Biomedical Genomics Analysis

Plugin

USER MANUAL

User manual for

Biomedical Genomics Analysis 25.1

Windows, macOS and Linux

April 4, 2025

This software is for research purposes only.

QIAGEN Aarhus AS
Kalkværksvej 5, 11.
DK - 8000 Aarhus C
Denmark

Contents

Part I

Introduction

Chapter 1

Introduction

Contents

Welcome to Biomedical Genomics Analysis 25.1 – a software package supporting your daily bioinformatics work.

1.1 The concept of Biomedical Genomics Analysis

The Biomedical Genomics Analysis plugin has primarily been developed for use in cancer and disease research to analyze next generation sequencing (NGS) data. The Biomedical Genomics Analysis plugin provides a variety of specialized tools, reference data for human and model species and a comprehensive collection of template workflows that cover all steps from the initial data processing and quality assurance through data analyses, annotation, and reporting.

Template workflows are provided for the following applications:

- SARS-CoV-2
- QIAseq Sample Analysis
- Whole Genome Sequencing
- TSO Panel Analysis
- Whole Exome Sequencing
- Targeted Amplicon Sequencing
- Whole Transcriptome Sequencing
- Small RNA Sequencing

For ease of use, these workflows are configured to use reference data available to download using the *CLC Genomics Workbench*. See section 3.1.

The Biomedical Genomics Analysis plugin is frequently updated. A detailed list of new features, improvements, bug fixes, and changes is available at https://digitalinsights.qiagen.com/biomedical-genomics-analysis-latest-improvements/.

1.2 System requirements

In addition to meeting the system requirements of the ***CLC Genomics Workbench*** or the ***CLC Genomics Server***, the following requirements must be met:

- 32 GB RAM recommended (16 GB RAM required).
- At least 100 GB free disk space in the temporary directory of the ***CLC Genomics Workbench*** or the ***CLC Genomics Server***.
- We recommend analyzing data generated with the following QIAseq kits and panels on a *CLC Genomics Server*:
  - QIAseq Human Exome Kits
  - QIAseq Multimodal Pan-Cancer Panel
  - QAIseq Multimodal DNA/RNA Library Kit
  - QIAseq Tumor Mutational Burden Panels
  - QIAseq Targeted DNA Human TMB and MSI Panel

Compatibility

The Biomedical Genomics Analysis 25.1 plugin and the Biomedical Genomics Analysis Server Plugin 25.1 can be installed on *CLC Genomics Workbench* 25.0 and *CLC Genomics Server* 25.0, respectively, and on later versions in the same major release line.

1.3 Contact information

Biomedical Genomics Analysis is developed by:

QIAGEN Aarhus A/S
Kalkværksvej 5, 11.
DK - 8000 Aarhus C
Denmark

https://digitalinsights.qiagen.com/

Email: ts-bioinformatics@qiagen.com

The QIAGEN Aarhus team continuously improves products with your interests in mind. We welcome feedback and suggestions for new features or improvements. How to contact us is described at: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Contact_information_citation.html

You can also make use of our online documentation resources, including:

- Core product manuals https://digitalinsights.qiagen.com/technical-support/manuals/
- Plugin manuals https://digitalinsights.qiagen.com/products-overview/plugins/
- Tutorials https://digitalinsights.qiagen.com/support/tutorials/
- Frequently Asked Questions https://qiagen.my.salesforce-sites.com/KnowledgeBase/KnowledgeNavigatorPage

Chapter 2

Getting started

Data import and export

Most import and export of data is done using the tools delivered as part of the CLC Genomics Workbench. This functionality is described in the *CLC Genomics Workbench* manual:

- Import: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Importing_data.html
- Export: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Exporting_data_graphics.html

Prepare raw data

The first thing to do after data import is to check the quality of the sequencing reads and perform the necessary trimming. Note that the CLC Workbench is able to trim automatically read-through adapters, but if you are not sure you have read-through reads, you will need to provide a Trim Adapter List. To learn how to create an adapter trim list, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_adapter_list.html.

A workflow is available for preparing the raw reads in the QIAGEN CLC Genomics Workbench, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Prepare_Raw_Data.html.

Batching workflows

Batch processing refers to running an analysis multiple times, using different inputs for each analysis run. For example, if you have 10 sequence lists and wish to run 10 mapping analyses, one per sequence list, then these 10 analyses could be launched by setting up one batch job. When a job is run in batch mode, parameter settings stay the same for each run. It is just the inputs that are changed.

See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html for details.

For launching a workflow with more than one input, where the contents of more than one

input should change in each batch, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batching_workflows_with_more_than_one_input_changing_per_run.html.

See section 14.4 for an example of how to launch a workflow in batch mode.

Chapter 3

Reference data management

Template workflows delivered by the Biomedical Genomics Analysis plugin are configured to use QIAGEN Reference Sets, making them simple to launch while helping ensure that the same reference data is used consistently. This reference data can be easily obtained using the Reference Data Manager in the *CLC Genomics Workbench*. Reference data for a specific workflow can also be downloaded via workflow launch wizards. These features are described in section 3.1.

3.1 QIAGEN Sets

QIAGEN provides access to much common reference data using functionality under the **QIAGEN Sets** tab of the Reference Data Manager. Data is distributed as **Reference Data Elements**, which can be individually downloaded, or downloaded as part of a **Reference Data Set**. Many template workflows are configured to make use of QIAGEN Reference Sets, making them simple to launch while helping to ensure that the same reference data is used consistently. Using such workflows, the relevant reference data can also be downloaded via the workflow launch wizard (figure 3.1). When logged into a CLC Server with a CLC_References location defined, you can choose whether to download the data to the Workbench or Server.

Using the Reference Data Manager for QIAGEN reference data

To access **QIAGEN Sets**, open the Reference Data Manager by clicking on the **Manage Reference Data ( )** button in the top Toolbar or go to the **Utilities** menu and select **Manage Reference Data ( )**. Then click on the **QIAGEN Sets** tab at the top left. Under this tab, there are subsections for **Reference Data Sets** and **Reference Data Elements** (figure 3.2).

When a Reference Data Set is selected, information about it is displayed in the right hand pane. This includes the size of the whole data set, and a table listing the workflow roles defined in the set, with information about the data element specified for each role. Further details about the element assigned to a role can be found by clicking on the link in the Version column. An icon to the left of each set indicates whether data for this set has already been downloaded ( ) or not ( ). The same icons are used to indicate the status of each element in a Reference Data Set (figure 3.3).

If you have permission to delete downloaded data, the **Delete** button will be enabled. When reference data is stored on a *CLC Server*, you need be logged in from the Workbench as an

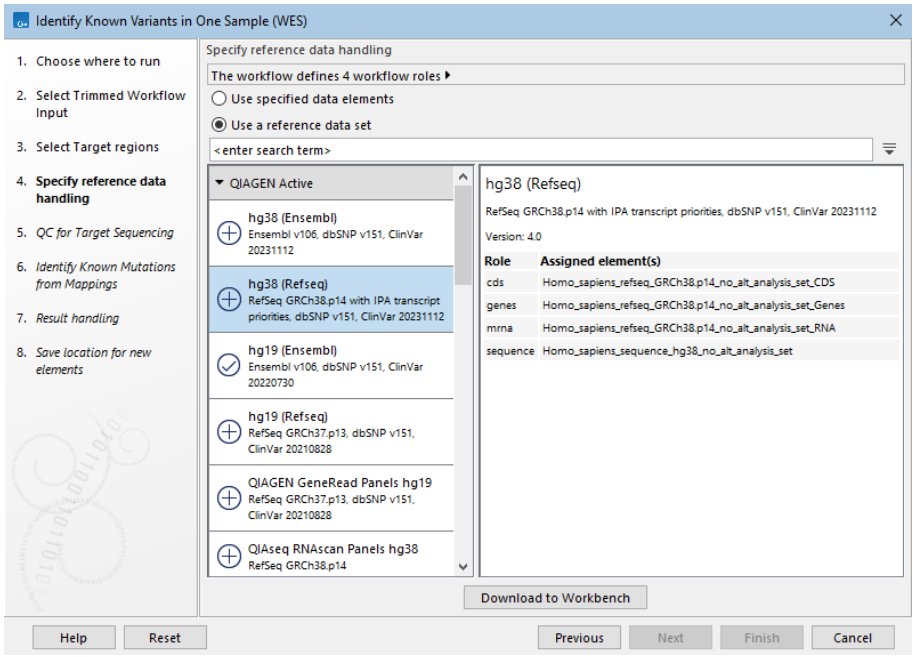


Figure 3.1: *When launching workflows configured to use data from Reference Data Sets, the relevant reference data can be downloaded via the workflow launch wizard.*

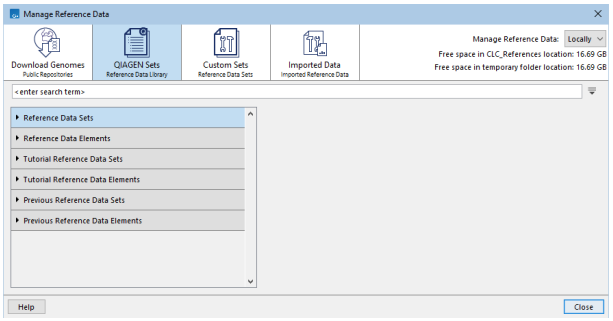


Figure 3.2: *Subheadings under the QIAGEN Sets tab provide access to Reference Data Sets and Reference Data Elements*

administrative user to delete reference data.

