).

Researchers often have a priori knowledge of which samples in a study should be similar (e.g. samples from the same experimental condition) and which should be different (samples from biological distinct conditions). Thus, researches have expectations about how they should cluster. Samples that are placed unexpectedly in the hierarchical clustering tree may be samples that have been wrongly allocated to a group, samples of unintended or unclean tissue composition or samples for which the processing has gone wrong. Unexpectedly placed samples, of course, could also be highly interesting samples.

There are a number of options to change the appearance of the heat map. At the top of the **Side Panel**, you find the **Heat map** preference group (see figure 27.76).

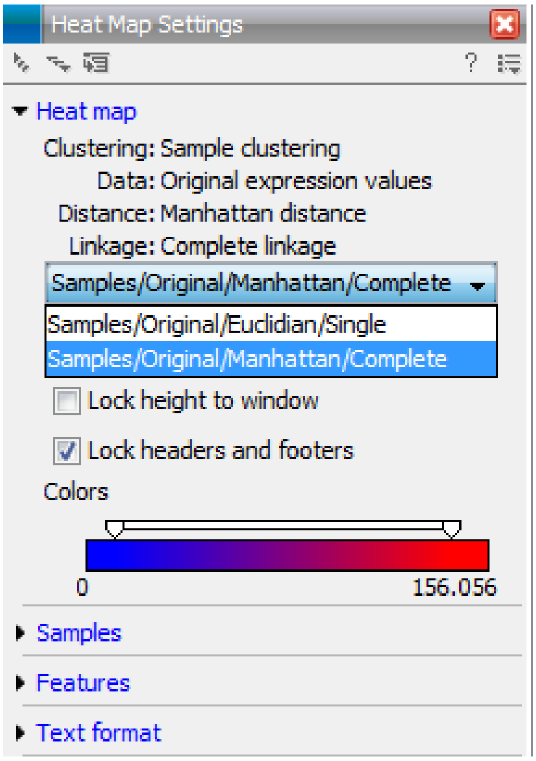


Clustering: Sample clustering
Data: Original expression values
Distance: Manhattan distance
Linkage: Complete linkage
Samples/Original/Manhattan/Complete

☑ Lock width to window
□ Lock height to window
☑ Lock headers and footers

Figure 27.76: *Side Panel of heat map.*

At the top, there is information about the heat map currently displayed. The information regards type of clustering, expression value used together with distance and linkage information. If you have performed more than one clustering, you can choose between the resulting heat maps in a drop-down box (see figure 27.87).



Clustering: Sample clustering
Data: Original expression values
Distance: Manhattan distance
Linkage: Complete linkage
Samples/Original/Manhattan/Complete
Samples/Original/Euclidian/Single
Samples/Original/Manhattan/Complete

□ Lock height to window
☑ Lock headers and footers

Figure 27.77: *When more than one clustering has been performed, there will be a list of heat maps to choose from.*

Note that if you perform an identical clustering, the existing heat map will simply be replaced.

Below this box, there is a number of settings for displaying the heat map.

- **Lock width to window**. When you zoom in the heat map, you will per default only zoom in on the vertical level. This is because the width of the heat map is locked to the window. If you uncheck this option, you will zoom both vertically and horizontally. Since you always have more features than samples, it is useful to lock the width since you then have all the samples in view all the time.
- **Lock height to window**. This is the corresponding option for the height. Note that if you check both options, you will not be able to zoom at all, since both the width and the height is fixed.
- **Lock headers and footers**. This will ensure that you are always able to see the sample and feature names and the trees when you zoom in.
- **Colors**. The expression levels are visualized using a gradient color scheme, where the right side color is used for high expression levels and the left side color is used for low expression levels. You can change the coloring by clicking the box, and you can change the relative coloring of the values by dragging the two knobs on the white slider above.

Below you find the **Samples** and **Features** groups. They contain options to show names, legend, and tree above or below the heatmap. Note that for clustering of samples, you find the tree options in the **Samples** group, and for clustering of features, you find the tree options in the **Features** group. With the tree options, you can also control the **Tree size**, from tiny to very large, and the option of showing the full tree, no matter how much space it will use.

Note that if you wish to use the same settings next time you open a heat map, you need to save the settings of the **Side Panel** (see section 4.5).

27.6.3 Principal component analysis

A principal component analysis is a mathematical analysis that identifies and quantifies the directions of variability in the data. For a set of samples, e.g. an experiment, this can be done either by finding the eigenvectors and eigenvalues of the *covariance matrix* of the samples or the *correlation matrix* of the samples (the correlation matrix is a 'normalized' version of the covariance matrix: the entries in the covariance matrix look like this $Cov(X,Y)$, and those in the correlation matrix like this: $Cov(X,Y)/(sd(X)*sd(Y))$. A covariance maybe any value, but a correlation is always between -1 and 1).

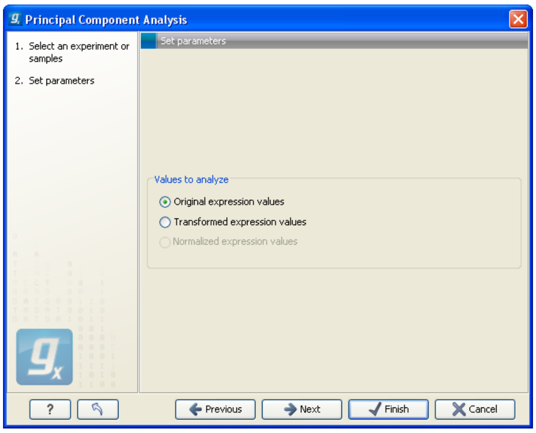
The eigenvectors are orthogonal. The first principal component is the eigenvector with the largest eigenvalue, and specifies the direction with the largest variability in the data. The second principal component is the eigenvector with the second largest eigenvalue, and specifies the direction with the second largest variability. Similarly for the third, etc. The data can be projected onto the space spanned by the eigenvectors. A plot of the data in the space spanned by the first and second principal component will show a simplified version of the data with variability in other directions than the two major directions of variability ignored.

To start the analysis:

> **Toolbox | Transcriptomics Analysis ( )| Quality Control | Principal Component Analysis ( )**

Select a number of samples (() or ()) or an experiment () and click **Next**.

This will display a dialog as shown in figure 27.78.



Principal Component Analysis

1. Select an experiment or samples
2. Set parameters

Set parameters

Values to analyze

- ● Original expression values
- ○ Transformed expression values
- ○ Normalized expression values

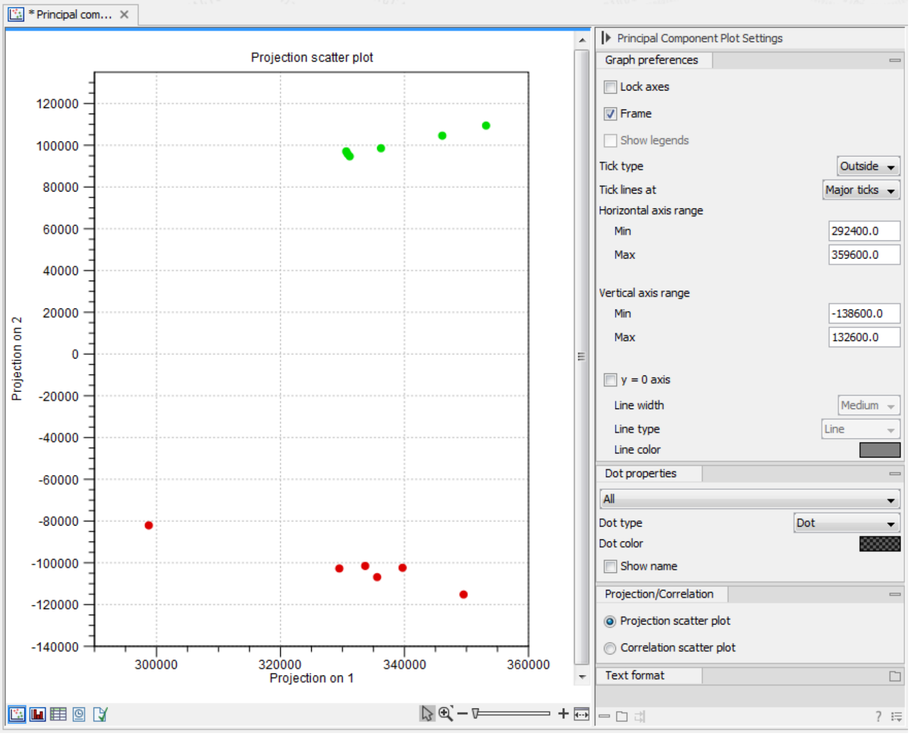
Figure 27.78: *Selcting which values the principal component analysis should be based on.*

In this dialog, you select the values to be used for the principal component analysis (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Principal component analysis plot

This will create a principal component plot as shown in figure 27.79.



Principal Component Plot Settings

Graph preferences

- □ Lock axes
- ☑ Frame
- □ Show legends

Tick type: Outside

Tick lines at: Major ticks

Horizontal axis range

Min: 292400.0

Max: 359600.0

Vertical axis range

Min: -138600.0

Max: 132600.0

- □ y = 0 axis

Line width: Medium

Line type: Line

Line color

Dot properties

All

Dot type: Dot

Dot color

- □ Show name

Projection/Correlation

- ● Projection scatter plot
- ○ Correlation scatter plot

Text format

Figure 27.79: *A principal component analysis colored by group.*

The plot shows the projection of the samples onto the two-dimensional space spanned by the first and second principal component of the covariance matrix. In the bottom part of the side-panel, the 'Projection/Correlation' part, you can change to show the projection onto the *correlation* matrix rather than the *covariance* matrix by choosing 'Correlation scatter plot'. Both plots will show how the samples separate along the two directions between which the samples exhibit the largest amount of variation. For the 'projection scatter plot' this variation is measured in absolute terms, and depends on the units in which you have measured your samples. The correlation scatter plot is a normalized version of the projection scatter plot, which makes it possible to compare principal component analysis between experiments, even when these have not been done using the same units (e.g an experiment that uses 'original' scale data and another one that uses 'log-scale' data).

The plot in figure 27.79 is based on a two-group experiment. The group relationships are indicated by color. We expect the samples from within a group to exhibit less variability when compared, than samples from different groups. Thus samples should cluster according to groups and this is what we see. The PCA plot is thus helpful in identifying outlying samples and samples that have been wrongly assigned to a group.

In the **Side Panel** to the left, there is a number of options to adjust the view. Under **Graph preferences**, you can adjust the general properties of the plot.

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **y = 0 axis**. Draws a line where y = 0. Below there are some options to control the appearance of the line:
  - **Line width**
    - Thin
    - Medium
    - Wide

  - **Line type**
    - None
    - Line
    - Long dash
    - Short dash
  - **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

Below the general preferences, you find the **Dot properties**:

- **Drop down menu** In this you choose which of the samples (that is, which 'dots') the choices you make below should apply to. You can choose between 'All', a particular group in your experiment, or a particular samples in your experiment.
- **Select sample or group**. When you wish to adjust the properties below, first select an item in this drop-down menu. That will apply the changes below to this item. If your plot is based on an experiment, the drop-down menu includes both group names and sample names, as well as an entry for selecting "All". If your plot is based on single elements, only sample names will be visible. Note that there are sometimes "mixed states" when you select a group where two of the samples e.g. have different colors. Selecting a new color in this case will erase the differences.
- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.
- **Show name**. This will show a label with the name of the sample next to the dot. Note that the labels quickly get crowded, so that is why the names are not put on per default.

Note that if you wish to use the same settings next time you open a principal component plot, you need to save the settings of the **Side Panel** (see section 4.5).

Scree plot

Besides the view shown in figure 27.79, the result of the principal component can also be viewed as a scree plot by clicking the **Show Scree Plot** (📊) button at the bottom of the view. The scree plot shows the proportion of variation in the data explained by the each of the principal components. The first principal component explains about 99 percent of the variability.

In the **Side Panel** to the left, there is a number of options to adjust the view. Under **Graph preferences**, you can adjust the general properties of the plot.

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.

The **Lines and plots** below contains the following parameters:

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.

- **Line width**
  - Thin
  - Medium
  - Wide
- **Line type**
  - None
  - Line
  - Long dash
  - Short dash
- **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

Note that the graph title and the axes titles can be edited simply by clicking them with the mouse. These changes will be saved when you **Save** ( ) the graph - whereas the changes in the **Side Panel** need to be saved explicitly (see section 4.5).

27.7 Statistical analysis - identifying differential expression

The *CLC Cancer Research Workbench* is designed to help you identify differential expression.

27.7.1 Empirical analysis of DGE

The Empirical analysis of DGE tool implements the 'Exact Test' for two-group comparisons developed by Robinson and Smyth [Robinson and Smyth, 2008] and incorporated in the EdgeR Bioconductor package [Robinson et al., 2010]. The test is applicable to count data only, and is designed specifically to deal with situations in which *many* features are studied simultaneously (e.g. genes in a genome) but where only a *few* biological replicates are available for each of the experimental groups studied. This is typically the case for RNA-seq expression analysis. The test is based on the assumption that the count data follows a Negative Binomial distribution, which in contrast to the Poisson distribution has the characteristic that it allows for a non-constant mean-variance relationship.

The 'Exact Test' of Robinson and Smyth is similar to Fisher's Exact Test, but also accounts for overdispersion caused by biological variability. Whereas Fisher's Exact Test compares the counts in one sample against those of another, the 'Exact Test' compares the counts in one *set* of count samples against those in another *set* of count samples. This is achieved by replacing the Hypergeometric distributions of Fisher's Exact Test by Negative binomial distributions, whereby the variability within each of the two groups of samples compared is taken into account. This only works if the dispersions in the two groups compared are identical. As this cannot generally be assumed to be the case for the *original* (nor for the normalized) data, pseudodata for which the dispersion *is* identical is generated from the original data, and the test is carried out on this pseudodata. The generation of the pseudodata is performed simultaneously with the estimation of the dispersion, in an iterative procedure called quantile-adjusted conditional maximum likelihood. Either a single common dispersion for all features may be assumed (as in [Robinson and Smyth, 2008]), or it may be assumed that the dispersion for each feature (e.g. gene) is a 'weighted average' of the common dispersion and feature (e.g. gene) specific dispersions (as suggested in [Robinson and Smyth, 2007]). The weight given to each of the components depends on the number of samples in the groups: the more samples there are in the groups, the higher the weight will be given to the gene-specific component.

The Exact Test in the EdgeR Bioconductor package provides the user with the option to set a large number of parameters. The implementation of the 'Empirical analysis of DGE' algorithm in the Genomics Workbench uses for the most parts the default settings in the edgeR package, version 3.4.0. A detailed outline of the parameter settings is given in section 27.7.1).

Empirical analysis of DGE - implementation parameters

The 'Empirical analysis of DGE' algorithm in the *CLC Cancer Research Workbench* is a re-implementation of the "Exact Test", available as part of the EdgeR Bioconductor package.

The parameter values used in the *CLC Cancer Research Workbench* implementation are the default values for the equivalent parameters in the EdgeR Bioconductor implementation in all but one case. The exception is the estimateCommonDisp parameter, where the default is more stringent than that of EdgeR. The advantage of using a more stringent value for this parameter is that the results will be more accurate. The disadvantage is that the algorithm will be slightly slower, however according to our performance tests, this change has only a marginal impact on the run time of the tool. Overall, the user has a somewhat compromised run time but gains greater confidence in the results at the end.

The parameter values used in the *CLC Cancer Research Workbench* implementation, with reference to the EdgeR function names for clarity, are provided in the table below.

Function in BioC package	Parameter name	Value used and comments
calcNormFactors	method	"TMM"
	refColumn	NULL (automatically selected)
	logratioTrim	0.3
	sumTrim	0.05
	doWeighting	TRUE
	Acutoff	-1e10
estimateCommonDisp	tol	1e-14 (default in edgeR: 1e-6)
	rowsum.filter	Set by user in wizard ("Total count filter cutoff", default 5)
estimateTagewiseDisp	prior.df	10
	trend	"movingave"
	span	NULL
	method	"grid"
	grid.length	11
	grid.range	c(-6, 6)
mglmOneGroup	maxit	50
	tol	1e-10
aveLogCPM	prior.count	2
	dispersion	0.05
exactTest	pair	Set by user in wizard ("Exact test comparisons")
	dispersion	"auto" (tagwise if available, otherwise common)
	rejection.region	"doubletail"
	big.count	900
	prior.count	0.125

Running the Empirical analysis of DGE

First, find the **Empirical analysis of DGE** tool:

> **Toolbox** | **Transcriptomics Analysis ()**| **Statistical Analysis** | **Empirical Analysis of DGE ()**

The original count data for a full expression experiment are the expected input to the Empirical Analysis of DGE tool.

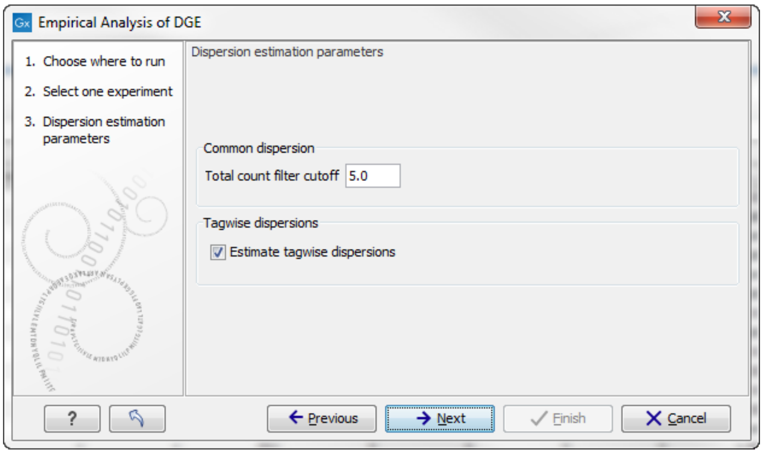
When Experiments created within the Workbench are used as input, the original count values are always used. Columns of such Experiments that contain transformed or normalized values are ignored.

If expression values are being imported from outside the Workbench for use with this test, the data should be original (non-transformed, non-normalized) counts.

Whether the data has been generated in the Workbench or outside the Workbench and imported, the full set of expression results should be used. Please do not run this test on a subset of values from the original sample data.

The reason that the complete set of original count data for samples should be used as input to this test is that the algorithm assumes that the counts on which it operates are Negative Binomially distributed. It implicitly normalizes and transforms these counts, so if the counts have been altered prior to submitting them to the Empirical Analysis of DGE tool, this assumption is likely to be compromised.

When running the Empirical analysis of DGE tool in the Genomics workbench, the user is asked to specify two parameters related to the estimation of the dispersion (figure 27.80). Of these, the 'Total count filter cut-off' specifies which features should be considered when estimating the common dispersion component. Features for which the counts across all samples are low are likely to contribute mostly with noise to the estimation, and features with a lower cummulative count across samples than the value specified will be ignored. When the check-box 'Estimate tag-wise dispersions' is checked, the dispersion estimate for each gene will be a weighted combination of the tag-wise and common dispersion, if the check-box is un-ticked the common dispersion will be used for all genes.



Dispersion estimation parameters

Common dispersion

Total count filter cutoff 5.0

Tagwise dispersions

☑ Estimate tagwise dispersions

Figure 27.80: *Empirical analysis of DGE: setting the parameters related to dispersion.*

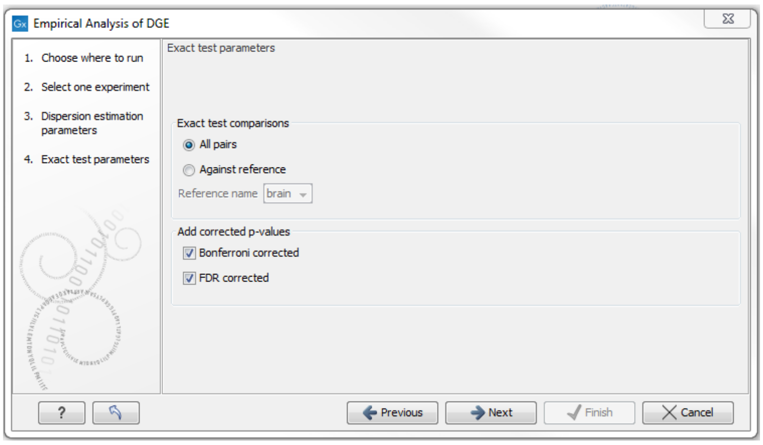
The Empirical analysis of DGE may be carried out between all pairs of groups (by clicking the 'All pairs' button) or for each group against a specified reference group (by clicking the 'Against reference' button) (figure 27.81). In the last case you must specify which of the groups you want to use as reference (the default is to use the group you specified as Group 1 when you set up the experiment). Below you can select to add two kinds of corrected p-values to the analysis (in addition to the standard p-value produced for the test statistic):

- **Bonferroni corrected.**
- **FDR corrected.**

Both are calculated from the original p-values, and aim in different ways to take into account the issue of multiple testing [Dudoit et al., 2003]. The problem of multiple testing arises because the original p-values are related to a single test: the p-value is the probability of observing a more extreme value than that observed in the test carried out. If the p-value is 0.04, we would expect an as extreme value as that observed in 4 out of 100 tests carried out among groups with no difference in means. Popularly speaking, if we carry out 10000 tests and select the features with original p-values below 0.05, we will expect about 0.05 times 10000 = 500 to be false positives.

The Bonferroni corrected p-values handle the multiple testing problem by controlling the 'family-wise error rate': the probability of making at least one false positive call. They are calculated by multiplying the original p-values by the number of tests performed. The probability of having at least one false positive among the set of features with Bonferroni corrected p-values below 0.05, is less than 5%. The Bonferroni correction is conservative: there may be many genes that are differentially expressed among the genes with Bonferroni corrected p-values above 0.05, that will be missed if this correction is applied.

Instead of controlling the family-wise error rate we can control the false discovery rate: FDR. The false discovery rate is the proportion of false positives among all those declared positive. We expect 5 % of the features with FDR corrected p-values below 0.05 to be false positive. There are many methods for controlling the FDR - the method used in *CLC Cancer Research Workbench* is that of [Benjamini and Hochberg, 1995].



Exact test comparisons

● All pairs

○ Against reference

Reference name brain

Add corrected p-values

☑ Bonferroni corrected

☑ FDR corrected

Figure 27.81: *Empirical analysis of DGE: setting comparisons and corrected p-value options.*

When the Empirical analysis of DGE is run three columns will be added to the experiment table for each pair of groups that are analyzed: the 'P-value', 'Fold change' and 'Weighted difference' columns. The 'P-value' holds the p-value for the Exact test. The 'Fold Change' and 'Weighted difference' columns are both calculated from the estimated 'average cpm (counts per million)' values of each of the groups. The estimated 'average cpm' values are values that are derived internally in the Exact Test algorithm. They depend on both the sizes of the samples, the magnitude of the counts and on the estimated negative binomial dispersion, so they cannot be obtained from the original counts by simple algebraic calculations. The 'Fold Change' will tell you how many times bigger the average cpm value of group 2 is relative to that of group 1. If the average cpm value of group 2 is bigger than that of group 1 the fold change is the average cpm value of group 2 divided by that of group 1. If the average cpm value of group 2 is smaller than that of group 1 the fold change is the average cpm value of group 1 divided by that of group 2 with a negative sign. The 'weighted difference' column contains the difference between the average cpm value of group 2 and the average cpm value of group 1. In addition to the three automatically added columns, columns containing the Bonferroni and FDR corrected p-values will be added if that was specified by the user.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Note that if you have already performed statistical analysis on the same values, the existing one will be overwritten.

27.7.2 Volcano plots - inspecting the result of the statistical analysis

The results of the statistical analysis are added to the experiment and can be shown in the experiment table (see section 27.3.2). Typically columns containing the differences (or weighted differences) of the mean group values and the fold changes (or weighted fold changes) of the mean group values will be added along with a column of p-values. Also, columns with FDR or Bonferroni corrected p-values will be added if these were calculated. This added information allows features to be sorted and filtered to exclude the ones without sufficient proof of differential expression (learn more in section 8.3).

If you want a more visual approach to the results of the statistical analysis, you can click the **Show Volcano Plot** ( ) button at the bottom of the experiment table view. In the same way as the scatter plot presented in section 27.3.5, the volcano plot is yet another view on the experiment. Because it uses the p-values and mean differences produced by the statistical analysis, the plot is only available once a statistical analysis has been performed on the experiment.

An example of a volcano plot is shown in figure 27.82.

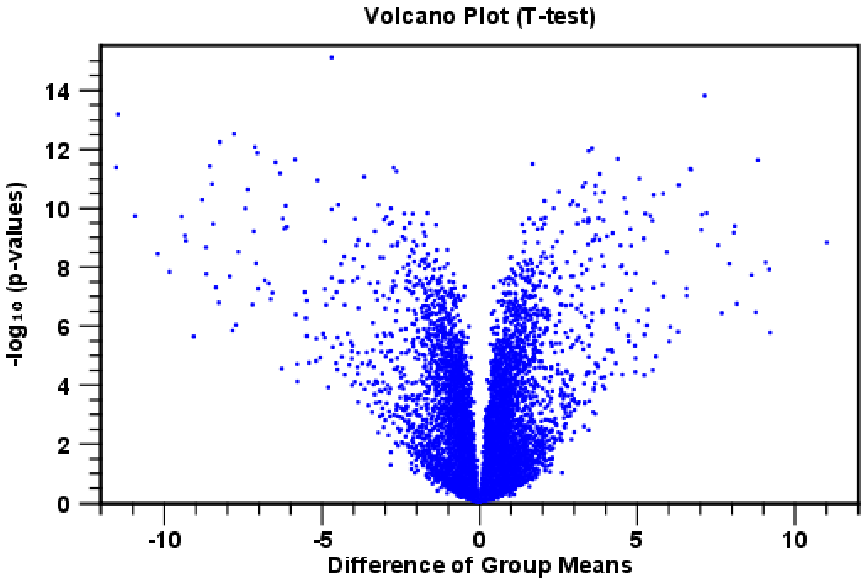


Figure 27.82: *Volcano plot.*

The volcano plot shows the relationship between the p-values of a statistical test and the magnitude of the difference in expression values of the samples in the groups. On the y-axis the $-\log_{10}$ p-values are plotted. For the x-axis you may choose between two sets of values by choosing either 'Fold change' or 'Difference' in the volcano plot side panel's 'Values' part. If you choose 'Fold change' the log of the values in the 'fold change' (or 'Weighted fold change') column for the test will be displayed. If you choose 'Difference' the values in the 'Difference' (or 'Weighted difference') column will be used. Which values you wish to display will depend upon the scale of you data (Read the note on fold change in section 27.3.2).

The larger the difference in expression of a feature, the more extreme it's point will lie on the X-axis. The more significant the difference, the smaller the p-value and thus the higher the $-\log_{10}(p)$ value. Thus, points for features with highly significant differences will lie high in the plot. Features of interest are typically those which change significantly and by a certain magnitude. These are the points in the upper left and upper right hand parts of the volcano plot.

If you have performed different tests or you have an experiment with multiple groups you need to specify for which test and which group comparison you want the volcano plot to be shown. You do this in the 'Test' and 'Values' parts of the volcano plot side panel.

Options for the volcano plot are described in further detail when describing the **Side Panel** below.

If you place your mouse on one of the dots, a small text box will tell the name of the feature. Note that you can zoom in and out on the plot (see section 2.2).

In the **Side Panel** to the right, there is a number of options to adjust the view of the volcano plot. Under **Graph preferences**, you can adjust the general properties of the volcano plot

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.

Below the general preferences, you find the **Dot properties**, where you can adjust coloring and appearance of the dots.

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.

At the very bottom, you find two groups for choosing which values to display:

- **Test**. In this group, you can select which kind of test you want the volcano plot to be shown for.
- **Values**. Under **Values**, you can select which values to plot. If you have multi-group experiments, you can select which groups to compare. You can also select whether to plot **Difference** or **Fold change** on the x-axis. Read the note on fold change in section 27.3.2.

Note that if you wish to use the same settings next time you open a box plot, you need to save the settings of the **Side Panel** (see section 4.5).

27.8 Feature clustering

Feature clustering is used to identify and cluster together features with similar expression patterns over samples (or experimental groups). Features that cluster together may be involved in the same biological process or be co-regulated. Also, by examining annotations of genes within a cluster, one may learn about the underlying biological processes involved in the experiment studied.

27.8.1 Hierarchical clustering of features

A hierarchical clustering of features is a tree presentation of the similarity in expression profiles of the features over a set of samples (or groups). The tree structure is generated by

1. letting each feature be a cluster
2. calculating pairwise distances between all clusters
3. joining the two closest clusters into one new cluster
4. iterating 2-3 until there is only one cluster left (which will contain all samples).

The tree is drawn so that the distances between clusters are reflected by the lengths of the branches in the tree. Thus, features with expression profiles that closely resemble each other have short distances between them, those that are more different, are placed further apart.

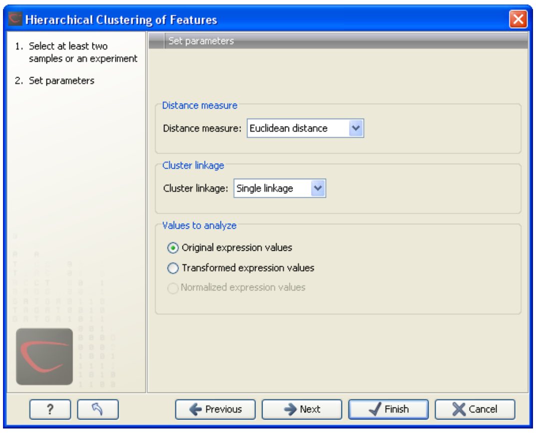
To start the clustering of features:

Toolbox | Transcriptomics Analysis ( )| Feature Clustering | Hierarchical Clustering of Features ( )

Select at least two samples (( ) or ( )) or an experiment ( ).

Note! If your data contains many features, the clustering will take very long time and could make your computer unresponsive. It is recommended to perform this analysis on a subset of the data (which also makes it easier to make sense of the clustering. Typically, you will want to filter away the features that are thought to represent only noise, e.g. those with mostly low values, or with little difference between the samples). See how to create a sub-experiment in section 27.3.2.

Clicking **Next** will display a dialog as shown in figure 27.83. The hierarchical clustering algorithm requires that you specify a distance measure and a cluster linkage. The distance measure is used specify how distances between two features should be calculated. The cluster linkage specifies how you want the distance between two clusters, each consisting of a number of features, to be calculated.



1. Select at least two samples or an experiment
2. Set parameters

Set parameters

Distance measure

Distance measure: Euclidean distance

Cluster linkage

Cluster linkage: Single linkage

Values to analyze

● Original expression values

○ Transformed expression values

○ Normalized expression values

Figure 27.83: *Parameters for hierarchical clustering of features.*

At the top, you can choose three kinds of **Distance measures**:

- **Euclidean distance**. The ordinary distance between two points - the length of the segment connecting them. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Euclidean distance between u and v is

$$|u - v| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}.$$

- **1 - Pearson correlation.** The Pearson correlation coefficient between two elements $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is defined as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{s_x}) * (\frac{y_i - \overline{y}}{s_y})$$

  where $\overline{x}/\overline{y}$ is the average of values in x/y and s_x/s_y is the sample standard deviation of these values. It takes a value $\in [-1, 1]$. Highly correlated elements have a high absolute value of the Pearson correlation, and elements whose values are un-informative about each other have Pearson correlation 0. Using $1 - |Pearsoncorrelation|$ as distance measure means that elements that are highly correlated will have a short distance between them, and elements that have low correlation will be more distant from each other.

- **Manhattan distance**. The Manhattan distance between two points is the distance measured along axes at right angles. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Manhattan distance between u and v is

$$|u - v| = \sum_{i=1}^{n} |u_i - v_i|.$$

Next, you can select different ways to calculate distances between clusters. The possible cluster linkage to use are:

- **Single linkage**. The distance between two clusters is computed as the distance between the two closest elements in the two clusters.
- **Average linkage**. The distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster. The averaging is performed over all pairs (x, y), where x is an object from the first cluster and y is an object from the second cluster.
- **Complete linkage**. The distance between two clusters is computed as the maximal object-to-object distance $d(x_i, y_j)$, where x_i comes from the first cluster, and y_j comes from the second cluster. In other words, the distance between two clusters is computed as the distance between the two farthest objects in the two clusters.

At the bottom, you can select which values to cluster (see section 27.5.1). Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Result of hierarchical clustering of features

The result of a feature clustering is shown in figure 27.84.

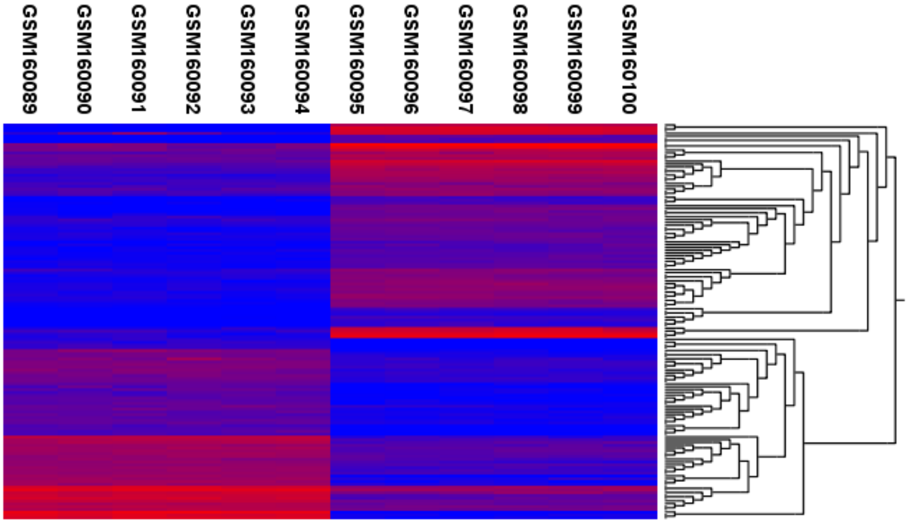


Figure 27.84: *Hierarchical clustering of features.*

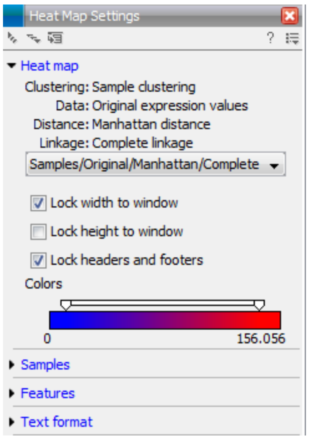
If you have used an **experiment** ( ) as input, the clustering is added to the experiment and will be saved when you save the experiment. It can be viewed by clicking the **Show Heat Map** ( ) button at the bottom of the view (see figure 27.85).

Figure 27.85: *Showing the hierarchical clustering of an experiment.*

If you have selected a number of **samples** (( ) or ( )) as input, a new element will be created that has to be saved separately.

Regardless of the input, a hierarchical tree view with associated heatmap is produced (figure 27.84). In the heatmap each row corresponds to a feature and each column to a sample. The color in the i'th row and j'th column reflects the expression level of feature i in sample j (the color scale can be set in the side panel). The order of the rows in the heatmap are determined by the hierarchical clustering. If you place the mouse on one of the rows, you will see the name of the corresponding feature to the left. The order of the columns (that is, samples) is determined by their input order or (if defined) experimental grouping. The names of the samples are listed at the top of the heatmap and the samples are organized into groups.

There are a number of options to change the appearance of the heat map. At the top of the **Side Panel**, you find the **Heat map** preference group (see figure 27.86).



Clustering: Sample clustering
Data: Original expression values
Distance: Manhattan distance
Linkage: Complete linkage
Samples/Original/Manhattan/Complete

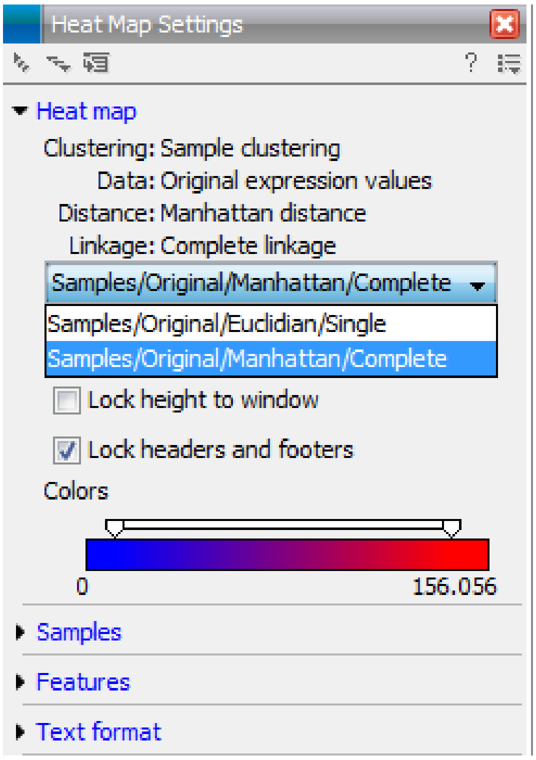
☑ Lock width to window
☐ Lock height to window
☑ Lock headers and footers

Figure 27.86: *Side Panel of heat map.*

At the top, there is information about the heat map currently displayed. The information regards type of clustering, expression value used together with distance and linkage information. If you have performed more than one clustering, you can choose between the resulting heat maps in a drop-down box (see figure 27.87).

Note that if you perform an identical clustering, the existing heat map will simply be replaced. Below this box, there is a number of settings for displaying the heat map.

- **Lock width to window**. When you zoom in the heat map, you will per default only zoom in on the vertical level. This is because the width of the heat map is locked to the window. If you uncheck this option, you will zoom both vertically and horizontally. Since you always have more features than samples, it is useful to lock the width since you then have all the samples in view all the time.
- **Lock height to window**. This is the corresponding option for the height. Note that if you check both options, you will not be able to zoom at all, since both the width and the height is fixed.
- **Lock headers and footers**. This will ensure that you are always able to see the sample and feature names and the trees when you zoom in.
- **Colors**. The expression levels are visualized using a gradient color scheme, where the right side color is used for high expression levels and the left side color is used for low expression levels. You can change the coloring by clicking the box, and you can change the relative coloring of the values by dragging the two knobs on the white slider above.



Heat Map Settings

Heat map

Clustering: Sample clustering

Data: Original expression values

Distance: Manhattan distance

Linkage: Complete linkage

Samples/Original/Manhattan/Complete

Samples/Original/Euclidian/Single

Samples/Original/Manhattan/Complete

□ Lock height to window

☑ Lock headers and footers

Colors

Figure 27.87: *When more than one clustering has been performed, there will be a list of heat maps to choose from.*

Below you find the **Samples** and **Features** groups. They contain options to show names, legend, and tree above or below the heatmap. Note that for clustering of samples, you find the tree options in the **Samples** group, and for clustering of features, you find the tree options in the **Features** group. With the tree options, you can also control the **Tree size**, from tiny to very large, and the option of showing the full tree, no matter how much space it will use.

Note that if you wish to use the same settings next time you open a heat map, you need to save the settings of the **Side Panel** (see section 4.5).

27.8.2 K-means/medoids clustering

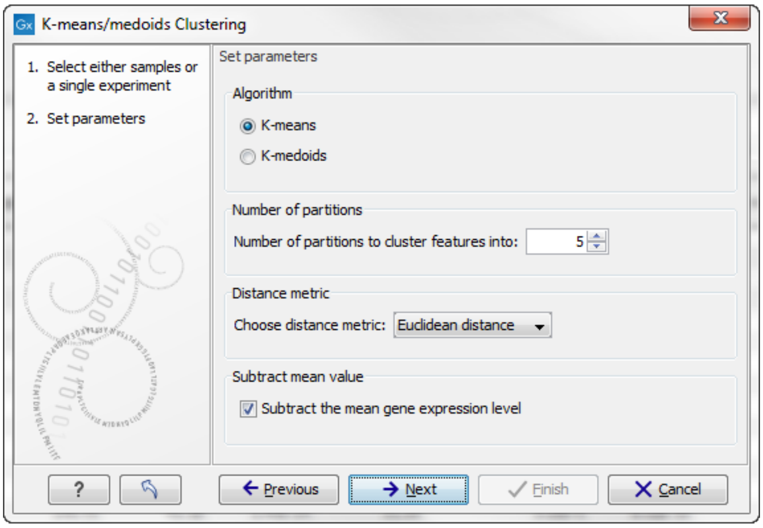
In a k-means or medoids clustering, features are clustered into k separate clusters. The procedures seek to find an assignment of features to clusters, for which the distances between features within the cluster is small, while distances between clusters are large.

> **Toolbox** | **Transcriptomics Analysis** ( )| **Feature Clustering** | **K-means/medoids Clustering** ( )

Select at least two samples (( ) or ( )) or an experiment ( ).

Note! If your data contains many features, the clustering will take very long time and could make your computer unresponsive. It is recommended to perform this analysis on a subset of the data (which also makes it easier to make sense of the clustering). See how to create a sub-experiment in section 27.3.2.

Clicking **Next** will display a dialog as shown in figure 27.88.


K-means/medoids Clustering

1. Select either samples or a single experiment
2. Set parameters

Set parameters

Algorithm

● K-means
○ K-medoids

Number of partitions

Number of partitions to cluster features into: 5

Distance metric

Choose distance metric: Euclidean distance

Subtract mean value

☑ Subtract the mean gene expression level

Figure 27.88: *Parameters for k-means/medoids clustering.*

The parameters are:

- **Algorithm**. You can choose between two clustering methods:
  - **K-means**. K-means clustering assigns each point to the cluster whose center is nearest. The center/centroid of a cluster is defined as the average of all points in the cluster. If a data set has three dimensions and the cluster has two points $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$, then the centroid Z becomes $Z = (z_1, z_2, z_3)$, where $z_i = (x_i + y_i)/2$ for $i = 1, 2, 3$. The algorithm attempts to minimize the intra-cluster variance defined by:

    $$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

    where there are k clusters $S_i, i = 1, 2, \ldots, k$ and μ_i is the centroid of all points $x_j \in S_i$. The detailed algorithm can be found in [Lloyd, 1982].
  - **K-medoids**. K-medoids clustering is computed using the PAM-algorithm (PAM is short for Partitioning Around Medoids). It chooses datapoints as centers in contrast to the K-means algorithm. The PAM-algorithm is based on the search for k representatives (called medoids) among all elements of the dataset. When having found k representatives k clusters are now generated by assigning each element to its nearest medoid. The algorithm first looks for a good initial set of medoids (the BUILD phase). Then it finds a local minimum for the objective function:

    $$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - c_i)^2$$

    where there are k clusters $S_i, i = 1, 2, \ldots, k$ and c_i is the medoid of S_i. This solution implies that there is no single switch of an object with a medoid that will decrease the objective (this is called the SWAP phase). The PAM-agorithm is described in [Kaufman and Rousseeuw, 1990].
- **Number of partitions**. The number of partitions to cluster features into.

- **Distance metric**. The metric to compute distance between data points.
  - **Euclidean distance**. The ordinary distance between two elements - the length of the segment connecting them. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Euclidean distance between u and v is

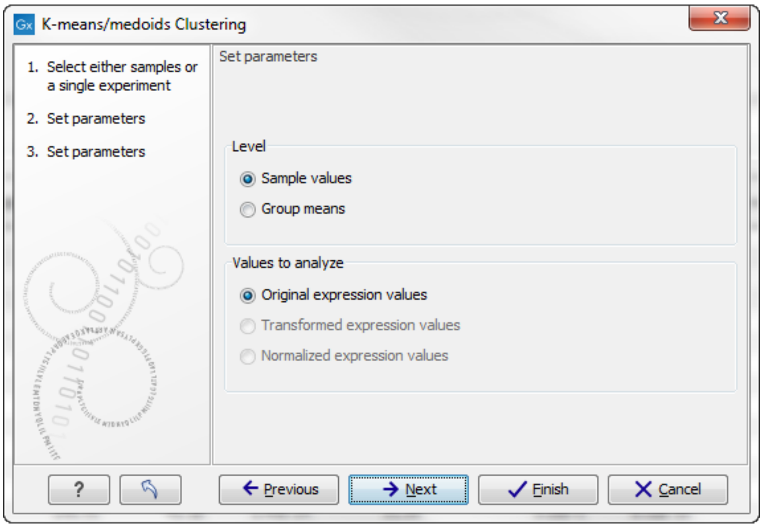
    $$|u - v| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}.$$

  - **Manhattan distance**. The Manhattan distance between two elements is the distance measured along axes at right angles. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Manhattan distance between u and v is

    $$|u - v| = \sum_{i=1}^{n} |u_i - v_i|.$$

- **Subtract mean value**. For each gene, subtract the mean gene expression value over all input samples.

Clicking **Next** will display a dialog as shown in figure 27.89.



K-means/medoids Clustering

1. Select either samples or a single experiment
2. Set parameters
3. Set parameters

Set parameters

Level

● Sample values
○ Group means

Values to analyze

● Original expression values
○ Transformed expression values
○ Normalized expression values

Figure 27.89: *Parameters for k-means/medoids clustering.*

At the top, you can choose the **Level** to use. Choosing 'sample values' means that distances will be calculated using all the individual values of the samples. When 'group means' are chosen, distances are calculated using the group means.

At the bottom, you can select which values to cluster (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Viewing the result of k-means/medoids clustering

The result of the clustering is a number of graphs. The number depends on the number of partitions chosen (figure 27.88) - there is one graph per cluster. Using drag and drop as explained in section 2.1.6, you can arrange the views to see more than one graph at the time.

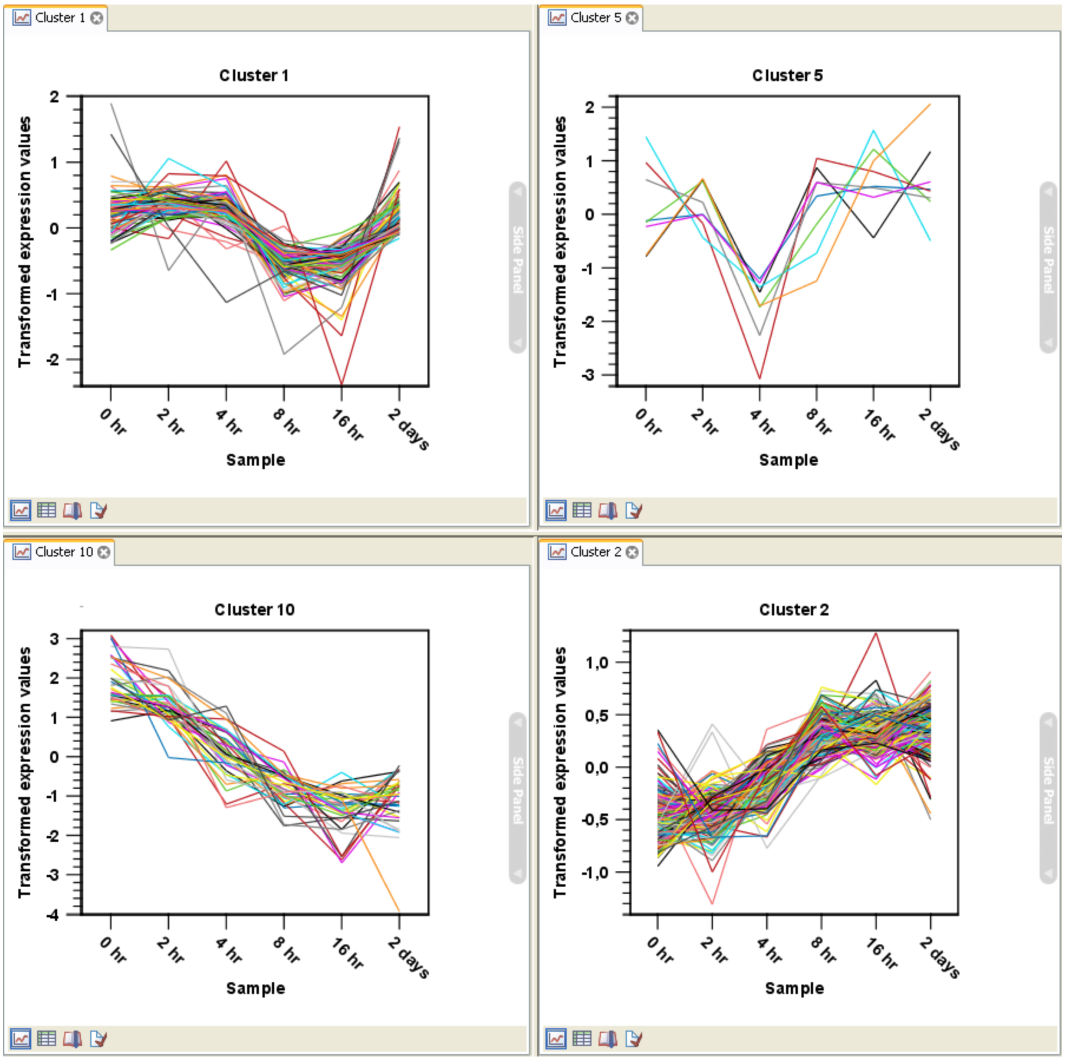


Figure 27.90: *Four clusters created by k-means/medoids clustering.*

Figure 27.90 shows an example where four clusters have been arranged side-by-side.

The samples used are from a time-series experiment, and you can see that the expression levels for each cluster have a distinct pattern. The two clusters at the bottom have falling and rising expression levels, respectively, and the two clusters at the top both fall at the beginning but then rise again (the one to the right starts to rise earlier that the other one).

Having inspected the graphs, you may wish to take a closer look at the features represented in each cluster. In the experiment table, the clustering has added an extra column with the name of the cluster that the feature belongs to. In this way you can filter the table to see only features from a specific cluster. This also means that you can select the feature of this cluster in a volcano or scatter plot as described in section 27.3.6.

27.9 Annotation tests

The annotation tests are tools for detecting significant patterns among features (e.g. genes) of experiments, based on their annotations. This may help in interpreting the analysis of the large numbers of features in an experiment in a biological context. Which biological context, depends on which annotation you choose to examine, and could e.g. be biological process, molecular function or pathway as specified by the Gene Ontology or KEGG. The annotation testing tools of course require that the features in the experiment you want to analyze are annotated. Learn how to annotate an experiment in section 27.3.4.

