



CLC **Server**

Administrator

USER MANUAL

Administrator Manual for
CLC Server 11.0
Windows, macOS and Linux

May 15, 2019

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark



Contents

1	Introduction	6
1.1	System requirements	6
1.2	Licensing	8
1.3	CLC Genomics Server	8
2	Installation	13
2.1	Quick installation guide	13
2.2	Installing and running the Server	14
2.3	Installation modes - console and silent	16
2.4	Upgrading an existing installation	16
2.5	Allowing access through your firewall	17
2.6	Downloading a license	18
2.7	Starting and stopping the server	20
2.8	Installing the database	22
3	Basic configuration	25
3.1	Logging into the administrative interface	25
3.2	Adding locations for saving data	25
3.3	Accessing files on, and writing to, areas of the server filesystem	30
3.4	Direct data transfer from client systems	32
3.5	Direct data transfer from client systems	32
3.6	Changing the listening port	34
3.7	Changing the tmp directory	34
3.8	Setting the amount of memory available for the JVM	35
3.9	Limiting the number of cpus available for use	35

3.10 HTTP settings	35
3.11 Deployment of server information to CLC Workbenches	35
4 Managing users and groups	36
4.1 Logging in the first time and changing the root password	36
4.2 User authentication using the web interface	37
5 Access privileges and permissions	43
5.1 Controlling access to CLC Server data	43
5.2 Controlling access to the server, server tasks and external data	46
5.3 Customized attributes on data locations	47
6 Job distribution	53
6.1 Introduction to servers setups	53
6.2 Model I: Master server with dedicated job nodes	54
6.3 Model II: Master server submitting to grid nodes	58
6.4 Model III: Single Server setup	71
6.5 Job queuing options	72
6.6 Job running options	74
7 Status and management	77
7.1 User statistics	77
7.2 System statistics	78
7.3 Server maintenance	78
8 Queue	80
9 Audit log	81
10 Server plugins	83
11 BLAST	86
11.1 Adding directories for BLAST databases on the Server	86
11.2 Adding and removing BLAST databases	87
12 External applications	89

12.1 General configuration of external applications	91
12.2 Stream handling	97
12.3 Environment	97
12.4 End user interface	99
12.5 Using consistent reference data in external applications	99
12.6 Velvet integration	100
12.7 Bowtie integration	104
12.8 Import and export of external application configurations	109
12.9 External applications in workflows	110
12.10 Running external applications	112
12.11 Troubleshooting external applications	114
13 Workflows	116
13.1 Installing and configuring workflows	116
13.2 Executing workflows	118
13.3 Updating workflows	118
14 Command line tools	122
15 Appendix	123
15.1 Use of multi-core computers	123
15.2 Troubleshooting	124
15.3 Database configurations	126
15.4 SSL and encryption	129
15.5 Non-exclusive Algorithms	131
15.6 DRMAA libraries	133
15.7 Consumable Resources	134
15.8 Third party libraries	136
15.9 External network connections	136
15.10 Monitoring	136
Bibliography	141

Chapter 1

Introduction

Welcome to *CLC Server 11.0*, a central element of the CLC product line enterprise solutions.

The latest version of the user manual can also be found in pdf format at <http://www.qiagenbioinformatics.com/support/manuals/>.

You can get an overview of the server solution in figure 1.1. The software depicted here is for research purposes only.

Using a server means that data can be stored centrally and analyses run on a central machine rather than a personal computer. After logging into the CLC Server from a Workbench, data on the server will be listed in the Workbench navigation area and analyses can be started as usual. The key difference is that when you are logged into a CLC Server from a Workbench, you will be get the choice of where to run the analysis: on the Workbench or on the CLC Server.

1.1 System requirements

The system requirements of *CLC Server* are:

Server system

- Windows 7, Windows 8, Windows 10, Windows Server 2012, and Windows Server 2016
- OS X 10.10, 10.11 and macOS 10.12, 10.13, 10.14
- Linux: RHEL 6.7 and later, SUSE Linux Enterprise Server 11 and later. The software is expected to run without problem on other recent Linux systems, but we do not guarantee this.
- 64 bit
- For CLC Server setups that include job nodes and grid nodes, those nodes must run the same type of operating system as the master CLC Server.
- File system that supports file locking. *Due to a bug in Java 10, CLC Genomics Server software in the 11.x release line should not be installed on a GPFS file system.*

Server hardware requirements



Figure 1.1: An overview of the server solution. Note that not all features are included with all license models.

- Intel or AMD CPU required
- Computer power: 2 cores required. 8 cores recommended.
- Memory: CLC Genomics Server: 4 GB RAM required, 16 GB RAM recommended.
- Disk space: 500 GB required. More needed if large amounts of data are analyzed.

Special memory requirements for working with genomes

The numbers below give minimum and recommended amounts for systems running mapping and analysis tasks. The requirements suggested are based on the genome size.

- **E. coli K12 (4.6 megabases)**
 - Minimum: 2 GB RAM
 - Recommended: 4 GB RAM
- **C. elegans (100 megabases) and Arabidopsis thaliana (120 megabases)**

- Minimum: 2 GB RAM
- Recommended: 4 GB RAM
- **Zebrafish (1.5 gigabases)**
 - Minimum: 2 GB RAM
 - Recommended: 4 GB RAM
- **Human (3.2 gigabases) and Mouse (2.7 gigabases)**
 - Minimum: 6 GB RAM
 - Recommended: 8 GB RAM

Special requirements for de novo assembly

De novo assembly may need more memory than stated above - this depends both on the number of reads and the complexity and size of the genome. See http://resources.qiagenbioinformatics.com/white-papers/White_paper_on_de_novo_assembly_4.pdf for examples of the memory usage of various data sets.

Special requirement for the shared filesystem used by the job node setup or grid integration

The file locking mechanism is required to ensure that all nodes see the latest version of the data stored on the shared filesystem.

1.2 Licensing

Three kinds of license can be involved in running analyses on the *CLC Server*.

- **A license for the server software itself.** This is needed for running analyses via the server. The license will allow a certain number of open sessions. This refers to the number of active, individual log-ins from server clients such as Workbenches, the Command Line Tools, or the web interface to the server. The number of sessions is part of the agreement with QIAGEN when you purchase a license. The manual chapter about installation provides information about how to obtain and deploy the license for the server.
- **A license for the Workbench software.** The Workbench is used to launch analyses on the server and to view the results. Find the user manuals and deployment manual for the Workbenches at <http://www.qiagenbioinformatics.com/support/manuals/>.
- **A network license if you will be submitting analyses to grid nodes.** This is explained in detail in section 6.3.5.

1.3 CLC Genomics Server

The *CLC Genomics Server* is shipped with the following tools and analyses that can all be started from *CLC Genomics Workbench* and *CLC Server Command Line Tools*:

- Import

- Export
- Search for Reads in SRA
- Download Genomes and References management
- Classical Sequence Analysis
 - Create Alignment
 - K-mer Based Tree Construction
 - Create Tree
 - Model Testing
 - Maximum Likelihood Phylogeny
 - Extract Annotations
 - Extract Sequences
 - Motif Search
 - Translate to Protein
 - Convert DNA to RNA
 - Convert RNA to DNA
 - Reverse Complement Sequence
 - Reverse Sequence
 - Find Open Reading Frames
 - Download Pfam Database
 - Pfam Domain Search
- Molecular Biology Tools
 - Assemble Sequences
 - Assemble Sequences to Reference
 - Secondary Peak Calling
 - Find Binding Sites and Create Fragments
 - Add attB Sites
 - Create Entry clone (BP)
 - Create Expression clone (LR)
- BLAST
 - BLAST
 - BLAST at NCBI
 - Download BLAST Databases
 - Create BLAST Database
- Track Tools
 - Merge Annotation Tracks

- Convert to Tracks
 - Convert from Tracks
 - Remove Orphan Reference Variants
 - Annotate with Overlap Information
 - Filter Annotations on Name
 - Filter Based on Overlap
 - Create GC Content Graph Tracks
 - Create Mapping Graph Tracks
 - Identify Graph Threshold Areas
- Prepare Sequencing Data
 - QC for Sequencing Reads
 - Trim Reads
 - Demultiplex Reads
- Resequencing Analysis
 - Map Reads to Reference
 - Local Realignment
 - Merge Read Mappings
 - Remove Duplicate Mapped Reads
 - Extract Consensus Sequence
 - Identify Known Mutations from Sample Mappings
 - Basic Variant Detection (Variant Detectors)
 - Fixed Ploidy Variant Detection (Variant Detectors)
 - Low Frequency Variant Detection (Variant Detectors)
 - Copy Number Variant Detection (CNVs)
 - InDels and Structural Variants
 - QC for Targeted Sequencing
 - QC for Read Mapping
 - Whole Genome Coverage Analysis
 - Filter Variants on Custom Criteria
 - Filter against Known Variants
 - Remove Marginal Variants
 - Remove Reference Variants
 - Remove Variants Present in Control Reads
 - Annotate from Known Variants
 - Remove Information from Variants
 - Annotate with Conservation Scores
 - Annotate with Exon Numbers

- Annotate with Flanking Sequences
 - Identify Shared Variants
 - Identify Enriched Variants in Case vs Control Samples
 - Trio Analysis
 - Amino Acid Changes
 - Predict Splice Site Effect
 - GO Enrichment Analysis
 - Download 3D Protein Structure Database
 - Link Variants to 3D Protein Structure
- RNA-Seq Analysis
 - RNA-Seq Analysis
 - Create Combined RNA-Seq Report
 - PCA for RNA-Seq
 - Differential Expression in Two Groups
 - Differential Expression for RNA-Seq
 - Create Heat Map for RNA-Seq
 - Create Expression Browser
 - Create Venn Diagram for RNA-Seq
 - Gene Set Test
- Microarray and Small RNA Analysis
 - Extract and Count
 - Annotate and Merge Counts
 - Create Box Plot
 - Hierarchical Clustering of Samples
 - Principal Component Analysis
 - Empirical Analysis of DGE
 - Proportion-based Statistical Analysis
 - Gaussian Statistical Analysis
 - Create MA Plot
 - Create Scatter Plot
 - Create Histogram
- Epigenomics Analysis
 - Transcription Factor ChIP-Seq
 - Annotate with Nearby Gene Information
 - Map Bisulfite Reads to Reference
 - Call Methylation Level
 - Create RRBS-fragment Track

- De Novo Sequencing
 - De Novo Assembly
 - Map Reads to Contigs
- Utility Tools
 - Extract Annotations
 - Sample Reads
 - Extract Reads
- Legacy Tools
 - Compare Sample Variant Tracks
 - Download Reference Genome Data
 - Identify Differentially Expressed Genes Groups and Pathways
 - Merge Overlapping Pairs
 - Add Fold Changes
 - Add Information from Overlapping Genes
 - Roche 454
 - Trim Primers of Mapped Single Reads
 - Create Fold Change Track
 - Create Track from Experiment
 - Trim Primers of Mapped Paired End Reads

The functionality of the *CLC Genomics Server* can be extended by installation of Server plugins. The available plugins can be found at <http://www.qiagenbioinformatics.com/plugins/>.

Latest improvements

CLC Genomics Server is under constant development and improvement. A detailed list that includes a description of new features, improvements, bugfixes, and changes for the current version of *CLC Genomics Server* can be found at:

<http://www.qiagenbioinformatics.com/products/clc-genomics-server/latest-improvements/current-line/>.

Chapter 2

Installation

2.1 Quick installation guide

The following describes briefly the steps needed to set up *CLC Genomics Server* with links out to more detailed explanations for each step.

If you are looking for how to set up a *CLC License Server*, instructions can be found in the *CLC License Server* manual.

If you are going to set up execution nodes as well, please read section 6 first.

1. Download and run the server software installer file. When prompted during the installation process, choose to start the server (section 2.2).
2. Run the license download script distributed with the server software. This script can be found in the installation area of the software. (section 2.6). The script will automatically download a license file and place it in the server installation directory under the folder called `licenses`.
3. Restart the server (section 2.7).
4. Ensure the necessary port is open for access by client software for the server. The default port is 7777 .
5. Log into the server web administrative interface using a web browser using the username **root** and password **default** (section 3).
6. Change the root password (section 4.1).
7. Configure the authentication mechanism and optionally set up users and groups (section 4.2).
8. Add data locations (section 3.2).
9. Check your server setup using the **Check set-up** link in the upper right corner as described in section 15.2.1.
10. Your server should now be ready for client software to connect to and use.

2.2 Installing and running the Server

Getting the *CLC Server* software installed and running involves, at minimum, these steps:

1. Install the software.
2. Ensure the necessary port in the firewall is open.
3. Download a license.
4. Start the Server and/or configure it as a service.

All these steps are covered in this section of the manual.

Further configuration information, including for job nodes, grid nodes, and External Applications, are provided in later chapters.

Installing and running the *CLC Server* is straightforward. However, if you do run into troubles, please refer to the troubleshooting section in Appendix 15.2, which provides tips on how to troubleshoot problems yourself, as well as how to get help.

2.2.1 Installing the Server software

The installation can only be performed by a user with administrative privileges. On some operating systems, you can double click on the installer file icon to begin installation. Depending on your operating system you may be prompted for your password (as shown in figure 2.1) or asked to allow the installation to be performed.

- On Windows 8 and Windows 7, you will need to right click on the installer file icon, and choose to **Run as administrator**.
- For the Linux-based installation script, you would normally wish to install to a central location, which will involve running the installation script as an administrative user - either by logging in as one, or by prefacing the command with `sudo`. Please check that the installation script has executable permissions before trying to execute it.



Figure 2.1: Enter your password.

Next, you will be asked where to install the server (figure 2.2). If you do not have a particular reason to change this, simply leave it at the default setting. The chosen directory will be referred to as the *server installation directory* throughout the rest of this manual.

The installer allows you to specify the maximum amount of memory the *CLC Server* will be able to utilize (figure 2.3). The range of choice depends on the amount of memory installed on your

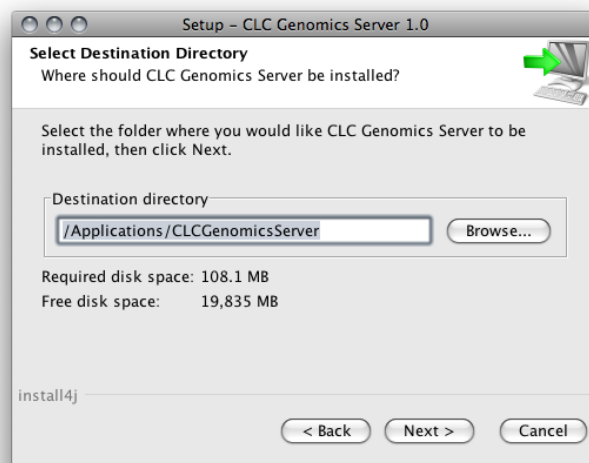


Figure 2.2: Choose where to install the server. Exemplified here with **CLC Genomics Server**

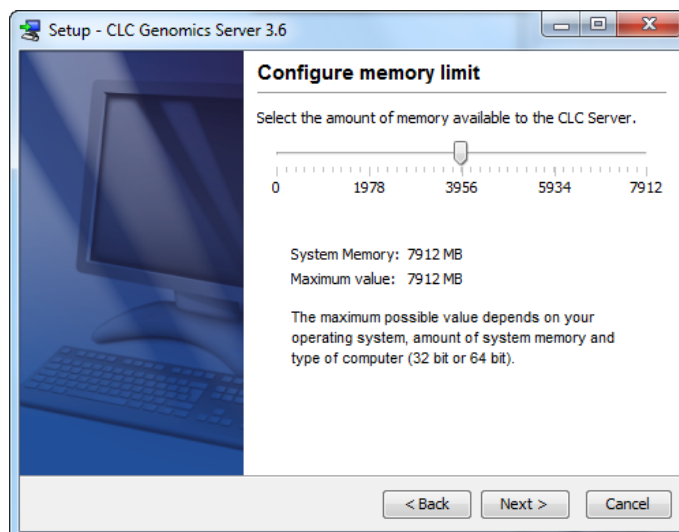


Figure 2.3: Choose the maximum amount of memory used by the server.

system and on the type of machine used. If you do not have a reason to change this value you should simply leave it at the default setting.

If you are installing the Server on a Linux or Mac system, you are offered the option to specify a user account that will be used to run the *CLC Server* process. Having a specific, non-root user for this purpose is generally recommended. On a standard setup, this would have the effect of adding this username to the service scripts, which can then be used for starting up and shutting down the *CLC Server* service and setting the ownership of the files in the installation area. Downstream, the user running the *CLC Server* process will own files created in File Locations, for example, after data import or data analyses.

If you are installing the server on a Windows system you will be able to choose if the service is started manually or automatically by the system.

The installer will now extract the necessary files.

On a Windows system, if you have chosen that the service should be started automatically, the service should also start running at this point. On Linux or Mac, if you have chosen the option to start the system at the end of installation, the service should also have started running. Please note that if you do not already have a license file installed, then the *CLC Server* process will be running in a limited capacity at this point. Downloading a license is described in section 2.6.

Information on stopping and starting the *CLC Server* service is provided in section 2.7.

2.3 Installation modes - console and silent

Two installation modes are available to support efficient installation of the Workbench software.

- **Console mode** This mode is particularly useful when installing Workbenches onto remote systems. On Linux, this mode is enabled by using the option `-c` when launching the installer from the command line. On Windows the option is `-console`.
- **Silent mode** This mode supports hands-off installation. Default answers to all prompts are used, although a non-default installation directory can be specified if desired (see below). Silent mode is activated using the `-q` parameter when launching the installer from the command line. On Windows, the `-console` option can be appended after `-q`, that is, as the second parameter, to ensure output to the console.

If desired, you can **specify the directory to install the software to** when running the installer in silent mode. Do this adding the `-dir` option to the command line.

On Windows, the `-console` and the `-dir` options only work when the installer is run in silent mode.

The following is an example of a command that would install a Workbench into the directory "c:\bioinformatics\clc" on a Windows system using silent mode with console output :

```
CLCMainWorkbench_8_0.exe -q -console -dir "c:\bioinformatics\clc"
```

On a Linux system, a similar command to install into the directory "/opt/clcgenomicsworkbench11" could look like:

```
./CLCGenomicsWorkbench_11_0_1_64.sh -c -q -dir /opt/clcgenomicsworkbench11
```

The `-q` and the `-console` options work for the uninstall program as well.

2.4 Upgrading an existing installation

For a single *CLC Server*, the steps we recommend when upgrading to a new version are:

- Stop the *CLC Server* service after making sure that nobody is using the server. Mechanisms to help with this, including sending a message to users logged into the *CLC Server*, can be found in section 7.3. Getting information about who is logged in is described in section 7.1.
- Install the *CLC Server* software in the same directory the existing version was installed in. All settings will be maintained, for example, the locations data are stored, Import/Export directories, BLAST database locations, Users and Groups, and External Application settings.

If you have a CLC Job Node setup, you will also need to upgrade the CLC Server software on each job node. Upgrading the software itself on each node is all you need to do. Configurations for job nodes, as well as new or updated plugins, are pushed to them by the master node.

When upgrading between major versions, there are extra steps to be taken. These are described in section [2.4.1](#). Major version lines are denoted by the first number in the version. For example, upgrading from software with version 10.0 to version 11.0 involves an upgrade to a new major version line.

2.4.1 Upgrading between major versions

There are a few extra steps to take beyond those outlined in section [2.4](#) when upgrading to a new major version line.

- An updated license file needs to be downloaded (see section [2.6](#)), and the service restarted.
- All users of client software (CLC Workbenches and the CLC Server Command Line Tools) must upgrade their software. Corresponding and compatible software versions are listed at the bottom of the Latest Improvement listings for a given server version. e.g. for the latest release, this can be found at: <http://www.qiagenbioinformatics.com/products/clc-genomics-server/latest-improvements/current-line/>.
- All plugins installed on the CLC Server need to be updated. See section [10](#). Plugins can be downloaded from <http://www.qiagenbioinformatics.com/plugins/>.
- **On job nodes**, any new tools included in the server upgrade will need to be enabled for the nodes you wish them to be run on. New tools are initially disabled on all job nodes to avoid interfering with a setup where certain nodes are dedicated to running specific types of jobs. Read more about enabling tools on job nodes in section [6.2.3](#).

Important note when upgrading on macOS to CLC Genomics Server 11.0 or higher from earlier major release lines A flag in the CLCGenomicsServer.vmoptions file must be removed when upgrading in place from CLC Genomics Server 10.x or earlier to CLC Genomics Server 11.x or later on macOS. Please delete "-d64" from the CLCGenomicsServer.vmoptions file, which can be found in the CLC Genomics Server installation area and then restart the CLC Genomics Server service. The -d64 option is not supported by recent versions of java. Its inclusion in the vmoptions file on macOS systems will stop the CLC Genomics Server 11.0, and any later versions, from starting up.

2.5 Allowing access through your firewall

By default, the server listens for TCP-connections on port 7777 (see section [3.6](#) for info about changing this).

If you are running a firewall on your server system you will have to allow incoming TCP-connections on this port before your clients can contact the server from a Workbench or web browser. Consult the documentation of your firewall for information on how to do this.

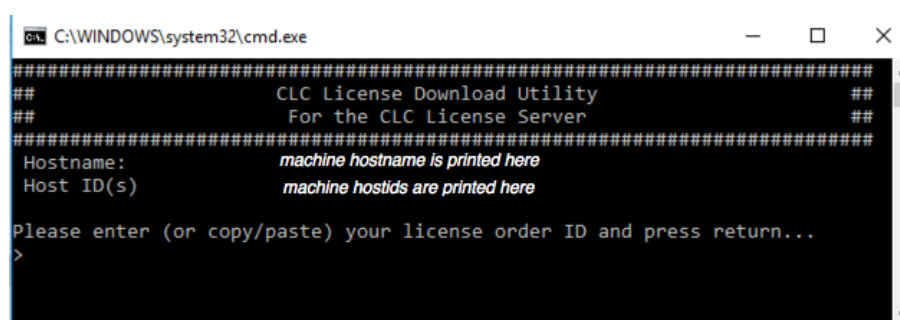
Besides the public port described above the server also uses an internal port on 7776. There is no need to allow incoming connections from client machines to this port.

2.6 Downloading a license

The CLC Server looks for licenses in the `licenses` folder in the installation area. Downloading and installing licenses is similar for all supported platforms, but varies in certain details. Please check the platform-specific instructions below for how to download a license file on the system you are running the CLC Server on or the section on downloading a license to a non-networked machine if the CLC Server is running on a machine without a direct connection to the external network.

2.6.1 Windows license download

License files are downloaded using the `licensedownload.bat` script. To run the script, right-click on the file and choose **Run as administrator**. This will present a window as shown in figure 2.4.



```

C:\WINDOWS\system32\cmd.exe
#####
##                  CLC License Download Utility                  ##
##                  For the CLC License Server                    ##
#####
Hostname:          machine hostname is printed here
Host ID(s):        machine hostids are printed here

Please enter (or copy/paste) your license order ID and press return...
>

```

Figure 2.4: Download a license based on the Order ID.

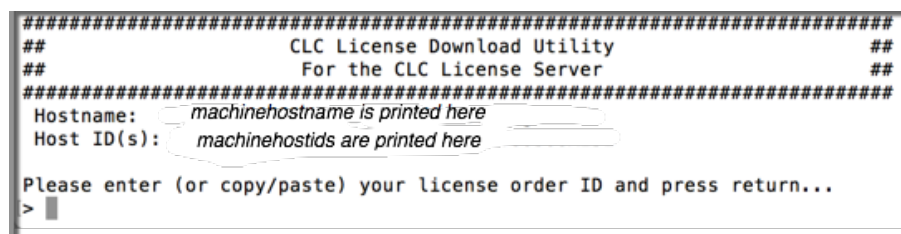
Paste the Order ID supplied by QIAGEN (right-click to **Paste**) and press Enter. Please contact ts-bioinformatics@qiagen.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

Restart the server for the new license to take effect (see how to restart the server in section 2.7.1).

2.6.2 macOS license download

License files are downloaded using the `downloadlicense.command` script. To run the script, double-click on the file. This will present a window as shown in figure 2.5.



```

#####
##                  CLC License Download Utility                  ##
##                  For the CLC License Server                    ##
#####
Hostname:          machinehostname is printed here
Host ID(s):        machinehostids are printed here

Please enter (or copy/paste) your license order ID and press return...
>

```

Figure 2.5: Download a license based on the Order ID.

Paste the Order ID supplied by QIAGEN and press Enter. Please contact ts-bioinformatics@qiagen.com

if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

Restart the server for the new license to take effect (see how to restart the server in section 2.7.2).

2.6.3 Linux license download

License files are downloaded using the `downloadlicense` script. Run the script and paste the Order ID supplied by QIAGEN Aarhus. Please contact ts-bioinformatics@qiagen.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense` script, it will create a new license file.

Restart the server for the new license to take effect (see how to restart the server in section 2.7.3).

2.6.4 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 after the *CLC Server* software has been installed.

- Determine the host ID of the machine the server software will be running on. This can be done by running the license download tool, which prints the host ID of the system to the terminal. The name of this tool depends on the system you are working on:
 - Linux: `downloadlicense`
 - Mac: `downloadlicense.command`
 - Windows: `licensedownload.bat`

In the case of a job or grid node setup, the host ID should be for the machine that will act as the *CLC Server* master node, as this is the machine the server license file will be stored on.

- 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 server license download web page¹:
<https://secure.clcbio.com/LmxWSv3/GetServerLicenseFile>
- Paste in your license order ID and the host ID that you noted down earlier into the relevant boxes on the webpage.
- Click on 'download license' and save the resulting `.lic` file.

¹For CLC Genomics Server 4.5.2 and lower, the URL is <http://licensing.clcbio.com/LmxWSv2/GetServerLicenseFile>. For server extensions, the URL is <https://secure.clcbio.com/LmxWSv3/GetLicenseFile>. Details about downloading licenses for server extensions can be found in the manual for each server extension product. To download a license file for a given product, the relevant URL must be used.

- On the machine with the host ID you specified when downloading the license file, place the license file in the folder called 'licenses' in the *CLC Server* installation directory. For job and grid node setups, this should be the machine acting as the master node.
- Restart the *CLC Server* software.

2.7 Starting and stopping the server

2.7.1 Microsoft Windows

On Windows based systems the *CLC Server* can be controlled through the *Services* control panel.

The service is named *CLC Genomics Server: CLCGenomicsServer*

Choose the service and click the start, stop or restart link as shown in figure 2.6.

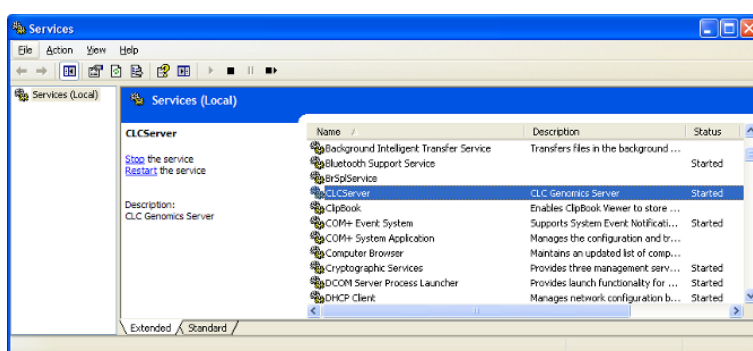


Figure 2.6: Stopping and restarting the server on Windows by clicking the blue links.

Once your server is started, you can use the Admin tab on the server web interface to manage your server operation (see section 7).

2.7.2 macOS

On macOS the server can be started and stopped from the command line.

Open a terminal and navigate to the *CLC Server* installation directory. Once there, the server can be controlled with the following commands.

Remember to replace *CLCServer*, in the commands listed below, with the name from the *CLC Genomics Server: CLCGenomicsServer*

To start the server run the command:

```
sudo ./CLCServer start
```

To stop the server run the command:

```
sudo ./CLCServer stop
```

To view the current status of the server run the command:

```
sudo ./CLCServer status
```

You will need to set this up as a service if you wish it to be run that way. Please refer to your operating system documentation if you are not sure how to do this.

Once your server is started, you can use the Admin tab on the server web interface to manage your server operation (see section 7).

2.7.3 Linux

You can start and stop the *CLC Server* service from the command line. You can also configure the service to start up automatically after the server machine is rebooted.

During installation of the *CLC Server* a service script is placed in `/etc/init.d/`.

This script will have a name reflecting the server solution, and it includes the name of the custom user account specified during installation for running the *CLC Server* process.

Starting and stopping the service using the command line:

To start the *CLC Server*:

```
sudo service CLCGenomicsServer start
```

To stop the *CLC Server*:

```
sudo service CLCGenomicsServer stop
```

To restart the *CLC Server*:

```
sudo service CLCGenomicsServer restart
```

To view the status of the *CLC Server*:

```
sudo service CLCGenomicsServer status
```

Start service on boot up:

On Red Hat Enterprise Linux and SuSE this can be done using the command:

```
sudo chkconfig CLCGenomicsServer on
```

How to configure a service to automatically start on reboot depends on the specific Linux distribution. Please refer to your system documentation for further details.

Troubleshooting

If the *CLC Server* is run as a service as suggested above, then the files in the installation area of the software and the data files created after installation in *CLC Server File Locations* will be owned by the user specified to run the *CLC Server* process. If someone starts up the *CLC Server* process as root (i.e. an account with super-user privileges) then the following steps are recommended to rectify the situation:

1. Stop the *CLC Server* process using the script located within the installation area of the *CLC Server* software. You can do that using the full path to this script, or by navigating to the installation area and running:

```
sudo ./CLCGenomicsServer stop
```

2. Change ownership recursively on all files in the installation area of the software and on all

areas specified as Server File Locations.

3. Start the *CLC Server* service as the specified user by using the service script:

```
sudo service CLCGenomicsServer start
```

4. In case the server still fails to start correctly it can be started in the foreground with output being written to the console to help identify the problem. It is done by running:



```
sudo ./CLCGenomicsServer start-launchd
```

Once your server is started, you can use the Admin tab of the web administrative interface to manage your server operation (see section 7).

If you want users to be able to use **External applications** (see chapter 12) on the server, the CLC External Applications Plugin needs to be installed in the Workbench.

Note: In order to install plugins and modules, the Workbench must be run in administrator mode. On Linux and Mac, it means you must be logged in as an administrator. On Windows, you can do this by right-clicking the program shortcut and choosing "Run as Administrator".

Plugins are installed and uninstalled using the plugin manager.

Help in the Menu Bar | Plugins... () or Plugins () in the Toolbar

From within the Plugins manager (see figure ??), choose the Download Plugins tab and click on CLC External Applications Plugin. Then click on the button labeled **Download and Install**.

2.8 Installing the database

This section pertains to the installation and configuration of an SQL database to store data imported into the *CLC Server*. This is relevant only if the database add-on has been purchased.

2.8.1 Download and install a Database Management System

If you do not have access to an existing installation of a Database Management System (*DBMS*) you will have to download and install one. *CLC Bioinformatics Database* can be used with a number of different DMBS implementations. Choosing the right one for you and your organization depends on many factors such as price, performance, scalability, security, platform-support, etc.

Information about the supported solutions are available via the links below.

- MySQL: <http://dev.mysql.com/downloads/>
- PostgreSQL: <http://www.postgresql.org/>
- Microsoft SQL Server: <http://www.microsoft.com/SQL/>
- Oracle: <http://www.oracle.com/>

You will need to make the appropriate JDBC driver available to the *CLC Server*. See section 15.3 for details and for additional configuration information for certain DBMSs.

2.8.2 Create a new database and user/role

Once your DBMS is installed and running you will need to create a database for containing your CLC data. We also recommend that you create a special database-user (sometimes called a database-role) for accessing this database.

Consult the documentation of your DBMS for information about creating databases and managing users/roles.

2.8.3 Initialize the database

Before you can connect to your database from a CLC Workbench or Server it must be initialized. The initialization creates the required tables for holding objects, and prepares an index used for searching. Initialization is performed with the CLC Bioinformatics Database Tool (see figure 2.7).

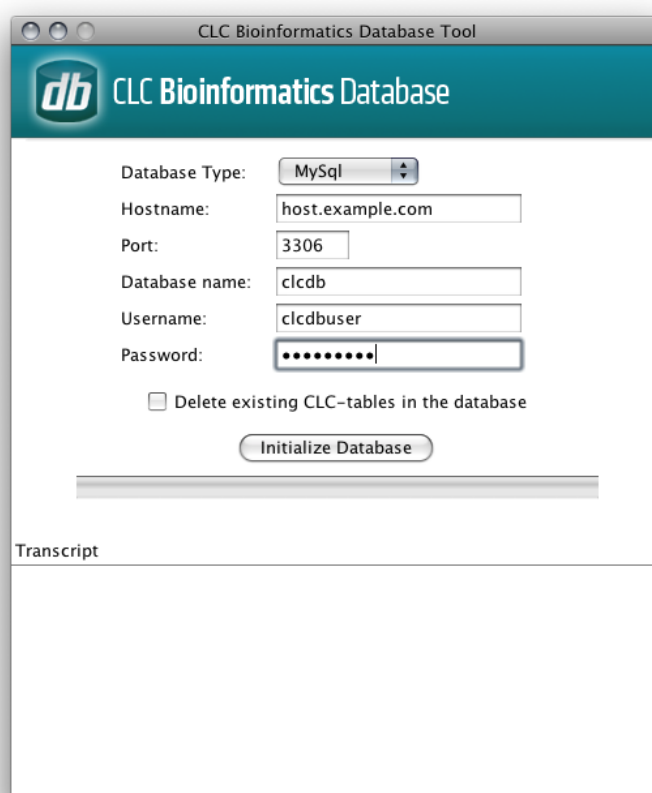


Figure 2.7: The CLC Bioinformatics Database tool

- Download the CLC Bioinformatics Database Tool from <https://www.qiagenbioinformatics.com/products/clc-bioinformatics-database-tool-direct-download/>
- Install the CLC Bioinformatics Database Tool on a client machine, and start the program.
- Fill in the fields with the required information.

- Hostname: The fully-qualified hostname of the server running the database.
NOTE: The same hostname must be used every time you connect to the database
 - Port: The TCP/IP listening port on the database server
 - Database name: The name of the database you created in the previous section
 - Username: the name of the user/role you created in the previous section
 - Password: the password for the user/role.
- To re-initializing an existing CLC database you must check the "Delete Existing..." checkbox.
NOTE: ANY DATA ALREADY STORED IN THE CLC DATABASE WILL BE DELETED.
 - Click the Initialize Database button to start the process.

While the program is working the progress-bar will show the status and the transcript will show a log of actions, events and problems. If anything goes wrong, please consult the transcript for more information. If you need assistance, please contact ts-bioinformatics@qiagen.com, and include the contents of transcript.

If the initialization is successful, the status bar will display this message: *Database successfully initialized*. You can now close the CLC Bioinformatics Database Tool.

Chapter 3

Basic configuration

3.1 Logging into the administrative interface

The administrative interface for a running *CLC Server* is accessed via a web browser. Most configuration occurs via this interface. Simply type the host name of the server machine you have installed the *CLC Server* software on, followed by the port it is listening on. Unless you change it, the port number is 7777. An example would be

```
http://clccomputer:7777/ or http://localhost:7777/
```

The default administrative user credentials are:

- **User name:** root
- **Password:** default

Use these details the first time you log in. We recommend that you change this password.

Details of how to change the administrative user password is covered in section [4.1](#).

3.2 Adding locations for saving data

Before using the server for analysis, the areas for storing data imported into or created by the software need to be specified.

On most server setups, this involves configuring a file system location, as described in section [3.2.1](#).

For *CLC Server* setups with a license that supports the add-on *CLC Bioinformatics Database*, database locations can also be added, as described in section [3.2.4](#).

3.2.1 File system locations

Data storage configuration is done via the administrative web interface. When logged in as a user with administrative privileges, navigate to the *Admin* tab, click on the *Main configuration* tab, and then click on the **File system locations** heading to expand that section. See figure [3.1](#).

The options available when configuring a file system location are described below. After making any changes, click on the **Save Configuration** button at the bottom of this area. Any file system locations that have been added should then appear in the list at the left hand side of the web client.