Searching for data available under the QIAGEN Sets tab

Use the search field under the top toolbar to search for terms in element and set names, workflow role names, and versions. To search for just an exact term, put the term in quotes.

The results include the name of the element or set the term was found in, followed in brackets by the tab it is listed under, e.g. (Reference Data Elements), (Tutorial Reference Data Sets), etc. Hover the cursor over a hit to see what aspect of the result matched the search term (figure 3.4). Double-click on a search result to open it.

Downloading resources

To download a Reference Data Element or a Reference Data Set (i.e. all elements in that set),

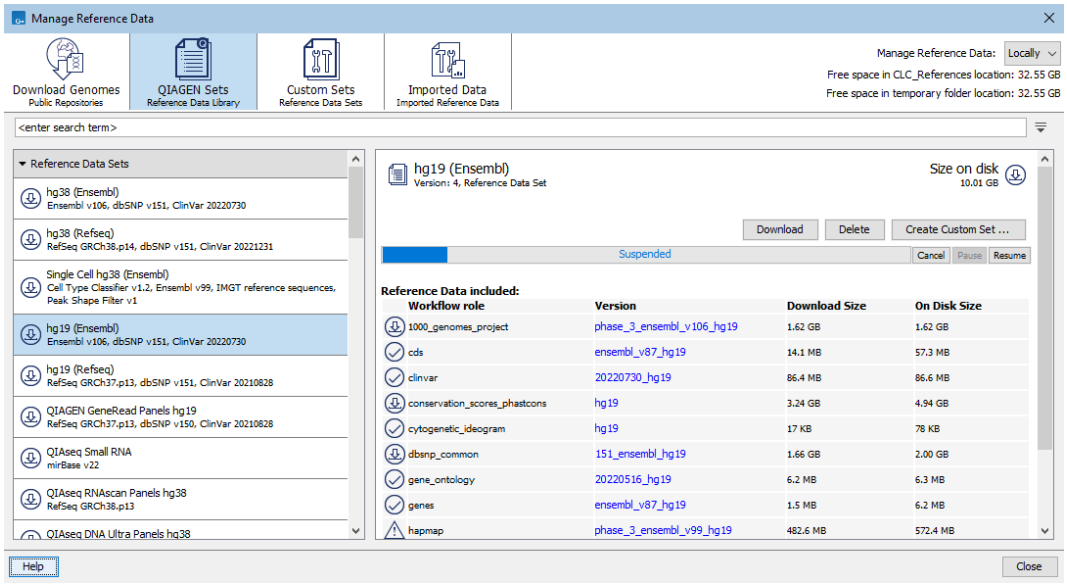


Figure 3.3: *The elements in a Reference Data Set are being downloaded. The full size of the data set is shown at the top, right hand side. The size of each element is reported in the "On Disk Size" column. Below the row of tabs at the top is a search field that can be used to search for data sets or elements.*

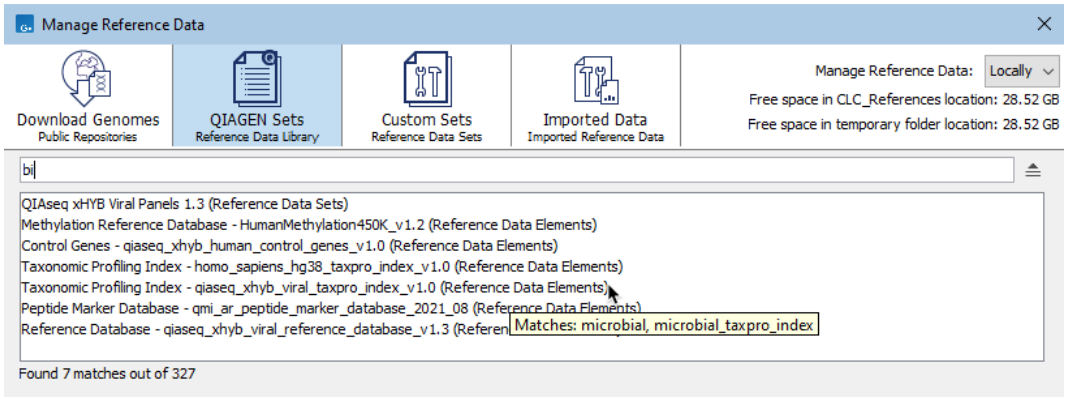


Figure 3.4: *Terms entered in the search field when the QIAGEN Sets tab is selected are searched for in element and set names, workflow role names, and versions of the resources available under that tab. Hovering the cursor over a hit reveals a tooltip with information about the match.*

select it and click on the **Download** button.

The progress of the download is indicated and you have the option to **Cancel**, **Pause** or **Resume** the download (figure 3.3).

When the "Manage Reference Data" option at the top of the Reference Data Manager is set to "Locally", data is downloaded to the CLC_References location in the *CLC Workbench*. When set to "On Server", the data is downloaded to the CLC_References location in the *CLC Server*.

Additional information

The HapMap (https://www.sanger.ac.uk/data/hapmap-3/) databases contain more

than one file. QIAGEN Reference Data Sets that include HapMap are initially configured with all the populations available. You can specify specific populations to use when launching a workflow, or you can create a custom reference set that contains only the populations of interest.

General information about Reference Data Sets, and creating Custom Sets, can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

General information about the Reference Data Manager is at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=References_management.html.

Part II

Biomedical Genomics Analysis Tools

Chapter 4

UMI tools

Contents

Most QIAseq library kits use Unique Molecular Indexes (UMIs) to improve performance. UMIs can be used to identify and correct sequencing errors to allow higher sensitivity in variant calling. UMIs can also be used to eliminate library amplification and sequencing biases in RNA quantification.

4.1 Remove and Annotate with Unique Molecular Index

During library preparation of the samples it is possible to add single or duplex UMI sequences to the reads, which are used towards correcting for sequencing errors and to help improve performance. Addition of UMI is often accompanied by a common sequence prefix that is also added before amplification and which can be very helpful when locating the exact UMI sequence. While the UMI is essential in identifying reads that originate from the same fragment, retaining it as such on the sequenced reads would hinder the subsequent read mapping efficiency and accuracy. Therefore, the **Remove and Annotate with Unique Molecular Index** tool removes the UMI and the common sequence prefix from the reads, while annotating each read with the UMI to retain the fragment identity as annotation.

Remove and Annotate with Unique Molecular Index is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Remove and Annotate with Unique Molecular Index** ()

In the first dialog, select sequence list(s) () containing the reads.