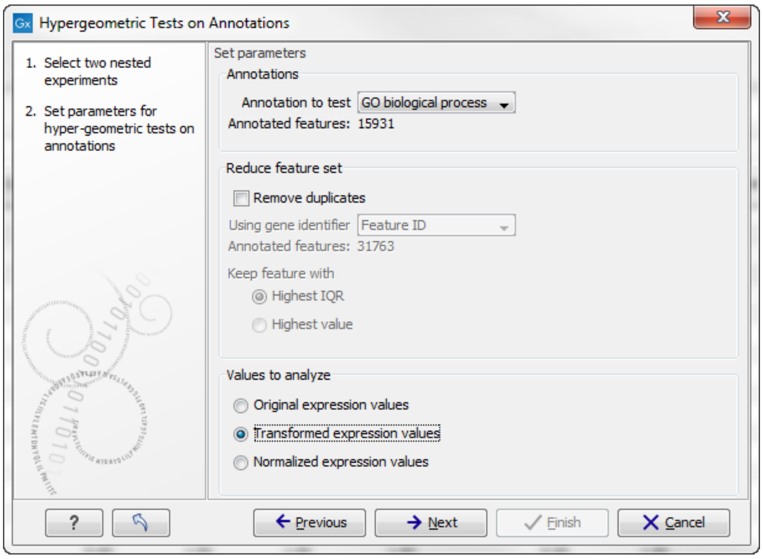
27.9.1 Hypergeometric tests on annotations

The first approach to using annotations to extract biological information is the hypergeometric annotation test. This test measures the extent to which the annotation categories of features in a smaller gene list, 'A', are over or under-represented relative to those of the features in larger gene list 'B', of which 'A' is a sub-list. Gene list B is often the features of the full experiment, possibly with features which are thought to represent only noise, filtered away. Gene list A is a sub-experiment of the full experiment where most features have been filtered away and only those that seem of interest are kept. Typically gene list A will consist of a list of candidate differentially expressed genes. This could be the gene list obtained after carrying out a statistical analysis on the experiment, and choosing to keep only those features with FDR corrected p-values <0.05 and a fold change larger than 2 in absolute value. The hyper geometric test procedure implemented is similar to the unconditional GOstats test of [Falcon and Gentleman, 2007].

> **Toolbox | Transcriptomics Analysis ( )| Annotation Test | Hypergeometric Tests on Annotations ( )**

This will show a dialog where you can select the two experiments - the larger experiment, e.g. the original experiment including the full list of features - and a sub-experiment (see how to create a sub-experiment in section 27.3.2).

Click **Next**. This will display the dialog shown in figure 27.91.


Hypergeometric Tests on Annotations

1. Select two nested experiments
2. Set parameters for hyper-geometric tests on annotations

Set parameters

Annotations

Annotation to test: GO biological process

Annotated features: 15931

Reduce feature set

□ Remove duplicates

Using gene identifier: Feature ID

Annotated features: 31763

Keep feature with

● Highest IQR

○ Highest value

Values to analyze

○ Original expression values

● Transformed expression values

○ Normalized expression values

Figure 27.91: *Parameters for performing a hypergeometric test on annotations.*

At the top, you select which annotation to use for testing. You can select from all the annotations available on the experiment, but it is of course only a few that are biologically relevant. Once you have selected an annotation, you will see the number of features carrying this annotation below.

Annotations are typically given at the gene level. Often a gene is represented by more than one feature in an experiment. If this is not taken into account it may lead to a biased result. The standard way to deal with this is to reduce the set of features considered, so that each gene is represented only once. In the next step, **Remove duplicates**, you can choose the basis on which the feature set will be reduced:

- **Using gene identifier**.
- **Keep feature with**:
  - **Highest IQR**. The feature with the highest interquartile range (IQR) is kept.
  - **Highest value**. The feature with the highest expression value is kept.

First you specify which annotation you want to use as gene identifier. Once you have selected this, you will see the number of features carrying this annotation below. Next you specify which feature you want to keep for each gene. This may be either the feature with the highest inter-quartile range or the highest value.

At the bottom, you can select which values to analyze (see section 27.5.1). Only features that have a numerical value assigned to them will be used for the analysis. That is, any feature which has a value of plus infinity, minus infinity or NaN will not be included in the feature list taken into the test. Thus, the choice of value at this step can affect the features that are taken forward into the test in two ways:

- If there are features with values of plus infinity, minus infinity or NaN, those features will not be taken forward into the test. This can be a consideration when choosing transformed values, where the mathematical manipulations involved may lead to such values.
- If you chose to remove duplicates, then the value type you choose here is the value used for checking the highest IQR or value to determine which feature is taken forward into the test.

The final number of features used for the test is reported in this history view of the test results.

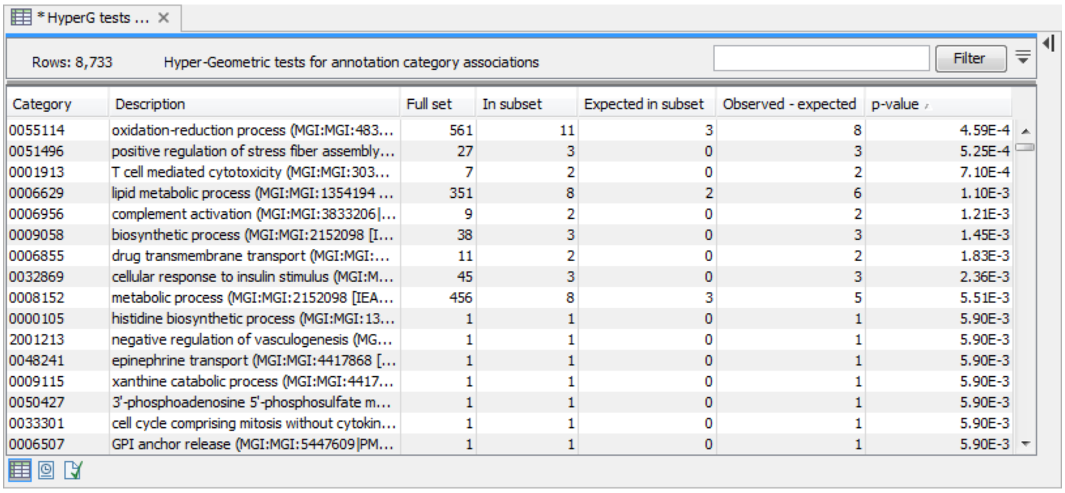
Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Result of hypergeometric tests on annotations

The result of performing hypergeometric tests on annotations using GO biological process is shown in figure 27.92.

The table shows the following information:

- **Category**. This is the identifier for the category.
- **Description**. This is the description belonging to the category. Both of these are simply extracted from the annotations.
- **Full set**. The number of features in the original experiment (not the subset) with this category. (Note that this is after removal of duplicates).



Rows: 8,733 Hyper-Geometric tests for annotation category associations

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
0055114	oxidation-reduction process (MGI:MGI:483...	561	11	3	8	4.59E-4
0051496	positive regulation of stress fiber assembly...	27	3	0	3	5.25E-4
0001913	T cell mediated cytotoxicity (MGI:MGI:303...	7	2	0	2	7.10E-4
0006629	lipid metabolic process (MGI:MGI:1354194 ...	351	8	2	6	1.10E-3
0006956	complement activation (MGI:MGI:3833206|...	9	2	0	2	1.21E-3
0009058	biosynthetic process (MGI:MGI:2152098 [I...	38	3	0	3	1.45E-3
0006855	drug transmembrane transport (MGI:MGI:...	11	2	0	2	1.83E-3
0032869	cellular response to insulin stimulus (MGI:M...	45	3	0	3	2.36E-3
0008152	metabolic process (MGI:MGI:2152098 [IEA...	456	8	3	5	5.51E-3
0000105	histidine biosynthetic process (MGI:MGI:13...	1	1	0	1	5.90E-3
2001213	negative regulation of vasculogenesis (MG...	1	1	0	1	5.90E-3
0048241	epinephrine transport (MGI:MGI:4417868 [...	1	1	0	1	5.90E-3
0009115	xanthine catabolic process (MGI:MGI:4417...	1	1	0	1	5.90E-3
0050427	3'-phosphoadenosine 5'-phosphosulfate m...	1	1	0	1	5.90E-3
0033301	cell cycle comprising mitosis without cytokin...	1	1	0	1	5.90E-3
0006507	GPI anchor release (MGI:MGI:5447609|PM...	1	1	0	1	5.90E-3

Figure 27.92: *The result of testing on GO biological process.*

- **In subset**. The number of features in the subset with this category. (Note that this is after removal of duplicates).
- **Expected in subset**. The number of features we would have expected to find with this annotation category in the subset, if the subset was a random draw from the full set.
- **Observed - expected**. 'In subset' - 'Expected in subset'
- **p-value**. The tail probability of the hyper geometric distribution This is the value used for sorting the table.

Categories with small p-values are categories that are over or under-represented on the features in the subset relative to the full set.

27.9.2 Gene set enrichment analysis

When carrying out a hypergeometric test on annotations you typically compare the annotations of the genes in a subset containing 'the significantly differentially expressed genes' to those of the total set of genes in the experiment. Which, and how many, genes are included in the subset is somewhat arbitrary - using a larger or smaller p-value cut-off will result in including more or less. Also, the magnitudes of differential expression of the genes is not considered.

The Gene Set Enrichment Analysis (GSEA) does NOT take a sublist of differentially expressed genes and compare it to the full list - it takes a single gene list (a single experiment). The idea behind GSEA is to consider a measure of association between the genes and phenotype of interest (e.g. test statistic for differential expression) and rank the genes according to this measure of association. A test is then carried out for each annotation category, for whether the ranks of the genes in the category are evenly spread throughout the ranked list, or tend to occur at the top or bottom of the list.

The GSEA test implemented here is that of [Tian et al., 2005]. The test implicitly calculates and uses a standard t-test statistic for two-group experiments, and ANOVA statistic for multiple group experiments for each feature, as measures of association. For each category, the test statistics for the features in than category are summed and a category based test statistic is calculated as this sum divided by the square root of the number of features in the category. Note that if a feature has the value NaN in one of the samples, the t-test statistic for the feature will be NaN. Consequently, the combined statistic for each of the categories in which the feature is included will be NaN. Thus, it is advisable to filter out any feature that has a NaN value before applying GSEA.

The p-values for the GSEA test statistics are calculated by permutation: The original test statistics for the features are permuted and new test statistics are calculated for each category, based on the permuted feature test statistics. This is done the number of times specified by the user in the wizard. For each category, the lower and upper tail probabilities are calculated by comparing the original category test statistics to the distribution of the permutation-based test statistics for that category. The lower and higher tail probabilities are the number of these that are lower and higher, respectively, than the observed value, divided by the number of permutations.

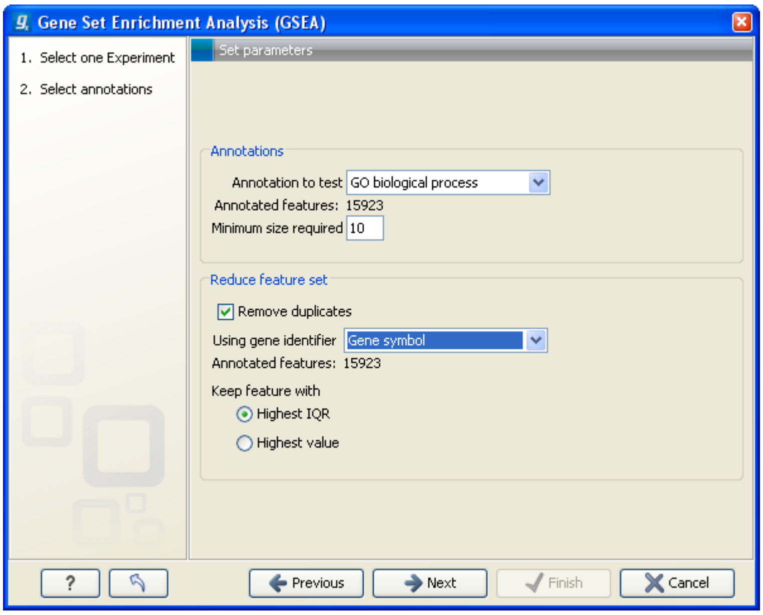
As the p-values are based on permutations you may some times see results where category x's test statistic is lower than that of category y and the categories are of equal size, but where the lower tail probability of category x is higher than that of category y. This is due to imprecision in the estimations of the tail probabilities from the permutations. The higher the number of permutations, the more stable the estimation.

You may run a GSEA on a full experiment, or on a sub-experiment where you have filtered away features that you think are un-informative and represent only noise. Typically you will remove features that are constant across samples (those for which the value in the 'Range' column is zero' — these will have a t-test statistic of zero) and/or those for which the inter-quantile range is small. As the GSEA algorithm calculates and ranks genes on p-values from a test of differential expression, it will generally not make sense to filter the experiment on p-values produced in an analysis if differential expression, prior to running GSEA on it.

Toolbox | Transcriptomics Analysis ( )| Annotation Test | Gene Set Enrichment Analysis (GSEA) ( )

Select an experiment and click **Next**.

Click **Next**. This will display the dialog shown in figure 27.93.



Annotations
Annotation to test: GO biological process
Annotated features: 15923
Minimum size required: 10

Reduce feature set
☑ Remove duplicates
Using gene identifier: Gene symbol
Annotated features: 15923
Keep feature with
● Highest IQR
○ Highest value

Figure 27.93: *Gene set enrichment analysis on GO biological process*

At the top, you select which annotation to use for testing. You can select from all the annotations available on the experiment, but it is of course only a few that are biologically relevant. Once you have selected an annotation, you will see the number of features carrying this annotation below.

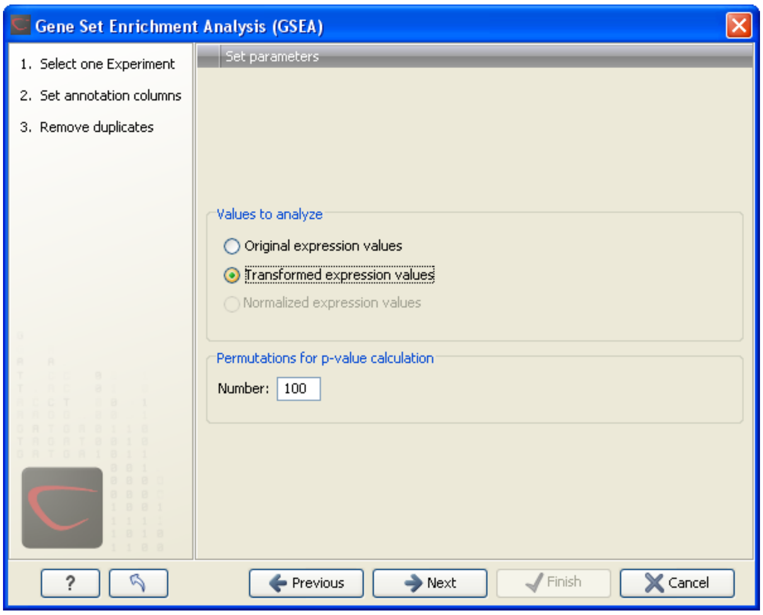
In addition, you can set a filter: **Minimum size required**. Only categories with more genes (i.e. features) than the specified number will be considered. Excluding categories with small numbers of genes may lead to more robust results.

Annotations are typically given at the gene level. Often a gene is represented by more than one feature in an experiment. If this is not taken into account it may lead to a biased result. The standard way to deal with this is to reduce the set of features considered, so that each gene is represented only once. Check the **Remove duplicates** check box to reduce the feature set, and you can choose how you want this to be done:

- **Using gene identifier**.
- **Keep feature with**:
  - **Highest IQR**. The feature with the highest interquartile range (IQR) is kept.
  - **Highest value**. The feature with the highest expression value is kept.

First you specify which annotation you want to use as gene identifier. Once you have selected this, you will see the number of features carrying this annotation below. Next you specify which feature you want to keep for each gene. This may be either the feature with the highest inter-quartile range or the highest value.

Clicking **Next** will display the dialog shown in figure 27.94.



Values to analyze

○ Original expression values

● Transformed expression values

○ Normalized expression values

Permutations for p-value calculation

Number: 100

Figure 27.94: *Gene set enrichment analsysis parameters.*

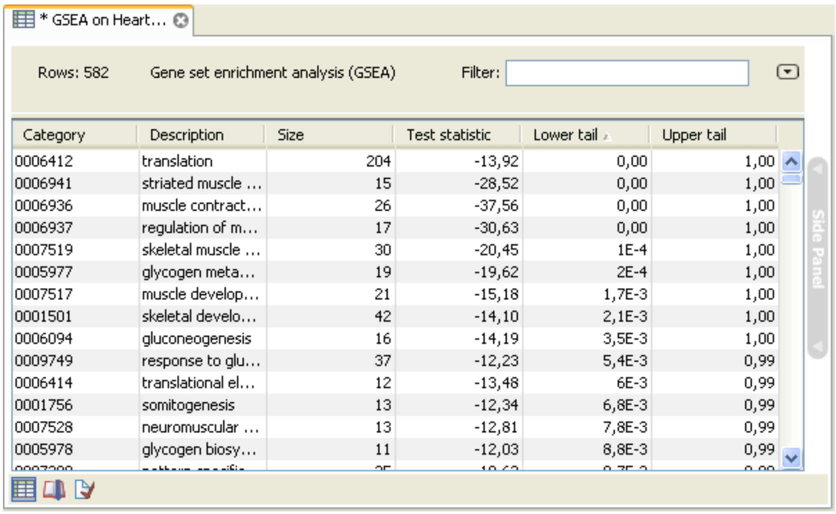
At the top, you can select which values to analyze (see section 27.5.1).

Below, you can set the **Permutations for p-value calculation**. For the GSEA test a p-value is calculated by permutation: p permuted data sets are generated, each consisting of the original features, but with the test statistics permuted. The GSEA test is run on each of the permuted data sets. The test statistic is calculated on the original data, and the resulting value is compared to the distribution of the values obtained for the permuted data sets. The permutation based p-value is the number of permutation based test statistics above (or below) the value of the test statistic for the original data, divided by the number of permuted data sets. For reliable permutation-based p-value calculation a large number of permutations is required (100 is the default).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Result of gene set enrichment analysis

The result of performing gene set enrichment analysis using GO biological process is shown in figure 27.95.



Rows: 582 Gene set enrichment analysis (GSEA) Filter:

Category	Description	Size	Test statistic	Lower tail	Upper tail
0006412	translation	204	-13,92	0,00	1,00
0006941	striated muscle ...	15	-28,52	0,00	1,00
0006936	muscle contract...	26	-37,56	0,00	1,00
0006937	regulation of m...	17	-30,63	0,00	1,00
0007519	skeletal muscle ...	30	-20,45	1E-4	1,00
0005977	glycogen meta...	19	-19,62	2E-4	1,00
0007517	muscle develop...	21	-15,18	1,7E-3	1,00
0001501	skeletal develo...	42	-14,10	2,1E-3	1,00
0006094	gluconeogenesis	16	-14,19	3,5E-3	1,00
0009749	response to glu...	37	-12,23	5,4E-3	0,99
0006414	translational el...	12	-13,48	6E-3	0,99
0001756	somitogenesis	13	-12,34	6,8E-3	0,99
0007528	neuromuscular ...	13	-12,81	7,8E-3	0,99
0005978	glycogen biosy...	11	-12,03	8,8E-3	0,99

Figure 27.95: *The result of gene set enrichment analysis on GO biological process.*

The table shows the following information:

- **Category**. This is the identifier for the category.
- **Description**. This is the description belonging to the category. Both of these are simply extracted from the annotations.
- **Size**. The number of features with this category. (Note that this is after removal of duplicates).
- **Test statistic**. This is the GSEA test statistic.
- **Lower tail**. This is the mass in the permutation based p-value distribution below the value of the test statistic.
- **Upper tail**. This is the mass in the permutation based p-value distribution above the value of the test statistic.

A small lower (or upper) tail p-value for an annotation category is an indication that features in this category viewed as a whole are perturbed among the groups in the experiment considered.

27.10 General plots

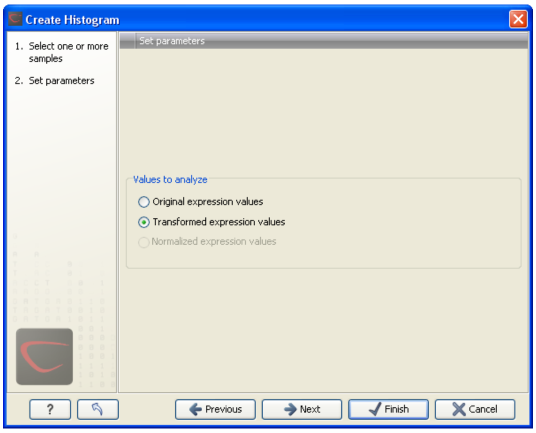
The last folder in the **Expression Analysis** ( ) folder in the **Toolbox** is **General Plots**. Here you find three general plots that may be useful at various point of your analysis work flow. The plots are explained in detail below.

27.10.1 Histogram

A histogram shows a distribution of a set of values. Histograms are often used for examining and comparing distributions, e.g. of expression values of different samples, in the quality control step of an analysis. You can create a histogram showing the distribution of expression value for a sample:

Toolbox | **Transcriptomics Analysis ( )**| **General Plots** | **Create Histogram ( )**

Select a number of samples (( ) or ( )) or a graph track. When you have selected more than one sample, a histogram will be created for each one. Clicking **Next** will display a dialog as shown in figure 27.96.


Create Histogram

1. Select one or more samples
2. Set parameters

Set parameters

Values to analyze

○ Original expression values

● Transformed expression values

○ Normalized expression values

Figure 27.96: *Selcting which values the histogram should be based on.*

In this dialog, you select the values to be used for creating the histogram (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Viewing histograms

The resulting histogram is shown in a figure 27.97

The histogram shows the expression value on the x axis (in the case of figure 27.97 the transformed expression values) and the counts of these values on the y axis.

In the **Side Panel** to the left, there is a number of options to adjust the view. Under **Graph preferences**, you can adjust the general properties of the plot.

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.

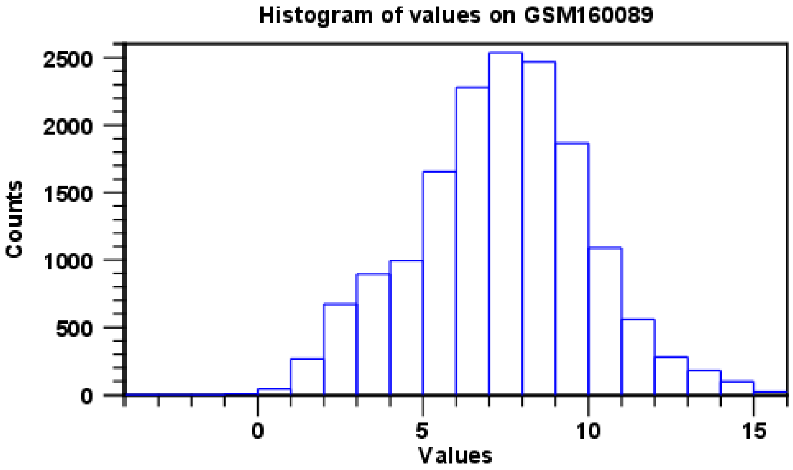


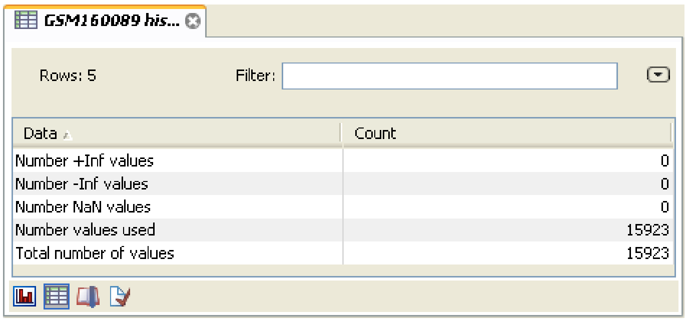
Figure 27.97: *Histogram showing the distribution of transformed expression values.*

- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Break points**. Determines where the bars in the histogram should be:
  - **Sturges method**. This is the default. The number of bars is calculated from the range of values by Sturges formula [Sturges, 1926].
  - **Equi-distanced bars**. This will show bars from **Start** to **End** and with a width of **Sep**.
  - **Number of bars**. This will simply create a number of bars starting at the lowest value and ending at the highest value.

Below the graph preferences, you find **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

Note that if you wish to use the same settings next time you open a principal component plot, you need to save the settings of the **Side Panel** (see section 4.5).

Besides the histogram view itself, the histogram can also be shown in a table, summarizing key properties of the expression values. An example is shown in figure 27.98.



Rows: 5 Filter:

Data	Count
Number +Inf values	0
Number -Inf values	0
Number NaN values	0
Number values used	15923
Total number of values	15923

Figure 27.98: *Table view of a histogram.*

The table lists the following properties:

- **Number +Inf values**
- **Number -Inf values**
- **Number NaN values**
- **Number values used**
- **Total number of values**

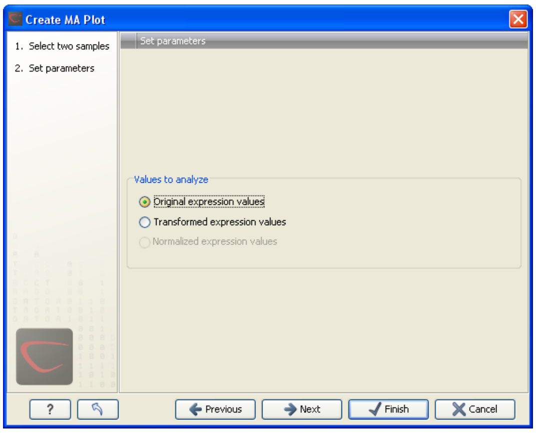
27.10.2 MA plot

The MA plot is a scatter rotated by $45°$. For two samples of expression values it plots for each gene the difference in expression against the mean expression level. MA plots are often used for quality control, in particular, to assess whether normalization and/or transformation is required.

You can create an MA plot comparing two samples:

Toolbox | **Transcriptomics Analysis ( )**| **General Plots** | **Create MA Plot ( )**

Select two samples (( ) or ( )). Clicking **Next** will display a dialog as shown in figure 27.99.



1. Select two samples
2. Set parameters

Set parameters

Values to analyze
● Original expression values
○ Transformed expression values
○ Normalized expression values

Figure 27.99: *Selcting which values the MA plot should be based on.*

In this dialog, you select the values to be used for creating the MA plot (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

Viewing MA plots

The resulting plot is shown in a figure 27.100.

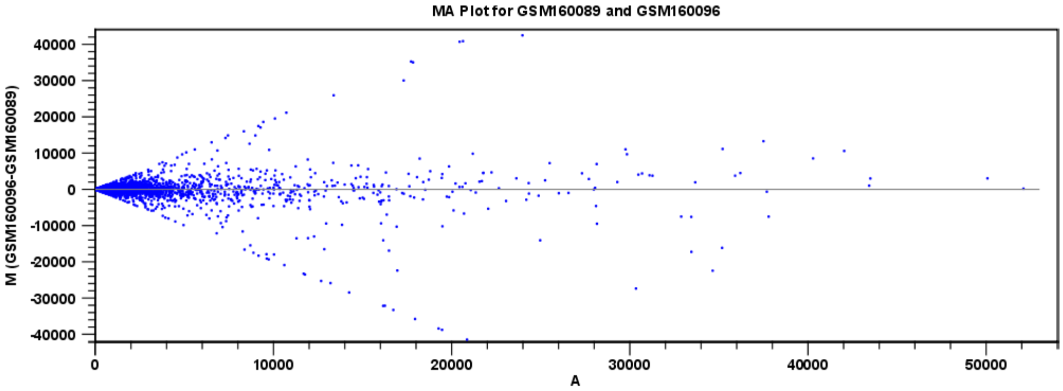


Figure 27.100: *MA plot based on original expression values.*

The X axis shows the mean expression level of a feature on the two samples and the Y axis shows the difference in expression levels for a feature on the two samples. From the plot shown in figure 27.100 it is clear that the variance increases with the mean. With an MA plot like this, you will often choose to transform the expression values (see section 27.5.2).

Figure 27.101 shows the same two samples where the MA plot has been created using log2 transformed values.

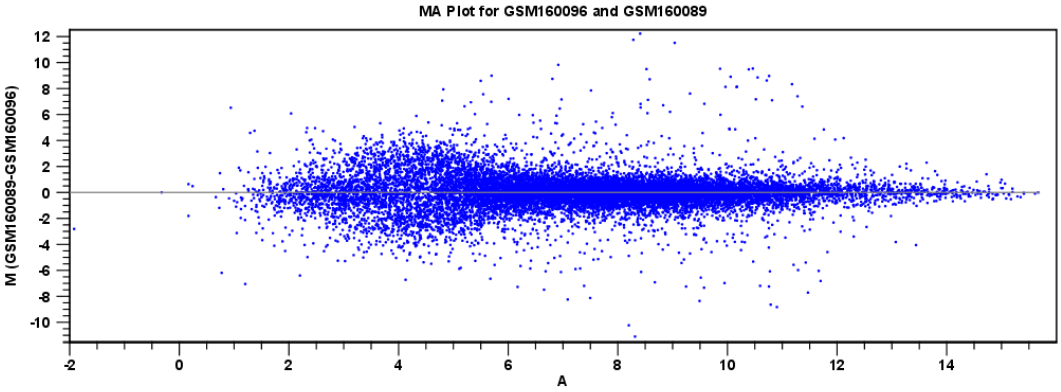


Figure 27.101: *MA plot based on transformed expression values.*

The much more symmetric and even spread indicates that the dependance of the variance on the mean is not as strong as it was before transformation.

In the **Side Panel** to the left, there is a number of options to adjust the view. Under **Graph preferences**, you can adjust the general properties of the plot.

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.

- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **y = 0 axis**. Draws a line where y = 0. Below there are some options to control the appearance of the line:
  - **Line width**
    - Thin
    - Medium
    - Wide
  - **Line type**
    - None
    - Line
    - Long dash
    - Short dash
  - **Line color.** Allows you to choose between many different colors. Click the color box to select a color.
- **Line width**
  - Thin
  - Medium
  - Wide
- **Line type**
  - None
  - Line
  - Long dash
  - Short dash
- **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

Below the general preferences, you find the **Dot properties** preferences, where you can adjust coloring and appearance of the dots:

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.

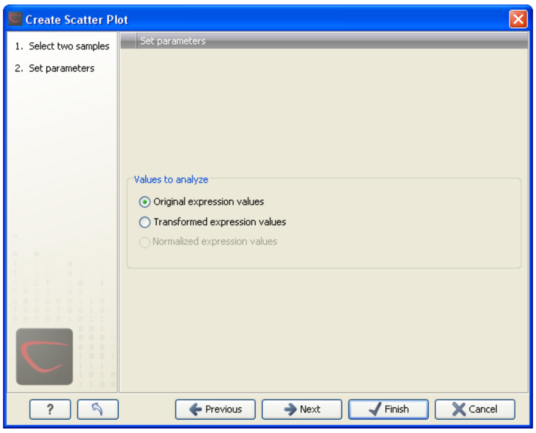
Note that if you wish to use the same settings next time you open a scatter plot, you need to save the settings of the **Side Panel** (see section 4.5).

27.10.3 Scatter plot

As described in section 27.3.5, an experiment can be viewed as a scatter plot. However, you can also create a "stand-alone" scatter plot of two samples:

Toolbox | Transcriptomics Analysis ()| General Plots | Create Scatter Plot ( )

Select two samples (() or ()). Clicking **Next** will display a dialog as shown in figure 27.102.



Create Scatter Plot

1. Select two samples
2. Set parameters

Set parameters

Values to analyze

● Original expression values

○ Transformed expression values

○ Normalized expression values

Figure 27.102: *Selcting which values the scatter plot should be based on.*

In this dialog, you select the values to be used for creating the scatter plot (see section 27.5.1).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

For more information about the scatter plot view and how to interpret it, please see section 27.3.5.

Chapter 28

Helper tools

Contents

28.1 Extract sequences

This tool allows the extraction of sequences from other types of data in the Workbench, such as sequence lists or alignments. The data types you can extract sequences from are:

- Alignments ()
- BLAST result ()
- BLAST overview tables ()
- sequence lists ()
- Contigs and read mappings ()
- Read mapping tables ()
- Read mapping tracks ()
- RNA-Seq mapping results ()

Note! When the Extract Sequences tool is run via the Workbench toolbox on an entire file of one of the above types, **all** sequences are extracted from the data used as input. If only a **subset** of the sequences is desired, for example, the reads from just a small area of a mapping, or the sequences for only a few blast results, then a data set containing just this subsection or subset should be created and the Extract Sequences tool should be run on that.

For extracting a subset of a mapping, please see section 29.7.5 that describes the function "Extract from Selection" that also can be selected from the right click menu (see figure 28.1).

For extracting a subset of a sequence list, you can highlight the sequences of interest in the table view of the sequence list, right click on the selection and launch the Extract Sequences tool.

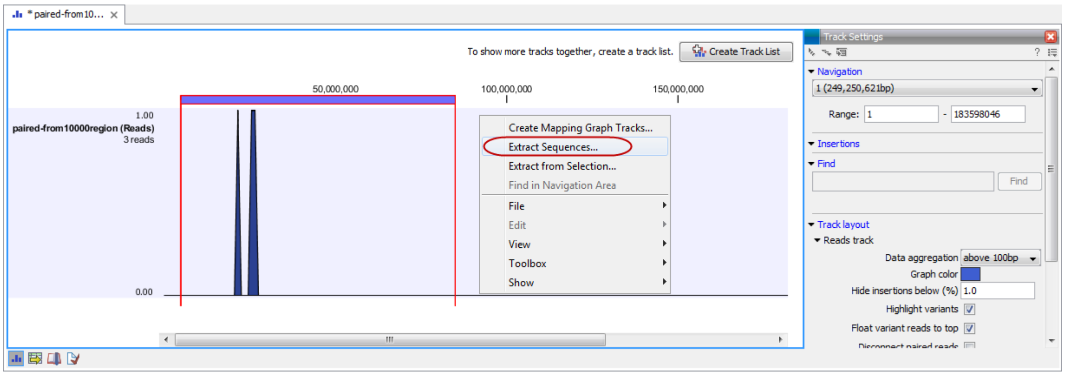
The Extract Sequences tool can be launched via the Toolbox menu, by going to:

Toolbox | **General Sequence Analysis** ()| **Extract Sequences** ()

Alternatively, on all the data types listed above except sequence lists, the option to run this tool appears by right clicking in the relevant area; a row in a table or in the read area of mapping data. An example is shown in figure 28.1.

Please note that for mappings, only the read sequences are extracted. Reference and consensus sequences are not extracted using this tool. Similarly, when extracting sequences from BLAST results, the sequence hits are extracted, not the original query sequence or a consensus sequence.

"Note also, that paired reads will be extracted in accordance with the read group settings, which is specified during the original import of the reads. If the orientation has since been changed (e.g. using the Element Info tab for the sequence list) the read group information will be modified and reads will be extracted as specified by the modified read group. The default read group orientation is forward-reverse."



To show more tracks together, create a track list.

Navigation: 1 (249,250,621bp)

Range: 1 - 183598046

Data aggregation: above 100bp

Hide insertions below (%): 1.0

Highlight variants ☑

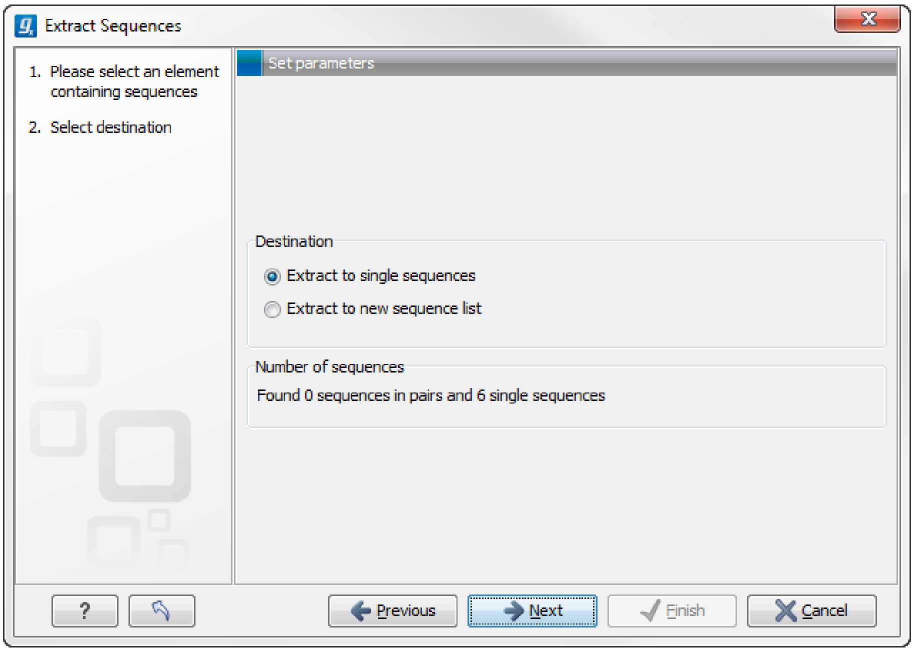
Float variant reads to top ☑

Disconnect paired reads □

Figure 28.1: *Right click somewhere in the reads track area and select "Extract Sequences".*

The dialog allows you to select the **Destination**. Here you can choose whether the extracted sequences should be extracted as single sequences or placed in a new sequence list. For most data types, it will make most sense to choose to extract the sequences into a sequence list. The exception to this is when working with a sequence list, where choosing to extract to a sequence list would create a copy of the same sequence list. In this case, the other option would generally be chosen. This would then result in the generation of individual sequence objects for each sequence in the sequence list.

Below these options, in the dialog, you can see the number of sequences that will be extracted.



Destination

● Extract to single sequences

○ Extract to new sequence list

Number of sequences

Found 0 sequences in pairs and 6 single sequences

Figure 28.2: *Choosing whether the extracted sequences should be placed in a new list or as single sequences.*

Part IX

Sanger Sequencing

Chapter 29

Sequencing Data Analysis

Contents

CLC Cancer Research Workbench lets you import, trim and assemble DNA sequence reads from automated sequencing machines. A number of different formats are supported (see section 6.1). This chapter explains the features in *CLC Cancer Research Workbench* for handling data analysis of low-throughput conventional Sanger sequencing data.

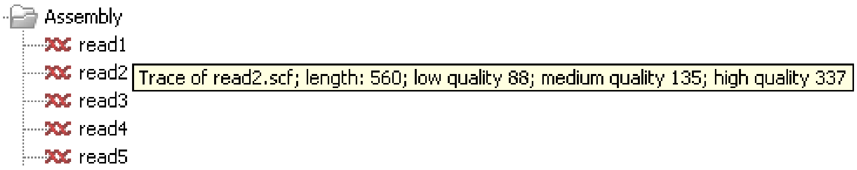
This chapter first explains how to trim sequence reads. Next follows a description of how to assemble reads into contigs both with and without a reference sequence. In the final section, the options for viewing and editing contigs are explained.

29.1 Importing and viewing trace data

A number of different binary trace data formats can be imported into the program, including *Standard Chromatogram Format (.SCF)*, *ABI sequencer data files (.ABI and .AB1)*, *PHRED output files (.PHD)* and *PHRAP output files (.ACE)* (see section 6.1).

After import, the sequence reads and their trace data are saved as DNA sequences. This means that all analyses that apply to DNA sequences can be performed on the sequence reads, including e.g. BLAST and open reading frame prediction.

You can see additional information about the quality of the traces by holding the mouse cursor on the imported sequence. This will display a tool tip as shown in figure 29.1.



Assembly
read1
read2 Trace of read2.scf; length: 560; low quality 88; medium quality 135; high quality 337
read3
read4
read5

Figure 29.1: *A tooltip displaying information about the quality of the chromatogram.*

The qualities are based on the phred scoring system, with scores below 19 counted as low quality, scores between 20 and 39 counted as medium quality, and those 40 and above counted as high quality.

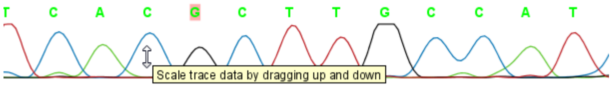
If the trace file does not contain information about quality, only the sequence length will be shown.

To view the trace data, open the sequence read in a standard sequence view (ACT).

29.1.1 Scaling traces

The traces can be scaled by dragging the trace vertically as shown in figure figure 29.2. The Workbench automatically adjust the height of the traces to be readable, but if the trace height varies a lot, this manual scaling is very useful.

The height of the area available for showing traces can be adjusted in the **Side Panel** as described insection 29.1.2.



Scale trace data by dragging up and down

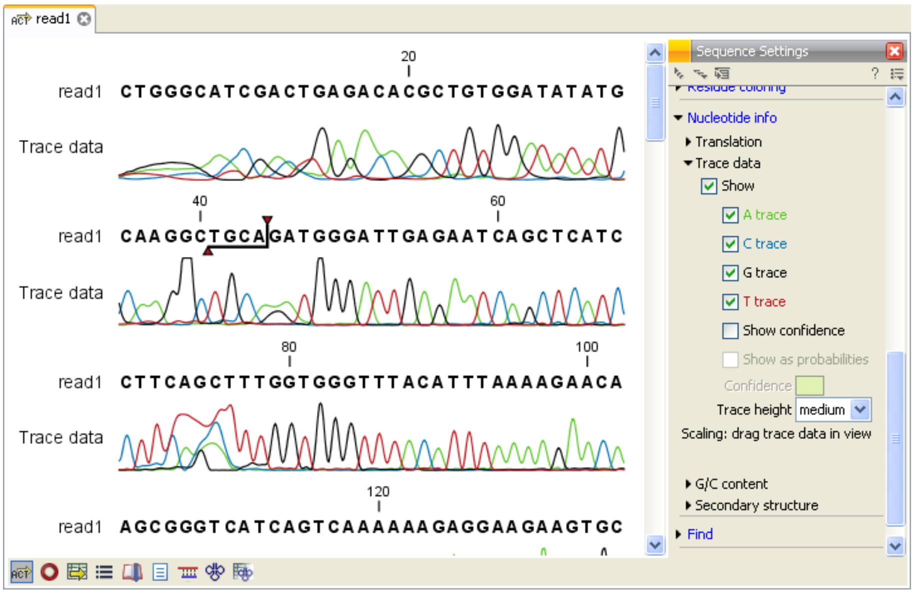
Figure 29.2: *Grab the traces to scale.*

29.1.2 Trace settings in the Side Panel

In the Nucleotide info preference group the display of trace data can be selected and unselected. When selected, the trace data information is shown as a plot beneath the sequence. The appearance of the plot can be adjusted using the following options (see figure 29.3):

- **Nucleotide trace.** For each of the four nucleotides the trace data can be selected and unselected.

- **Scale traces.** A slider which allows the user to scale the height of the trace area. Scaling the traces individually is described in section 29.1.1.


read1
20
read1 CTGGGCATCGACTGAGACACGCTGTGGATATATG
Trace data
40
60
read1 CAAGGCTGCAGATGGGATTGAGAATCAGCTCATC
Trace data
80
100
read1 CTTCAGCTTTGGTGGGTTTACATTTAAAAGAACA
Trace data
120
read1 AGCGGGTCATCAGTCAAAAAAGAGGAAGAAGTGC
Sequence Settings
Residue coloring
Nucleotide info
Translation
G/C content
Secondary structure
Find

Trace data
☑ Show
☑ A trace
☑ C trace
☑ G trace
☑ T trace
☐ Show confidence
☐ Show as probabilities
Confidence
Trace height: medium
Scaling: drag trace data in view

Figure 29.3: *A sequence with trace data. The preferences for viewing the trace are shown in the Side Panel.*

When working with stand along mappings containing reads with trace data, you can view the traces by turning on the trace setting options as described here **and** choosing **Not compact** in the Read layout setting for the mapping. Please see section 29.7.1.

29.2 Trim sequences

Trimming as described in this section involves marking of low quality and/or vector sequence with a Trim annotation as shown in figure 29.4). Such annotated regions are then ignored when using downstream analysis tools located in the same section of the Workbench toolbox, for example Assembly (see section 29.3). The trimming described here annotates, but does not remove data, allowing you to explore the output of different trimming schemes easily.

Trimming as a separate task can be done manually or using a tool designed specifically for this task.

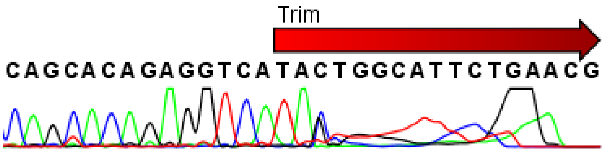


Figure 29.4: *Trimming creates annotations on the regions that will be ignored in the assembly process.*

Note! If you wish to remove regions that are trimmed, you should instead use the NGS trim tool (see section 19.2).

When exporting sequences in fasta format, there is an option to remove the parts of the sequence covered by trim annotations.

29.2.1 Trimming using the Trim tool

Sequence reads can be trimmed based on a number of different criteria. Using a trimming tool for this is particularly useful if:

- You have many sequences to trim.
- You wish to trim vector contamination from sequencing reads.
- You wish to ensure that consistency when trimming. That is, you wish to ensure the same criteria are used for all the sequences in a set.

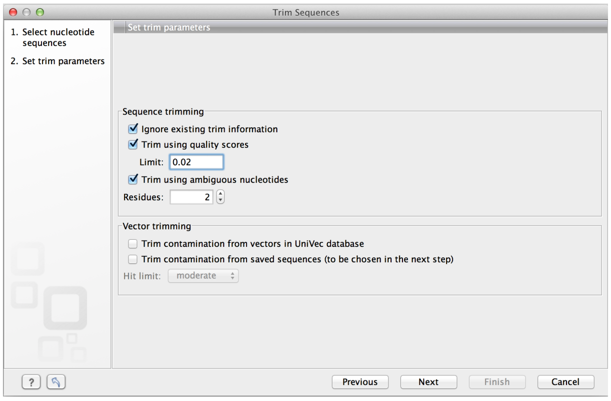
To start up the Trim tool in the Workbench, go to the menu option:

Toolbox | **Sequencing Data Analysis** ()| **Trim Sequences** ()

This opens a dialog where you can choose the sequences to trim, by using the arrows to move them between the Navigation Area and the 'Selected Elements' box.

When the sequences are selected, click **Next**.

This opens the dialog displayed in figure 29.5.



Trim Sequences

1. Select nucleotide sequences
2. Set trim parameters

Set trim parameters

Sequence trimming

☑ Ignore existing trim information

☑ Trim using quality scores

Limit: 0.02

☑ Trim using ambiguous nucleotides

Residues: 2

Vector trimming

☐ Trim contamination from vectors in UniVec database

☐ Trim contamination from saved sequences (to be chosen in the next step)

Hit limit: moderate

Figure 29.5: *Setting parameters for trimming.*

The following parameters can be adjusted in the dialog:

- **Ignore existing trim information.** If you have previously trimmed the sequences, you can check this to remove existing trimming annotation prior to analysis.
- **Trim using quality scores.** If the sequence files contain quality scores from a base-caller algorithm this information can be used for trimming sequence ends. The program uses the modified-Mott trimming algorithm for this purpose (Richard Mott, personal communication):

  Quality scores in the Workbench are on a Phred scale, and formats using other scales will be converted during import. The Phred quality scores (Q), defined as: $Q = -10log10(P)$, where P is the base-calling error probability, can then be used to calculate the error probabilities, which in turn can be used to set the limit for, which bases should be trimmed.

  Hence, the first step in the trim process is to convert the quality score (Q) to an error probability: $p_{error} = 10^{\frac{Q}{-10}}$. (This now means that low values are high quality bases.)

Next, for every base a new value is calculated: $Limit - p_{error}$. This value will be negative for low quality bases, where the error probability is high.

For every base, the Workbench calculates the running sum of this value. If the sum drops below zero, it is set to zero. The part of the sequence not trimmed will be the region ending at the highest value of the running sum and starting at the last zero value before this highest score. Everything before and after this region will be trimmed. A read will be completely removed if the score never makes it above zero.

At http://www.clcbio.com/files/usermanuals/trim.zip you find an example sequence and an Excel sheet showing the calculations done for this particular sequence to illustrate the procedure described above.

- **Trim ambiguous nucleotides.** This option trims the sequence ends based on the presence of ambiguous nucleotides (typically N). Note that the automated sequencer generating the data must be set to output ambiguous nucleotides in order for this option to apply. The algorithm takes as input the *maximal number of ambiguous nucleotides allowed in the sequence after trimming*. If this maximum is set to e.g. 3, the algorithm finds the maximum length region containing 3 or fewer ambiguities and then trims away the ends not included in this region.

- **Trim contamination from vectors in UniVec database.** If selected, the program will match the sequence reads against all vectors in the UniVec database and mark sequence ends with significant matches with a 'Trim' annotation (the database is included when you install the *CLC Cancer Research Workbench*). A list of all the vectors in the UniVec database can be found at http://www.ncbi.nlm.nih.gov/VecScreen/replist.html.

  - **Hit limit.** Specifies how strictly vector contamination is trimmed. Since vector contamination usually occurs at the beginning or end of a sequence, different criteria are applied for terminal and internal matches. A match is considered terminal if it is located within the first 25 bases at either sequence end. Three match categories are defined according to the expected frequency of an alignment with the same score occurring between random sequences. The *CLC Cancer Research Workbench* uses the same settings as VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html):

    - **Weak.** Expect 1 random match in 40 queries of length 350 kb
      - Terminal match with Score 16 to 18.
      - Internal match with Score 23 to 24.
    - **Moderate.** Expect 1 random match in 1,000 queries of length 350 kb
      - Terminal match with Score 19 to 23.
      - Internal match with Score 25 to 29.
    - **Strong.** Expect 1 random match in 1,000,000 queries of length 350 kb
      - Terminal match with Score $\geq$ 24.
      - Internal match with Score $\geq$ 30.

    Note that selecting e.g. **Weak** will also include matches in the **Moderate** and **Strong** categories.

- **Trim contamination from saved sequences.** This option lets you select your own vector sequences that you have imported into the Workbench. If you select this option, you will be able to select one or more sequences when you click **Next**.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. This will start the trimming process. Views of each trimmed sequence will be shown, and you can inspect the result by looking at the "Trim" annotations (they are colored red as default). Note that the trim annotations are used to signal that this part of the sequence is to be ignored during further analyses, hence the trimmed sequences are not deleted. If there are no trim annotations, the sequence has not been trimmed.

29.2.2 Manual trimming

Sequence reads can be trimmed manually while inspecting their trace and quality data.

> **double-click the sequence to trim in the Navigation Area | select the region you want to trim | right-click the selection | Trim sequence left/right to determine the direction of the trimming**

This will add a trimming annotation to the end of the sequence in the selected direction. No sequence is being deleted here. Rather, the regions covered by trim annotations are noted by downstream analyses (in the same section of the Workbench Toolbox as the Trim tool) as regions to be ignored.