Add a new file system location Click on the Add New File Location button and then specify the path to the folder where data imported into or created by the *CLC Server* will be stored. The path provided should point to an *existing* folder on the server machine that the user running the server process has read and write access to.

If a file system location with the name *CLC_References* is configured, users logged into a *CLC Server* from a CLC Genomics Workbench will be able to download data directly to this server area using the Workbench's Reference Data Manager tool. Special conditions apply to this file system location. These are outlined in section 3.2.2.

Enable or disable access for all users The checkbox to the left of each file system location is used to control whether or not it should be available to users. Access is enabled by default. For example, in a CLC Workbench connected to the *CLC Server*, each enabled location is visible in the **Navigation Area**. Unchecking this box and saving the configuration makes the location unavailable for use. Disabled locations are not be visible in a CLC Workbench **Navigation Area**.

Remove a file system location Clicking on the **Remove Location** button beside a particular file system location removes it from the *CLC Server*. The underlying folder and its contents are **not** deleted. To re-enable access via the *CLC Server*, simply configure the same folder as a file system location again.

Rebuild the index The *CLC Server* maintains an index of all the elements in a data location. This is used when searching for data. There is no need to re-index when adding a new area as a file system location. Rebuilding the index is described in more detail in section 3.2.5.

Enable permissions Permissions can be configured for all file system locations, except if they are named *CLC_References*. The first step in doing this for a location is to check the **Permissions enabled** box just underneath it. After saving this change, that file system location and its contents will initially be available only to admin users. Read and write permissions can then be enabled for each group using a CLC Workbench client, as described in section 5.1.

3.2.2 Reference data management

Using the Reference Data Manager of the CLC Genomics Workbench, data can be downloaded directly to a CLC Genomics Server file system location called *CLC_References*.

To enable this, a folder named *CLC_References* must be available on a file system the *CLC Server* has access to and which the user running the *CLC Server* process has read and write access to. That folder must then be configured as a file system location, as described in section 3.2.1.

Special conditions exist for a *CLC_References* file system location:

The screenshot displays the 'Main configuration' page of the CLC Genomics Server. At the top, there are tabs for 'Element Info', 'History', and 'Admin'. The 'Main configuration' section is expanded, showing 'File system locations' and 'Data compression'. The 'File system locations' section contains a list of configured locations. Two locations are shown: '/path/to/CLCServerLoc1' and '/path/to/CLCServerLoc2'. Each location has a checkmark in the 'Path' field, a 'Permissions enabled' checkbox, and buttons for 'Remove Location' and 'Rebuild Index'. The 'Data compression' section has a checkbox for 'Save CLC data elements in a compressed format' which is checked. Below these sections are links for 'Import/export directories', 'Automatic recycle bin cleanup', 'Direct data transfer from client systems', and 'HTTP settings'. A 'Save Configuration' button is at the bottom.

Figure 3.1: File system location settings. The checkmark to the left of a configured location indicates it is available for use by those logged into the server. Internal data compression is enabled by default. This setting applies to all configured file system locations.

- All users logged into the CLC Genomics Server can see and use all data stored in *CLC_References*.
- All users logged into the CLC Genomics Server from a CLC Genomics Workbench can download data to this area using the Workbench Reference Data Manager.
- Only administrative users can delete data in this area using the CLC Genomics Workbench Reference Data Manager.
- No user, including admin users, can delete data stored in this area via the Navigation Area of the CLC Genomics Workbench.
- Data in this area can be deleted via the Element Info tab of the CLC Genomics Server web administrative interface.
- Custom permissions cannot be set on a *CLC_References* file system location. The checkbox enabling permissions should not be selected. If it is, only administrative users will be able to read or write to this area using the CLC Genomics Workbench Reference Data Manager.

Further information about the Reference Data Manager, including how to select to download to a server file location, can be found in the CLC Genomics Workbench manual: <http://>

resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=References_management.html.

Important points about CLC Server file system locations

- Folders added as file system locations should be **dedicated for use** by the CLC Server and should be directly accessed only by the CLC Server.
- The underlying file system must support file locking.
- All the data written to file system locations by the CLC Server will be in clc format and will be owned by the user that runs the CLC Server process.
- Files should **not** be moved manually into folders designated as CLC Server file system locations, or their subfolders, for example using standard operating system's command tools, drag and drop, and so on.
- Areas designated as file system locations should **not** overlap. That is, if a folder has been designated as a file system location, it should not be a subfolder of another area also designated as a file system location, or vice versa. Overlapping file system locations lead to problems with data indexing, which in turn leads to problems finding the stored data. Further information about indexing can be found in section 3.2.5.

File locations for job node set-ups

When you have a job node set-up, all the job node computers need to have access to the same data location folder. This is because the job nodes will write files directly to the folder rather than passing through the master node (which would be a bottleneck for big jobs). Furthermore, the user running the server must be the same for all the job nodes and it needs to act as the same user when accessing the folder no matter whether it is a job node or a master node.

The data location should be added **after** the job nodes have been configured and attached to the master node. In this way, all the job nodes will inherit the configurations made on the master node.

One relatively common problem faced in this regard is *root squashing* which often needs to be disabled, because it prevents the servers from writing and accessing the files as the same user - read more about this at http://nfs.sourceforge.net/#faq_b11.

You can read more about job node setups in section 6.

3.2.3 Enabling and disabling internal compression of CLC data

CLC data stored using internal compression takes less space. An option is provided under the **Data compression** heading to turn off internal data compression. See figure 3.1. Enabling data compression may impose a performance penalty depending on the characteristics of the hardware used. This penalty is typically small, and we generally recommend that this option remains enabled.

This setting applies to all configured file system locations. Any change will apply to data imported into or created after the change is made. Existing data is not affected.

Internal data compression and data compatibility Internal compression was introduced with the CLC Genomics Server 11.0 and corresponding client software, the CLC Genomics Workbench 12.0 and CLC Main Workbench 8.1. Data imported into or created by these systems and newer versions are compressed by default. However, software older than this, including retired products, cannot read the compressed data format.

Data created using older versions of CLC software will not be stored in compressed format.

To facilitate sharing particular datasets with people using older versions of the software, an option is available when exporting the data to CLC or zip format to export without this compression.

3.2.4 Database locations

Adding an SQL database location for use by the *CLC Server* is possible after the relevant add-on has been purchased and a license supporting it has been installed.

Before adding a database location, you need to set up the database itself. This is described in section 2.8.

To set up a database location on the CLC Server,

- Open a web browser and navigate to the web administrative interface.
- Go to the *Admin* tab and open the *Main configuration* section.
- Under the **Database locations** heading, click the **Add New Database Location** button. This pops a window like that shown in figure 3.2.
- To configure the database location, enter the host and port information and select the database type. The *Database type* drop down list contains the types for which drivers are available. A connection string is generated from this. A custom connection string can be entered instead. Add the user name and password information for the user role on your Database Management System (DBMS), see section 2.8.

If an Oracle database driver is available to the *CLC Server*, two items will be presented in the Database type drop down list, as shown in figure 3.2. The one shown as "Oracle" is the traditional one, which uses the SID style (e.g. `jdbc:oracle:thin:@[HOST][:PORT]:SID`). The other, "Oracle Service", uses the thin-style service name (e.g. `jdbc:oracle:thin:@//[HOST][:PORT]/SERVICE`).

- Click the *Save Configuration* button to save the configuration.

The newly added database location should now appear in the **Navigation Area** in the left hand side of the window.

3.2.5 Rebuilding the index

The server maintains an index of all the elements in the data locations. The index is used when searching for data. For all locations you can choose to **Rebuild Index**. This should be done only when a new location is added or if you experience problems while searching (e.g. something is missing from the search results). This operation can take a long time depending on how much data is stored in this location.

Add new database location

☒ ▼ -MySQL-@

Host

Database type ☒ MySQL
Oracle
Oracle Service

Port

Database name

☐ Use connection string

Connection string

Username

Password

☒ Rebuild index when adding location (recommended)

To connect to a database,
please **install** and **select the appropriate JDBC driver**.

Figure 3.2: Adding a new database location. Here, two drivers are available to the CLC Server, a MySQL driver and an Oracle driver.

If you move the server from one computer to another, you need to move the index as well. Alternatively, you can re-build the index on the new server (this is the default option when you add a location). If the rebuild index operation takes too long and you would prefer to move the old index, simply copy the folder called `searchindex` from the old server installation folder to the new server.

The status of the index server can be seen in the **User Statistics** pane found in the **Status and Management** tab page showing information on where the index server resides and the number of locations currently being serviced.

3.3 Accessing files on, and writing to, areas of the server filesystem

There are situations when it is beneficial to be able to interact with (non-CLC) files directly on your server filesystem.

A common use case would be importing high-throughput sequencing data or large molecule libraries from folders where it is stored on the same system that your *CLC Server* is running on. This could eliminate the need for each user to copy large data files to the machine the CLC Workbench is running on before importing the data into a *CLC Server* data area.

Another example is if you wish to export data from CLC format to other formats and save those files on your server machine's filesystem (as opposed to saving the files in the system your Workbench is running on).

From the administrator's point of view, this is about configuring folders that are safe for the *CLC Server* to read and write to on the server machine system.

This means that users logged into the *CLC Server* from their Workbench will be able to access files in that area, and potentially write files to that area. Note that the *CLC Server* will be accessing the file system as the *user running the server process* - not as the user logged into the Workbench. This means that you should be careful when opening access to the server filesystem

in this way. Thus, only folders that do not contain sensitive information should be added.

Folders to be added for this type of access are configured in the web administration interface **Admin** tab. Under **Main configuration**, open the **Import/export directories** (Figure 3.3) to list and/or add directories.

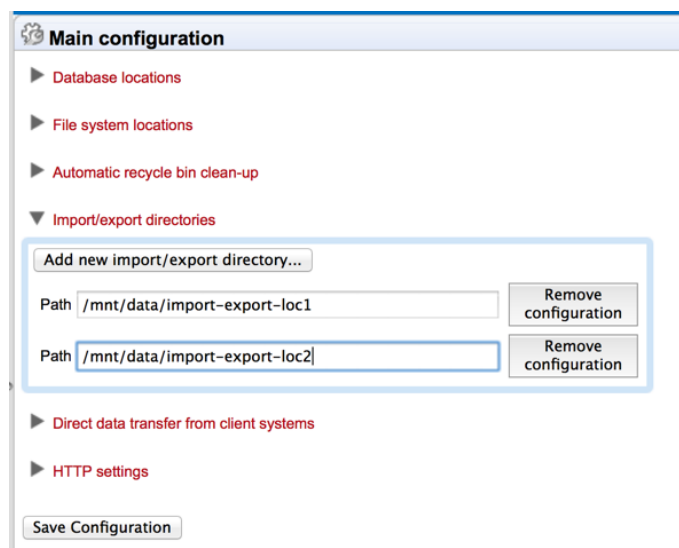


Figure 3.3: Folders on the server that should be available from the Workbench for browsing, importing data from and exporting data to.

Press the **Add new import/export directory** button to specify a path to a folder on the server. This folder and all its subfolders will then be available for browsing in the Workbench for certain activities (e.g. importing data functions).

The import/export directories can be accessed from the Workbench via the Import function in the Workbench. If a user that is logged into the CLC Server via their CLC Workbench wishes to import high throughput sequencing data, an option like the one shown in figure 3.4 will appear.

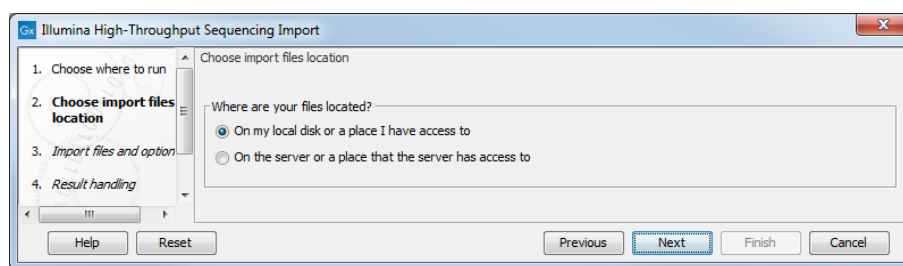


Figure 3.4: Deciding source for high-throughput sequencing data files.

On my local disk or a place I have access to means that the user will be able to select files from the file system of the machine their CLC Workbench is installed on. These files will then be transferred over the network to the server and placed as temporary files for importing. If the user chooses instead the option *On the server or a place the server has access to*, the user is presented with a file browser for the selected parts of the server file system that the administrator has configured as an Import/export location (an example is shown in figure 3.5).

Note: Import/Export locations should NOT be set to subfolders of any defined CLC file or data location. CLC file and data locations should be used for CLC data, and data should only be added or removed from these areas by CLC tools. By definition, an Import/Export folder is meant for

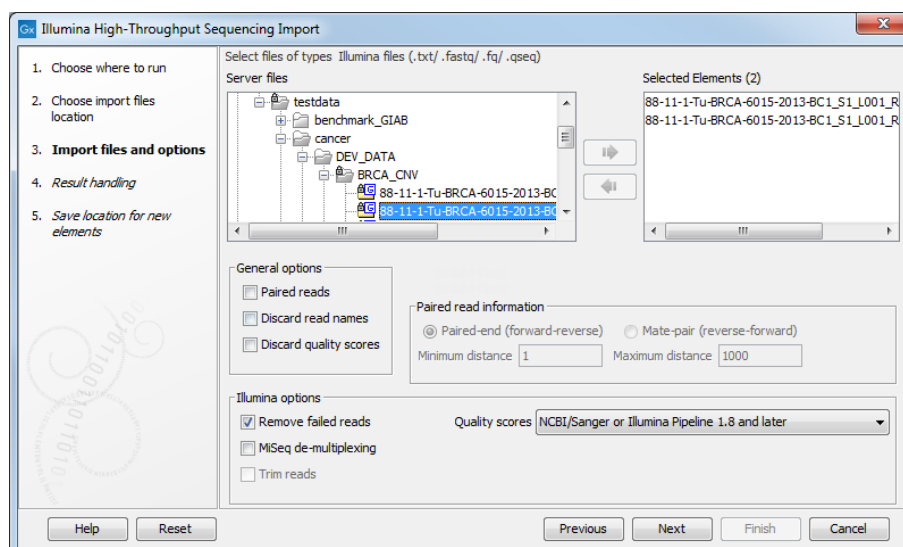


Figure 3.5: Selecting files on server file system.

holding non-CLC data, for example, sequencing data that will be imported, data that you export from the CLC Server, or BLAST databases.

3.4 Direct data transfer from client systems

Users of client systems are able, by default, to import data from a client system that the CLC Workbench or CLC Command Line Tools is installed on directly into a Server file or data location. The settings shown in figure 3.8 control whether this facility should be allowed and, if it should, then how the temporary data associated with Server data import should be handled.

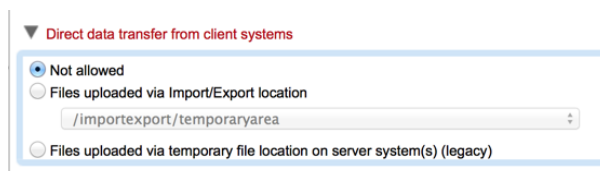


Figure 3.6: Specify whether data from client systems can be directly imported into Server locations and if so, where temporary data associated with these actions should be located.

If this facility is allowed on a system, we recommend the option to use a specified Import/Export location. To use this option, an Import/Export area must first be defined. It will then be available from the drop down list of areas to choose from (figure 3.9).

The use of default system temporary areas is deprecated and may be retired in future. We recommend that one of the other two options is chosen.

3.5 Direct data transfer from client systems

Users of client systems are able, by default, to import data from a client system that the CLC Workbench or CLC Command Line Tools is installed on directly into a Server file or data location. The settings shown in figure 3.8 control whether this facility should be allowed and, if it should, then how the temporary data associated with Server data import should be handled.

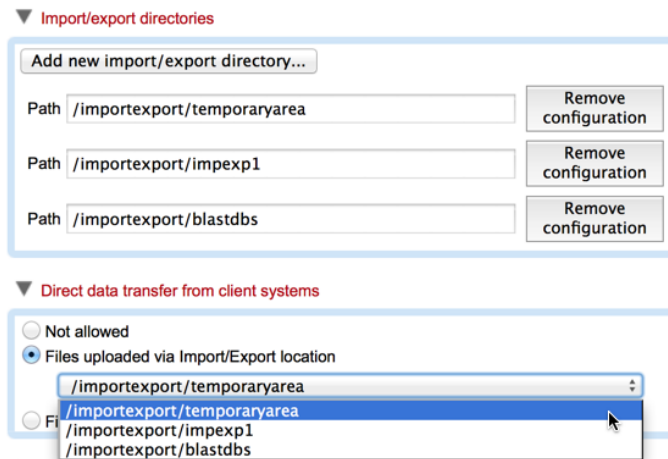


Figure 3.7: Specifying an Import/Export area for temporary data associated with the direct transfer of data from a client system into a CLC Server location.

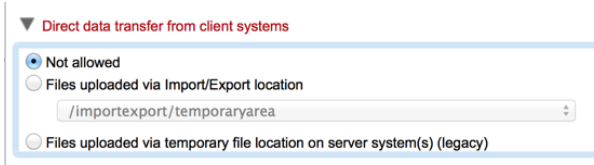


Figure 3.8: Specify whether data from client systems can be directly imported into Server locations and if so, where temporary data associated with these actions should be located.

If this facility is allowed on a system, we recommend the option to use a specified Import/Export location. To use this option, an Import/Export area must first be defined. It will then be available from the drop down list of areas to choose from (figure 3.9).

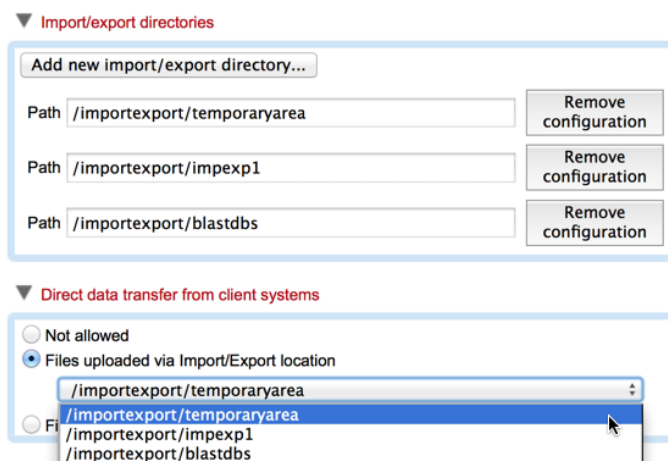


Figure 3.9: Specifying an Import/Export area for temporary data associated with the direct transfer of data from a client system into a CLC Server location.

The use of default system temporary areas is deprecated and may be retired in future. We recommend that one of the other two options is chosen.

3.6 Changing the listening port

The default listening port for the CLC Server is 7777. This has been chosen to minimize the risk of collisions with existing web-servers using the more familiar ports 80 and 8080. If you would like to have the server listening on port 80 in order to simplify the URL, this can be done in the following way.

- Navigate to the CLC Server installation directory.
- Locate the file called *server.xml* in the conf directory.
- Open the file in a text editor and locate the following section

```
<Connector port="7777" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

- Change the port value to desired listening port (80 in the example below)

```
<Connector port="80" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

- Restart the service for the change to take effect (see how to restart the server in section 2.7).
- Once the service is restarted, please log into the administrative interface and change the default port number in the "Master node port" field under Admin | Job distribution | Server setup, then click on **Save Configuration** button to save the new setting.

3.7 Changing the tmp directory

The CLC Server often uses a lot of disk space for temporary files. These are files needed during an analysis, and they are deleted when no longer needed. By default, these temporary files are written to your system default temporary directory. Due to the amount of space that can be required for temporary files, it can be useful to specify an alternative, larger, disk area where temporary files created by the CLC Server can be written.

In the *server installation directory* you will find a file called `CLCServer.vmoptions`, where `CLCServer` will be the name of your particular CLC server: `CLCGenomicsServer`

Open the file in a text editor and add a new line like this: `-Djava.io.tmpdir=/path/to/tmp` with the path to the new tmp directory. Restart the server for the change to take effect (see how to restart the server in section 2.7).

We highly recommend that the tmp area is set to a file system local to the server machine. Having tmp set to a file system on a network mounted drive can substantially affect the speed of performance.

3.7.1 Job node setup

The advice about having a tmp area being set on a local file system is true also for job nodes. Here, the tmp areas for nodes should **not** point to a shared folder. Rather, each node should have a tmp area with an identical name and path, but situated on a drive local to each node.

You will need to edit the `CLCServer.vmoptions` file on each job node, as well as the master node, as described above. This setting is **not** pushed out from the master to the job nodes.

3.8 Setting the amount of memory available for the JVM

When running the *CLC Server*, the Java Virtual Machine (JVM) needs to know how much memory it can use. This depends on the amount of physical memory (RAM) and can thus be different from computer to computer. Therefore, the installer investigates the amount of RAM during installation and sets the amount of memory that the JVM can use.

On **Windows** and **Linux**, this value is stored in a property file called `ServerType.vmoptions` (e.g. `CLCGenomicsServer.vmoptions`) which contains a text like this:

```
-Xmx8192m
```

The number (8192) is the amount of memory in megabytes the *CLC Server* is allowed to use. This file is located in the installation folder of the *CLC Server* software.

By default, the value is set to 50% of the available RAM on the system you have installed the software on.

You can manually change the number contained in the relevant line of the `vmoptions` file for your *CLC Server* if you wish to raise or lower the amount of RAM allocated to the Java Virtual Machine.

3.9 Limiting the number of cpus available for use


A number of the algorithms in the *CLC Server* will, in the case of large jobs, use all the cpu available on your system to make the analysis as fast as possible. The maximum number of cpu that can be used can be configured:

Master nodes Configured via the web administrative interface as described in section 6.4¹.

Job nodes Configured via the web administrative interface as described in section 6.2.

Grid nodes Controlled by the grid scheduler. Further information is available in section 6.3.8.

3.10 HTTP settings

Under the **Admin**  tab, click **Configuration**, and you will be able to specify HTTP settings. Here you can set the time out for the user HTTP session and the maximum upload size (when uploading files through the web interface).

3.11 Deployment of server information to CLC Workbenches

See the *Deployment manual* at <http://www.qiagenbioinformatics.com/support/manuals/> for information on pre-configuring the server log-in information when Workbench users log in for the first time.

¹The method of using of a `cpu.properties` file to limit CPU usage on master nodes and job nodes is deprecated.

Chapter 4

Managing users and groups

4.1 Logging in the first time and changing the root password

When the server is installed, you will be able to log in via the web interface using the following credentials:

- **User name:** `root`
- **Password:** `default`

Once logged in, you should as a minimum set up user authentication (see section 4.2) and data locations (see section 3.2) before you can start using the server.

For security reasons, you should change the root password (see figure 4.1):

Admin (🔧) | Authentication (🔑) Change root password

Note that if you are going to use job nodes, it makes sense to set these up before changing the authentication mechanism and root password (see section 6).

The screenshot shows a web interface with a top navigation bar containing 'Element Info', 'History', and 'Admin'. The 'Admin' tab is selected. Below the navigation bar is a 'Main configuration' section with an 'Authentication' sub-section. The 'Change root password' form is expanded, showing three password input fields: 'Current password', 'New password', and 'Verify password'. A green checkmark is next to the 'Verify password' field. A 'Change Root Password' button is located below the fields. A message below the button reads: 'If you are performing this change as root, please log out and log in again after changing the password.' Below this message is a section for 'Authentication mechanism'.

Figure 4.1: We recommend changing the root password. The verification of the root password is shown with the green checkmark.

4.2 User authentication using the web interface

When the server is installed, you can log in using the default root password (username=root, password=default).

Once logged in, you can specify which of the three modes of authentication should be used by going to:

Admin (⚙️) | Authentication (🔑) Authentication mechanism

The three different modes of authentication are shown in figure 4.2.

If LDAP or Active Directory are selected, a settings panel is revealed, where the details of the integration are specified. An example for LDAP settings is shown in figure 4.3.

Members of a group specified as an administrative group with login rights to the CLC Server will be configure the CLC Server using the functionality under the Admin tab of the web administrative interface, as well as set permissions on folders of data, as described in section 5. For the built-in authentication method, this means adding particular users to the built-in **admin** group. For Active Directory or LDAP, this means designating a group in the box labeled **Admin group name** and adding any users who should be administrators of the CLC Server to this group.

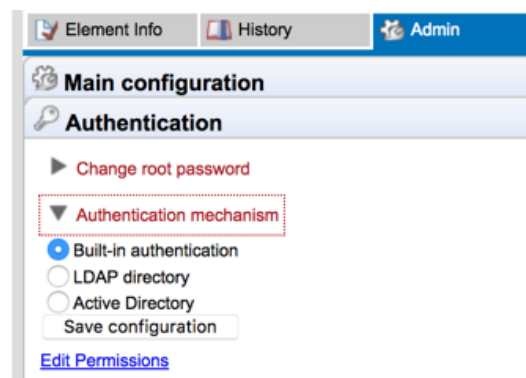


Figure 4.2: Three modes of user authentication are available. Clicking on the *Edit Permissions* link at the bottom opens up the *Global Permissions* tab, where access to the server and its functionality can be configured, as described in section 5.2.

4.2.1 Authentication options

Built-in authentication

This option will enable you to set up user authentication using the server's built-in user management system. This means that you create users, set passwords, assign users to groups and manage groups using the web interface (see section 4.2.2) or using the Workbench (see section 4.2.4). All the user information is stored on the server and is not accessible from other systems.

LDAP directory

This option will allow you to use an existing LDAP directory. This means that all information needed during authentication and group memberships is retrieved from the LDAP directory (figure 4.3).

Authentication

▼ Authentication mechanism

☐ Built-in authentication

☒ LDAP directory

☐ Active Directory

Hostname: host.example.com

Port: 389 Default if "ldaps://" is selected: 636. Else: 389

Encryption: ☒ Plain text Default: "Plain text"

☐ Forced Start TLS

☐ ldaps://

Disable SSL certificate check: ☐

Base DN: dc=example,dc=com

Admin group name: Default: admins

Cache timeout: 3600 Default: 3600 (seconds)

Users DN: ou=users (Base DN will be appended)

Groups DN: ou=groups (Base DN will be appended)

UID attribute: Default: uid

Group name attribute: Default: cn

Membership attribute: Default: memberUid

Bind DN: Leave empty to use anonymous bind for the initial lookup, used to get a user DN

Bind password:

DN to use for lookups: ☒ User DN Select the DN to be used for all search and read operations except for the initial one, for example user and email lookups. The Bind DN option is enabled for selection when Bind DN and password details have been entered above.

☐ Bind DN

Kerberos/GSSAPI Authentication: ☐

Kerberos realm: Leave empty to use default realm

Kerberos config file: Default: /etc/krb5.conf

Figure 4.3: LDAP settings panel.

If needed, the LDAP integration can use Kerberos/GSSAPI. Encryption options (Start TLS and LDAP over SSL) are available. If your LDAP server uses a certificate that is not generally trusted by the server system that the CLC Server software is running on, then it must be added to the truststore of the CLC Server installation (`CLC_SERVER_BASE/jre/lib/security/cacerts`, where `CLC_SERVER_BASE` is the server installations root location). This can be done with the `keytool` shipped with Java installations (also available in the `CLC_SERVER_BASE/jre/bin/keytool`), with a command like:

```
CLC_SERVER_BASE/jre/bin/keytool -import -alias \
  ldap_certificate -file LDAP_CERTIFICATE.cer -keystore \
  CLC_SERVER_BASE/jre/lib/security/cacerts -storepass changeit
```

Replace `LDAP_CERTIFICATE` with the path to the certificate your LDAP server uses for Start TLS/LDAPS connections. Replace `CLC_SERVER_BASE` with the path to the servers installation location.

For a node setup, this must be done for all job nodes as well.

Caution: If you update the server installation or reinstall the server, all imported certificates will be removed, and have to be imported again. You should also be aware that certificates have an expiration date, and will not be trusted after this date. Make sure to add a new certificate in advance of the expiration date.

The **DN to use for lookups** configuration allows you to choose which bind should be used for read

and search operations. If no bind DN have been entered an unauthenticated bind will be used to do the initial lookup (lookup of users DN based on the username), and all other read and search operations will be performed with users binds. If the **Bind DN** and **Bind password** have been filled in, you have the choice between using the 'Bind DN' or the 'User DN' for read and search operations, the 'Bind DN' will in this case always be used for the initial lookup.

Active Directory

This option will allow you to use an existing Active directory. This means that all information needed during authentication and group memberships is retrieved from the Active directory. Encryption options (Start TLS and LDAP over SSL) are available. Please see the notes about certificates in the LDAP section (see section 4.2.1) above for details.

4.2.2 Managing users and groups using built-in authentication

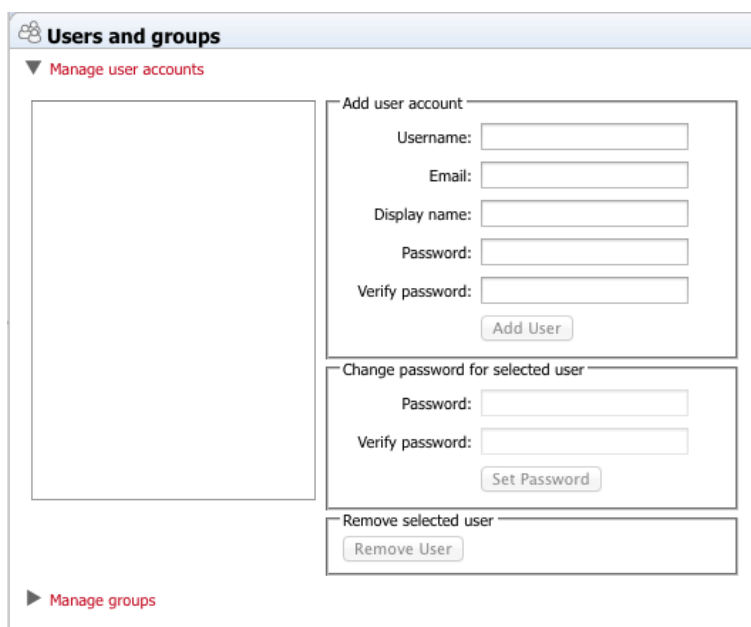
This information only applies if built-in authentication is being used.

Managing users via the web administrative interface

To create or remove users or change their password:

Admin (🔑) | Users and groups (👤) Manage user accounts

This will display the panel shown in figure 4.4.



The screenshot shows a web interface titled 'Users and groups' with a sub-section 'Manage user accounts'. On the left is a large empty box for a user list. On the right are three forms: 'Add user account' with fields for Username, Email, Display name, Password, and Verify password, plus an 'Add User' button; 'Change password for selected user' with fields for Password and Verify password, plus a 'Set Password' button; and 'Remove selected user' with a 'Remove User' button. At the bottom, there is a link to 'Manage groups'.

Figure 4.4: Managing users.

Managing groups via the web administrative interface

To create or remove groups or change group membership for users when using built-in authentication, go to:

Admin (🔑) | Users and groups (👤) Manage groups

This will display the panel shown in figure 4.5.

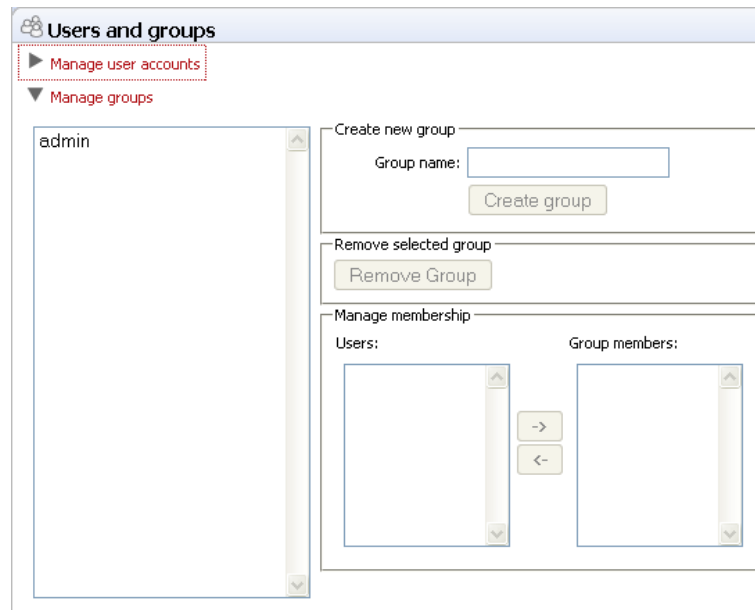


Figure 4.5: Managing users.

The same user can be a member of several groups.

Users who should have access to the administrative part of the server should be part of the "admin" group which is the only group with special permissions by default. The admin group already exists in a default setup of the CLC Server.

You will always be able to log in as the CLC Server root user, and this user has administrative level access rights.

4.2.3 User authentication using the Workbench

This information only applies if built-in authentication is being used. If LDAP or AD is being used, the menus described here will be disabled.

Users and groups can be managed through the Workbench by logging into the CLC Server as an administrative user and then going to the Workbench menu:

File | Manage Users and Groups

This will display the dialog shown in figure 4.6.

4.2.4 Managing users through the Workbench

This information only applies if built-in authentication is being used. If LDAP or AD is being used, the menus described here will be disabled.

Click the **Add** (+) button to create a new user. Enter the name of the user and enter a password. You will be asked to re-type the password. If you wish to change the password at a later time, select the user in the list and click **Change password** (🔑).

To delete a user, select the user in the list and click **Delete** (🗑).

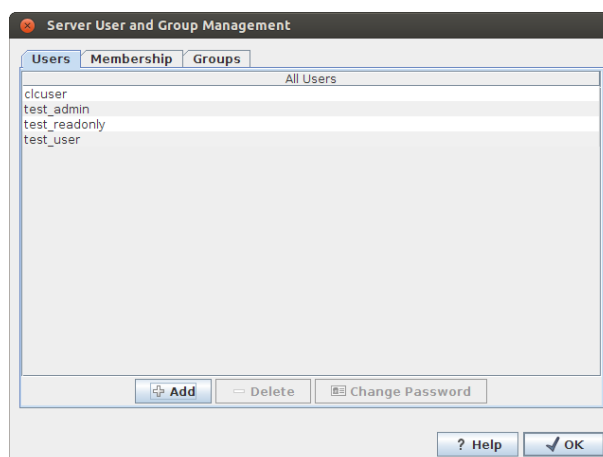


Figure 4.6: Managing users.

4.2.5 Managing groups through the Workbench

This information only applies if built-in authentication is being used. If LDAP or AD is being used, the menus described here will be disabled. Access rights are granted to groups, not users, so a user has to be a member of one or more groups to get access to the data location. Here you can see how to add and remove groups, and next you will see how to add users to a group.

Adding and removing groups is done in the **Groups** tab (see figure 4.7).

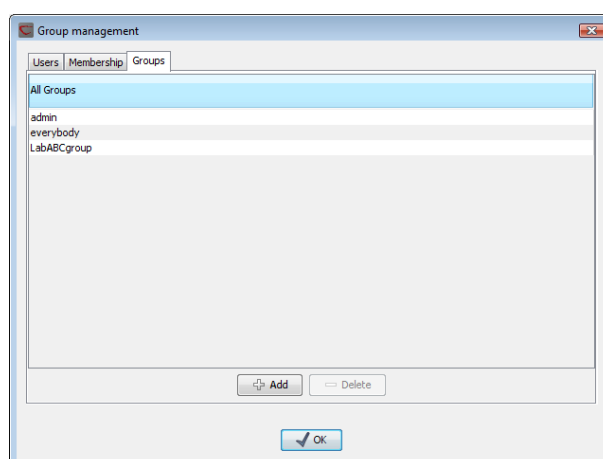


Figure 4.7: Managing groups.

To create a new group, click the **Add** (+) button and enter the name of the group. To delete a group, select the group in the list and click the **Delete** (=) button.

When a new group is created, it is empty. To assign users to a group, click the **Membership** tab. In the **Selected group** box, you can choose among all the groups that have been created. When you select a group, you will see its members in the list below (see figure 4.8). To the left you see a list of all users.

To add or remove users from a group, click the **Add** (➡) or **Remove** (⬅) buttons. To create new users, see section 4.2.4.

The same user can be a member of several groups.