In the Settings dialog (figure 4.1), the following options are available:

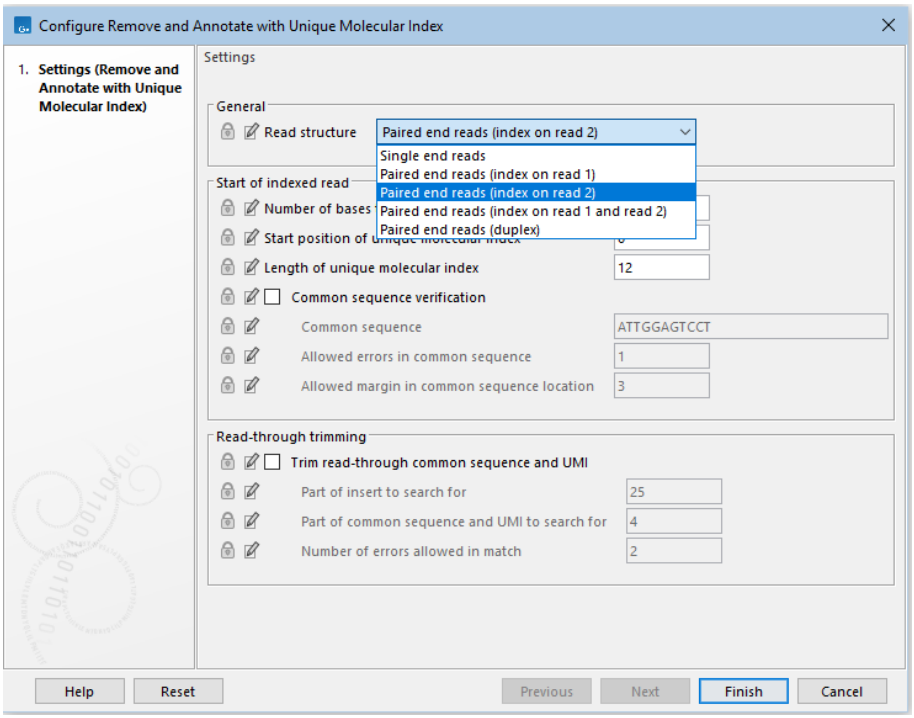


Figure 4.1: *Settings.*

- **Read structure** Specify the read type and, for paired data, which read the UMI is present on or if there are duplex UMI. The duplex option assumes the TruSight Oncology, Illumina (TSO) UMI protocol was used.
- **Number of bases to remove** The length of the UMI, or duplex UMI, plus the length of the common sequence. The default value is 8 for duplex UMI and 23 for other cases.
- **Start position of Unique Molecular Index** The initial position of the UMI on the read. The default is 0, indicating that the UMI is at right at the start of the reads.
- **Length of Unique Molecular Index** The length of the UMI. The default value is 7 for duplex UMI and 12 for other cases.
- **Common sequence verification** When enabled, a common sequence to be searched for, next to the UMI on each read, can be specified. By default, this option is unchecked, and the tool will output the same number of reads as were present in the input, but with UMI and common sequences trimmed away. When checked, only reads containing the specified common sequence, within the given error margins, will be retained. The following fields must be filled in when using this option:
  - **Common sequence** The common sequence to be searched for.
  - **Allowed errors in common sequence** The number of insertion/deletion/mismatches allowed between the common sequence defined, and the read sequence, for that read to be retained.
  - **Allowed margin in common sequence location** The number of base positions that the actual location of the common sequence can differ from its intended location.

  This option cannot be used with reads generated with an Illumina MiSeq sequencer.

- **Trim read through common sequence and UMI**: If this option is enabled, then for each read pair, first a sequence is extracted from the indexed read consisting of a part of the insert sequence and a part of the adjacent common sequence and UMI. Then, the reverse complement of this sequence is used to search the non-indexed read of a read pair, and if a match is found, the non-indexed read will be trimmed at the boundary between the insert and the common sequence.
  - **Part of insert to search for** Number of nucleotides from the sample sequence insert used to identify read-through. Increase this value to get more specific matches, decrease it if the indexed reads are very short, or to improve speed.
  - **Part of common sequence and UMI to search for** Number of nucleotides from the common sequence and UMI used to identify read-through. Increase this value to get more specific matches and avoid truncation at repetitive instances, decrease it to trim off shorter partial occurrences of common sequence and UMI.
  - **Number of errors allowed in match** Number of insertion, deletion, or mismatch errors allowed when looking for read-through sequences.

A report can be generated that contains information about the number of reads processed, and the number and fraction of reads found to have UMIs. It also includes a plot of the nucleotide distribution per position of the UMI barcode.

4.2 Calculate Unique Molecular Index Groups

The **Calculate Unique Molecular Index Groups** tool annotates the mapped reads with a "Unique Molecular Index group ID", that is identical for reads that are determined to belong to the same UMI.

Calculate Unique Molecular Index Groups is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Calculate Unique Molecular Index Groups** ()

In the first dialog (figure 4.2), select a read mapping of reads that were previously annotated with UMI annotations.

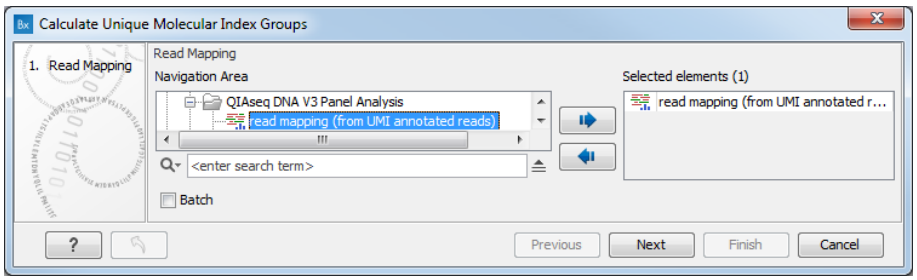


Figure 4.2: *Select a read mapping made from reads whose UMI was removed and annotated on the sequences.*

The grouping of reads into UMI groups works as follows:

1. The tool groups reads that

- start at the same position based on the end of the read to which the UMI is ligated. This can either be defined in the Remove and Annotate with Unique Molecular Index tool or directly in the wizard, (If the UMI was removed from the start of read 2 using the Remove and Annotate with Unique Molecular Index tool, this tool considers grouping reads where the start of read 2 map to the same position)
- are from the same strand, and
- have identical UMIs.

The tool then merges smaller groups into larger groups if

2. Their start positions are sufficiently close as defined by the *Window size* parameter.
3. Their UMIs are similar enough as defined by the *Fuzzy match Unique Molecular Indices* parameter.

   Merging is only done if the larger group is sufficiently large compared to the smaller group as defined by the parameters described below. If a smaller group can be merged into multiple larger groups that are equally good in terms of similarity of UMI and start position as well as group size, the group will not be merged.

Duplex groups are created if input reads were defined as duplex data in the Remove and Annotate with Unique Molecular Index tool, or if the UMI location setting has been set to duplex. Duplex groups consist of two paired end UMI groups from different strands of the original fragment. Two UMI groups, A and B, are grouped to a duplex group when:

1. Both are paired reads.
2. One must consist of forward paired end reads (referred to as group A) and the other must consist of reverse paired end reads (referred to as group B)
3. The genomic positions of read 1 in group A and read 2 in group B are the same, and the genomic positions of read 2 in group A and read 1 in group B are the same.
4. The UMI is the same for read 1 of group A and read 2 of group B, and the UMI is the same for read 2 of group A and read 1 of group B.

It is possible to change the following parameters (figure 4.3):

- **UMI location** Define if the UMI was sequenced as part of read 1 or read 2. The UMI location determines if the R1 or R2 start position must be the same for reads to be grouped.
  - **Defined in Remove and Annotate with Unique Molecular Index** Groups reads based on the mapped genomic position of the start of the read that originally had the UMI as defined in Remove and Annotate with Unique Molecular Index.
  - **Read 1** Groups reads based on the mapped genomic position of the start of read 1. Use if the UMI sequence was annotated on the reads during import, and the UMI was originally sequenced as part of read 1. Read about annotating with UMIs during import here: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=General_notes_on_UMIs.html

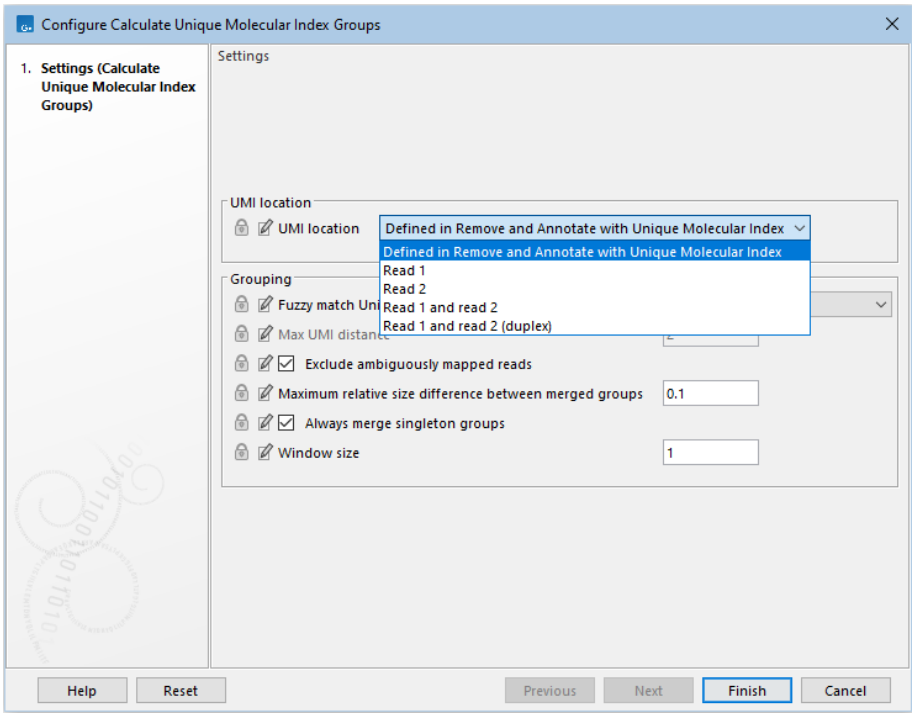


Figure 4.3: *Select a read mapping made from reads whose UMI was removed and annotated on the sequences.*