29.3 Assemble sequences

This section describes how to assemble a number of sequence reads into a contig without the use of a reference sequence (a known sequence that can be used for comparison with the other sequences, see section 29.5). To perform the assembly:

> **Toolbox | Sequencing Data Analysis ( )| Assemble Sequences ( )**

This will open a dialog where you can select sequences to assemble. If you already selected sequences in the Navigation Area, these will be shown in 'Selected Elements'. You can alter your choice of sequences to assemble, or add others, by using the arrows to move sequences between the Navigation Area and the 'Selected Elements' box. You can also add sequence lists.

Note! You can assemble a maximum of 2000 sequences at a time.

To assemble more sequences, please use the **De Novo Assembly** ( ) tool under **De Novo Sequencing ( )** in the **Toolbox** instead.

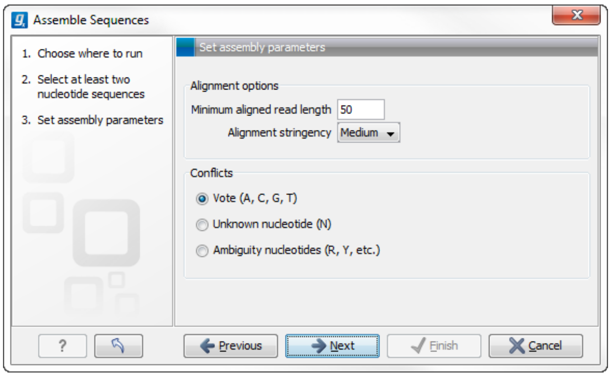
To assemble more sequences, you need the *CLC Genomics Workbench* (see `http://www.clcbio.com/genomics`).

When the sequences are selected, click **Next**. This will show the dialog in figure 29.6

This dialog gives you the following options for assembly:

- **Minimum aligned read length.** The minimum number of nucleotides in a read which must be successfully aligned to the contig. If this criteria is not met by a read, the read is excluded from the assembly.

- **Alignment stringency.** Specifies the stringency of the scoring function used by the alignment step in the contig assembly algorithm. A higher stringency level will tend to produce contigs with fewer ambiguities but will also tend to omit more sequencing reads and to generate more and shorter contigs. Three stringency levels can be set:



1. Choose where to run
2. Select at least two nucleotide sequences
3. Set assembly parameters

Set assembly parameters

Alignment options

Minimum aligned read length 50

Alignment stringency Medium

Conflicts

● Vote (A, C, G, T)

○ Unknown nucleotide (N)

○ Ambiguity nucleotides (R, Y, etc.)

Figure 29.6: *Setting assembly parameters.*

  - **Low.**
  - **Medium.**
  - **High.**

- **Conflicts.** If there is a conflict, i.e. a position where there is disagreement about the residue (A, C, T or G), you can specify how the contig sequence should reflect the conflict:
  - **Vote (A, C, G, T).** The conflict will be solved by counting instances of each nucleotide and then letting the majority decide the nucleotide in the contig. In case of equality, ACGT are given priority over one another in the stated order.
  - **Unknown nucleotide (N).** The contig will be assigned an 'N' character in all positions with conflicts (conflicts are registered already when two nucleotides differ).
  - **Ambiguity nucleotides (R, Y, etc.).** The contig will display an ambiguity nucleotide reflecting the different nucleotides found in the reads (nucleotide ambiguity is registered already when two nucleotides differ). For an overview of ambiguity codes, see Appendix D.

  Note, that conflicts will always be highlighted no matter which of the options you choose. Furthermore, each conflict will be marked as annotation on the contig sequence and will be present if the contig sequence is extracted for further analysis. As a result, the details of any experimental heterogeneity can be maintained and used when the result of single-sequence analyzes is interpreted. Read more about conflicts in section 29.7.4.

- **Create full contigs, including trace data.** This will create a contig where all the aligned reads are displayed below the contig sequence. (You can always extract the contig sequence without the reads later on.) For more information on how to use the contigs that are created, see section 29.7.

- **Show tabular view of contigs.** A contig can be shown both in a graphical as well as a tabular view. If you select this option, a tabular view of the contig will also be opened (Even if you do not select this option, you can show the tabular view of the contig later on by clicking **Table** (▦) at the bottom of the view.) For more information about the tabular view of contigs, see section 29.7.6.

- **Create only consensus sequences.** This will not display a contig but will only output the assembled contig sequences as single nucleotide sequences. If you choose this option it is not possible to validate the assembly process and edit the contig based on the traces.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

When the assembly process has ended, a number of views will be shown, each containing a contig of two or more sequences that have been matched. If the number of contigs seem too high or low, try again with another **Alignment stringency** setting. Depending on your choices of output options above, the views will include trace files or only contig sequences. However, the calculation of the contig is carried out the same way, no matter how the contig is displayed.

See section 29.7 on how to use the resulting contigs.

29.4 Sort Sequences By Name

With this functionality you will be able to group sequencing reads based on their file name. A typical example would be that you have a list of files named like this:

```
...
A02__Asp_F_016_2007-01-10
A02__Asp_R_016_2007-01-10
A02__Gln_F_016_2007-01-11
A02__Gln_R_016_2007-01-11
A03__Asp_F_031_2007-01-10
A03__Asp_R_031_2007-01-10
A03__Gln_F_031_2007-01-11
A03__Gln_R_031_2007-01-11
...
```

In this example, the names have five distinct parts (we take the first name as an example):

- **A02** which is the position on the 96-well plate
- **Asp** which is the name of the gene being sequenced
- **F** which describes the orientation of the read (forward/reverse)
- **016** which is an ID identifying the sample
- **2007-01-10** which is the date of the sequencing run

To start mapping these data, you probably want to have them divided into groups instead of having all reads in one folder. If, for example, you wish to map each sample separately, or if you wish to map each gene separately, you cannot simply run the mapping on all the sequences in one step.

That is where **Sort Sequences by Name** comes into play. It will allow you to specify which part of the name should be used to divide the sequences into groups. We will use the example described above to show how it works:

Toolbox | **Molecular Biology Tools ( ) | Sort Sequences by Name ( )**

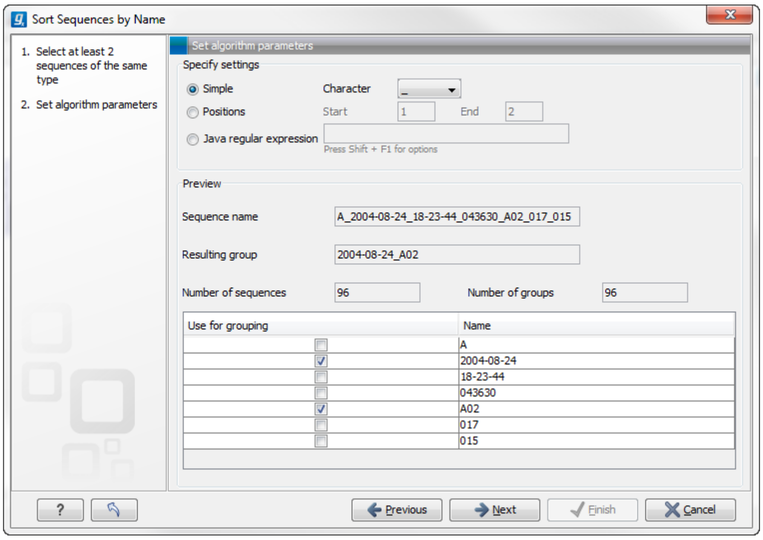
This opens a dialog where you can add the sequences you wish to sort, by using the arrows to move them between the Navigation Area and 'Selected Elements'. You can also add sequence lists or the contents of an entire folder by right-clicking the folder and choose: **Add folder contents**.

When you click **Next**, you will be able to specify the details of how the grouping should be performed. First, you have to choose how each part of the name should be identified. There are three options:

- **Simple**. This will simply use a designated character to split up the name. You can choose a character from the list:
  - Underscore _
  - Dash -
  - Hash (number sign / pound sign) #
  - Pipe |
  - Tilde ~
  - Dot .
- **Positions**. You can define a part of the name by entering the start and end positions, e.g. from character number 6 to 14. For this to work, the names have to be of equal lengths.
- **Java regular expression**. This is an option for advanced users where you can use a special syntax to have total control over the splitting. See more below.

In the example above, it would be sufficient to use a simple split with the underscore _ character, since this is how the different parts of the name are divided.

When you have chosen a way to divide the name, the parts of the name will be listed in the table at the bottom of the dialog. There is a checkbox next to each part of the name. This checkbox is used to specify which of the name parts should be used for grouping. In the example above, if we want to group the reads according to date and analysis position, these two parts should be checked as shown in figure 29.7.



Sort Sequences by Name

1. Select at least 2 sequences of the same type
2. Set algorithm parameters

Set algorithm parameters

Specify settings

● Simple Character: _
○ Positions Start: 1 End: 2
○ Java regular expression
Press Shift + F1 for options

Preview

Sequence name: A_2004-08-24_18-23-44_043630_A02_017_015
Resulting group: 2004-08-24_A02
Number of sequences: 96
Number of groups: 96

Use for grouping	Name
□	A
☑	2004-08-24
□	18-23-44
□	043630
☑	A02
□	017
□	015

Figure 29.7: *Splitting up the name at every underscore (_) and using the date and analysis position for grouping.*

At the middle of the dialog there is a preview panel listing:

- **Sequence name**. This is the name of the first sequence that has been chosen. It is shown here in the dialog in order to give you a sample of what the names in the list look like.
- **Resulting group**. The name of the group that this sequence would belong to if you proceed with the current settings.
- **Number of sequences**. The number of sequences chosen in the first step.
- **Number of groups**. The number of groups that would be produced when you proceed with the current settings.

This preview cannot be changed. It is shown to guide you when finding the appropriate settings.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. A new sequence list will be generated for each group. It will be named according to the group, e.g. *2004-08-24_A02* will be the name of one of the groups in the example shown in figure 29.7.

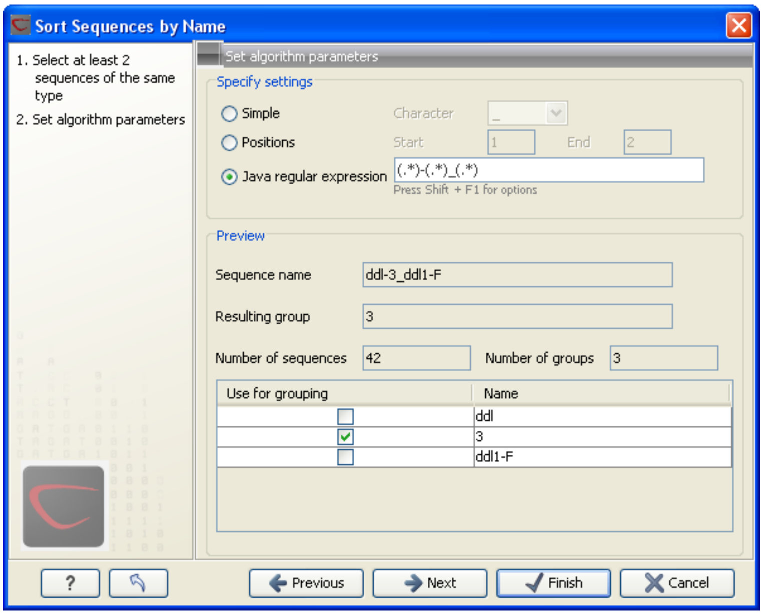
Advanced splitting using regular expressions

In this section you will see a practical example showing how to create a regular expression. Consider a list of files as shown below:

```
...
adk-29_adk1n-F
adk-29_adk2n-R
adk-3_adk1n-F
adk-3_adk2n-R
adk-66_adk1n-F
adk-66_adk2n-R
atp-29_atpA1n-F
atp-29_atpA2n-R
atp-3_atpA1n-F
atp-3_atpA2n-R
atp-66_atpA1n-F
atp-66_atpA2n-R
...
```

In this example, we wish to group the sequences into three groups based on the number after the "-" and before the "_" (i.e. 29, 3 and 66). The simple splitting as shown in figure 29.7 requires the same character before and after the text used for grouping, and since we now have both a "-" and a "_", we need to use the regular expressions instead (note that dividing by position would not work because we have both single and double digit numbers (3, 29 and 66)).

The regular expression for doing this would be `(.*)-(.*)_(.*)` as shown in figure 29.8. The round brackets () denote the part of the name that will be listed in the groups table at the bottom of the dialog. In this example we actually did not need the first and last set of brackets, so the expression could also have been `.*-(.*)_.*` in which case only one group would be listed in the table at the bottom of the dialog.



Sort Sequences by Name

1. Select at least 2 sequences of the same type
2. Set algorithm parameters

Set algorithm parameters

Specify settings

○ Simple — Character: _
○ Positions — Start: 1 — End: 2
● Java regular expression: (.*)-(.*)_(.*)
Press Shift + F1 for options

Preview

Sequence name: ddl-3_ddl1-F

Resulting group: 3

Number of sequences: 42 — Number of groups: 3

Use for grouping	Name
□	ddl
☑	3
□	ddl1-F

Figure 29.8: *Dividing the sequence into three groups based on the number in the middle of the name.*

29.5 Assemble sequences to reference

This section describes how to assemble a number of sequence reads into a contig using a reference sequence. A reference sequence can be particularly helpful when the objective is to characterize SNP variation in the data.

To start the assembly:

Toolbox | **Sequencing Data Analysis** ()| **Assemble Sequences to Reference** ()

This opens a dialog where you can alter your choice of sequences to assemble. If you have already selected sequences in the Navigation Area, these will be shown in Selected Elements, however you can remove these or add others, by using the arrows to move sequences between the Navigation Area and Selected Elements boxes. You can also add sequence lists.

Note! You can assemble a maximum of 2000 sequences at a time.

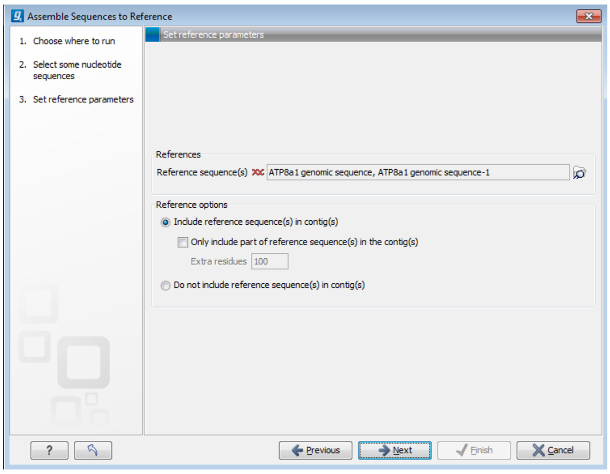
To assemble more sequences, please use the **Map Reads to Reference** () under **NGS Core Tools** () in the **Toolbox**.

To assemble more sequences, you need the *CLC Genomics Workbench* (see http://www.clcbio.com/genomics).

When the sequences are selected, click **Next**, and you will see the dialog shown in figure 29.9

This dialog gives you the following options for assembling:

- **Reference sequence.** Click the **Browse and select element** icon () in order to select one or more sequences to use as reference(s).
- **Include reference sequence(s) in contig(s).** This will create a contig for each reference with the corresponding reference sequence at the top and the aligned sequences below. This



1. Choose where to run
2. Select some nucleotide sequences
3. Set reference parameters

References

Reference sequence(s): ATP8a1 genomic sequence, ATP8a1 genomic sequence-1

Reference options

● Include reference sequence(s) in contig(s)

□ Only include part of reference sequence(s) in the contig(s)

Extra residues: 100

○ Do not include reference sequence(s) in contig(s)

Figure 29.9: *Parameters for how the reference should be handled when assembling sequences to a reference sequence.*

option is useful when comparing sequence reads to a closely related reference sequence e.g. when sequencing for SNP characterization.

  - **Only include part of reference sequence(s) in the contig(s).** If the aligned sequences only cover a small part of a reference sequence, it may not be desirable to include the whole reference sequence in a contig. When this option is selected, you can specify the number of residues from reference sequences that should be included on each side of regions spanned by aligned sequences using the **Extra residues** field.

- **Do not include reference sequence(s) in contig(s).** This will produce contigs without any reference sequence where the input sequences have been assembled using reference sequences as a scaffold. The input sequences are first aligned to the reference sequence(s). Next, the consensus sequence for regions spanned by aligned sequences are extracted and output as contigs. This option is useful when performing assembling sequences where the reference sequences that are not closely related to the input sequencing.

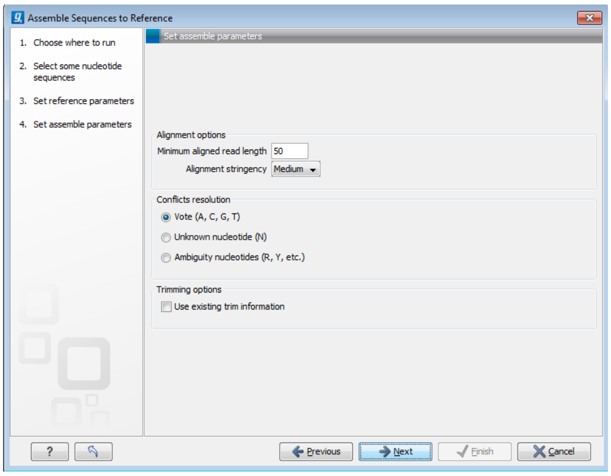
When the reference sequence has been selected, click **Next**, to see the dialog shown in figure 29.10

In this dialog, you can specify the following options:

- **Minimum aligned read length.** The minimum number of nucleotides in a read which must match a reference sequence. If an input sequence does not meet this criteria, the sequence is excluded from the assembly.

- **Alignment stringency.** Specifies the stringency of the scoring function used for aligning the input sequences to the reference sequence(s). A higher stringency level often produce contigs with lower levels of ambiguity but also reduces the ability to align distant homologs or sequences with a high error rate to reference sequences. The result of a higher stringency level is often that the number of contigs increases and the average length of contigs decreases while the quality of each contig increases. Three stringency levels can be set:

  - **Low.**



1. Choose where to run
2. Select some nucleotide sequences
3. Set reference parameters
4. Set assemble parameters

Alignment options

Minimum aligned read length: 50

Alignment stringency: Medium

Conflicts resolution

● Vote (A, C, G, T)

○ Unknown nucleotide (N)

○ Ambiguity nucleotides (R, Y, etc.)

Trimming options

□ Use existing trim information

Figure 29.10: *Options for how the input sequences should be aligned and how nucleotide conflicts should be handled.*

- **Medium.**
- **High.**

The stringency settings Low, Medium and High are based on the following score values (mt=match, ti=transition, tv=transversion, un=unknown):

Score values			
	Low	Medium	High
Match (mt)	2	2	2
Transversion (tv)	-6	-10	-20
Transition (ti)	-2	-6	-16
Unknown (un)	-2	-6	-16
Gap	-8	-16	-36

Score Matrix					
	A	C	G	T	N
A	mt	tv	ti	tv	un
C	tv	mt	tv	ti	un
G	ti	tv	mt	tv	un
T	tv	ti	tv	mt	un
N	un	un	un	un	un

- **Conflicts resolution.** If there is a conflict, i.e. a position where aligned sequences disagreement about the residue (A, C, T or G), you can specify how the contig sequence should reflect this conflict:
  - **Unknown nucleotide (N).** The contig will be assigned an 'N' character in all positions with conflicts (conflicts are registered already when two nucleotides differ).
  - **Ambiguity nucleotides (R, Y, etc.).** The contig will display an ambiguity nucleotide reflecting the different nucleotides found in the aligned sequences (nucleotide ambiguity is registered when two nucleotides differ). For an overview of ambiguity codes, see Appendix D.

  - **Vote (A, C, G, T).** The conflict will be solved by counting instances of each nucleotide and then letting the majority decide the nucleotide in the contig. In case of equality, ACGT are given priority over one another in the stated order.

  Note, that conflicts will be highlighted for all options. Furthermore, conflicts will be marked with an annotation on each contig sequence which are preserved if the contig sequence is extracted for further analysis. As a result, the details of any experimental heterogeneity can be maintained and used when the result of single-sequence analyzes is interpreted.

- **Trimming options.** When aligning sequences to a reference sequence, trimming is generally not necessary, but if you wish to use trimming you can check this box. It requires that the sequence reads have been trimmed beforehand (see section 29.2 for more information about trimming).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. This will start the assembly process. See section 29.7 on how to use the resulting contigs.

29.6 Add sequences to an existing contig

This section describes how to assemble sequences to an existing contig. This feature can be used for example to provide a steady work-flow when a number of exons from the same gene are sequenced one at a time and assembled to a reference sequence.

Note that the new sequences will be added to the existing contig which will not be extended. If the new sequences extend beyond the existing contig, they will be cut off.

To start the assembly:

Toolbox in the Menu Bar | Sequencing Data Analysis ()| Add Sequences to Contig ()

or **right-click in the empty white area of the contig | Add Sequences to Contig ()**

This opens a dialog where you can select one contig and a number of sequences to assemble. If you have already selected sequences in the Navigation Area, these will be shown in the 'Selected Elements' box. However, you can remove these, or add others, by using the arrows to move sequences between the Navigation Area and Selected Elements boxes. You can also add sequence lists.

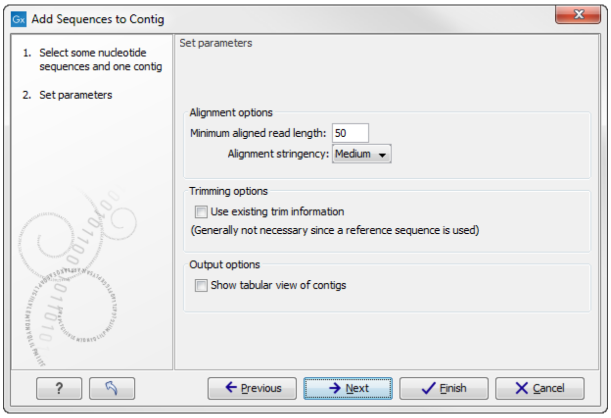
Often, the results of the assembly will be better if the sequences are trimmed first (see section 29.2.1).

When the elements are selected, click **Next**, and you will see the dialog shown in figure 29.11

The options in this dialog are similar to the options that are available when assembling to a reference sequence (see section 29.5).

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. This will start the assembly process. See section 29.7 on how to use the resulting contig.

Note that the new sequences will be added to the existing contig which will not be extended. If the new sequences extend beyond the existing contig, they will be cut off.



Add Sequences to Contig

1. Select some nucleotide sequences and one contig
2. Set parameters

Set parameters

Alignment options

Minimum aligned read length: 50

Alignment stringency: Medium

Trimming options

☐ Use existing trim information

(Generally not necessary since a reference sequence is used)

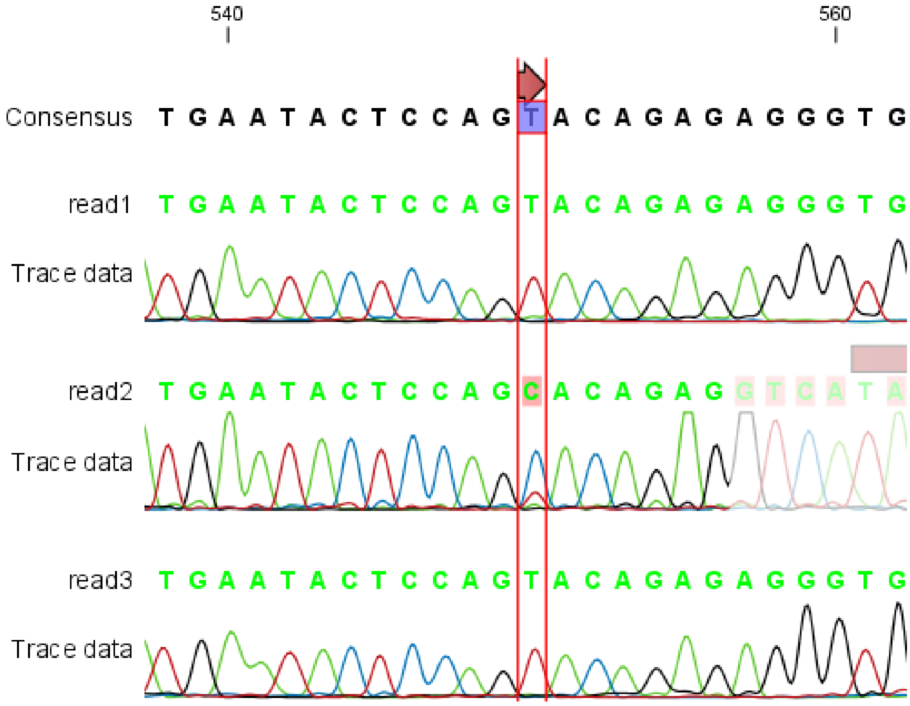
Output options

☐ Show tabular view of contigs

Figure 29.11: *Setting assembly parameters when assembling to an existing contig.*

29.7 View and edit read mappings

The result of the mapping process is one or more read mappings where the sequence reads have been aligned (see figure 29.12). If multiple reference sequences were used, this information will be in a table where the actual visual mapping can be opened by double-clicking. The result of the assembly process is one or more contigs where the sequence reads have been aligned (see figure 29.12).



540
560
Consensus T G A A T A C T C C A G T A C A G A G A G G G T G
read1 T G A A T A C T C C A G T A C A G A G A G G G T G
Trace data
read2 T G A A T A C T C C A G C A C A G A G G T C A T A
Trace data
read3 T G A A T A C T C C A G T A C A G A G A G G G T G
Trace data

Figure 29.12: *The view of a read mapping. Notice that you can zoom to a very detailed level in read mappings.*

You can see that color of the residues and trace at the end of one of the reads has been faded. This indicates, that this region has not contributed to the mapping. This may be due to trimming before or during the assembly or due to misalignment to the other reads.

You can easily adjust the trimmed area to include more of the read in the mapping: simply drag the edge of the faded area as shown in figure 29.13.

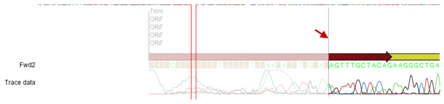


Figure 29.13: *Dragging the edge of the faded area.*

Note! This is only possible when you can see the residues on the reads. This means that you need to have zoomed in to 100% or more and chosen **Compactness** levels "Not compact", "Low" or "Packed". Otherwise the handles for dragging are not available (this is done in order to make the visual overview more simple).

If reads have been reversed, this is indicated by red. Otherwise, the residues are colored green. The colors can be changed in the **Side Panel** as described in section 29.7.1

If you find out that the reversed reads should have been the forward reads and vice versa, you can reverse complement the whole mapping (imagine flipping the whole mapping):

right-click in the empty white area of the mapping | **Reverse Complement**

29.7.1 View settings in the Side Panel

Apart from this the view resembles that of alignments but has some extra preferences in the **Side Panel**:

[1]

- **Read layout.** This section appears at the top of the **Side Panel** when viewing a stand alone read mapping:
  - **Compactness.** The compactness setting options let you control the level of detail to be displayed. This setting affects many of the other settings in the **Side Panel** as well as the general behavior of the view. For example: if the compactness is set to **Compact**, you will not be able to see quality scores or annotations on the reads, even if these are turned on via the Nucleotide info section of the Side Panel. You can change the Compactness setting in the Side Panel directly, or you can use the shortcut: press and hold the Alt key while you scroll with the mouse wheel or touchpad.
    - **Not compact.** This allows the mapping to be viewed full detail, including quality scores and trace data for the reads, where this is relevant. To view such information, additional viewing options under the **Nucleotide info** view settings must also selected. For further details on these, please see section 29.1.2 and section 9.1.
    - **Low.** Hides trace data, quality scores and puts the reads' annotations on the sequence.
    - **Medium.** The labels of the reads and their annotations are hidden, and the residues of the reads cannot be seen.
    - **Compact.** Even less space between the reads.
    - **Packed.** All the other compactness settings will stack the reads on top of each other, but the packed setting will use all space available for displaying the reads. When zoomed in to 100%, you can see the residues but when zoomed

[1] Note that for interpretation of mappings with large amounts of data, have a look at section 20.6

out the reads will be represented as lines just as with the Compact setting. The packed mode is very useful when viewing large amounts of data. However certain functionality possible with other views are not available in packed view. For example, no editing of the read mapping or selections of it can be done and color coding changes are not possible. An example of the packed setting is shown in figure 29.14.

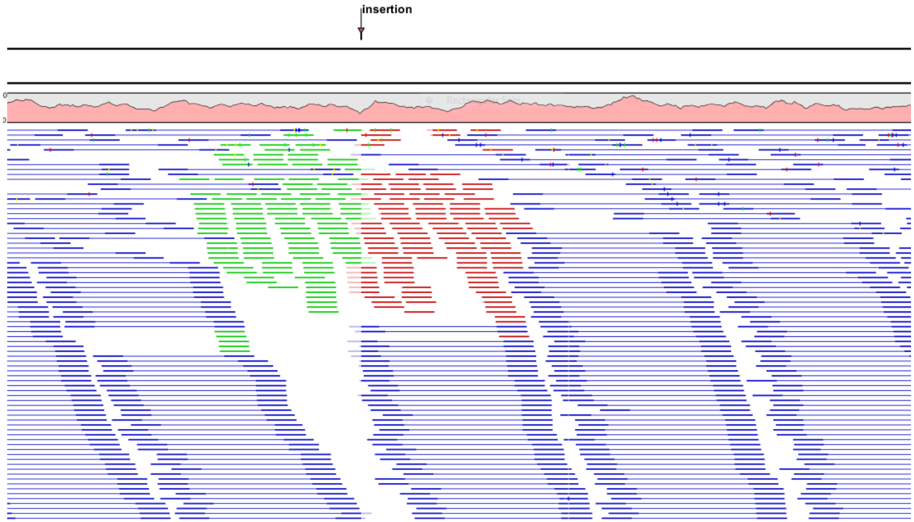


Figure 29.14: *An example of the packed compactness setting.*

- **Gather sequences at top.** Enabling this option affects the view that is shown when scrolling horizontally. If selected, the sequence reads which did not contribute to the visible part of the mapping will be omitted whereas the contributing sequence reads will automatically be placed right below the reference. This setting is not relevant when the compactness is packed.
- **Show sequence ends.** Regions that have been trimmed are shown with faded traces and residues. This illustrates that these regions have been ignored during the assembly.
- **Show mismatches.** When the compactness is packed, you can highlight mismatches which will get a color according to the Rasmol color scheme. A mismatch is whenever the base is different from the reference sequence at this position. This setting also causes the reads that have mismatches to be floated at the top of the view.
- **Disconnect paired reads.** This option will break up the paired reads in the display (they are still marked as pairs - this just affects the visualization). The reads are marked with colors for the direction (default red and green) instead of the color for pairs (default blue). This is particularly useful when investigating overlapping pairs in packed view and when the strand / read orientation is important.
- **Packed read height.** When the compactness is set to "packed", you can choose the height of the visible reads. When there are more reads than the height specified, an overflow graph will be displayed below the reads. The overflow graph is shown in the same colors as the sequences, and mismatches in reads are shown as narrow horizontal lines in. The colors of the small lines represent the mismatching residue. The color codes for the horizontal lines correspond to the color used for highlighting mismatches in the sequences (red = A, blue = C, yellow = G, and green = T). E.g.

a red line with half the height of the blue part of the overflow graph will represent a mismatching "A" in half of the paired reads at this particular position.

- **Find Conflict.** Clicking this button selects the next position where there is an conflict between the sequence reads. Residues that are different from the reference are colored (as default), providing an overview of the conflicts. Since the next conflict is automatically selected it is easy to make changes. You can also use the Space key to find the next conflict.
- **Low coverage threshold.** All regions with coverage up to and including this value are considered low coverage. When clicking the 'Find low coverage' button the next region in the read mapping with low coverage will be selected.

- **Alignment info.** There is one additional parameter:
  - **Coverage**: Shows how many sequence reads that are contributing information to a given position in the mapping. The level of coverage is relative to the overall number of sequence reads.
    - **Foreground color.** Colors the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage.
    - **Background color.** Colors the background of the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage
    - **Graph.** The coverage is displayed as a graph (Learn how to export the data behind the graph in section 6.7).
      - **Height.** Specifies the height of the graph.
      - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
      - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. If a Color bar is chosen, the color box is replaced by a gradient color box as described under Foreground color.
- **Residue coloring.** There is one additional parameter:
  - **Sequence colors.** This option lets you use different colors for the reads.
    - **Main**. The color of the consensus and reference sequence. Black per default.
    - **Forward**. The color of forward reads (single reads). Green per default.
    - **Reverse**. The color of reverse reads (single reads). Red per default.
    - **Paired**. The color of paired reads. Blue per default. Note that reads from **broken pairs** are colored according to their Forward/Reverse orientation or as a Non-specific match, but with a darker nuance than ordinary single reads.
    - **Non-specific matches**. When a read would have matched equally well another place in the mapping, it is considered a non-specific match. This color will "overrule" the other colors. Note that if you are mapping with several reference sequences, a read is considered a double match when it matches more than once *across all the contigs/references*. A non-specific match is yellow per default.
- **Sequence layout.** At the top of the **Side Panel**:
  - **Matching residues as dots** Matching residues will be presented as dots. Only the top sequence will be preserved in its original format.

There are many other viewing options available, both general and aimed as specifice elements of a mapping, which can be adjusted in the View settings. Those covered here were the key ones relevant standard review of mapping results.

29.7.2 Editing the read mapping

When editing mappings, you are typically interested in confirming or changing single bases, and this can be done simply by:

selecting the base | typing the right base

Some users prefer to use lower-case letters in order to be able to see which bases were altered when they use the results later on. In *CLC Cancer Research Workbench* all changes are recorded in the history log (see section 7) allowing the user to quickly reconstruct the actions performed in the editing session.

There are three shortcut keys for easily finding the positions where there are conflicts:

- Space bar: Finds the *next* conflict.
- "." (punctuation mark key): Finds the *next* conflict.
- "," (comma key): Finds the *previous* conflict.

In the mapping view, you can use **Zoom in** (🔍) to zoom to a greater level of detail than in other views (see figure 29.12). This is useful for discerning the trace curves.

If you want to replace a residue with a gap, use the **Delete** key.

If you wish to edit a selection of more than one residue:

right-click the selection | Edit Selection (✎)

This will show a warning dialog, but you can choose never to see this dialog again by clicking the checkbox at the bottom of the dialog.

Note that for mappings with more than 1000 reads, you can only do single-residue replacements (you can't delete or edit a selection). When the compactness is **Packed**, you cannot edit any of the reads.

29.7.3 Sorting reads

If you wish to change the order of the sequence reads, simply drag the label of the sequence up and down. Note that this is not possible if you have chosen **Gather sequences at top** or set the compactness to **Packed** in the **Side Panel**.

You can also sort the reads by right-clicking a sequence label and choose from the following options:

- **Sort Reads by Alignment Start Position.** This will list the first read in the alignment at the top etc.
- **Sort Reads by Name.** Sort the reads alphabetically.
- **Sort Reads by Length.** The shortest reads will be listed at the top.

29.7.4 Read conflicts

When the mapping is created, conflicts between the reads are annotated on the consensus sequence. The definition of a conflict is *a position where at least one of the reads have a different residue*.

A conflict can be in two states:

- **Conflict**. Both the annotation and the corresponding row in the Table (▦) are colored **red.**
- **Resolved**. Both the annotation and the corresponding row in the Table (▦) are colored **green**.

The conflict can be resolved by correcting the deviating residues in the reads as described above.

A fast way of making all the reads reflect the consensus sequence is to select the position in the consensus, right-click the selection, and choose **Transfer Selection to All Reads**.

The opposite is also possible: make a selection on one of the reads, right click, and **Transfer Selection to Contig Sequence**.

29.7.5 Extract parts of a mapping

Sometimes it is useful to extract part of a mapping for in-depth analysis. This could be the case if you have performed an analysis of a whole genome data set and have found a region that you are particularly interested in analyzing further. Rather than running all further analysis on your full data, you may prefer to run only on a subset of the data. You can extract a subset of your mapping data by running the **Extract from Selection** tool on a selected region in your mapping. The result of running this tool is a new mapping which contains only the reads (and optionally only those that are of a particular type) in your selected region.

To select a region, use the **Selection mode** ( ) (see Section 2.2.3 for a detailed description of the different modes) and select you region of interest in your mapping, then right-click. You are now presented with the dialog shown in Figure 29.15.

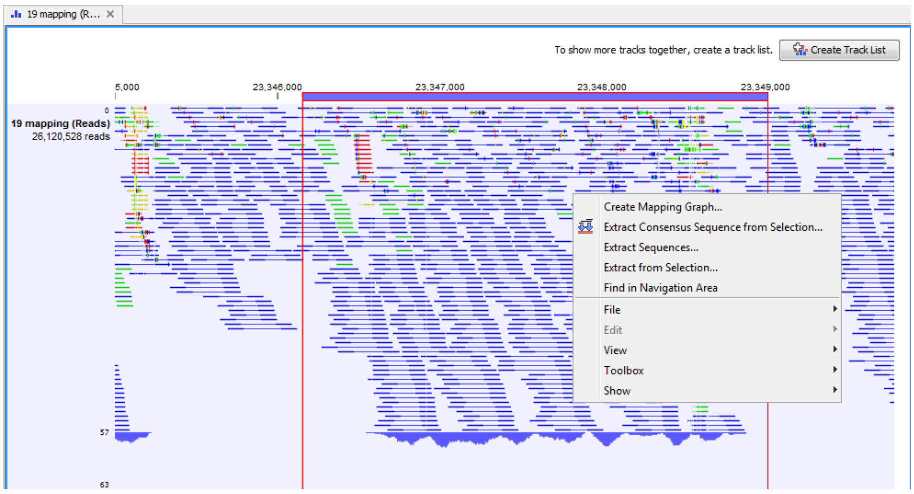
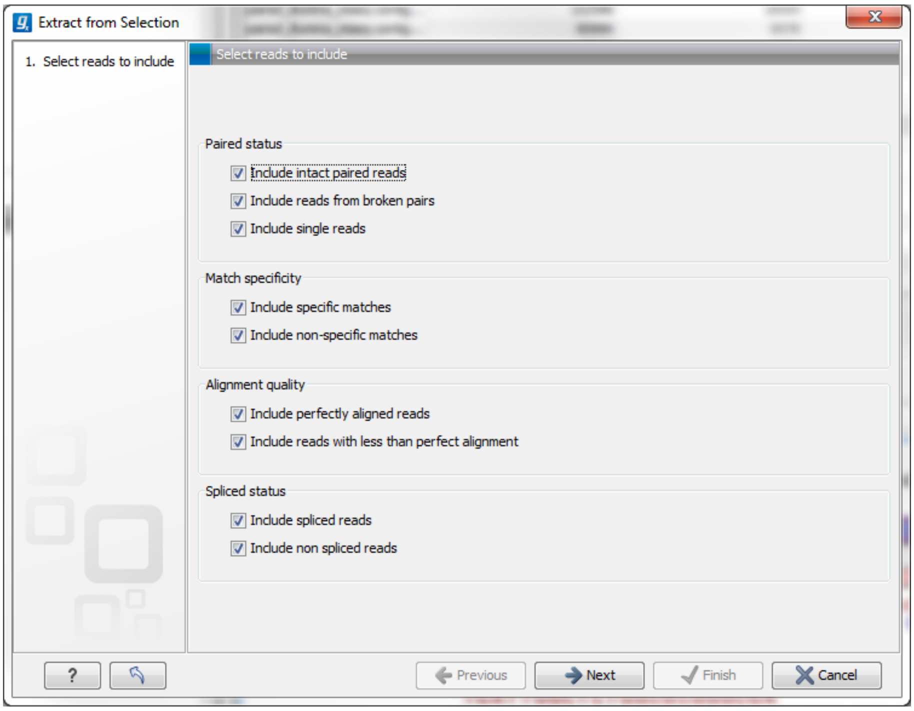


Figure 29.15: *Extracting parts of a mapping.*

When you choose the **Extract from Selection** option you are presented by the dialog shown in figure 29.16.



Paired status

☑ Include intact paired reads

☑ Include reads from broken pairs

☑ Include single reads

Match specificity

☑ Include specific matches

☑ Include non-specific matches

Alignment quality

☑ Include perfectly aligned reads

☑ Include reads with less than perfect alignment

Spliced status

☑ Include spliced reads

☑ Include non spliced reads

Figure 29.16: *Selecting the reads to include.*

The purpose of this dialog is to let you specify what kind of reads you want to include. Per default all reads are included. The options are:

Paired status **Include intact paired reads** When paired reads are placed within the paired distance specified, they will fall into this category. Per default, these reads are colored in blue.

Include paired reads from broken pairs When a pair is broken, either because only one read in the pair matches, or because the distance or relative orientation is wrong, the reads are placed and colored as single reads, but you can still extract them by checking this box.

Include single reads This will include reads that are marked as single reads (as opposed to paired reads). Note that paired reads that have been broken during assembly are not included in this category. Single reads that come from trimming paired sequence lists are included in this category.

Match specificity **Include specific matches** Reads that only are mapped to one position.

Include non-specific matches Reads that have multiple equally good alignments to the reference. These reads are colored yellow per default.

Alignment quality **Include perfectly aligned reads** Reads where *the full read* is perfectly aligned to the reference sequence (or consensus sequence for de novo assemblies). Note that at the end of the contig, reads may extend beyond the contig (this is not visible unless you make a selection on the read and observe the position numbering in the status bar). Such reads are not considered perfectly aligned reads because they don't align in their entire length.

Include reads with less than perfect alignment Reads with mismatches, insertions or deletions, or with unaligned nucleotides at the ends (the faded part of a read).

Spliced status **Include spliced reads** Reads that are across an intron.

Include non spliced reads Reads that are not across an intron.

Note that only reads that are completely covered by the selection will be part of the new contig.

One of the benefits of this is that you can actually use this tool to extract subset of reads from a contig. An example work flow could look like this:

1. Select the whole reference sequence
2. Right-click and **Extract from Selection**
3. Choose to include only paired matches
4. Extract the reads from the new file (see section 28.1)

You will now have all paired reads from the original mapping in a list.

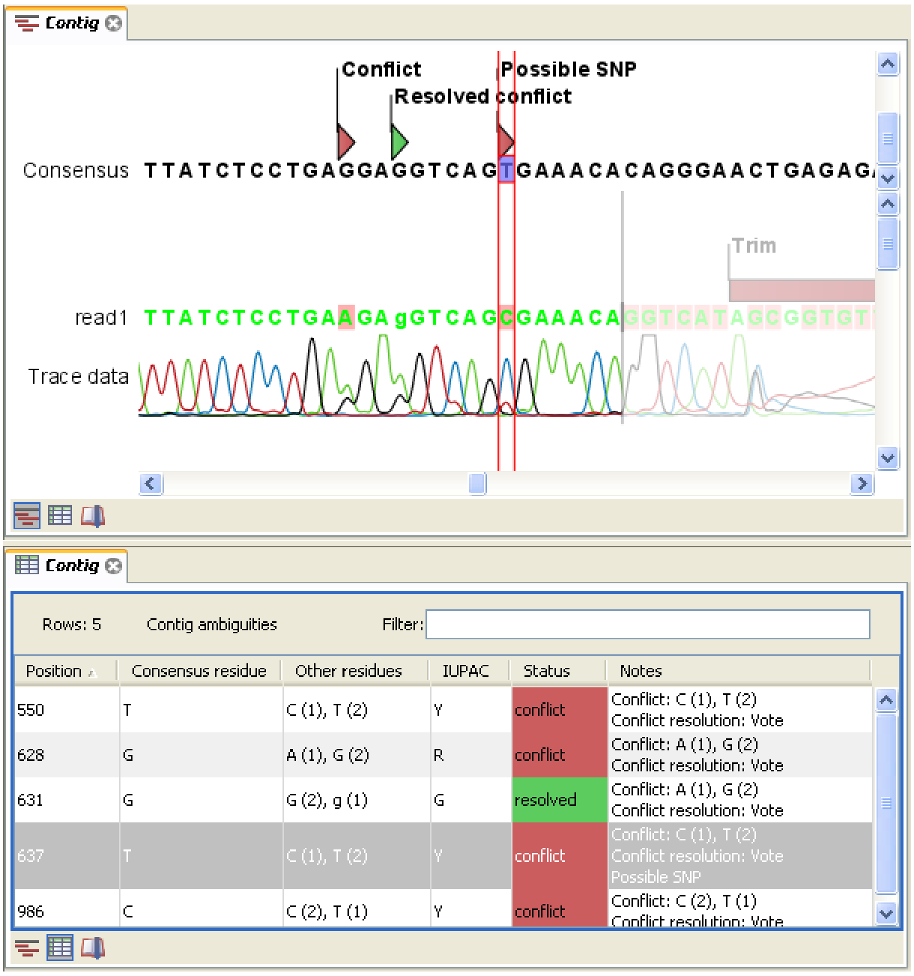
29.7.6 Variance table

In addition to the standard graphical display of a mapping as described above, you can also see a tabular overview of the conflicts between the reads by clicking the **Table (▦)** icon at the bottom of the view.

This will display a new view of the conflicts as shown in figure 29.17.

The table has the following columns:

- **Reference position.** The position of the conflict measured from the starting point of the reference sequence.
- **Consensus position.** The position of the conflict measured from the starting point of the consensus sequence.
- **Consensus residue.** The consensus's residue at this position. The residue can be edited in the graphical view, as described above.
- **Other residues.** Lists the residues of the reads. Inside the brackets, you can see the number of reads having this residue at this position. In the example in figure 29.17, you can see that at position 637 there is a 'C' in the top read in the graphical view. The other two reads have a 'T'. Therefore, the table displays the following text: 'C (1), T (2)'.
- **IUPAC.** The ambiguity code for this position. The ambiguity code reflects the residues in the reads - not in the consensus sequence. (The IUPAC codes can be found in section D.)
- **Status.** The status can either be conflict or resolved:
  - **Conflict.** Initially, all the rows in the table have this status. This means that there is one or more differences between the sequences at this position.
  - **Resolved.** If you edit the sequences, e.g. if there was an error in one of the sequences, and they now all have the same residue at this position, the status is set to *Resolved*.


Contig
Conflict
Possible SNP
Resolved conflict
Consensus TTATCTCCTGAGGAGGTCAGTGAAACACAGGGAACTGAGAG
Trim
read1 TTATCTCCTGAAGAgGTCAGCGAAACAGGTCATAGCGGTGT
Trace data
Contig

Rows: 5 Contig ambiguities Filter:

Position	Consensus residue	Other residues	IUPAC	Status	Notes
550	T	C (1), T (2)	Y	conflict	Conflict: C (1), T (2) Conflict resolution: Vote
628	G	A (1), G (2)	R	conflict	Conflict: A (1), G (2) Conflict resolution: Vote
631	G	G (2), g (1)	G	resolved	Conflict: A (1), G (2) Conflict resolution: Vote
637	T	C (1), T (2)	Y	conflict	Conflict: C (1), T (2) Conflict resolution: Vote Possible SNP
986	C	C (2), T (1)	Y	conflict	Conflict: C (2), T (1) Conflict resolution: Vote

Figure 29.17: *The graphical view is displayed at the top. At the bottom the conflicts are shown in a table. At the conflict at position 637, the user has entered a comment in the table. This comment is now also reflected on the tooltip of the conflict annotation in the graphical view above.*

- **Note.** Can be used for your own comments on this conflict. Right-click in this cell of the table to add or edit the comments. The comments in the table are associated with the conflict annotation in the graphical view. Therefore, the comments you enter in the table will also be attached to the annotation on the consensus sequence (the comments can be displayed by placing the mouse cursor on the annotation for one second - see figure 29.17). The comments are saved when you **Save** ( ).

By clicking a row in the table, the corresponding position is highlighted in the graphical view. Clicking the rows of the table is another way of navigating the mapping, apart from using the **Find Conflict** button or using the **Space bar**. You can use the up and down arrow keys to navigate the rows of the table.

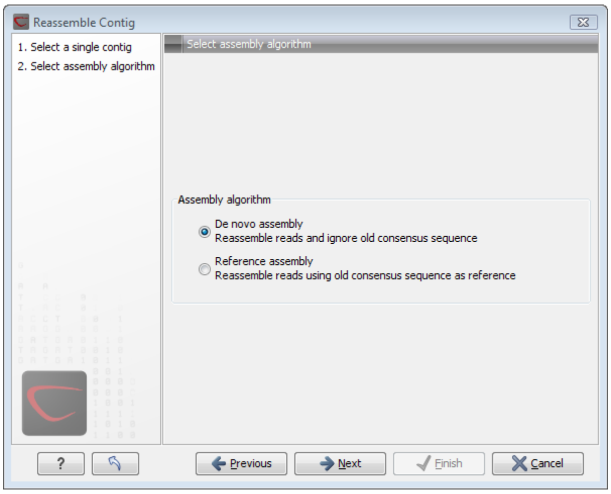
29.8 Reassemble contig

If you have edited a contig, changed trimmed regions, or added or removed reads, you may wish to reassemble the contig. This can be done in two ways:

Toolbox | **Sequencing Data Analysis** ()| **Reassemble Contig** () | **select the contig from Navigation Area, move to 'Selected Elements' and click Next**

or **right-click in the empty white area of the contig** | **Reassemble contig** ()

This opens a dialog as shown in figure 29.18



Reassemble Contig

1. Select a single contig
2. Select assembly algorithm

Select assembly algorithm

Assembly algorithm

● De novo assembly
Reassemble reads and ignore old consensus sequence

○ Reference assembly
Reassemble reads using old consensus sequence as reference

Figure 29.18: *Re-assembling a contig.*

In this dialog, you can choose:

- **De novo assembly**. This will perform a normal assembly in the same way as if you had selected the reads as individual sequences. When you click **Next**, you will follow the same steps as described in section 29.3. The consensus sequence of the contig will be ignored.
- **Reference assembly**. This will use the consensus sequence of the contig as reference. When you click **Next**, you will follow the same steps as described in section 29.5.

When you click **Finish**, a new contig is created, so you do not lose the information in the old contig.

29.9 Secondary peak calling

CLC Cancer Research Workbench is able to detect secondary peaks - a peak within a peak - to help discover heterozygous mutations. Looking at the height of the peak below the top peak, the *CLC Cancer Research Workbench* considers all positions in a sequence, and if a peak is higher than the threshold set by the user, it will be "called".

The peak detection investigates any secondary high peaks in the same interval as the already called peaks. The peaks must have a peak shape in order to be considered (i.e. a fading signal from the previous peak will be ignored).

Regions that are trimmed (i.e. covered by trim annotations) are ignored in the analysis (section 29.2).

When a secondary peak is called, the residue is change to an ambiguity character to reflect that two bases are possible at this position, and optionally an annotation is added at this position.