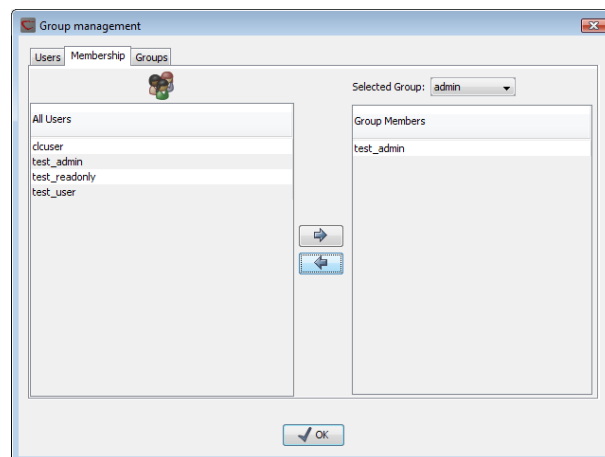


Figure 4.8: Listing members of a group.

Chapter 5

Access privileges and permissions


Server administrators can restrict access to members of specified groups at various levels:


- **Access to the CLC Server**, described in section 5.2.
- **Access to data in the server's file and data locations**, described in section 5.1.
- **Launching jobs** on the server, described in section 5.2. This includes setting permissions for data import, data export and for running specified analysis tasks, including built-in analyses, installed workflows and configured External Applications.
- **Access to import/export directories**, described in section 5.2.
- **Access to individual grid presets**, as described in section 5.2.

5.1 Controlling access to CLC Server data

The *CLC Server* uses folders as the basic unit for controlling access to data, and access is granted (or denied) to groups of users.

Members of groups can be granted two types of access on folders within a server location:

Read access Members of the designated group(s) can see the elements in the folder, open them and copy from them. Access can be through any client software, for example, via the CLC Server Command Line Tools or via a CLC Workbench, for example when browsing in the **Navigation Area**, searching, or when clicking the "Originates from" link in the **History** () of a data element.

Write access Members of the designated group(s) can make and **Save** () changes to an element, and new elements and subfolders can be created in that area.

For a user to be able to access a folder, they must have read access to all the folders above it in the hierarchy. In the example shown in figure 5.1, to access the *Sequences* folder, the user must have access to both the *Example Data* and *Protein* folders.

It is fine to just give write access to the final folder. For example, read access only could be granted to the *Example Data* and *Protein* folders, with read and write access granted to the *Sequences* folder.

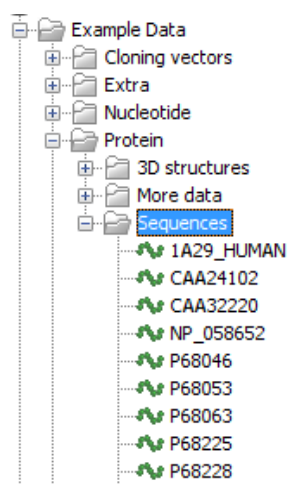


Figure 5.1: A folder hierarchy on the server.

At the point when permissions are enabled on a file system location, i.e. after the **Enable permissions** option described in section 3.2.1, has been checked, only the *CLC Server* root user or users in a configured admin group will have access to data stored in that file system location. Permissions are then set by the root or other admin user on the folders in that area, via a *CLC Workbench* acting as a client for the *CLC Server*, as described in the next section.

Please see 5.1.3 for further details about the system behavior if permissions are not enabled and configured.

5.1.1 Setting permissions on a folder

This step is done from within a *CLC Workbench*. Start up a copy of a *CLC Workbench* click on the menu option:

File | CLC Server Connection (S)

Log into the *CLC Server* as an administrative user.

You can then set permissions on folders within *File Locations* that have had permissions enabled, or on *Database Locations*, if you have a *CLC Bioinformatics Database*.

right-click the folder (F) | Permissions (P)

This will open the dialog shown in figure 5.2.

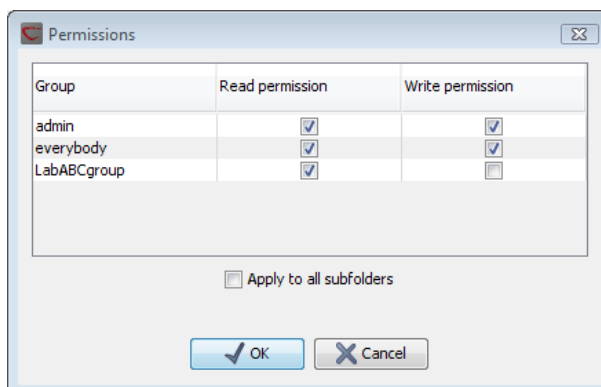


Figure 5.2: Setting permissions on a folder.

Set the relevant permissions for each of the groups and click **OK**.

If you wish to apply the permissions recursively, that is to all subfolders, check **Apply to all subfolders** in the dialog shown in figure 5.2. **Note** that this operation is usually only relevant if you wish to clean up the permission structure of the subfolders. **It should be applied with caution**, since it can potentially destroy valuable permission settings in the subfolder structure.

5.1.2 Recycle bin

When users delete data in the **Navigation Area** of the Workbench, it is placed in the recycle bin. When the data is situated in a *CLC Server* data location, the data will be placed in a recycle bin for that data location. Each user has an individual recycle bin containing the data deleted by that particular user that cannot be accessed by other users, except server administrators. This means that any permissions applied to the data prior to deletion are no longer in effect, and it is not possible to grant other users permission to see it while it is in the recycle bin. In summary, the recycle bin is a special concept that is not included in the permission control system.

An exception: Deletion of data held in a *CLC Server* file system location named *CLC_References* is different than for other file system locations. Please refer to section 3.2.2 for details.

Server administrators can access the recycle bins of other users through the Workbench:

right-click the data location (📁) | Location | Show All Recycle Bins

This will list all the recycle bins at the bottom of the location as shown in figure 5.3.

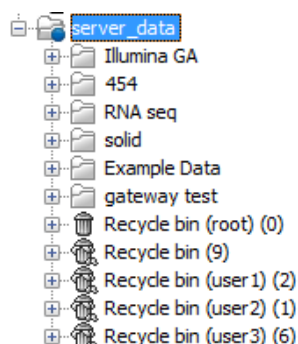


Figure 5.3: Showing all recycle bins.

The recycle bin without a name contains all the data that was deleted in previous versions of the *CLC Server* before the concept of a per-user recycle bin was introduced. This recycle bin can only be accessed by server administrators by selecting **Show All Recycle Bins**.

The administrator is also able to empty the recycle bin of a user:

right-click the recycle bin (🗑️) | Empty

All recycle bins can be emptied in one go:

right-click the data location (📁) | Location | Empty All Recycle Bins

Please note that these operations cannot be undone.

CLC Server can be set to automatically empty recycle bins when the data has been there for more than 100 days. This behavior can be controlled for each data location: Under the **Main configuration** heading, click the **Automatic recycle bin clean-up** header and click the **Configure**

button. This will allow you to disable the automatic clean-up completely or specify when it should be performed as shown in figure 5.4.

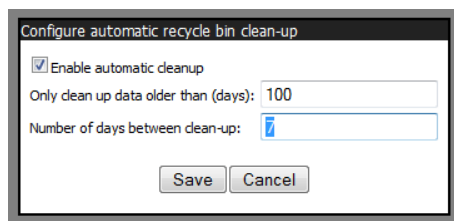


Figure 5.4: Automatic clean-up of the recycle bin.

Data deleted before the per-user recycle bin concept was introduced will be ignored by the automatic clean-up (this is the data located in the general recycle bin that is not labeled with a user name).

5.1.3 Technical notes about permissions and security

All data stored in *CLC Server* file system locations are owned by the user that runs the *CLC Server* process. Changing the ownership of the files using standard system tools is not recommended and will usually lead to serious problems with data indexing and hamper your work on the *CLC Server*.

One implication of the above ownership setup is that by default, (i.e. without permissions enabled), all users logging into the *CLC Server* are able to access all data within that file system location, and write data to that file system locations. All files created within such a file system location are then also accessible to all users of the *CLC Server*.

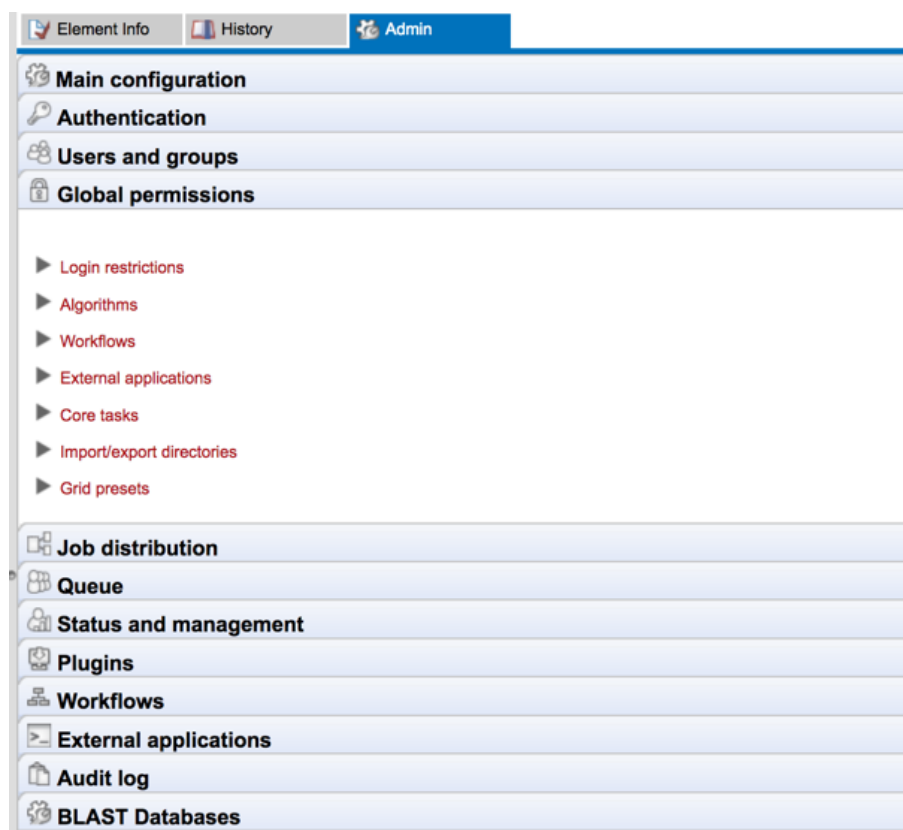
Group permissions on file system locations is an additional layer within the *CLC Server*, and is not part of your operating system's permission system. This means that enabling permissions, and setting access restrictions on *CLC* file system locations only affects users accessing data through *CLC* tools (e.g. using a Workbench, the *CLC* Command Line Tools, the *CLC Server* web interface or the Server API). If users have direct access to the data, using for example general system tools, the permissions set on the data in *CLC Server* has no effect.

5.2 Controlling access to the server, server tasks and external data

The configurations discussed in this section refer to settings under the **Global Permissions** section of the Admin tab in the *CLC Server* web administrative interface (figure 5.5).

Permissions can be set to restrict access to just the members of specified groups for the following areas:

- **Login restrictions** The ability to log into the *CLC Server*.
- **Algorithms** The analysis algorithms.
- **Workflows** Workflows installed on the server.
- **External applications** External tools configured as External Applications.

Figure 5.5: *Global permissions.*

- **Core tasks** Currently covers setting permissions on actions associated with the Standard Import tools. (High throughput sequence data import is handled via tools listed in the Algorithms section.)
- **Import/export directories** File system areas not part of the CLC data setup, which the CLC Server is able to access. These are described in section 3.3.
- **Grid presets** For grid node setups only: presets for sending jobs to a particular queue with particular parameters. Note that grid presets are identified by name. If you change the name of a preset under the Job Distribution settings section, then this, in effect, creates a new preset. In this situation, if you had access permissions previously set, you would need to reconfigure those settings for this, now new, preset.

You can specify which groups should have access to each of the above by opening the relevant section and then clicking the **Edit Permissions** button for each relevant element listed. A dialog appears like that in figure 5.6. If you choose **Only authorized users from selected groups**, you will be offered a list of groups that you can select (or de-select) to grant or restrict access to that functionality.

The default configuration is that all users have access to everything.

5.3 Customized attributes on data locations

Location-specific attributes can be set on all elements stored in a given data location. Attributes could be things like company-specific information such as LIMS id, freezer position etc. Attributes

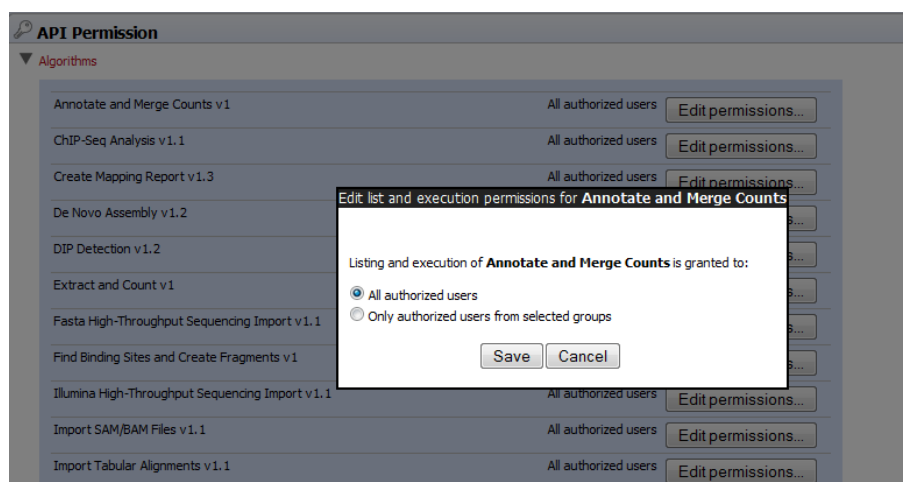


Figure 5.6: Setting permissions for an algorithm.

are set using a CLC Workbench acting as a client to the CLC Server.

Note that the attributes scheme belongs to a particular data location, so if there are multiple data locations, each will have its own set of attributes.

To configure which fields that should be available¹ go to the Workbench:

right-click the data location | Location | Attribute Manager

This will display the dialog shown in figure 5.7.

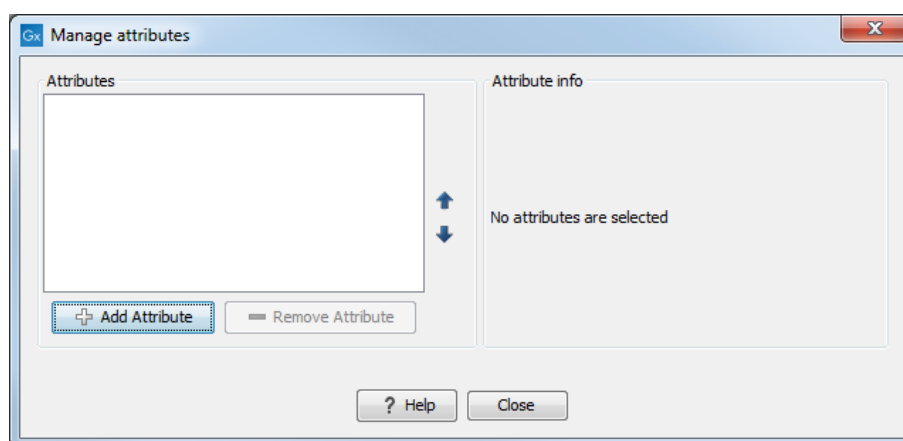


Figure 5.7: Adding attributes.

Click the **Add Attribute** (+) button to create a new attribute. This will display the dialog shown in figure 5.8.

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).

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

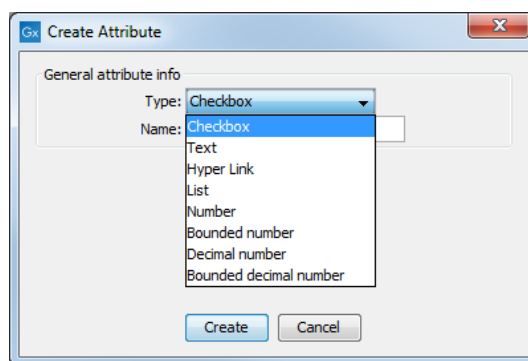


Figure 5.8: The list of attribute types.

- **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 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.

Lists are a little special, since you have to define the items in the list. When you choose to add the list attribute 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 5.9).

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

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.

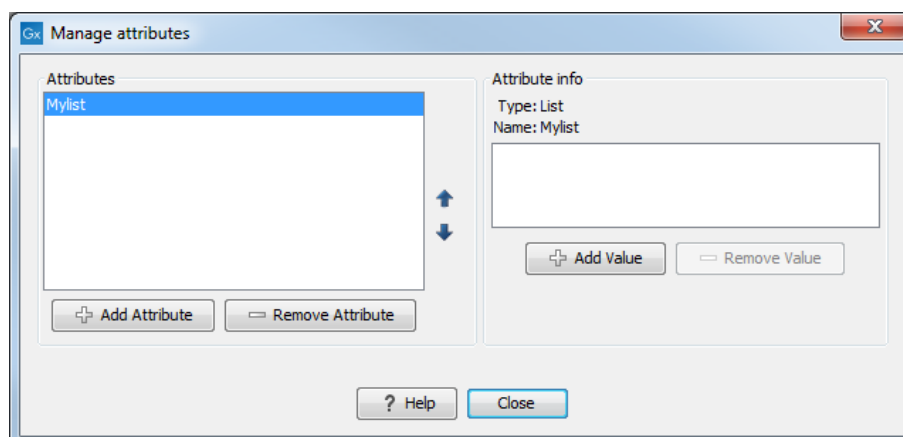


Figure 5.9: Defining items in a list.

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.

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.

5.3.1 Filling in values

When a set of attributes has been created (as shown in figure 5.10), the end users can start filling in information.

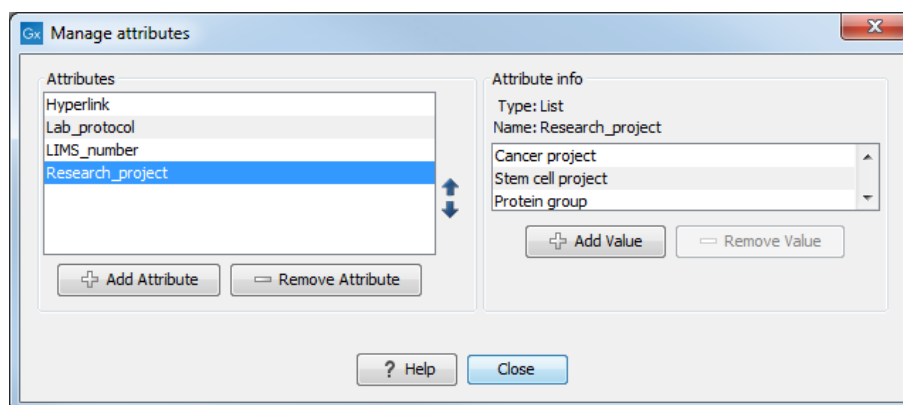


Figure 5.10: A set of attributes defined in the attribute manager.

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 5.11.

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).

The screenshot shows a web-based form for editing an element named '1WCL_A'. The form is organized into sections: 'Fixed Fields' (Name, Description, Metadata), 'Local Attribute Fields', and 'Research_project'. The 'Research_project' dropdown is set to 'Cancer project'. The 'Hyper_link' is 'http://www.clcbio.com'. The 'Is_confirmed' checkbox is checked. The 'Lab_protocol' is 'P123'. The 'LIMS_number' is '412'. The 'Location' is 'Lab 23'. The 'Patient_number' is '651'. Each attribute has an 'Edit' or 'Clear' link next to it.

Figure 5.11: Adding values to the attributes.

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 5.12).

This screenshot shows a zoomed-in view of the 'Local Attribute Fields' section. The 'Research_project' dropdown is set to 'Cancer project'. The 'Hyper_link' attribute is shown with the text 'Not set' in red next to it, indicating it has not been assigned a value.

Figure 5.12: An attribute which has not been set.

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 cannot search for it, even if it looks like it has a value. In figure 5.12, 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 the Workbench manuals at <https://www.qiagenbioinformatics.com/support/manuals/>).

5.3.2 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.

5.3.3 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 5.13).

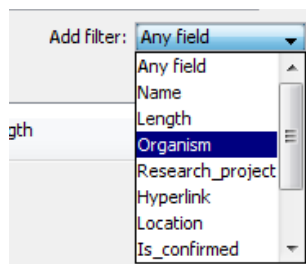


Figure 5.13: The attributes from figure 5.10 are now listed in the search filter.

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 5.14).

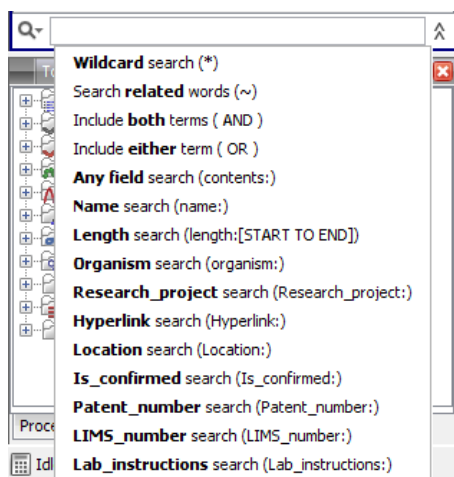


Figure 5.14: The attributes from figure 5.10 are now available in the Quick Search as well.

Read more about search here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Local_search.html.

Chapter 6

Job distribution

The *CLC Server* has the concept of *distributing jobs to nodes*. This means having a master server with the primary purpose of handling user access, serving data to users and starting jobs, and one or more nodes, which execute the jobs submitted to them. This chapter describes the server setups that are available for the *CLC Server* as well as some job running options available for single servers and those running *CLC* job nodes.

6.1 Introduction to servers setups

The three models for running the *CLC Server* are:

- **Model I: Master server with dedicated job nodes.** In this model, a master server submits *CLC* jobs directly to machines running the *CLC Server* for execution. In this setup, a group of machines (from two upwards) have the *CLC Server* software installed on them. The system administrator assigns one of them as the *master node*. The master controls the queue and distribution of jobs and compute resources. The other nodes are *job nodes*, which execute the computational tasks they are assigned. This model is simple to set up and maintain, with no other software required. However, it is not well suited to situations where the compute resources are shared with other systems because there is no mechanism for managing the load on the computer. This setup works best when the execute nodes are machines dedicated to running a *CLC Server*. Further details about this setup can be found in section 6.2
- **Model II: Master server submitting to grid nodes.** In this model, a master server submits tasks to a local third party scheduler. That scheduler controls the resources on a local computer cluster (grid) where the job will be executed. This means that it is the responsibility of the native grid job scheduling system to start the job. When the job is started on one of the grid nodes, a **CLC Grid Worker**, which is a stand-alone executable including all the algorithms on the server, is started with a set of parameters specified by the user. Further details about this setup can be found in section 6.3.
- **Model III: Single Server setup.** In this model, the master and execution node functionality is carried out by a single *CLC Server* instance.

Figure 6.1 shows a schematic overview.

For models I and II, the master server and job nodes, or master server and grid nodes must run on the same type of operating system. It is not possible to have a master server running Linux and a job node running Windows, for example.

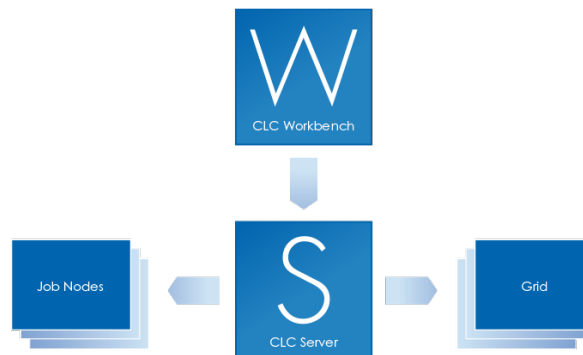


Figure 6.1: An overview of the job distribution possibilities.

6.2 Model I: Master server with dedicated job nodes

6.2.1 Overview: Model I

This setup consists of two types of CLC Server instances:

1. **A master node** - a machine that accepts jobs from users and then passes them to job nodes for execution.
2. **Job nodes** - machines running the *CLC Server* that accept jobs directly from a master node.

The general steps for setting up this model are:

1. Install the *CLC Server* software on all the machines involved. (See section 2.2.)
2. Install the license on the machine that will act as the master node. (See section 2.6.)
3. Start up the *CLC Server* software on the master server. Then start up the software on the job nodes. (See section 2.7.)
4. Log in to the web administrative interface for the *CLC Server* of the master node. (See section 9.)
5. Configure the master node, attach the job nodes and configure the job nodes via the administrative interface on the master node.

Almost all configuration for the CLC Server cluster is done via the web administrative interface for the *CLC Server* on the **master node**. This includes the installation of plugins. See section 6.2.4.

The only work done directly on the machines that will run as job nodes is

- Installation of the *CLC Server* software.
- Starting up the software up on each node.

- The changing of the built-in administrative login credentials under certain circumstances. See section 6.2.2.
- If using a CLC Bioinformatics Database, installing the relevant database driver on each job node.

6.2.2 User credentials on a master-job node setup

When initially installed, all instances of the *CLC Server* will have the default admin user credentials.

If you have a brand new installation, and you are happy to use the default administrative login credentials (see section 9) during initial setup, you do not need to change anything.

Once the *CLC Server* software on all machines is up and running and the job nodes have been attached to the master, changes to passwords for the built-in authentication system, which includes the default admin user, root, will be pushed from the master to the job nodes. You do not need to manually change the password on each job node.

If you wish to change the default administrative password before attaching the job nodes to the master, then please log into the web administrative interface of each machine running the *CLC Server* software and setup *identical* details on each one.

The master node needs access to the job nodes to be able to push configurations to them. Thus, if you change the admin password on the master server and later wish to attach a new job node, you will need to log into the web administrative interface of the job node and set the root password for the *CLC Server* software to match that of the master server. Until that is done, the master will not be able to communicate with the job node because the root admin passwords are different. Once the master can communicate with the job node, it can push other configurations to the job node.

6.2.3 Configuring your setup

If you have not already, please download and install your license to the master node. (See section 2.6.) Do **not** install license files on the job nodes. The licensing information, including how many job nodes you can run, are all included in the license on the master node.

To configure your master/execution node setup, go to the Job distribution tab in the web administrative interface on the master node:

Admin (⚙️) | **Job distribution** (👤)

Then enter the following information:

- **Server mode** - select `MASTER_NODE`.
- **Master node host** - Enter the master node host name. Click on the "Show suggestions" text next to this field to see information about the server that can be useful when configuring this option.
- **Master node port** - usually 7777
- **Master node display name** - the name shown in the top bar of the web interface for the *CLC Server*

- **CPU limit** - The maximum number of CPU the CLC Server should use. This is set to unlimited by default, meaning that up to all cores of the system can be used.

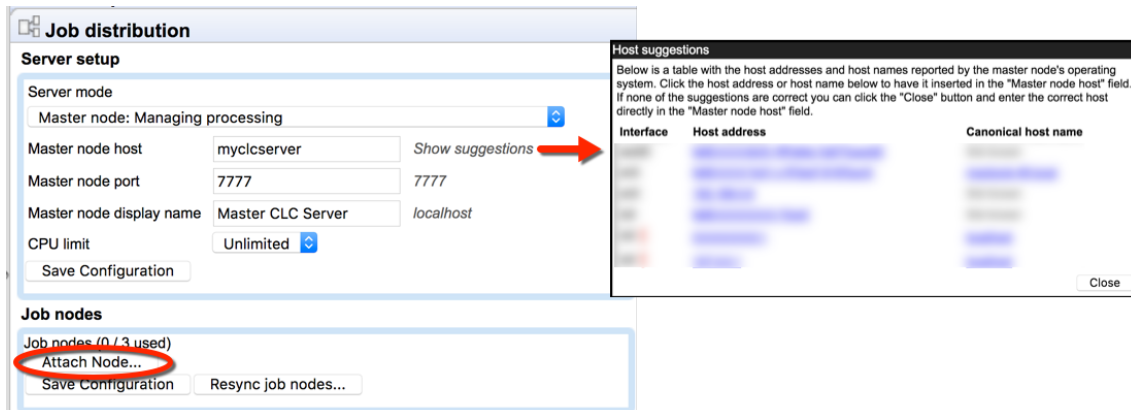


Figure 6.2: Setting up a master server.

Click on the button **Save Configuration** to register the changes just made.

If the **Attach Node** button in the Job nodes section is greyed out, please ensure that the server mode selected is `MASTER_NODE` and that you have clicked on the **Save Configuration** button to save your configuration changes.

Then, for each job node:

- Click on the **Attach Node** button to specify a job node to attach. See figure 6.2.
- Enter information about the node in the fields (see figure 6.3).

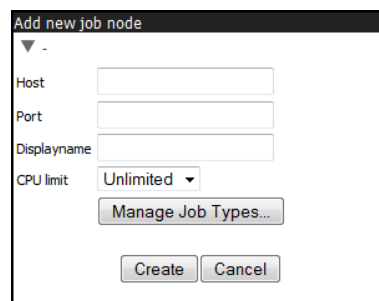


Figure 6.3: Add new job node.

- Optionally, click on the button labeled "Manage Job Types" to specify the types of jobs that can be run on each node. See figure 6.4. The default is "Any installed Server Command". If you choose instead "Only selected Server Commands", a search field and a list of all server command names and types will appear. The search field can be used to narrow down the server commands by name or type. Only the tools selected can then be run on that particular job node. Click on the button labeled **Modify** when you are done.
- Click on the button labeled **Create**.

Repeat this process for each job node you wish to attach and click **Save Configuration** when you are done.

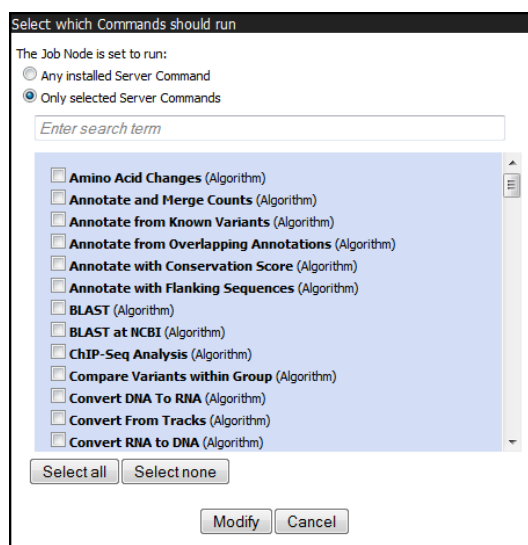


Figure 6.4: Select Server Commands to run on the job node.

You will get a warning dialog if there are types of jobs that are not enabled on any of the nodes.

Note that when a node has finished a job, it will take the first job in the queue that is of a type the node is configured to process. This then means that, depending on how you have configured your system, the job that is number one in the queue will not necessarily be processed first.

To test that access works for both job nodes and the master node, you can click "check setup" in the upper right corner as described in section 15.2.1.

One relatively common problem that can arise is *root squashing*. This often needs to be disabled, because it prevents the servers from writing and accessing the files as the same user - read more about this at http://nfs.sourceforge.net/#faq_b11.

Once set up, job nodes automatically inherit all configurations on the master node. If one of the job nodes gets out of sync with the master, click the **Resync job nodes** button, shown in figure 6.2. This should not be done while jobs are running on any nodes. The server can be put into Maintenance Mode to allow current jobs to complete before maintenance tasks are carried out. See section 7.3.

6.2.4 Installing Server plugins on job nodes

You **only** need to install or uninstall plugins on the master server. The *CLC Server* master and all job nodes need to be restarted to complete the installation or removal of plugins. This is easily accomplished by using the **Restart** option in the **Server Maintenance** section under the **Status and management** area on the master server, which restarts the master and all the job nodes. See section 7.

Server plugin installation and removal is described in section 10.

Once a plugin is installed, you should **check that the plugin is enabled** for **each job node** you wish to make available for users to run that particular task on. To do this:

- Go to the Job Distribution tab in the master nodes web administrative interface
- Click on the link to each job node

- Click in the box by the relevant task, marking it with a check mark if you wish to enable it.

6.3 Model II: Master server submitting to grid nodes

6.3.1 Overview: Model II

The CLC Grid Integration allows jobs to be offloaded from a master server onto grid nodes using the local grid submission/queuing software to handle job scheduling and submission.

A given CLC algorithm will only run on a single machine, i.e., a single job will run on one node. Each grid node employed for a given task must have enough memory and space to carry out that entire task.

The grid system uses two locations for its deployment:

- **Path to CLC Grid Worker** in the grid preset editor. This location is used for plugins, a grid version of the server, i.e., the code that is executed when a grid job is started, licences, libraries and more. Grid workers are redeployed in two situations: 1) When the server starts up and 2) If the configuration of one of the grid presets is changed, in which case the grid workers of all presets are redeployed. In addition, the grid workers are updated when a plugin is installed or removed.
- **Shared work directory.** This location is where each grid job gets its own sub directory in which it places its temporary data, e.g., the description of the job to be executed, the configurations files that the grid version of the server needs in order to setup persistence models and log files. This location is only deployed once, either when the grid job starts executing (in case of workflow jobs) or when the grid job is queued (in all other cases).

6.3.2 Requirements for CLC Grid Integration

- A functional grid submission system must already be in place. Please also see the section on supported job submission systems below.
- The DRMAA library for the grid submission system to be used. See Appendix section [15.6](#) for more information about DRMAA libraries for the supported grid submission systems.
- The *CLC Server* must be installed on a Linux based system configured as a *submit host* in the grid environment.
- The user running the *CLC Server* process is seen as the submitter of the grid job, and thus this user must exist on all the grid nodes.
- *CLC Server* file locations holding data that will be used must be mounted with the same path on the grid nodes as on the master *CLC Server* and accessible to the user that runs the *CLC Server* process.
- If a *CLC Bioinformatics Database* is in use, all the grid nodes must be able to access that database using the user that runs the *CLC Server* process.
- A *CLC License Server* with one or more available *CLC Grid Worker* licenses must be reachable from the *execution hosts* in the grid setup.

Supported grid scheduling systems

Grid integration in CLC Genomics Server is done using DRMAA, described further in the appendix 15.6. QIAGEN Bioinformatics has verified grid integration functionality using the following third party scheduling systems:

- SLURM 16.05.2
- UNIVA 8.4.1
- LSF 9.1.1 and 10.1
- PBS Pro 14.2.1

Integration using other grid scheduling systems is anticipated to work as long there is a working DRMAA library and a mechanism to limit the number of CLC jobs launched for execution such that when this number exceeds the number of CLC Grid Worker licenses, excess tasks are held in the queue until a license becomes available.

For SLURM, the number of CLC Grid Worker licenses can be configured as described on <https://slurm.schedmd.com/licenses.html>. For LSF and UNIVA, a "Consumable Resource" would be configured, as described in section 6.3.6. Relevant information about configuring consumable resources for PBS Pro can be found in the administrator's guide for that scheduling software.

TORQUE from Adaptive Computing is an example of a system that works for submitting CLC jobs, but that cannot be supported because it does not provide a means of limiting the number of CLC jobs. As far as we know, there is no way to limit the number of CLC jobs sent simultaneously to the cluster to match the number of CLC Grid Worker licenses. So, with TORQUE, if you had three Grid Worker licenses, up to three jobs could be run simultaneously. However, if three jobs are already running and you launch a fourth job, then this fourth job will fail because there would be no license available for it. This limitation can be overcome, allowing you to work with systems such as TORQUE, if you control the job submission in some other way so the license number is not exceeded. One possible setup for this is if you have a one-node-runs-one-job setup. You could then set up a queue where jobs are only sent to a certain number of nodes, where that number matches the number of CLC Grid Worker licenses you have.

6.3.3 Technical overview

Figure 6.5 shows an overview of the communication involved in running a job on the grid, using OGE as the example submission system.

The steps of this figure are in detail:

1. From the Workbench the user invokes an algorithm to be run on the grid. This information is sent to the master server running the *CLC Server*.
2. The master server writes a file with job parameters to a shared work directory of the grid execution nodes. The job parameters contain identifiers mapping to the job data placed in the CLC server data location. The job parameters file is automatically deleted when it is no longer used by the grid node.