  - **Read 2** Groups reads based on the mapped genomic position of the start of read 2. Use if the UMI sequence was annotated on the reads during import, and the UMI was originally sequenced as part of read 2. Read about annotating with UMIs during import here: `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=General_notes_on_UMIs.html`
  - **Read 1 and read 2** Groups reads based on the mapped genomic positions of the start of read 1 and read 2. Does not calculate duplex groups.
  - **Read 1 and read 2 (duplex)** Groups reads based on the mapped genomic positions of the start of read 1 and read 2. And also calculates duplex groups.

- **Fuzzy match Unique Molecular Indices** Method for deciding which UMIs are considered similar enough for merging:
  - **Do not fuzzy match** Groups will be merged only if they match exactly.
  - **Allow one mismatch** Groups will be merged if they are at most one mismatch apart.
  - **Allow one mismatch/deletion/insertion** Groups will be merged if they are at most one mismatch, deletion or insertion apart.
  - **Distance** Groups will be merged if their edit distance (also called Levenshtein distance), is smaller than the value given in **Max UMI distance**. Note that if the distance is greater than 1, the groups also have to satisfy a stricter requirement for ratio between their sizes.

- **Max UMI distance** The maximum edit distance allowed if Distance is selected as Fuzzy match Unique Molecular Indices.

- **Exclude ambiguously mapped reads** is checked by default.
- **Maximum relative size difference between merged groups** will merge small groups into bigger ones if the size ratio between the two groups is smaller than a certain value (set at 0.1 per default). We define the distance between two groups as the number of differences in their UMIs (which can only be greater than one if "Distance" is chosen as Fuzzy match Unique Molecular Indices) plus one if their start positions are not the same. The size ratio parameter is taken to the power of the distance, i.e. if the distance is two, the smaller group size should be at most $0.1^2 = 0.01$ the size of the larger group.
- **Always merge singleton groups** When this option is checked, a singleton UMI group, a group that contain only one read, is merged with a non-singleton group with distance 1 even if the "Maximum relative size difference between merged groups" threshold is not met.
- **Window size** Groups will be merged if the difference between their start positions are less than this.

Click **Next** to choose whether to **Open** or **Save** the resulting read mapping of reads which now have a "UMI group ID" annotation.

A report can also be generated. It contains:

- A summary table with the following information:
  - Reads in input: Reads that were aligned to the reference
  - Reads mapped multiple places (discarded): Reads that aligned to the reference in multiple places, and thus discarded
  - Groups merged
  - Groups not merged due to >1 candidate of equal size
- Group size table and plots described in section 4.6.
- Duplex summary and plots if calculating duplex groups.

Note: When the group sizes (the number of reads in UMI groups) are very large (in most cases more than 10 reads in a UMI group is not desirable), this can indicate problems, such as quality issues with the sample. It can also indicate that the sequencing depth could be reduced.

4.3 Create UMI Reads from Grouped Reads

The tool **Create UMI Reads from Grouped Reads** generates a single consensus read, called a UMI read, from reads which belong to the same group, as determined by the Calculate Unique Molecular Index Groups tool. The consensus reads are placed in a read mapping at the location of the original reads. Therefore, the output of the tool is a read mapping of generated UMI reads.

Create UMI Reads from Grouped Reads is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Create UMI Reads from Grouped Reads** ()

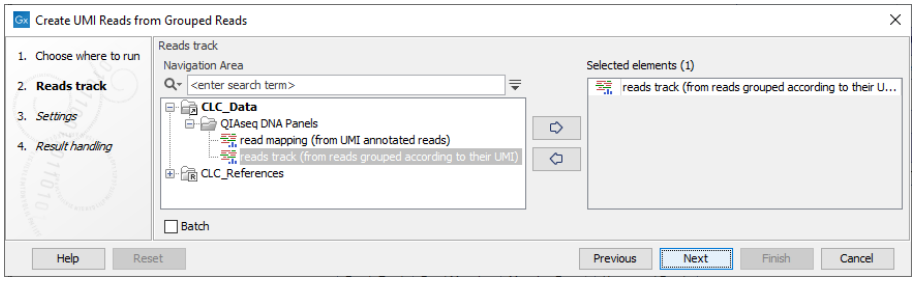


Figure 4.4: *Select a read mapping of the original reads with UMI annotations.*

In the first dialog (figure 4.4), select a read mapping of the original reads *with UMI annotations* that was previously handled with the Calculate Unique Molecular Index Groups tool.

The second dialog of the wizard (figure 4.5) offers the following options:

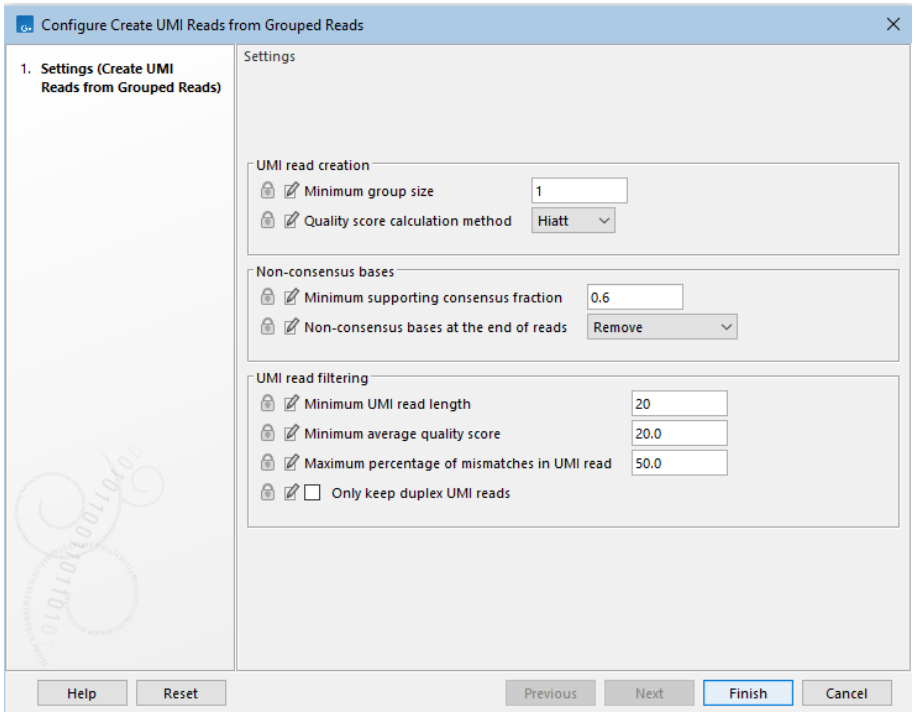


Figure 4.5: *Settings for the Create UMI Reads from Grouped Reads tool.*

- **UMI read creation**
  - **Minimum group size**: The tool will only create a UMI read if the number of reads in the UMI is at least "Minimum group size".
  - **Quality score calculation method**: Choose between two methods for computing Q-scores for UMI consensus reads (for more details see section 4.3.1):
    1. **Hiatt**: The Q-score is calculated following the method described in Hiatt et al., 2013. The quality score is $-10\log_{10}$ of the posterior probability of the consensus base.
    2. **MAGERI**: The Q-score is calculated using a slightly modified version of the method described in Shugay et al., 2017.