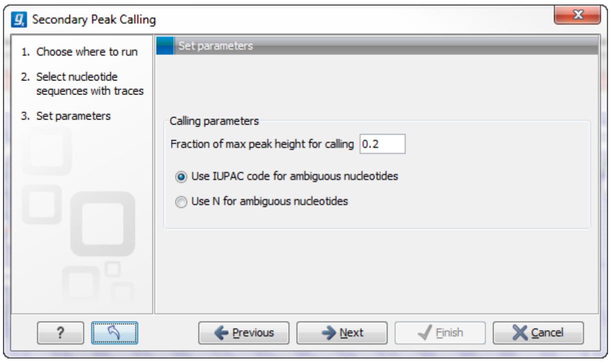
To call secondary peaks:

Toolbox | **Sequencing Data Analysis** (**)**| **Call Secondary Peaks** (**)**

This opens a dialog where you can add the sequences to be analysed. If you had already selected sequence in the Navigation Area, these will be shown in the 'Selected Elements' box. However you can remove these, or add others, by using the arrows to move sequences between the Navigation Area and Selected Elements boxes.

When the sequences are selected, click **Next**.

This opens the dialog displayed in figure 29.19.


Secondary Peak Calling

1. Choose where to run
2. Select nucleotide sequences with traces
3. Set parameters

Set parameters

Calling parameters

Fraction of max peak height for calling 0.2

● Use IUPAC code for ambiguous nucleotides

○ Use N for ambiguous nucleotides

Figure 29.19: *Setting parameters secondary peak calling.*

The following parameters can be adjusted in the dialog:

- **Fraction of max peak height for calling.** Adjust this value to specify how high the secondary peak must be to be called.
- **Use IUPAC code / N for ambiguous nucleotides.** When a secondary peak is called, the residue at this position can either be replaced by an N or by a ambiguity character based on the IUPAC codes (see section D).

Clicking **Next** allows you to add annotations. In addition to changing the actual sequence, annotations can be added for each base that has been called. The annotations hold information about the fraction of the max peak height.

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. This will start the secondary peak calling. A detailed history entry will be added to the history specifying all the changes made to the sequence.

Chapter 30

Primers

Contents

CLC Cancer Research Workbench offers graphically and algorithmically advanced design of primers and probes for various purposes. This chapter begins with a brief introduction to the general concepts of the primer designing process. Then follows instructions on how to adjust parameters for primers, how to inspect and interpret primer properties graphically and how to interpret, save and analyze the output of the primer design analysis. After a description of the different reaction types for which primers can be designed, the chapter closes with sections on how to match primers with other sequences and how to create a primer order.

30.1 Primer design - an introduction

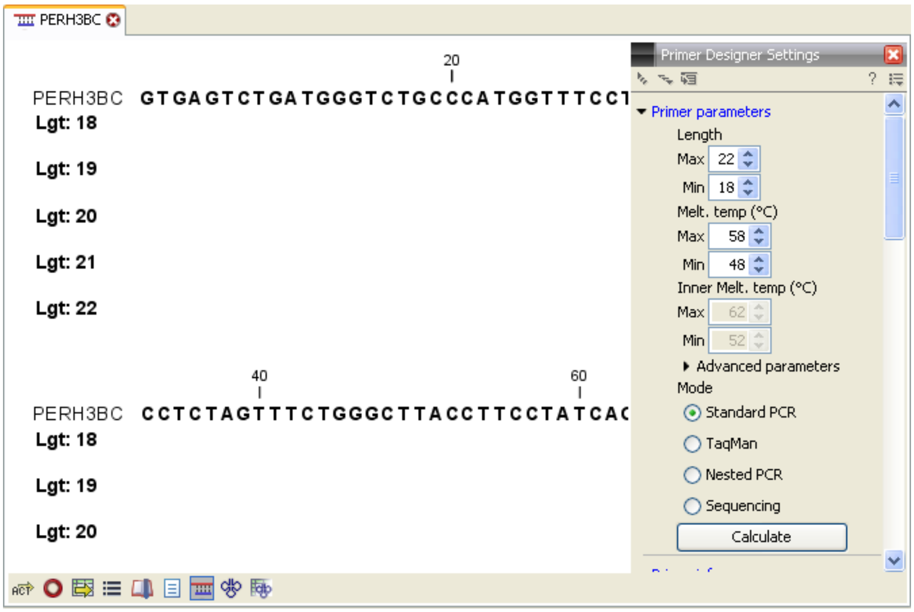
Primer design can be accessed in two ways:

Toolbox | **Primers and Probes** ()| **Design Primers** () | **OK**

or **right-click sequence in Navigation Area** | **Show** | **Primer** ()

In the primer view (see figure 30.1), the basic options for viewing the template sequence are the same as for the standard sequence view. See section 9.1 for an explanation of these options.

Note! This means that annotations such as e.g. known SNPs or exons, can be displayed on the template sequence to guide the choice of primer regions. Also, traces in sequencing reads can be shown along with the structure to guide e.g. the re-sequencing of poorly resolved regions.



PERH3BC GTGAGTCTGATGGGTCTGCCCATGGTTTCCT
Lgt: 18
Lgt: 19
Lgt: 20
Lgt: 21
Lgt: 22

PERH3BC CCTCTAGTTTCTGGGCTTACCTTCCTATCAC
Lgt: 18
Lgt: 19
Lgt: 20

Primer parameters
Length
Max 22
Min 18
Melt. temp (°C)
Max 58
Min 48
Inner Melt. temp (°C)
Max 62
Min 52
Advanced parameters
Mode
● Standard PCR
○ TaqMan
○ Nested PCR
○ Sequencing
Calculate

Figure 30.1: *The initial view of the sequence used for primer design.*

30.1.1 General concept

The concept of the primer view is that the user first chooses the desired reaction type for the session in the Primer Parameters preference group, e.g. *Standard PCR*. Reflecting the choice of reaction type, it is now possibly to select one or more regions on the sequence and to use the right-click mouse menu to designate these as primer or probe regions (see figure 30.2).

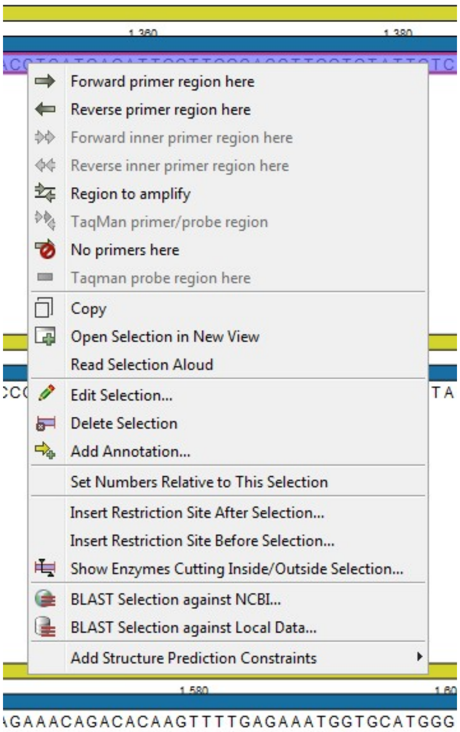

Forward primer region here
Reverse primer region here
Forward inner primer region here
Reverse inner primer region here
Region to amplify
TaqMan primer/probe region
No primers here
Taqman probe region here
Copy
Open Selection in New View
Read Selection Aloud
Edit Selection...
Delete Selection
Add Annotation...
Set Numbers Relative to This Selection
Insert Restriction Site After Selection...
Insert Restriction Site Before Selection...
Show Enzymes Cutting Inside/Outside Selection...
BLAST Selection against NCBI...
BLAST Selection against Local Data...
Add Structure Prediction Constraints
GAAACAGACACAAGTTTTGAGAAATGGTGCATGGG

Figure 30.2: *Right-click menu allowing you to specify regions for the primer design*

When a region is chosen, graphical information about the properties of all possible primers in this region will appear in lines beneath it. By default, information is showed using a compact mode but the user can change to a more detailed mode in the Primer information preference group.

The number of information lines reflects the chosen length interval for primers and probes. In the compact information mode one line is shown for every possible primer-length and each of these lines contain information regarding all possible primers of the given length. At each potential primer starting position, a circular information point is shown which indicates whether the primer fulfills the requirements set in the primer parameters preference group. A green circle indicates a primer which fulfils all criteria and a red circle indicates a primer which fails to meet one or more of the set criteria. For more detailed information, place the mouse cursor over the circle representing the primer of interest. A tool-tip will then appear on screen, displaying detailed information about the primer in relation to the set criteria. To locate the primer on the sequence, simply left-click the circle using the mouse.

The various primer parameters can now be varied to explore their effect and the view area will dynamically update to reflect this allowing for a high degree of interactivity in the primer design process.

After having explored the potential primers the user may have found a satisfactory primer and choose to export this directly from the view area using a mouse right-click on the primers information point. This does not allow for any design information to enter concerning the properties of primer/probe pairs or sets e.g. primer pair annealing and T_m difference between primers. If the latter is desired the user can use the **Calculate** button at the bottom of the Primer parameter preference group. This will activate a dialog, the contents of which depends on the chosen mode. Here, the user can set primer-pair specific setting such as allowed or desired T_m

difference and view the single-primer parameters which were chosen in the Primer parameters preference group.

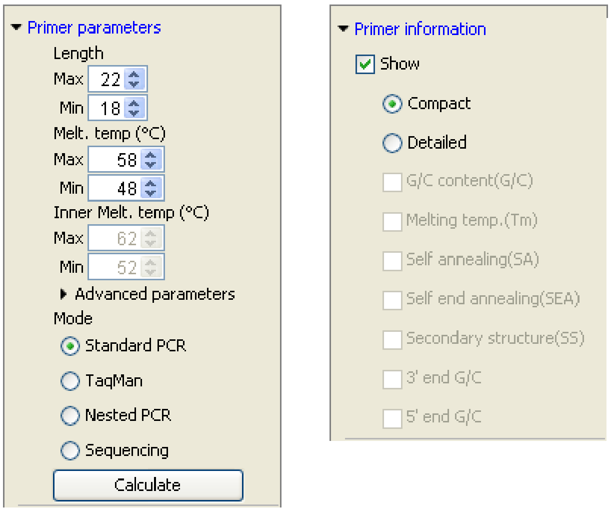
Upon pressing finish, an algorithm will generate all possible primer sets and rank these based on their characteristics and the chosen parameters. A list will appear displaying the 100 most high scoring sets and information pertaining to these. The search result can be saved to the navigator. From the result table, suggested primers or primer/probe sets can be explored since clicking an entry in the table will highlight the associated primers and probes on the sequence. It is also possible to save individual primers or sets from the table through the mouse right-click menu. For a given primer pair, the amplified PCR fragment can also be opened or saved using the mouse right-click menu.

30.1.2 Scoring primers

CLC Cancer Research Workbench employs a proprietary algorithm to rank primer and probe solutions. The algorithm considers both the parameters pertaining to single oligos, such as e.g. the secondary structure score and parameters pertaining to oligo-pairs such as e.g. the oligo pair-annealing score. The ideal score for a solution is 100 and solutions are thus ranked in descending order. Each parameter is assigned an ideal value and a tolerance. Consider for example oligo self-annealing, here the ideal value of the annealing score is 0 and the tolerance corresponds to the maximum value specified in the side panel. The contribution to the final score is determined by how much the parameter deviates from the ideal value and is scaled by the specified tolerance. Hence, a large deviation from the ideal and a small tolerance will give a large deduction in the final score and a small deviation from the ideal and a high tolerance will give a small deduction in the final score.

30.2 Setting parameters for primers and probes

The primer-specific view options and settings are found in the **Primer parameters** preference group in the **Side Panel** to the right of the view (see figure 30.3).



Primer parameters

Length
Max 22
Min 18
Melt. temp (°C)
Max 58
Min 48
Inner Melt. temp (°C)
Max 62
Min 52

Advanced parameters

Mode
● Standard PCR
○ TaqMan
○ Nested PCR
○ Sequencing

Primer information

☑ Show
● Compact
○ Detailed
□ G/C content(G/C)
□ Melting temp.(Tm)
□ Self annealing(SA)
□ Self end annealing(SEA)
□ Secondary structure(SS)
□ 3' end G/C
□ 5' end G/C

Figure 30.3: *The two groups of primer parameters (in the program, the Primer information group is listed below the other group).*

30.2.1 Primer Parameters

In this preference group a number of criteria can be set, which the selected primers must meet. All the criteria concern *single primers*, as primer pairs are not generated until the **Calculate** button is pressed. Parameters regarding primer and probe sets are described in detail for each reaction mode (see below).

- **Length.** Determines the length interval within which primers can be designed by setting a maximum and a minimum length. The upper and lower lengths allowed by the program are 50 and 10 nucleotides respectively.
- **Melting temperature.** Determines the temperature interval within which primers must lie. When the *Nested PCR* or *TaqMan* reaction type is chosen, the first pair of melting temperature interval settings relate to the outer primer pair i.e. not the probe. Melting temperatures are calculated by a nearest-neighbor model which considers stacking interactions between neighboring bases in the primer-template complex. The model uses state-of-the-art thermodynamic parameters [SantaLucia, 1998] and considers the important contribution from the dangling ends that are present when a short primer anneals to a template sequence [Bommarito et al., 2000]. A number of parameters can be adjusted concerning the reaction mixture and which influence melting temperatures (see below). Melting temperatures are corrected for the presence of monovalent cations using the model of [SantaLucia, 1998] and temperatures are further corrected for the presence of magnesium, deoxynucleotide triphosphates (dNTP) and dimethyl sulfoxide (DMSO) using the model of [von Ahsen et al., 2001].
- **Inner melting temperature.** This option is only activated when the *Nested PCR* or *TaqMan* mode is selected. In *Nested PCR* mode, it determines the allowed melting temperature interval for the inner/nested pair of primers, and in *TaqMan* mode it determines the allowed temperature interval for the TaqMan probe.
- **Advanced parameters.** A number of less commonly used options
  - **Buffer properties.** A number of parameters concerning the reaction mixture which influence melting temperatures.
    - **Primer concentration.** Specifies the concentration of primers and probes in units of nanomoles (nM)
    - **Salt concentration.** Specifies the concentration of monovalent cations ($[NA^+]$, $[K^+]$ and equivalents) in units of millimoles (mM)
    - **Magnesium concentration.** Specifies the concentration of magnesium cations ($[Mg^{++}]$) in units of millimoles (mM)
    - **dNTP concentration.** Specifies the combined concentration of all deoxynucleotide triphosphates in units of millimoles (mM)
    - **DMSO concentration.** Specifies the concentration of dimethyl sulfoxide in units of volume percent ($vol.\%$)
  - **GC content.** Determines the interval of CG content (% C and G nucleotides in the primer) within which primers must lie by setting a maximum and a minimum GC content.
  - **Self annealing.** Determines the maximum self annealing value of all primers and probes. This determines the amount of base-pairing allowed between two copies of

the same molecule. The self annealing score is measured in number of hydrogen bonds between two copies of primer molecules, with A-T base pairs contributing 2 hydrogen bonds and G-C base pairs contributing 3 hydrogen bonds.

- **Self end annealing.** Determines the maximum self end annealing value of all primers and probes. This determines the number of consecutive base pairs allowed between the 3' end of one primer and another copy of that primer. This score is calculated in number of hydrogen bonds (the example below has a score of 4 - derived from 2 A-T base pairs each with 2 hydrogen bonds).

```
AATTCCCTACAATCCCCAAA
                  ||
  AAACCCCTAACATCCCTTAA
```

.

- **Secondary structure.** Determines the maximum score of the optimal secondary DNA structure found for a primer or probe. Secondary structures are scored by the number of hydrogen bonds in the structure, and 2 extra hydrogen bonds are added for each stacking base-pair in the structure.

- **3' end G/C restrictions.** When this checkbox is selected it is possible to specify restrictions concerning the number of G and C molecules in the 3' end of primers and probes. A low G/C content of the primer/probe 3' end increases the specificity of the reaction. A high G/C content facilitates a tight binding of the oligo to the template but also increases the possibility of mispriming. Unfolding the preference groups yields the following options:
  - **End length.** The number of consecutive terminal nucleotides for which to consider the C/G content
  - **Max no. of G/C.** The maximum number of G and C nucleotides allowed within the specified length interval
  - **Min no. of G/C.** The minimum number of G and C nucleotides required within the specified length interval

- **5' end G/C restrictions.** When this checkbox is selected it is possible to specify restrictions concerning the number of G and C molecules in the 5' end of primers and probes. A high G/C content facilitates a tight binding of the oligo to the template but also increases the possibility of mis-priming. Unfolding the preference groups yields the same options as described above for the 3' end.

- **Mode.** Specifies the reaction type for which primers are designed:
  - **Standard PCR.** Used when the objective is to design primers, or primer pairs, for PCR amplification of a single DNA fragment.
  - **Nested PCR.** Used when the objective is to design two primer pairs for nested PCR amplification of a single DNA fragment.
  - **Sequencing.** Used when the objective is to design primers for DNA sequencing.
  - **TaqMan.** Used when the objective is to design a primer pair and a probe for TaqMan quantitative PCR.

  Each mode is described further below.

- **Calculate.** Pushing this button will activate the algorithm for designing primers

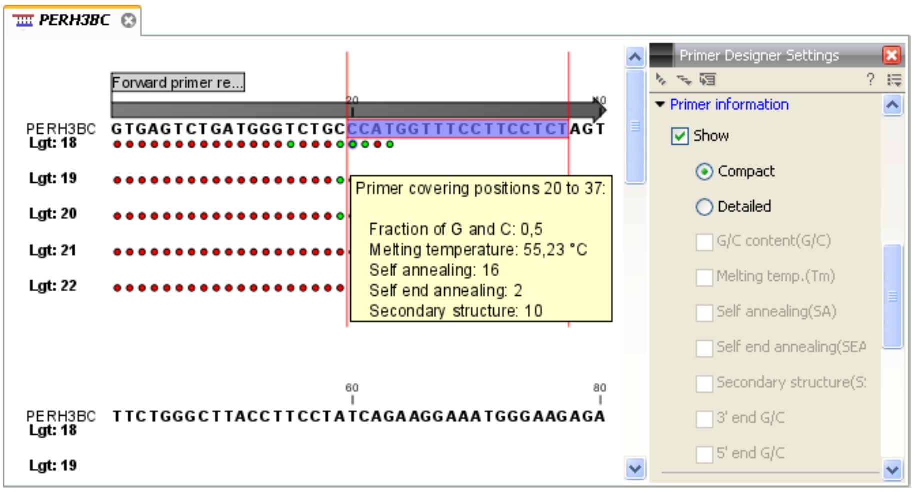
30.3 Graphical display of primer information

The primer information settings are found in the **Primer information** preference group in the **Side Panel** to the right of the view (see figure 30.3).

There are two different ways to display the information relating to a single primer, the detailed and the compact view. Both are shown below the primer regions selected on the sequence.

30.3.1 Compact information mode

This mode offers a condensed overview of all the primers that are available in the selected region. When a region is chosen primer information will appear in lines beneath it (see figure 30.4).



Primer covering positions 20 to 37:

Fraction of G and C: 0,5
Melting temperature: 55,23 °C
Self annealing: 16
Self end annealing: 2
Secondary structure: 10

☑ Show
● Compact
○ Detailed
□ G/C content(G/C)
□ Melting temp.(Tm)
□ Self annealing(SA)
□ Self end annealing(SEA
□ Secondary structure(S:
□ 3' end G/C
□ 5' end G/C

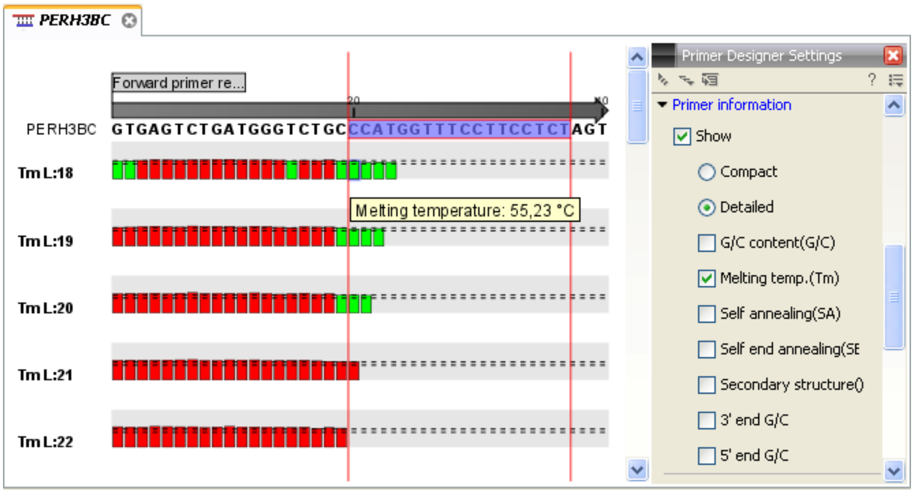
Figure 30.4: *Compact information mode*

The number of information lines reflects the chosen length interval for primers and probes. One line is shown for every possible primer-length, if the length interval is widened more lines will appear. At each potential primer starting position a circle is shown which indicates whether the primer fulfills the requirements set in the primer parameters preference group. A green primer indicates a primer which fulfils all criteria and a red primer indicates a primer which fails to meet one or more of the set criteria. For more detailed information, place the mouse cursor over the circle representing the primer of interest. A tool-tip will then appear on screen displaying detailed information about the primer in relation to the set criteria. To locate the primer on the sequence, simply left-click the circle using the mouse.

The various primer parameters can now be varied to explore their effect and the view area will dynamically update to reflect this. If e.g. the allowed melting temperature interval is widened more green circles will appear indicating that more primers now fulfill the set requirements and if e.g. a requirement for 3' G/C content is selected, rec circles will appear at the starting points of the primers which fail to meet this requirement.

30.3.2 Detailed information mode

In this mode a very detailed account is given of the properties of all the available primers. When a region is chosen primer information will appear in groups of lines beneath it (see figure 30.5).



Primer information

☑ Show

○ Compact

● Detailed

□ G/C content(G/C)

☑ Melting temp.(Tm)

□ Self annealing(SA)

□ Self end annealing(SE

□ Secondary structure()

□ 3' end G/C

□ 5' end G/C

Figure 30.5: *Detailed information mode*

The number of information-line-groups reflects the chosen length interval for primers and probes. One group is shown for every possible primer length. Within each group, a line is shown for every primer property that is selected from the checkboxes in the primer information preference group. Primer properties are shown at each potential primer starting position and are of two types:

Properties with numerical values are represented by bar plots. A green bar represents the starting point of a primer that meets the set requirement and a red bar represents the starting point of a primer that fails to meet the set requirement:

- G/C content
- Melting temperature
- Self annealing score
- Self end annealing score
- Secondary structure score

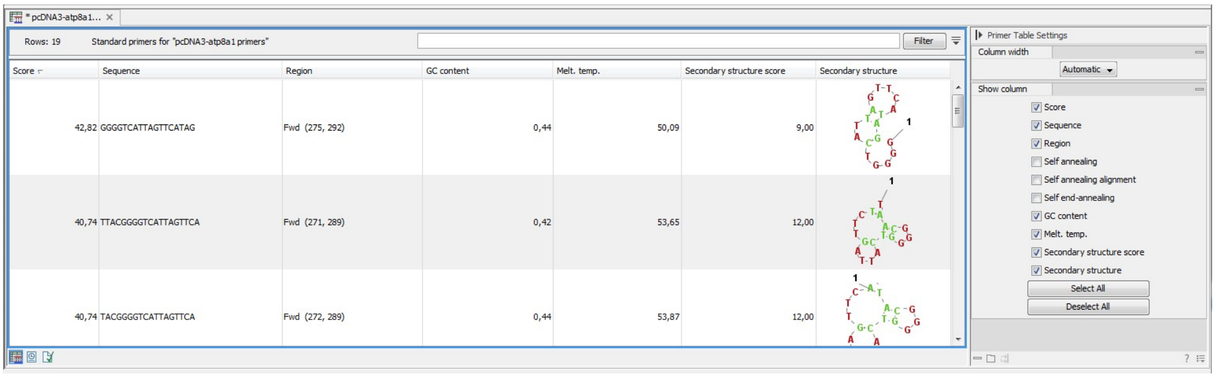
Properties with Yes - No values. If a primer meets the set requirement a green circle will be shown at its starting position and if it fails to meet the requirement a red dot is shown at its starting position:

- C/G at 3' end
- C/G at 5' end

Common to both sorts of properties is that mouse clicking an information point (filled circle or bar) will cause the region covered by the associated primer to be selected on the sequence.

30.4 Output from primer design

The output generated by the primer design algorithm is a table of proposed primers or primer pairs with the accompanying information (see figure 30.6).



Rows: 19 Standard primers for "pcDNA3-atp8a1 primers"

Score	Sequence	Region	GC content	Melt. temp.	Secondary structure score	Secondary structure
42,82	GGGGTCATTAGTTCATAG	Fwd (275, 292)	0,44	50,09	9,00	
40,74	TTACGGGGTCATTAGTTCA	Fwd (271, 289)	0,42	53,65	12,00	
40,74	TACGGGGTCATTAGTTCA	Fwd (272, 289)	0,44	53,87	12,00	

Primer Table Settings

Column width: Automatic

Show column

- ☑ Score
- ☑ Sequence
- ☑ Region
- □ Self annealing
- □ Self annealing alignment
- □ Self end-annealing
- ☑ GC content
- ☑ Melt. temp.
- ☑ Secondary structure score
- ☑ Secondary structure

Figure 30.6: *Proposed primers*

In the preference panel of the table, it is possible to customize which columns are shown in the table. See the sections below on the different reaction types for a description of the available information.

The columns in the output table can be sorted by the present information. For example the user can choose to sort the available primers by their score (default) or by their self annealing score, simply by right-clicking the column header.

The output table interacts with the accompanying primer editor such that when a proposed combination of primers and probes is selected in the table the primers and probes in this solution are highlighted on the sequence.

30.4.1 Saving primers

Primer solutions in a table row can be saved by selecting the row and using the right-click mouse menu. This opens a dialog that allows the user to save the primers to the desired location. Primers and probes are saved as DNA sequences in the program. This means that all available DNA analyzes can be performed on the saved primers, including BLAST. Furthermore, the primers can be edited using the standard sequence view to introduce e.g. mutations and restriction sites.

30.4.2 Saving PCR fragments

The PCR fragment generated from the primer pair in a given table row can also be saved by selecting the row and using the right-click mouse menu. This opens a dialog that allows the user to save the fragment to the desired location. The fragment is saved as a DNA sequence and the position of the primers is added as annotation on the sequence. The fragment can then be used for further analysis and included in e.g. an in-silico cloning experiment using the cloning editor.

30.4.3 Adding primer binding annotation

You can add an annotation to the template sequence specifying the binding site of the primer: Right-click the primer in the table and select **Mark primer annotation on sequence**.

30.5 Standard PCR

This mode is used to design primers for a PCR amplification of a single DNA fragment.

30.5.1 User input

In this mode the user must define either a *Forward primer region*, a *Reverse primer region*, or both. These are defined by making a selection on the sequence and right-clicking the selection. It is also possible to define a *Region to amplify* in which case a forward- and a reverse primer region are automatically placed so as to ensure that the designated region will be included in the PCR fragment. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

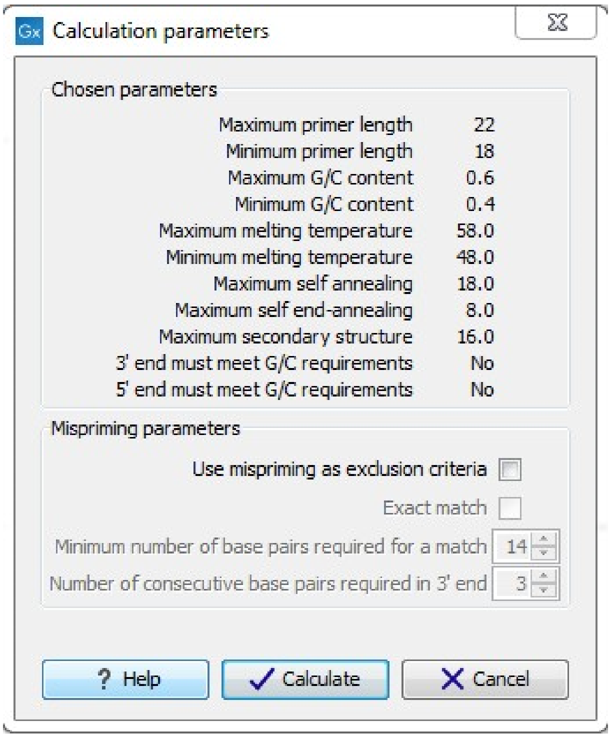
If two regions are defined, it is required that at least a part of the *Forward primer region* is located upstream of the *Reverse primer region*.

After exploring the available primers (see section 30.3) and setting the desired parameter values in the Primer Parameters preference group, the **Calculate** button will activate the primer design algorithm.

When a single primer region is defined

If only a single region is defined, only *single primers* will be suggested by the algorithm.

After pressing the **Calculate** button a dialog will appear (see figure 30.7).



Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Use mispriming as exclusion criteria □

Exact match □

Minimum number of base pairs required for a match 14

Number of consecutive base pairs required in 3' end 3

Figure 30.7: *Calculation dialog for PCR primers when only a single primer region has been defined.*

The top part of this dialog shows the parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm.

Mispriming: The lower part contains a menu where the user can choose to include mispriming as an exclusion criteria in the design process. If this option is selected the algorithm will search for competing binding sites of the primer within the rest of the sequence, to see if the primer would match to multiple locations. If a competing site is found (according to the parameters set), the primer will be excluded.

The adjustable parameters for the search are:

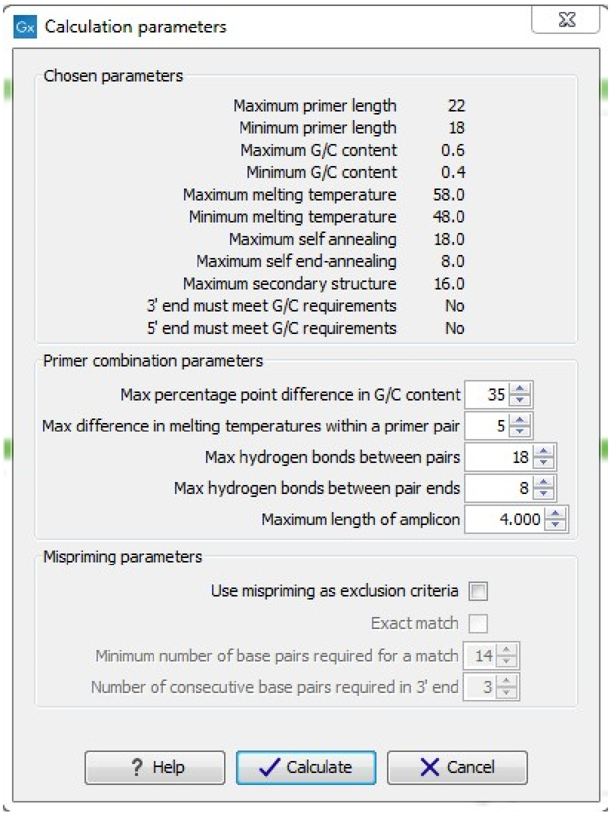
- **Exact match**. Choose only to consider exact matches of the primer, i.e. all positions must base pair with the template for mispriming to occur.
- **Minimum number of base pairs required for a match**. How many nucleotides of the primer that must base pair to the sequence in order to cause mispriming.
- **Number of consecutive base pairs required in 3' end**. How many consecutive 3' end base pairs in the primer that MUST be present for mispriming to occur. This option is included since 3' terminal base pairs are known to be essential for priming to occur.

Note! Including a search for potential mispriming sites will prolong the search time substantially if long sequences are used as template and if the minimum number of base pairs required for a match is low. If the region to be amplified is part of a very long molecule and mispriming is a concern, consider extracting part of the sequence prior to designing primers.

When both forward and reverse regions are defined

If both a forward and a reverse region are defined, *primer pairs* will be suggested by the algorithm.

After pressing the **Calculate** button a dialog will appear (see figure 30.8).



Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Primer combination parameters

Max percentage point difference in G/C content	35
Max difference in melting temperatures within a primer pair	5
Max hydrogen bonds between pairs	18
Max hydrogen bonds between pair ends	8
Maximum length of amplicon	4.000

Mispriming parameters

Use mispriming as exclusion criteria □

Exact match □

Minimum number of base pairs required for a match 14

Number of consecutive base pairs required in 3' end 3

Figure 30.8: *Calculation dialog for PCR primers when two primer regions have been defined.*

Again, the top part of this dialog shows the parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm. The lower part again contains a menu where the user can choose to include mispriming of both primers as a criteria in the design process (see section 30.5.1). The central part of the dialog contains parameters pertaining to primer pairs. Here three parameters can be set:

- Maximum percentage point difference in G/C content - if this is set at e.g. 5 points a pair of primers with 45% and 49% G/C nucleotides, respectively, will be allowed, whereas a pair of primers with 45% and 51% G/C nucleotides, respectively will not be included.
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ.
- Max hydrogen bonds between pairs - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair.
- Max hydrogen bonds between pair ends - the maximum number of hydrogen bonds allowed in the consecutive ends of the forward and the reverse primer in a primer pair.
- Maximum length of amplicon - determines the maximum length of the PCR fragment.

30.5.2 Standard PCR output table

If only a single region is selected the following columns of information are available:

- Sequence - the primer's sequence.
- Score - measures how much the properties of the primer (or primer pair) deviates from the optimal solution in terms of the chosen parameters and tolerances. The higher the score, the better the solution. The scale is from 0 to 100.
- Region - the interval of the template sequence covered by the primer
- Self annealing - the maximum self annealing score of the primer in units of hydrogen bonds
- Self annealing alignment - a visualization of the highest maximum scoring self annealing alignment
- Self end annealing - the maximum score of consecutive end base-pairings allowed between the ends of two copies of the same molecule in units of hydrogen bonds
- GC content - the fraction of G and C nucleotides in the primer
- Melting temperature of the primer-template complex
- Secondary structure score - the score of the optimal secondary DNA structure found for the primer. Secondary structures are scored by adding the number of hydrogen bonds in the structure, and 2 extra hydrogen bonds are added for each stacking base-pair in the structure
- Secondary structure - a visualization of the optimal DNA structure found for the primer

If both a forward and a reverse region are selected a table of primer pairs is shown, where the above columns (excluding the score) are represented twice, once for the forward primer (designated by the letter F) and once for the reverse primer (designated by the letter R).

Before these, and following the score of the primer pair, are the following columns pertaining to primer pair-information available:

- Pair annealing - the number of hydrogen bonds found in the optimal alignment of the forward and the reverse primer in a primer pair
- Pair annealing alignment - a visualization of the optimal alignment of the forward and the reverse primer in a primer pair.
- Pair end annealing - the maximum score of consecutive end base-pairings found between the ends of the two primers in the primer pair, in units of hydrogen bonds
- Fragment length - the length (number of nucleotides) of the PCR fragment generated by the primer pair

30.6 Nested PCR

Nested PCR is a modification of Standard PCR, aimed at reducing product contamination due to the amplification of unintended primer binding sites (mispriming). If the intended fragment can not be amplified without interference from competing binding sites, the idea is to seek out a larger outer fragment which can be unambiguously amplified and which contains the smaller intended fragment. Having amplified the outer fragment to large numbers, the PCR amplification of the inner fragment can proceed and will yield amplification of this with minimal contamination.

Primer design for nested PCR thus involves designing two primer pairs, one for the outer fragment and one for the inner fragment.

In *Nested PCR* mode the user must thus define four regions a *Forward primer region* (the outer forward primer), a *Reverse primer region* (the outer reverse primer), a *Forward inner primer region*, and a *Reverse inner primer region*. These are defined by making a selection on the sequence and right-clicking the selection. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

It is required that the *Forward primer region*, is located upstream of the *Forward inner primer region*, that the *Forward inner primer region*, is located upstream of the *Reverse inner primer region*, and that the *Reverse inner primer region*, is located upstream of the *Reverse primer region*.

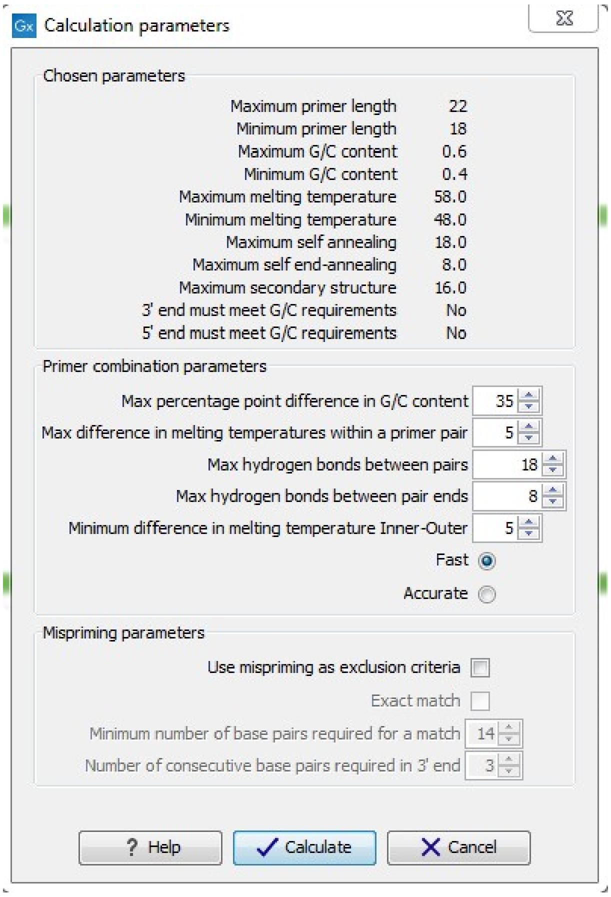
In *Nested PCR* mode the *Inner melting temperature* menu in the Primer parameters panel is activated, allowing the user to set a separate melting temperature interval for the inner and outer primer pairs.

After exploring the available primers (see section 30.3) and setting the desired parameter values in the Primer parameters preference group, the **Calculate** button will activate the primer design algorithm.

After pressing the **Calculate** button a dialog will appear (see figure 30.9).

The top and bottom parts of this dialog are identical to the *Standard PCR* dialog for designing primer pairs described above.

The central part of the dialog contains parameters pertaining to primer pairs and the comparison between the outer and the inner pair. Here five options can be set:



Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Primer combination parameters

Max percentage point difference in G/C content 35

Max difference in melting temperatures within a primer pair 5

Max hydrogen bonds between pairs 18

Max hydrogen bonds between pair ends 8

Minimum difference in melting temperature Inner-Outer 5

Fast ●

Accurate ○

Mispriming parameters

Use mispriming as exclusion criteria □

Exact match □

Minimum number of base pairs required for a match 14

Number of consecutive base pairs required in 3' end 3

Figure 30.9: *Calculation dialog*

- Maximum percentage point difference in G/C content (described above under Standard PCR) - this criteria is applied to both primer pairs independently.

- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ. This criteria is applied to both primer pairs independently.

- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair. This criteria is applied to all possible combinations of primers.

- Minimum difference in the melting temperature of primers in the inner and outer primer pair - all comparisons between the melting temperature of primers from the two pairs must be at least this different, otherwise the primer set is excluded. This option is applied to ensure that the inner and outer PCR reactions can be initiated at different annealing temperatures. Please note that to ensure flexibility there is no directionality indicated when setting parameters for melting temperature differences between inner and outer primer pair, i.e. it is not specified whether the inner pair should have a lower or higher T_m. Instead this is determined by the allowed temperature intervals for inner and outer primers that are set in the primer parameters preference group in the side panel. If a higher T_m of inner primers is desired, choose a T_m interval for inner primers which has higher values than the interval for outer primers.

- Two radio buttons allowing the user to choose between a fast and an accurate algorithm for primer prediction.

30.6.1 Nested PCR output table

In nested PCR there are four primers in a solution, forward outer primer (FO), forward inner primer (FI), reverse inner primer (RI) and a reverse outer primer (RO).

The output table can show primer-pair combination parameters for all four combinations of primers and single primer parameters for all four primers in a solution (see section on Standard PCR for an explanation of the available primer-pair and single primer information).

The fragment length in this mode refers to the length of the PCR fragment generated by the inner primer pair, and this is also the PCR fragment which can be exported.

30.7 TaqMan

CLC Cancer Research Workbench allows the user to design primers and probes for TaqMan PCR applications.

TaqMan probes are oligonucleotides that contain a fluorescent reporter dye at the 5' end and a quenching dye at the 3' end. Fluorescent molecules become excited when they are irradiated and usually emit light. However, in a TaqMan probe the energy from the fluorescent dye is transferred to the quencher dye by fluorescence resonance energy transfer as long as the quencher and the dye are located in close proximity i.e. when the probe is intact. TaqMan probes are designed to anneal within a PCR product amplified by a standard PCR primer pair. If a TaqMan probe is bound to a product template, the replication of this will cause the Taq polymerase to encounter the probe. Upon doing so, the 5'exonuclease activity of the polymerase will cleave the probe. This cleavage separates the quencher and the dye, and as a result the reporter dye starts to emit fluorescence.

The TaqMan technology is used in Real-Time quantitative PCR. Since the accumulation of fluorescence mirrors the accumulation of PCR products it can can be monitored in real-time and used to quantify the amount of template initially present in the buffer.

The technology is also used to detect genetic variation such as SNP's. By designing a TaqMan probe which will specifically bind to one of two or more genetic variants it is possible to detect genetic variants by the presence or absence of fluorescence in the reaction.

A specific requirement of TaqMan probes is that a G nucleotide can not be present at the 5' end since this will quench the fluorescence of the reporter dye. It is recommended that the melting temperature of the TaqMan probe is about 10 degrees celsius higher than that of the primer pair.

Primer design for TaqMan technology involves designing a primer pair and a TaqMan probe.

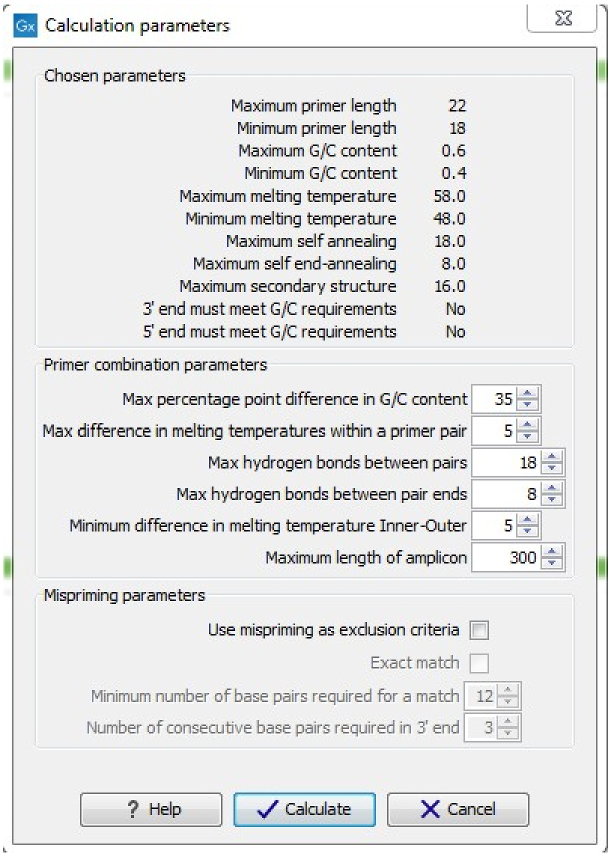
In *TaqMan* the user must thus define three regions: a *Forward primer region*, a *Reverse primer region*, and a *TaqMan probe region*. The easiest way to do this is to designate a *TaqMan primer/probe region* spanning the sequence region where TaqMan amplification is desired. This will automatically add all three regions to the sequence. If more control is desired about the placing of primers and probes the *Forward primer region*, *Reverse primer region* and *TaqMan probe region* can all be defined manually. If areas are known where primers or probes must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined. The regions are defined by making a selection on the sequence and right-clicking the selection.

It is required that at least a part of the *Forward primer region* is located upstream of the *TaqMan Probe region*, and that the *TaqMan Probe region*, is located upstream of a part of the *Reverse primer region*.

In *TaqMan* mode the *Inner melting temperature* menu in the primer parameters panel is activated allowing the user to set a separate melting temperature interval for the TaqMan probe.

After exploring the available primers (see section 30.3) and setting the desired parameter values in the Primer Parameters preference group, the **Calculate** button will activate the primer design algorithm.

After pressing the **Calculate** button a dialog will appear (see figure 30.10) which is similar to the *Nested PCR* dialog described above (see section 30.6).



Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Primer combination parameters

Max percentage point difference in G/C content	35
Max difference in melting temperatures within a primer pair	5
Max hydrogen bonds between pairs	18
Max hydrogen bonds between pair ends	8
Minimum difference in melting temperature Inner-Outer	5
Maximum length of amplicon	300

Mispriming parameters

Use mispriming as exclusion criteria	☐
Exact match	☐
Minimum number of base pairs required for a match	12
Number of consecutive base pairs required in 3' end	3

Figure 30.10: *Calculation dialog*

In this dialog the options to set a minimum and a desired melting temperature difference between outer and inner refers to primer pair and probe respectively.

Furthermore, the central part of the dialog contains an additional parameter

- Maximum length of amplicon - determines the maximum length of the PCR fragment generated in the TaqMan analysis.

30.7.1 TaqMan output table

In TaqMan mode there are two primers and a probe in a given solution, forward primer (F), reverse primer (R) and a TaqMan probe (TP).

The output table can show primer/probe-pair combination parameters for all three combinations of primers and single primer parameters for both primers and the TaqMan probe (see section on Standard PCR for an explanation of the available primer-pair and single primer information).

The fragment length in this mode refers to the length of the PCR fragment generated by the primer pair, and this is also the PCR fragment which can be exported.

30.8 Sequencing primers

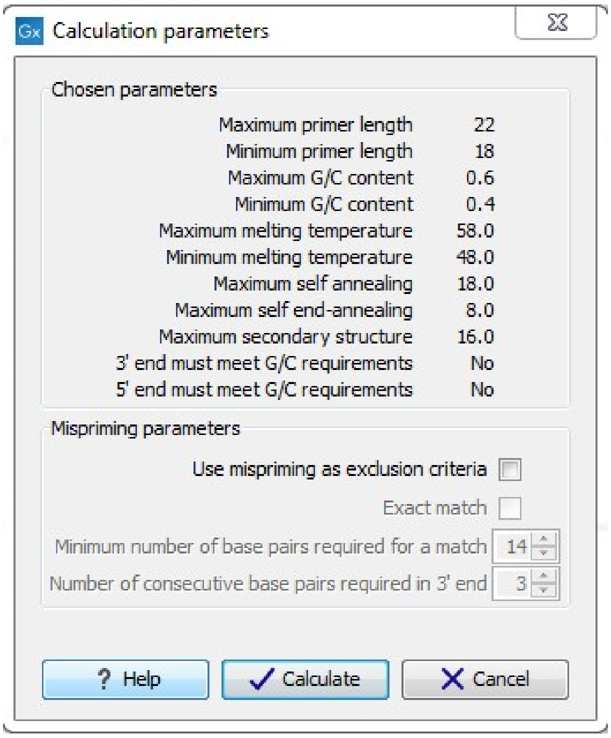
This mode is used to design primers for DNA sequencing.

In this mode the user can define a number of *Forward primer regions* and *Reverse primer regions* where a sequencing primer can start. These are defined by making a selection on the sequence and right-clicking the selection. If areas are known where primers must not bind (e.g. repeat rich areas), one or more *No primers here* regions can be defined.

No requirements are instated on the relative position of the regions defined.

After exploring the available primers (see section 30.3) and setting the desired parameter values in the Primer Parameters preference group, the **Calculate** button will activate the primer design algorithm.

After pressing the **Calculate** button a dialog will appear (see figure 30.11).



Calculation parameters

Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Mispriming parameters

Use mispriming as exclusion criteria □

Exact match □

Minimum number of base pairs required for a match 14

Number of consecutive base pairs required in 3' end 3

Figure 30.11: *Calculation dialog for sequencing primers*

Since design of sequencing primers does not require the consideration of interactions between primer pairs, this dialog is identical to the dialog shown in *Standard PCR* mode when only a single primer region is chosen. See the section 30.5 for a description.

30.8.1 Sequencing primers output table

In this mode primers are predicted independently for each region, but the optimal solutions are all presented in one table. The solutions are numbered consecutively according to their position on the sequence such that the forward primer region closest to the 5' end of the molecule is designated F1, the next one F2 etc.

For each solution, the single primer information described under Standard PCR is available in the table.

30.9 Alignment-based primer and probe design

CLC Cancer Research Workbench allows the user to design PCR primers and TaqMan probes based on an alignment of multiple sequences.

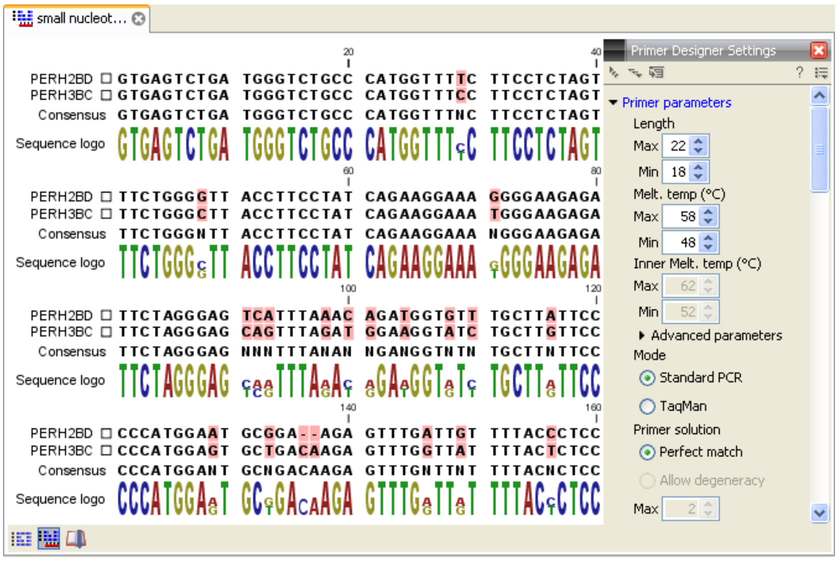
The primer designer for alignments can be accessed in two ways:

Toolbox | Sanger Sequencing (**) | **Primers and Probes ()| Design Primers ()

or **If the alignment is already open: | Click Primer Designer () in the lower left part of the view**

In the alignment primer view (see figure 30.12), the basic options for viewing the template alignment are the same as for the standard view of alignments.