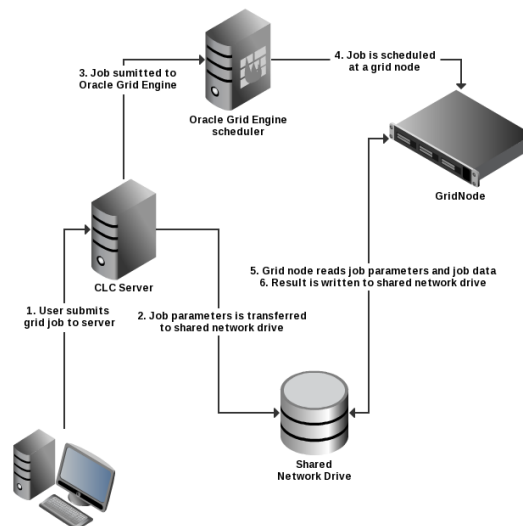


Figure 6.5: An overview of grid integration, using OGE as the example submission system.

3. Now the server invokes *qsub* through the specified DRMAA native library. Then *qsub* transfers the job request to the grid scheduler. Since the user that runs the CLC Server process has invoked *qsub*, the grid node will run the job as this CLC-Server user.
4. The job scheduler will choose a grid node based on the parameters given to *qsub* and the user that invoked *qsub*.
5. The chosen grid node will retrieve CLC Grid Worker executable and the job parameters from the shared file system and start performing the given task.
6. After completion of the job, the grid node will write the results to the server's data location. After this step the result can be accessed by the Workbench user through the master server.

6.3.4 Setting up the grid integration

CLC jobs are submitted to a local grid via a special, stand-alone executable called **clcgriidworker**. In the documentation, this executable is also referred to as the CLC Grid Worker.

The following steps are taken to setup grid integration for CLC jobs. These steps are described in more detail in the sections that follow. It is assumed that your CLC Server software is already installed on the machine that is to act as the master.

1. Set up the licensing of the grid workers, as described in section 6.3.5
2. Configure the CLC grid licenses as a consumable resource in the local grid system, as described in section 6.3.6
3. Configure and save grid presets, as described in section 6.3.7
4. Optionally, create and edit a `clcgriidworker.vmoptions` file in each deployed CLC Grid Worker area, as described in section 6.3.10. This is usually desirable and would be done if you wished to customize settings such as maximum memory to dedicate to the java process.

5. Test your setup by submitting some small tasks to the grid via a *CLC Server* client, such as a *CLC Genomics Workbench* or the *CLC Server Command Line Tools*. Ideally, these would be tasks already known to run smoothly directly on your *CLC Server*.

6.3.5 Licensing of grid workers

There are two main steps involved in setting up the licenses for your *CLC Grid Workers*.

Step 1: Installing network licenses and making them available for use

Generally, a pool of *CLC* grid licenses are purchased and are served by the *CLC License Server* software. For information on how to install the *CLC License Server* and download and install your *CLC Grid Worker* licenses please follow the instructions in the *CLC License Server* user manual, which can be found at

<http://resources.qiagenbioinformatics.com/manuals/clclicenseserver/current/>.

A pdf version is available at

http://www.resources.qiagenbioinformatics.com//manuals/clclicenseserver/User_Manual.pdf.

Step 2: Configuring the location of your *CLC License Server* for your *CLC Grid Workers*

One license is used for each *CLC Grid Worker* script launched. When the *CLC Grid Worker* starts running on a node, it will attempt to get a license from the license server. Once the job is complete, the license will be returned. Thus, your *CLC Grid Worker* needs to know where it can contact your *CLC License Server* to request a license.

To configure this, use a text editor and open the file: `gridres/settings/license.properties` under the installation of your *CLC Server*.

The file will look something like this:

```
#License Settings

serverip=host.example.com
serverport=6200
disableborrow=false

autodiscover=false
useserver=true
```

You can leave `autodiscover=true` to use UDP-based auto discovery of the license server. However, for grid usage it is recommended that you set `autodiscover=false` and use the `serverip` property to specify the host name or IP-address of your *CLC License Server*.

After you have configured your grid presets, see section 6.3.7, and have saved them, those presets are deployed to the location you specify in the **Path to *CLC Grid Worker*** field of the preset. Along with the `clcgridworker` script, the license settings file is also deployed.

If you need to change your license settings, we recommend that you edit the `license.properties` file under `gridres/settings/license.properties` of your *CLC Server* installation, and

then re-save each of your grid presets. This will re-deploy the CLC Grid Workers, including the changed license.properties file.

6.3.6 Configuring licenses as a consumable resource

Since there is a limitation on the number of licenses available, it is important that the local grid system is configured so that the number of CLC Grid Worker scripts launched is never higher than the maximum number of licenses installed. If the number of CLC Grid Worker scripts launched exceeds the number of licenses available, jobs unable to find a license will fail when they are executed.

Some grid systems support the concept of a *consumable resource*. Using this, you can set the number of CLC grid licenses available. This will restrict the number of CLC jobs launched to run on the grid at any one time to this number. Any job that is submitted while all licenses are already in use should sit in the queue until a license becomes available. **We highly recommend that CLC grid licenses are configured as a consumable resource on the local grid submission system.**

Information about how a consumable resource can be configured for LSF has been provided by IBM and can be found in Appendix section [15.7](#)

6.3.7 Configure grid presets

A **grid preset** contains information needed for jobs to be handled by the CLC Server and submitted to the grid scheduling system. Multiple grid presets can be configured. Users specify the relevant grid preset when submitting a job. See also section [6.3.12](#).

To configure a grid preset, log into the web interface of the CLC Server master and navigate to:

Admin (⚙️) | **Job distribution** (👤)

In the **Grid setup** section, under **Grid Presets**, click on the **Create New Preset** button.

The screenshot shows a web form titled "Edit preset LSF". It has several input fields and a "Submit test job..." button. The "Preset name" field contains "LSF". The "Native library path" field contains "/usr/lib/libdmaa.so". The "Shared work directory" field contains "/mnt/shared/tmp/gridworker". The "Path to CLC Grid Worker" field contains "/mnt/shared/gridworker/clcgridworker". The "Job category" field is empty. The "Grid Mode" section has two radio buttons: "Legacy" and "Resource Aware", with "Resource Aware" selected. Below this is a "Native specification f(x)..." field which is empty. To its right is a "Submit test job..." button. Below that is a "Shared native specification f(x)..." field containing the text "-n {COMMAND_THREAD_MIN},{COMMAND_THREAD_MAX}", with another "Submit test job..." button to its left. At the bottom of the form are "Cancel" and "Save Configuration" buttons.

Figure 6.6: The grid preset configuration form. The Shared native specification field is only visible when the Resource Aware grid mode is selected.

Preset name The preset name will be specified by users when submitting a job to the grid. See

section 6.3.12). Preset names can contain alphanumeric characters and hyphens. Hyphens cannot be used at the start of preset names.

Native library path The full path to the grid-specific DRMAA library.

Shared work directory The path to a directory that can be accessed by both the CLC Server and the Grid Workers. Temporary directories are created within this area during each job run to hold files used for communication between the CLC Server and Grid Worker.

Path to CLC Grid Worker The path to a directory on a shared file system that is readable from all execution hosts. If this directory does not exist, it will be created.

The CLC Grid Worker and associated settings files are extracted from the installation area of the CLC Server software and are deployed to this location when the grid preset is saved and whenever plugins are updated on the CLC Server.

In versions of CLC Server earlier than 5.0, this path needed to point at the clcgridworker script itself. To support backwards compatibility with existing setups, we ask that you do not use the name "clcgridworker" for a directory you wish a CLC Grid Worker to be deployed to.

Job category The name of the job category - a parameter passed to the underlying grid system.

Grid mode There are two grid modes for backwards compatibility reasons. The "Resource Aware" mode is generally recommended. Choosing this mode allows jobs that require few resources to run concurrently on a given node. For this, the field **Shared native specification** must also be configured. This is described further below. If "Legacy mode" is selected, all jobs submitted to the grid from the CLC Server will request use of the entire node. "Legacy Mode" is the default, but this may change in the future.

Native specification and Shared native specification Parameters to be passed to the grid scheduler are specified here. For example, a specific grid queue or limits on numbers of cores. Clicking on the f(x) next to the field name pops up a box containing the variables that will be evaluated at run time. These are described further below.

The **Native specification** field contains the information to be passed to the grid scheduler for exclusive jobs, those where the whole execution node will be used.

The **Shared native specification** contains the information to be passed to the grid scheduler for jobs classified as non-exclusive. Such jobs can share the execution node with other jobs. This specification is only visible and configurable if the "Resource Aware" grid mode is selected.

<i>Grid mode</i>	Exclusive Jobs	Non-exclusive jobs
<i>Legacy</i>	Native specification	NA - all jobs treated as exclusive
<i>Resource aware</i>	Native specification	Shared native specification

Table 6.1: Summary of grid modes and the specifications used for exclusive jobs, requiring a whole node, and non-exclusive jobs, which can share a node with other jobs.

Exclusive, streaming and non-exclusive tasks are described in section 6.6.1 and further configuration for running concurrent jobs is described in the section on Multi job processing on grid 6.3.9

Below are examples of OGE-specific arguments one could provide in the native specification field of a grid preset. Please refer to your grid scheduler documentation for information on the options available for your grid system.

Example 1: To redirect standard output and error output, this in a native specification field:

```
-o <path/to/standard_out> -e <path/to/error_out>
```

would result in the following *qsub* command being generated:

```
qsub my_script -o <path/to/standard_out> -e <path/to/error_out>
```

Example 2: To use a specific OGE queue for all jobs, this in a native specification field:

```
-hard -l qname=<name_of_queue>
```

would lead to the following *qsub* command:

```
qsub my_script -q queue_name
```

f(x) - adding variables to be evaluated at run-time

Grid Presets are essentially static in nature, with most options being defined directly in the preset itself. There are, however, 5 variables that can be specified that will be evaluated at runtime.

To aid with the required syntax, click on the f(x) link and choose the variable to insert. The variables can also be entered directly into the specification by typing the variable name between a pair of curly brackets.

The available variables are:

USER_NAME The name of the user who submitted the job. This variable might be added to log usage statistics or to send an email to an address that includes the contents of this variable. For example, something like the following could be put into a native specification field:

```
-M {USER_NAME}@yourmailserver.com
```

COMMAND_NAME The name of the CLC Server command to be executed on the grid by the clcgridworker executable. One example of the use of this variable is to specify `-q {COMMAND_NAME}` if there were certain commands to be submitted to queues of the same name as the command.

COMMAND_ID The ID of the CLC Server command to be executed on the grid.

COMMAND_THREAD_MIN A value passed by non-exclusive jobs indicating the minimum number of threads required to run the command being submitted. This variable is only valid for Shared native specifications.

COMMAND_THREAD_MAX A value passed by non-exclusive jobs indicating the maximum number of threads supported by the command being submitted. This variable is only valid for Shared native specifications.

Using functions in native specifications

Two functions can be used in native specifications `take_lower_of` and `take_higher_of`. These are invoked with the syntax: `{#function arg1, arg2, [... argn]}`: These functions are anticipated to be primarily of use in Shared native specifications when limiting the number of threads or cores that could be used by a non-exclusive job, and where the grid system requires a fixed number to be specified, rather than a range.

take_lower_of Evaluates to the lowest integer value of its argument.

take_higher_of Evaluates to the highest integer-value of its argument.

In both cases, the allowable arguments are integers or variable names. If an argument provided is a string that is not a variable name, or if the variable expands to a non-integer, the argument is ignored. For instance `{#take_lower_of 8,4,FOO}` evaluates to 4 and ignores the non-integer, non-variable "FOO" string. Similarly, `{#take_higher_of 8,4,FOO}` evaluates to 8 and ignores the non-integer, non-variable "FOO" string.

An example of use of the `take_lower_of` function in the context of running concurrent jobs on a given grid node is provided in section [6.3.9](#).

6.3.8 Controlling the number of cores utilized

This section covers basics related to setting limits on core or thread usage via a CLC Server grid preset. Parameter details are specific to the grid scheduler being used. We provide some information below for some schedulers, but please refer to the grid scheduler documentation for full details.

When using "Legacy mode" grid mode, or when running exclusive jobs with the "Resource Aware" grid mode, the default for all jobs is to assume they have access to all cores on the node they are run on. Details about grid modes can be found in section [6.3.7](#).

When configuring a core or thread limit for exclusive jobs, i.e. those that will require use of a whole node, the relevant parameter(s) and integer value(s) to be used by the grid scheduler are entered directly in the Native specification field. To specify different core limits for different types of tasks, one could set up multiple presets with different values supplied in the Native specification field of each.

Configuration of core requirements is central to supporting concurrent execution of non-exclusive jobs on a grid node. This is done by specifying core or thread requirements in the Shared native specification, making use of the variables `COMMAND_THREAD_MIN`, `COMMAND_THREAD_MAX` and optionally the functions `take_lower_of` and `take_higher_of`. Further information about this is provided in section [6.3.9](#).

Configuration of OGE

1) CPU Core usage when not using parallel environment

In the CLC Server, there is an environmental variable, which when set to 1, specifies that the

number of allocated slots should be interpreted as the maximum number of cores a job should be run on. To set this environmental variable, add the following to the Native specification of the grid preset:

```
-v CLC_USE_OGE_SLOTS_AS_CORES=1
```

In this case, the maximum number of cores the job should use will be set to the number of slots allocated by OGE for the job.

2) Limiting CPU core usage when using the parallel environment feature

The parallel environment feature can be used to limit the number of cores used by the *CLC Server* jobs when running on the grid. When the parallel environments feature is used, the number of allocated slots is interpreted as the maximum number of cores to be used by the job. The parallel environment must be setup by the grid administrator in such a way that the number of slots corresponds to the number of cores.

The syntax in the Native specification for using parallel environments is:

```
-pe <pe-name> <min-cores>-<max-cores>
```

where *pe-name* is the name of the parallel environment, and a range of cores is specified with integers, e.g. 1-4.

An example as might be entered into a Native specification when using parallel environments is:

```
-l cfl=1 -l qname=64bit -pe clc 1-3.
```

Here, the *clc* parallel environment is selected and 1 to 3 cores are requested.

Configuration of PBS Pro

With PBS Pro the number of cores to use is specified with a single number. This request can be granted (the process is scheduled) or denied (the process is not scheduled). The number of cores are requested as a resource: `-l nodes=1:ppn=X`, where *X* is the number of cores. Please ensure that *the number of nodes requested is 1*.

An example as might be entered into a Native specification is: `-q bit64 -l nodes=1:ppn=2`. This would request 2 cores and the job would be put in the *bit64* queue.

Configuration of LSF

With LSF the number of cores to use is specified with the `-n` option. This parameter can accept a single argument or two arguments. A single argument, `-n X`, is a request for exactly *X* cores. Two arguments, `-n X, Y`, (separated by a comma), is a request for between *X* and *Y* cores.

6.3.9 Multi-job processing on grid

Certain types of *CLC Server* jobs, known as "non-exclusive" jobs, can be scheduled to run concurrently on the same grid node when appropriate. Non-exclusive jobs are those that have reasonably low demands for system resources. A list of such jobs is provided in the Appendix, section 15.5.

There are two ways non-exclusive jobs can be configured to run concurrently on a grid node:

1. In the context of a workflow executed on a single grid node

When the "Single entity" job queueing option is chosen, as described in section 6.5, all tasks that are part of a given workflow run are executed on a single grid node. If a workflow design includes parallel non-exclusive tasks, these can run concurrently on the grid node. By default, up to 10 such non-exclusive jobs can be run concurrently. This value can be changed in the "Maximum number of concurrent jobs" field, available in the "Server setup" area of the Job Distribution tab. That field is visible when the server mode "Single server" has been selected, as shown in figure 6.7. The value configured is passed through to the CLC Server queue of the grid worker.

Figure 6.7: The Maximum number of concurrent jobs setting is visible when the Single server mode is selected.

A grid setup can be run with the Single server or Master server mode set. If you prefer to have the Master server option set, then to alter the "Maximum number of concurrent jobs" setting, change the mode to Single server, set the desired value, and then change the mode back to Single server. This this will, however, have the downside that the "Maximum number of concurrent jobs" value will not be obvious when reviewing the CLC Server configuration.

Limitation: The maximum value that can be entered in the "Maximum number of concurrent jobs" field is the number of cores on the master server. Setting that maximum value is the equivalent of checking the unlimited box. If this value is smaller than the desired value, as might happen when grid nodes have many more cpu than the master server, then we recommend leaving this field blank so that the default value of 10 is used.

Licensing note: When using the "Single entity" job queueing option, only a single CLC Grid Worker is launched for a given workflow run. Thus, irrespective of the number of concurrently running jobs in such a workflow run, only a single license is used. For further details on the "Single entity" job queueing option, see section 6.5.

2. In cases other than workflows executed on a single grid node

To support the concurrent execution of non-exclusive jobs submitted to the CLC Server outside workflows, or within workflows when the "Classic" job queueing option has been selected, information about the CPU or thread requirements of these jobs must be passed to the grid scheduler.

Non-exclusive algorithms expose their CPU or thread usage, and this information can be passed on to the grid scheduler via the `COMMAND_THREAD_MIN` and `COMMAND_THREAD_MAX` variables in the Shared native specification of a grid preset. The variable `COMMAND_THREAD_MAX`

would be used alone as an argument when a single value should be specified, or both the variables `COMMAND_THREAD_MIN` and `COMMAND_THREAD_MAX` can be provided when a range is required. An example of specifying a range is shown in the image of a grid present in section 6.3.7.

One can also use the functions `take_lower_of` and `take_higher_of` for settings relevant to configuring multiple job processing. For example, to specify 4 as the maximum number of cores to be used by a non-exclusive job, the following could be used as the argument to the relevant parameter in the Shared native specification of a grid preset: `{#take_lower_of COMMAND_THREAD_MAX, 4}`. As the non-exclusive job passes on its thread usage requirements via the `COMMAND_THREAD_MAX` variable, this evaluates to 4 if that requirement is higher than 4 or the value specified by the job if is lower than 4.

Further details about grid preset configuration, including Shared native specifications, functions and variables, can be found in section 6.3.7.

Licensing note: Each CLC Grid Worker launched, whether it is to run alone on a node or run alongside a job already running on a particular node, will attempt to get a license from the CLC License Server. Once the job is complete, the license will be returned.

6.3.10 Other grid worker options

Additional java options can be set for grid workers by creating a file called:

```
clcgridworker.vmoptions
```

in the same folder as the *deployed* `clcgridworker` script, that is, the `clcgridworker` script within the folder specified in the **Path to CLC Grid Worker** field of the grid preset.

For example, if a `clcgridworker.vmoptions` was created, containing the following two lines, it would, for the CLC Grid Worker specified in a given preset, set memory limits for the CLC Server java process and a temporary directory available from the grid nodes, overriding the defaults that would otherwise apply:

```
-Xmx1000m  
-Djava.io.tmpdir=/path/to/tmp
```

For each grid preset you created, you can create a `clcgridworker.vmoptions` file within the folder you specified in the **Path to CLC Grid Worker** field. So for example, if you had two grid presets, you could set two quite different memory limits for the CLC Server java process.

This might be a useful idea in the case where you wished to provide two queues, one for tasks with low overheads, such as import jobs and trimming jobs in the case of CLC Genomics Server, and one for tasks with higher overheads, such as de novo assemblies or read mappings in the case of CLC Genomics Server.

6.3.11 Testing a Grid Preset

There are two types of tests that can be run to check a Grid Preset. The first runs automatically whenever the *Save Configuration* button in the Grid Preset configuration window is pressed. This is a basic test that looks for the native library you have specified. The second type of test is optional, and is launched if the *Submit test job...* button is pressed. This submits a small test job

to your grid and the information returned is checked for things that might indicate problems with the configuration. While the job is running, a window is visible highlighting the jobs progression as shown in figure 6.8.

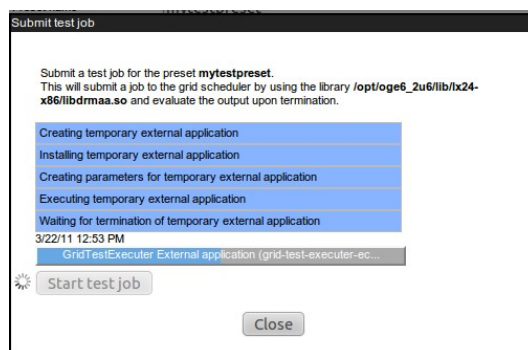


Figure 6.8: Testing a Grid Preset.

6.3.12 Client-side: starting CLC jobs on the grid

Starting grid jobs

Submitting jobs to grid nodes from a CLC Workbench

Once the CLC Server is configured and you are logged into it via a CLC Workbench, an extra option will appear in the first dialog box presented when setting up a task that could be executed on the CLC Server. At this stage, you can choose to execute the task on the machine the Workbench is running on, the CLC Server machine, or to submit the job to one of the available grid presets. To submit to the grid is as simple as choosing from among the grid presets in the drop down box, as shown in figure 6.9.

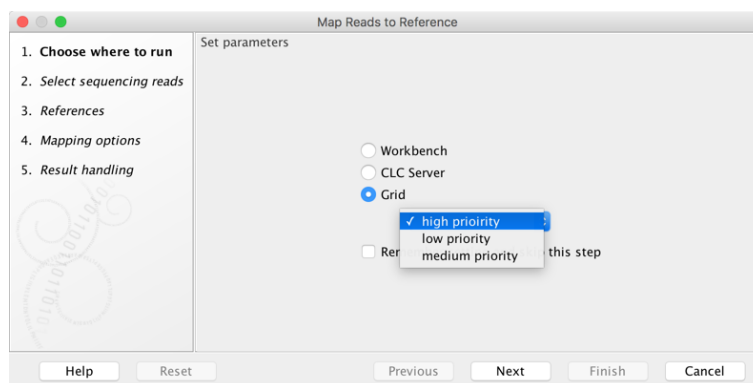


Figure 6.9: Starting the job on the grid.

Submitting jobs to grid nodes using the CLC Server Command Line Tools

Submitting jobs for execution on grid nodes using the **clcserver** command of the CLC Server Command Line Tools involves adding the -G parameter to the command, followed by the name of the grid preset to send the job to as the parameter value. A list of available grid presets is returned, along with other information about the clcserver command, if the command is run with just the server address, port and user credentials specified. Further details about the **clcserver** command can be found in the CLC Server Command Line Tools manual at http://resources.qiagenbioinformatics.com/manuals/clcservercommandlinetools/current/index.php?manual=Basic_usage.html.

6.3.13 Grid Integration Tips

If you are having problems with your CLC Grid Integration, please check the following points:

- Does your system meets the requirements of the CLC Grid Integration tool [6.3.2](#)? For example, please check that the machine the *CLC Server* is running on is configured as a submit host for your grid system, and please check that you are running Sun/Oracle Java 1.7 on all execution hosts.
- The user running the *CLC Server* process is the same user seen as the submitter of all jobs to the grid. Does this user exist on your grid nodes? Does it have permission to submit to the necessary queues, and to write to the shared directories identified in the Grid Preset(s) and any `clcgridworker.vmoptions` files?
- Are your *CLC Server* file locations mounted with the same path on the grid nodes as on the master *CLC Server* and accessible to the user that runs the *CLC Server* process?
- If you store data in a database, are all the grid nodes able to access that database, using the user account of the user running the *CLC Server* process?
- If you store data in a database, did you enter a machine name in the Host box of the Database Location field when setting up the Database Location using the *CLC Server* web administration form? In particular, a generic term such as `localhost` will not work, as the grid nodes will not be able to find the host with the database on it using that information.
- If you installed the *CLC Server* as root, and then later decided to run it as a non-privileged user, please ensure that you stop the server, recursively change ownership on the *CLC Server* installation directory and any data locations assigned to the *CLC Server*. Please restart the server as the new user. You may need to re-index your CLC data locations (section [3.2.5](#)) after you restart the server.
- Is your java binary on the PATH? If not, then either add it to PATH, or edit the `clcgridworker` script in the *CLC Server* installation area, with the relative path from this location: `gridres/dist/clcgridworker`, and set the JAVA variable to the full path of your java binary. Then re-save each of your grid presets, so that this altered `clcgridworker` script is deployed to the location specified in the **Path to CLC Grid Worker** field of your preset.
- Is the `SGE_ROOT` variable set early enough in your system that it is included in the environment of services? Alternatively, did you edit the *CLC Server* startup script to set this variable? If so, the script is overwritten on upgrade - you will need to re-add this variable setting, either to the startup script, or system wide in such a way that it is available in the environment of your services.
- Is your DRMAA library 64-bit? Both Java and DRMAA must be for 64-bit systems for it to work.

6.3.14 Understanding memory settings

Most work done by the *CLC Server* is done via its java process.

However, some tools involving de novo assembly or mapping phases (e.g. read mappings, RNA-seq analyses, smallRNA analyses, etc.) on *CLC Genomics Server* use external native binaries for the computational phases.

Java process

For the grid worker **java process**, if there is a memory limit set in your `clcgridworker.vmoptions` file, this is the value that will be used. See section 6.3.10.

If there is no memory setting in your grid worker's `clcgridworker.vmoptions` file, then the following sources are referred to, in the order stated. As soon as a valid setting is found, that is the one that will be used:

1. Any virtual memory settings given in the grid preset, or if that is not set, then
2. Any physical memory settings given in the grid preset, or if that is not set, then
3. Half of the total memory present, with 50GB being the maximum set in this circumstance.

External binaries

For the computationally intensive tools that include a phase using an external binary, the binary phase is not restricted by the amount of memory set for the java process. For this reason, we highly recommend caution if you plan to submit more jobs of these types to nodes that are being used simultaneously for other work.

6.4 Model III: Single Server setup

In this model, the master and execution node functionality is carried out by a single CLC Server instance. Here, the *CLC Server* software is installed on a single machine. Jobs submitted to the server are executed on this same machine.

To designate the system as a single server, after installation and starting the server, select the option `SINGLE_SERVER` from the drop down list of Server modes.

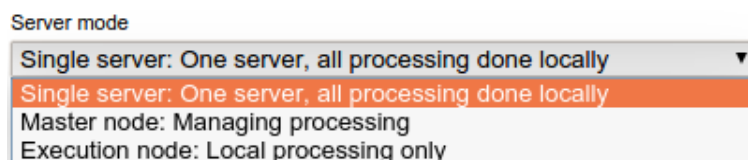


Figure 6.10: The configuration options for the types of machines running the **CLC Server**. The choices of relevance under normal circumstances are `SINGLE_SERVER` and `MASTER_NODE`. An administrator will not usually need to manually choose the Execution Node option. This option is there primarily to allow for troubleshooting.

You can then configure aspects of the server:

- **Master node host** - usually set to localhost for Single server mode. Click on the "Show suggestions" text next to this field to see information about the server that can be useful if localhost is not a suitable option for your setup.
- **Master node port** - usually 7777.
- **Master node display name** - the name shown in the top bar of the web interface for the server.

- **CPU limit** - The maximum number of CPU the CLC Server should use. This is set to unlimited by default, meaning that up to all cores of the system can be used.
- **Maximum number of concurrent jobs** - Limit the maximum number of jobs that are allowed to run concurrently on the single server. For further information about this setting see section 6.6.3.

Figure 6.11: Add new job node.

6.5 Job queuing options

The way Workflows should be handled on server setups with nodes can be configured using the Job queuing options found under the Job distribution tab (figure 6.12).

Figure 6.12: Job queuing options are available that determine how Workflow jobs are handled.

The options are:

- **Classic** Each task of a Workflow is scheduled separately for execution. For example, a Workflow with 10 tasks would result in 10 jobs being submitted. Each of those jobs can be sent to any available node with adequate resources when that step is ready to be run. Classic is the default setting.
- **Single entity** A single job submission is made for a Workflow, regardless of how many tasks that Workflow consists of. All tasks of that Workflow are run on the same node.

6.5.1 Choosing between Classic and Single entity options

Choosing the best queuing option involves consideration of the types of analyses being run and the system the jobs are being run on. The following are some considerations to help guide this

decision.

Classic

The Classic option would be beneficial in the following situations:

- **Workflows containing parallel branches are commonly launched.** Elements of such Workflows can be scheduled on multiple nodes. This potential for parallel execution of Workflow steps can yield shorter average running times, depending on other aspects of the setup (e.g. node or grid license availability). Time savings are most noticeable on systems with spare capacity when running Workflows with computationally intensive elements on parallel branches.
- **Single tools are predominantly submitted,** as opposed to Workflows. There is no need to change the default setting in this case, as the Single entity option only affects the way Workflows are handled.
- **Job node setups only: Nodes have been dedicated to certain types of analyses** If a job node has been configured to run only certain tasks (Server commands) as described in section 6.2.3, then, with the Classic option, this node can be used for those configured tasks, whether or not they are elements of a Workflow. This is not the case with the Single entity option where Workflows cannot be run on the job nodes configured like this.

Single entity

The Single entity option would be beneficial when the execution of Workflows is common, and:

- **Workflows being submitted often consist of large number of steps and each step has small computational requirements.** This is common when working with data from organisms with small genomes, such as bacterial or viral samples, or when working with enriched data from organisms with large genomes. Where scheduling overhead outweighs the net analysis time of a given Workflow step, the Single entity setting can yield many-fold improvements in overall sample throughput.
- **Nodes frequently run at capacity.** By running an entire Workflow on a single node, the Single entity option can leverage caching mechanisms, aiding performance. When all nodes are busy much of the time, the opportunity for gains through concurrent processing of parallel Workflow branches on multiple nodes is much lower than on a system with spare capacity. On such a system, the Single entity option may thus yield overall performance gains, even when running Workflows containing elements with computationally intensive steps.
- **The node hardware is homogeneous.** All the nodes should be of a size that could handle all tasks in the Workflows being submitted.
- **Resource allocation is a focus.** On setups where many users are sharing the resources and are running Workflows, the Single entity option may help with resource access for different users. For example, consider a grid node setup with 20 nodes, where one user submits 15 Workflows with 10 tasks in each Workflow. With the Classic option, this would lead to 150 jobs, which can be sent across all 20 nodes. When the next user submits a job, it would be queued behind all of those 150 jobs from the first user. With the Single entity option,

the 15 Workflows would have been submitted to 15 nodes, leaving 5 nodes available on which the other user's job could be run.

- **Large numbers of Workflows are submitted during limited periods of time, each Workflow consisting of several or many tasks.** Such a situation leads to thousands of jobs in the queue using the Classic option. Using the Single entity option is a way to keep the scheduling load on the master node with reasonable limits.
- **Grid node setups only: The number of grid worker licenses is limited relative to the number of job submissions.** Using the Single entity option, a Workflow is submitted as a single job and thus consumes a single grid worker license. If a Workflow has 10 steps and is submitted using the classic option, 10 jobs are created. Each of these jobs will consume a license, making license availability a limiting factor, along with node availability, for when jobs can be run.
- **Grid node setups only: Resource tracking is a focus.** Workflows involving many steps would run as a single job with a single license consumed by that job. This potentially decreases the complexity of resource use tracking of users running Workflows.

6.6 Job running options

There are three general categories of tools on the CLC Server: non-exclusive, streaming and exclusive. These are described below. Non-exclusive or streaming jobs can run concurrently alongside others of the non-exclusive type on a given machine. Those defined as exclusive cannot be run on the same server or node at the same time as other jobs of any type.

- **Non-exclusive algorithms** Tools with low demands on system resources. They can be alongside others in this category, as well with a job of the streaming category, described below. An example of a non-exclusive algorithm is Convert from tracks.
- **Streaming algorithms** Tools with high I/O demands, that is, much reading from and writing to disk is needed. These cannot be run with others in the streaming category but can be run alongside jobs running non-exclusive algorithms. An example of streaming algorithms are the NGS data import tools.
- **Exclusive algorithms** Tools optimized to utilize the machine they are running on. They have very high I/O bandwidth, memory, or CPU requirements and therefore should not be run at the same time as other jobs on the same machine. An example of this sort is Map Reads to Reference.

See Appendix section [15.5](#) for a list of *CLC Genomics Server* algorithms that can be run alongside others on a given machine.

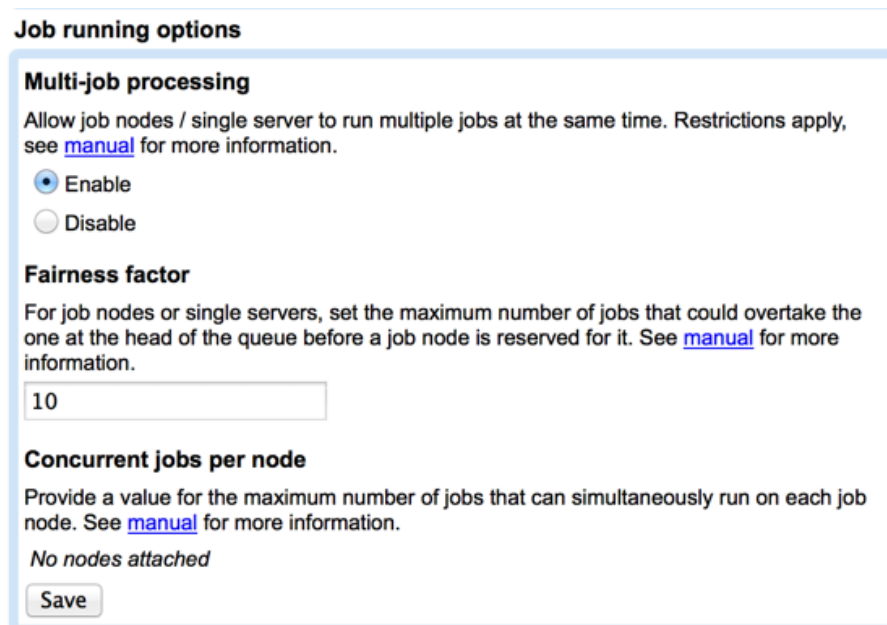
The rest of this section discusses the configuration options relevant for concurrent jobs on a **single server** or on a **job node**. Information about concurrent job processing on grid nodes is in section [6.3.9](#).

6.6.1 Multi-job processing

Allowing more than one analysis to run on a CLC Server in single server mode or on a job node is enabled by default. This feature can be disabled in the **Multi-job processing** section of the

interface by setting the "Multi-job Processing" option to "Disable" (figure 6.13). Click on the button labeled "Save" to save this change.

When this feature is disabled jobs will be executed sequentially only.



Job running options

Multi-job processing
Allow job nodes / single server to run multiple jobs at the same time. Restrictions apply, see [manual](#) for more information.

☒ Enable
☐ Disable

Fairness factor
For job nodes or single servers, set the maximum number of jobs that could overtake the one at the head of the queue before a job node is reserved for it. See [manual](#) for more information.

10

Concurrent jobs per node
Provide a value for the maximum number of jobs that can simultaneously run on each job node. See [manual](#) for more information.

No nodes attached

Save

Figure 6.13: The default status is to enable Multi-job processing. Select "Disable" to require that only a single job is executed at a given time on a job node or single server.

6.6.2 Fairness factor

The fairness factor defines the number of times that a job in the queue can be overtaken by other jobs before resources are reserved for it to run. So, for example, in a situation where there are many non-exclusive jobs and some exclusive jobs being submitted, it is desirable to be able to clear the queue at some point to allow the exclusive job to have a system to itself so it can run. The fairness factor setting is used to determine how many jobs can move ahead of an exclusive job in the queue before the exclusive job will get priority and a system will be reserved for it. The same fairness factor applies to streaming jobs being overtaken in the queue by non-exclusive jobs.

The default value for this setting is 10. With this value set, a job could be overtaken by 10 others before resources are reserved for it that will allow it to run. A fairness factor of 0 means that a node will be reserved for the job at the head of the queue.