- **Non-consensus bases**
  - **Minimum supporting consensus fraction** set at 0.6 by default. At each position in the UMI read, the consensus nucleotide is chosen to be the nucleotide with the highest probability of being correct (see the Consensus nucleotide calculation paragraph below). If this probability is higher than "Minimum supporting consensus fraction", a Q score for the consensus nucleotide is calculated. A position in UMI reads that does not have a consensus nucleotide will be considered a N with Q score 0.
  - There is a choice between 3 methods of handling non-consensus bases (N with a Q score of 0) that are located at the end of the reads: **Remove** removes the bases, **Keep as unaligned** keeps the bases as unaligned ends, and **Keep as aligned** keeps the bases as aligned bases.
- **UMI read filtering**
  - **Minimum UMI read length**: UMI shorter than this value will be discarded.
  - **Minimum average quality score**: UMI reads will be discarded, if their average Q-score is lower than "Minimum average quality score".
  - **Maximum percentage of mismatches in UMI read**: UMI reads will be discarded, if more than this percentage of the bases are mismatches.
  - **Only keep duplex UMI reads**: Only duplex UMI reads will be kept. Simplex UMI reads and UMI reads from broken pairs will be discarded.

Click **Next** to **Open** or **Save** the resulting read mapping of UMI reads, i.e., a read mapping of the merged UMI groups. UMI reads are named `umi[UMI_ID]_count[UMI_group_size]`, where `UMI_ID` is a unique UMI group number, and `UMI_group_size` is the number of reads that are in that UMI group.

It is also possible to generate a report that will indicate how many reads were ignored and the reason why they were not included in a UMI read.

4.3.1 Consensus nucleotide calculation

The consensus nucleotide calculation is performed following the method described in Hiatt et al., 2013. The consensus base is chosen so that the posterior probability of the observed read bases is maximized.

In order to maximize the posterior probability of calling a base, i.e.,

$$P(C|O_1O_2\dots O_k) = \frac{P(O_1O_2\dots O_k|C)P(C)}{P(O_1O_2\dots O_k)} = \frac{P(O_1O_2\dots O_k|C)P(C)}{\sum_{x\in B} P(O_1O_2\dots O_k|x)P(x)}$$

where O_i is the observed base at a given position, C the base in question, and where all possible bases are summed up in the denominator, e.g. B=A,T,C,G.

Assuming that the prior for observing any base is equal, i.e., P(A)=P(T)=P(C)=P(G), then the posterior probability is:

$$P(C|O_1O_2\dots O_k) = \frac{P(O_1O_2\dots O_k|C)}{\sum_{x\in B} P(O_1O_2\dots O_k|x)}$$

And by assuming each read base observation is independent,

$$P(C|O_1O_2\dots O_k) = \frac{P(O_1|C)P(O_2|C)\dots P(O_k|C)}{\sum_{x\in B} P(O_1|x)P(O_2|x)\dots P(O_k|x)}$$

To obtain the consensus base we only need to maximize the numerator.

Consensus Q-score

The Hiatt Q-score is $-10\log_{10}$ of the probability of making a wrong call, i.e.

$$1 - P(C|O_1O_2\dots O_k)$$

which means that the Hiatt Q-score is

$$-10\log_{10}\left(1 - P(C|O_1O_2\dots O_k)\right)$$

Q-scores are capped at 60.

The probabilistic model outlined above and used in the Hiatt Q-score, assumes the only source of errors are independent sequencing errors. While PCR errors are typically rarer than sequencing errors, PCR errors are not independent and they can affect a large fraction of the reads in an UMI group. For this reason, the Hiatt quality scores will often attain the maximum value of 60, even in situations where the reads constituting the UMI group do not unanimously agree on the base call.

The Fixed Ploidy and Low Frequency Variant Detection tools both rely on statistical models for the sequencing error rates, which is estimated for each value of the Q-score and each substitution type, for details see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_Detection_error_model_estimation.html. If most quality scores are 60, the variant callers can not differentiate between, i.e. reads with unanimous agreement and reads without, or between small groups with unanimous agreement and large groups with unanimous agreement.

MAGERI Q-scores does not have a probabilistic interpretation as Hiatt Q-scores, but they a more distributed in the set of possible Q-scores, allowing the variant callers to differentiate between qualities. The MAGERI Q-scores is an adaption of the method described in Shugay et al., 2017.

First, the frequency, f, of the consensus base is computed, only bases with a Q-score above 25 contribute to the frequency computation. A pseudo-count is applied to the denominator, so that larger groups automatically get higher Q-scores:

$$f = \frac{c}{n + 0.9},$$

where c is the count of the consensus base and n is the total count.

The MAGERI Q-score is then computed as

$$Q = \frac{60}{3} \cdot (4f - 1).$$

4.4 Create UMI Reads from Reads

The tool **Create UMI Reads from Reads** generates a single consensus read, called a UMI read, from reads that have the same or similar UMI, and similar sequences. It can be used to process single end reads and paired end reads, including reads generated by duplex sequencing methods. Reads must be preprocessed by the Remove and Annotate with Unique Molecular Index tool before running **Create UMI Reads from Read**. It is recommended that the input reads are also trimmed for adapters and homopolymers.

This tool outputs a list of consensus UMI reads. Optional outputs include a QC Report and a list of discarded reads.

The algorithm is loosely inspired by Mash (Ondov 2016), Linclust (Steinegger 2017) and Calib (Orabi 2017) and involves clustering similar reads, merging reads in a cluster with the same UMI, and then filtering the merged reads. These steps are described in detail below:

Grouping reads

The tool makes extensive use of the minHash concept for Locality-Sensitive Hashing (LSH) to find clusters of similar reads. Briefly, all k-mers of each read are hashed with a number of hash functions, and for each of these hash functions the lowest hash value over all the k-mers is recorded. Two reads that share a lowest hash value are likely to share a k-mer and are linked together. In practice, a single hash function is not sufficiently specific to link related reads, and so sets of hash functions are used. Reads are only linked if each hash function in the set has the same lowest value for the two reads. The linked reads can be thought of as forming a graph, where each read is a node, and edges are made between the linked reads. Clusters of reads form disconnected subgraphs i.e. they are only linked to each other.

The tool uses LSH in three rounds:

1) In the first round reads are 'coarse clustered', so that it is very likely that similar reads are in the same cluster. The clusters are 'coarse' because it is also likely that unrelated reads are in the same cluster. The UMI is not used during coarse clustering.

2) In the second round each coarse cluster is 'fine clustered' by applying LSH again, but using different hash settings in order to get a more precise clustering. Reads in the resulting clusters are 'merged' if they have the same UMI.

3) In the third round all reads for a coarse cluster (merged or not) are again 'fine clustered' by LSH. Each disconnected subgraph is then pruned to only keep links between reads where (i) the UMIs differ by one mismatch, and (ii) either one of the reads is not merged, or the 'weight' of one of the reads is more than twice the weight of the other (this is the directional method in Smith 2017). The 'weight' of a merged read is the number of raw reads that were used to form the merged read.

Merging reads

The merging of similar UMI reads is performed using Multiple Sequence Alignment (MSA) based on the SPOA C++ library (https://github.com/rvaser/spoa). In general terms, the POA method creates a graph whose nodes represent the bases of the first sequence, and aligns the second sequence to the graph. Aligned bases of the first/second sequence are then merged in the graph. Every new sequence is aligned to the graph (using a modified version of the usual dynamic programming algorithms) and merged into the graph. As bases are merged into the graph, the edges between the nodes are updated to reflect their 'weight', which corresponds to the number

of times they occurred in multiple sequences. Finally, when all sequences are aligned, the graph is traversed to find the 'heaviest' path, which is the consensus sequence.

When sequences are aligned, mismatching nodes are connected by a special type of edge. Using these edges, it is possible to find all mismatches of a base in the consensus read. With this information in hand, the quality consensus is calculated using a method similar to Hiatt et al., 2013, which is also used in the other UMI tools. Every node keeps a match and a mismatch quality value (because it is not possible to know at the time of graph construction if it will be a match or not). Every time a new sequence is added and a new node is merged the quality values are updated using the quality scores of the new sequence. After calculating the consensus, the values in each node reflect the improved quality scores due to multiple sequences calling the same base. The quality consensus is obtained by processing the match and a mismatch quality value accordingly as described in Hiatt et al., 2013.

Duplex UMI

Reads produced by the TSO (TruSight Oncology 500, Illumina) duplex sequencing protocol can be processed to create a duplex consensus if reads from both strands are available. To create this consensus, reads probably originating from the same molecule and strand are grouped based on their similarity, UMI sequence and strand information. Single-strand consensus sequences are generated from these groups. Then, the consensus sequences of the opposing strands are combined to generate the duplex consensus sequences.