Note! This means that annotations such as e.g. known SNPs or exons, can be displayed on the template sequence to guide the choice of primer regions. Since the definition of groups of sequences is essential to the primer design, the selection boxes of the standard view are shown as default in the alignment primer view.



```
                                            20                      40
    PERH2BD □ GTGAGTCTGA TGGGTCTGCC CATGGTTTTC TTCCTCTAGT
    PERH3BC □ GTGAGTCTGA TGGGTCTGCC CATGGTTTCC TTCCTCTAGT
  Consensus   GTGAGTCTGA TGGGTCTGCC CATGGTTTNC TTCCTCTAGT
                                            60                      80
    PERH2BD □ TTCTGGGGTT ACCTTCCTAT CAGAAGGAAA GGGGAAGAGA
    PERH3BC □ TTCTGGGCTT ACCTTCCTAT CAGAAGGAAA TGGGAAGAGA
  Consensus   TTCTGGGNTT ACCTTCCTAT CAGAAGGAAA NGGGAAGAGA
                                           100                     120
    PERH2BD □ TTCTAGGGAG TCATTTAAAC AGATGGTGTT TGCTTATTCC
    PERH3BC □ TTCTAGGGAG CAGTTTAGAT GGAAGGTATC TGCTTGTTCC
  Consensus   TTCTAGGGAG NNNTTTANAN NGANGGTNTN TGCTTNTTCC
                                           140                     160
    PERH2BD □ CCCATGGAAT GCGGA--AGA GTTTGATTGT TTTACCCTCC
    PERH3BC □ CCCATGGAGT GCTGACAAGA GTTTGGTTAT TTTACTCTCC
  Consensus   CCCATGGANT GCNGACAAGA GTTTGNTTNT TTTACNCTCC
```

Primer Designer Settings

Primer parameters

Length

Max 22

Min 18

Melt. temp (°C)

Max 58

Min 48

Inner Melt. temp (°C)

Max 62

Min 52

Advanced parameters

Mode

● Standard PCR

○ TaqMan

Primer solution

● Perfect match

○ Allow degeneracy

Max 2

Figure 30.12: *The initial view of an alignment used for primer design.*

30.9.1 Specific options for alignment-based primer and probe design

Compared to the primer view of a single sequence the most notable difference is that the alignment primer view has no available graphical information. Furthermore, the selection boxes found to the left of the names in the alignment play an important role in specifying the oligo design process. This is elaborated below. The **Primer Parameters** group in the **Side Panel** has the same options for specifying primer requirements, but differs by the following (see figure 30.12):

- In the **Mode** submenu which specifies the reaction types the following options are found:
  - **Standard PCR.** Used when the objective is to design primers, or primer pairs, for PCR amplification of a single DNA fragment.
  - **TaqMan.** Used when the objective is to design a primer pair and a probe set for TaqMan quantitative PCR.
- The **Primer solution** submenu is used to specify requirements for the match of a PCR primer against the template sequences. These options are described further below. It contains the following options:
  - **Perfect match.**
  - **Allow degeneracy.**
  - **Allow mismatches.**

The work flow when designing alignment based primers and probes is as follows:

- Use selection boxes to specify groups of included and excluded sequences. To select all the sequences in the alignment, right-click one of the selection boxes and choose **Mark All**.
- Mark either a single forward primer region, a single reverse primer region or both on the sequence (and perhaps also a TaqMan region). Selections must cover all sequences in the included group. You can also specify that there should be no primers in a region (No Primers Here) or that a whole region should be amplified (Region to Amplify).
- Adjust parameters regarding single primers in the preference panel.
- Click the **Calculate** button.

30.9.2 Alignment based design of PCR primers

In this mode, a single or a pair of PCR primers are designed. *CLC Cancer Research Workbench* allows the user to design primers which will specifically amplify a group of *included* sequences but **not** amplify the remainder of the sequences, the *excluded* sequences. The selection boxes are used to indicate the status of a sequence, if the box is checked the sequence belongs to the included sequences, if not, it belongs to the excluded sequences. To design primers that are general for all primers in an alignment, simply add them all to the set of included sequences by checking all selection boxes. Specificity of priming is determined by criteria set by the user in the dialog box which is shown when the **Calculate** button is pressed (see below).

Different options can be chosen concerning the match of the primer to the template sequences in the included group:

- **Perfect match.** Specifies that the designed primers must have a perfect match to all relevant sequences in the alignment. When selected, primers will thus only be located in regions that are completely conserved within the sequences belonging to the included group.
- **Allow degeneracy.** Designs primers that may include ambiguity characters where heterogeneities occur in the included template sequences. The allowed fold of degeneracy is user defined and corresponds to the number of possible primer combinations formed by a degenerate primer. Thus, if a primer covers two 4-fold degenerate site and one 2-fold degenerate site the total fold of degeneracy is $4 * 4 * 2 = 32$ and the primer will, when supplied from the manufacturer, consist of a mixture of 32 different oligonucleotides. When scoring the available primers, degenerate primers are given a score which decreases with the fold of degeneracy.
- **Allow mismatches.** Designs primers which are allowed a specified number of mismatches to the included template sequences. The melting temperature algorithm employed includes the latest thermodynamic parameters for calculating T_m when single-base mismatches occur.

When in Standard PCR mode, clicking the **Calculate** button will prompt the dialog shown in figure 30.13.

The top part of this dialog shows the single-primer parameter settings chosen in the Primer parameters preference group which will be used by the design algorithm.

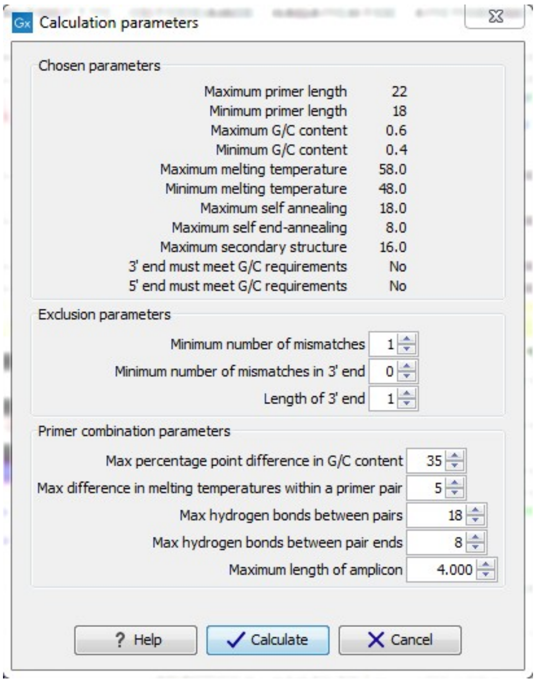
The central part of the dialog contains parameters pertaining to primer specificity (this is omitted if all sequences belong to the included group). Here, three parameters can be set:

- Minimum number of mismatches - the minimum number of mismatches that a primer must have against all sequences in the excluded group to ensure that it does not prime these.
- Minimum number of mismatches in 3' end - the minimum number of mismatches that a primer must have in its 3' end against all sequences in the excluded group to ensure that it does not prime these.
- Length of 3' end - the number of consecutive nucleotides to consider for mismatches in the 3' end of the primer.

The lower part of the dialog contains parameters pertaining to primer pairs (this is omitted when only designing a single primer). Here, three parameters can be set:

- Maximum percentage point difference in G/C content - if this is set at e.g. 5 points a pair of primers with 45% and 49% G/C nucleotides, respectively, will be allowed, whereas a pair of primers with 45% and 51% G/C nucleotides, respectively will not be included.
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in a pair are all allowed to differ.
- Max hydrogen bonds between pairs - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in a primer pair.
- Maximum length of amplicon - determines the maximum length of the PCR fragment.

The output of the design process is a table of single primers or primer pairs as described for primer design based on single sequences. These primers are specific to the included sequences in the alignment according to the criteria defined for specificity. The only novelty in the table, is that melting temperatures are displayed with both a maximum, a minimum and an average value to reflect that degenerate primers or primers with mismatches may have heterogeneous behavior on the different templates in the group of included sequences.



Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Exclusion parameters

Minimum number of mismatches	1
Minimum number of mismatches in 3' end	0
Length of 3' end	1

Primer combination parameters

Max percentage point difference in G/C content	35
Max difference in melting temperatures within a primer pair	5
Max hydrogen bonds between pairs	18
Max hydrogen bonds between pair ends	8
Maximum length of amplicon	4.000

Figure 30.13: *Calculation dialog shown when designing alignment based PCR primers.*

30.9.3 Alignment-based TaqMan probe design

CLC Cancer Research Workbench allows the user to design solutions for TaqMan quantitative PCR which consist of four oligos: a general primer pair which will amplify all sequences in the alignment, a specific TaqMan probe which will match the group of *included* sequences but **not** match the *excluded* sequences and a specific TaqMan probe which will match the group of *excluded* sequences but **not** match the *included* sequences. As above, the selection boxes are used to indicate the status of a sequence, if the box is checked the sequence belongs to the included sequences, if not, it belongs to the excluded sequences. We use the terms included and excluded here to be consistent with the section above although a probe solution is presented for both groups. In TaqMan mode, primers are not allowed degeneracy or mismatches to any template sequence in the alignment, variation is only allowed/required in the TaqMan probes.

Pushing the **Calculate** button will cause the dialog shown in figure 30.14 to appear.

The top part of this dialog is identical to the *Standard PCR* dialog for designing primer pairs described above.

The central part of the dialog contains parameters to define the specificity of TaqMan probes. Two parameters can be set:

- Minimum number of mismatches - the minimum total number of mismatches that must exist between a specific TaqMan probe and all sequences which belong to the group not recognized by the probe.

- Minimum number of mismatches in central part - the minimum number of mismatches in the central part of the oligo that must exist between a specific TaqMan probe and all sequences which belong to the group not recognized by the probe.

The lower part of the dialog contains parameters pertaining to primer pairs and the comparison between the outer oligos(primers) and the inner oligos (TaqMan probes). Here, five options can be set:

- Maximum percentage point difference in G/C content (described above under Standard PCR).
- Maximal difference in melting temperature of primers in a pair - the number of degrees Celsius that primers in the primer pair are all allowed to differ.
- Maximum pair annealing score - the maximum number of hydrogen bonds allowed between the forward and the reverse primer in an oligo pair. This criteria is applied to all possible combinations of primers and probes.
- Minimum difference in the melting temperature of primer (outer) and TaqMan probe (inner) oligos - all comparisons between the melting temperature of primers and probes must be at least this different, otherwise the solution set is excluded.
- Desired temperature difference in melting temperature between outer (primers) and inner (TaqMan) oligos - the scoring function discounts solution sets which deviate greatly from this value. Regarding this, and the minimum difference option mentioned above, please note that to ensure flexibility there is no directionality indicated when setting parameters for melting temperature differences between probes and primers, i.e. it is not specified whether the probes should have a lower or higher T_m. Instead this is determined by the allowed temperature intervals for inner and outer oligos that are set in the primer parameters preference group in the side panel. If a higher T_m of probes is required, choose a T_m interval for probes which has higher values than the interval for outer primers.

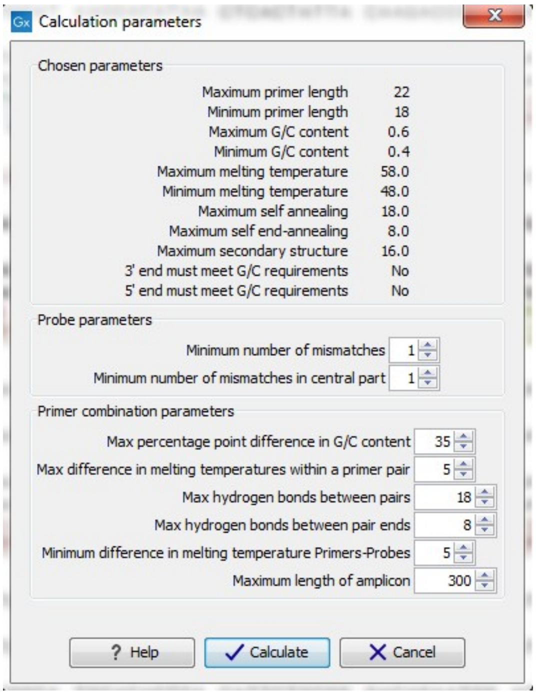
The output of the design process is a table of solution sets. Each solution set contains the following: a set of primers which are general to all sequences in the alignment, a TaqMan probe which is specific to the set of included sequences (sequences where selection boxes are checked) and a TaqMan probe which is specific to the set of excluded sequences (marked by *). Otherwise, the table is similar to that described above for TaqMan probe prediction on single sequences.

30.10 Analyze primer properties

CLC Cancer Research Workbench can calculate and display the properties of predefined primers and probes:

Toolbox | **Primers and Probes** ()| **Analyze Primer Properties** ()

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove a sequence from the selected elements. (Primers are represented as DNA sequences in the Navigation Area).



Chosen parameters

Maximum primer length	22
Minimum primer length	18
Maximum G/C content	0.6
Minimum G/C content	0.4
Maximum melting temperature	58.0
Minimum melting temperature	48.0
Maximum self annealing	18.0
Maximum self end-annealing	8.0
Maximum secondary structure	16.0
3' end must meet G/C requirements	No
5' end must meet G/C requirements	No

Probe parameters

Minimum number of mismatches 1

Minimum number of mismatches in central part 1

Primer combination parameters

Max percentage point difference in G/C content 35

Max difference in melting temperatures within a primer pair 5

Max hydrogen bonds between pairs 18

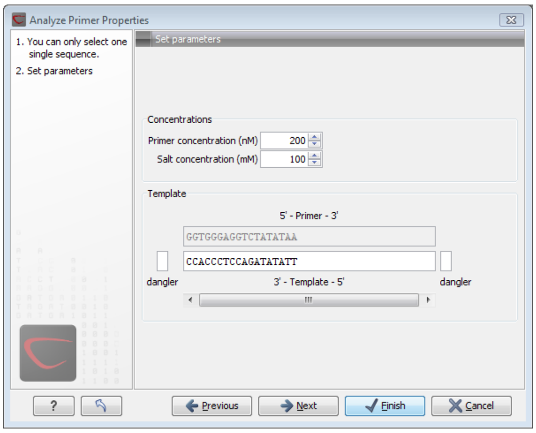
Max hydrogen bonds between pair ends 8

Minimum difference in melting temperature Primers-Probes 5

Maximum length of amplicon 300

Figure 30.14: *Calculation dialog shown when designing alignment based TaqMan probes.*

Clicking **Next** generates the dialog seen in figure 30.15:



1. You can only select one single sequence.
2. Set parameters

Concentrations

Primer concentration (nM) 200

Salt concentration (mM) 100

Template

5' - Primer - 3'

GGTGGGAGGTCTATATAA

CCACCCTCCAGATATATT

dangler 3' - Template - 5' dangler

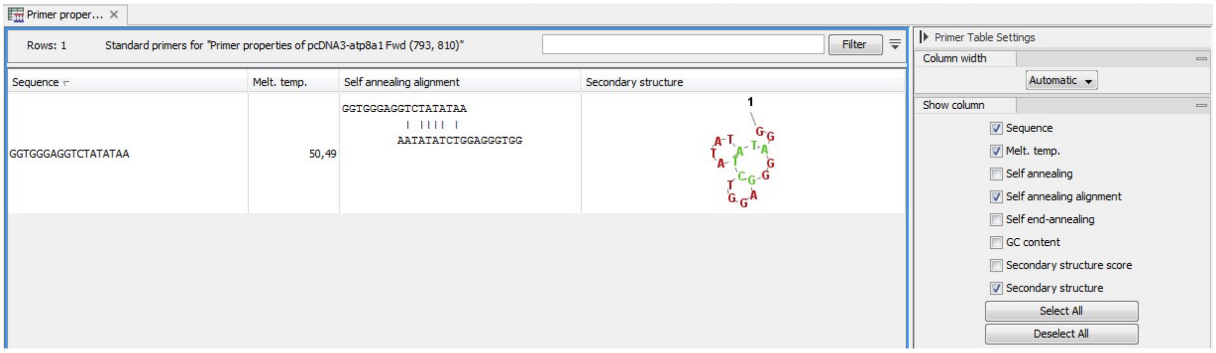
Figure 30.15: *The parameters for analyzing primer properties.*

In the *Concentrations* panel a number of parameters can be specified concerning the reaction mixture and which influence melting temperatures

- **Primer concentration.** Specifies the concentration of primers and probes in units of nanomoles (nM)
- **Salt concentration.** Specifies the concentration of monovalent cations ($[NA^+]$, $[K^+]$ and equivalents) in units of millimoles (mM)

In the *Template panel* the sequences of the chosen primer and the template sequence are shown. The template sequence is as default set to the reverse complement of the primer sequence i.e. as perfectly base-pairing. However, it is possible to edit the template to introduce mismatches which may affect the melting temperature. At each side of the template sequence a text field is shown. Here, the dangling ends of the template sequence can be specified. These may have an important affect on the melting temperature [Bommarito et al., 2000]

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**. The result is shown in figure 30.16:



Rows: 1 Standard primers for "Primer properties of pcDNA3-atp8a1 Fwd (793, 810)"

Sequence	Melt. temp.	Self annealing alignment	Secondary structure
GGTGGGAGGTCTATATAA	50,49	GGTGGGAGGTCTATATAA | |||| | AATATATCTGGAGGGTGG	

Column width: Automatic

Show column

☑ Sequence
☑ Melt. temp.
□ Self annealing
☑ Self annealing alignment
□ Self end-annealing
□ GC content
□ Secondary structure score
☑ Secondary structure

Figure 30.16: *Properties of a primer from the Example Data.*

In the **Side Panel** you can specify the information to display about the primer. The information parameters of the primer properties table are explained in section 30.5.2.

30.11 Find binding sites and create fragments

In *CLC Cancer Research Workbench* you have the possibility of matching known primers against one or more DNA sequences or a list of DNA sequences. This can be applied to test whether a primer used in a previous experiment is applicable to amplify e.g. a homologous region in another species, or to test for potential mispriming. This functionality can also be used to extract the resulting PCR product when two primers are matched. This is particularly useful if your primers have extensions in the 5' end.

To search for primer binding sites:

> **Toolbox | Sanger Sequencing ( ) |Primers and Probes ( )| Find Binding Sites and Create Fragments ( )**

If a sequence was already selected in the Navigation Area, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** when all the sequence have been added.

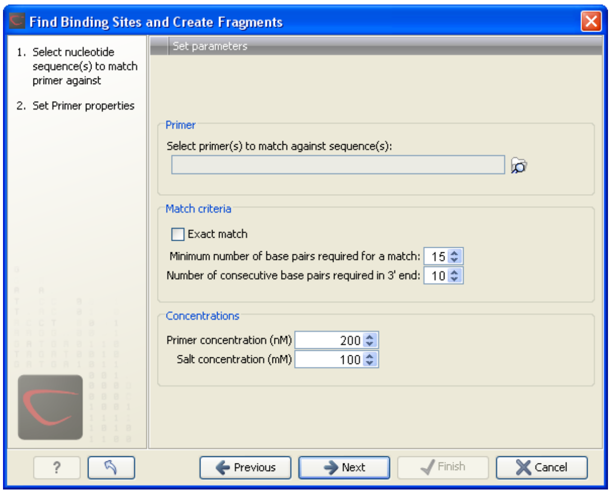
Note! You should not add the primer sequences at this step.

30.11.1 Binding parameters

This opens the dialog displayed in figure 30.17:

At the top, select one or more primers by clicking the browse ( ) button. In *CLC Cancer Research Workbench*, primers are just DNA sequences like any other, but there is a filter on the length of the sequence. Only sequences up to 400 bp can be added.

The **Match criteria** for matching a primer to a sequence are:



Find Binding Sites and Create Fragments

1. Select nucleotide sequence(s) to match primer against
2. Set Primer properties

Set parameters

Primer

Select primer(s) to match against sequence(s):

Match criteria

□ Exact match

Minimum number of base pairs required for a match: 15

Number of consecutive base pairs required in 3' end: 10

Concentrations

Primer concentration (nM) 200

Salt concentration (mM) 100

Figure 30.17: *Search parameters for finding primer binding sites.*

- **Exact match**. Choose only to consider exact matches of the primer, i.e. all positions must base pair with the template.
- **Minimum number of base pairs required for a match**. How many nucleotides of the primer that must base pair to the sequence in order to cause priming/mispriming.
- **Number of consecutive base pairs required in 3' end**. How many consecutive 3' end base pairs in the primer that MUST be present for priming/mispriming to occur. This option is included since 3' terminal base pairs are known to be essential for priming to occur.

Note that the number of mismatches is reported in the output, so you will be able to filter on this afterwards (see below).

Below the match settings, you can adjust **Concentrations** concerning the reaction mixture. This is used when reporting melting temperatures for the primers.

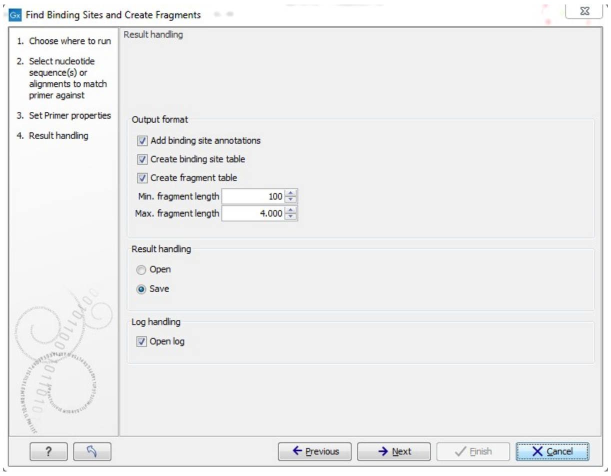
- **Primer concentration.** Specifies the concentration of primers and probes in units of nanomoles (nM)
- **Salt concentration.** Specifies the concentration of monovalent cations ($[NA^+]$, $[K^+]$ and equivalents) in units of millimoles (mM)

30.11.2 Results - binding sites and fragments

Click **Next** to specify the output options as shown in figure 30.18:

The output options are:

- **Add binding site annotations**. This will add annotations to the input sequences (see details below).
- **Create binding site table**. Creates a table of all binding sites. Described in details below.
- **Create fragment table**. Showing a table of all fragments that could result from using the primers. Note that you can set the minimum and maximum sizes of the fragments to be shown. The table is described in detail below.


Find Binding Sites and Create Fragments
1. Choose where to run
2. Select nucleotide sequence(s) or alignments to match primer against
3. Set Primer properties
4. Result handling
Result handling
?
Previous
Next
Finish
Cancel

Output format

☑ Add binding site annotations

☑ Create binding site table

☑ Create fragment table

Min. fragment length 100

Max. fragment length 4.000

Result handling

○ Open

● Save

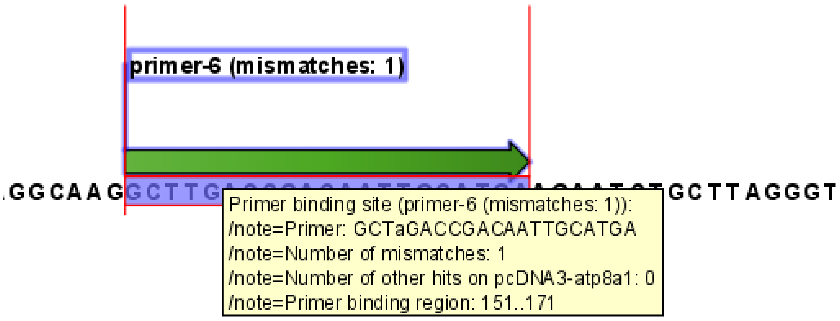
Log handling

☑ Open log

Figure 30.18: *Output options include reporting of binding sites and fragments.*

Click **Next** if you wish to adjust how to handle the results (see section 8.2). If not, click **Finish**.

An example of a **binding site annotation** is shown in figure 30.19.



Primer binding site (primer-6 (mismatches: 1)):
/note=Primer: GCTaGACCGACAATTGCATGA
/note=Number of mismatches: 1
/note=Number of other hits on pcDNA3-atp8a1: 0
/note=Primer binding region: 151..171

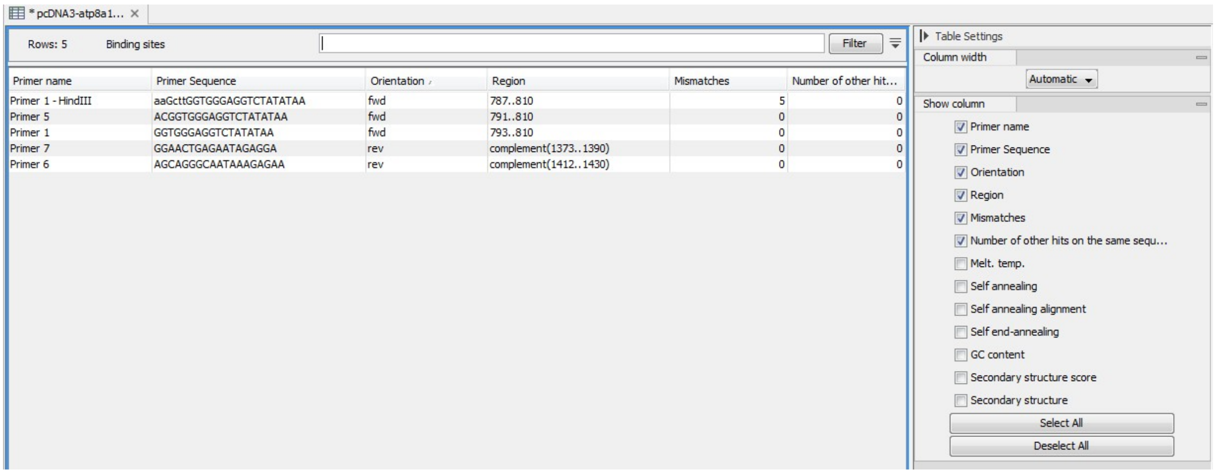
Figure 30.19: *Annotation showing a primer match.*

The annotation has the following information:

- **Sequence of the primer.** Positions with mismatches will be in lower-case (see the fourth position in figure 30.19 where the primer has an `a` and the template sequence has a `T`).
- **Number of mismatches.**
- **Number of other hits on the same sequence.** This number can be useful to check specificity of the primer.
- **Binding region.** This region ends with the 3' exact match and is simply the primer length upstream. This means that if you have 5' extensions to the primer, part of the binding region covers sequence that will actually not be annealed to the primer.

An example of the **primer binding site table** is shown in figure 30.20.

The information here is the same as in the primer annotation and furthermore you can see additional information about melting temperature etc. by selecting the options in the **Side Panel**. See a more detailed description of this information in section 30.5.2. You can use this table to browse the binding sites. If you make a split view of the table and the sequence (see



Rows: 5 Binding sites

Primer name	Primer Sequence	Orientation	Region	Mismatches	Number of other hit...
Primer 1 - HindIII	aaGcttGGTGGGAGGTCTATATAA	fwd	787..810	5	0
Primer 5	ACGGTGGGAGGTCTATATAA	fwd	791..810	0	0
Primer 1	GGTGGGAGGTCTATATAA	fwd	793..810	0	0
Primer 7	GGAACTGAGAATAGAGGA	rev	complement(1373..1390)	0	0
Primer 6	AGCAGGGCAATAAAGAGAA	rev	complement(1412..1430)	0	0

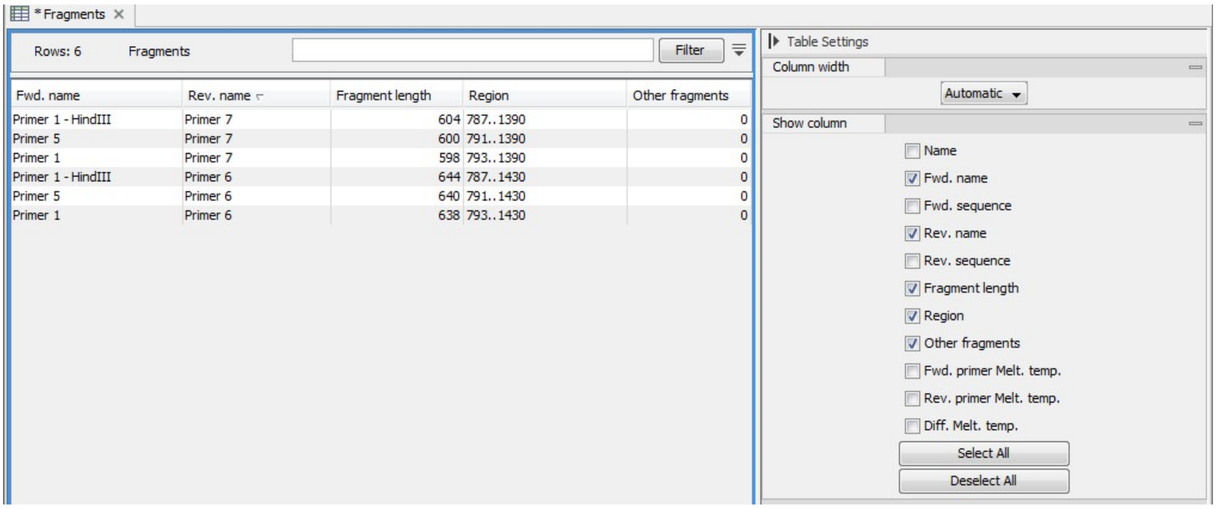
Column width: Automatic

Show column
- ☑ Primer name
- ☑ Primer Sequence
- ☑ Orientation
- ☑ Region
- ☑ Mismatches
- ☑ Number of other hits on the same sequ...
- □ Melt. temp.
- □ Self annealing
- □ Self annealing alignment
- □ Self end-annealing
- □ GC content
- □ Secondary structure score
- □ Secondary structure

Figure 30.20: *A table showing all binding sites.*

section 2.1.6), you can browse through the binding positions by clicking in the table. This will cause the sequence view to jump to the position of the binding site.

An example of a **fragment table** is shown in figure 30.21.



Rows: 6 Fragments

Fwd. name	Rev. name	Fragment length	Region	Other fragments
Primer 1 - HindIII	Primer 7	604	787..1390	0
Primer 5	Primer 7	600	791..1390	0
Primer 1	Primer 7	598	793..1390	0
Primer 1 - HindIII	Primer 6	644	787..1430	0
Primer 5	Primer 6	640	791..1430	0
Primer 1	Primer 6	638	793..1430	0

Column width: Automatic

Show column
- □ Name
- ☑ Fwd. name
- □ Fwd. sequence
- ☑ Rev. name
- □ Rev. sequence
- ☑ Fragment length
- ☑ Region
- ☑ Other fragments
- □ Fwd. primer Melt. temp.
- □ Rev. primer Melt. temp.
- □ Diff. Melt. temp.

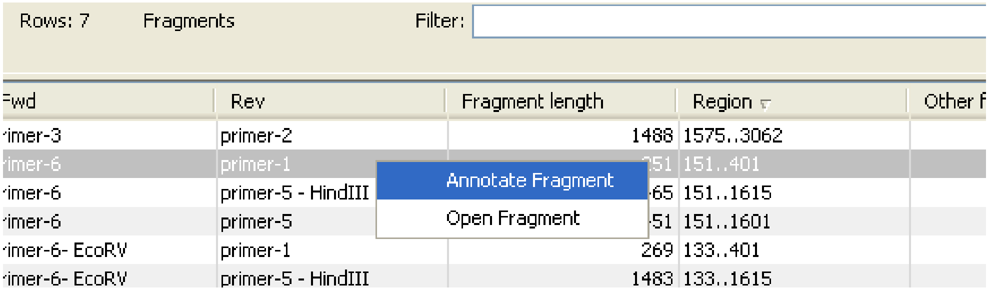
Figure 30.21: *A table showing all possible fragments of the specified size.*

The table first lists the names of the forward and reverse primers, then the length of the fragment and the region. The last column tells if there are other possible fragments fulfilling the length criteria on this sequence. This information can be used to check for competing products in the PCR. In the **Side Panel** you can show information about melting temperature for the primers as well as the difference between melting temperatures.

You can use this table to browse the fragment regions. If you make a split view of the table and the sequence (see section 2.1.6), you can browse through the fragment regions by clicking in the table. This will cause the sequence view to jump to the start position of the fragment.

There are some additional options in the fragment table. First, you can annotate the fragment on the original sequence. This is done by right-clicking (Ctrl-click on Mac) the fragment and choose **Annotate Fragment** as shown in figure 30.22.

This will put a *PCR fragment* annotations on the input sequence covering the region specified in



Rows: 7 Fragments Filter:

Fwd	Rev	Fragment length	Region	Other f
imer-3	primer-2	1488	1575..3062	
imer-6	primer-1	51	151..401	
imer-6	primer-5 - HindIII	65	151..1615	
imer-6	primer-5	51	151..1601	
imer-6- EcoRV	primer-1	269	133..401	
imer-6- EcoRV	primer-5 - HindIII	1483	133..1615	

Figure 30.22: *Right-clicking a fragment allows you to annotate the region on the input sequence or open the fragment as a new sequence.*

the table. As you can see from figure 30.22, you can also choose to **Open Fragment**. This will create a new sequence representing the PCR product that would be the result of using these two primers. Note that if you have extensions on the primers, they will be used to construct the new sequence.

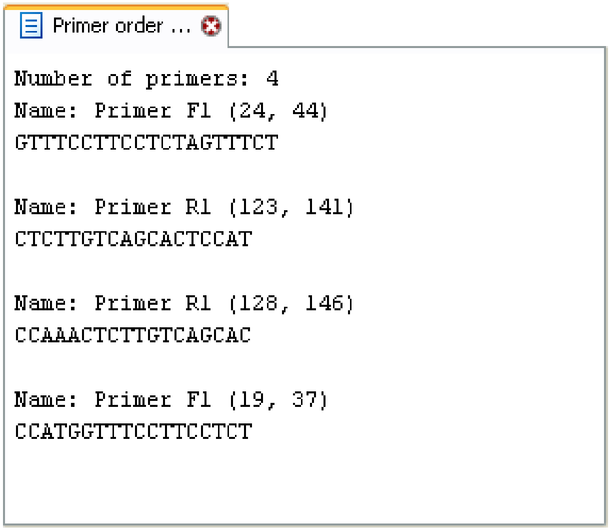
30.12 Order primers

To facilitate the ordering of primers and probes, *CLC Cancer Research Workbench* offers an easy way of displaying and saving a textual representation of one or more primers:

Toolbox | **Sanger Sequencing** ( ) | **Primers and Probes** ( )| **Order Primers** ( )

This opens a dialog where you can choose additional primers. Clicking **OK** opens a textual representation of the primers (see figure 30.23). The first line states the number of primers being ordered and after this follows the names and nucleotide sequences of the primers in 5'-3' orientation. From the editor, the primer information can be copied and pasted to web forms or e-mails. The created object can also be saved and exported as a text file.

See figure 30.23



```
Number of primers: 4
Name: Primer F1 (24, 44)
GTTTCCTTCCTCTAGTTTCT

Name: Primer R1 (123, 141)
CTCTTGTCAGCACTCCAT

Name: Primer R1 (128, 146)
CCAAACTCTTGTCAGCAC

Name: Primer F1 (19, 37)
CCATGGTTTCCTTCCTCT
```

Figure 30.23: *A primer order for 4 primers.*

Part X

Workflows

Chapter 31

Workflows

Contents

The *CLC Cancer Research Workbench* provides a framework for creating, distributing, installing and running workflows. Workflows created in the Workbench can also be installed on a *CLC Genomics Server*.

A workflow consists of a series of connected tools where the output of one tool is used as input for another tool. In this way you create a workflow that for example makes a read mapping, uses the mapped reads as input for variant detection, and performs filtering of the variant track. Once the workflow is set up, it can be installed (either in your own Workbench or on a Server or it can be sent to a colleague). In that way it becomes possible to analyze lots of samples using the same standard pipeline, the same reference data and the same parameters.

This chapter will first explain how to create a new workflow, and next go into details about the installation and execution of a workflow. For information about installing a workflow on the *CLC Genomics Server*, please see the user manual at http://www.clcbio.com/usermanuals.

Note that the examples below are using tools from the *CLC Genomics Workbench* that are not available in the *CLC Cancer Research Workbench*. But the principles and workflow framework can be used in the same way with tools from *CLC Cancer Research Workbench*.

31.1 Creating a workflow

A workflow can be created by pressing the "Workflows" button ( ) in the toolbar and then selecting "New Workflow..." ( ).

Alternatively, a workflow can be created via the menu bar:

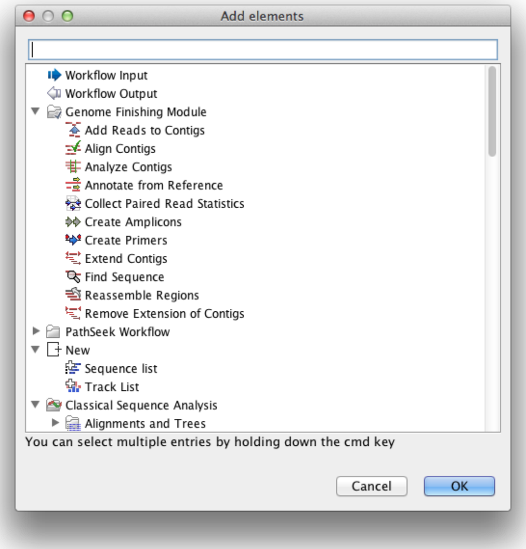
File | **New** | **Workflow ( )**

This will open a new view with a blank screen where a new workflow can be created.

31.1.1 Adding workflow elements

First, click the **Add Element** ( ) button at the bottom (or use the shortcut Shift + Alt + E). This will bring up a dialog that lists the elements and tools, which can be added to a workflow (see figure 31.1).

Alternatively elements can be dragged directly from the **Toolbox** into the workflow. Not all elements are workflow enabled. This means that only workflow enabled elements can be dropped in the workflow.



You can select multiple entries by holding down the cmd key

Figure 31.1: *Adding elements in the workflow.*

Elements that can be selected in the dialog are mostly tools from the Toolbox. However, there are two special elements on the list; the elements that are used for input and output. These two elements are explained in section 31.1.5.

You can select more than one element in the dialog by pressing Ctrl (⌘ on Mac) while selecting. Click OK when you have selected the relevant tools (you can always add more later on).

You will now see the selected elements in the editor (see figure 31.2).

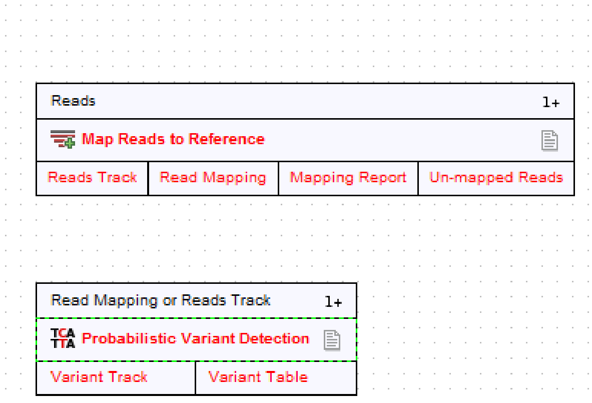


Figure 31.2: *Read mapping and variant calling added to the workflow.*

Once added, you can move and re-arrange the elements by dragging with the mouse (grab the box with the name of the element).

31.1.2 Configuring workflow elements

Each of the tools can be configured by right-clicking the name of the tool as shown in figure 31.3.

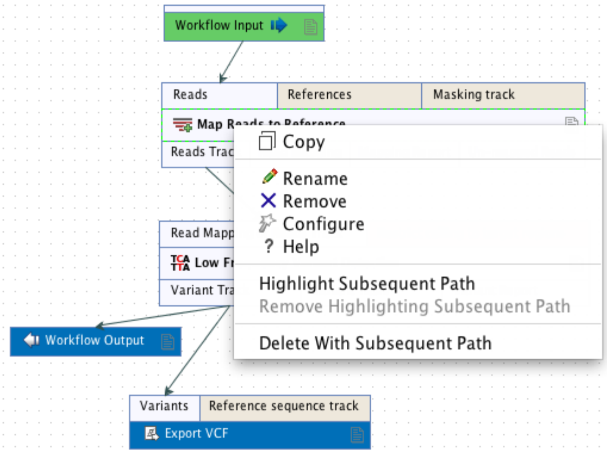


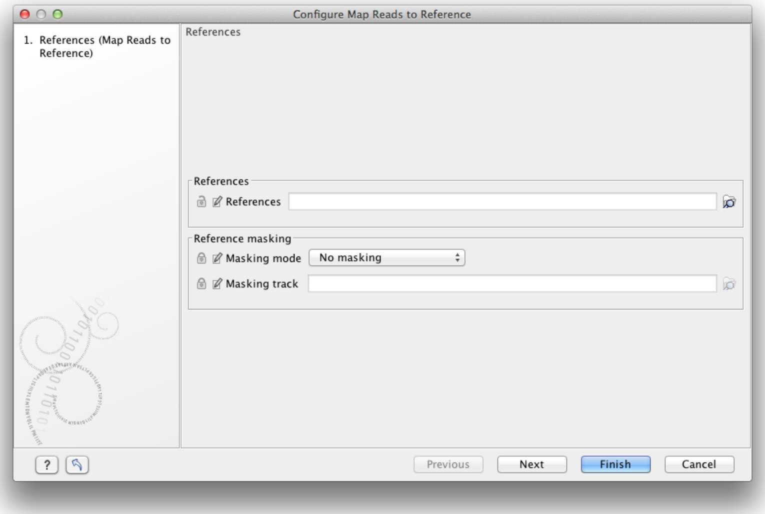
Figure 31.3: *Configuring a tool.*

The first option you are presented with is the option to **Rename** the element. This is for example useful when you wish to discriminate several copies of the same tool in a workflow. The name of the element is also visible as part of the process description when the workflow is executed. Right click on the tool in the workflow and select "Rename" or click on the tool in the workflow and use the F2 key as a shortcut.

With the **Remove** option, elements can be removed from the workflow. The shortcut Alt + Shift + R removes all elements from the workflow.

You can also **Configure** the tool from the right click menu or alternatively it can be done by double-clicking the element. This will open a dialog with options for setting parameters, selecting reference data etc. An example is shown in figure 31.4.

Click through the dialogs using **Next** and press **Finish** when you are done. This will save the



Configure Map Reads to Reference

1. References (Map Reads to Reference)

References

References

References

Reference masking

Masking mode: No masking

Masking track

Figure 31.4: *Configuring read mapper parameters.*

parameter settings that will then be applied when the workflow is executed.

You can also change the name of the parameter into something that fits the vocabulary of the users that are intended to execute the workflow. This is done by clicking the edit icon (✎) and entering a new name.

Note that reference data are a bit special. In the example with the read mapper in figure 31.3, you have to define a reference genome. This is done by pointing to data in the **Navigation Area**. If you distribute the workflow and install it in a different setting where this data is not accessible, the installation procedure will involve defining the new reference data to use (e.g. the reference genome sequence for read mapping). This is explained in more detail in section 31.2.

In some workflows, many elements use the same reference data, and there is a quick way of configuring all these: right-click the empty space and choose **Configure All References**. This will show a dialog listing all reference data needed by the workflow.

The lock icons in the dialog are used for specifying whether the parameter should be locked and unlocked as described in the next section. Hereby it is possible to lock so the workflow runs with the same parameters like references(s) every time.

Once an element has been configured, the workflow element gets a darker color to make it easy to see which elements have been configured.

With **Highlight Subsequent Path** the path from Name of the tool that was clicked on and further downstream will be highlighted whereas all other elements will be grayed out (figure 31.5). The **Remove Highlighting Subsequent Path** reverts the highlighting to the normal workflow layout.

Instead of configuring the various tools individually, the **Configuration Editor** enable specification of all settings, references, masking parameters etc. through a single wizard window (figure 31.6). The Editor is accessed through the (▤) icon located lower left corner.

31.1.3 Locking and unlocking parameters

Figure 31.7 shows the different stages in a workflow.

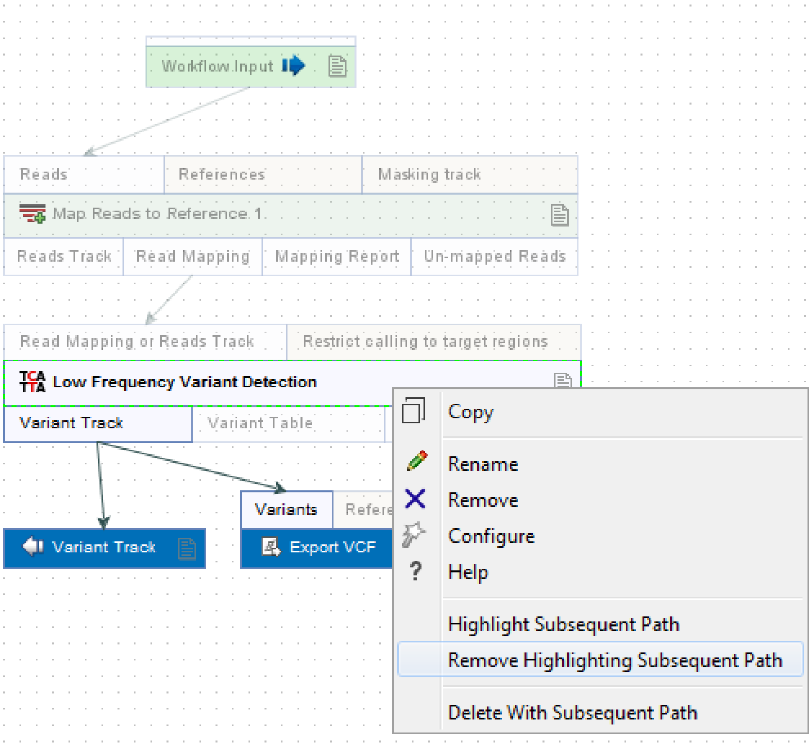


Figure 31.5: *Highlight path from the selected tool and downstream.*

At the top, the workflow creation is illustrated. Workflow creation is explained above. Next, the workflow can be installed in a Workbench or Server (explained in section 31.2). Subsequently, the workflow can be executed as any other tool in the **Toolbox**.

At the creation step, the workflow creator can specify which parameters should be locked or unlocked. If a parameter is locked, it means that it cannot be changed neither in the installation nor the execution step. The lock icons shown in figure 31.4 specifies whether the parameter should be open or locked.

If the parameter is left open, it is possible to adjust it as part of the installation (see section 31.2). Furthermore, it can also be locked at this stage.

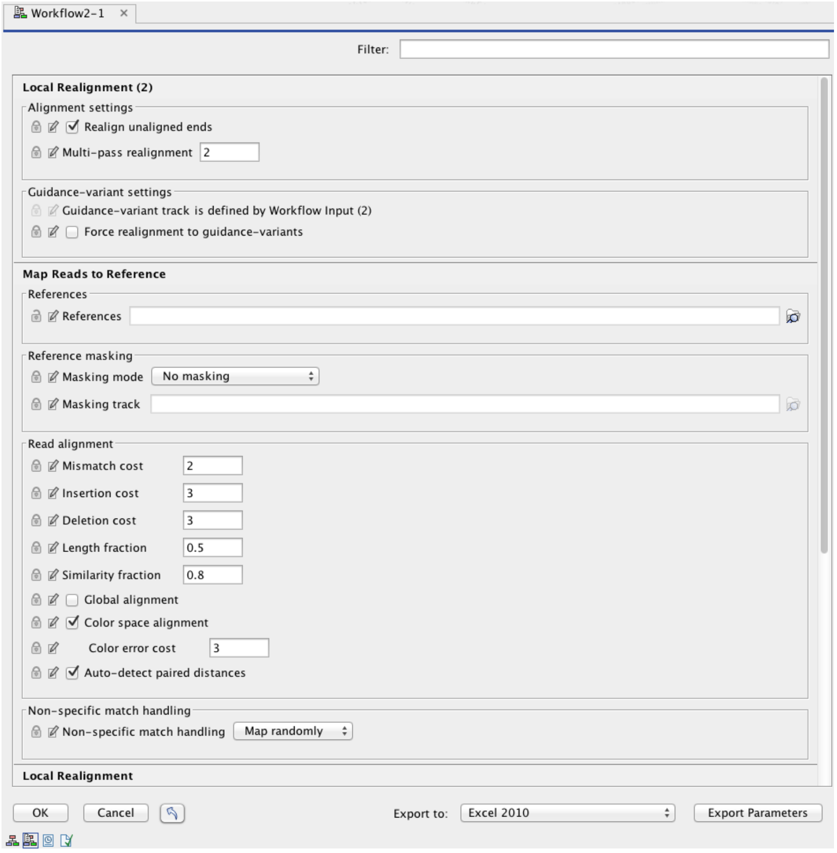
Parameters that are left open both from the workflow creation and installation, will be available for adjustment when the workflow is executed.

Please note that data parameters per default are marked as unlocked. When installing the workflow somewhere else, the connection to the data needs to be re-established, and this is only possible when the parameter is unlocked. Data parameters should only be locked if they should not be set, or if the workflow will only be installed in a setting where there is access to the same data.

31.1.4 Connecting workflow elements

Figure 31.8 explains the different parts of a workflow element.

At the top of each element a description of the required type of input is found. In the right-hand side, a symbol specifies whether the element accepts multiple incoming connections, e.g. +1



Workflow2-1 ×

Filter:

Local Realignment (2)

Alignment settings

☑ Realign unaligned ends

Multi-pass realignment: 2

Guidance-variant settings

Guidance-variant track is defined by Workflow Input (2)

□ Force realignment to guidance-variants

Map Reads to Reference

References

References:

Reference masking

Masking mode: No masking

Masking track:

Read alignment

Mismatch cost: 2

Insertion cost: 3

Deletion cost: 3

Length fraction: 0.5

Similarity fraction: 0.8

□ Global alignment

☑ Color space alignment

Color error cost: 3

☑ Auto-detect paired distances

Non-specific match handling

Non-specific match handling: Map randomly

Local Realignment

Figure 31.6: *The Configuration Editor enable setting of all configurable tools through a single window.*

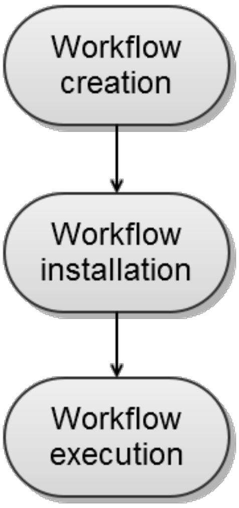


Figure 31.7: *The life cycle of a workflow.*

means that more than one output can be connected, and no symbol means that only one can be connected. At the bottom of each element there are a number of small boxes that represent the different kinds of output that is produced. In the example with the read mapper shown in figure 31.2, the read mapper is able to produce a reads track, a report etc.