This value can be configured under the **Fairness factor** section of:

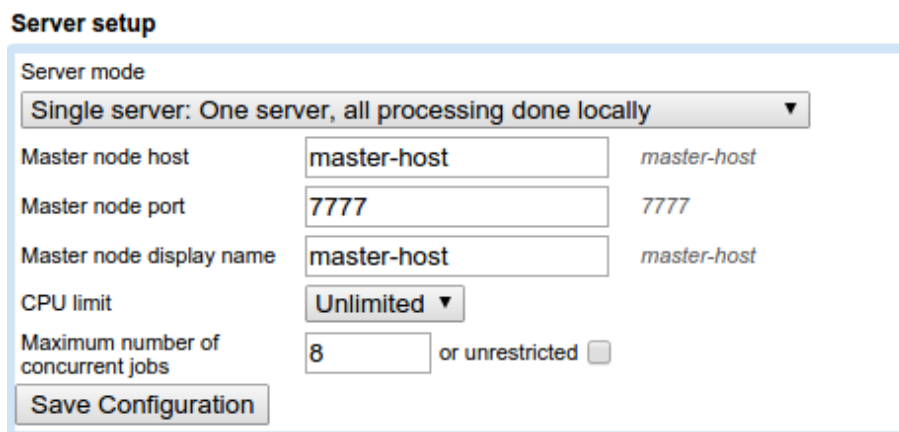
Admin (⚙️) | Job distribution (📊) | Job queuing options

6.6.3 Concurrent jobs per node

The maximum number of jobs that can be run on a single server or on each job node can be configured. The maximum allowable value is equal the number of cores on the relevant system. The default value is 10 or the number of cores on the system, whichever is lowest. If a CPU limit has been set on the single server or node, then the default is 10 or that CPU limit value,

whichever is lowest.

Single server setup To configure the maximum number of jobs that can be run concurrently on a single server, go to the section **Server setup** under the Job Distribution tab of the web administration page. Enter the desired value in the box labeled "Maximum number of concurrent jobs" (figure 6.14).



Server setup

Server mode
Single server: One server, all processing done locally ▼

Master node host master-host master-host

Master node port 7777 7777

Master node display name master-host master-host

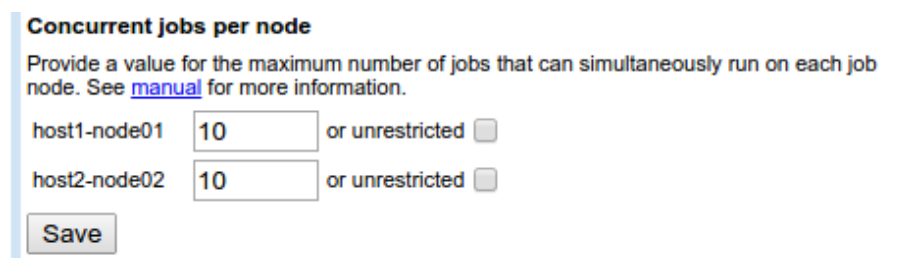
CPU limit Unlimited ▼

Maximum number of concurrent jobs 8 or unrestricted ☐

Save Configuration

Figure 6.14: Set the maximum number of concurrent jobs or check the "unrestricted" box. If unrestricted is chosen, the maximum number of jobs that can be run concurrently is equal to the number of cores on the system.

Job node setup To configure the maximum number of jobs that can be run concurrently on a given job node, go to the section **Job queuing options** under the Job Distribution tab of the web administration page. Enter a value for each job node listed under the "Concurrent jobs per node" section (figure 6.15).



Concurrent jobs per node

Provide a value for the maximum number of jobs that can simultaneously run on each job node. See [manual](#) for more information.

host1-node01 10 or unrestricted ☐

host2-node02 10 or unrestricted ☐

Save

Figure 6.15: Set the maximum number of concurrent jobs or check the "unrestricted" box for any job node. If unrestricted is chosen, the maximum number of jobs that can be run concurrently is equal to the number of cores on that job node.

Chapter 7

Status and management

Server operation can be managed from the Admin tab, under Status and Management (figure 7.1).

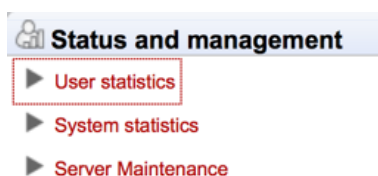


Figure 7.1: The Status and management tab.

7.1 User statistics

The User statistics section contains information about the number of users logged in, the number of active sessions (number of logins), and information about each active session. An example is shown in figure 7.2.

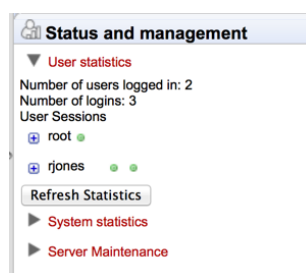


Figure 7.2: Information about the number of users and active sessions (logins) is provided in the User statistics area. Here, two users are logged in. rjones has two active sessions and root has one.

A green dot by a user's name indicates that they are logged into the server. Two green dots indicate that this user is logged in twice. For example, perhaps they have Workbenches running on two different systems and are logged in via both Workbenches. A grey dot means they have previously logged in but are not at this time.

Click on the small button with a plus to the left of a username to expand the information about that user's sessions (figure 7.3). You can also log users off the server by clicking on the **Invalidate Session...** button. This opens a confirmation dialog where a message to the user can

be written. This message is displayed via the user's active session. For example, if they are logged into a Workbench, a window will pop up saying they have been logged out of the server and also containing the message written in this field. This action forcibly logs the user out of the CLC Server. This action does **not** stop jobs already submitted or running on the server. Optionally, you can send a message to the user whose session is being terminated (see figure 7.4). If the user is logged into the CLC Server from a CLC Workbench, then the message entered will appear in a warning box that pops up via the CLC Workbench.



Figure 7.3: Details about *jbloggs*' session can be seen by clicking on the small button to the left of that username.

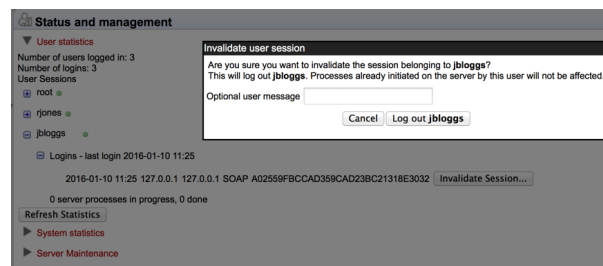


Figure 7.4: Clicking on the *Invalidate Session* button will forcibly log a user out of the CLC Server. The admin can optionally provide a message to the user when doing this.

7.2 System statistics

Crashed threads, suggesting system level problems, are reported in this area. In some instances, a system restart may be needed to resolve the issue.

The message "No system level problems detected" is shown in this area if no problems have been detected. An example of the information provided when a problem is detected is shown in figure 7.5. In the case shown, the job submission threads were dead, with the problem reported here and in more detail in the CLC Server log files.

7.3 Server maintenance

Settings under the Server maintenance tab allow a server administrator to change the operating mode of the server and send out messages to users of the CLC Server (see figure 7.6).

- **Normal Operation** The CLC Server is running.
- **Maintenance Mode** Current jobs are allowed to run and complete, but submission of new jobs is restricted. While the server is in maintenance mode, users already logged

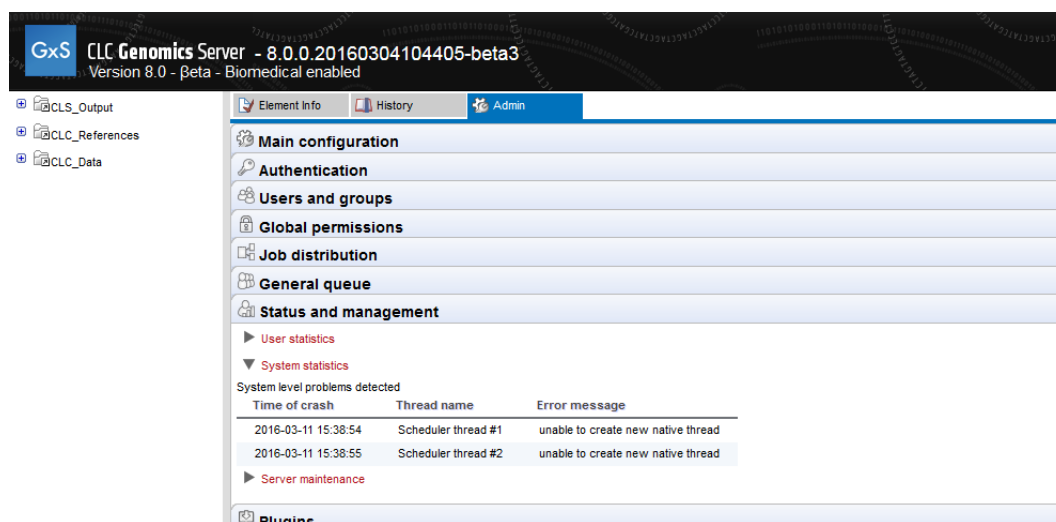


Figure 7.5: System level problems detected and reported in the system statistics area.

in can check the progress of their jobs or view their data, but they cannot submit new jobs. Users not already logged in cannot log in. An administrator can write a warning message, for example, to inform users about the expected period of time the server will be in maintenance mode.

- **Log Out Users** All users currently logged in will be logged out. All running jobs will be allowed to complete. No users can log in while in this mode. An administrator can also write a warning message for the users.
- **Shut down** The CLC Server and any attached job nodes will shut down.
- **Restart** The CLC Server and any attached job nodes will be shut down and restarted.

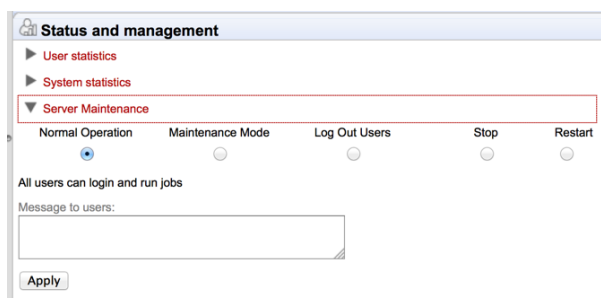


Figure 7.6: The server administrator can control the operating mode of the CLC Server from under the Server maintenance tab.

Chapter 8

Queue

Clicking the **Queue** panel will show a list of all the processes that are currently in the queue including jobs in progress.

An example is shown in figure 8.1.

Users and groups

Global permissions

Job distribution

Queue

Master process

Workflow: QW (160270.master-release - queued and in user hold, 160273.master-release - queued and in...

root

Add Conservation Scores (160276.master-release - queued and in user hold)

root

Create New Genome Browser View (160278.master-release - queued and in user hold)

root

Add Information from Variant Databases (160275.master-release - queued and in user hold)

root

Add Information from Variant Databases (160272.master-release - queued and in user hold)

root

Remove False Positives (160267.master-release - running)

root

Add Information from Overlapping Genes (160269.master-release - queued and in user hold)

root

Add Information from Variant Databases (160274.master-release - queued and in user hold)

root

Add Information from Variant Databases (160271.master-release - queued and in user hold)

root

Remove Information from Variants (160277.master-release - queued and in user hold)

root

Merge reports for the QW workflow (160279.master-release - queued and in user hold)

root

Remove Variants Outside Targeted Regions (160268.master-release - queued and in user hold)

root

Figure 8.1: *The process queue.*

For each process, you are able to **Cancel** (✖) the processes. At the top, you can see the progress of the process that is currently running.

Chapter 9

Audit log

The audit log records the actions performed on the *CLC Server*. Included are actions like logging in, logging out, import, and the launching and running of analysis tasks. Data management operations such as copying, deleting and adding files are not Server actions and are thus not recorded.

Audit log information is available via the web administrative interface under the **Audit log** tab. The information presented here is stored in a database. Once a month, and when the *CLC Server* is started up, entries in the audit log older than 3 months are deleted.

The limit the audit log database can grow to is 64 GB. If a new entry will push the size past this limit, the system will remove some of the oldest entries so that it is possible for newer entries to be added.

Audit information is also written to text-based log files. Upon the first activity on a given date, a new log file called `audit.log` is created. This file is then used for logging that activity and subsequent Server activities on that day. When this new `audit.log` file is created, the file that previously had that name is renamed to `audit.<actual events date>.log`. These log files are retained for 31 days. When the creation of a new `audit.log` file is triggered, audit log files older than 31 days are checked for and deleted.

The audit log files can be found under the Server installation area under `webapps/CLCServer/WEB-INF`.

The audit log text files are tab delimited and have the following fields:

- Date and time
- Log level
- Operation: Login, Logout, Command queued, Command done, Command executing, Change server configuration, Server lifecycle; more may be added and existing may be changed or removed.
- Users
- IP Address
- Process name (when operation is one of the Command values) or description of server lifecycle (when operation is Server lifecycle)

- Process identifier - can be used to differentiate several processes of the same type.
- Status - can be used to identify whether the entry was successful or not, e.g. if a job execution failed it will be marked here. Any number other than 0 means failed.

Chapter 10

Server plugins

Plugins can be installed using the functionality found under the **Plugins** (🔌) area of the **Admin** (⚙️) tab (see figure 10.1).

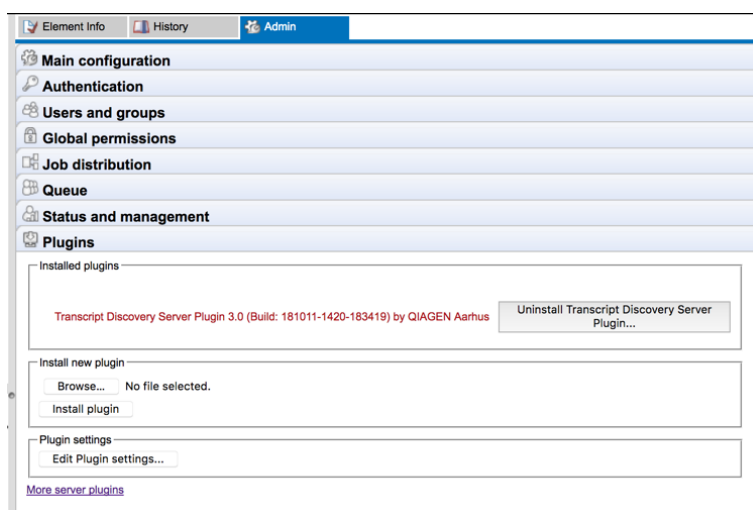


Figure 10.1: Installing and uninstalling server plugins is done in the *Plugins* area under the *Admin* tab. Installing is done in the "Install new plugin" section. Uninstalling is done by pressing the button beside the installed plugin name.

To install a server plugin:

- Download the plugin .cpa file from Server Plugins section of: <https://www.qiagenbioinformatics.com/plugins>. The "More server plugins" link at the bottom of the Plugins area leads to this web page.
- Click the **Browse** button in the "Install new plugin" area, select the plugin .cpa file and click on the "Open" button.
- Click on the **Install plugin** button in the "Install new plugin" area.

To uninstall a plugin: Click on the button beside the plugin in the "Installed plugins" area and then confirm that you wish to uninstall the plugin when prompted.

To complete plugin installation or removal, the CLC Server must be restarted. When the server is restarted, all jobs still in the queue at the time the server is shut down will be dropped and would need to be resubmitted.

To minimize the impact on users, the server can be put into Maintenance Mode. Maintenance Mode is described in section 7, but in brief: running in this mode allows current jobs to run, but no new jobs to be submitted, and users cannot log in. The CLC Server can then be restarted when desired. Each time you install or remove a plugin, you will be offered the opportunity to enter Maintenance Mode. You will also be offered the option to restart the CLC Server. If you choose not to restart when prompted, you can restart later using the option under the **Status and Management** tab.

For further information about plugins on job node setups, please refer to section 6.2.4.

Grid workers will be re-deployed when a plugin is installed on the master server. Thus, no further action is needed to enable the newly installed plugin to be used on grid nodes. See section 6.3 for further details about grid worker re-deployment.

Listing the tools delivered by a plugin

The list of tools delivered with a server plugin can be seen by clicking on the name of the plugin within the **Plugins** (🔧) area of the **Admin** (🔑) tab, as illustrated in figure 10.4.

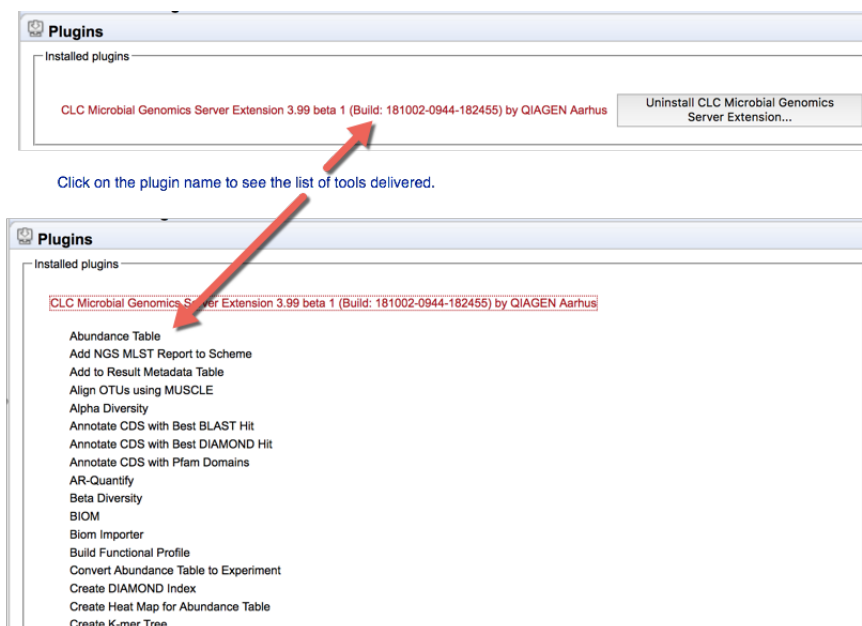


Figure 10.2: The tools delivered with a server plugin are shown in a list that opens when you click on the name of the plugin in the Plugins area. Clicking on the name again collapses the list.

Workflows delivered with a server plugin are not shown in this listing.

Plugin compatibility

Plugins must be compatible with the version of the CLC Server being run. A message is written under an installed plugin's name if it is not compatible with the version of the CLC Server software running. See figure 10.3 for an example of such warning messages.

When upgrading to a new major version of the CLC Server, all plugins will need to be updated. This means removing the old version and installing a new version. Clicking on the Download

link for a plugin on <https://www.qiagenbioinformatics.com/plugins> opens a window where you can select the version of a plugin according to the version of the CLC Server it is compatible with.

As incompatibilities can also arise when updating to a new bugfix or minor release of the CLC Server, it can be a good idea to open the **Plugins** area after any server software upgrade to check for any messages about the installed plugins.

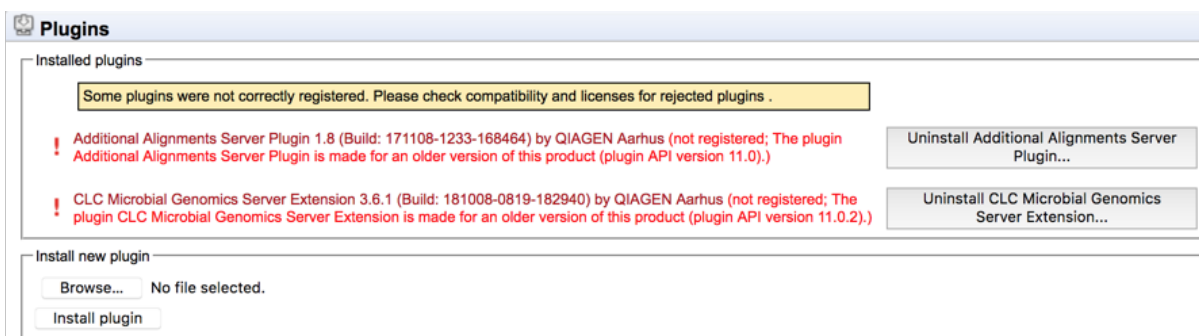


Figure 10.3: A warning in bright red text appears in the *Plugins* area when plugins are not compatible with the CLC Server software version being run. Just after an upgrade to a new major version, all plugins needed will need to be updated, as shown here.

Plugin licensing

Commercial plugins, known as modules, require a license to be installed in the CLC Server before analysis tools delivered by the module can be used. The license only needs to be present on the master server on grid and job node setups¹. If a license file is present, but it is valid only for an older version of the plugin, or it has expired, a warning will be shown in the **Plugins** area of the web administrative interface, as illustrated in figure 10.4.

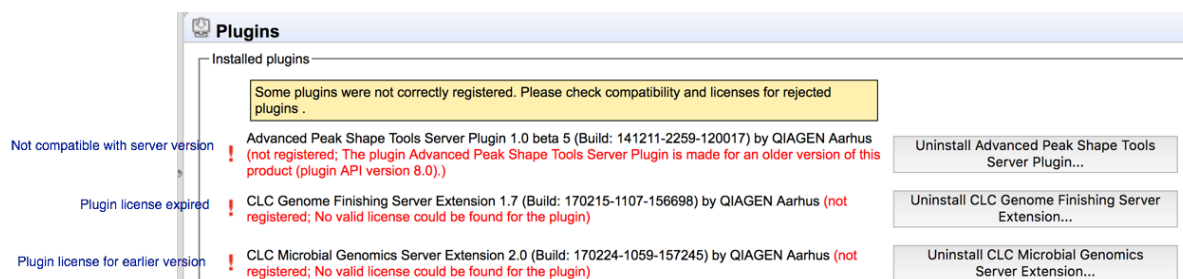


Figure 10.4: Warning text appears in the *Plugins* area when licenses are missing or expired, as well as if plugins are not compatible with the version of the server software being run.

¹Prior to CLC Genomics Server 11.0, module licenses were also required on nodes.

Chapter 11

BLAST

The *CLC Server* supports running BLAST jobs submitted from the workbenches that have BLAST tools and from *CLC Server Command Line Tools*. Users will be able to select data from Server **data locations** (see section 3.2.1) to search against other sequences held in Server data locations, or against BLAST databases stored in an area configured as an **import/export directory** (see section 3.3).

11.1 Adding directories for BLAST databases on the Server

In the web interface of the server, you can configure your Server for BLAST databases:

Admin  | **BLAST Databases** 

Here, you can add a folder where you want the Server to look for BLAST databases. However, before doing this, please ensure that the folder you will be adding has been configured as an **import/export directory** (see section 3.3). This is necessary because BLAST databases are not truly CLC data, and thus are stored outside data locations specified for CLC data. They need, however, to be stored somewhere accessible to *CLC Server* process though, hence the need to put them in a directory configured as an import/export directory.

After the folder holding BLAST databases is configured as an Import/Export directory, it can be configured as a location that the *CLC Server* will look in for BLAST databases.

Do this by clicking on the **Edit BLAST Database Locations** button at the bottom of the area under the BLAST Databases area in the administrative interface.

This will bring up a dialog as shown in figure 11.1 where you can select which of the import/export directories you wish to use for storing BLAST databases.

Once added as a BLAST Database Location, the *CLC Server* will search this directory for any BLAST databases and list them under the BLAST tab in the web interface (see a section of this as an example in figure 11.2).

This overview is similar to the one you find in the Workbench BLAST manager for local databases including the following information:

- **Name.** The name of the BLAST database.

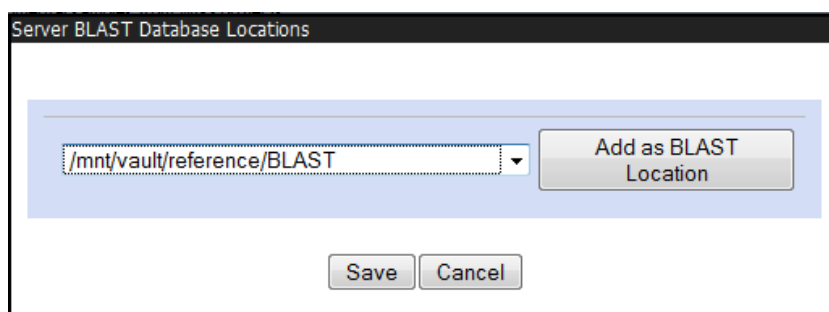


Figure 11.1: Adding import/export directories as BLAST database locations.

BLAST Databases				
BLAST databases overview				
Name	Description	Date	Sequences	Type
NC_000001	Human makeDB test	2011-11-14	19	DNA
all_contig	Homo sapiens build 37.3 genome database (reference assembly GRCh37.p5 [GCF_000001405.17] and alternate assemblies HuRef [GCF_000002125.1] and CRA_TCAGchr7v2 [GCF_000002135.2])	2011-10-07	4900	DNA
allcontig_and_rna	mouse build 37 RNA, reference and alternate assemblies	2011-05-25	35640	DNA
alt_CRA_TCAGchr7v2_contig	alt_CRA_TCAGchr7v2_contig	2011-10-07	6	DNA
alt_HuRef_contig	alt_HuRef_contig	2011-10-07	4530	DNA
alt_contig	Mus musculus build 37 genome database (alternate assembly Mm_Celera only)	2010-11-09	13033	DNA

Figure 11.2: Selecting database to BLAST against.


- **Description.** Detailed description of the contents of the database.
- **Date.** The date the database was created.
- **Sequences.** The number of sequences in the database.
- **Type.** The type can be either nucleotide (DNA) or protein.
- **Total size (1000 residues).** The number of residues in the database, either bases or amino acid.
- **Location.** The location of the database.

To the right of the Location information is a link labeled Delete that can be used to delete a BLAST database.

11.2 Adding and removing BLAST databases

Databases can be added in two ways:

- Place pre-formatted databases in the directory selected as BLAST database location on the server file system. The CLC Server will automatically detect the database files and list the database as target when running BLAST. You can download pre-formatted database from e.g. <ftp://ftp.ncbi.nih.gov/blast/db/>.

- Run the **Create BLAST Database** () tool via your Workbench, and choose to run the function on the Server when offered the option in the Workbench Wizard. You will get a list of the BLAST database locations that are configured on your Server. The final window of the wizard offers you a location to save the output to. The output referred to is the log file for the BLAST database creation. The BLAST databases themselves are stored in the designated BLAST database folder you chose earlier in the setup process.

A note on permissions: To create BLAST databases on the Server, using the Workbench interface, the user **running the CLC Server process** must have file system level write permission on the import/export directory that you have configured to hold BLAST database.

By default, if you do not change any permissions within *CLC Server*, all users logging into the *CLC Server* (e.g., via their Workbench, or via the Command Line Tools), will be able to create BLAST databases in the areas you have configured to hold BLAST databases.

If you wish to restrict the ability to create BLAST databases to these areas completely, but still wish your users to be able to access the BLAST databases to search against, then set the file system level permissions on the import/export directory so they are read-only.

When listing the databases as shown in figure 11.2, it is possible to delete the databases by clicking the **Delete** link at the far right-hand side of the database information.

Chapter 12

External applications

Command line applications on a server machine can be made available for use by CLC Workbench or CLC Server Command Line Tools users. This involves configuring an external application in the *CLC Server* web administrative interface.

Tools configured as external applications will be available via the graphical menu system of Workbenches logged into the *CLC Server*, and also as commands that can be run using the CLC Server Command Line Tools. Input data and parameter settings are specified in the same way as for standard tools provided via the client software. This is described further in section 12.10.

Key facts about external applications:

- If you plan to configure external applications, we highly recommend that you run the **CLC Server** software as an un-privileged user. Like other *CLC Server* tasks, tools configured as an External Application are run as the **same logical user** that runs the *CLC Server* process itself. In other words, if your system's root user is running the *CLC Server* process, then tasks run via the External Applications functionality will also be executed by the root system user. This is usually undesirable.
- The tools configured as external applications must be available on all the systems where an external application can be run by the *CLC Server*.
- The *External Applications Client Plugin* must be installed on a CLC Workbench for external applications to be accessible via the Workbench menu system. This plugin can be found in the Workbench Plugin Manager.
- Updates made to existing external applications configurations are registered in the CLC Workbench during a single login session, but to discover new external applications, you need to log out of and back into the *CLC Server*¹.

Figure 12.1 shows an overview of the actions and data flow that occur when an external application is launched on the *CLC Server* via a CLC Workbench.

In general terms the basic work flow is:

¹Prior to CLC Genomics Server 11.0, users had to log out and log back into the server to discover updates to existing external application configurations.

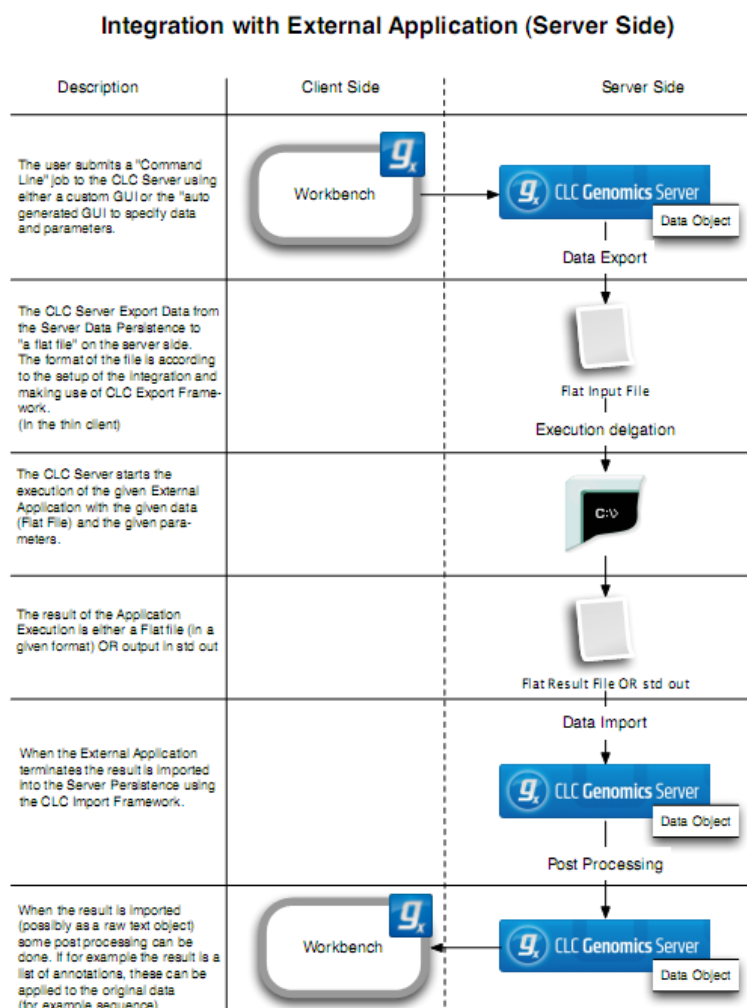


Figure 12.1: An overview of the external applications integration in this example illustrated with CLC Genomics Workbench and CLC Genomics Server.

1. The end user selects input data, sets values for parameters and then launches the external application.
2. The CLC Server exports the input data to a temporary file.
3. The CLC Server starts the command line application, using the values specified by the user and the temporary file(s) as inputs.
4. When the command line application is finished, the CLC Server imports the output into the CLC environment, saving it in the location on the CLC Server specified by the user.
5. Results are then available for viewing and further analysis, for example using a Workbench.

All files saved during the execution of an external application are saved within the CLC environment. Temporary files are created outside the CLC environment during the execution of a third party tool and are deleted after the process completes.

Configurations must be saved by clicking on the Save button at the bottom of the editing section before they can be tested via a CLC Workbench or the CLC Server Command Line Tools.

After a configuration has been saved, the external application name will appear in a list under the External applications tab of the web administrative interface. A small checkbox appears to the left of each name. When checked it means that application is accessible to CLC Workbenches that are logged into the *CLC Server* with the External Applications Client Plugin installed, and also to users of the CLC Server Command Line Tools. To remove access to an external application by end-users, without deleting it, deselect the checkbox beside it.

We will describe integration of third party command lines tools through a series of examples. We start with a basic example and work towards more complex situations.

12.1 General configuration of external applications

Here, we describe the configuration of an external application that uses the system *cp* command to create a copy of the sequence data that a user selects. This is an artificial, but simple example, allowing us to explain many aspects of external application configuration using a command we expect will be on server systems.

Figure 12.2 shows a configuration for the *cp* external application. It is named "Copy sequence data", and would result the data selected by a user being exported from the CLC Genomics Server in fasta format, copied, and then the copied file being imported back into the CLC Genomics Server.

Overview of configuration steps

1. Log into the web administrative interface for the CLC Genomics Server and click on the **External applications** tab.
2. Click on the **New configuration...** button.
3. Add a name into the **External application name** field. This is the name of the external application that a user will see via a Workbench or the CLC Server Command Line Tools.
4. Enter the command to be launched when a user runs the external application into the **Command line** field. This means the command itself, its parameters and values. Any value that the *CLC Server* uses, e.g. to define exports or imports, and any values that should be available to a user to configure, must be enclosed by *{curly brackets}*.

Each time a set of curly brackets is entered into the **Command line** field, a corresponding entry is added into the **General configuration** area. The name of that entry will be the text entered between the brackets. The same name will be presented to end users for the values they have access to, either as a parameter label in a Workbench wizard, or as a CLC Server Command Line Tools parameter name.

5. Configure each parameter using the options offered within the **General configuration** area. All entries in the command line that are enclosed by curly brackets will be substituted at run time by values. The following sections describe how to configure an external application so the relevant information is substituted.

delete... ☒ ▼ Copy sequence data v1.0

This external application can be run on single server or job node setups only. For execution on grid nodes, configure a shared temp-dir un

External application name
Copy sequence data

Command line
cp {Sequence to copy} {Name of copied sequence}

General configuration

Sequence to copy

User-selected input data (CLC data location) FASTA (.fa/.fsa/.fasta) Edit parameters

Name of copied sequence

Output file from CL No standard import or map to high-throughput

▼ High-throughput sequencing import / Post processing

Add new

► Stream handling

► Environment

► End user interface

Save as... Save

Figure 12.2: Setting up the `cp` command as an external application. Here, two values are specified. The type chosen for the first one results in the end user being prompted to select the data to copy. The type chosen for the second one specifies that the output from the `cp` command will be in fasta format and should be imported back into the CLC Genomics Server using the appropriate importer.

12.1.1 Parameter value configuration

A description of the value types that can be configured is given below. A tool tip providing a description of each type is also available by hovering the mouse cursor over it.

Some types allow a default value to be entered. This is indicated by the presence of an empty field beside it. Default values are displayed via the Workbench wizard and in the CLC Server Command Line Tools help. A default value specified this way will be used if the user does not change it, or in the case of the CLC Server Command Line Tools, when the user does not specify the parameter the value is associated with.

Another level of configuration is available for exporters, high throughput sequencing importers and post-processing tools, where the administrator can set default values and also choose if these should be visible and editable by end users. This is described in section 12.1.2 and section 12.1.3 respectively.

Value types

- **Text** - The end user can provide text that will be substituted into the command at runtime. A default value can be configured.
- **Integer** - The end user can provide a whole number that will be substituted into the command at runtime. A default value can be configured. If no value is set, then 0 is the default used.
- **Double** - The end user can provide a number that will be substituted into the command at runtime. A default value can be configured. If no value is set, then 0 is the default used.
- **Boolean text** - A checkbox is shown in the Workbench wizard interface. If the user checks the box, the given text will be substituted into the command at runtime. If the box is unchecked, this means that no value will be substituted.

- **CSV enum** - A drop down list is presented to a Workbench end user, from which they can choose a desired option. A corresponding value that will be substituted into the command at runtime. To configure this parameter type, enter a comma delimited list of the values to be substituted at runtime into the first box, and a comma delimited list of corresponding labels to display to end users in the second box. Each entry in a given list should be unique and the two lists should be of equal length.
For an example of this, please see section 12.6 on setting up Velvet as an external application.
- **User-selected input data (CLC data location)** - The end user should specify one or more input files from those stored on the CLC Server. The data selected will be exported from the CLC Server, so which exporter, and any additional parameter settings for that exporter, need to be configured. This is described in detail in section 12.1.2.
- **User-selected files (Import/Export directory)** - The end user should specify one or more input files from those stored in an Import/Export area on the CLC Server. These would typically not be .clc files. Files can be configured so they are pre-selected for the end user. The end user can deselect pre-configured files when launching the external application via the Workbench.
- **Output file from CL** - This option specifies how the output of the external tool should be handled. This is described in detail in section 12.1.3.
- **File** - The end user should specify an input file from their local machine. This would typically not be a .clc file
- **Context substitute** - The options are:
 - *CPU limit max cores* The core limit defined for the server that executes the command will be substituted.
 - *Name of user* The name of the user who launched the external application will be substituted.

This parameter is not visible to the end user.