Background information: A protocol is considered duplex if the unique molecular barcode(s) attached to the target molecules include enough information to determine if a read originates from the same or from the opposite strand of the same target molecule. For the TSO protocol, the dual stranded reads have different barcodes attached to each end of the fragments, resulting in each read having two different barcodes named alpha and beta. Reads with the same orientation have identical alpha and beta tags, whereas reverse complements of the reads have alpha reads matching beta reads and vice versa (see section 4.1 for further information).

Running the tool

Create UMI Reads from Reads is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Create UMI Reads from Reads** ()

In the first dialog, select the sequence list containing the reads (figure 4.6).

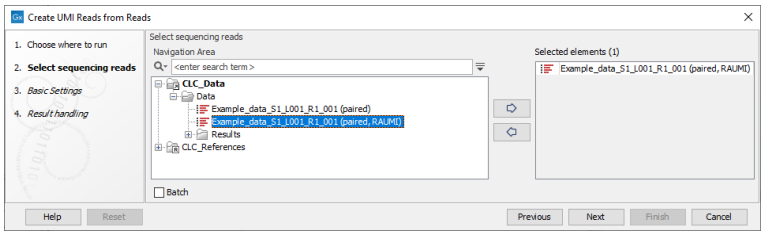


Figure 4.6: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

The next dialog allows you to configure basic settings for this tool, as shown in figure 4.7 and

described below.

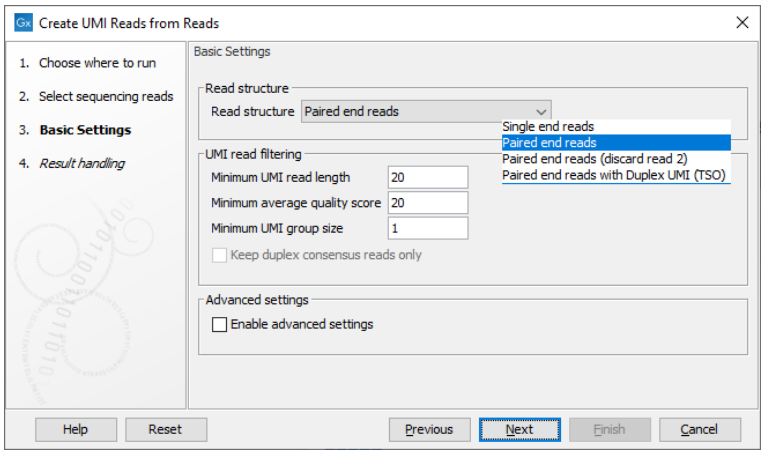


Figure 4.7: *Basic settings of the Create UMI Reads from Reads tool.*

- **Read structure** The read type and how they should be processed. The options are:
  - **Single end reads** When selected, each step of the algorithm described above is performed on each read.
  - **Paired end reads** When selected, each step of the algorithm described above is performed on read 1 (R1) and read 2 (R2) separately. During the hashing step, specific hash nodes are created only for R1. The purpose is to keep track of the UMI barcode and the LSH hashes. To link R1 and R2 a paired UMI reads object is created. In the final part of merging, reads are only combined in paired reads if R1 and R2 are similar. Strand specificity is not considered when this option is selected.
  - **Paired end reads (discard read 2)** Only the R1 member of a pair is processed. Discarding R2 is useful when it does not contain any biological information.
  - **Paired end reads with Duplex UMI (TSO)** Appropriate for reads generated using the TSO protocol. A duplex consensus will be generated. Here, both reads are hashed and added to the list for grouping. R1 hashes together with "alpha + beta" UMI, and R2 hashes together with "beta + alpha" UMI. In this way, reads from both strands end up in the same "coarse" groups. In the final part of merging, reads are only combined in paired reads if R1 and R2 are similar, the same as for the "Paired end reads" case, described above.
- **Minimum UMI read length**: UMI reads shorter than this value will be discarded.
- **Minimum average quality score**: UMI reads will be discarded, if their average Q-score is lower than the value specified here.
- **Minimum group size**: The tool will only create a UMI read if the number of reads in the UMI group is at least the size specified here.
- **Keep duplex consensus reads only**: When selected, only UMI reads built from reads representing both strands are retained. This is only relevant when a duplex UMI option has been selected for the "Read structure" setting.

- **Enable advanced settings**: When enabled, the advanced settings, available in the next wizard step, can be configured.

The next wizard step lists the advanced settings, which are organized in three categories: "Consensus options", "Coarse grouping" options, used for the coarse clustering step, and "Fine grouping" options, using for the fine clustering step (figure 4.8).

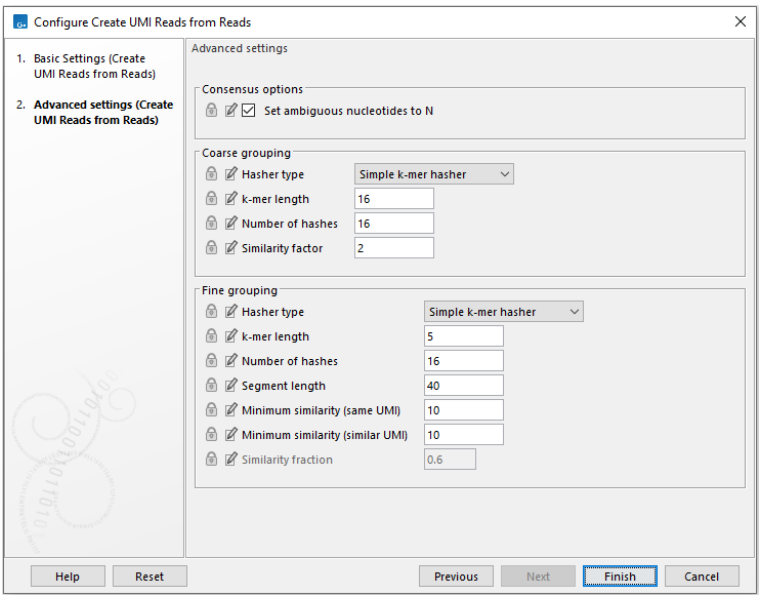


Figure 4.8: *Advanced settings of the Create UMI Reads from Reads tool.*

As both clustering steps use the same algorithm, their parameters are similar.

- **Hasher type**:
  - **Simple k-mer hasher**: Hashes are computed for every k-mer of a read.
  - **Spaced seed hasher**: Hashes are computed over multiple subsets of positions within the k-mer. For example, a spaced seed hasher of length 5 might make shorter 3-mers of positions 1,2,5 and 1,3,5.
  - **Rolling min k-mer hasher**: Hashes are the minimum k-mer hashes within k-mer length of each other for all read bases. This hasher can group reads that do not map to the same position, because the overlapping parts of the reads will have the same hashes.

  Note that for single primer extension protocols, reads representing the same DNA fragment can originate from different primers. This can happen if primers in the same direction are located near each other, making it possible for a downstream primer to amplify a PCR product generated from an upstream primer.

  These types of reads will not be grouped using the **Simple k-mer hasher option**, but they will be grouped when using the **Rolling min k-mer hasher** option.

  When reads originating from different primers are not grouped, this can lead to too high coverage. When there are few reads per UMI, this effect is expected to be negligible.

  When reads originating from different primers are grouped, primer sequence from reads starting downstream of another read is incorporated in the consensus UMI read. Because

the primer sequence is always the reference sequence, this can lead to correction of variants to reference sequence when the consensus is calculated. This can cause variant frequencies to become skewed, because a subset of UMI reads will have been corrected from variant to reference. Fusion detection and RNA expression results are unlikely to be affected by skewed variant frequencies. When there are few reads per UMI, this effect is expected to be negligible.

- **K-mer length**:
  - **Simple k-mer hasher**: A length in the range 2-32. Shorter k-mers lead to coarser clusters.
  - **Spaced seed hasher**: A length of 5, 8, 12, or 16. Shorter k-mers lead to coarser clusters.
  - **Rolling min k-mer hasher**: A length below 27. Shorter k-mers lead to coarser clusters.
- **Number of hashes**: the total number of hash functions applied to each k-mer of the read
- **Similarity factor / Minimum similarity (same UMI) / Minimum similarity (similar UMI)**: hash functions are divided into groups of this size. For two reads to be linked, all the hash functions in the group must have the same minimum hash values for the two reads. Therefore the number of hashes must be an exact multiple of this number. The higher this value, the finer the clusters.
- **Similarity fraction**: Minimum fraction of equal hashes between two reads required to group them. This is used by the rolling min k-mer hasher to handle reads with different number of hashes.