Each of the output boxes can be connected to further analysis in three ways:

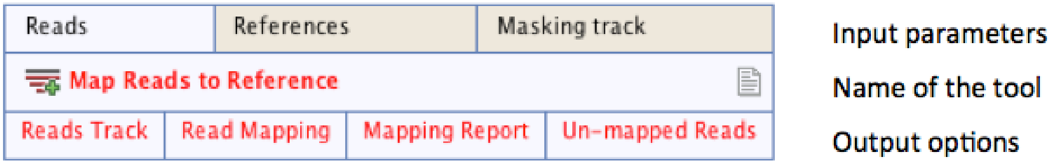


Figure 31.8: *A workflow element consists of three parts: input, name of the tool, and output.*

- By dragging with the mouse from the output into the input box of the next element. This is shown in figure 31.9. A green border around the box will tell you when the mouse button can be released, and an arrow will connect the two elements (see figure 31.10).
- Right-clicking the output box will display a list of the possible elements that this output could be connected to. You can also right-click the input box of an element and connect this to a matching output of another element.
- Alternatively, if the element to connect to is not already added, you can right-click the output and choose **Add Element to be Connected**. This will bring up the dialog from figure 31.1, but only showing the tools that accepts this particular output. Selecting a tool will both add it to the workflow and connect with the output you selected. You can also add an upstream element of workflow in the same way by right-clicking the input box.

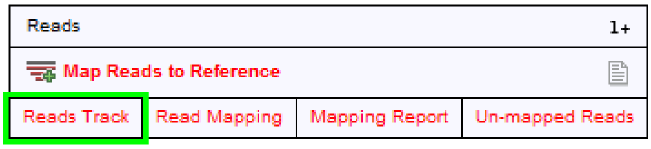


Figure 31.9: *Dragging the reads track output with the mouse.*

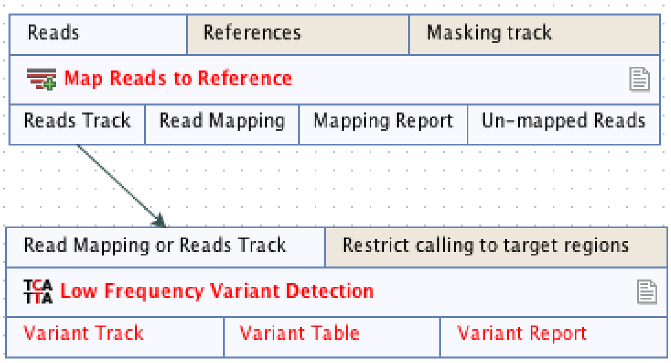


Figure 31.10: *The reads track is now used for variant calling.*

All the logic of combining output and input is based on matching the type of input. So the read mapper creates a reads track and a report as output. The variant caller accepts reads tracks as input but not mapping reports. This means that you will not be able to connect the mapping report to the variant caller.

Figure 31.11 demonstrates how one tool can receive input from two different sources; 1) a reads track that is the input that hold the data that is to be analyzed (in this case reads that is to be locally realigned), and 2) a parameter that can have different functions depending on the tool that it is connected to (in this case the InDel track is used as a guidance track for the local realignment. In other situations the parameter track could be used for e.g. annotation or could provide a reference sequence).

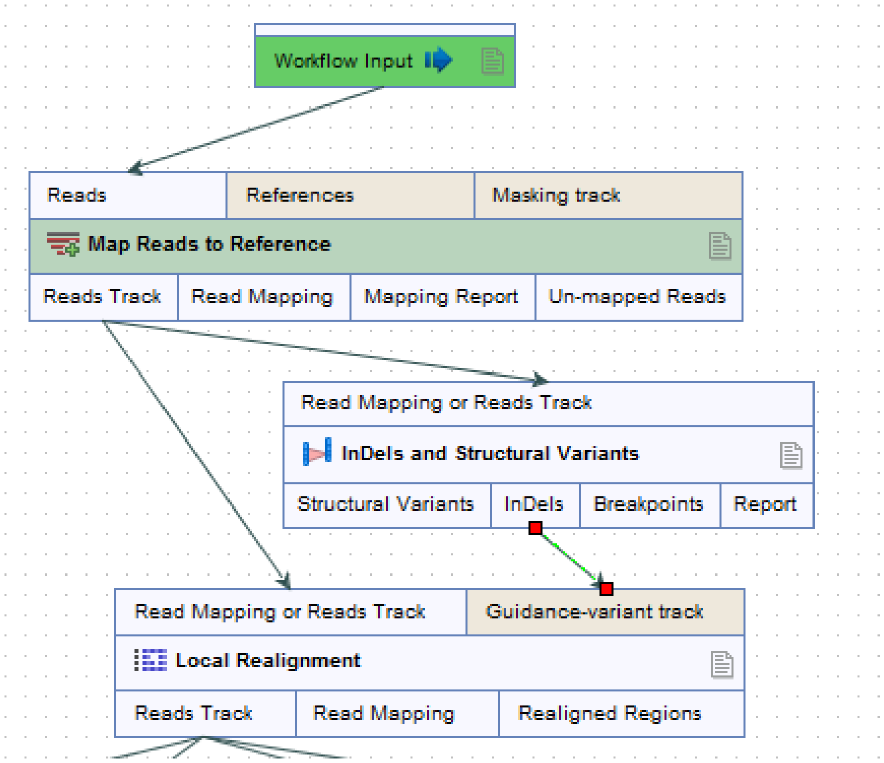


Figure 31.11: *A tool can receive input from both the generated output from another tool (in this example a reads track) and from a parameter (in this case InDels detected with the InDels and Structural Variants tool).*

31.1.5 Input and output

Besides connecting the elements together, you have to decide what the input and the output of the workflow should be. We will first look at specification of the output, which is done by right-clicking the output box of any tool and selecting **Use as Workflow Output** as shown in figure 31.12.

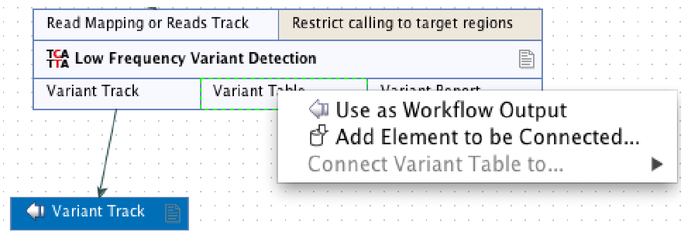
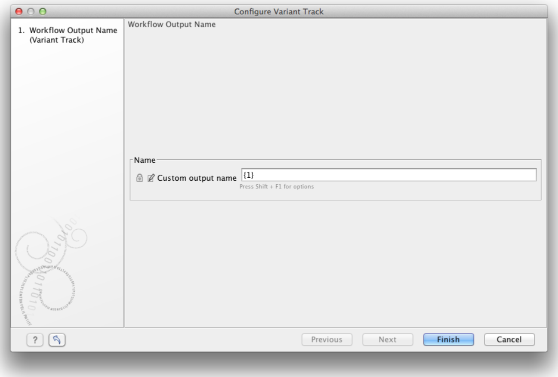

Read Mapping or Reads Track
Restrict calling to target regions
Low Frequency Variant Detection
Variant Track
Variant Table
Variant Report
Use as Workflow Output
Add Element to be Connected...
Connect Variant Table to...
Variant Track

Figure 31.12: *Selecting a workflow output.*

You can mark several outputs this way throughout the workflow. Note that no intermediate results are saved unless they are marked as workflow output[1].

By double-clicking the output box, you can specify how the result should be named as shown in figure 31.13.



Configure Variant Track

1. Workflow Output Name (Variant Track)

Workflow Output Name

Name

Custom output name: {1}

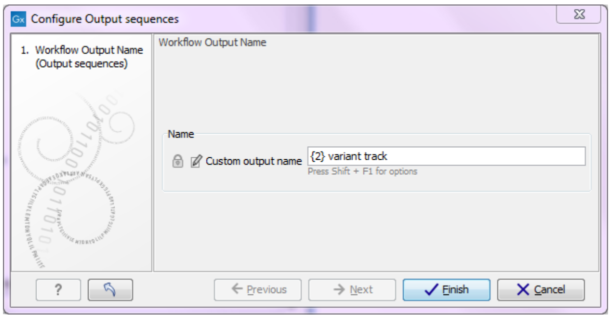
Press Shift + F1 for options

Figure 31.13: *Specifying naming of a workflow output.*

In this dialog you can enter a name for the output result, and you can make use of two dynamic placeholders for creating this name (press Shift + F1 to get assistance):

- {1} Represents the default name of the result. When running the tool outside of a workflow, this is the name given to the result.
- {2} Represents the name of the workflow input (not the input to this particular tool but the input to the entire workflow).

An example of a meaningful name to a variant track could be `{2} variant track` as shown in figure 31.14. If your workflow input is named `Sample 1`, the result would be `Sample 1 variant track`.



Configure Output sequences

1. Workflow Output Name (Output sequences)

Workflow Output Name

Name

Custom output name: {2} variant track

Press Shift + F1 for options

Figure 31.14: *Providing a custom name for the result.*

In addition to output, you also have to specify where the data should go into the workflow by adding an element called **Workflow Input**. This can be done by:

[1] When the workflow is executed, all the intermediate results are indeed saved temporarily but they are automatically deleted when the workflow is completed. If a part of the workflow fails, the intermediate results are not deleted.

- Right-clicking the input box of the first tool and choosing **Connect to Workflow Input**. By dragging from the workflow input box to other input boxes several tools can use the input data directly.
- Pressing the button labeled **Add Element** (or right-click somewhere in the workflow background area and select **Add Element** from the menu that appears). The input box must then be connected to the relevant tool(s) in the workflow by dragging from the Workflow Input box to the "input description" part of the relevant tool(s) in the workflow.

At this point you have only prepared the workflow for receiving input data, but not specified which data to use as input. To be able to do this you must first save the workflow. When this has been done, the button labeled **Run** is enabled which allows you to start executing the workflow. When you click on the button labeled **Run** you will be asked to provide the input data.

Multiple input files can be used when

- Data is generated within the Workflow
- Data is held within the Workbench
- Data is a combination of the two above situations

However, when once the multiple input feature is used it is not possible to run such a workflow in batch mode. It is furthermore possible to Rename the input elements in order discriminate the elements as part of the process description when the workflow is executed.

The example in 31.15 shows how to generated a track list as part of a workflow. It is possible to include reference tracks and also tracks from any step, or multiple steps within the Workflow as long output Read Tracks are defined and linked to.

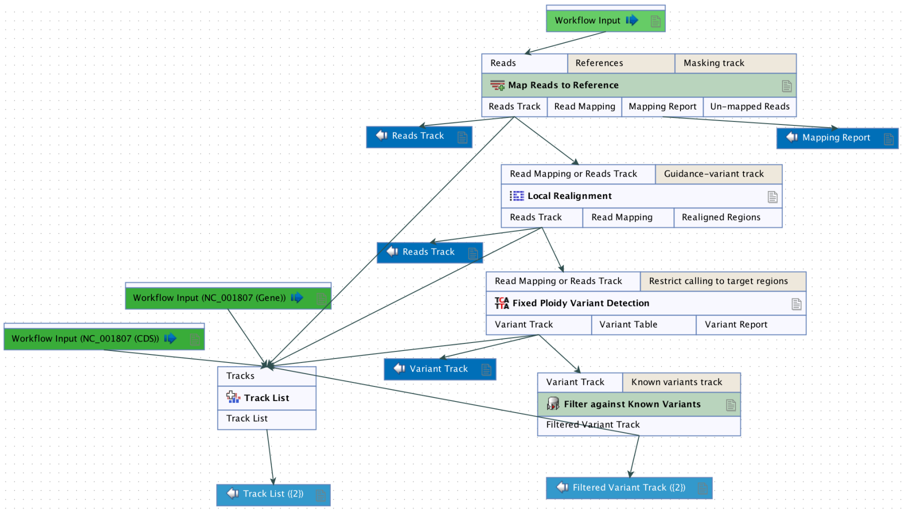


Figure 31.15: *Generation of a track list including data generated within the Workflow, as well as data held in the Workbench.*

31.1.6 Layout

The workflow layout can be adjusted automatically. Right clicking in the workflow editor will bring up a pop-up menu with the option "Layout". Click on "Layout" to adjust the layout of the selected elements (Figure 31.16). Only elements that have been connected will be adjusted.

Note! The layout can also be adjusted with the quick command Shift + Alt + L.

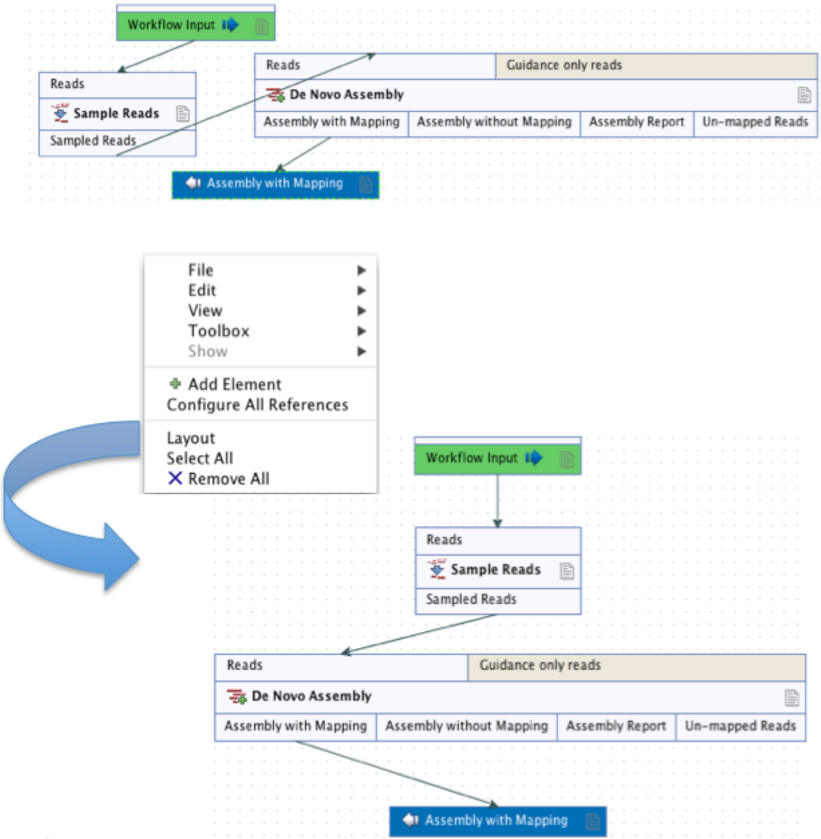


Figure 31.16: *A workflow layout can be adjusted automatically with the "Layout" function.*

Note! It is very easy to make an image of the workflow. Simply select the elements in the workflow (this can be done pressing Ctrl + A, by dragging the mouse around the workflow while holding down the left mouse button, or by right clicking in the editor and then selecting "Select All"), then press the Copy button in the toolbar (⧉) or CTRL + C. Press Ctrl + V to paste the image into the wanted destination e.g. an email or a text or presentation program.

31.1.7 Input modifying tools

An input modifying tool is a tool that manipulates its input objects (e.g. adds annotations) without producing a new object. This behavior differs from the rest of the tools and requires special handling in the workflow.

In the workflow an input modifying tool is marked with the symbol (Ⓜ) (figure 31.17).

Restrictions apply to workflows that contain input modifying tools. For example, branches are not allowed where one of the elements is a modifying tool (see figure 31.18), as it cannot be guaranteed which workflow branch will be executed first, which in turn means that different runs can result in production of different objects. Hence, if a workflow is constructed with a branch where one of the succeeding elements is a modifying tool, a message in red letters will appear saying "Branching before a modifying tool can lead to non-deterministic behavior". In such a

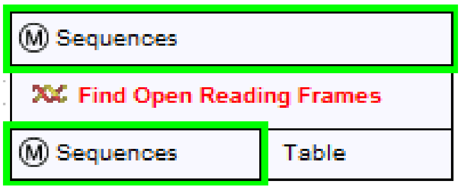
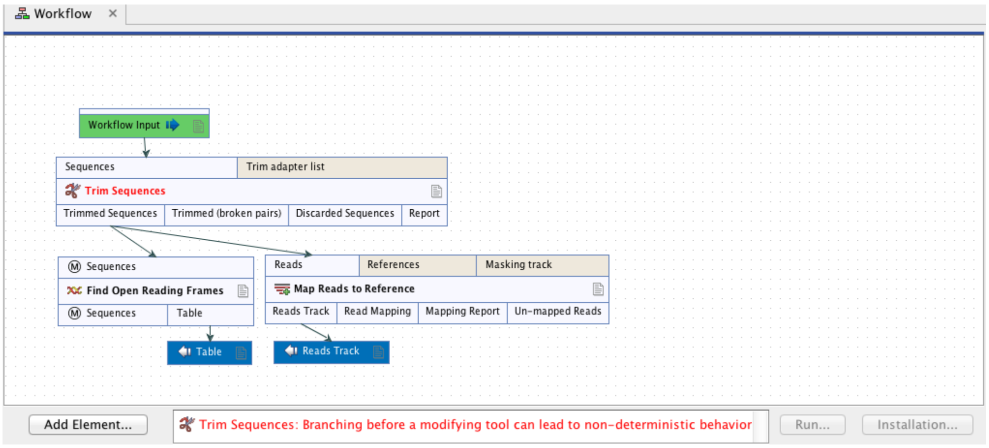


Figure 31.17: *Input modifying tools are marked with the letter M.*

situation the "Run" and "Create Installer" buttons will be disabled (figure 31.18).



Trim Sequences: Branching before a modifying tool can lead to non-deterministic behavior

Figure 31.18: *A branch containing an input modifying tool is not allowed in a workflow.*

The problem can be solved by resolving the branch by putting the elements in the right order (with respect to order of execution). This is shown in figure 31.19 that also shows that the "Run" and "Create Installer" buttons are now enabled. In addition, a message in green letters has appeared saying "Validation successful".

As input modifying tools only modify existing objects without producing a new object, it is not possible to add a workflow output element directly after an input modifying tool (figure 31.20). A workflow output element can only be added when other tools than input modifying tools are included in the workflow.

If the situation occur where more input modifying tools are used succeedingly, a copy of the object will be created in addition to using the modified object as input at the next step of the chain (see figure 31.21). In order to see this output you must right click on the output option (marked with a red arrow in figure 31.21) and select "Use as Workflow Output".

When running a workflow where a workflow output has been added after the first input modifying tool in the chain (see figure 31.22) the output arrow is marked with "copy" to indicate that this is a copy of the result that is used as input at the next level in the chain. When running this workflow you will be able to see the copy of the output from the first input modifying tool in the **Navigation Area** (at the destination that you selected when running the workflow).

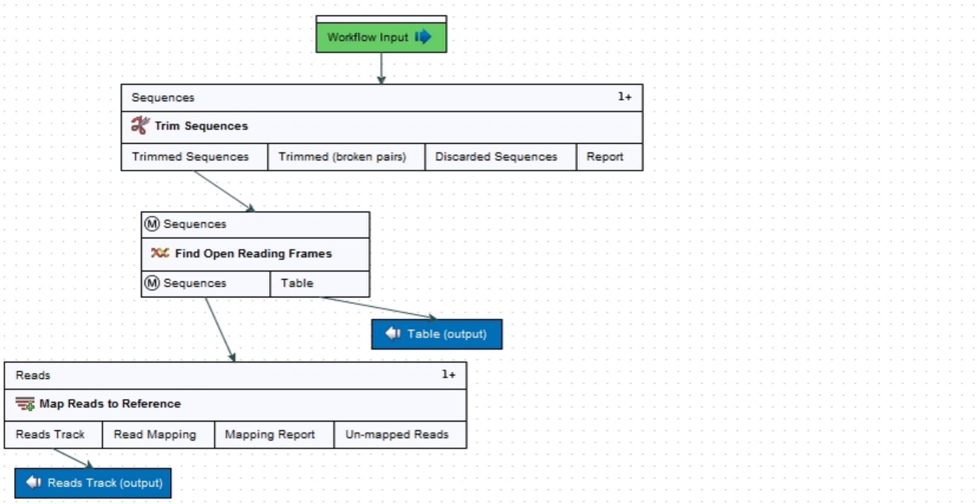


Figure 31.19: *A branch containing an input modifying tool has been resolved and the workflow can now be run or installed.*

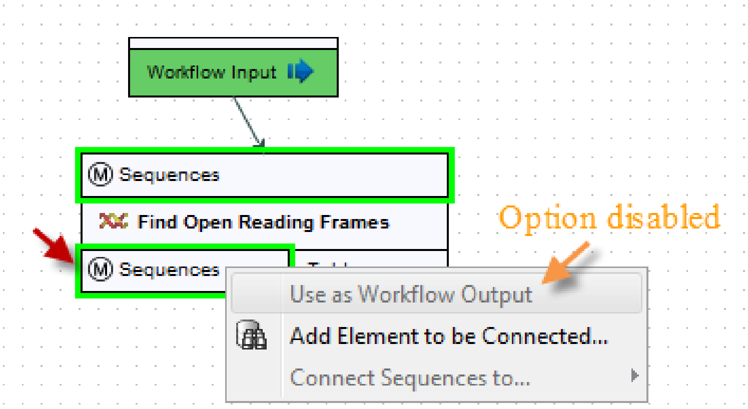


Figure 31.20: *A workflow output element cannot be added if the workflow only contains an input modifying tool.*

31.1.8 Workflow validation

At the bottom of the view, there is a text with a status of the workflow (see figure 31.23). It will inform about the actions you need to take to finalize the workflow.

The validation may contain several lines of text. Scroll the list to see more lines. If one of the errors pertain to a specific element in the workflow, clicking the error will highlight this element.

The following needs to be in place before a workflow can be executed:

- All input boxes need to be connected either to the workflow input or to the output of other tools.
- At least one output box from each tool needs to be connected to either a workflow output or to the input box of another tool.
- Additional checks that the workflow is consistent.

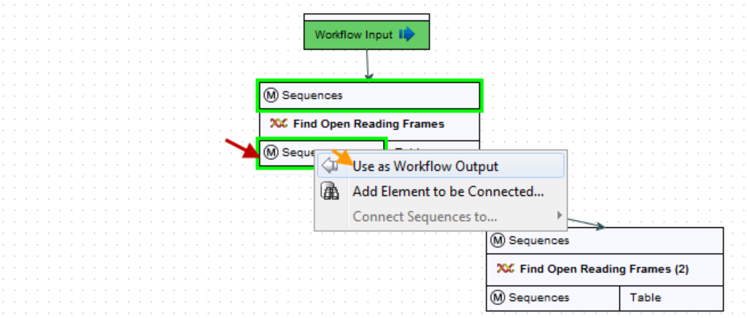


Figure 31.21: *A workflow output element can be added when more than one input modifying tool is used succeedingly (despite that the workflow only contains input modifying tools). Select "Use as Workflow Output" to make a copy of the output.*

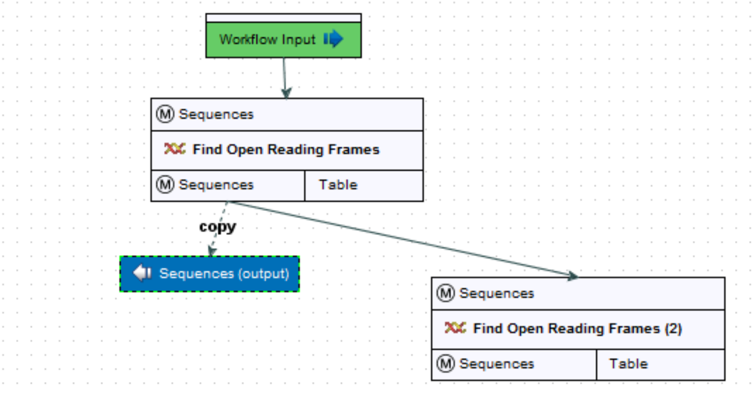


Figure 31.22: *A workflow output element can be added when more than one input modifying tool is used succeedingly (despite that the workflow only contains input modifying tools). Note that this output is marked with "copy" to indicate that this is a copy of the result that is used as input at the next level in the chain.*

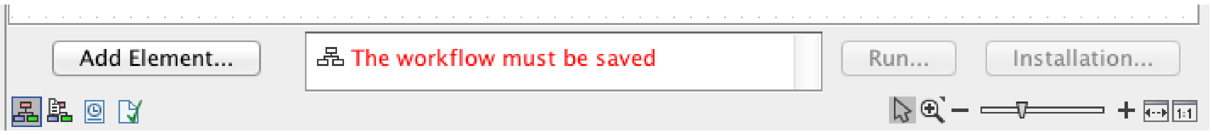


Figure 31.23: *A workflow is constantly validated at the bottom of the view.*

Once these conditions are fulfilled, the status will be "Validation successful", the **Run** button is enabled. Clicking this button will enable you to try running a data set through the workflow to test that it produces the expected results. If reference data has not been configured (see section 31.1.2), there will be a dialog asking for this as part of the test run.

31.1.9 Workflow creation helper tools

In the workflow editor **Side Panel**, you will find the following workflow display settings that can be useful to know (figure 31.24):

Grid

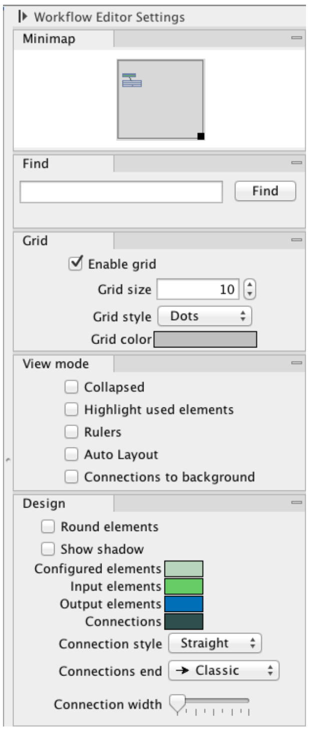
- Enable grid You can display a grid and control the spacing and color of the grid. Per default, the grid is shown, and the workflow elements snap to the grid when they are moved around.

View mode

- Collapsed The elements of the workflow can be collapsed to allow a cleaner view and especially for large workflows this can be useful.
- Highlight used elements Ticking **Highlight used elements** (or using the shortcut Alt + Shift + U) will show all elements that are used in the workflow whereas unused elements are grayed out.
- Rulers Vertical and horizontal rules can be visualised
- Auto Layout Ticking **Auto Layout** will ensure rearrangement of elements once new elements are added.
- Connections to background Connecting arrows are shown behind elements. This may easy reading of element names and accessible parameters.

Design

- Round elements Enable rounding of the element boxes.
- Show shadow Shadows of element boxes can be added.
- Configured elements Background color can be customized.
- Inpupt elements Background color can be customized.
- Edges Color of connecting arrows can be customized.



Workflow Editor Settings

Minimap

Find

Find

Grid

☑ Enable grid

Grid size 10

Grid style Dots

Grid color

View mode

□ Collapsed

□ Highlight used elements

□ Rulers

□ Auto Layout

□ Connections to background

Design

□ Round elements

□ Show shadow

Configured elements

Input elements

Output elements

Connections

Connection style Straight

Connections end → Classic

Connection width

Figure 31.24: *The Side Panel of the workflow editor.*

31.1.10 Adding to workflows

Additional elements can be added to an already existing workflow by dragging it from the navigation area into the workflow editor and joining more elements as necessary. The new workflow must be saved and validated before it can be executed. Two or more workflows can be joined by dragging and dropping one from the Navigation Area, into another that is already open in the main viewing area. The output of one must be connected to the input of the next to allow the whole workflow to run in one go.

Workflows do not need to be valid to be dragged in to the workflow editor, but they must have been migrated to the current version of the workbench.

31.1.11 Supported data flows

The current version of the workflow framework supports single-sample workflows. This means processing one sample through various analysis steps. When it comes to comparative analysis, this has to be done outside the workflow.

A typical example that would explain how this works is a trio analysis study where you want to compare variants found in a child with those from the mother and father. For this, you would create a workflow including mapping, variant detection, variant annotation and maybe some quality control. All three samples would be processed through this workflow in batch mode (see section 31.3). At the end, you can manually create a track list with all the relevant tracks (reads and variants) and run the trio analysis tool manually.

Since all the comparative tools are relatively quick, the bulk of the computation work can usually be incorporated into the workflow which can take care of the more tedious parts of the manual work involved.

CLC bio is planning further improvements to the workflow framework that allows you to model this kind of study as a workflow.

31.2 Distributing and installing workflows

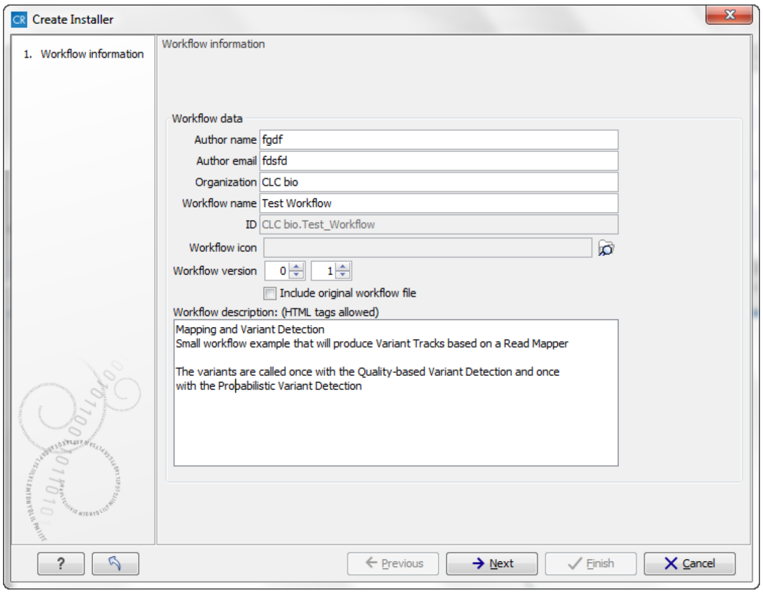
Once the workflow has been configured, you can use the **Run** button (see section 31.1.8) to process data through the workflow, but the real power of the workflow is its ability to be distributed and installed in the **Toolbox** alongside the other tools that come with the *CLC Cancer Research Workbench*, as well as the ability to install the same workflow on a *CLC Genomics Server*. The mechanism for distributing the workflow is a workflow installer file which can be created from the workflow editor and distributed and installed in any Workbench or Server.

31.2.1 Creating a workflow installation file

At the bottom of the workflow editor, click the **Create Installer** button (or use the shortcut Shift + Alt + I) to bring up a dialog where you provide information about the workflow to be distributed (see an example with information from a CLC bio workflow in figure 31.25).

Author information Providing name, email and organization of the author of the workflow. This will be visible for users installing the workflow and will enable them to look up the source of the workflow any time. The organization name is important because it is part of the workflow id (see more in section 31.2.3)

Workflow name The workflow name is based on the name used when saving the workflow in the **Navigation Area**. The workflow name is essential because it is used as part of the workflow



Author name fgdf
Author email fdsfd
Organization CLC bio
Workflow name Test Workflow
ID CLC bio.Test_Workflow
Workflow icon
Workflow version 0 1
□ Include original workflow file
Workflow description: (HTML tags allowed)
Mapping and Variant Detection
Small workflow example that will produce Variant Tracks based on a Read Mapper

The variants are called once with the Quality-based Variant Detection and once
with the Probabilistic Variant Detection

Figure 31.25: *Workflow information for the installer.*

id (see more in section 31.2.3). The workflow name can be changed during the installation of the workflow. This is useful whenever you have a workflow that you would like to use e.g. with small variations. The original workflow name will remain the same in the **Navigation Area** - only the installed workflow will receive the customized name.

ID The final id of the workflow.

Workflow icon An icon can be provided. This will show up in the installation overview and in the **Toolbox** once the workflow is installed. The icon should be a 16 x 16 pixels gif or png file. If the icon is larger, it will automatically be resized to fit 16 x 16 pixels.

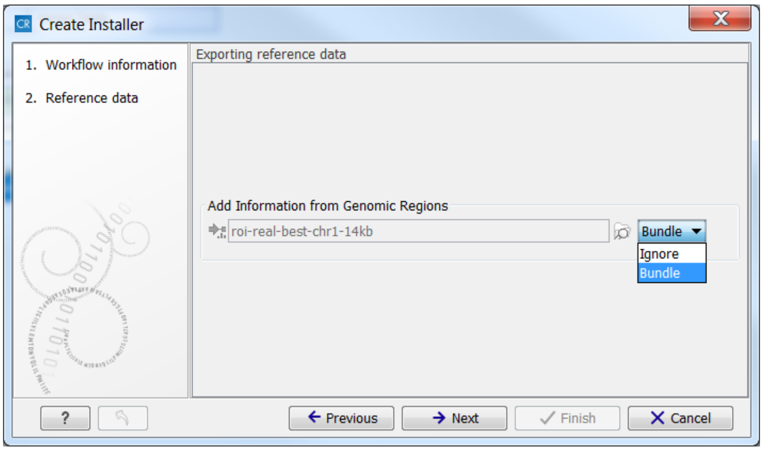
Workflow version A major and minor version can be provided.

Include original workflow file This will include the design file to be included with the installer. Once the workflow is installed in a workbench, you can extract the original workflow file and modify it.

Workflow description Provide a textual description of the workflow. This will be displayed for users when they have installed the workflow. Simple HTML tags are allowed (should be HTML 3.1 compatible, see `http://www.w3.org/TR/REC-html32`).

If you configured any of the workflow elements with data, clicking **Next** will ask you if you want to bundle your workflow installer with this data. For each piece of data, select 'Bundle' to include this data with the workflow distribution, or 'Ignore' to exclude it, leaving it empty (see figure 31.26). **Note!** Bundling data should only be used to bundle small data sets with the workflow installer.

Click **Next** and you will be asked to specify where to install the workflow (figure 31.27). You can install your workflow directly on your local computer. If you are logged on a server and are the administrator, the option "Install the workflow on the current server" will be enabled. Finally, you can select to save the workflow as a .cpw file that can be installed on another computer. Click **Finish**. This will install the workflow directly on the selected destination. If you have selected to



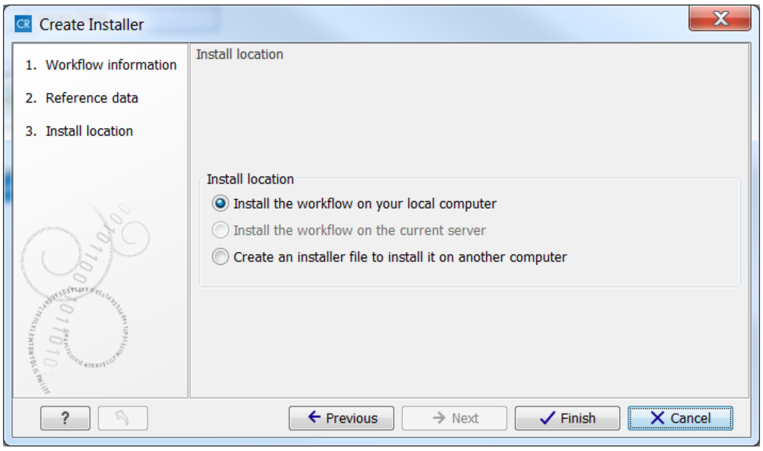
Exporting reference data

Add Information from Genomic Regions

roi-real-best-chr1-14kb: Bundle

Figure 31.26: *Bundling data with the workflow installer.*

save the workflow for installation on another computer, you will be asked where to save the file after clicking **Finish**. If you chose to bundle data with your workflow installation, you will be asked for a location to put the bundled data on the workbench. Installing a workflow with bundled data on a server, the data will be put in a folder created in the first writeable persistence location. Should this location not suit your needs, you can always move it afterwards, using the normal persistence operations.



Install location

Install location

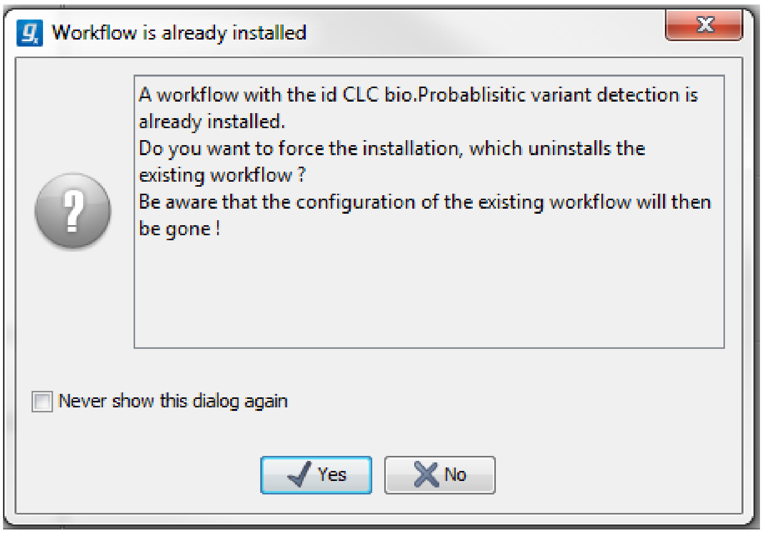
● Install the workflow on your local computer

○ Install the workflow on the current server

○ Create an installer file to install it on another computer

Figure 31.27: *Select whether the workflow should be installed on your local computer or on the current server. A third option is to create an installer file (.cpw) that can be installed on another computer.*

In cases where an existing workflow, that has already been installed, is modified, the workflow must be reinstalled. This can be done by first saving the workflow after it has been modified and then pressing the **Create Installer** button. Click through the wizard and select whether you wish to install the modified workflow on your local computer or on a server. Press **Finish**. This will open a pop-up dialog "Workflow is already installed" (figure 31.28) with the option that you can force the installation. This will uninstall the existing workflow and install the modified version of the workflow. **Note!** When forcing installation of the modified workflow, the configuration of the original workflow will be lost.



A workflow with the id CLC bio.Probablisitic variant detection is already installed.
Do you want to force the installation, which uninstalls the existing workflow ?
Be aware that the configuration of the existing workflow will then be gone !

□ Never show this dialog again

Figure 31.28: *Select whether you wish to force the installation of the workflow or keep the original workflow.*

31.2.2 Installing a workflow

Workflows are installed in the workflow manager (for information about installing a workflow on the *CLC Genomics Server*, please see the user manual at http://www.clcbio.com/usermanuals):

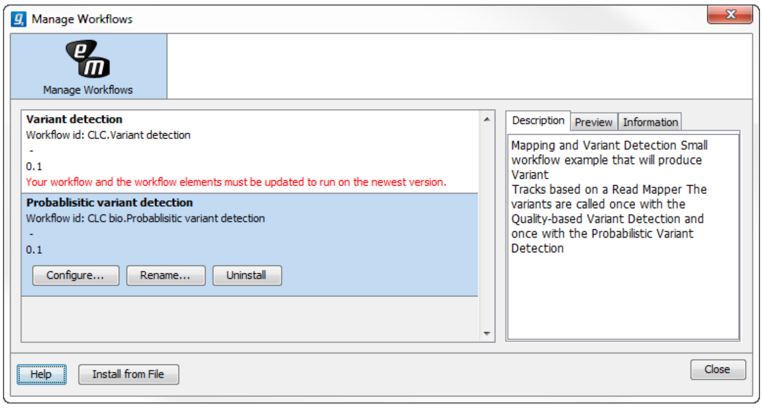
Help | **Manage Workflows** ( )

or press the "Workflows" button ( ) in the toolbar and then select "Manage Workflow..." ( ).

This will display a dialog listing the installed workflows. To install an existing workflow, click **Install from File** and select a workflow .cpw file .

Once installed, it will appear in the workflow manager as shown in figure 31.29.

If the workflow was bundled with data, installing it on the workbench will ask you for a location to put the bundled data. Installing a workflow with bundled data on a server, the data will be put in a folder created in the first writeable persistence location.



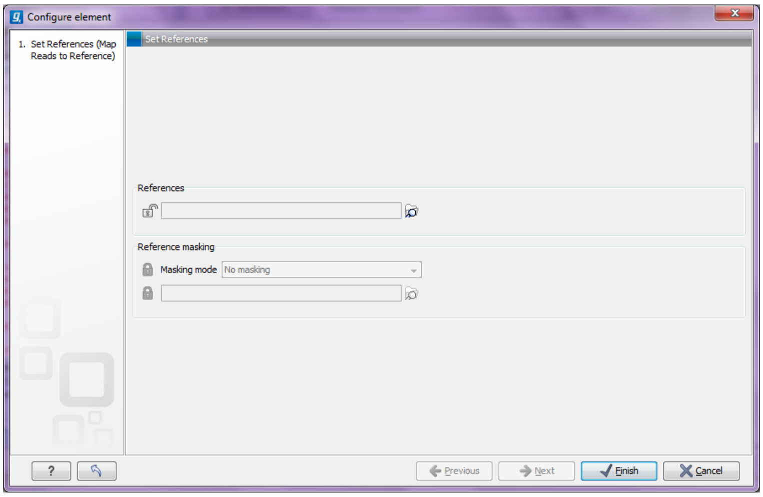
Variant detection
Workflow id: CLC.Variant detection
-
0.1
Your workflow and the workflow elements must be updated to run on the newest version.

Probablisitic variant detection
Workflow id: CLC bio.Probablisitic variant detection
-
0.1

Mapping and Variant Detection Small workflow example that will produce Variant
Tracks based on a Read Mapper The variants are called once with the Quality-based Variant Detection and once with the Probabilistic Variant Detection

Figure 31.29: *Workflows available in the workflow manager. Note the alert on the "Variant detection" workflow, that means that this workflow needs to be updated.*

Click **Configure** and you will be presented with a dialog listing all the reference data that need to be selected. An example is shown in figure 31.30.



Configure element

1. Set References (Map Reads to Reference)

Set References

References

Reference masking

Masking mode: No masking

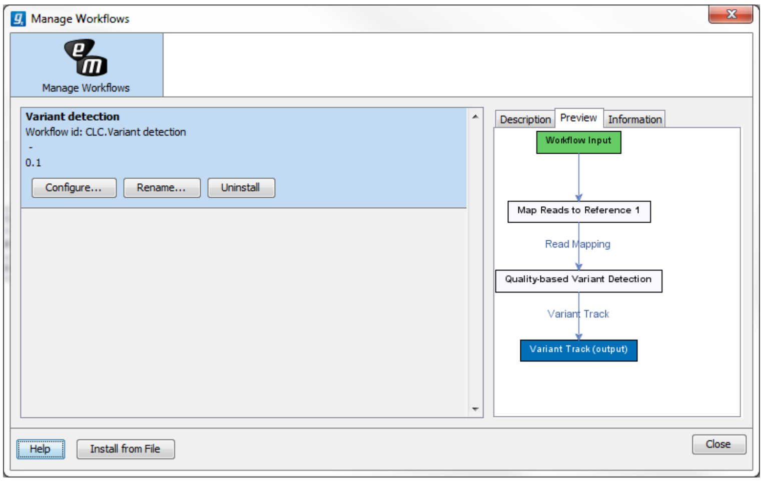
Figure 31.30: *Configuring parameters for the workflow.*

This dialog also allows you to further lock parameters of the workflow (see more about locking in section 31.1.3).

If the workflow is intended to be executed on a server as well, it is important to select reference data that is located on the server.

In addition to the configuration option, it is also possible to rename the workflow. This will change the name of the workflow in the **Toolbox**. The workflow id (see below) remains the same. To rename an element right click on the element name in the Navigation Area and select "Rename" or click on the F2 button.

In the right side of the window, you will find three tabs. **Description** contains the description that was entered when creating the workflow installer (see figure 31.25), the **Preview** shows a graphical representation of the workflow (figure 31.31), and finally you can get **Information** about the workflow (figure 31.32).



Manage Workflows

Variant detection
Workflow id: CLC.Variant detection
-
0.1

Figure 31.31: *Preview of the workflow.*

The "Information" field (figure 31.32) contains the following:

Build id The date followed by the time

Download href The name of the workflow .cpw file

Id The unique id of a workflow, by which the workflow is identified

Major version The major version of the workflow

Minor version The minor version of the workflow

Name Name of workflow

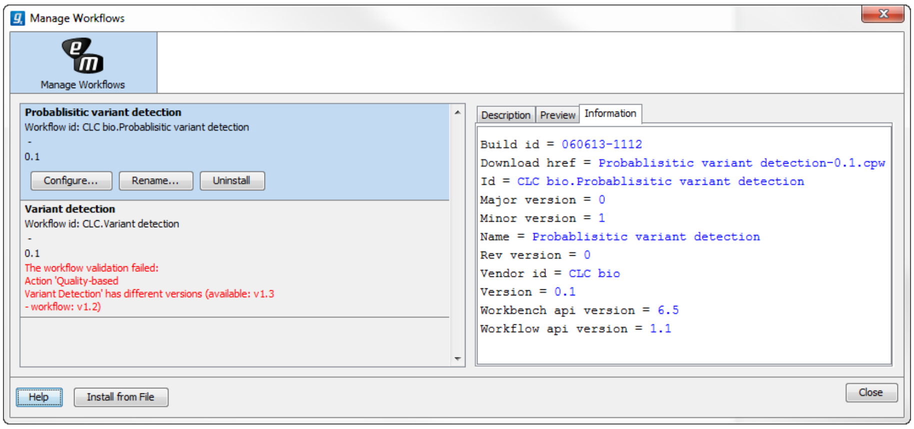
Rev version Revision version. The functionality is activated but currently not in use

Vendor id ID of vendor that has created the workflow

Version <Major version>.<Minor version>

Workbench api version Workbench version

Workflow api version Workflow version (a technical number that can be used for troubleshooting)


Manage Workflows
Manage Workflows
Configure... Rename... Uninstall
Description Preview Information
Help Install from File Close

Probablisitic variant detection
Workflow id: CLC bio.Probablisitic variant detection
-
0.1

Variant detection
Workflow id: CLC.Variant detection
-
0.1
The workflow validation failed:
Action 'Quality-based
Variant Detection' has different versions (available: v1.3
- workflow: v1.2)

```
Build id = 060613-1112
Download href = Probablisitic variant detection-0.1.cpw
Id = CLC bio.Probablisitic variant detection
Major version = 0
Minor version = 1
Name = Probablisitic variant detection
Rev version = 0
Vendor id = CLC bio
Version = 0.1
Workbench api version = 6.5
Workflow api version = 1.1
```

Figure 31.32: *With "Manage Workflows" it is possible to configure, rename and uninstall workflows.*

31.2.3 Workflow identification and versioning

A workflow has a version. The version is used to make it easy to distribute an improved version of the same workflow. To do this, create a new installer with an incremented version number. In order to install a new and updated version, the old one has to be uninstalled.

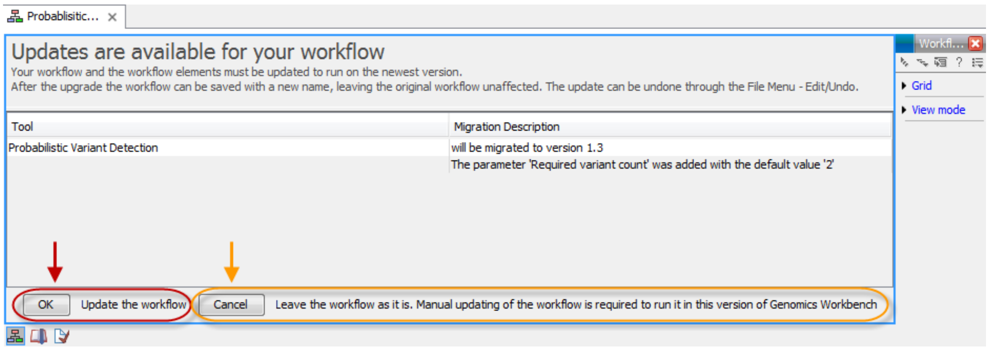
The way the *CLC Cancer Research Workbench* checks whether a workflow already exists in a previous version is by looking at the workflow id. The id is a combination of the organization name and the name of the workflow itself as it is shown in the dialog shown in figure 31.25. Once installed this information is also available in the workflow manager (in figure 31.29 this is `CLC bio.Simple variant detection and annotation-1.2`).

If you create two different workflows with the same name and using the same organization name when creating the installer, they cannot both be installed.

31.2.4 Automatic update of workflow elements

When new versions of the *CLC Cancer Research Workbench* are released, some of the tools that are part of a workflow may change. When this happens, the workflow may no longer be valid. This will happen both to the workflow configurations saved in the **Navigation Area** and the installed workflows.

When a workflow is opened from the **Navigation Area**, an editor will appear, if tools used in the workflow have been updated (see figure 31.33).



Updates are available for your workflow

Your workflow and the workflow elements must be updated to run on the newest version.
After the upgrade the workflow can be saved with a new name, leaving the original workflow unaffected. The update can be undone through the File Menu - Edit/Undo.

Tool	Migration Description
Probabilistic Variant Detection	will be migrated to version 1.3 The parameter 'Required variant count' was added with the default value '2'

OK Update the workflow

Cancel Leave the workflow as it is. Manual updating of the workflow is required to run it in this version of Genomics Workbench

Figure 31.33: *When updates are available an editor appears with information about which tools should be updated. Press "OK" to update the workflow. The workflow must be updated to be able to run the workflow on the newest version of the Workbench.*

Updating a workflow means that the tools in your workflow is updated with the most recent version of these particular tools. To update your workflow, press the **OK** button at the bottom of the page.

There may be situations where it is important for you to keep the workflow in its original form. This could be the case if you have used a workflow to generate results for a publication. In such cases it may be necessary for you to be able to go back to the original workflow to e.g. repeat an analysis.

You have two options to keep the old workflow:

- If you do not wish to update the workflow at all, press the **Cancel** button. This will keep the workflow unchanged. However, the next time you open the workflow, you will again be asked whether you wish to update the workflow. Please note that only updated workflows can run on the newest versions of the Workbench.

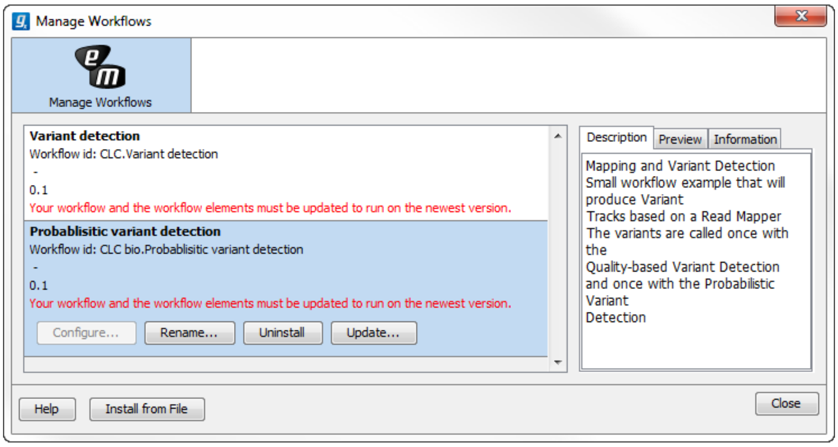
- Another option is to update the workflow and save the updated workflow with a new name. This will ensure that the old workflow is kept rather than being overwritten.