- **Boolean compound** - This enables the creation of a checkbox. If checked, the end user is presented with another option as configured by the administrator. If not checked, the option associated with the checkbox is grayed out. Whether the box is checked or unchecked by default can be configured.

12.1.2 Configuring export from the CLC Server for an external application

The first step when running an external application is the export of data from the CLC Server to the system the software is running on. Starting with CLC Genomics Server 10.0, export parameters can be configured.

Click on the button labeled **Edit parameters** beside the relevant export. In figure 12.2, the "Sequence to copy" parameter will be exported in fasta format. Clicking on the Edit parameters button leads to a window opening, as shown in figure 12.3.

Parameters can be configured by:

- Changing the values of default fields. To edit fields that are locked by default, click on the symbol of the lock image to open the lock, then make changes.

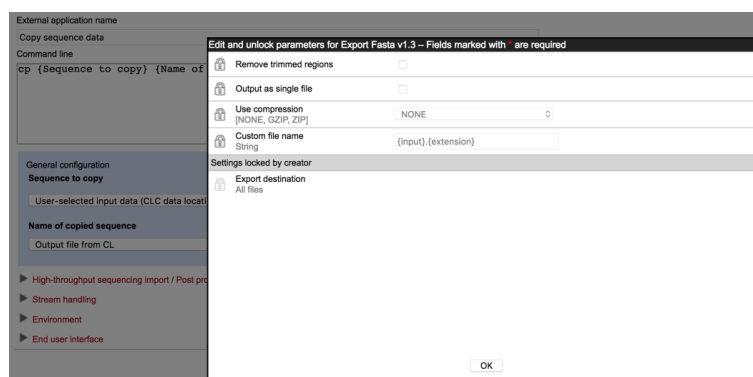


Figure 12.3: The parameters for exporters are configured in a window like the one shown. Here, the *Edit* parameter button for the "Sequence to copy" parameter has been clicked. All the parameters here have a locked symbol beside them, so none of them will be offered to the end user to view or change.

- Choosing the parameters to be visible and editable by end users when they run the tool. A parameter with an unlocked symbol beside it will be shown to the end user and will be editable by them. Locked parameters will not be shown and cannot be changed by end users.

If an exporter is configured in a way that will lead to multiple output files, then the full path to each output file will be substituted in the command at runtime. The external application itself must be able to handle the outputs generated.

Tips for configuring export parameters:

- The parameters listed in the configuration window are *all* those available for the exporter. Unlocking all of these can result in a confusing, and sometimes conflicting, set of options for end users. We recommend that for parameters that are locked by default, you unlock only those you know to be of interest to your users.
- A simple way to explore how many files an exporter will generate with a given configuration is to set up an external application using the echo command and a single parameter linked to the exporter of interest. Set up the Standard out handling to Plain text. This is described in section 12.2. The output from such an external application is a file, which is re-imported into the CLC Server as a text file. This file contains the full paths to the files the exporter created.

12.1.3 Handling the output of an external tool

When **Output file from CL** is chosen for a parameter in the **General configuration** area, a drop down list appears beside it. The possibilities, and how those are configured, are described below.

In the last field presented, the name of the file the external process is expected to produce can be entered. If this field is left blank, the base name of the file produced by the external tool will be used as the base name for the data element imported into the CLC Server.

- **Do nothing with the output of the external command.** Choose the option "No standard

import or map to high throughput sequencing importer". Then do no further configuration of this parameter.

- **Link the output to a high throughput sequencing importer or a tool on the CLC Server that should post-process the results.** Choose the option "No standard import or map to high throughput sequencing importer" and then configure the relevant high throughput sequencing or post processing tool as described in section [12.1.3](#)
- **Use a standard importer to import the results into the CLC Server.** Select the relevant importer from the list displayed. Specifying a default filename, including the relevant suffix (e.g. .fasta, .xlsx), in the final field is recommended.

If the import type **Automatic** is selected, the importer used is determined by the filename suffix in combination with a check of the format of the elements in the file. If the file type is not recognized, it will be imported as an external file. A list of file formats, including the expected filename suffix for each format, can be found in the appendix of the CLC Genomics Workbench manual: Read more about search here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Local_search.html.

When **Output file from CL** is chosen for at least one parameter, the end user will need to provide a location on the CLC Server to store the results. This will be the case even if the output of the external file will not be imported, as log files will still be written to the location selected.

12.1.4 High throughput sequencing importers and post processing tools

Data generated by the external tool can be imported using a high throughput sequencing importer or processed by a post processing tool on the CLC Server. A general description of this is given in this section. For further details, please refer to section [12.7](#), which covers configuring Bowtie as an external application. There, two post processing tools are configured for the import of the results.

To configure a high throughput sequencing importer or post processing tool:

- Choose the option **Output file from CL** for the relevant parameter.
- Choose the option "No standard import or map to high throughput sequencing importer". Despite the name, this will also allow connections to be made to post processing tools.
- Click on the small triangle beside the **High-throughput sequencing import / Post processing** heading to reveal that section.
- Click on the **Add new** button.
- Select the high throughput sequencing import or post processing tool of interest from the list.
- Click on the button below the list called **Edit and map parameters....** This button appears when a tool is selected.

A new window showing all the parameters for that tool will appear. See figure [12.4](#) for an example.

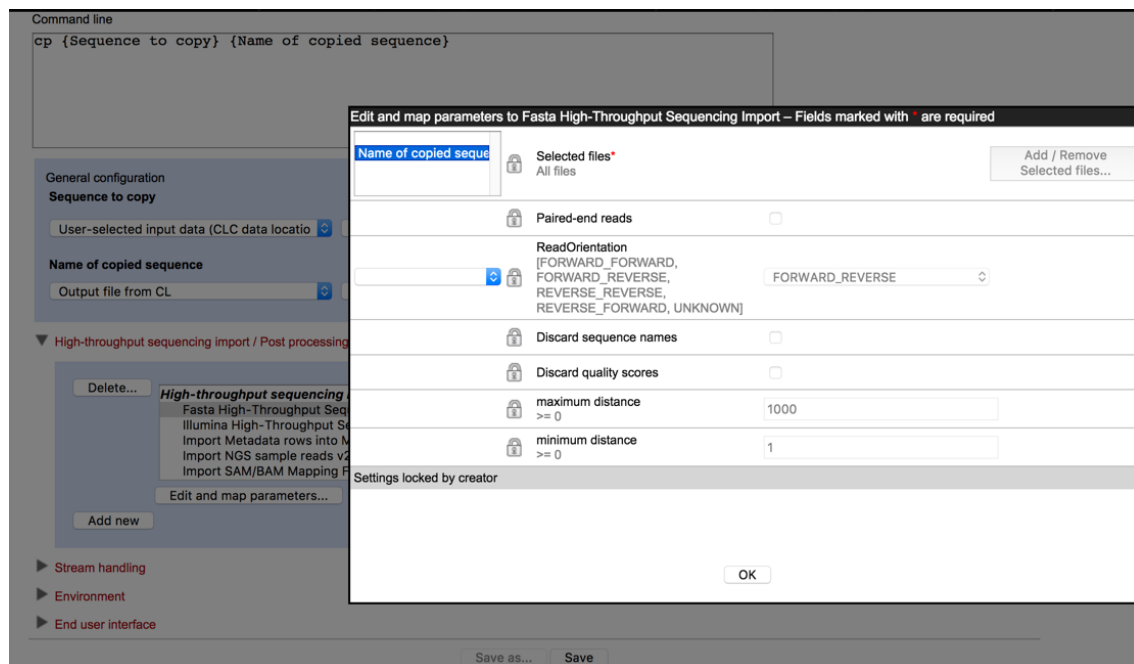


Figure 12.4: The parameters for high throughput sequencing importers and post processing tools are configured in a window like this. Here, the external application parameter "Name of copied sequence" has been clicked on next to "Selected files" to select it as input to the tool. Other parameters have been left with their default settings. As all the parameters show a locked symbol, none of them will be offered to the end user to view or change.

- Select the relevant external application parameter to be used as input to the tool by clicking on it in the list.
For high throughput sequencing tools, the input parameter is usually called "Selected files". For post processing tools, the name of the input field varies, but the word "(input)" is usually present by that parameter name.
- Configure any of the other parameters that you wish to. This includes:
 - Changing the values of default fields. To edit fields that are locked by default, click on the symbol of the lock image to open the lock, then make changes.
 - Choosing the parameters to be visible and editable by end users when they run the tool. A parameter with an unlocked symbol beside it will be shown to the end user and will be editable by them. Locked parameters will not be shown and cannot be changed by end users.

After configuration is complete, the text displayed in the General Configuration panel for the second drop down field of the relevant parameter will indicate which tool it has been linked to. For example, if the post processing tool: "Import SAM/BAM Mapping Files" was configured and the relevant output mapped to it, the text in the General configuration area for the relevant output would show: "Linked with Import SAM/BAM Mapping Files" in the second field.

If you change your mind about linking to a high throughput sequencing importer or post processing tool from an external application parameter, you must remove the connection between the two. This would usually be done by deleting the relevant configuration in the High-throughput sequencing import / Post processing area.

12.2 Stream handling

There is a general configuration of stream handling available.

The stream handling shown in figure 12.5 allows you to specify where standard out and standard error for the external application should be handled.



Figure 12.5: *Stream handling.*

Basically, you can choose to ignore it, or you can import it using one of the importers available on the server. For some applications, standard out produces the main result, so here it makes sense to choose an appropriate importer. But also for debugging purposes it can be beneficial to import standard out and standard error as text so that you can see it in the Workbench after a run.

12.3 Environment

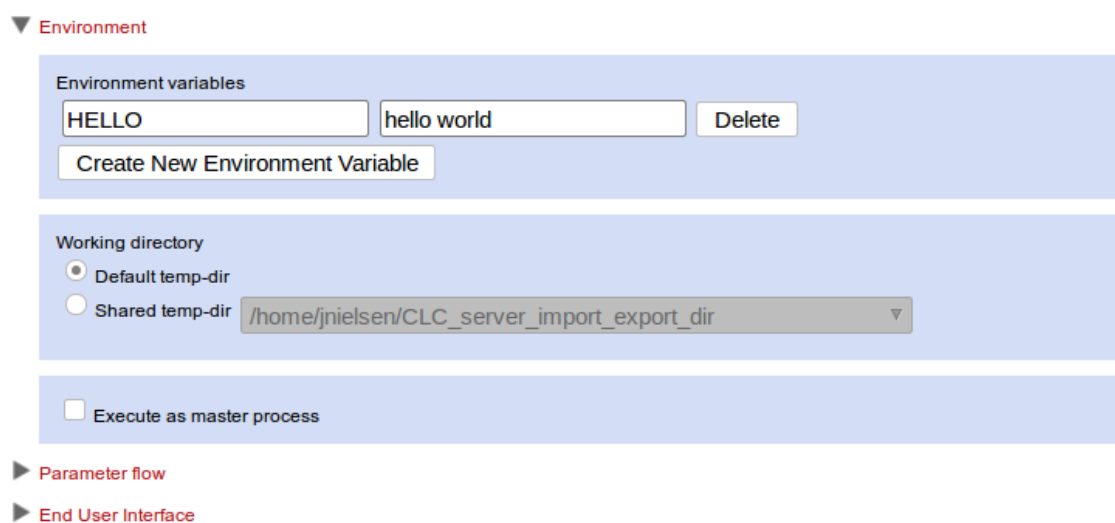


Figure 12.6: *The menu for configuring the execution environment of the external application.*

In the *Environment* sub-menu, it is possible to configure a few central aspects of the environment in which the external application will execute. A screenshot of the sub-menu can be seen in Fig. 12.6.

12.3.1 Environmental variables

In this section it is possible to create and set a default value for environment variables that should be present for the external application when it executes. In the example of Fig. 12.6 an

environment variable named "HELLO" with value "hello world" will be present in the execution environment of the external application.

12.3.2 Working directory

Define the area where temporary files will be stored. The *Default temp-dir* option uses the directory specified by the `java.io.tmpdir` setting for your system. The *Shared temp-dir* option allows you to set one of the directories you have already specified as an *Import/export directory* as the area to be used for temporary files created.

Choosing the *Shared temp-dir* option means that temporary files created will be accessible to all execution nodes and the master server, without having to move them between machines.

For an external application to be executable in a grid environment, the *Share temp-dir* option **must** be chosen for the working directory.

For external applications that will be run on job nodes, one can choose either the *Shared temp-dir* or the *Default temp-dir* option. Here, the *default temp-dir* would not normally be an area shared between machines, and thus the choice of the *Default temp-dir* means that files will be moved between the master and job node(s).

If you configure the *Shared temp-dir* for an external application, this area must:

- Be configured in the Import/Export directories area under the Main Configuration tab (see section 3.3).
- Be a shared directory, accessible to your all machines that will execute the external application.

12.3.3 Execute as master process

When the **Execute as master process** option is enabled, third party application(s) are executed on the master machine of a job or grid node setup. Export, import and post-processing steps are still run on the execution environment selected by the user when launching the external application. This setting has no effect for single server setups, where all execution happens on the same system.

With this option unselected, all stages of the external application are run on the execution environment selected when launching the external application. This is the default.

Execute as master process is not recommended for use with memory or cpu intensive tasks as the third party application will be launched on the master system without consideration of how busy that system is, or what processing capabilities it has.

For grid setups: Whether or not the **Execute as master process** option is enabled, export, import and post-processing steps will be run on a grid node if a grid execution environment is selected when launching an external application. Thus the Working directory option must be set to *Shared temp-dir* either way. See section 12.3.2 for more information about the Working directory settings.

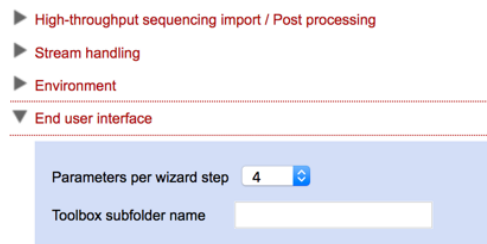


Figure 12.7: *End user interface* settings alter how the external application appears in a Workbench client.

12.4 End user interface

Settings that affect how an external application will look in a Workbench client are in the *End user interface* section. See figure 12.7.

- **Parameters per wizard step:** Enter the number of parameters to present in a given wizard step when the application is launched from the Workbench Toolbox. The default value is 4. With this value, up to 4 parameters will appear on each wizard page that the user steps through when launching the job via a Workbench.
- **Toolbox subfolder name:** Text entered in this field specifies the name of a subfolder under Toolbox | External Applications in Workbenches. The application will be listed under this subfolder. If a subfolder of the specified name does not already exist, it will be created. When this field is empty, the application will be listed directly under Toolbox | External Applications.

Additional notes:

- The text entered is case sensitive; "foldername" and "Foldername" refers to 2 different subfolders.
- The creation of subfolders of subfolders is not supported. If you enter text like "myfolder/subfolder", a single folder of exactly that name would be listed under Toolbox | External Applications in the Workbench.

12.5 Using consistent reference data in external applications

For external applications that include third party tools that use reference data, such as reference genomes, annotations, primer sets, etc., the same reference data should be used for both the third party and CLC processing steps. To achieve this, reference data must be in locations and formats that can be used by each relevant application. This usually means either importing reference data into a *CLC Server* or exporting it from a *CLC Server* and placing it where the third party application can access it when it is run.

Reference data is imported into a *CLC Server* using client software, either a *CLC Workbench* or the *CLC Server Command Line Tools*. Data to be imported should be copied into an Import/export directory first, from where it can be imported.

If you do not already have the reference data you need, the Reference Data Manager of the *CLC Genomics Workbench* can be used to download data from QIAGEN servers and some public repositories such as Ensembl to the *CLC Server*. Further details about this can be found in section 3.2.2.

Reference data is exported from a CLC Server using standard export functionality of the client software, where the relevant data elements are selected in the Navigation Area of CLC Workbenches, or specified using CLC URLs when using the CLC Server Command Line Tools. Exported data is put into an Import/export directory, from where it should be moved to a location that can be accessed by the third party tool when it is run.

Import/export directories are configured under the Main tab of the server web administrative interface, as described in section 3.3.

Data import and export are described in the client software manuals: <https://www.qiagenbioinformatics.com/support/manuals/>.

12.6 Velvet integration

Velvet [Zerbino and Birney, 2008] is a popular de novo assembler for next-generation sequencing data.



The Velvet package includes two programs that need to be run consecutively. The external applications system on the CLC Server is designed to call one program, so a wrapper script is needed to make the needed consecutive calls to the Velvet applications.

An example script as well as a configuration file are available in a zip file at <http://www.resources.qiagenbioinformatics.com/external-applications/velvet-example.zip>. These will be used to illustrate how this sort of application can be configured as an external application on a CLC Server.

12.6.1 Installing Velvet

To get started, you need to do the following:

- Install Velvet on the server computer. The program and installation information is available from <https://github.com/dzerbino/velvet/tree/master>. If you have job nodes, Velvet will need to be installed on all the nodes that will be configured to run it.
- Download the scripts and configuration files from <http://www.resources.qiagenbioinformatics.com/external-applications/velvet-example.zip>. These files have been created assuming that Velvet is installed in `/usr/local/velvet`. If it is installed elsewhere, please update the files with the correct path to the program on your server.
- Check to ensure execute permissions are set on the `velvetg` and `velveth` executable files in the Velvet installation directory. These must be executable by the user that owns the CLC Server process.
- Unzip the `velvet-example.zip` file and place the `clcbio` folder and its contents in the Velvet installation directory. This contains a script (`velvet.sh`) that links the two Velvet programs, `velvetg` and `velveth`, together. If the Velvet binaries are not in the folder `/usr/local/velvet`, you will need to edit the line that starts with `exe=` to include the correct path.
- Set the permissions on the `velvet.sh` script in the `clcbio` subfolder so that it can be executed by the user that owns the CLC Server process.

- Use the `velvet.xml` file as a new configuration on the server: Log into the server via the web interface and go to the **External applications** () tab under **Admin** () and click **Import Configuration**.

When the configuration has been imported, click the **CLC bio Velvet** header and you should see a configuration as shown in figure 12.8.

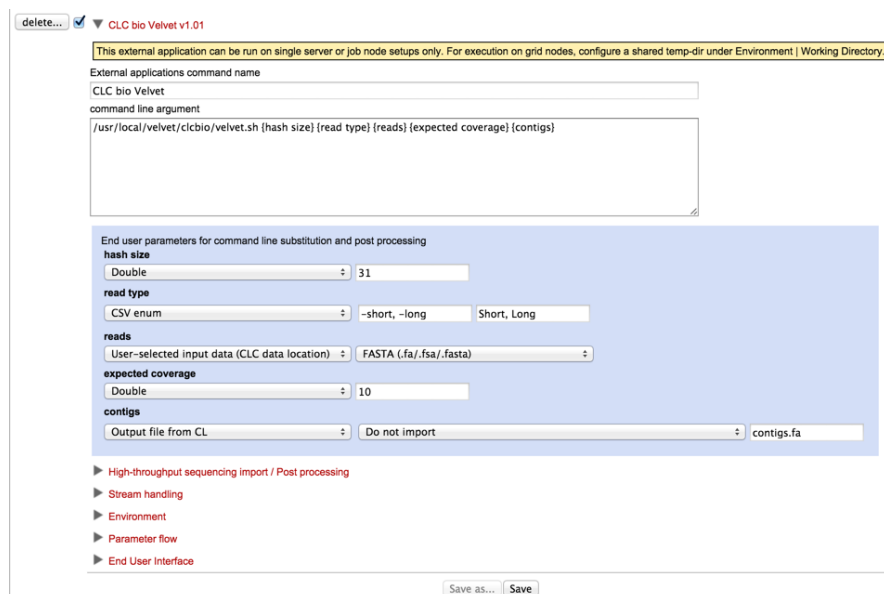


Figure 12.8: The Velvet configuration has been imported.

- Update the path to the Velvet installation at the very top if necessary.



If you wish to execute this job on grid nodes, then a shared temp-dir must be specified in the Working directory section of the Execution area of the configuration. See section 12.3.2 for details.

If you see a small red exclamation point beside the external application name, then something is wrong and needs to be attended to. The specific area where there is a problem should also be identified by a red exclamation mark.

This is seen, for example, when the versions of a High-throughput sequencing import or Post-processing step specified in the configuration file is different to the version on the CLC Server. This can occur when the configuration was set up on an older version of the CLC Server than the one running. This situation is shown in figure 12.9. It is easily resolved by expanding the High-throughput sequencing import / Post-processing section, selecting the relevant tool and then saving the configuration.

12.6.2 Running Velvet from the Workbench

To run Velvet, open a Workbench with the **External Applications Client Plugin** installed. Then:

- Go to:
Toolbox | External Applications () | CLC bio Velvet ()
- Confirm where you wish to run the job.

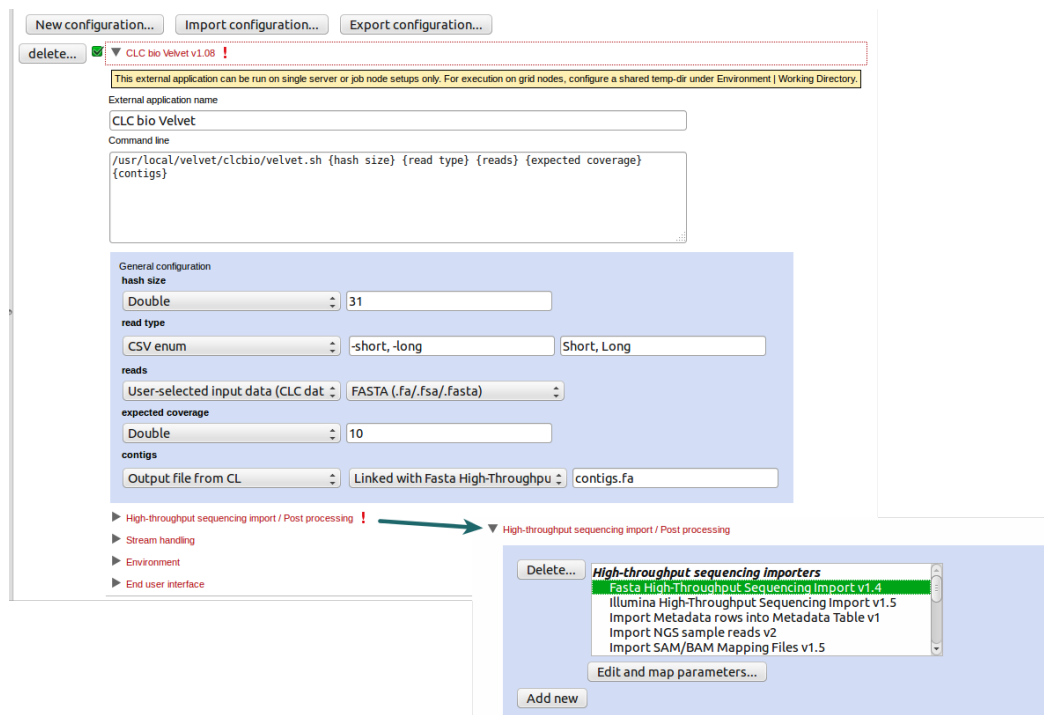



Figure 12.9: Problems with the configuration are indicated by red exclamation marks, as can be seen here by the external application name and by the area with the problem, the High-throughput sequencing import/ Post-processing section. Expanding this section, the problem can be resolved by selecting the relevant tool.

- Select  the reads to assemble and configure the Velvet parameters, as shown in figure 12.10.

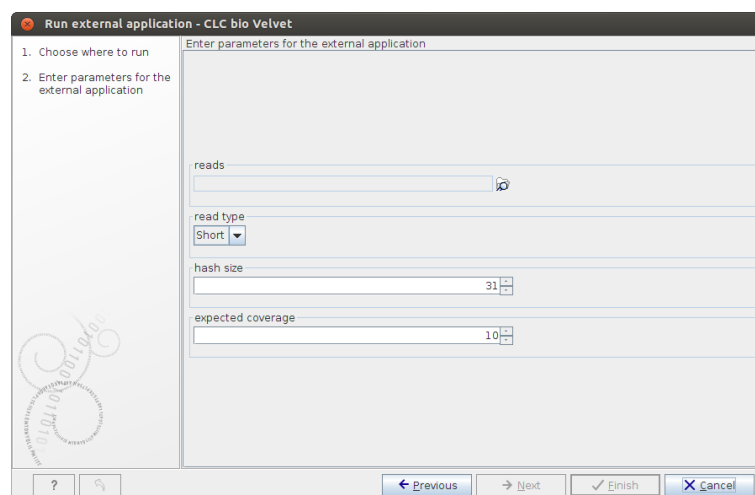


Figure 12.10: Configuring Velvet parameters from a Workbench.

- Click on the button labeled **Next**.
- Specify where to save the results.
- Click on the button labeled **Finish**.

The process that follows has four steps:

1. The sequencing reads are exported by the server to a fasta format file. This is a temporary file that will be deleted when the process is done.
2. The velvet script is executed using the fasta file and the user-specified parameters as input.
3. The resulting output file is imported into the save location specified in the save step of the Workbench dialog, and the user is notified that the process is done.
4. All temporary files are deleted

12.6.3 Understanding the Velvet configuration

The Velvet configuration file is explained here as a specific example of all external application configuration files.

Going back to figure 12.8, there is a text field at the top. This is where the command expression is created, in this case:

```
/opt/local/velvet/clcbio/velvet.sh {hash size} {read type}  
{reads} {expected coverage} {contigs}
```

The first part is the path to the script. The following parts are parameters that are interpreted by the server when calling the script. Parameters to be interpreted are surrounded by curly brackets { }. Note that each parameter entered in curly brackets gets an entry in the panel below the command line expression.

The first parameter, `hash size`, can be entered as a **Double** (which is a number that can take decimal values in computer parlance). The user provides a value when they launch Velvet. A default value is provided in the configuration (31).

The second parameter is `read type`, which has been configured as a **CSV enum** which means a list of possible values that the user can choose from. The first part of the configuration information consists of the parameters to be used when calling the velvet script (`-short`, `-shortPaired`, `-long`, `-longPaired`), and the second part is the more human-readable representation that is to be shown in the Workbench (`Short`, `Short Paired`, `Long`, `Long Paired`).

The third parameter is `reads` which is used for the input data. When the **User-selected input data** option is chosen, a list of all the available export formats is presented. In this case, Velvet expects a fasta file. When a user starts Velvet from the Workbench, the server starts exporting the selected input data from `clc` format to a temporary fasta file before running the script.

The `expected coverage` parameter is similar to `hash size`.

The last parameter, `contigs`, represents the output file, as indicated by the choice of "Output file from CL" as the type. Here, you specify how the output from velvet should be handled. A list of standard import formats is provided, as well as the option not to import using those tools. Choosing not to import using those tools means that you can choose a high throughput importer instead from the section High-throughput sequencing import/ Post-processing section.

For this example, Do not import is the action set for the contigs parameter. Then, below, in the High-throughput sequencing import/ Post-processing section, the Fasta High-throughput Sequencing Import tool has been selected. Thus, when the results from velvet are ready, they are imported into the CLC Server using that tool and saved where the user indicates when they run the job.



12.7 Bowtie integration

In this example, we show how to integrate Bowtie [Langmead et al., 2009], a popular tool for mapping short sequencing reads to a reference sequence, using the External Applications functionality. Here, two post processing tools will be configured, allowing us to import more than one output generated by bowtie into the CLC Server.

Importing the configuration file provided as an example, and following the instructions in this section, leads to three tools being made available to users logged into the CLC Server via their Workbench or the Command Line Tools: CLC bio Bowtie Build Index, CLC bio Bowtie List Indices and CLC bio Bowtie Map.

12.7.1 Installing Bowtie

To get started:

- Install Bowtie from <http://bowtie-bio.sourceforge.net/index.shtml>. We assume that Bowtie is installed in `/usr/local/bowtie` but you can just update the paths if it is placed elsewhere.
- Download the scripts and configuration files from <http://www.resources.qiagenbioinformatics.com/external-applications/bowtie-pp2.zip>
- Place the `clcbio` folder and contents in the Bowtie installation directory. This folder contains the scripts used to wrap the Bowtie functionality. Those wrapper scripts are what is then configured via the External Applications folder.
- Make sure execute permissions are set on these wrapper scripts and on the executable files located in the Bowtie installation directory. The user that will execute these files will be the user that started the CLC Server process.
- Import the `bowtie-pp2.xml` file as a new configuration on the server by going to the **External applications**  tab under **Admin**  in the web administrative interface and clicking on the button labeled **Import Configuration**.

The `bowtie-pp2.xml` file contains configurations for three tools associated with Bowtie: CLC Bowtie build index, CLC Bowtie list indices and Bowtie Map. If you already have a set of indices you wish to use and the location of these is known to the system via the `BOWTIE_INDEXES`, then you can just use the Bowtie Map tool via the Workbench and specify the index to use by name.

Otherwise, you can build the index to use using the CLC Bowtie build index tool. Here, unless you edit the wrapper scripts in the files you download from CLC bio, the indices will be written to the directory indicated by the `BOWTIE_INDEXES` environmental variable. If you have

not specified anything for this, indices will likely be written into the folder called `indexes` in the installation area of Bowtie. Please ensure that your users have appropriate write access to the area indices should be written to.

From ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/ you can download pre-built index files of many model organisms. Download the index files relevant for you and extract them into the `indexes` folder in the Bowtie installation directory.

When configuring Bowtie to run as an external application on master-node setups, the **Environment** configuration will need to be edited. See 12.7.3 for details. In addition, Bowtie indices will have to be placed somewhere accessible to all nodes. One option could be areas configured as Import/Export areas on the CLC Server.

The rest of this section focuses on understanding the integration of the Bowtie Map tool in particular.

12.7.2 Understanding the Bowtie configuration

Once the `bowtie-pp2.xml` configuration file has been imported, you can click the **CLC bio Bowtie Map** header to see the configuration as shown figure 12.11.

The screenshot shows the 'CLC bio Bowtie Map v1.04' configuration window. At the top, there is a 'delete...' button and a green checkmark icon. Below this is a yellow warning box: 'This external application can be run on single server or job node setups only. For execution on grid nodes, configure a shared temp-dir under Environment | Working Directory.' The 'External application name' field is 'CLC bio Bowtie Map'. The 'Command line' field contains: `/usr/local/bowtie/bowtie/clcbio/bowtie_map.sh {reads} {bowtie index} {sam file} {max number of multimatches} {multimatch filename} {max number of mismatches} {report all matches}`. The 'General configuration' section is highlighted in blue and contains the following settings:

- reads**: User-selected input data (CLC dat) dropdown, FASTA (.fa/.fsa/.fasta) dropdown.
- bowtie index**: Text field with 'coli'.
- sam file**: Output file from CL dropdown, Linked with Import SAM/BAM Ma dropdown, sam_output text field.
- max number of multimatches**: CSV enum dropdown, 2,3,4 text field, 2,3,4 text field.
- multimatch filename**: Output file from CL dropdown, Linked with Fasta High-Throughput dropdown, multimapped.fasta text field.
- max number of mismatches**: CSV enum dropdown, 0,1,2,3 text field, 0,1,2,3 text field.
- report all matches**: Boolean text dropdown, -a text field.

At the bottom, there are four expandable sections: 'High-throughput sequencing import / Post processing', 'Stream handling', 'Environment', and 'End user interface'. At the very bottom are 'Save as...' and 'Save' buttons.

Figure 12.11: The Bowtie configuration has been imported.

From an end-user perspective, when the configuration on the CLC Server is complete, they will be able to launch the CLC bio Bowtie Map tool via their Workbench Toolbox. A wizard will appear, within which they will select the sequencing reads to be mapped, identify the pre-built index file of the reference sequence to use and set a few parameters. The bowtie executable will then be

executed on the server system and the results generated will be imported into the CLC Server using post processing tools. The sam mapping file is imported using the Import SAM/BAM Mapping Files tool. A fasta file of sequences mapping to multiple locations is imported using the Fasta High-Throughput Sequencing Import tool.

Below, we step through the General configuration panel and then explain the configuration of the post processing tools that handle the outputs from the bowtie analysis.

General configuration panel

Each of the parameters (items within curly brackets) written into the "Command line" box is presented as an item in the General configuration panel. There, we define the type of information each parameter expects or represents and default values, where relevant.

To understand how these parameters relate to the information that will be passed to the native bowtie executable, please refer to the bowtie_map.sh script in the clcbio folder that should now be in place in the bowtie distribution folder.

Stepping through the parameters in the order they appear in the Command line area of the configuration, and thus the order they appear in the General config panel:

- The `reads` parameter refers to the data that will be provided to bowtie to map. The **User-selected input data** option means the user will be able to select data in a CLC File or Data Location. This data will be exported from the CLC File or Data Location such that the bowtie tool can use it. The second element in this line specifies the format the data should be exported in. This is set to **FASTA (.fa/.fsa/.fasta)** as this is the format the bowtie tool expects sequencing read data.
- The `index` parameter is expecting the name of a bowtie index. Specifying the type **Text** for this parameter means a user will see a box in the Workbench Wizard that they can type text into. Here, a default name, "coli" has been specified, which can be changed by a user launching CLC bio Bowtie Map.

When setting up a tool like this, it would be simpler for users, and much less subject to error, if the type **CSV enum** were selected, and a specified set of indices were listed. Then, a drop down list of options would be provided to the user in a Workbench Wizard, when launching the external application, rather than relying on users typing in the correct names of available bowtie indices.

- The `sam file` parameter refers to the sam mapping file that bowtie will generate as one of its results file. Thus, the type is set to **Output file from CL**. Import of sam files into the CLC Server involves a tool that requires user input. Thus, a post processing tool is configured. This can be seen immediately by the text in the second drop-down box: "Linked with Import SAM/BAM Mapping Files".

If a parameter with type **Output file from CL** is not mapped to a parameter of a post processing step, the text displayed is "Do not standard import / map to high-throughput sequencing importer". Mapping of outputs to post processing tools is described in more detail below.

The last entry in the configuration of the `sam file` parameter is the name of the sam output file that bowtie should generate. Here it is set to **sam_output**. This file name is used by the bowtie command. The Workbench or Command Line Tools user never sees it.

- The `max number of multimatches` parameter allows a user to select the maximum number of locations a read can map equally well to for it to be included in the mapping. The type is set to **CSV enum**, which means a user will be able to select a value from a drop down list of the 3 values listed in the last field (2,3,4). The first value will be the default. The values in the middle field are those passed to the bowtie wrapper script and then onto bowtie. So, for example, if "2,3,4" were entered in the middle field, and "two, three, four" in the last field, a user could select the option "two", and bowtie would be sent the value 2.
- The `multimatch filename` parameter refers to another output from bowtie, this one containing fasta formatted reads that match to multiple locations of the reference equally well. Since it is a result file, the type is set to **Output file from CL**. We have decided to use a post processing tool to bring the results back into the CLC Server, the Fasta High-Throughput Sequencing Import tool.
- The `max number of mismatches` parameter allows a user to select the maximum number of mismatches to be allowed between a read and the reference in order for a read to be considered as matching the reference at that location. The type is set to **CSV enum**, and is presented to a user in the same way as the `max number of multimatches` parameter described above.
- The `report all matches` option is one that can be turned on or off. Thus it is set to type **Boolean text**. A user will be presented with a checkbox they can select or deselect in the Workbench Wizard. The value in the text field, here "-a", is the one bowtie will be passed if the user selects the checkbox. If the user does not select the checkbox, this parameter will not be sent to bowtie.

Post processing - importing the results from Bowtie

If you expand click on the **High-throughput sequencing import /Post-processing** link below the General configuration area, you will see that there are two post-processing tools selected: the **Import SAM/BAM Mapping Files** tool and the **Fasta High-Throughput Sequencing Import** tool.

In each case, clicking on the **Edit and map paramaters** button below it will bring up the configuration window for that tool. Here, several types of configuration can be carried out.

1. Mapping of outputs of the external application to inputs of the post processing tool.
2. Locking or unlocking of parameters, determining which parameters users can alter when launching the tool via the Workbench or Command Line Tools.
3. Setting default values for parameters of the external application.

Here, we step through the configuration of the **Import SAM/BAM Mapping Files** tool. The configuration of the **Fasta High-Throughput Sequencing Import** is similar.