The remaining parameters are:

- **Segment length**: Only compute hashes over this number of bases at the start of a read. By only clustering on the start of the read, we reduce the chance of merging reads with different start positions (and which therefore likely come from different fragments) into a consensus UMI read.
- **Set ambiguous nucleotide to N (checked by default)**: This option determines whether to attempt to error correct UMI reads, or to replace conflict positions with 'N' to indicate that there is uncertainty about the true nucleotide. Error correction works by majority vote, and is useful when the predominant source of error is expected to be sequencing error. Replacing conflict positions with 'N' is more conservative, effectively discarding a position. This can be desired when errors are introduced by library preparation.

Click **Next** to **Open** or **Save** the sequence list of merged UMI reads. It is also possible to generate a report that will indicate how many reads were ignored and the reason why they were not included in a UMI read. The report also contains group size (see section 4.6) and quality score statistics useful for QC.

4.5 Create UMI Reads for miRNA

UMI reads are created with the **Create UMI Reads for miRNA** tool. This tool takes a sequence list as an input (reads including UMI sequences), and outputs a new sequence list where UMI reads have been merged. In the resulting sequence list, only the small RNA sequences are present annotated with (rather than containing) the UMIs. The output of this tool can be directly used in the small RNA quantification tool.

The expected read structure of the original (untrimmed) input is illustrated in (figure 4.9). It is therefore important to trim the 5' adapter on Ion Torrent reads before running the Create UMI Reads for miRNA tool.

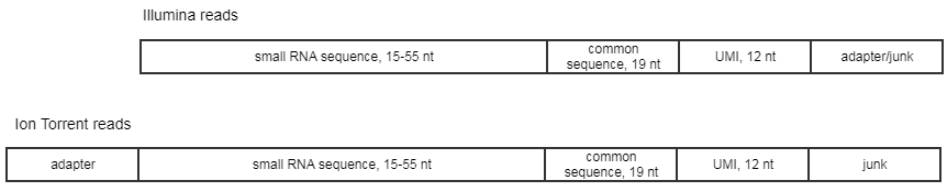


Figure 4.9: *Illumina and Ion Torrent expected read structure.*

The steps for merging UMI reads are as follows:

1/ The structure of the reads is analyzed. The common sequence is identified, and the small RNA sequence (preceding the common sequence) as well as the UMI (12 nucleotides following the common sequence) are identified. Reads where the common sequence is not found, or where the lengths of the small RNA or UMI do not fulfill the criteria are discarded. Note that it can be configured whether the common sequence should match exactly or whether mismatches - and how many of them, and whether these include indels - are allowed.

2/ Reads, stripped of common sequence and 3' adapter/junk, are grouped into UMI read groups based on exact identity of the small RNA sequence and the UMI. Each UMI read group keeps track of the number of reads merged into it, as well as the average nucleotide-level quality scores (if any) both for the small RNA sequence and the UMI part.

3/ We then attempt to merge "singleton" UMI read groups (containing only 1 read) into one of the existing UMI read groups based on how close the UMI and sequence match. The max number of mismatches for UMIs is set to 1. In addition, as is the case with the Create UMI Reads tool (section 4.3):

- UMIs are matched by SNVs first, then indels if enabled.
- If there are multiple best groups to merge into, the one with the most reads will be picked. In case of a tie an arbitrary existing group is selected, unless the option "Only merge into unique UMI group matches" is checked - in which case the singleton read is not merged.
- A perfect match on UMIs with an acceptable match in the sequence supersedes always perfect match in sequence with an acceptable match on UMIs. The reason to prefer the perfect UMI over perfect sequence is that we expect the UMI is only affected by random errors (sequencing errors), whereas the miRNA sequence is affected by both random errors and the systematic/biological variations in the small RNA sequence that we actually want

to be able to detect. Therefore we also expect that the "mismatch rate" will be higher in the sequence part than in the UMI part.

- We do not try to merge into groups with ambiguous bases, but we do allow them in the singleton groups to be merged. In this case they are just treated like any other SNV/indel.

Each resulting UMI read group produces one read without the UMI fragment in the output sequence list. Details on the statistics can be studies in the generated report.

Create UMI Reads for miRNA is available under the Tools menu at:

Tools | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Create UMI Reads for miRNA** ()

In the first dialog, choose the sequence list containing miRNA reads including UMI sequences as input. Then click **Next** to configure the following parameters (figure 4.10):

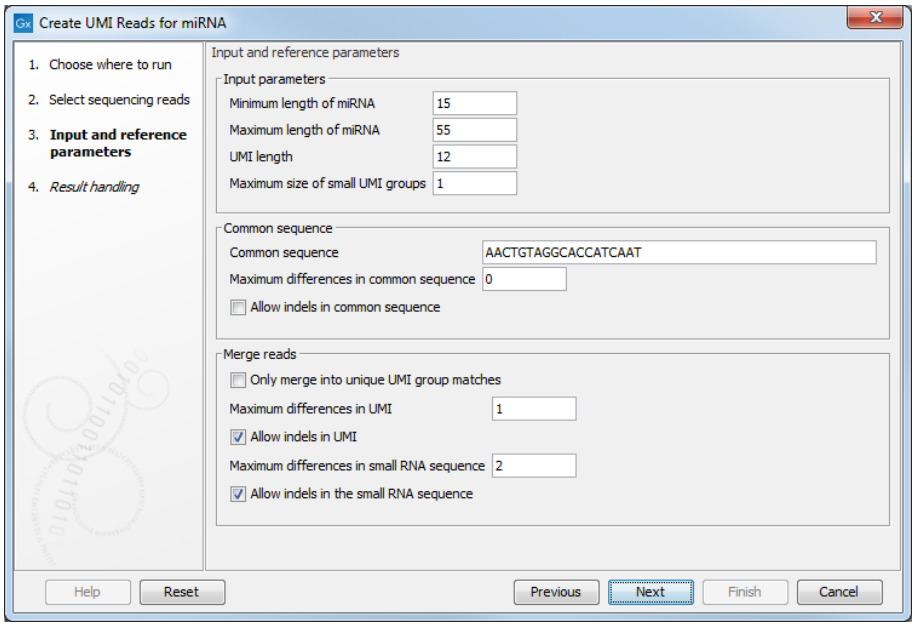


Figure 4.10: *Input and reference parameters for the Create UMI Reads for miRNA tool.*

- **Input parameters**
  - Minimum length of miRNA: miRNA shorter than this value will be discarded.
  - Maximum length of miRNA: miRNA longer than this value will be discarded.
  - UMI length: set to 12, the size expected when using a QIAseq miRNA protocol.
  - Maximum size of small UMI groups: UMI read groups are split into two categories. "Small" UMI groups either contain any number of reads with ambiguous nucleotides, or contain at most the number of reads specified by this parameter; the remaining groups are considered "large" UMI groups. The algorithm will merge small UMI groups into large ones whenever possible. Large UMI groups will not be merged with other groups.
- **Common sequence**

  - Maximum differences in common sequence: maximum number of mismatches allowed in the common sequence.
  - Allow indels in common sequence

- **Merge reads**
  - Only merge into unique UMI group matches: When enabled, this option means that a small UMI group will only be merged into a large one if there is only one candidate with the best score. Unchecking this option means that a small UMI group will be merged into one of the large one even if multiple candidates are found with the same score.
  - Maximum differences in UMI: Number of allowed differences in the UMI sequence when merging UMI groups. Note that this value can only be set to zero or one. As indels also count as a variation, it does not make sense to allow indels and have the number of variations be zero.
  - Allow indels in UMI
  - Maximum differences in small RNA sequence: Number of allowed differences in the miRNA when merging UMI groups.
  - Allow indels in small RNA sequence

The tool will output a read mapping of UMI reads, i.e., a read mapping of the merged UMI groups. In the last dialog, choose whether you would like to output a report that will indicate how many reads were ignored and the reason why they were not included in a UMI read. The report also contain group size statistics (see section 4.6) useful for QC. You can also output a file containing the discarded reads before opening or saving your results.