Note! In cases where new parameters have been added, these will be used with their default settings.

If you have used the toolbar "Workflow" button ( ) and "Manage Workflow..." ( ) to access a specific workflow in order to e.g. change the workflow configuration or are going to use the "Install from File" function, a button labeled "Update..." will appear whenever tools have been changed and the workflow needs to be updated (figure 31.34). When you click the button labeled "Update...", your workflow will be updated and the existing workflow will be overwritten.

31.3 Executing a workflow

Once installed and configured, a workflow will appear in the **Toolbox** under **Workflows** ( ). If an icon was provided with the workflow installer this will also be shown (see figure 31.35).



Variant detection
Workflow id: CLC.Variant detection
-
0.1
Your workflow and the workflow elements must be updated to run on the newest version.

Probablisitic variant detection
Workflow id: CLC bio.Probablisitic variant detection
-
0.1
Your workflow and the workflow elements must be updated to run on the newest version.

Mapping and Variant Detection
Small workflow example that will produce Variant
Tracks based on a Read Mapper
The variants are called once with the
Quality-based Variant Detection
and once with the Probabilistic Variant
Detection

Figure 31.34: *Workflow migration.*

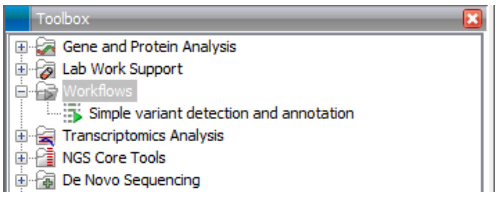


Figure 31.35: *A workflow is installed and ready to be used.*

The workflow is executed just as any other tool in the **Toolbox** by double-clicking or selecting it in the menu (or with the shortcut Ctrl + Enter). This will open a dialog where you provide input data and with options to run the workflow in batch mode (see section 8.1). In the last page of the dialog, you can preview all the parameters of the workflow, as well as the input data, before clicking "Next" to choose where to save the output, and then "Finish" to execute the workflow.

If you are connected to a *CLC Genomics Server*, you will be presented with the option to run the workflow locally on the Workbench or on the Server. When you are selecting where to run the workflow, you should also see a message should there be any missing configurations. There are more details about running Workflows on the Server in the Server manual (http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Workflows.html).

When the workflow is started, you can see the log file with detailed information from each step in the process.

If the workflow is not properly configured, you will see that in the dialog when the workflow is started [2].

[2] If the workflow uses a tool that is part of a plugin, a missing plugin can also be the reason why the workflow is not enabled. A workflow can also become outdated because the underlying tools have changed since the workflow was created (see section 31.2.3)

Chapter 32

Legacy tools

Contents

32.1 Quality-based variant detection

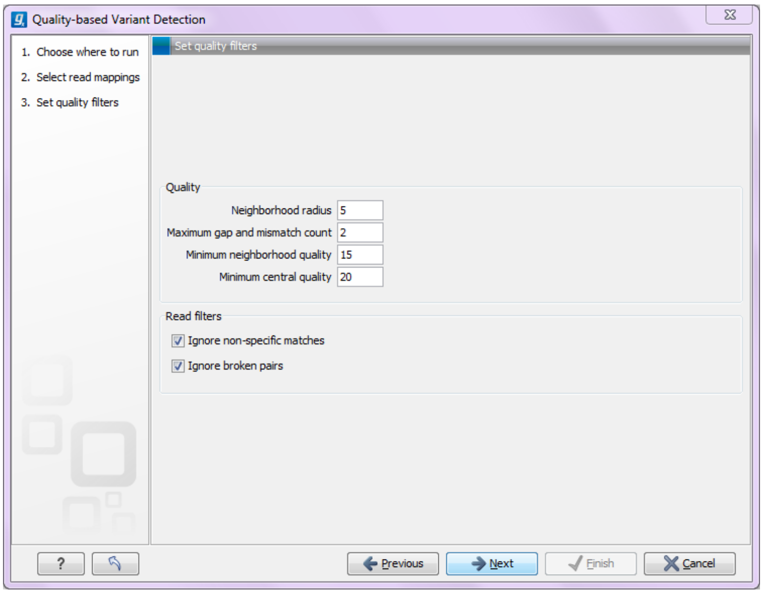
The quality-based variant detection in *CLC Cancer Research Workbench* is based on the *Neighborhood Quality Standard (NQS)* algorithm of [Altshuler et al., 2000] (also see [Brockman et al., 2008] for more information). Using a combination of quality filters and user-specified thresholds for coverage and frequency, this tool finds all variants that are covered by aligned reads.

To run the variant detection:

Toolbox | Legacy Tools ( ) | Quality-based Variant Detection (legacy) ( )

This opens a dialog where you can select mapping results ( )/ ( )/ ( ) or RNA-Seq analysis results ( ).

Clicking **Next** will display the dialog shown in figure 32.1



Quality-based Variant Detection

1. Choose where to run
2. Select read mappings
3. Set quality filters

Set quality filters

Quality

Neighborhood radius	5
Maximum gap and mismatch count	2
Minimum neighborhood quality	15
Minimum central quality	20

Read filters

☑ Ignore non-specific matches

☑ Ignore broken pairs

Figure 32.1: *Quality filtering.*

32.1.1 Assessing the quality of the neighborhood bases

The variant detection will look at each position in the mapping to determine if there is an SNV, MNV, replacement, deletion or insertion at this position.

Variants that are adjacent are reported as one. E.g. two SNVs next to each other will be reported as one MNV. Similarly, an SNV and an adjacent deletion will be reported as one replacement. Note that variants are only reported as one when they are supported by the same reads.

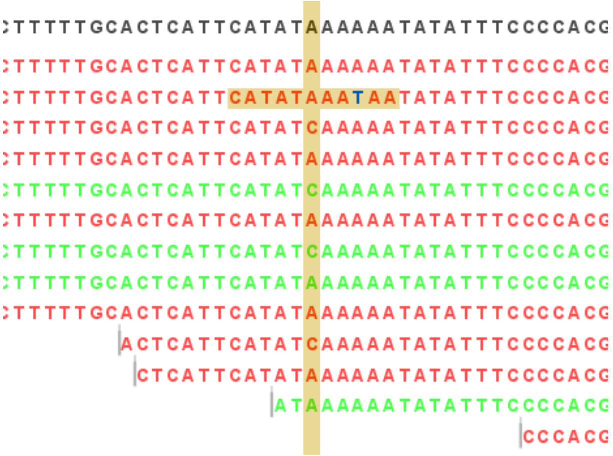
The size of insertions and deletions that can be found depend on how the reads are mapped: Only indels that are spanned by reads will be detected. This means that the reads have to align both before and after the indel. In order to detect larger insertions and deletions, please use the structural variation tool described in section 20.17 instead.

Please note that the variants reported by the structural variation tool can be fed into the local realignment tool (see section 20.7) to re-adjust the alignment of the reads to span the indels, making some of the indels detected by the structural variation ready to be picked up by the quality-based variant detection.

In order to make a qualified assessment, the quality-based variant detection also considers the general quality of the neighboring bases. The **Neighborhood radius** is used to determine how far away from the current variant this quality assessment should extend, and it can be specified in the upper part of the dialog. Note that at the ends of the read, an asymmetric window of the specified length is used.

If the mapping is based on local alignment of the reads, there will be some reads with un-aligned ends (these ends are faded when you look at the mapping). These unaligned ends are not included in the scanning for variants but they are included in the quality filtering (elaborated below).

In figure 32.2, you can see an example with a neighborhood radius of 5. The current position is high-lighted, and the horizontal high-lighting marks the nucleotides considered for a read with the radius set to 5.



```
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAATAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
        ACTCATTCATATCAAAAATATATTTCCCCACG
         CTCATTCATATAAAAAATATATTTCCCCACG
                  ATAAAAAATATATTTCCCCACG
                                  CCCACG
```

Figure 32.2: *An example of a neighborhood radius of 5 nucleotides.*

For each read and within the given radius,[1] the following two parameters are used to assess the quality:

- **Minimum neighborhood quality**. The average quality score of the nucleotides in a read within the specified radius has to exceed this threshold for the base to be included in the calculation for this position (learn more about importing quality scores from different sequencing platforms in section 6.3).
- **Maximum gap and mismatch count**. The number of gaps and mismatches allowed within the window length of the read. Note that this is excluding the "mismatch" or gap that is considered a potential variant. If there are more gaps or mismatches than this threshold within the radius, this read will not be included in the variant calculation at this position. Unaligned regions (the faded parts of a read) also count as mismatches, *even if some of the bases match*.

Note that for sequences without quality scores, the quality score settings will have no effect. In this case only the gap/mismatch threshold will be used for filtering low quality reads.

Figure 32.2 shows an example of a read with a mismatch, marked in dark blue. The mismatch is inside the radius of 5 nucleotides.

When looking at a position near the end of a read (like the read at the bottom in figure 32.2), the window will be asymmetric as shown in figure 32.3.

Besides looking horizontally within a window for each read, the quality of the central base is also examined: **Minimum quality of central base**. This is the quality score for the central base, i.e. the bases in the column high-lighted in figure 32.4. Bases with a quality score below this value are not considered in the variant calculation at this position.

In addition to low-quality reads, reads can also be filtered further:

Ignore non-specific matches This will ignore all reads that are marked as non-specific matches (see section 20.3.3). This is generally recommended, since there is no way of knowing whether the reads and thereby the variant are mapped to the correct position.

[1] The radius is defined as the number of positions in the local alignment between that particular read and the reference sequence (for de novo assembly it would be the consensus sequence).)

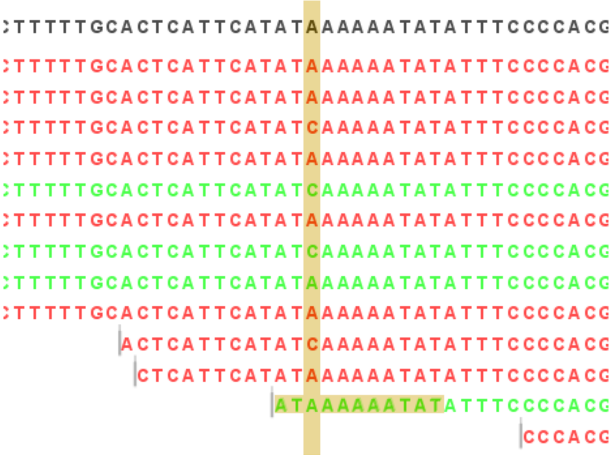


Figure 32.3: *A window near the end of a read.*

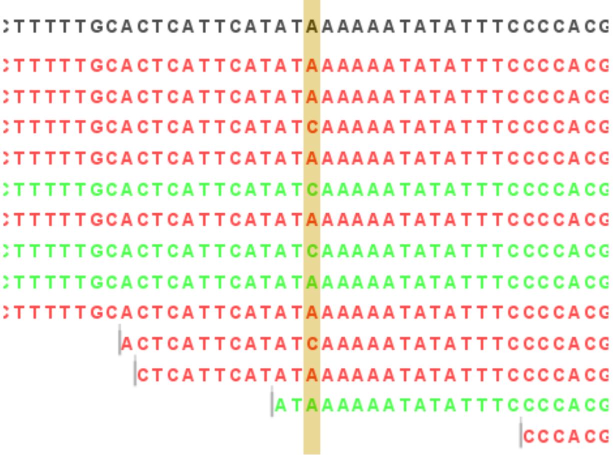


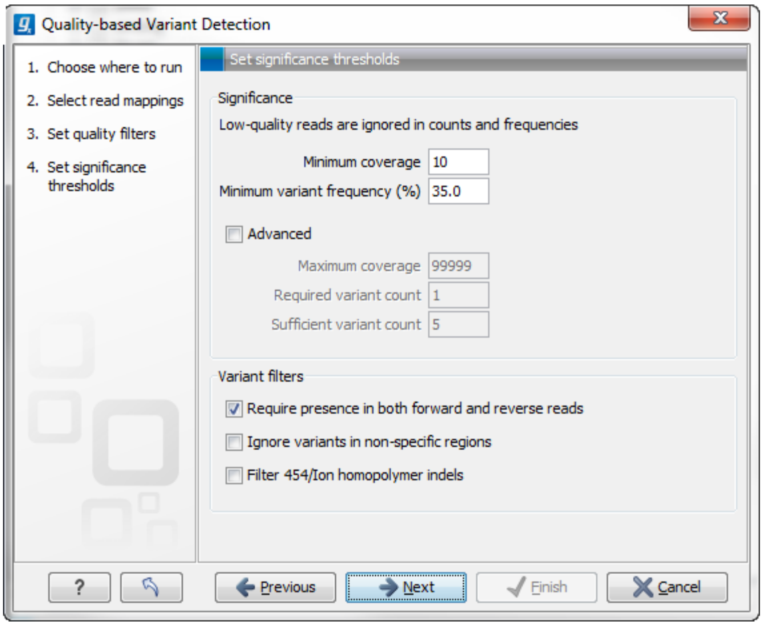
Figure 32.4: *A column of central bases in the neighborhood.*

Ignore broken pairs This will ignore all reads that come from broken pairs (see section 20.3.3). We recommend to switch on the 'Ignore broken reads' filter in case data included paired-reads. As paired-reads have a larger overall alignment with the reference genome, the alignment is more trustworthy than an alignment with a single read, because the probability that the pair could map somewhere else is lower. However, variants in regions with larger deletions, insertions or rearrangements will be ignored, as broken pairs are often indicators for these kinds of events. Note that if you have mapped a combination of single and paired reads, the reads that were marked as single when running the mapping will still be part of the variant detection, even if you have chosen to ignore broken pairs.

Please note that all the filtering described here means that sometime there is a difference between the coverage of the mapping and the actual counts reported for a variant. The difference would be the number of reads that have been filtered before variant calling.

32.1.2 Significance of variant

At a given position, when the reads have been filtered, the remaining reads will be compared to the reference sequence to see if they are different at this position (for *de novo* assembly the consensus sequence is used for comparison). For a variant to be reported, it has to comply with the significance threshold specified in the dialog shown in figure 32.5.



Significance
Low-quality reads are ignored in counts and frequencies
Minimum coverage 10
Minimum variant frequency (%) 35.0
□ Advanced
Maximum coverage 99999
Required variant count 1
Sufficient variant count 5

Variant filters
☑ Require presence in both forward and reverse reads
□ Ignore variants in non-specific regions
□ Filter 454/Ion homopolymer indels

Figure 32.5: *Significance thresholds.*

- **Minimum coverage**. If variants were called in areas of low coverage, you would get a higher amount of false positives. Therefore you can set the minimum coverage threshold. Note that the coverage is counted as the number of valid reads at the current position (i.e. the reads remaining when the quality assessment has filtered out the bad ones).
- **Minimum variant frequency**. This option is the threshold for the number of reads that display a variant at a given position, or in other words, the reported zygosity depends on the setting of the variant frequency parameter. Setting the percentage at 35% means that at least 35% of the validated reads at this position should have a different base than the reference in order to be considered heterozygous rather than homozygous. This means that if, in one reference position, A is represented in more than 35% of the reads and C is also represented in more than 35% of the reads, the variant would be considered heterozygous because two different alleles were called for the same variant. If one of these bases (A and C in this example) is the reference base, then it will be reported in the variant track as the reference allele variant, but not in the annotated table.

Below, there is an **Advanced** option letting you specify additional requirements. These will only take effect if the **Advanced** checkbox is checked.

- **Maximum coverage**. Although it sounds counter-intuitive at first, there is also a good reason to be suspicious about high-coverage regions. Read coverage often displays peaks in repetitive regions where the alignment is not very trustworthy. Setting the maximum coverage threshold higher than the expected average coverage (allowing for some variation in coverage) can be helpful in ruling out false positives from such regions.

  You can see the distribution of coverage by creating a QC for read mapping report (see section 18.3).

  The result table, created by the variant detection, includes information about coverage, so you can specify a high threshold in this dialog, check the coverage in the result afterwards, and then run the variant detection again with an adjusted threshold.

- **Required variant count**. This option is the threshold for the number of reads that display a variant at a given position. In addition to the percentage setting in the simple panel above, this setting is based on absolute counts. If the count required is set to 3, it means that **even though** the required percentage of the reads has a variant base, it will still not be reported if there are less than 3 reads supporting the variant.
- **Sufficient variant count**. This option can be used for deep sequencing data where you have very high coverage and many different alleles. In this case, the percentage threshold is not suitable for finding valid variants only present in a small number of alleles. If the sufficient variant count is set to 5, it means that as long as there are 5 reads supporting a variant, it will be called irrespective of the frequency setting (it still has to be above the required variant count which should always be lower than the sufficient variant count).

When there are **ambiguity** bases in the reads, they will be treated as separate variants. This means that e.g. a `Y` will not be collapsed with `C` or `T` in other reads. Rather, the Ys will be counted separately.

Variant filters

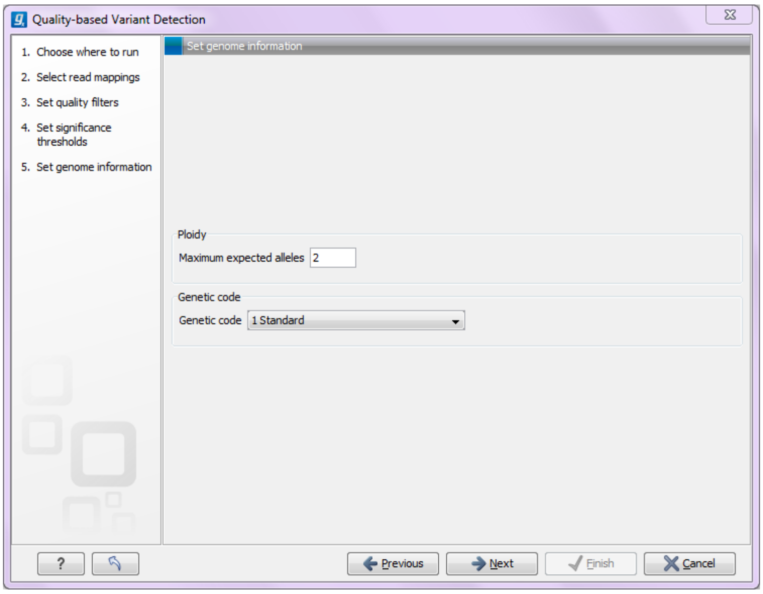
Below the significance settings, there are filters that can be useful for removing false positives:

- **Require presence in both forward and reverse reads**. Some systematic sequencing errors can be triggered by a certain combination of bases. This means that sequencing one strand may lead to sequencing errors that are not seen when sequencing the other strand (see [Nguyen et al., 2011] for a recent study with Illumina data). This can easily lead to false positive variant calls, and by checking this filter, the minimum ratio between forward and reverse reads supporting the variant should be at least 0.05. In this way, systematic sequencing errors of this kind can be eliminated. The forward/reverse reads balance is also reported for each variant in the result (see section 20.18).
- **Ignore variants in non-specific regions**. Variants in regions covered by one or more non-specific reads are ignored.
- **Filter 454/Ion homopolymer indels**. The 454 and Ion Torrent/Proton sequencing platforms exhibit weaknesses when determining the correct number of the same kind of nucleotides in a homopolymer region (e.g. AAA). This leads to a high false positive rate for calling InDels in these regions. This filter is very basic: it removes all indels that are found within or just next to a homopolymer region. A homopolymer region is defined as at least two consecutive identical bases in the reference.

32.1.3 Ploidy and genetic code

Clicking **Next** offers options for setting ploidy and genetic code (see figure 32.7:

- **Maximum expected alleles**. Allows the user to flag variants that fall in locations with an unexpectedly high number of observed alleles. For a given variant, the entry in the 'hyper-allelic' column of the variant table will contain 'yes', if more than the user-specified 'maximum expected alleles' is observed at the variant position, other observations will result in 'no'.



Quality-based Variant Detection

1. Choose where to run
2. Select read mappings
3. Set quality filters
4. Set significance thresholds
5. Set genome information

Set genome information

Ploidy

Maximum expected alleles: 2

Genetic code

Genetic code: 1 Standard

Figure 32.6: *Ploidy and genetic code.*

Note, that with this interpretation the "yes" flag holds true regardless of whether the sequencing data are generated from a population sample or from an individual sample. For example, using a minimum variant frequency of 30% with a diploid organism, you are allowing variants with up to 3 different alleles within the sequencing reads, and by then setting the maximum expected variants count to 2 (the default), any variant with 3 different alleles will be marked as "yes".

- **Genetic code**. For the table report, the variant's effect on the protein level is calculated, and the translation table specified here is used. When reporting the variant as a track, this setting has no effect, since the amino acid consequences are calculated separately (see section 21.6).

32.1.4 Reporting the variants

When you click **Next**, you will be able to specify how the variants should be reported (see figure 32.7).

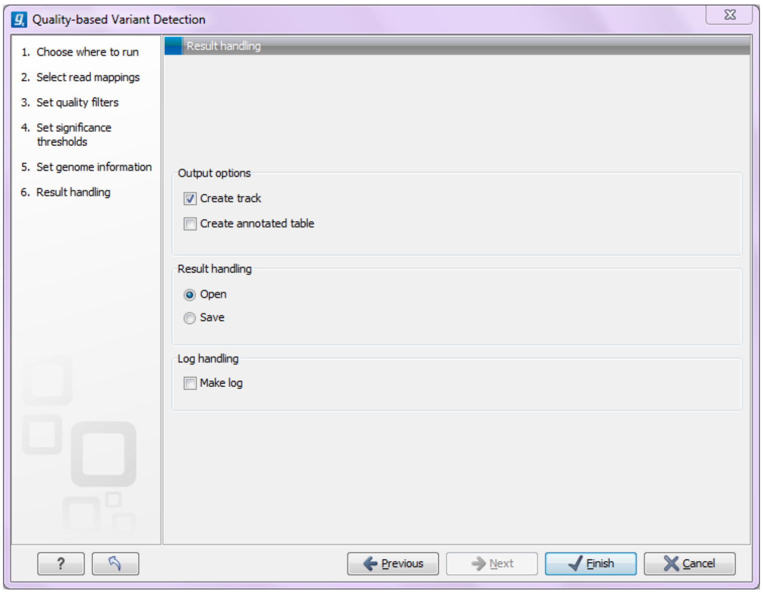
- **Create track.** This will create a variant track that can be further annotated (functional consequences, annotation overlap etc) and used for comparative analysis and visualization.

  Note that the track can be displayed in a table view (▦) as well. See a description of the output in section 20.18.1.

- **Create annotated table.** This will create a table showing all the variants including information about overlapping annotations and amino acid changes. See a description of the output in section 20.18.2.

32.2 Probabilistic variant detection

The purpose of the Probabilistic Variant Caller is to identify variants in a sample by using a probabilistic model built from read mapping data. This tool can detect variants in data sets from



Quality-based Variant Detection

1. Choose where to run
2. Select read mappings
3. Set quality filters
4. Set significance thresholds
5. Set genome information
6. Result handling

Result handling

Output options

☑ Create track

□ Create annotated table

Result handling

● Open

○ Save

Log handling

□ Make log

Figure 32.7: *Output options.*

haploid (e.g. Bacteria), diploid (e.g. Human) and polyploid organisms (e.g. Cancer and higher plants) with a high sensitivity and specificity.

The algorithm used is a combination of a Bayesian model and a Maximum Likelihood approach to calculate prior and error probabilities for the Bayesian model.

Parameters are calculated on the mapped reads alone. The reference sequence is not considered at this stage. After observing a certain combination of nucleotides from the reads at every position in the genome, the probability for each combination of alleles is calculated. These probabilities are then used to determine which one of the allele combinations is the most likely combination for each position. In the case where the ploidy is expected to be 2, the types of cases considered would be homozygous A/A, heterozygous A/G, heterozygous A/C and so on. In the case where the ploidy is expected to be 3, the cases considered would be homozygous A/A/A, heterozygous A/G/C, heterozygous A/C/C and so on. This can then be compared with the reference allele to find out if it is different from the reference sequence and therefore can be called as a variant. Please refer to the white paper at http://www.clcbio.com/white-paper/ for more information including benchmarks.

Variants that are adjacent are reported as one. E.g. two SNVs next to each other will be reported as one MNV. Similarly, an SNV and an adjacent deletion will be reported as one replacement. Note that variants are only reported as one when they are supported by the same reads.

The size of insertions and deletions that can be found depend on how the reads are mapped: Only indels that are spanned by reads will be detected. This means that the reads have to align both before and after the indel. In order to detect larger insertions and deletions, please use the structural variation tool described in section 20.17 instead.

Please note that the variants reported by the structural variation tool can be fed into the local realignment tool (see section 20.7) to re-adjust the alignment of the reads to span the indels, making some of the indels detected by the structural variation ready to be picked up by the probabilistic variant detection.

Probabilistic Variant Detection is not designed to detect low frequency alleles. More specifically, it was not designed for calling variants that are not present in allelic frequencies that are in accordance with the ploidy assumption. Increasing the coverage will lead to higher confidence in the variants that are called, but will not result in the calling of low frequency variants. We recommend using the Low Frequency Variant Detection tool for this purpose (see section 20.13) .

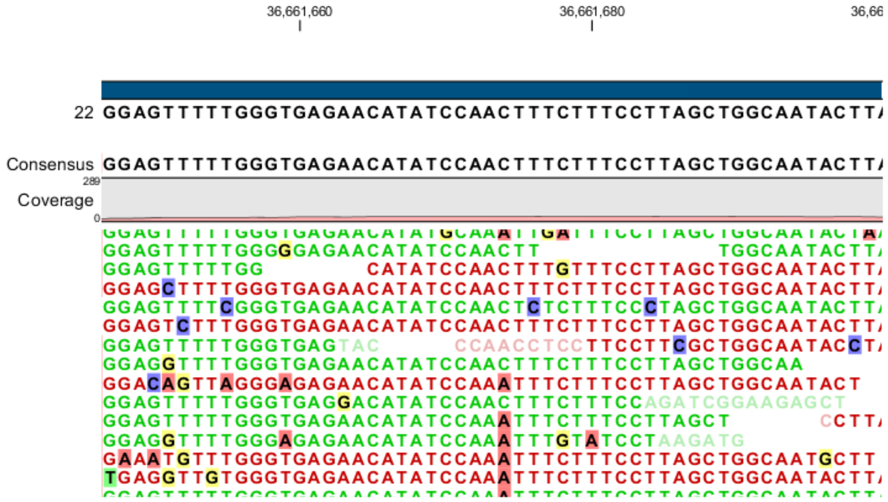

36,661,660
36,661,680
36,66
22 GGAGTTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
Consensus GGAGTTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
289
Coverage
0

Figure 32.8: *An example of a heterozygous variant surrounded by a lot of noise from sequencing errors.*

32.2.1 Calculation of the prior and error probabilities

The prior probabilities are estimated using only the mapped reads through four rounds of Expectation Maximization and are calculated for each potential combination of alleles (site types). Thus, the prior probabilities reflect the likelihood of observing each combination of alleles in the genome studied. The reference sequence is not taken into account during the first part of the analysis. More about the Maximum Likelihood estimation (MLE) can be found at http://en.wikipedia.org/wiki/Maximum_likelihood.

For a diploid organism, the initial parameters for the priors, which are then updated, are shown in Table 32.1. The sum of the probabilities for all site types is always 1.

If the expected ploidy level is set to 1, analogous values to table 32.1 are calculated. Here, only the values for the homozygous site types like A, C, G, T and - would be calculated.

If the expected ploidy is set to 3, the analogous values are calculated, which here would be values for site types like A|A|A, A|C|G, G|G|- and so on.

Error probabilities are calculated alongside the priors for each observed allele and assumed reference allele, before the reference sequence is incorporated into the analysis. Table 32.2 illustrates an example of the values calculated in an error probability matrix.

If quality values are available, an error matrix is calculated for each quality value.

32.2.2 Calculation of the likelihood

After the prior and error probabilities have been estimated, the calculation of the likelihood is undertaken. For every combination of reference allele (site types) and nucleotide in every read,

Site Type	**Prior probability**
A/A	0.2475
A/C	0.001
A/G	0.001
A/T	0.001
T/C	0.001
T/G	0.001
T/T	0.2475
G/C	0.001
C/C	0.2475
G/G	0.2475
G/-	0.001
A/-	0.001
C/-	0.001
T/-	0.001

Table 32.1: Site Types for a diploid organism with example probabilities.

	A	**C**	**G**	**T**	**-**
A	0.90	0.025	0.025	0.025	0.025
C	0.025	0.90	0.025	0.025	0.025
G	0.025	0.025	0.90	0.025	0.025
T	0.025	0.025	0.025	0.90	0.025
-	0.025	0.025	0.025	0.025	0.90

Table 32.2: Error probability matrix - observed sequenced nucleotide in read versus actual nucleotide at this position.

the probability of the observed allele being the same as the reference is calculated. These probabilities are then multiplied for all nucleotides in the reads at that position.

Here is an example:

Assumed reference allele: A/C

Read 1: C [$\frac{1}{2}$ (P(C|A)) + $\frac{1}{2}$(P(C|C))] *

Read 2: C [$\frac{1}{2}$(P(C|A)) + $\frac{1}{2}$(P(C|C))] *

Read 3: A [$\frac{1}{2}$(P(A|A)) + $\frac{1}{2}$(P(A|C))] *

Read 4: A [$\frac{1}{2}$(P(A|A)) + $\frac{1}{2}$(P(A|C))] *

Read 5: T [$\frac{1}{2}$(P(T|A)) + $\frac{1}{2}$(P(T|C))]

Here, P(X|Y) is the probability that we will observe nucleotide X in a read when the true reference sequence is Y.

32.2.3 Calculation of the posterior probability for each site type at each position in the genome

Based on the probabilities calculated, one can determine which of the site types is the best fit at each position in the genome. The site type determined to be the most likely at each position can then be compared with the allele in the reference sequence at the same position. If it is likely to be different, it suggests the presence of a variation.

Therefore the posterior probability is formed as follows:

$$P(site\ type|Obs) = \frac{P(Obs|site\ type) * P(site\ type)}{P(Obs)}$$

where

$$P(Obs) = \sum_{Site\ types} P(Obs|site\ type) * P(site\ type)$$

32.2.4 Comparison with the reference sequence and identification of candidate variants

Once we have all of the probabilities for each combination of alleles for all positions in the reference sequence, the next step is to determine which of them have the highest probability of existing in the sample. These are the candidate variations. Nucleotide combinations that are the same as the reference sequence are not reported. At this point in the algorithm, a probability threshold is taken into consideration, utilizing a threshold provided by the user.

The threshold provided by the user indicates how sure one would like to be that the candidate variant differs from the reference type. The threshold is applied by the Probabilistic Variant Caller by considering the inverse situation: is the probability of the candidate variant being the same as the reference position lower than 1 minus the threshold. So, for a user-provided threshold of 90%, the Probabilistic Variant Caller requires that any given site type has a probability of less than or equal to 0.1 (i.e. 1 - 0.9) of being the same as the reference type. For example, if a user gave a threshold of 90%, and a particular position was found to have a probability of 15%, or 0.15, of being the same as the reference (equivalently, having a probability of 85% of being different than the reference), then this position would not be called as a variant. If the threshold had been set to 80%, then this position would have been called as a variant, as 0.15 is less then 0.20, or in other words, the position has a high enough probability of being different than the reference according to the user-defined threshold, to be reported as a variant.

If a variant is called at a given position, the second step performed by the algorithm is to determines the allele combination (type site) with the highest probability. This type site, together with the corresponding probability, will be reported as the candidate variant.

32.2.5 Posterior filtering and reporting of variants

The algorithm includes several filters to reduce the rate of false positive variants. These filters can be activated or deactivated by the user.

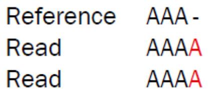
Filtering of variants in homopolymeric regions

Different sequencing platforms generate different types of sequencing errors, which can cause incorrectly called variants. The most common source of sequencing errors across platforms is the determination of nucleotides in so-called homopolymeric regions. These are regions that include stretches of the same nucleotide (e.g. AAAAA or TTTTTTTT). As a result of the internal chemistry used on platforms such as 454 and Ion Torrent, the number of identical nucleotides in such regions is often not accurately reported. This causes variant-callers to identify within homopolymer regions, insertions and deletions not actually present in the sample. The Illumina platform has a similar problem in which one nucleotide is surrounded by other nucleotides of the same type (e.g. AAAAGAAAA). Such cases are sometimes misread, with the different base identified as being the same as the surrounding nucleotides. This can lead to incorrect SNV calls. For example, a region of AAAAGAAAA in the sample may appear as AAAAAAAAA in the read. This could lead to a variant allele, A, being called where the G appears in the reference, when in fact the sample itself did contain a G at that position.

The Probabilistic Variant Caller includes an internal filter to recognize and prevent variants being reported in homopolymeric regions.

The 454/Ion Torrent homopolymer filter does not report insertion or deletion variants found at the ends of regions of two or more nucleotides of the same kind (e.g. AA, TT, GGG).

An example is given in figure 32.9:



Reference	AAA -
Read	AAAA
Read	AAAA

Figure 32.9: *Example of insertions filtered out using the 454/Ion Torrent homopolymer filter.*

The red A will not be reported as a variant when the 454/Ion Torrent filter is applied, as it is characteristic of sequencing errors frequently observed on those platforms.

Forward/reverse reads support

This filter is recommended in all cases where an even distribution of forward and reverse reads at every position is expected. However, it should not be used for data sets such as large amplicons, where the ends of an amplicon are likely to be covered by only forward or reverse reads.

Due to sequencing or PCR artifacts and mapping issues, there can be some positions in the reference genome where only forward or only reverse reads are aligned. This can lead to certain alleles being present on one strand only.

If there is a strand bias from sequencing visible in the quality output check after sequencing, these should be regarded as suspicious regions that should be ignored during variant calling. If the user has selected the forward/reverse read support option, only variants that have a forward/reverse read balance of at least 0.05 are reported.

The forward/reverse balance is calculated as:

$$Min((\#forward/\#total)(\#reverse/\#total))$$

where

#forward = number of forward reads supporting the variant
#reverse = number of reverse reads supporting the variant
#total = all reads supporting the variant

32.2.6 Running the variant detection

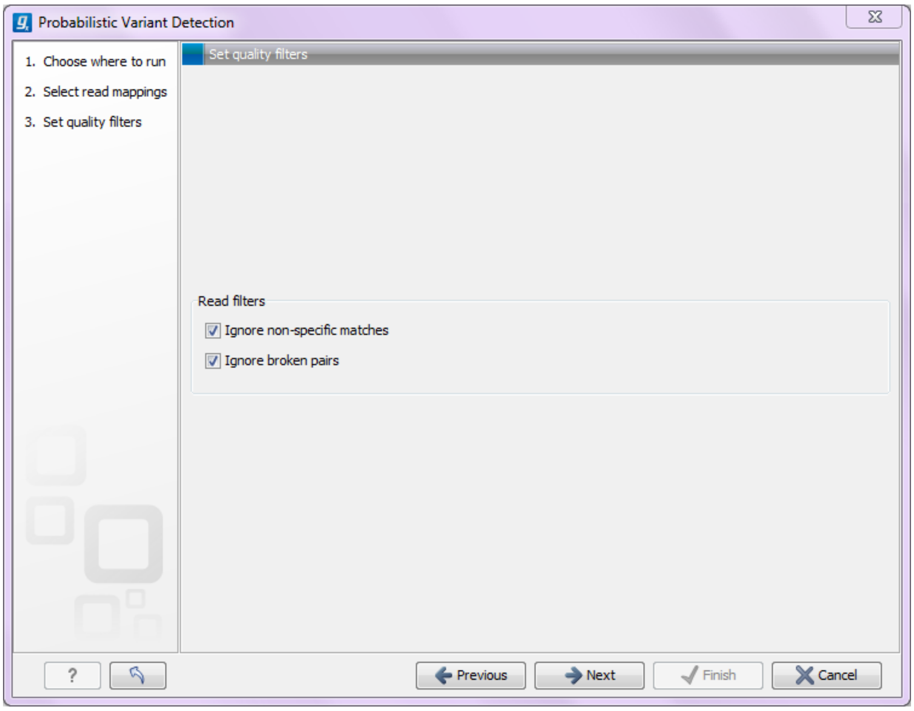
To start the variant calling:

Toolbox | **Legacy Tools** ( ) | **Probabilistic Variant Detection (legacy)** ( )

This opens a dialog where you can select mapping results ( )/ ( )/ ( ) or RNA-Seq analysis results ( ).

Read filters

Clicking **Next** will display the dialog shown in figure 32.10.



Probabilistic Variant Detection

1. Choose where to run
2. Select read mappings
3. Set quality filters

Set quality filters

Read filters

☑ Ignore non-specific matches

☑ Ignore broken pairs

Figure 32.10: *Read filters for the variant detection.*

In this dialog, you can specify reads to be filtered away before variant calling:

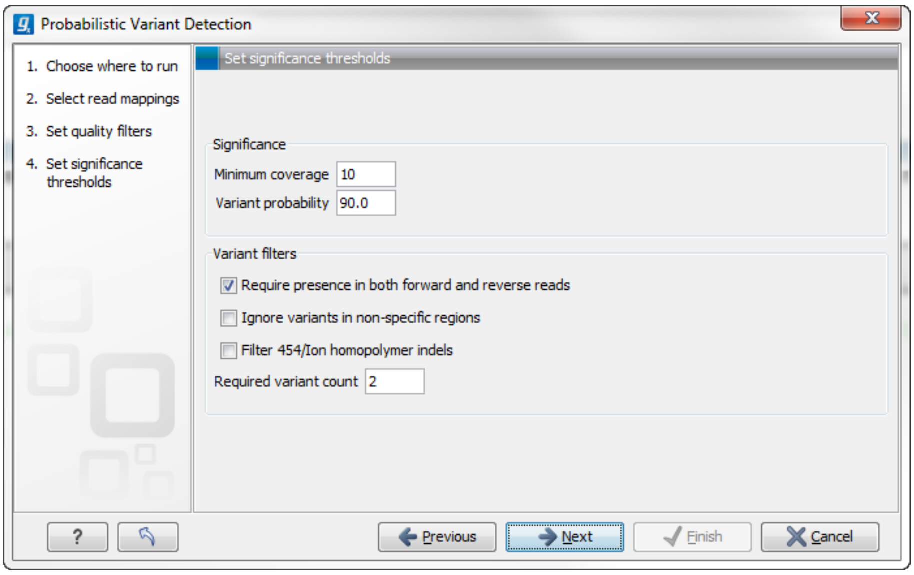
Ignore non-specific matches This will ignore all reads that are marked as non-specific matches (see section 20.3.3). This is generally recommended, since there is no way of knowing whether the reads and thereby the variant are mapped to the correct position.

Ignore broken pairs This will ignore all reads that come from broken pairs (see section 20.3.3). We recommend to switch on the 'Ignore broken reads' filter in case data included paired-reads. As paired-reads have a larger overall alignment with the reference genome, the alignment is more trustworthy than an alignment with a single read, because the probability that the pair could map somewhere else is lower. However, variants in regions with larger deletions, insertions or rearrangements will be ignored, as broken pairs are often indicators for these kinds of events. Note that if you have mapped a combination of single and paired reads, the reads that were marked as single when running the mapping will still be part of the variant detection, even if you have chosen to ignore broken pairs.

Please note that all the filtering described here means that sometime there is a difference between the coverage of the mapping and the actual counts reported for a variant. The difference would be the number of reads that have been filtered before variant calling.

Significance thresholds

Clicking **Next** will display the dialog shown in figure 32.11.



Probabilistic Variant Detection

1. Choose where to run
2. Select read mappings
3. Set quality filters
4. Set significance thresholds

Set significance thresholds

Significance

Minimum coverage 10

Variant probability 90.0

Variant filters

☑ Require presence in both forward and reverse reads

□ Ignore variants in non-specific regions

□ Filter 454/Ion homopolymer indels

Required variant count 2

Figure 32.11: *Significance thresholds.*

The follow parameters can be set:
Significance

- **Minimum coverage** The minimum number of reads aligned to the site to be considered a potential variant.
- **Variant probability** The minimum total probability that a variant is different from the reference for that position to be reported.

Variant filters

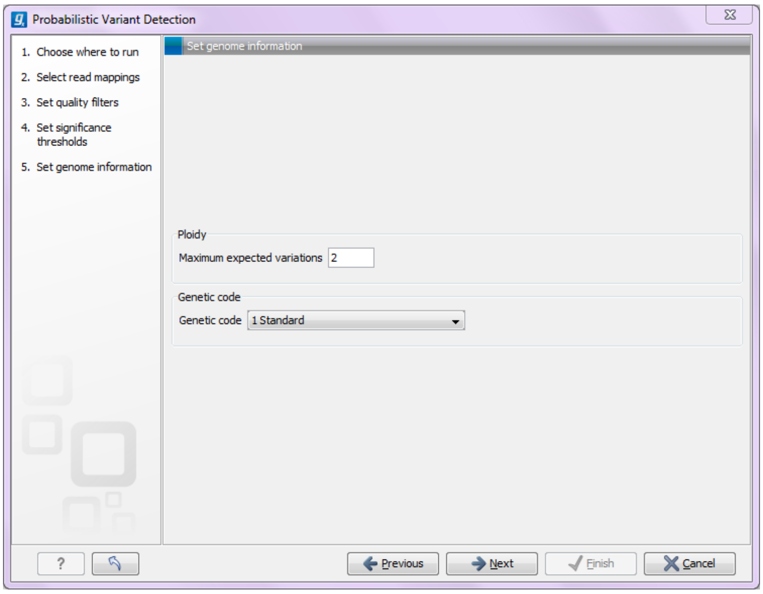
Below the significance settings, there are filters that can be useful for removing false positives:

- **Require presence in both forward and reverse reads**. Some systematic sequencing errors can be triggered by a certain combination of bases. This means that sequencing one strand may lead to sequencing errors that are not seen when sequencing the other strand (see [Nguyen et al., 2011] for a recent study with Illumina data). This can easily lead to false positive variant calls, and by checking this filter, the minimum ratio between forward and reverse reads supporting the variant should be at least 0.05. In this way, systematic sequencing errors of this kind can be eliminated. The forward/reverse reads balance is also reported for each variant in the result (see section 20.18).

- **Ignore variants in non-specific regions**. Variants in regions covered by one or more non-specific reads are ignored.
- **Filter 454/Ion homopolymer indels**. The 454 and Ion Torrent/Proton sequencing platforms exhibit weaknesses when determining the correct number of the same kind of nucleotides in a homopolymer region (e.g. AAA). This leads to a high false positive rate for calling InDels in these regions. This filter is very basic: it removes all indels that are found within or just next to a homopolymer region. A homopolymer region is defined as at least two consecutive identical bases in the reference.
- **Required Variant Count**. This option is the threshold for the number of reads that display a variant at a given position and is based on absolute counts. If the count required is set to 3, it means that even though the required percentage of the reads has a variant base, it will still not be reported if there are less than 3 reads supporting the variant.

32.2.7 Setting ploidy and genetic code

Clicking **Next** offers options for setting ploidy and genetic code (see figure 32.12):



Ploidy

Maximum expected variations 2

Genetic code

Genetic code 1 Standard

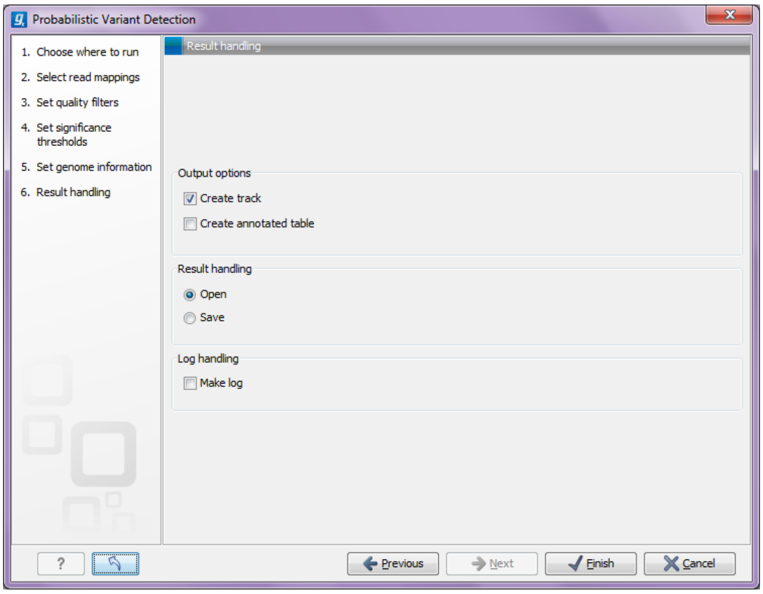
Figure 32.12: *Ploidy and genetic code.*

- **Maximum expected alleles**. This is the ploidy of your organism (or actually "the maximum expected number of alleles"). If set to 1, only homozygous alleles are reported even if another allele is present as well. For cancer samples, which often have a lot of genome duplications, we recommend a setting of 3. For polyploid organisms like plants, a setting of 4 should be used.
- **Genetic code**. For the table report, the variant's effect on the protein level is calculated, and the translation table specified here is used. When reporting the variant as a track, this

setting has no effect, since the amino acid consequences are calculated separately (see section 21.6).

32.2.8 Reporting the variants found

When you click **Next**, you will be able to specify how the variants should be reported (see figure 32.13).



Output options

☑ Create track

□ Create annotated table

Result handling

● Open

○ Save

Log handling

□ Make log

Figure 32.13: *Output options.*

- **Create track.** This will create a variant track that can be further annotated (functional consequences, annotation overlap etc) and used for comparative analysis and visualization.

  Note that the track can be displayed in a table view (▦) as well. See a description of the output in section 20.18.1.

- **Create annotated table.** This will create a table showing all the variants including information about overlapping annotations and amino acid changes. See a description of the output in section 20.18.2.

Part XI

Appendix

Appendix A

Use of multi-core computers

Many tools in CLC Workbenches and Servers can make use of multi-core CPUs. This does not necessarily mean that all available CPU cores are used throughout the analysis. It means that these tools benefit from running on computers with multiple CPU cores.

Tools available differ between CLC Workbenches. In the table, the availability of these tools in different CLC Workbench Toolbox menus is indicated with an X.

Use of multi-core computers	Genomics	Drug Discovery	Cancer Research
Basic Variant Detection	X		X
BLAST (will not scale well on many cores)	X		
Create Alignment	X	X	X
Create Detailed Mapping Report	X		X
Create Sequencing QC Report (will not scale well on more than four cores)	X		
De Novo Assembly	X		
Dock Ligands		X	
Download Reference Genome Data	X		
Extract and Count	X		X
Fixed Ploidy Variant Detection	X		X
Import Molecules from SMILES or 2D		X	
K-mer Based Tree Construction	X		
Large Gap Read Mapper (currently in beta)	X		
Local Realignment	X		X
Low Frequency Variant Detection	X		X
Map Reads to Contigs	X		
Map Reads to Reference	X		X
Maximum Likelihood Phylogeny	X		
Model Testing	X		
Probabilistic Variant Detection (legacy)	X		X
QC for Sequencing Reads (will not scale well on more than four cores)			X
Quality-based Variant Detection (legacy)	X		X
RNA-Seq Analysis	X		X
Screen Ligands		X	
Trim Sequences	X		X

Please note that a static license has a limitation on the maximum number of cores, see section 1.3.1.

Appendix B

Reference data overview

Data	Provider	URL to the original file	Description
Human reference sequence	ENSEMBL	ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens/dna/	Chromosomes 1-22, X, Y and M human reference DNA sequence GRCh37(HG19).
Human genes, coding sequences and transcripts	ENSEMBL	ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/	All annotated protein coding genes for human reference sequence GRCh37(HG19). The annotation was done by ENSEMBL and includes annotations from RefSeq, CCDS as well as ENSEMBL itself.
HapMap variants	ENSEMBL	ftp://ftp.ensembl.org/pub/current_variation/gvf/homo_sapiens/	The goal of the International HapMap Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation (for more information about HapMap see http://hapmap.ncbi.nlm.nih.gov/). Please note that there are 12 different files (tracks) to be downloaded (one file for each population). It is recommended that you configure your workflows with the file from this population that best matches the ethnicity of the patient from which the sample was taken. You can find more about the population codes, which are part of the filename here: http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html.

Data	Provider	URL to the original file	Description
Variants found by the 1000 Genomes Project	ENSEMBL	ftp://ftp.ensembl.org/pub/current_variation/gvf/homo_sapiens/	The 1000 Genomes Project Phase 1 created an integrated map of genetic variations from 1092 human genomes [et al., 2012]. Please note that there are 4 different files (tracks) to be downloaded (one file for each population). It is recommended that you configure your workflows with the file from the population that bests matches the ethnicity of patient from which the sample was taken. You can learn more about the population codes that are part of the filename here: http://www.ensembl.org/Help/Faq?id=328.
COSMIC variants	Sanger Institute	ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/data_export/CosmicMutantExport_*.tsv.gz	The mutation data was obtained from the Sanger Institute Catalogue Of Somatic Mutations In Cancer web site, http://www.sanger.ac.uk/cosmic Bamford et al (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website [Bamford et al., 2004]. The COSMIC database is a human, curated database.
dbSNP variants	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp*.txt.gz	Human variants present in the Single Nucleotide Polymorphism Database (dbSNP), which includes smaller insertions, deletions, replacements, SNPs and MNVs. Please note that most variants in dbSNP are not validated and everybody can submit data to dbSNP. The collection of variants includes clinical relevant as well as common variants.

Data	Provider	URL to the original file	Description
dbSNP variants	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp*Common.txt.gz	Uniquely mapped variants that appear in at least 1% of the population or are 100% non-reference
ClinVar database variants	NCBI	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf/clinvar_00-latest.vcf.gz	ClinVar is designed to provide a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.
PhastCons Conservation Scores	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/	Conservation track of UCSC from a multiple alignments of 100 species and measurements of evolutionary conservation using the phastCons algorithm from the PHAST package.
Human Gene Ontology (GO slim) file	EBI	http://www.ebi.ac.uk/QuickGO/GMultiTerm	Gene Ontology file in slim format (only high level GO terms annotated) for the GO categories Molecular Function, Biological Process and Cellular Component annotated on human genes. The file was made using the QuickGO tool from the EBI (http://www.ebi.ac.uk/QuickGO/ GMultiTerm).