The parameters in this configuration window are the **Import SAM/BAM Mapping Files** tool parameters, just as would be offered when that tool is launched directly in a CLC Workbench.

A locked lock symbol by a parameter means that the user will not be given access to this option when launching the tool. Default settings for lock parameters are used. The locked parameters

shown in figure 12.12 indicate that a track will be output rather than a stand-alone read mapping, unmapped reads will be saved, references will not be downloaded from an external source and, had they been, downloaded references will not be saved. Quality scores and sequence names will be kept (not discarded).

By contrast, the References parameter is unlocked. When using the Import SAM/BAM Mapping Files tool, users need to specify where the relevant reference sequences are. Thus, this option should be made available for users to configure when the tool is being launched.

The input to the Import SAM/BAM Mapping Files also needs to be defined. This is done by mapping the relevant output from the bowtie command to the input parameter for the Import SAM/BAM Mapping Files tool. The output from bowtie is defined by the "sam file" parameter, and the relevant input parameter in the import tool is "Selected files". A drop down list of potentially relevant parameters appears to the left of the "Selected files" parameter. In our example, this has already been mapped to the "sam file" parameter of the command, as shown in figure 12.12.

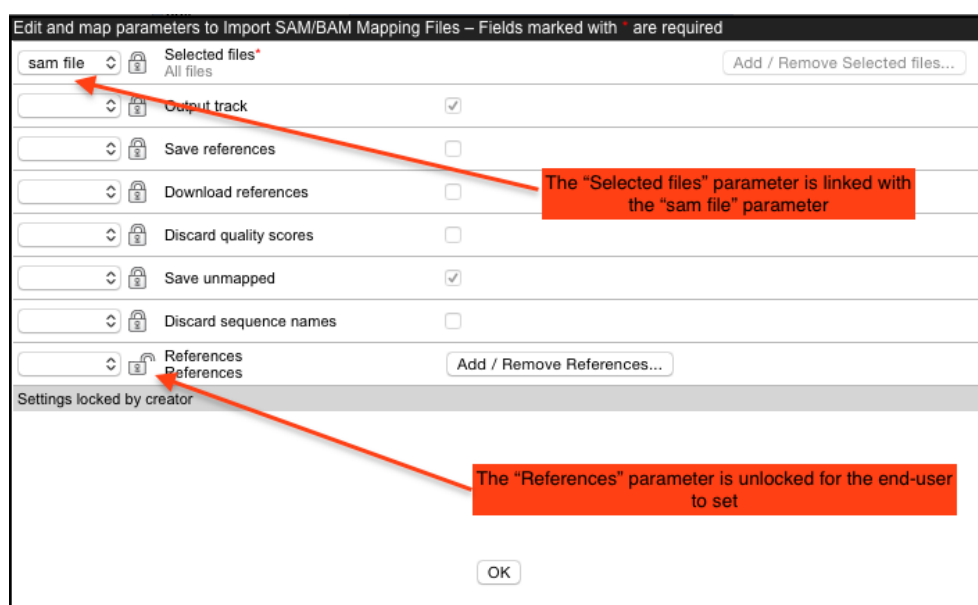


Figure 12.12: Configuration of Import SAM/BAM Mapping Files for import of a sam file after mapping using Bowtie.

Note: The drop-down lists of possibly relevant parameters provided in the post processing tool configuration window are populated based only on the types of parameters (in the General config pane). Any parameters of a type that could be relevant are presented. This means that some parameters appearing in these lists may not make sense contextually.

12.7.3 Setting the path for temporary data

The **Environment** handling shown in figure 12.13 allows you to specify a folder for temporary data and add additional environment variables to be set when running the external application.

Post-processing steps need to access the results files of the external application. Thus, **if you are running on a master-node setup, the directory you choose for these results files must be shared**, that is, accessible to all nodes you plan to have as execution nodes for this external application. This is because different stages of your task could be run on different nodes. For example, the export process could run on a different node than the actual execution of the Bowtie

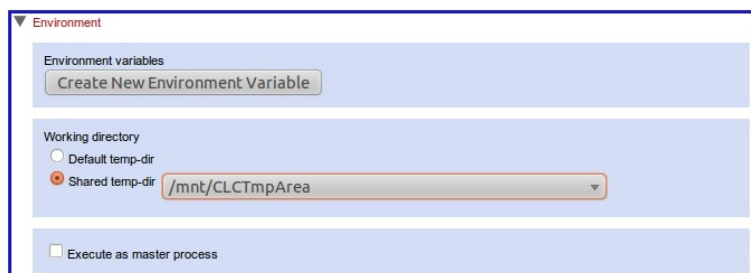


Figure 12.13: Where to save temporary data needs to be defined for Bowtie.

script and the post processing. Thus, in a master-node setup, be it using grid or CLC execution nodes, having this shared temporary area eliminates the overhead of transferring the temporary files between nodes.

12.7.4 Tools for building index files

We have also included scripts and configurations for building index files using the external applications on *CLC Server*. This also includes the possibility of listing the index files available. To get these to work, please make sure the path to the Bowtie installation directory is correct.

The Bowtie distribution itself also includes scripts to download index files of various organisms.

12.8 Import and export of external application configurations

External application configuration files are XML file containing configuration information about one or more external applications. These can be exported from or imported into a *CLC Server*, facilitating backup and exchange.

Exporting external application configurations

To export the configuration information for external applications, click the **Export configuration. . .** button. Select the applications to export the configurations for and click on the **Export** button. See figure 12.14.

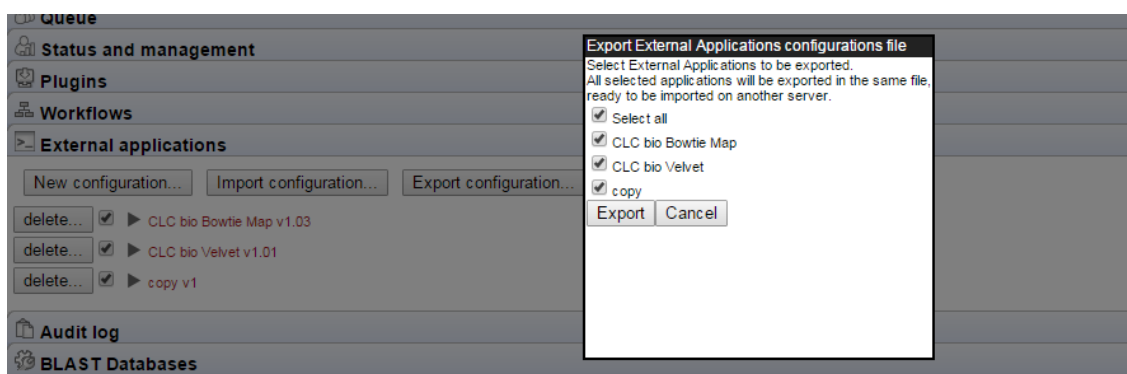


Figure 12.14: Exporting external applications configuration.

Importing external application configurations External application configuration files can be imported *CLC Server* by clicking on the **Import configuration. . .** button. In the window that appears, click on the **Browse** button and select the configuration to import. Then click on the **Import and add External Applications configuration** button.

A dialog is then presented confirming the import. If the imported file included configuration information for an external application with the same internal ID as one already on the CLC Server, the copy on the server will be overwritten. The confirmation dialog will include the names of any external applications overwritten for this reason.

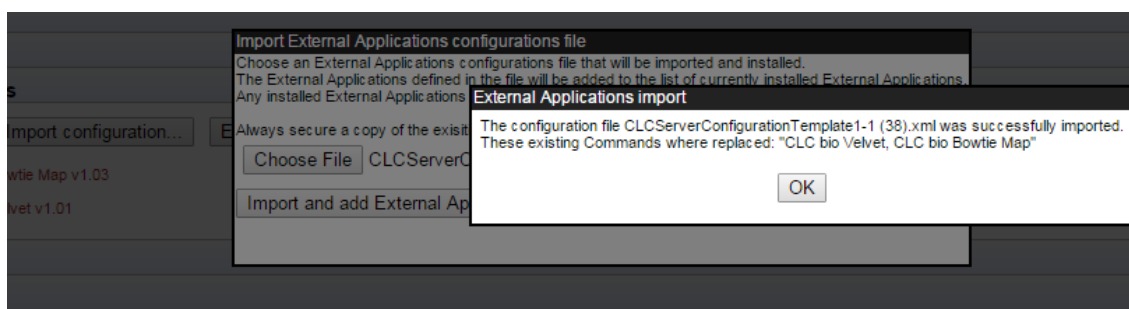


Figure 12.15: Importing an external application configuration file.

12.9 External applications in workflows

Like most other tools available for execution on the CLC Server, tools configured as external applications can be included in workflows. Parameters that are locked in a workflow element will not be offered to the user for editing. Parameters that are unlocked will be.

The inputs and outputs are as configured in the external application itself. In figure 12.16, the external application, CLC bio Bowtie Map is used as a workflow element. In this particular case, the configuration had a single post processing step, the Import SAM/BAM Mapping Files tool. That tool can output a stand-alone read mapping or a read mapping track, and outputs a sequence list containing unmapped reads. Thus these are the output options you see in the workflow element.

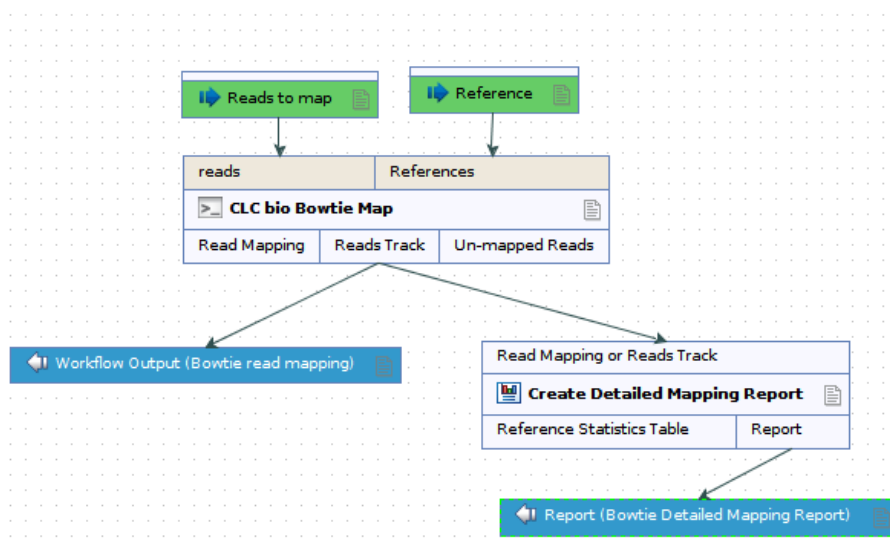


Figure 12.16: The CLC bio Bowtie Map external application used as an element in a workflow. Here, the outputs visible are those provided by the (single) post processing tool configured: Import SAM/BAM Mapping Files

When additional post processing tools are configured, their outputs will be added to those available in the workflow element. See figure 12.17, where a workflow element for the external

application, CLC bio Bowtie Map is present, but in this case, the configuration specified two post processing tools: the Import SAM/BAM Mapping Files tool and the Fasta High-Throughput Sequencing Import tool. As earlier, the three outputs associated with the Import SAM/BAM Mapping Files tool are present. In addition, the output from the Fasta High-Throughput Sequencing Import tool, "Imported reads" is present.

The output channel names make sense for individual tools, but may not make sense in the context of a given external application. For example "Imported Reads" makes sense when you run the Fasta High-Throughput Sequencing Import tool by itself. However, in the context of the CLC bio Bowtie Map external application, it is not indicative of what is really being output. Meaningful names can be set on the workflow output elements themselves though, providing more information to the workflow user about what a particular output is. See for example figure 12.17, where the output from the Fasta High-Throughput Sequencing Import tool, "Imported reads", has been renamed "bowtie multimapped reads".

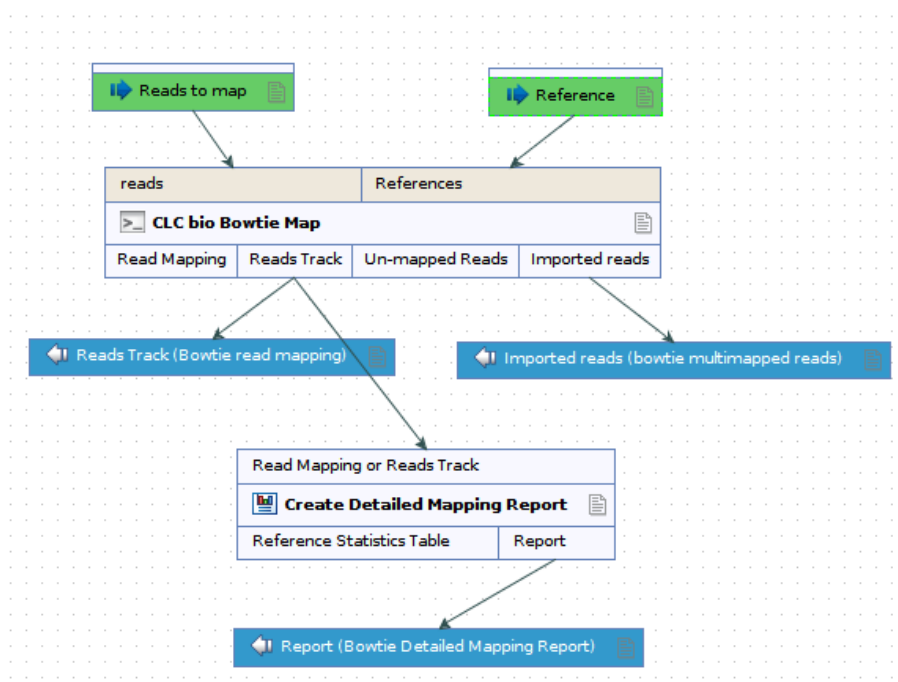


Figure 12.17: The CLC bio Bowtie Map external application used as an element in a workflow. Here, the outputs visible are those provided by the post processing tools configured: Import SAM/BAM Mapping Files and Fasta High-Throughput Sequencing Import.

Limitation: External applications can only be used in a workflow if they are configured with at least one "User-selected input data" and at least one import of the data generated. If the external application does not generate data that is suitable for import then importing the output from standard out or standard error as plain text is recommended. In general it is recommended to always import standard error to help identify potential problems with an external application.

For more details about designing and running workflows, see <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

12.10 Running external applications

External applications can be executed using a *CLC Workbench* or using the *CLC Server Command Line Tools*.

12.10.1 Using a CLC Workbench to launch external applications

After installing the *External Applications Client Plugin* on a CLC Workbench², external applications can be executed from the Workbench menu system by going to:

Toolbox | External Applications (📁)

External applications are listed as individual tools in the Workbench Toolbox, as shown in figure 12.18. Depending on how they were configured, they can also be located within subfolders of the External Applications folder.

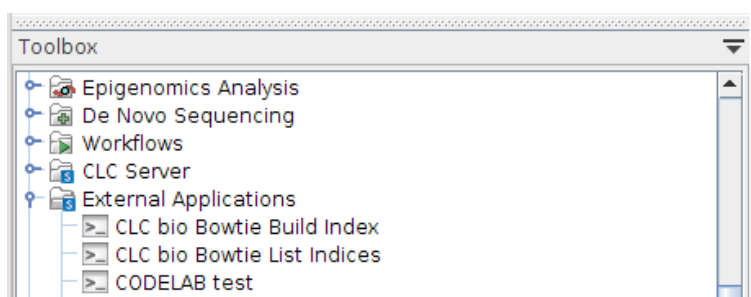


Figure 12.18: Selecting the external application to run.

When an external application is launched, the dialog shown in figure 12.19 is displayed. Depending on your server setup, there may be one or two types of execution environment: The *CLC Server* environment, and grid presets. The CLC Server environment is always present, while grid presets are only shown if they have been configured as described in section 6.3.

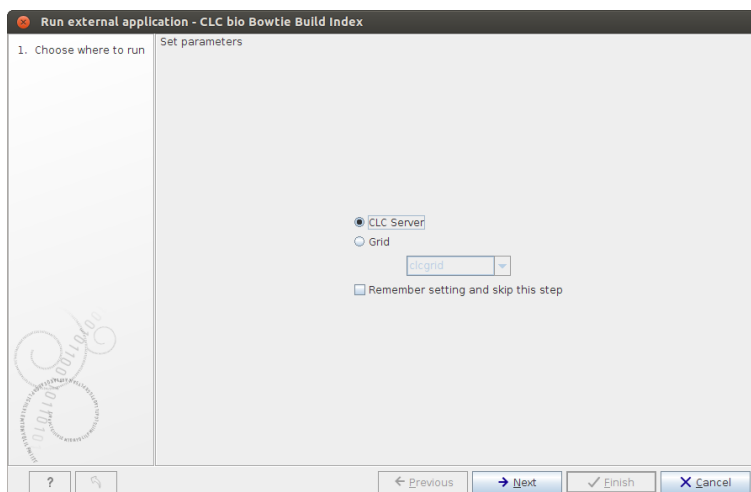


Figure 12.19: Selecting execution environment.

Clicking on the *Next* button starts the progress through wizard steps where values can be entered for any parameters that need to be configured, illustrated in figure 12.20.

²Information about installing plugins on a CLC Workbench can be found in the CLC Workbench manuals, available from <https://www.qiagenbioinformatics.com/support/manuals/>

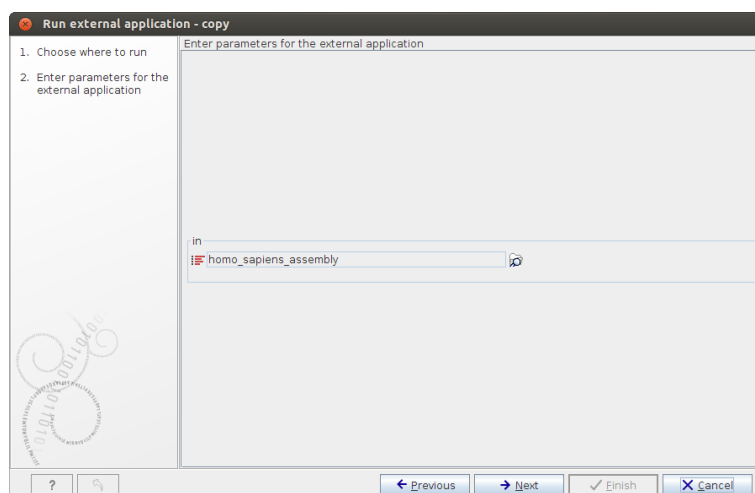


Figure 12.20: Providing values for configurable parameters of an external application. Here, a sequence list has been selected for the in parameter.

12.10.2 Using CLC Server Command Line Tools to launch external applications

The CLC Server Command Line Tools can be used to launch CLC tools, workflows and external applications on a CLC Server. General information about using the CLC Server Command Line Tools can be found in the manual for that software: <https://www.qiagenbioinformatics.com/support/manuals/>.

When launching an external application, the CLC Server execution context will be used unless the `-G` option is present, specifying a grid preset.

Running a CLC Server Command Line Tools command with missing or invalid parameters results in messages about the problem being returned. For example, trying to invoke the `copy` external application described earlier in this chapter with no arguments yields the output below. The final line makes clear the problem: the `-d` and `-in` parameters need to be specified in the command for the job to be executed.

```
clcserver -S <HOSTNAME> -U <USER> -W <PASSWORD> -A copy
Message: Trying to log on to server
Message: Login successful
The following options are available through the command line and the types are as follows:
```

Type	Valid input
----	-----
<Integer>	A decimal number in the range
[-2147483648;2147483647]	
Example: 42	
<Boolean>	The string true or false
Example: true	
<String>	Any valid string. It is recommended
to enclose all strings in '' to	
avoid issues with the shell	
misinterpreting spaces or double	
quotes	
Example: 'text="My text"'	
<ClcFileUrl>	A valid path to a file on the server
or in the local file system	
Example: clc://serverfile/tmp/export	
<ClcObjectUrl>	A valid path to a Clc object on the
server or locally	
Example: clc://server/pstore1/Variant1	
Option	Description
-----	-----

-A <Command>	Command currently set to 'copy'
-C <Integer>	Specify column width of help output.
-D <Boolean>	Enable debug mode (default: false)
-G <Grid preset names>	Specify to execute on grid.
-H	Display general help.
-O <File>	Output file.
-P <Integer>	Server port number. (default: 7777)
-Q <Boolean>	Quiet mode. No progress output.
(default: false)	
-S <String>	Server hostname or IP-address of the
CLC Server.	
-U <String>	Valid username for logging on to the
CLC Server	
-V	Display version.
-W <String>	Clear text password or domain
specific password token.	
-d, --destination <ClcServerObjectUrl>	Destination for import from External
Application	
--in <ClcServerObjectUrl>	Model object(s) to be exported to
FASTA (.fa/.fsa/.fasta)	
Error: Missing required options: d, in	

12.11 Troubleshooting external applications

12.11.1 Checking the configuration

There is no check for the consistency of the configuration when it has been set up, so errors will only be seen on runtime when the application is executed. In order to help with troubleshooting, there are a few things that can be done:

Users of an external application launched via a CLC Workbench will see an error window if something goes wrong. Information in the Message tab may help to identify the issue. If not, try opening the **Advanced** tab, where the error message from the system should be visible.

Another aid when debugging is to import standard out and standard error as text. This will make it possible to check error messages posted by the external application (see figure 12.21).

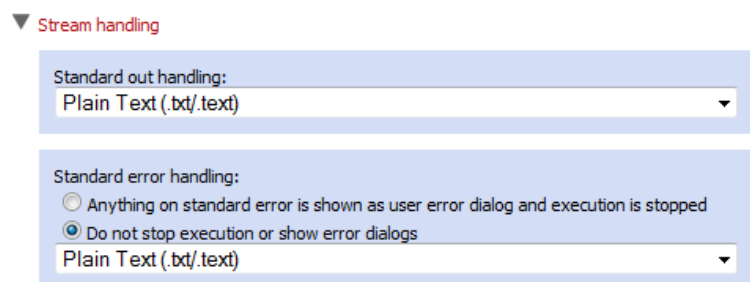


Figure 12.21: Importing the result of standard error and standard out.

Once the set-up is running and stable, you can deselect these options.

If your external application was previously working and then stops working, you can also see if the configuration has been recently changed, for example, by another administrator. Clues that this may have happened include

- The version number of the external application has changed. The version number is bumped each time the configuration is saved.

- When mousing over the name of the external application in the thin client, the timestamp of the most recent change to the configuration is presented beside the name of the user who made the change.

A "Change server configuration" operation is recorded in the audit log when changes to external application configurations are made. Clicking on the links in the log entry in the Data In column, you can see details related to the configuration change made.

12.11.2 Check your third party application

- Is your third party application being found? Perhaps try giving the full path to it.
- If you are using someone else's configuration file, make sure the location to the third party application is correct for your system.
- If you are using someone else's wrapper scripts, make sure all locations referred to inside the script are correct for your system.
- Is your third party application executable?
- If there was a wrapper script being used to call the third party application, is that wrapper script executable?

12.11.3 Is your Import/Export directory configured?

For certain setups, you need to have Import/Export directories configured. Please refer to section [12.3.2](#) for more details on this.

12.11.4 Check for conflicts in the naming of the external application

If your users will only access external applications via the Workbench, then you do not have to worry about what name you choose when setting up the configuration. However, if they plan to use the **clcserver** program, from the CLC Server Command Line Tools, to interact with your *CLC Server*, then please ensure that you do not use the same name as any of the *CLC Server* internal commands. You can get a list of these by running the `clcserver` command, with your *CLC Server* details, and using the flag with no argument. I.e. a command of the form:

```
clcserver -S <host> -P <port> -U <username> -W <password or token>
```

Chapter 13

Workflows

The *CLC Server* supports workflows created using *CLC Workbenches*. A workflow consists of a series of tools where the output of one tool is connected as the input to another tool. For a workflow to be executable on a *CLC Server*, all the tools in the workflow must be available on the server. General information about workflows can be found here: <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

13.1 Installing and configuring workflows

After a workflow is created in a *CLC Workbench*, an installer file can be created, as described on http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Creating_workflow_installation_file.html.

Workflow installer files can be used to install the workflow onto other *CLC Workbenches* or onto the *CLC Server*. When logged into the *CLC Server* from a *Workbench* as a user with administrative rights, workflows can also be installed onto the server directly from the *Workbench*.

When working logged into the *CLC Server* web interface as an administrative user, workflows can be installed by going to:

Admin (⚙️) | **Workflows** (📁)

Click the **Install Workflow** button and select a workflow installer file.

Once installed, a validated (✅) or attention (⚠️) status icon will be shown to the left of the workflow name, as shown in figure 13.1.

The workflow elements with red text are ones that can be configured. Click on such an element to bring up a dialog with a listing the parameters that can be configured, as well as an overview of the locked parameters. An example is shown in figure 13.2.

Open parameters can be configured, and can also be locked if desired, so that the parameter cannot be changed when executing the workflow. Locking and unlocking parameters is described further here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Locking_unlocking_parameters.html.

If changes to the parameter settings of an installed workflow are made, the timestamp of the most recent change and the name of the administrator who made those changes are reported at

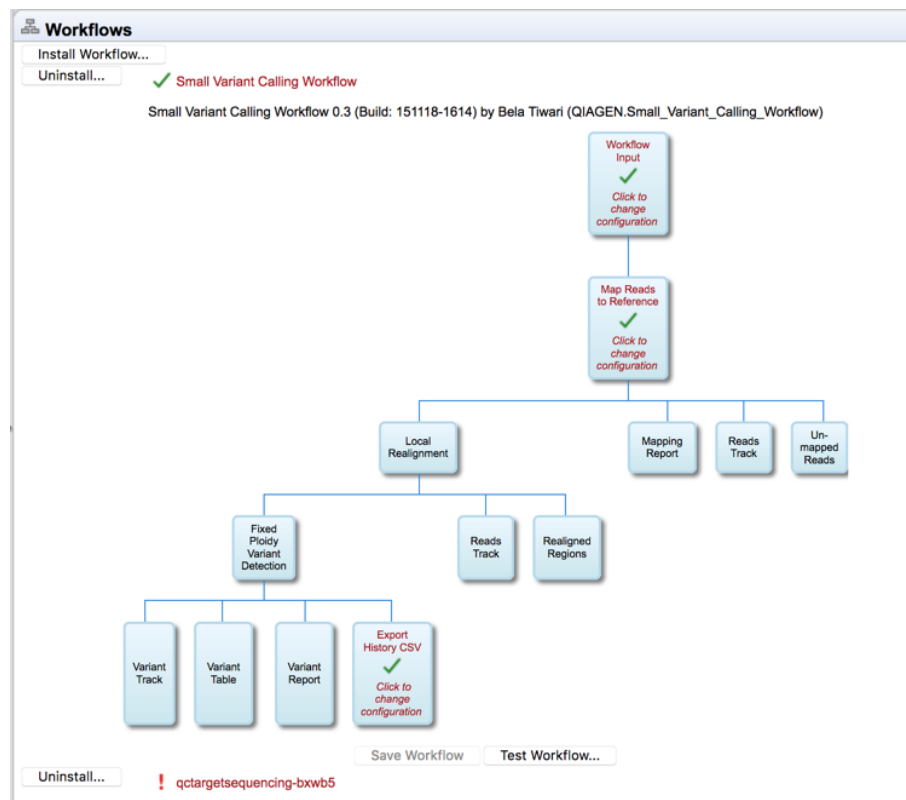


Figure 13.1: The workflow named Small Variant Calling Workflow is installed on the server and has been validated. The workflow below, named qctargetsequencing-bxwb5 is installed, but there are issues that need to be addressed.

Edit parameters for Map Reads to Reference – Fields marked with * are required

References*
Single genome track or reference sequences

Add / Remove References...

Settings locked by creator

Create report ☐

Auto-detect paired distances ☒

Masking track
Annotation track

Masking mode
[NO_MASKING, EXCLUDE, INCLUDE] No masking

Collect un-mapped reads ☐

Deletion cost
1 <= x <= 3 3

Non-specific match handling
[RANDOM, IGNORE] Map randomly

Color space alignment ☒

Insertion cost
1 <= x <= 3 3

Mismatch cost
1 <= x <= 3 2

Color error cost 2

OK

Figure 13.2: In this example only one parameter can be configured, the rest of the parameters are locked for the user.

the top of the workflow configuration view.

13.2 Executing workflows

Once a workflow is installed and validated, it is available for execution via client software. When you log in on the server using a CLC Workbench, workflows installed on the server automatically become available in the **Installed Workflows** folder of the **Toolbox** (see figure 13.3).

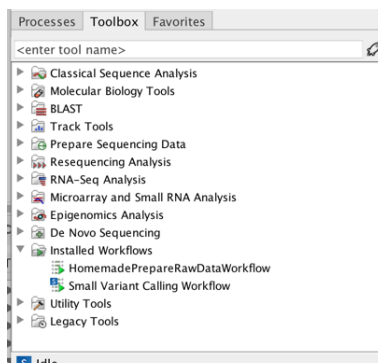


Figure 13.3: The Small Variant Calling Workflow is installed on the CLC Genomics Server, as signified by the blue S in the icon. The other installed workflow is installed locally, on the Workbench.

When you launch an installed workflow, you are presented with a dialog as shown in figure 13.4 with options of where to run it.

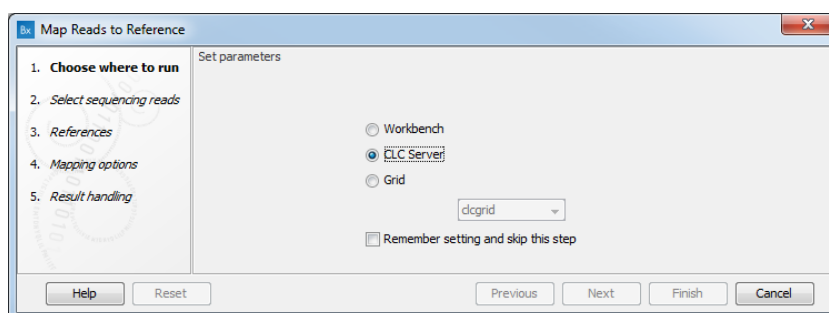


Figure 13.4: Selecting where to run the workflow.

Workflows installed on the CLC Server can be executed on the server or on a Workbench. Similarly, workflows installed on a Workbench can be executed there, or on a CLC Server. The only requirement is that the tools included in the workflow and any reference data configured for use by the workflow are available in the environment you wish to run the workflow on.

An important benefit of installing workflows on the CLC Server is that it provides the administrator with an easy way to update and deploy new versions of the workflow. Changes made centrally become immediately available for all Workbench users logged into the server.

13.3 Updating workflows

Tools included in a workflow are versioned. They will initially be the same version as in the software that was used to design the workflow. If one or more tools are updated through upgrading the CLC Server or plugins installed on it, then any workflow containing one or more such tools must be updated.

An exclamation mark (!) is presented beside any workflow that needs to be updated. Clicking

on the workflow name opens up a view where each element that needs to be updated is also indicated with an exclamation mark. See figure 13.5. Workflows that can be updated directly in the web administrative interface will have a button at the bottom labeled "Update Workflow" enabled. For workflows that cannot be updated this way, a message will appear stating this, and providing some tips of how to proceed. Details of both these situations are outlined below.

Updating workflows via the server web administrative interface

A button labeled "Update Workflow" just under the workflow is enabled for workflows that can be updated directly in the web administrative interface. Clicking on this button, or any of the elements with exclamation marks, starts the update. See figure 13.5.

Note! If a tool has been updated with a new parameter, then an updated workflow that includes that tool will have that new parameter configured with the default value.

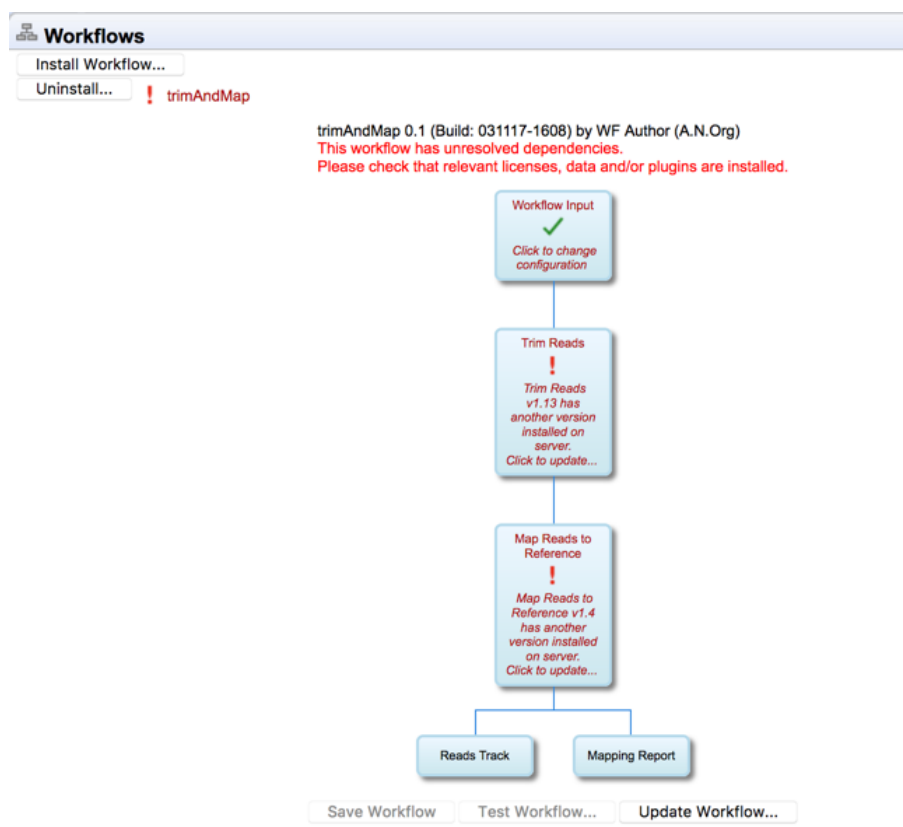


Figure 13.5: Click on the "Update Workflow" button, or on an element marked with an exclamation mark to start the update.

When updating, a window appears containing information about the changes to be enacted if you proceed. If errors have occurred these will also be displayed. See figure 13.6. Accept the changes by pressing the "Update" button. The update can also be canceled at this point if desired.

After pressing the "Update" button, the updated workflow will be marked with a green check mark (✓). A copy of the original workflow is also kept. It is disabled and has the original name with "-backup (disabled)" appended. An example is shown in figure 13.7.

If you click on the copy of the original workflow, a button labeled "Re-enable Workflow" appears (figure 13.8). Clicking on this button re-enables the original workflow and uninstalls the updated

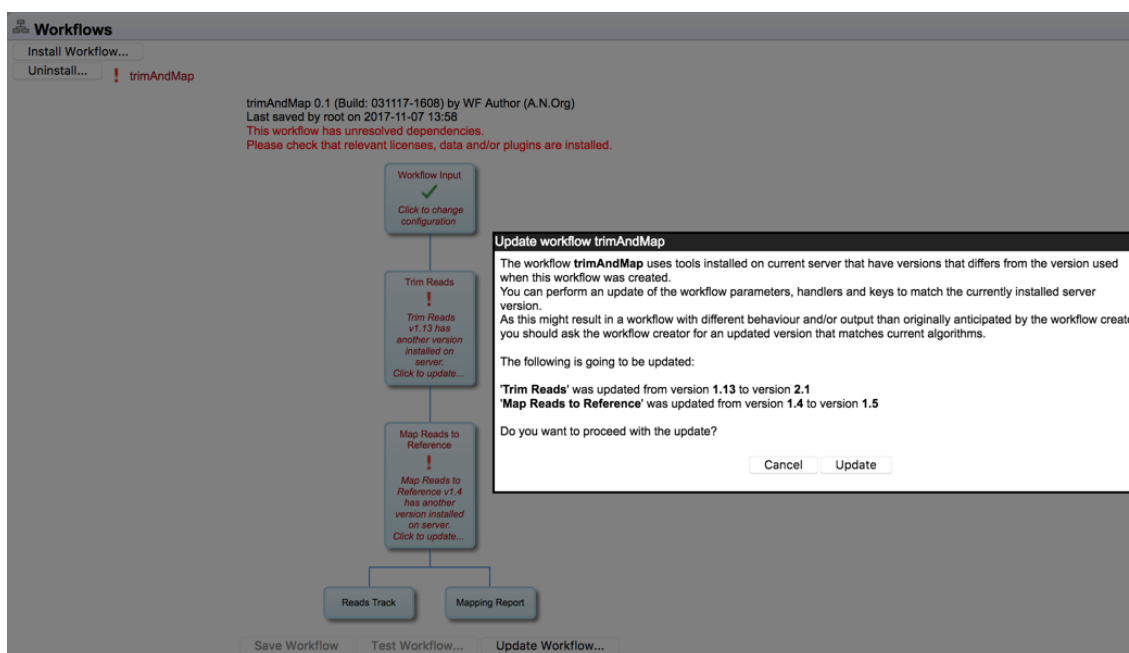


Figure 13.6: Details about the upgrade are presented, and you can choose whether to proceed with the update, or cancel it.

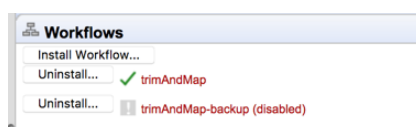


Figure 13.7: In addition to the updated version of the workflow, marked with a green check mark, a copy of the original workflow is kept. It is disabled and has the original name with "-backup (disabled)" appended.

version of the workflow.

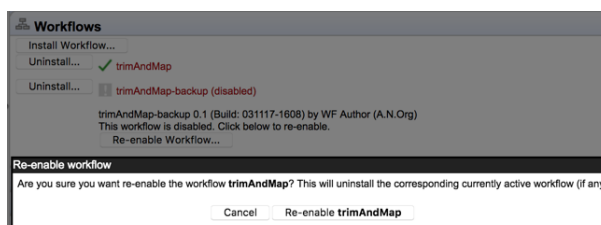


Figure 13.8: After a workflow has been updated, it is possible to re-enable the original workflow.

Updating workflows that cannot be updated via the server web administrative interface

Some installed workflows cannot be updated directly in the web administrative client. Common situations where this can occur include:

1. Workflows containing tools provided by plugins not installed on the CLC Server.
2. Workflows containing tools from server extensions (commercial plugins) that require a license, but either the license is not present or it does not support the version of the server extension that is installed.
3. Workflows containing tools not on the version of the CLC Server running.

4. Workflows containing tools that cannot be upgraded directly due to the nature of the changes made to them in the updated CLC Server.

To resolve the first 2 circumstances, check install any needed plugins and licenses, restart the CLC Server, and check the status of workflows under the **Workflows** tab of the web administrative interface.

To address the third and fourth issues, new versions of the workflows must be made on a CLC Workbench and then installed on the CLC Server. For this, a Workbench version that the installed workflow can be run from is needed, as well as the latest version of the Workbench.

To start, open a copy of the installed workflow in a version of the Workbench it can currently be run on. This is done by selecting the workflow in the **Installed Workflows** folder of the **Toolbox** in the bottom left side of the Workbench, then right-clicking on the workflow name and choosing the option "Open Copy of Workflow" (figure 13.9).

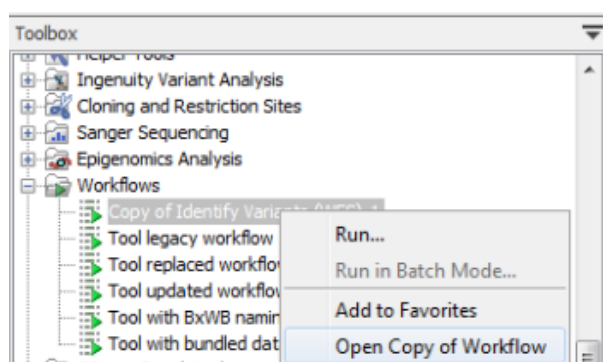


Figure 13.9: Open a copy of an installed workflow by right-clicking on its name in the Workbench Toolbox.

Save the copy of the workflow in the Navigation Area. To do this, you can simply drag and drop the tab to the location of your choice in the Navigation Area.

Open the new version of the Workbench and there, open the workflow that was saved in the Navigation Area. Click on the **OK** button if you are prompted to update the workflow.

You can now check that the workflow has been updated correctly, including that any reference data is configured as expected. Then save the updated version of the workflow. Finally, click the **Installation** button to install the workflow, if desired.

If the above process does not work when upgrading directly from a much older Workbench version, it may be necessary to upgrade step-wise by upgrading the workflow in sequentially higher major versions of the Workbench.

The updated workflow can now be installed on the CLC Server as described in section [13.1](#).

Chapter 14

Command line tools

CLC Server Command Line Tools is a command-line client for the *CLC Server*. Installation and basic usage information can be found in the *CLC Server Command Line Tools* manual. Links are provided below. The html version also contains extensive usage information for the commands that can be run.

- **html:** <http://resources.qiagenbioinformatics.com/manuals/clcservercommandlinetools/current/>
- **pdf:** http://resources.qiagenbioinformatics.com/manuals/clcservercommandlinetools/current/User_Manual.pdf

Chapter 15

Appendix

15.1 Use of multi-core computers

The tools listed below can make use of multi-core CPUs. This does not necessarily mean that all available CPU cores are used throughout the analysis, but that these tools benefit from running on computers with multiple CPU cores.

- Amino Acid Changes
- Annotate and Merge Counts
- Annotate from Known Variants
- Annotate with Conservation Scores
- Annotate with Exon Numbers
- Annotate with Flanking Sequences
- Basic Variant Detection
- BLAST (will not scale well on many cores)
- Call Methylation Levels
- Compare Sample Variant Tracks
- Copy Number Variant Detection
- Create Alignment
- Create RRBS-fragment Track
- De Novo Assembly
- Differential Expression
- Differential Expression in Two Groups
- Extract and Count
- Filter against Known Variants
- Filter based on Overlap
- Fixed Ploidy Variant Detection
- GO Enrichment Analysis

- Identify enriched Variants in Case vs Control Samples
- InDels and Structural Variants
- K-mer Based Tree Construction
- Link Variants to 3D Protein Structure
- Local Realignment
- Low Frequency Variant Detection
- Map Bisulfite Reads to Reference
- Map Reads to Contigs
- Map Reads to Reference
- Maximum Likelihood Phylogeny
- Merge Annotation Tracks
- Model Testing
- Predict Splice Site Effect
- QC for Read Mapping
- QC for Sequencing Reads
- QC for Targeted Sequencing
- Remove Variants Present in Control Reads
- Remove Marginal Variants
- Remove Reference Variants
- RNA-Seq Analysis
- Trim Sequences
- Trio Analysis

15.2 Troubleshooting

15.2.1 Check setup

To check your server is set up correctly, run the **check setup** tool. To do this:

- Log in on the web interface of the server as an administrator.
- Click the **check setup** link in the top right hand corner.
- Click on the **Generate Diagnostics Report** button in the window that appears.

When the report is ready, a list of tests performed is provided, as shown in figure 15.1.

Tests that passed are marked with a green check mark. Tests that failed are marked with a red X. Click on any of the tests listed to see more information about the test.

Additional notes:

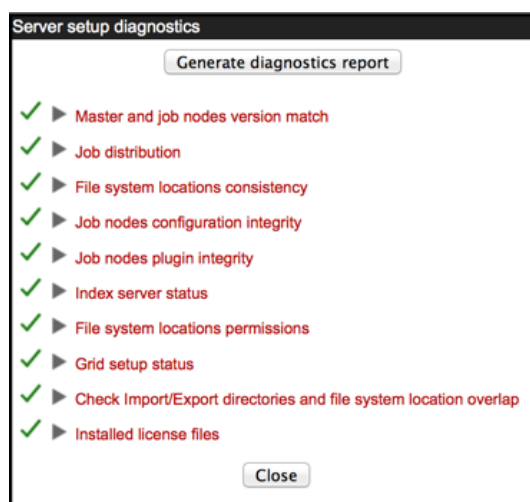


Figure 15.1: Check system. Tests that passed are marked with a green check mark. Tests that failed are marked with a red X.

- A green check mark is presented beside "List license files" when the contents of the "licenses" folder in the installation area of the CLC Server could be listed. Click on this item to see a list of the licenses found. The products and versions supported by the licenses found are reported. Information about expired licenses is also presented. See figure 15.2.
- A green check mark is presented beside "Grid setup status" in two cases:
 - You have configured a grid setup and it is configured correctly.
 - You have not configured a grid setup.



Figure 15.2: Click on the List license files item in the report to see the list of the license files found in the licenses subfolder of the installation area. Products and versions supported are reported, and any expired license is noted with red text.

15.2.2 Bug reporting

If there are problems regarding the installation and configuration of the server, please contact ts-bioinformatics@qiagen.com.

When contacting ts-bioinformatics@qiagen.com regarding problems on the CLC Server, you will often be asked for additional information about the server setup. You can easily send the necessary information by submitting a bug report:

Log in to the web interface of the server as administrator | report a bug (at the top right corner) | Enter relevant information with as much detail as possible | Submit Bug Report

You can see the bug report dialog in 15.3.

Figure 15.3: Submitting a bug report.

The bug report includes the following information:

- Log files
- A subset of the audit log showing the last events that happened on the server
- Configuration files of the server configuration

In a job node setup you can include the information from the job nodes by checking the **Include comprehensive job node info** checkbox in the **Advanced** part of the dialog.

If the server has access to the internet, you can **Submit Bug Report** to send the report to QIAGEN Bioinformatics Support. If the server does not have access to the internet, click on **Download bug report** to create a zip file containing the same information that you can attach to an email you send to ts-bioinformatics@qiagen.com from a machine connected to the network.

Note that the process of gathering the information for the bug report can take a while, especially for job node setups. If a Workbench user experiences a server-related error, it is also possible to submit a bug report from a Workbench error dialog if they are presented with one. The same archive is included as when submitting a bug report from the server web interface.

No password information is included in the bug report.

All data sent to ts-bioinformatics@qiagen.com is treated confidentially.

15.3 Database configurations

This section pertains to the configuration of an SQL database to store data imported into the *CLC Server*. This is relevant only if the database add-on has been purchased.

To use an SQL database for data management with the *CLC Server*, the appropriate JDBC driver for your database system must be installed. This is described in section [15.3.1](#).

For MySQL installations, please also refer to section [15.3.2](#) for information specific to MySQL database configuration.

15.3.1 Installing JDBC drivers

The general steps for installing JDBC drivers for use with CLC software are:

1. Download the appropriate JDBC driver for your database system from the provider. Further details about this are provided below.
2. Place the driver into the `userlib` directory in the installation area of the *CLC Server*.
 - For a CLC job node setup, the JDBC driver file must be placed into the `userlib` folder under the *CLC Server* installation area on **the master node and also on each of the job nodes**.
 - For a grid node setup, the driver file only needs to be placed into the `userlib` folder of the master *CLC Server* installation area.
3. Restart the *CLC Server* software.
 - For a job node setup, restart the master server and each of the job nodes.
 - For a grid setup, only the master CLC Server needs to be restarted. This will cause the changes to be deployed to the grid workers.

Getting JDBC drivers

Information on obtaining JDBC drivers for supported DBMS systems is provided in this section. After obtaining the driver, the general instructions given above should be followed to complete the driver installation.

MySQL JDBC Drivers

1. Go to <http://dev.mysql.com/downloads/connector/j/> to download the driver.
2. Choose the option **Platform Independent** when selecting a platform.
3. After clicking on the button to Download, you can login if you already have an Oracle Web account, or you can just click on the link that says `No thanks, just start my download` further down the page.
4. Uncompress the downloaded file and move the driver file, which will have a name of this form: `mysql-connector-java-X.X.XX-bin.jar`, to the folder called `userlib`, found in the installation area of the *CLC Server*.

PostgreSQL JDBC Drivers

1. Go to <https://jdbc.postgresql.org/download.html> and download the relevant driver.

2. Place it in the folder called `userlib`, found in the installation area of the *CLC Server*.

Microsoft SQL Server

1. Go to <https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server> and download the relevant driver.
2. Follow the installation instructions given for the driver to the point where the downloaded file has been unpacked.
3. Move the driver jar file compatible with java 1.8 to the folder called `userlib`, found in the installation area of the *CLC Server*.

Oracle JDBC Drivers

1. Go to <http://www.oracle.com/technetwork/database/application-development/jdbc/downloads/index.html>.
2. Select the relevant version for your Oracle database version.
You will need an Oracle account to download the driver.
3. Move the driver jar file to the folder called `userlib`, found in the installation area of the *CLC Server*.

15.3.2 Configuring MySQL

We recommend basing your MySQL configuration on the example configuration file `my-large.cnf` included in the MySQL distribution.

In addition the following changes should be made:

- Increase the value of the `max_allowed_packet` setting to support transfer of large binary objects to and from the database will be supported. We recommend this setting:
`max_allowed_packet = 64M`
- Ensure InnoDB is available and configured. This is necessary for the MySQL instance to work properly with the *CLC Server*.
- Enable the options in the InnoDB section of the configuration as outlined below:

```
# You can set .._buffer_pool_size up to 50 - 80 %
# of RAM but beware of setting memory usage too high
innodb_buffer_pool_size = 256M
innodb_additional_mem_pool_size = 20M
# Set .._log_file_size to 25 % of buffer pool size
innodb_log_file_size = 64M
innodb_log_buffer_size = 8M
innodb_flush_log_at_trx_commit = 1
innodb_lock_wait_timeout = 50
```

Additionally, there appears to be a bug in certain versions of MySQL that can cause the cleanup of the query cache to take a very long time. If you experience this, please disable the query log by setting the `query_cache_size` option to 0: `query_cache_size= 0`

15.4 SSL and encryption

The *CLC Server* supports SSL communication between the Server and its clients (i.e. Workbenches or the *CLC Server Command Line Tools*). This is particularly relevant if the server is accessible over the internet as well as on a local network.

The default configuration of the server does not use SSL.

15.4.1 Enabling SSL on the server

A **server certificate** is required before SSL can be enabled on the *CLC Server*. This is usually obtained from a *Certificate Authority* (CA) like Thawte or Verisign (see http://en.wikipedia.org/wiki/Certificate_authorities).

A **signed certificate** in a `pkcs12` keystore file is also needed. The keystore file is either provided by the CA or it can be generated from the private key used to request the certificate and the signed-certificate file from the CA (see section 15.4.1).

Copy the keystore file to the `conf` subdirectory of the *CLC Server* installation folder.

Next, the `server.xml` file in the `conf` subdirectory of the *CLC Server* installation folder has to be edited to enable SSL-connections. Add text like the following text to the `server.xml` file:

```
<Connector port="8443" protocol="HTTP/1.1" SSLEnabled="true"
           maxThreads="150" scheme="https" secure="true"
           clientAuth="false" sslProtocol="TLS"
           keystoreFile="conf/keystore.pkcs12" keystorePass="tomcat"
           keystoreType="PKCS12"
/>
```

Replace `keystore.pkcs12` with the name of your keystore file, and replace `tomcat` with the password for your keystore.

The above settings make SSL available on port 8443. The standard (non-SSL) port would still be 7777, or whatever port number you have configured it to.

Self-signed certificates can be generated if only connection encryption is needed. See http://www.akadia.com/services/ssh_test_certificate.html for further details.

Creating a PKCS12 keystore file

If the certificate is not supplied in a `pkcs12` keystore file, it can be put into one by combining the private key and the signed certificate obtained from the CA by using `openssl`:

```
openssl pkcs12 -export -out keystore.pkcs12 -inkey private.key -in certificate.crt -name "tomcat"
```

This will take the private key from the file `private.key` and the signed certificate from `certificate.crt` and generate a `pkcs12`-store in the `keystore.pkcs12` file.

15.4.2 Logging in using SSL from the Workbench

When the Workbench connects to the CLC Server it automatically detects if Secure Socket Layer (SSL) should be used on the port it is connecting to or not.

If SSL is detected, the server's certificate will be verified and a warning is displayed if the certificate is not signed by a recognized Certificate Authority (CA) as shown in figure 15.4.

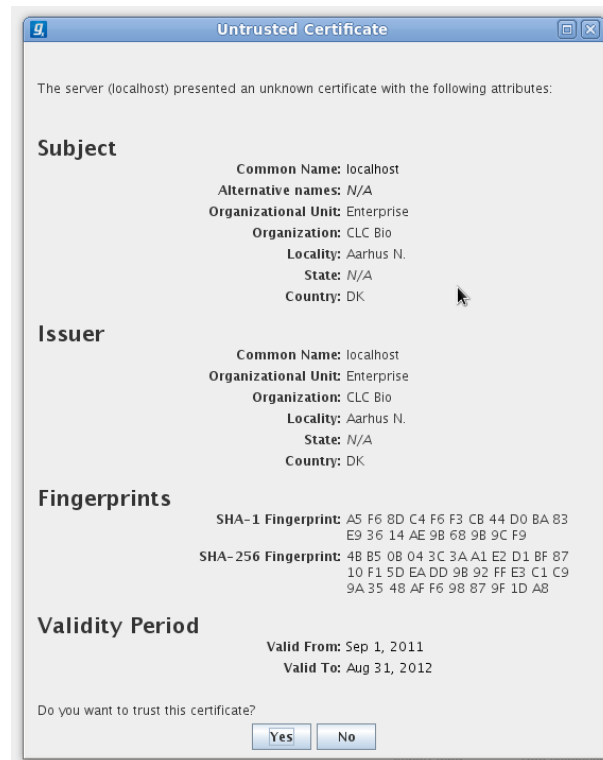


Figure 15.4: A warning is shown when the certificate is not signed by a recognized CA.

When such an "unknown" certificate has been accepted once, the warning will not appear again. It is necessary to log in again once the certificate has been accepted.

When logged into a server, information about the connection can be viewed by hovering the connection icon on the status-panel as shown in figure 15.5.

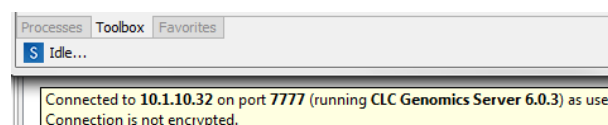


Figure 15.5: Showing details on the server connection by placing the mouse on the globe.

The icon is gray when the user is not logged in, and a pad lock is overlaid when the connection is encrypted via SSL.

15.4.3 Logging in using SSL from the CLC Server Command Line Tools

The CLC Server Command Line Tools will also automatically detect and use SSL if present on the port it connects to. If the certificate is untrusted the `clcserver` program will refuse to login:

```
./clcserver -S localhost -U root -W default -P 8443
```

```

Message: Trying to log into server
Error: SSL Handshake failed. Check certificate.
Option          Description
-----
-A <Command>    Command to run. If not specified the list of commands on the server will be returned.
-C <Integer>    Specify column width of help output.
-D <Boolean>    Enable debug mode (default: false)
-G <Grid Preset value> Specify to execute on grid.
-H             Display general help.
-I <Algorithm Command> Get information about an algorithm
-O <File>       Output file.
-P <Integer>    Server port number. (default: 7777)
-Q <Boolean>    Quiet mode. No progress output. (default: false)
-S <String>     Server hostname or IP-address of the CLC Server.
-U <String>     Valid username for logging on to the CLC Server
-V            Display version.
-W <String>     Clear text password or domain specific password token.

```

In order to trust the certificate the `clcsserversslstore` tool must be used:

```

./clcsserversslstore -S localhost -U root -W default -P 8443
The server (localhost) presented an untrusted certificate with the following attributes:
SUBJECT
=====
Common Name       : localhost
Alternative Names : N/A
Organizational Unit: Enterprise
Organization      : CLC Bio
Locality          : Aarhus N.
State             : N/A
Country           : DK

ISSUER
=====
Common Name       : localhost
Organizational Unit: Enterprise
Organization      : CLC Bio
Locality          : Aarhus N.
State             : N/A
Country           : DK

FINGERPRINTS
=====
SHA-1             : A5 F6 8D C4 F6 F3 CB 44 D0 BA 83 E9 36 14 AE 9B 68 9B 9C F9
SHA-256           : 4B B5 0B 04 3C 3A A1 E2 D1 BF 87 10 F1 5D EA DD 9B 92 FF E3 C1 C9 9A 35 48 AF F6 98 87 9F 1D A8

VALIDITY PERIOD
=====
Valid From        : Sep 1, 2011
Valid To          : Aug 31, 2012
Trust this certificate? [yn]

```

Once the certificate has been accepted, the `clcserver` program is allowed to connect to the server.

15.5 Non-exclusive Algorithms

Below is a list of algorithms which are non-exclusive, meaning that multiple jobs of these types can be run concurrently on a given node, be that a single server, job or grid node.

Algorithms marked as *Streaming* are I/O intensive and two streaming algorithms will not be run at the same time. When running on grid, *Streaming* algorithms are treated as exclusive, meaning that they will never run in conjunction with other algorithms (or themselves).

Algorithm	Streaming
Add attB Sites	
Amino Acid Changes	X
Annotate and Merge Counts	
Annotate from Known Variants	X
Annotate with Conservation Scores	X
Annotate with Exon Numbers	X
Annotate with Flanking Sequences	X

Algorithm	Streaming
Annotate with Nearby Gene Information	
Annotate with Overlap Information	X
Assemble Sequences	
Assemble Sequences to Reference	
BLAST at NCBI	
ChIP-Seq Analysis	
ChIP-Seq Analysis (legacy)	
Compare Sample Variant Tracks	X
Convert DNA To RNA	X
Convert from Tracks	X
Convert RNA to DNA	X
Convert to Tracks	X
Count-based statistical analysis	
Whole Genome Coverage Analysis	
Create Alignment	
Create BLAST Database	
QC for Read Mapping	
Create Entry Clone (BP)	
Create Expression Clone (LR)	
Create GC Content Graph Track	
Create Histogram	
Create Mapping Graph Tracks	
QC for Targeted Sequencing	
Create Track List	
Create Tree	
Demultiplex Reads	
Download 3D Protein Structure Database	X
Empirical Analysis of DGE	
Extract and Count	
Extract Annotations	
Extract Consensus Sequence	
Extract Reads	
Extract Sequences	X
Fasta High-Throughput Sequencing Import	X
Remove Variants Present in Control Reads	X
Filter against Known Variants	X
Filter Annotations on Name	X
Filter Based on Overlap	X
Remove Marginal Variants	X
Remove Reference Variants	X
Find Binding Sites and Create Fragments	
Find Open Reading Frames	
Identify enriched Variants in Case vs Control Samples	X
GO Enrichment Analysis	
Identify Graph Threshold Areas	
Illumina High-Throughput Sequencing Import	X
Import SAM/BAM Mapping Files	X

Algorithm	Streaming
Import Tracks from File	
InDels and Structural Variants	
Ion Torrent High-Throughput Sequencing Import	X
Link Variants to 3D Protein Structure	X
Merge Annotation Tracks	X
Merge Overlapping Pairs	
Merge Read Mappings	X
Motif Search	
Gaussian Statistical Analysis	
Predict Splice Site Effect	
Probabilistic Variant Detection	
Quality-based Variant Detection	
Reverse Complement Sequence	
Reverse Sequence	
Roche 454 High-Throughput Sequencing Import	X
Sanger High-Throughput Sequencing Import	X
Secondary Peak Calling	
Translate to Protein	
Trim Sequences	
TRIO analysis	

15.6 DRMAA libraries

Distributed Resource Management Application API (DRMAA) libraries are provided by third parties. Please refer to the distributions for instructions for compilation and installation, and contact the DRMAA providers if problems arise. QIAGEN Bioinformatics Support is not able to troubleshoot DRMAA library issues.

Information in this section of the manual is provided as a courtesy, but should not be considered a replacement for reading the documentation that accompanies the DRMAA distribution itself.

15.6.1 DRMAA for SLURM

Information about DRMAA for SLURM can be found at <https://slurm.schedmd.com/download.html>.

15.6.2 DRMAA for LSF

The source code for this library can be downloaded from <https://github.com/PlatformLSF/lsf-drmaa>. Please refer to the documentation that comes with the distribution for full instructions. Of particular note are the configure parameters "--with-lsf-inc" and "--with-lsf-lib" parameters, used to specify the path to LSF header files and libraries respectively.

15.6.3 DRMAA for PBS Pro

Source code for this library can be downloaded from <http://sourceforge.net/projects/pbspro-drmaa/>.

Please refer to the documentation that comes with the distribution for full instructions. Of particular note is the inclusion of the "--with-pbs=" parameter, used to specify the path to the PBS installation root. The configure script expects to be able to find lib/libpbs.a and include/pbs_ifl.h in the given root area, along with other files.

Please note that SSL is needed. The configure script expects that linking with "ssl" will work, thus libssl.so must be present in one of the system's library paths. On Red Hat and SUSE you will have to install openssl-devel packages to get that symlink (or create it yourself). The install procedure will install libdrmaa.so to the provided prefix (configure argument), which is the file the CLC Server needs to know about.

The PBS DRMAA library can be configured to work in various modes as described in the README file of the pbs-drmaa source code. We have experienced the best performance, when the CLC Server has access to the PBS log files and pbs-drmaa is configured with wait_thread 1.

15.6.4 DRMAA for OGE or SGE

OGE/SGE comes with a DRMAA library.

15.7 Consumable Resources

Setting up Consumable Resources with LSF

The following information was provided by IBM.

If you have questions or issues with setting up a consumable resource for LSF, please refer to your LSF documentation. For questions not covered there, please contact ruzhuchen@us.ibm.com and achristi@ca.ibm.com.

LSF has the ability to do "license scheduling" and ensure that CLC Server jobs running under LSF are only dispatched when there are available CLC Grid Worker licenses. When such scheduling is configured, CLC jobs for which no free licenses are available would stay in "pend" status, waiting for a CLC Grid Worker license to become available.

There are two parts to making use of this type of scheduling:

1. Configure the consumable resource in LSF.
2. Specify a clcbio license reservation when jobs are submitted to LSF.

Configuring the consumable resource in LSF

Add a consumable resource called clcbio in \$LSF_ENVDIR/lsf.shared:

```
Begin Resource
RESOURCENAME  TYPE   INTERVAL  INCREASING  DESCRIPTION # Keywords
    mips        Boolean  ()         ()          (MIPS architecture)
...
...
    clcbio      Numeric  ()         N           (clcbio license)
End Resource
```

Add the number of clcbio licenses in `$LSF_ENVDIR/lsf.cluster.<clustername>`:

```
Begin ResourceMap
  RESOURCENAME  LOCATION
  # CLCBIO license resource
  clcbio        (14@[all]) # 14 clcbio licenses can be used
  #....
End ResourceMap
```

This example shows a configuration for 14 CLC Grid Worker licenses, which means that up to 14 CLC jobs can be running on the LSF cluster at the same time. This integer needs to be changed to the number of licenses you own.

The configuration shown here assumes the CLC Grid Worker licenses can only be use in the LSF cluster as LSF will manage the free token count from the scheduling side.

In this context, LSF does not replace or directly talk to the LMX license server for CLC licenses. Rather, LSF manages the CLC Grid Worker license reservations internally.

Specify a clcbio license reservation when jobs are submitted to LSF

CLC jobs submitted to LSF need to have a clcbio license reservation specified.

This can be done in several different ways:

- via the CLC Grid Preset "Native Specification" field. (This is the most convenient method.) Simply add:

```
-R "rusage[clcbio=1]"
```

to this field.

- via the batch job submission command line
- using the RES_REQ line inside the lsb.queues file
- via an application profile (lsb.applications)

Important: After any LSF configuration file changes, one needs to reconfigure LSF for the changes to take effect. That is, run:

```
lsadmin reconfig
badmin reconfig
```

These are "safe" commands to run. That is, pending LSF jobs will continue to "pend" in status and running LSF jobs will continue to run.

15.8 Third party libraries

The CLC Server includes a number of third party libraries.

Please consult the files named NOTICE and LICENSE in the server installation directory for the legal notices and acknowledgements of use.

For the code found in this product that is subject to the Lesser General Public License (LGPL) you can receive a copy of the corresponding source code by sending a request to our support team at ts-bioinformatics@qiagen.com.

15.9 External network connections

Some functionality available in CLC software requires access to specific addresses on the Internet. A list of the internet sites accessed and a listing of the tools involved can be found at <https://secure.clcbio.com/helpspot/index.php?pg=kb.page&id=242>. The list of sites there can be referred to when configuring firewall settings for networks that utilize a whitelist approach. For CLC Servers with nodes, any nodes that need to execute functionality listed on that webpage will need access to the relevant sites.

15.9.1 Proxy settings

To add proxy settings to the CLC server, you will need to add lines to the server `.vmoptions` file. This file can be found in the installation area of the CLC server software. The name will reflect the specific product. For example, for the CLC Genomics Server, it would be called `CLCGenomicsServer.vmoptions`.

The following lines must be added to that file:

```
-Dhttp.proxyHost=  
-Dhttp.proxyPort=  
-Dhttp.nonProxyHosts="localhost|127.0.0.1|11...|.foo.com|etc"
```

where the `proxyHost` IP address and the `proxyPort` port number need to be added to complete the top two lines.

For job nodes, the settings described here must be applied to each node that will need to submit jobs that need access to the internet. For the `nonProxyHosts` values in this case, specify the names or IP addresses/subnets for *all* job nodes (if applicable with your server configuration) that should not be using the proxy service, if relevant.

For grid node setups, a `clcgridworker.vmoptions` file must be created in each deployed gridworker area, if such a file does not already exist and the settings described above added. See also section [6.3.10](#).

15.10 Monitoring

We recommend that you monitor the health and performance of the servers that the CLC Server software is running on and also monitor some key metrics of the CLC Server itself. This will

enable you to react quickly to any problems that occur and can aid in optimization of server performance.

With regards to the servers' physical resources, monitoring the amount of free memory and disk space is recommended, as consumption of these can be substantial depending on types and numbers of analyses being run.

Monitoring metrics of the *CLC Server* itself enables you to keep tabs on how well the jobs processing is going. The metrics the *CLC Server* provides are available as JMX¹ attributes. Software from a third party will be necessary to set up the monitoring of these attributes. Numerous software products for this are available and most support JMX. If your monitoring software supports the raising of alarms, you can set up triggers based on these metrics to receive alerts when a situation arises that needs attention.

15.10.1 Setting up JMX monitoring

No special configuration is necessary if monitoring will take place locally. However, JMX must be enabled for remote monitoring. This is done by adding a few settings to the server `vmoptions` file. The relevant `.vmoptions` file is located in the root of the server installation folder of a single server, or the root of the installation folder of the master server in the case of a node setup (Hereafter this folder will be referred to as `CLC_SERVER_BASE`).

Enable JMX with no security Add the following to the `.vmoptions` file:

```
-Dcom.sun.management.jmxremote.port=9999
-Dcom.sun.management.jmxremote.ssl=false
-Dcom.sun.management.jmxremote.authenticate=false
```

Enable JMX with authentication

1. Add the following to the `.vmoptions` file, instead of the lines provided in the section above.

```
-Dcom.sun.management.jmxremote.port=9999
-Dcom.sun.management.jmxremote.ssl=true
-Dcom.sun.management.jmxremote.authenticate=true
-Dcom.sun.management.jmxremote.password.file=../conf/jmxremote.password
-Dcom.sun.management.jmxremote.access.file=../conf/jmxremote.access
```

2. Create the access authorization file `CLC_SERVER_BASE/conf/jmxremote.access` and write the following in it:

```
monitorRole readonly
controlRole readwrite
```

3. Create the password file `CLC_SERVER_BASE/conf/jmxremote.password` and write the following in it:

¹https://en.wikipedia.org/wiki/Java_Management_Extensions

```
monitorRole clcserver  
controlRole clcserver
```

Further information about setting up JMX monitoring can be found in the Oracle guide on Monitoring and Management Using JMX Technology <http://docs.oracle.com/javase/8/docs/technotes/guides/management/agent.html>.

15.10.2 Completed process metrics

Object name: `com.clcbio.server:type=CompletedProcesses`

The "completed process" metrics can be used to evaluate the processes that are complete either because they were successful or because they failed. Canceled processes are ignored. The CLC server provides a temporal view of the latest completed processes, with two different temporal views available: time frame view or a history view that includes a fixed number of the most recently completed processes.

The size of the time frame to view and the number of entries for the history view can be configured directly through JMX or by editing the `Monitoring.properties` file. This properties file is located in this folder: `CLC_SERVER_BASE/settings/`. Please change the default settings to values that fit your specific setup. Lower values will generally result in a more reactive monitoring solution, but values that are too low may result lead to false alarms.

The following is a list of all the process related metrics that are available:

Number of processes in history

Attribute name: `NumberOfProcessesInHistory`

The number of processes currently in the history. Successful and failed processes are included. Canceled processes are not included. The maximum size of the history can be set using the `SizeOfHistory` attribute, available through JMX and the configuration file.

Number of failed processes in history

Attribute name: `NumberOfFailedProcessesInHistory`

The number of failed processes currently in the history.

Fraction of failed processes in history

Attribute name: `FractionOfFailedProcessesInHistory`

The fraction of the processes in the history that have failed. This fraction is not of much value if the number of processes in the history is very low.

Number of processes within time frame

Attribute name: `NumberOfProcessesWithinTimeFrame`

The number of processes that have been completed between now and a variable number of milliseconds earlier. Both successful and failed processes are included. Canceled processes

are not included. The time frame can be set using the `TimeFrameInMilliseconds` attribute, that is available through JMX and the configuration file.

Number of failed processes within time frame

Attribute name: `NumberOfFailedProcessesWithinTimeFrame`

The number of failed processes that have been completed between now and a variable number of milliseconds earlier.

Fraction of failed processes within time frame

Attribute name: `FractionOfFailedProcessesWithinTimeFrame`

The fraction of failed processes compared to the total number of processes that have been completed between now and a variable number of milliseconds from now. This fraction is not of much value if the number of processes in the time frame is very low.

15.10.3 Process execution metrics

Object name: `com.clcbio.server:type=ProcessExecution`

Process execution metrics allow measurement of how many jobs are being processed and how many are in a queue because they are waiting for available processing resources.

On job node setups, the object names are available on all nodes, but it only makes sense to monitor them on the master node since this is where the jobs are managed. If a grid is used to process the jobs, the actual queue will be a part of the grid system, which results in the processes being moved almost instantly to the "currently processing" state and the *CLC Server* queue itself is then empty.

Currently processing

Attribute name: `CurrentlyProcessing`

Number of processes being executed at the moment.

Waiting for resources

Attribute name: `WaitingForResources`

On a job node setup, the number of processes that are queued and are waiting for a node to be available for processing. This does not include processes that are waiting for output from another process. It includes only processes that have the input they need and are ready to be processed, but where no resources are available at that moment.

15.10.4 Job node metrics

Object name: `com.clcbio.server:type=JobNodes`

`com.clcbio.server:type=JobNodes,name=<job node name>`

With the job node metrics you can monitor a master node's connection with its job nodes. The object name is available on all nodes, but it only makes sense to monitor this on the master node. There are two sets of attributes to monitor. One set provides an aggregated view of all the job nodes while the other provides individual attributes for each job node.

Communication errors are only reported if the server uses the General queue. If the High throughput queue is used, the master node never contacts the job nodes on its own initiative and therefore there is no support for monitoring the job nodes through JMX in the current version of the server.

Apart from communication error attributes, the set of individual attributes also includes information about the host and port of a given job node. This information is not meant for monitoring as such, but is included for convenience, as when an error does arise identification of the job node involved is usually necessary.

Communication failed

Attribute name: `CommunicationFailed`

This attribute is set to true if the master node is currently having problems communicating with this particular job node.

Seconds since communication failed

Attribute name: `SecondsSinceCommunicationFailed`

The number of seconds since the communication with the job node first failed. As soon as the master node succeeds in connecting with the job node again, this value returns to zero. This attribute can be useful if you want to avoid reacting to very short fallouts in communication.

Max seconds since communication failed

Attribute name: `MaxSecondsSinceCommunicationFailed`

The maximum number of seconds since communication with any of the job nodes first failed. The advantage of this metric is that monitoring of it can be set up once and does not need to be changed if a job node is attached or detached.

Number of job nodes

Attribute name: `NumberOfJobNodes`

The number of job nodes currently attached to the master server.

Number of failed job nodes

Attribute name: `NumberOfFailedJobNodes`

The number of job nodes the master server currently has problems communicating with.

Bibliography

- [Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25.
- [Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829.