Appendix C

IUPAC codes for amino acids

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.insdc.org/documents/feature_table.html

One-letter abbreviation	Three-letter abbreviation	Description
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
J	Xle	Leucine or Isoleucineucine
L	Leu	Leucine
I	ILe	Isoleucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
O	Pyl	Pyrrolysine
U	Sec	Selenocysteine
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine Asparagine
Z	Glx	Glutamic acid or Glutamine Glutamine
X	Xaa	Any amino acid

Appendix D

IUPAC codes for nucleotides

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.iupac.org and http://www.insdc.org/documents/feature_table.html.

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Appendix E

Formats for import and export

E.1 List of bioinformatic data formats

Below is a list of bioinformatic data formats, i.e. formats for importing and exporting molecule structures, sequences, alignments and trees.

E.1.1 Molecule structure formats

File type	Suffix	Import	Export	Description
PDB	.pdb	X		
Tripos Mol2	.mol2	X	X	
MDL Mol	.sdf	X		
CLC	.clc	X	X	Rich format including all information

E.1.2 Sequence data formats

File type	Suffix	Import	Export	Description
Annotation CSV export	.csv		X	Annotations in csv format
CLC	.clc	X	X	Rich format including all information
FASTA	.fsa/.fasta	X	X	Simple format, name & description
GCG sequence	.gcg	X		Rich information incl. annotations
Raw sequence	any	X		Only sequence (no name)
Sequence CSV	.csv	X	X	Simple format. One seq per line: name, description(optional), sequence
Tab delimited text	.txt		X	Annotations in tab delimited text format
PIR(NBRF)	.pir	X	X	Simple format, name and description
Swiss-Prot	.swp	X		Rich information incl. annotations (only peptides)

Note that high-throughput sequencing data formats from Illumina, SOLiD, IonTorrent, 454 and also high-throughput fasta and trace files are imported using a special import as described in section 6.3. These data can also be exported in fastq format (using NCBI/Sanger Phred quality scores).

When exporting in fasta format, it is possible to remove sequence ends covered by annotations of type "Trim" (read more in section 29.2).

E.1.3 Read mapping formats

File type	Suffix	Import	Export	Description
ACE	.ace	X	X	No chromatogram or quality score
AGP	.agp/.fa		X	Exports scaffolded contigs (see below)
BAM	.bam	X	X	Compressed version of SAM. See details in section 6.3.9
CLC	.clc	X	X	Rich format including all information
CLC Assembly File	.cas	X		Output from the CLC Assembly Cell
SAM	.sam	X	X	Sequence Alignment/Map. See details in section 6.3.9
Mapping coverage	.tsv		X	Detailed per-base info on coverage (see below)

Special note about AGP format Both sequence lists and contigs with reads mapped can be used. Based on annotations of type **Scaffold** (which are automatically added when running the *de novo* assembly with the scaffold option), the contigs are broken up before exported as fasta. The agp file produced holds information about how the contigs relate to each other.

Export of coverage information from sequence alignments Coverage information from read mappings can be exported in a tabular format using the **Mapping coverage** export. The output contains information on the number of nucleotides aligned to positions in reference sequences. Insertions are also reported as described below while deletions are reported as reference regions without read coverage. Both stand-alone read mappings and read tracks can be used as input.

The exported file contains the following columns:

Column	Description
1	Reference name
2	Reference position
3	Reference sub-position (insertion)
4	Reference symbol
5	Number of A's
6	Number of C's
7	Number of G's
8	Number of T's
9	Number of N's
10	Number of Gaps
11	Total number of reads covering the position

The **Reference sub-position** column is empty (indicated by a - symbol) when the reference is defined at a given position. In case of an insertion this column contains an index into the insertion (a number between 1 and the length of the insertion) while the **Reference symbol** column is empty and the **Reference position** column contains the position of the last reference.

E.1.4 Contig formats

File type	Suffix	Import	Export	Description
ACE	.ace	X	X	No chromatogram or quality score
CLC	.clc	X	X	Rich format including all information

E.1.5 Alignment formats

File type	Suffix	Import	Export	Description
Aligned fasta	.fa	X	X	Simple fasta-based format with – for gaps
CLC	.clc	X	X	Rich format including all information
ClustalW	.aln	X	X	
GCG Alignment	.msf	X	X	
Nexus	.nxs/.nexus	X	X	
Phylip Alignment	.phy	X	X	

E.1.6 Expression data formats

Read about technical details of these data formats in section G.

File type	Suffix	Import	Export	Description
Affymetrix CHP	.chp/.psi	X		Expression values and annotations
Affymetrix pivot/metric	.txt/.csv	X		Gene-level expression values
Affymetrix NetAffx	.csv	X		Annotations
CLC	.clc	X	X	Rich format including all information
Excel	.xls/.xlsx		X	All tables and reports
Generic	.txt/.csv	X		Expression values
Generic	.txt/.csv	X		Annotations
GEO soft sample/series	.txt/.csv	X		Expression values
Illumina	.txt	X		Expression values and annotations
Table CSV	.csv		X	Samples and experiments
Tab delimited	.txt		X	Samples and experiments

E.1.7 Annotation and variant formats

Please note that all of the annotation and variant formats can be imported as tracks (see section 6.2). GFF, GVF and GTF formats can also be imported as annotations on a standard sequence or sequence list using functionality provided by the Annotate with GFF plugin (http://www.clcbio.com/clc-plugin/annotate-sequence-with-gff-file/).

File type	Suffix	Import	Export	Description
VCF	.vcf	X	X	See note below
GFF	.gff	X	X	To import as annotation track, see section 6.2. To annotated sequence or sequence list, see plugin: http://www.clcbio.com/clc-plugin/annotate-sequence-with-gff-file/
GVF	.gvf	X	X	Special version of GFF for variant data. See GFF entry above.
GTF	.gtf	X	X	Special version of GFF for gene annotation data. See GFF entry above.
COSMIC variation database	.tsv	X		Special format for COSMIC data
BED	.bed	X		See section 6.2
Wiggle	.wig	X		See section 6.2
UCSC variant database table dump	.txt	X		See section 6.2
Complete genomics master var files	masterVar	X		Complete genomics variant data format

Special note on VCF export

For VCF export, counts from the variant track are put in CLCAD2 tags and coverage in DP tags. The values of the CLCAD2 tag follow the order of REF and ALT, with one value for the REF and for each ALT. For example if there has been a homozygote variant identified at a certain position, the value of the GT field is 1/1 and the corresponding CLCAD2 value for the reference allele will be 0, which is always the first number in the CLCAD2 field. Please note that this does not mean the original *mapping* did not have any reads with that sequence, but it means that the *variant track* being exported does not contain the reference allele.

When exporting VCF files, there are three options:

Reference sequence track Since the VCF format specifies that reference and allele sequences cannot be empty, deletions and insertions have to be padded with bases from the reference sequence. The export needs access to the reference sequence track in order to find the neighboring bases.

Enforce diploid export The *CLC Cancer Research Workbench* option will generate a VCF file in which the allele values in the Genotype (GT) field for haploid variants are reported following the format for diploid variants (i.e. the GT allele values reported are 1/1). This is to ensure compatibility of the exported VCF file with programs for downstream variant analysis that expect strictly diploid genomes. The user can specify that the Enforce diploid option is only applied to certain chromosomes, while others may be reported as haploid. If you export a variant track that has been filtered, there can be situations where there is only one heterozygous variant at a given position. In this case, the *CLC Cancer Research Workbench* will use a "." to denote an unknown genotype, so the GT field will be "1/.".

Note: the "Enforce diploid" option does NOT enforce diploidy for polyploid variant loci. Regardless of this setting, all variant alleles reported during variant calling are included in the exported VCF file.

It is important to note that this **Enforce diploid export** option will create a diploid format of the VCF file, but it is not able to recover any inconsistencies in the variant track used as input. If the variant track has three variants at a given position, three genotypes will be output. Or if the variant track has two variants at the same position that both postulate to be homozygous, they will be output as two heterozygous variants. When exporting data created by the variant callers of *CLC Cancer Research Workbench*, this is usually not a problem, but when applying this diploid scheme to data that has been imported into the *CLC Cancer Research Workbench* from other sources, the data can be inconsistent with a diploid model.

Exceptions Some chromosomes can be excepted from the enforced diploid export. For a human genome, that would be relevant for the mitochondrion and for male X and Y chromosomes. For this option, you can select which chromosomes should be excepted. They will be exported in the standard way without assuming there should be two genotypes, and homozygous calls will just have one value in the GT field.

Special note on former VCF export

In CLC Genomics Workbench 6.5 instead of the CLCAD2, the CLCAD field had been reported. The difference between CLCAD and CLCAD2 is that the former is following the order in the GT (genotype) field in VCF, while the latter is following the order of the REF and ALT fields in VCF in is therefore more in line with the AD field reported from GATK and other sources.

Special notes on VCF import Note! Please also see section 6.2.

The import process for VCF files into the CLC Genomics Workbench currently work as follows:

1. For VCF rows that are reporting the reference base no variants are imported
2. In cases where GT = 0/0, GT=./., GT=0/. or GT=./0 no variants are imported at all
3. In cases where GT = X/. or GT = ./X , and where X is not zero, a single variant is imported depending on the actual value of X
4. In cases where GT = X/X and X is not zero, in Genomics Workbench 6.5 this will result in two independent variants. In version 6.5.1 they will be reported as a single homozygous variant
5. In cases where GT = X/Y and X and Y are different but either one may be zero, two independent variants are created

Please note that some replacements cannot be interpreted in versions that are older than CLC Genomics Workbench 7.0; such replacements will therefore not be imported in previous versions of *CLC Cancer Research Workbench*.

An example of these types of replacements are the following:

chr2 32843292 . TTTA T,TTT 100 PASS DP=44

Due to the VCF interpretation, the initial T base is removed from all alternatives.

In version 6.5.x, only the reference variant (TTA -> TTA) and the first deletion in ALT (TTA -> -) will be imported. The replacements TTA -> TT will not be imported.

In version 7.0 and later versions, all variants will be imported, but the replacements TTA -> TT will be imported as one deletion A -> - .

To get a variant count as part of your imported variant, one of the following VCF tags have to be present in your VCF file: CLCAD2, AD, or AO.

The import of CLCAD2/AD/AO tags are prioritized in the following order:

1. CLCAD2
2. AD
3. AO

If the CLCAD2 is missing, and only AD is present, then AD is used in the "count" column.

The consequence of this, if the file for example has CLCAD2:AD, and in a sample for three possible variants the values are 2,3,4:5,6,7, then the CLCAD2 tag will be imported as count, so each of the three variants will have just one count value (2, 3, and 4 respectively). At the same time, the AD tag will be imported as an annotation so all of the three variants will have "5,6,7" under the AD column, like for any unknown format tag.

Special notes on chromosome names synonyms used during import

When importing annotations as tracks, we try to make things simple for the user by having a set of chromosome names that are recognized as synonyms. The check on the chromosome name comparison is made by looking through the chromosomes in the order in which they are registered in the genome. The first match with any of the synonym names for a given chromosome is the chromosome to which the information will be added.

The synonyms applied are:

For any number N between (including) 1 and 22:

N, chrN, chromosome_N, and NC_00000N are seen as meaning the same thing. As concrete examples:

1 == chr1 == chromosome_1 == NC_000001

22 == chr22 == chromosome_22 == NC_000022

For any number N larger than 23:

N, chrN, chromosome_N are seen as meaning the same thing. As a concrete example:

26 == chr26 == chromsome_26

For chromsome names with letters, not numbers:

X, chrX, and chromosome_X and NC_000023 are synonyms.

Y, chrY, chromosome_Y and NC_000024 are synonyms.

M, MT, chrM, chrMT, chromosome_M, chromosome_MT and NC_001807 are synonyms.

The accession numbers in the listings above (NC_XXXXXX) allow for the matching against NCBI hg19 human reference names against the names used by USCS and vitally, the names used by Ensembl. Thus, in this case, if you have the correct number of chromosomes in a human reference (i.e. 25 references, including the hg19 mitochondria), that set of tracks can be used as the basis for downloading/importing annotations via Download Genomes, for example.

Note: These rules only apply for importing annotations as tracks, whether that is directly or via Download Genomes. Synonyms are not applied when doing BAM imports or when using the Annotate with GFF plugin. There, your reference names in the Workbench must exactly match the references names used in your BAM file or GFF/GTF/GVF file respectively.

E.1.8 Table and text formats

File type	Suffix	Import	Export	Description
Excel	.xls/.xlsx	X	X	All tables and reports
Table CSV	.csv	X	X	All tables
Tab delimited	.txt		X	All tables
Text	.txt	X	X	All data in a textual format
CLC	.clc	X	X	Rich format including all information
HTML	.html		X	All tables
PDF	.pdf		X	Export reports in Portable Document Format

Please see table E.1.6 **Expression data formats** for special cases of table imports.

E.1.9 File compression formats

File type	Suffix	Import	Export	Description
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gz/.tar	X		Contained files/folder structure (.tar and .zip not supported for NGS data)

Note! It is possible to import 'external' files into the Workbench and view these in the **Navigation Area**, but it is only the above mentioned formats whose *contents* can be shown in the Workbench.

E.2 List of graphics data formats

Below is a list of formats for exporting graphics. All data displayed in a graphical format can be exported using these formats. Data represented in lists and tables can only be exported in .pdf format (see section 6.6 for further details).

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

See the complete list including download links at http://www.geneontology.org/GO.current.annotations.shtml.

Appendix F

SAM/BAM export format specification

SAM Specification The workbench aims to import and export SAM and BAM files according to the v1.4-r962 version of the SAM specification (see http://samtools.sourceforge.net/SAM1.pdf). This appendix describes how the workbench exports SAM and BAM files along with known limitations.

SAM and BAM Export - General notes The SAM exporter writes unsorted SAM and BAM files.

If the reference name contains spaces, the spaces are removed. Each occurrence of '=' (equals sign) and '@' (at sign) in a reference name is replaced by an '_' (underscore).

The SAM importer and exporter support the ID, SM, PI and PL read group tags. All other read group tags are ignored.

SAM Alignment Section A few remarks on the exported alignment section:

- Unmapped reads are not exported.
- If pairs are not on the same contig, the mates will be exported as single reads.
- Multi segment mappings will be imported as a paired data set.
- If a read name contains spaces, the spaces are replaced by an underscore '_'.
- The exported CIGAR string uses 'M' to indicate match or mismatch and does not use '=' (equals sign) or 'X'.
- CLC software does not support or record mapping quality for read mappings. To fulfill the requirement in the BAM format specifications that a read mapping quality is recorded for all mapped reads, the values 0 and 60 are used when mappings are exported from the Workbench. The value 60 is given to reads that mapped uniquely. The value 0 is given to reads that could map equally well to other locations besides the one being reported in the BAM file.

  This scoring system is based on a recommendation provided in the the SAM FAQ:

  http://sourceforge.net/apps/mediawiki/samtools/index.php?title=SAM_FAQ#How_to_make_my_aligner_work_best_with_samtools.3F

Optional fields in the alignment section The following is true for the export of optional fields:

- The NH tag is exported.
- The NM tag is not exported.
- The workbench exports color space information in the CS tag.
- The colors of a right mate are incorrect since the colors of a paired read are stored as a single color string.
- For hard clipped sequence reads, the color space is incorrect, since the color space string is not hard clipped.
- SAM files contain sequence quality score and color quality scores. The workbench only have color quality scores and these are stored and exported as sequence quality scores.

F.1 Flags

The workbench's use of the alignment flags is shown in the following table and subsequent examples.

Bit	SAM description	Usage in Workbench
0x1	template having multiple segments in sequencing	set if the segment is part of a pair
0x2	each segment properly aligned according to the aligner	set if the pair is not broken
0x4	segment unmapped	never set since the exporter does not export unmapped reads
0x8	next segment in the template unmapped	never set by the exporter. If a segment has an unmapped mate, the flag 0x1 is not set for the segment, i.e. it is not output as part of a pair
0x10	SEQ being reverse complemented	set if and only if the segment was reverse complemented during mapping
0x20	SEQ of the next segment in the template being reversed	set if and only if the mate was reverse complemented during mapping
0x40	the first segment in the template	this mate is the first segment of the pair
0x80	the last segment in the template	this mate is the second segment of the pair
0x100	secondary alignment	never set by the exporter. No reads with this flag set are imported[1].
0x200	not passing quality controls	never set by the exporter and ignored by the importer
0x400	PCR or optical duplicate	never set by the exporter and ignored by the importer

Flag Examples

The following table illustrates some of the possible flags in the workbench.

Description of the example	Bits	Flag	Illustration
The first mate of a non-broken paired read	`0x1, 0x2, 0x20, 0x40`	**99**	See Figure F.1
The second mate of a non-broken paired read	`0x1, 0x2, 0x10, 0x80`	**147**	See Figure F.2
A single, forward read (or paired read, where only one mate of the pair is mapped)	No set bits	**0**	see Figure F.3
A single, reversed read (or paired read, where only one mate of the pair is mapped)	`0x10`	**16**	See Figure F.4
The first, forward segment from a broken pair with forward mate	`0x1, 0x40`	**65**	See Figure F.5
The second, forward segment from broken pair with reversed mate	`0x1, 0x20, 0x80`	**161**	See Figure F.6
The first, reversed segment from broken pair with forward mate	`0x1, 0x10, 0x40`	**81**	See Figure F.7
The second, reversed segment from broken pair with reversed mate	`0x1, 0x10, 0x20, 0x80`	**177**	See Figure F.8

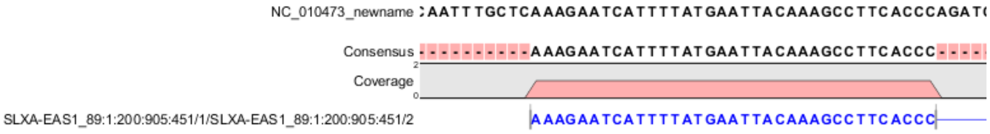

NC_010473_newname :AATTTGCTCAAAGAATCATTTTATGAATTACAAAGCCTTCACCCAGAT(
Consensus ----------AAAGAATCATTTTATGAATTACAAAGCCTTCACCC----
2
Coverage
0
SLXA-EAS1_89:1:200:905:451/1/SLXA-EAS1_89:1:200:905:451/2
AAAGAATCATTTTATGAATTACAAAGCCTTCACCC

Figure F.1: *The read is paired, both reads are mapped and the mate of this read is reversed*

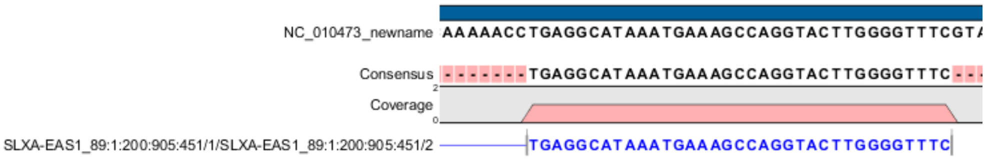

NC_010473_newname AAAAACCTGAGGCATAAATGAAAGCCAGGTACTTGGGGTTTCGTA
Consensus -------TGAGGCATAAATGAAAGCCAGGTACTTGGGGTTTC---
2
Coverage
0
SLXA-EAS1_89:1:200:905:451/1/SLXA-EAS1_89:1:200:905:451/2
TGAGGCATAAATGAAAGCCAGGTACTTGGGGTTTC

Figure F.2: *The read is paired, both mates are mapped, and this segment is reversed*

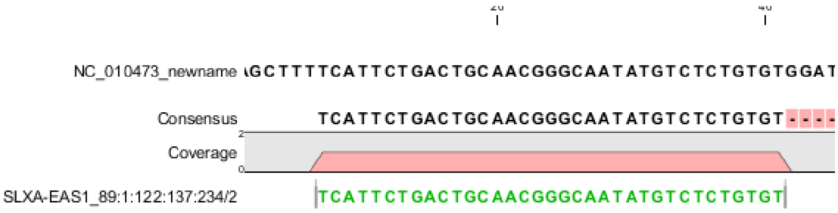

20
40
NC_010473_newname AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGAT
Consensus TCATTCTGACTGCAACGGGCAATATGTCTCTGTGT----
2
Coverage
0
SLXA-EAS1_89:1:122:137:234/2
TCATTCTGACTGCAACGGGCAATATGTCTCTGTGT

Figure F.3: *A single, forward read, or a paired read where the mate is not mapped*

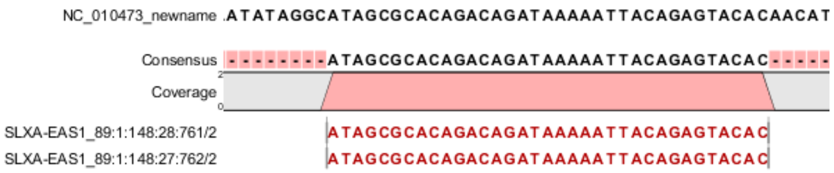

NC_010473_newname .ATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACAT
Consensus --------ATAGCGCACAGACAGATAAAAATTACAGAGTACAC-----
Coverage
SLXA-EAS1_89:1:148:28:761/2 ATAGCGCACAGACAGATAAAAATTACAGAGTACAC
SLXA-EAS1_89:1:148:27:762/2 ATAGCGCACAGACAGATAAAAATTACAGAGTACAC

Figure F.4: *The read is a single, reversed read, or a paired read where the mate is not mapped*

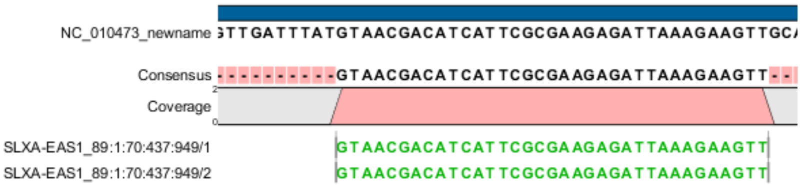

NC_010473_newname GTTGATTTATGTAACGACATCATTCGCGAAGAGATTAAAGAAGTTGCA
Consensus ----------GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT--
Coverage
SLXA-EAS1_89:1:70:437:949/1 GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT
SLXA-EAS1_89:1:70:437:949/2 GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT

Figure F.5: *These forward reads are paired. They map to the same place, so the pair is broken*

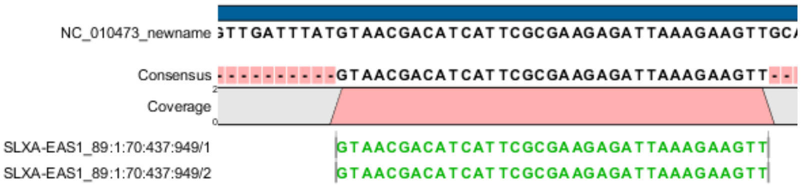

NC_010473_newname GTTGATTTATGTAACGACATCATTCGCGAAGAGATTAAAGAAGTTGCA
Consensus ----------GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT--
Coverage
SLXA-EAS1_89:1:70:437:949/1 GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT
SLXA-EAS1_89:1:70:437:949/2 GTAACGACATCATTCGCGAAGAGATTAAAGAAGTT

Figure F.6: *Forward read that is part of a broken read where the mate is reversed*

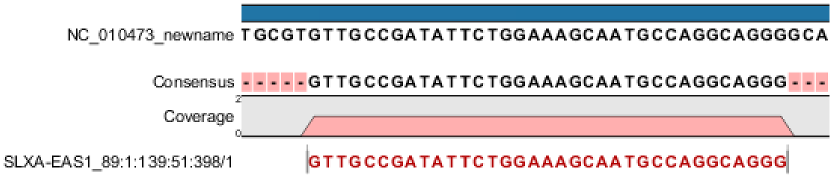

NC_010473_newname TGCGTGTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGGGCA
Consensus -----GTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGG---
Coverage
SLXA-EAS1_89:1:139:51:398/1 GTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGG

Figure F.7: *Reversed read that is part of a broken pair, where the mate is forward*

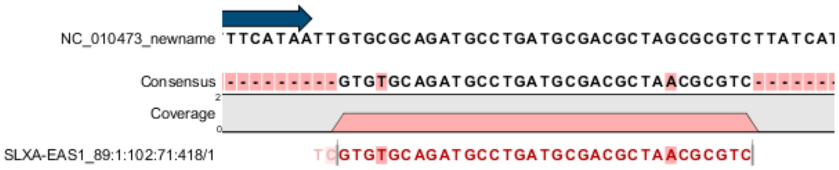

NC_010473_newname TTCATAATTGTGCGCAGATGCCTGATGCGACGCTAGCGCGTCTTATCAT
Consensus ---------GTGTGCAGATGCCTGATGCGACGCTAACGCGTC-------
Coverage
SLXA-EAS1_89:1:102:71:418/1 TCGTGTGCAGATGCCTGATGCGACGCTAACGCGTC

Figure F.8: *Reversed read that is part of a broken pair, where the mate is also reversed.*

Appendix G

Gene expression annotation files and microarray data formats

The workbench supports analysis of one-color expression arrays. These may be imported from GEO soft sample- or series- file formats, or for Affymetrix arrays, tab-delimited pivot or metrics files, or from Illumina expression files. Expression array data from other platforms may be imported from tab, semi-colon or comma separated files containing the expression feature IDs and levels in a tabular format (see section G.5).

The workbench assumes that expression values are given at the gene level, thus probe-level analysis of e.g. Affymetrix GeneChips and import of Affymetrix CEL and CDF files is currently not supported. However, the workbench allows import of txt files exported from R containing processed Affymetrix CEL-file data (see see section G.2).

Affymetrix NetAffx annotation files for expression GeneChips in csv format and Illumina annotation files can also be imported. Also, you may import your own annotation data in tabular format see section G.5).

Below you find descriptions of the microarray data formats that are supported by *CLC Cancer Research Workbench*. Note that we for some platforms support both expression data and annotation data.

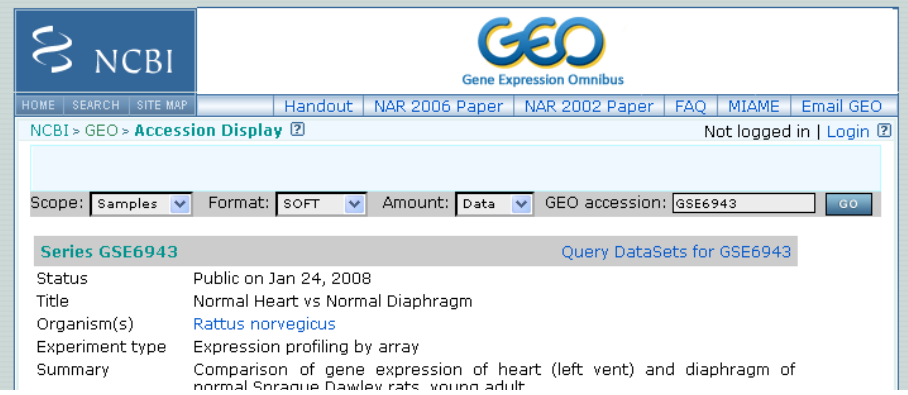
G.1 GEO (Gene Expression Omnibus)

The GEO (Gene Expression Omnibus) sample and series formats are supported. Figure G.1 shows how to download the data from GEO in the right format. GEO is located at `http://www.ncbi.nlm.nih.gov/geo/`.

The GEO sample files are tab-delimited .txt files. They have three required lines:

```
^SAMPLE = GSM21610
!sample_table_begin
...
!sample_table_end
```

The first line should start with `^SAMPLE =` followed by the sample name, the line `!sample_table_begin`


NCBI
GEO
Gene Expression Omnibus
HOME | SEARCH | SITE MAP
Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO
NCBI > GEO > Accession Display
Not logged in | Login
GO
Query DataSets for GSE6943

Scope: Samples Format: SOFT Amount: Data GEO accession: GSE6943

Series GSE6943

Status	Public on Jan 24, 2008
Title	Normal Heart vs Normal Diaphragm
Organism(s)	Rattus norvegicus
Experiment type	Expression profiling by array
Summary	Comparison of gene expression of heart (left vent) and diaphragm of normal Sprague Dawley rats, young adult

Figure G.1: *Selecting Samples, SOFT and Data before clicking go will give you the format supported by the* **CLC Cancer Research Workbench***.*

and the line `!sample_table_end`. Between the `!sample_table_begin` and `!sample_table_end`, lines are the column contents of the sample.

Note that GEO sample importer will also work for concatenated GEO sample files — allowing multiple samples to be imported in one go. Download a sample file containing concatenated sample files here:
http://www.clcbio.com/madata/GEOSampleFilesConcatenated.txt

Below you can find examples of the formatting of the GEO formats.

G.1.1 GEO sample file, simple

This format is very simple and includes two columns: one for feature id (e.g. gene name) and one for the expression value.

```
^SAMPLE = GSM21610
!sample_table_begin
ID_REF    VALUE
id1       105.8
id2       32
id3       50.4
id4       57.8
id5       2914.1
!sample_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSampleFileSimple.txt

G.1.2 GEO sample file, including present/absent calls

This format includes an extra column for absent/present calls that can also be imported.

```
^SAMPLE = GSM21610
!sample_table_begin
ID_REF    VALUE    ABS_CALL
```

```
id1      105.8   M
id2      32      A
id3      50.4    A
id4      57.8    A
id5      2914.1  P
!sample_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSampleFileAbsentPresent.txt

G.1.3 GEO sample file, including present/absent calls and p-values

This format includes two extra columns: one for absent/present calls and one for absent/present call p-values, that can also be imported.

```
^SAMPLE = GSM21610
!sample_table_begin
ID_REF    VALUE     ABS_CALL     DETECTION P-VALUE
id1       105.8     M            0.00227496
id2       32        A            0.354441
id3       50.4      A            0.904352
id4       57.8      A            0.937071
id5       2914.1    P            6.02111e-05
!sample_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSampleFileAbsentPresentCallAndPValue.txt

G.1.4 GEO sample file: using absent/present call and p-value columns for sequence information

The workbench assumes that if there is a third column in the GEO sample file then it contains present/absent calls and that if there is a fourth column then it contains p-values for these calls. This means that the contents of the third column is assumed to be text and that of the fourth column a number. As long as these two basic requirements are met, the sample should be recognized and interpreted correctly.

You can thus use these two columns to carry additional information on your probes. The absent/present column can be used to carry additional information like e.g. sequence tags as shown below:

```
^SAMPLE = GSM21610
!sample_table_begin
ID_REF      VALUE     ABS_CALL
id1        105.8      AAA
id2        32         AAC
id3        50.4       ATA
id4        57.8       ATT
```

```
id5         2914.1     TTA
!sample_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSampleFileSimpleSequenceTag.txt

Or, if you have multiple probes per sequence you could use the present/absent column to hold the sequence name and the p-value column to hold the interrogation position of your probes:

```
^SAMPLE = GSM21610
!sample_table_begin
ID_REF    VALUE    ABS_CALL    DETECTION P-VALUE
probe1    755.07   seq1        1452
probe2    587.88   seq1        497
probe3    716.29   seq1        1447
probe4    1287.18  seq2        1899
!sample_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSampleFileSimpleSequenceTagAndProbe.txt

G.1.5 GEO series file, simple

The series file includes expression values for multiple samples. Each of the samples in the file will be represented by its own element with the sample name. The first row lists the sample names.

```
!Series_title "Myb specificity determinants"
!series_matrix_table_begin
"ID_REF" "GSM21610" "GSM21611" "GSM21612"
"id1"      2541       1781.8     1804.8
"id2"      11.3       621.5     50.2
"id3"      61.2       149.1     22
"id4"      55.3       328.8     97.2
"id5"        183.8       378.3      423.2
!series_matrix_table_end
```

Download the sample file here:
http://www.clcbio.com/madata/GEOSeriesFile.txt

G.2 Affymetrix GeneChip

For Affymetrix, three types of files are currently supported: Affymetrix .CHP files, Affymetrix NetAffx annotation files and tab-delimited pivot or metrics files. Affymetrix .CEL files are currently not supported. However, the Bioconductor R package 'affy' allows you to preprocess the .CEL files and export a txt file containing a table of estimated gene-level expression values in three lines of code:

```
library(affy) # loading Bioconductor library 'affy'
data=ReadAffy() # probe-level data import
eset=rma(data) # probe-level data pre-processing using 'rma'
write.exprs(eset,file="evals.txt") # writing gene expression levels to 'evals-txt'
```

The exported txt file (evals.txt) can be imported into the workbench using the Generic expression data table format importer (see section G.5; you can just 'drag-and-drop' it in). In R, you should have all the CEL files you wish to process in your working directory and the file 'evals.txt' will be written to that directory.

G.2.1 Affymetrix CHP expression files

The Affymetrix scanner software produces a number of files when a GeneChip is scanned. Two of these are the .CHP and the .CEL files. These are binary files with native Affymetrix formats. The Affymetrix GeneChips contain a number of probes for each gene (typically between 22 and 40). The .CEL file contains the probe-level intensities, and the .CHP file contains the gene-level information. The gene-level information has been obtained by the scanner software through postprocessing and summarization of the probe-level intensities.

In order to interpret the probe-level information in the .CEL file, the .CDF file for the type of GeneChip that was used is required. Similarly for the .CHP file: in order to interpret the gene-level information in the .CHP file, the .PSI file for the type of GeneChip that was used is required.

In order to import a .CHP file it is required that the corresponding .PSI file is present in the same folder as the .CHP file you want to import, and furthermore, this must be the only .PSI file that is present there. There are no requirements for the name of the .PSI file. Note that the .PSI file itself will not be imported - it is only used to guide the import of the .CHP file which contains the expression values.

Download example .CHP and .PSI files here (note that these are binary files):
http://www.clcbio.com/madata/AffymetrixCHPandPSI.zip

G.2.2 Affymetrix metrix files

The Affymetrix metrics or pivot files are tab-delimited files that may be exported from the Affymetrix scanner software. The metrics files have a lot of technical information that is only partly used in the Workbench. The feature ids (Probe Set Name), expression values (Used Signal), absent/present call (Detection) and absent/present p-value (Detection p-value) are imported into the Workbench.

Download a small example sample file here:
http://www.clcbio.com/madata/AffymetrixMetrics.txt

G.2.3 Affymetrix NetAffx annotation files

The NetAffx annotation files for Whole-Transcript Expression Gene arrays and 3' IVT Expression Analysis Arrays can be imported and used to annotate experiments as shown in section 27.3.4.

Download a small example annotation file here which includes header information:
http://www.clcbio.com/madata/AffymetrixNetAffxAnnotationFile.csv

G.3 Illumina BeadChip

Both BeadChip expression data files from Illumina's BeadStudio software and the corresponding BeadChip annotation files are supported by *CLC Cancer Research Workbench*. The formats of the BeadStudio and annotation files have changed somewhat over time and various formats are supported.

G.3.1 Illumina expression data, compact format

An example of this format is shown below:

```
TargetID        AVG_Signal      BEAD_STDEV     Detection
GI_10047089-S   112.5           4.2            0.16903226
GI_10047091-S   127.6           4.8            0.76774194
```

All this information is imported into the Workbench. The `AVG_Signal` is used as the expression measure.

Download a small sample file here:
http://www.clcbio.com/madata/IlluminaBeadChipCompact.txt

G.3.2 Illumina expression data, extended format

An example of this format is shown below:

```
TargetID        MIN_Signal  AVG_Signal  MAX_Signal  NARRAYS  ARRAY_STDEV  BEAD_STDEV  Avg_NBEADS  Detection
GI_10047089-S   73.7        73.7        73.7        1        NaN          3.4         53          0.05669084
GI_10047091-S   312.7       312.7       312.7       1        NaN          11.1        50          0.99604483
```

All this information is imported into the Workbench. The `AVG_Signal` is used as the expression measure.

Download a small sample file here:
http://www.clcbio.com/madata/IlluminaBeadChipExtended.txt

G.3.3 Illumina expression data, with annotations

An example of this format is shown below:

```
TargetID Accession Symbol Definition Synonym Signal-BG02 DCp32  Detection-BG02 DCp32
GI_10047089-S NM_014332.1 SMPX "Homo sapiens small muscle protein, X-linked (SMPX), mRNA."  -17.6  0.03559657
GI_10047091-S NM_013259.1 NP25 "Homo sapiens neuronal protein (NP25), mRNA." NP22  32.6  0.99604483
GI_10047093-S NM_016299.1 HSP70-4 "Homo sapiens likely ortholog of mouse heat shock protein, 70 kDa 4 (HSP70-4), mRNA."  228.1 1
```

Only the TargetID, Signal and Detection columns will be imported, the remaining columns will be ignored. This means that the annotations are not imported. The `Signal` is used as the expression measure.

Download a small example sample file here:
http://www.clcbio.com/madata/IlluminaBeadStudioWithAnnotations.txt

G.3.4 Illumina expression data, multiple samples in one file

This file format has too much information to show it inline in the text. You can download a small example sample file here:
http://www.clcbio.com/madata/IlluminaBeadStudioMultipleSamples.txt

This file contains data for 18 samples. Each sample has an expression value (the value in the AVG_Signal column), a detection p-value, a bead standard deviation and an average bead number column. The workbench recognizes the 18 samples and their columns.

G.3.5 Illumina annotation files

The Workbench supports import of two types of Illumina BeadChip annotation files. These are either comma-separated or tab-delimited .txt files. They can be used to annotate experiments as shown in section 27.3.4.

This file format has too much information to show it inline in the text.

Download a small example annotation file of the first type here:
http://www.clcbio.com/madata/IlluminaBeadChipAnnotation.txt

G.4 Gene ontology annotation files

The Gene ontology web site provides annotation files for a variety of species which can all be downloaded and imported into the *CLC Cancer Research Workbench*. This can be used to annotate experiments as shown in section 27.3.4. See the complete list including download links at http://www.geneontology.org/GO.current.annotations.shtml.

This is an easy way to annotate your experiment with GO categories.

G.5 Generic expression and annotation data file formats

If you have your expression or annotation data in e.g. Excel and can export the data as a txt file, or if you are able to do some scripting or other manipulations to format your data files, you will be able to import them into the *CLC Cancer Research Workbench* as a 'generic' expression or annotation data file. There are a few simple requirements that need to be fulfilled to do this as described below.

G.5.1 Generic expression data table format

The *CLC Cancer Research Workbench* will import a tab, semicolon or comma-separated .txt or .csv file as expression array samples if the following requirements are met:

1. the first non-empty line of the file contains text. All entries, except the first, will be used as sample names

2. the following (non-empty) lines contain the same number of entries as the first non-empty line. The requirements to these are that the first entry should be a string (this will be used as the feature ID) and the remaining entries should contain numbers (which will be used as

expression values — one per sample). Empty entries are not allowed, but NaN values are allowed.

3. the file contains at least two samples.

An example of this format is shown below:

```
FeatureID;sample1;sample2;sample3
gene1;200;300;23
gene2;210;30;238
gene3;230;50;23
gene4;50;100;235
gene5;200;300;23
gene6;210;30;238
gene7;230;50;23
gene8;50;100;235
```

This will be imported as three samples with eight genes in each sample.

Download a this example as a file here:
http://www.clcbio.com/madata/CustomExpressionData.txt

G.5.2 Generic annotation file for expression data format

The *CLC Cancer Research Workbench* will import a tab, semicolon or comma-separated .txt or .csv file as an annotation file if the following is met:

1. It has a line which can serve as a valid header line. In order to do this, the line should have a number of headers where at least two are among the valid column headers in the **Column header** column below.

2. It contains one of the PROBE_ID headers (that is: 'Probe Set ID', 'Feature ID', 'ProbeID' or 'Probe_Id').

The importer will import an annotation table with a column for each of the valid column headers (those in the **Column header** column below). Columns with invalid headers will be ignored.

Note that some column headers are alternatives so that only one of the alternative columns headers should be used.

When adding annotations to an experiment, you can specify the column in your annotation file containing the relevant identifiers. These identifiers are matched to the feature ids already present in your experiment. When a match is found, the annotation is added to that entry in the experiment. In other words, at least one column in your annotation file must contain identfiers matching the feature identifiers in the experiment, for those annotations to be applied.

A simple example of an annotation file is shown here:

```
"Probe Set ID","Gene Symbol","Gene Ontology Biological Process"
"1367452_at","Sumo2","0006464 // protein modification process //  not recorded"
"1367453_at","Cdc37","0051726 // regulation of cell cycle //  not recorded"
"1367454_at","Copb2","0006810 // transport //  ///  0016044 // membrane organization // "
```

Download this example plus a more elaborate one here:
`http://www.clcbio.com/madata/SimpleCustomAnnotation.csv`
`http://www.clcbio.com/madata/FullCustomAnnotation.csv`

To meet requirements imposed by special functionalities in the workbench, there are a number of further restrictions on the contents in the entries of the columns:

Download sequence functionality In the experiment table, you can click a button to download sequence. This uses the contents of the `PUBLIC_ID` column, so this column must be present for the action to work and should contain the NCBI accession number.

Annotation tests The annotation tests can make use of several entries in a column as long as a certain format is used. The tests assume that entries are separated by /// and it interprets all that appears before // as the actual entry and all that appears after // within an entry as comments. Example:

```
/// 0000001 //  comment1  /// 0000008 // comment2 /// 0003746 //  comment3
```

The annotation tests will interpret this as three entries (0000001, 0000008, and 0003746) with the according comments.

The most common column headers are summarized below:

Column header in imported file (alternatives separated by commas)	**Label in experiment table**	**Description (tool tip)**
Probe Set ID, Feature ID, ProbeID, Probe_Id, transcript_cluster_id	Feature ID	Probe identifier tag
Representative Public ID, Public identifier tag, GenbankAccession	Public identifier tag	Representative public ID
Gene Symbol, GeneSymbol	Gene symbol	Gene symbol
Gene Ontology Biological Process, Ontology_Process, GO_biological_process	GO biological process	Gene Ontology biological process
Gene Ontology Cellular Component, Ontology_Component, GO_cellular_component	GO cellular component	Gene Ontology cellular component
Gene Ontology Molecular Function, Ontology_Function, GO_molecular_function	GO molecular function	Gene Ontology molecular function
Pathway	Pathway	Pathway

The full list of possible column headers:

Column header in imported file (alternatives separated by commas)	**Label in experiment table**	**Description (tool tip)**
Species Scientific Name, Species Name, Species	Species name	Scientific species name
GeneChip Array	Gene chip array	Gene Chip Array name
Annotation Date	Annotation date	Date of annotation
Sequence Type	Sequence type	Type of sequence
Sequence Source	Sequence source	Source from which sequence was obtained
Transcript ID(Array Design), Transcript	Transcript ID	Transcript identifier tag
Target Description	Target description	Target description
Archival UniGene Cluster	Archival UniGene cluster	Archival UniGene cluster
UniGene ID, UniGeneID, Unigene_ID, unigene	UniGene ID	UniGene identifier tag
Genome Version	Genome version	Version of genome on which annotation is based
Alignments	Alignments	Alignments
Gene Title	Gene title	Gene title
geng_assignments	Gene assignments	Gene assignments
Chromosomal Location	Chromosomal location	Chromosomal location
Unigene Cluster Type	UniGene cluster type	UniGene cluster type
Ensemble Ensembl	Ensembl	
Entrez Gene, EntrezGeneID, Entrez_Gene_ID	Entrez gene	Entrez gene
SwissProt	SwissProt	SwissProt
EC	EC	EC
OMIM	OMIM	Online Mendelian Inheritance in Man
RefSeq Protein ID	RefSeq protein ID	RefSeq protein identifier tag
RefSeq Transcript ID	RefSeq transcript ID	RefSeq transcript identifier tag
FlyBase	FlyBase	FlyBase
AGI	AGI	AGI
WormBase	WormBase	WormBase
MGI Name	MGI name	MGI name
RGD Name	RGD name	RGD name
SGD accession number	SGD accession number	SGD accession number
InterPro	InterPro	InterPro
Trans Membrane	Trans membrane	Trans membrane
QTL	QTL	QTL
Annotation Description	Annotation description	Annotation description
Annotation Transcript Cluster	Annotation transcript cluster	Annotation transcript cluster
Transcript Assignments	Transcript assignments	Trancript assignments
mrna_assignments	mRNA assignments	mRNA assignments
Annotation Notes	Annotation notes	Annotation notes
GO, Ontology	Go annotations	Go annotations
Cytoband	Cytoband	Cytoband
PrimaryAccession	Primary accession	Primary accession
RefSeqAccession	RefSeq accession	RefSeq accession
GeneName	Gene name	Gene name
TIGRID	TIGR Id	TIGR Id
Description	Description	Description
GenomicCoordinates	Genomic coordinates	Genomic coordinates
Search_key	Search key	Search key
Target	Target	Target
Gid, GI	Genbank identifier	Genbank identifier
Accession	GenBank accession	GenBank accession
Symbol	Gene symbol	Gene symbol
Probe_Type	Probe type	Probe type
crosshyb_type	Crosshyb type	Crosshyb type
category	category	category
Start, Probe_Start	Start	Start
Stop	Stop	Stop
Definition	Definition	Definition
Synonym, Synonyms	Synonym	Synonym
Source	Source	Source
Source_Reference_ID	Source reference id	Source reference id
RefSeq_ID	Reference sequence id	Reference sequence id
ILMN_Gene	Illumina Gene	Illumina Gene
Protein_Product	Protein product	Protein product
protein_domains	Protein domains	Protein domains
Array_Address_Id	Array adress id	Array adress id
Probe_Sequence	Sequence	Sequence
seqname	Seqname	Seqname
Chromosome	Chromosome	Chromosome
strand	Strand	Strand
Probe_Chr_Orientation	Probe chr orientation	Probe chr orientation
Probe_Coordinates	Probe coordinates	Probe coordinates
Obsolete_Probe_Id	Obsolete probe id	Obsolete probe id

Bibliography

[et al., 2012] , . G. P. C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[Allison et al., 2006] Allison, D., Cui, X., Page, G., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *NATURE REVIEWS GENETICS*, 7(1):55.

[Altshuler et al., 2000] Altshuler, D., Pollara, V. J., Cowles, C. R., Etten, W. J. V., Baldwin, J., Linton, L., and Lander, E. S. (2000). An snp map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803):513–516.

[Bamford et al., 2004] Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., and Wooster, R. (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *Br J Cancer*, 91(2):355–358.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B*, 57:289–289.

[Bolstad et al., 2003] Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

[Bommarito et al., 2000] Bommarito, S., Peyret, N., and SantaLucia, J. (2000). Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res*, 28(9):1929–1934.

[Brockman et al., 2008] Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and snp detection in sequencing-by-synthesis systems. *Genome Res*, 18(5):763–770.

[Choi et al., 2009] Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*, 106(45):19096–19101.

[Cornette et al., 1987] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol*, 195(3):659–685.

[Creighton et al., 2009] Creighton, C. J., Reid, J. G., and Gunaratne, P. H. (2009). Expression profiling of micrornas by deep sequencing. *Brief Bioinform*, 10(5):490–497.

[Cronn et al., 2008] Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., and Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using solexa sequencing-by-synthesis technology. *Nucleic Acids Res*, 36(19):e122.

[Dudoit et al., 2003] Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple Hypothesis Testing in Microarray Experiments. *STATISTICAL SCIENCE*, 18(1):71–103.

[Eisen et al., 1998] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

[Eisenberg et al., 1984] Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*, 179(1):125–142.

[Emini et al., 1985] Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55(3):836–839.

[Engelman et al., 1986] Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353.

[Falcon and Gentleman, 2007] Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257.

[Guo et al., 2006] Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F. M., Hurban, P., Phillips, K. L., Xu, J., Deng, X., Sun, Y. A., Tong, W., Dragan, Y. P., and Shi, L. (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol*, 24(9):1162–1169.

[Heap et al., 2010] Heap, G. A., Yang, J. H. M., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., Franke, L., Dubois, P. C., Mein, C. A., Dobson, R. J., Albert, T. J., Rodesch, M. J., Clayton, D. G., Todd, J. A., van Heel, D. A., and Plagnol, V. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*, 19(1):122–134.

[Homer N, 2010] Homer N, N. S. (2010). Improved variant discovery through local re-alignment of short-read next-generation sequencing data using srma. *Genome Biol.*, 11(10):R99.

[Hopp and Woods, 1983] Hopp, T. P. and Woods, K. R. (1983). A computer program for predicting protein antigenic determinants. *Mol Immunol*, 20(4):483–489.

[Janin, 1979] Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492.

[Karplus and Schulz, 1985] Karplus, P. A. and Schulz, G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. (1990). Finding groups in data. an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990*.

[Kolaskar and Tongaonkar, 1990] Kolaskar, A. S. and Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172--174.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137.

[Martin and Wang, 2011] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, 12(10):671–682.

[Meyer et al., 2007] Meyer, M., Stenzel, U., Myles, S., Prï¿$\frac{1}{2}$fer, K., and Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res*, 35(15):e97.

[Morin et al., 2008] Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008). Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem cells. *Genome Res*, 18(4):610–621.

[Morrison, 1968] Morrison, D. R. (1968). Patricia – practical algorithm to retrieve information coded in alphanumeric. *J. ACM*, 15(4):514–534.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628.

[Ng et al., 2009] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

[Nguyen et al., 2011] Nguyen, P., Ma, J., Pei, D., Obert, C., Cheng, C., and Geiger, T. (2011). Identification of errors introduced during high throughput sequencing of the t cell receptor repertoire. *BMC genomics*, 12(1):106.

[Parkhomchuk et al., 2009] Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary dna. *Nucleic Acids Res*, 37(18):e123.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Robinson and Smyth, 2007] Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887.

[Robinson and Smyth, 2008] Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332.

[Rose et al., 1985] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838.

[SantaLucia, 1998] SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, 95(4):1460–1465.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

[Stark et al., 2010] Stark, M. S., Tyagi, S., Nancarrow, D. J., Boyle, G. M., Cook, A. L., Whiteman, D. C., Parsons, P. G., Schmidt, C., Sturm, R. A., and Hayward, N. K. (2010). Characterization of the melanoma mirnaome by deep sequencing. *PLoS One*, 5(3):e9685.

[Sturges, 1926] Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21:65–66.

[Tian et al., 2005] Tian, L., Greenberg, S., Kong, S., Altschuler, J., Kohane, I., and Park, P. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549.

[Tusher et al., 2001] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121.

[von Ahsen et al., 2001] von Ahsen, N., Wittwer, C. T., and Schï¿½tz, E. (2001). Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem*, 47(11):1956–1961.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

[Welling et al., 1985] Welling, G. W., Weijer, W. J., van der Zee, R., and Welling-Wester, S. (1985). Prediction of sequential antigenic regions in proteins. *FEBS Lett*, 188(2):215–218.

[Wyman et al., 2009] Wyman, S. K., Parkin, R. K., Mitchell, P. S., Fritz, B. R., O'Briant, K., Godwin, A. K., Urban, N., Drescher, C. W., Knudsen, B. S., and Tewari, M. (2009). Repertoire of micrornas in epithelial ovarian cancer as determined by next generation sequencing of small rna cdna libraries. *PLoS One*, 4(4):e5311.

Part XII

Index

Appendix H

Index

Index