Consensus nucleotide calculation is performed following the method described in Hiatt et al., 2013, and can be summarized as follow:

- The UMI is defined by the majority regardless of quality. So if there are two reads with UMI ACG and one with ATG the UMI becomes ACG (there is tolerance for one SNV).
- In case of a draw, whichever read is first wins. So ACG, ATG, ACG and ATG becomes ACG whereas shuffling the order to ATG, ACG, ATG and ACG gives a consensus ATG.
- The quality is taken from the max quality of a read with exactly that UMI, so ACG (q 20), ATG (q 40), ACG (q 30) and ATG (q 50) becomes ACG (q 30). On the other hand shuffling the order to ATG (q 40), ACG (q 20), ATG (q 50) and ACT (q 20) makes it ATG (q 50). Variations have no negative impact on quality, so in the two cases above the q-score would be the same on all three nucleotides.

4.6 UMI group sizes

The tools Calculate Unique Molecular Index Groups, Create UMI Reads from Reads and Create UMI Reads for miRNA all find UMI groups, i.e. reads originating from the same fragment. An important set of QC metrics which is common for all three tools regards the sizes of the UMI groups.

The reports from the tools all contain a Group table with the following information:

- Output groups: The total number of UMI groups
- Singleton groups: The number of singleton UMI groups
- Average, Median and Standard deviation of reads per group
- Reads in largest group
- Reads by group size. "Group size" is the number of raw reads in a UMI group. For each read, its group size is recorded and these values are then sorted. The group sizes for a set of percentiles are reported
- Groups with sizes >=x (% of groups) (% of reads): A series of values reporting the number of groups containing at least a particular number of reads, followed by the percentage of UMI groups this represents and the percentage of all reads included in these groups

In addition, the following plots are also available:

- Reads by group size. The first plot shows the number of reads in groups by group sizes. The second plot includes only groups with fewer than 50 reads
- Group Sizes graphs. The first plot shows the sizes of all UMI groups. The second includes the sizes of only groups with fewer than 50 reads

For most applications the ideal UMI group size will be around 2-4, larger UMI groups tend to have diminishing returns for the increased sequencing budget. Please refer to the kit handbook to see the suggested UMI group size for your application.

4.7 Annotate Variants with Unique Molecular Index Info

The **Annotate Variants with Unique Molecular Index Info** tool annotates the variants with UMI groups information generated by the Calculate Unique Molecular Index Groups, and produces the annotated variant track as output.

Annotate Variants with Unique Molecular Index Info is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **UMI Tools** () | **Annotate Variants with Unique Molecular Index Info** ()

In the first dialog (figure 4.11), select a variant track.

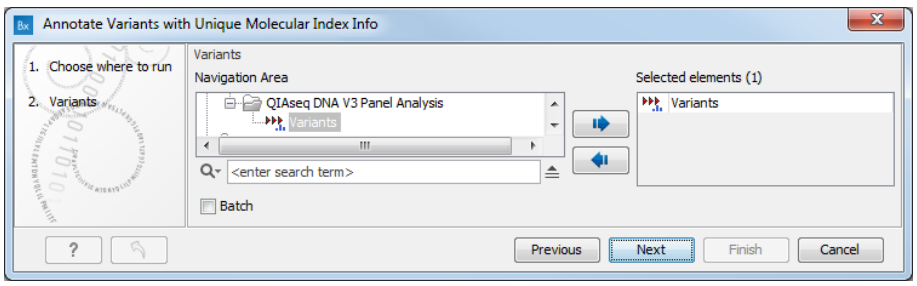


Figure 4.11: *Select a variant track.*

In the second dialog, select a read mapping. The tool works on any read mapping on which UMI groups have been calculated, i.e. a read mapping consisting of raw reads or a read mapping consisting of UMI consensus reads generated by the Create UMI Reads from Grouped Reads tool (as seen on figure 4.12). If the read mapping consists of UMI reads, check the "Mapping consists of UMI reads" option.

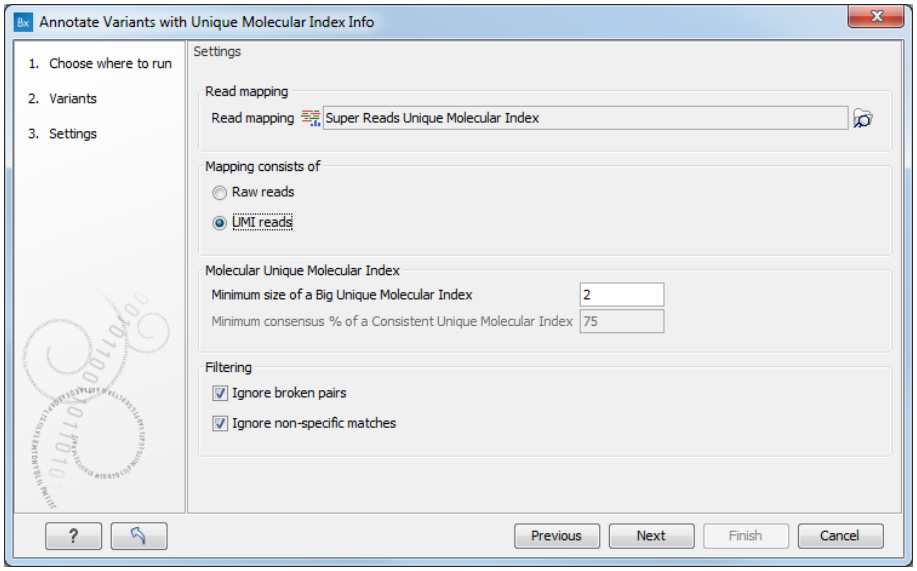


Figure 4.12: *Select a read mapping.*

The parameters below are used to calculate the annotations:

- **Minimum size of a Big Unique Molecular Index**: Minimum number of reads in a UMI group for it to be considered Big.
- **Minimum consensus % of a Consistent Unique Molecular Index**: Minimum percentage of reads in a UMI group that should support a variant for the UMI to be considered Consistent for that variant. This option is valid only if the read mapping chosen is made of raw reads.

Finally, it is possible to filter the data using the following options:

- **Ignore broken pairs**: reads from broken pairs will be ignored.
- **Ignore non-specific matches**: read that map in multiple places will be ignored.

Annotations The following annotations are added to the variants found using a read mapping consisting of raw reads, while only the three annotations indicated with a * are added when the read mapping consists of UMI reads. When using the Identify QIAseq DNA Variants workflows, the annotations are always based on UMI reads.

- Coverage (UMI): Number of UMI groups that overlap this variant. It is the coverage in the UMI reads track as seen by the Annotate Variants with Unique Molecular Index Info tool. Note that this value can be different form the Coverage value, which is based on the coverage in the UMI reads track as seen by the Low Frequency Variant Detection tool, where broken pairs, non-specific reads and reads with pyro-error variants are filtered out when using the default settings.

- Coverage (Big UMI): Number of big UMI groups that overlap this variant.
- Count (UMI): Number of UMI groups where at least one read has this variant.
- *Count (singleton UMIs): Number of singletons UMIs supporting the variant.
- *Count (big UMIs): Number of big UMIs supporting the variant.
- Count (Consistent and Big UMI): Number of Consistent and Big UMI groups that have this variant.
- *Proportion (singleton UMIs): Proportion of UMIs supporting the variant that are singleton UMIs.
- Freq (UMI): The percentage of UMI groups with this variant out of all UMI groups overlapping this variant.
- Freq (Consistent and Big UMI): The percentage of Consistent and Big UMI groups out of all UMI groups overlapping this variant.
- F/R (UMI coverage): Forward reverse balance of the UMI groups that overlap this variant.
- F/R (UMI count): Forward reverse balance of the UMI groups that have this variant.
- F/R (Big UMI coverage): Forward reverse balance of the Big UMI groups that overlap this variant.
- F/R (Consistent and Big UMI count): Forward reverse balance of the Big and Consistent UMI groups that have this variant.
- UMI info: A value of "24/29; 6/8; 1/40 (12 total)" means that there are 12 UMI groups with at least 1 read having this variant, the best of these groups consist of 29 read, where 24 of those reads have this variant, the second best group have 6 our of 8 reads with this variant. A variant can be overlapped by paired read that overlaps itself, where only the left or the right end has the variant. As long as at least one of the left or right ends of the paired read has the variant, we count the paired read as having the variant.

Note that the counts generated by the tool may differ from the counts generated by the variant callers. First, the tool ignores reads where the forward and reverse reads do not agree. Second, the variant callers take into account quality score and frequency of sequencing errors when computing counts and they may ignore broken pairs and/or non-specific matches based on user settings. For more details see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html and https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=General_filters.html.

Chapter 5

QIAseq tools

Contents

The following tools are developed specifically for QIAseq library kits.

5.1 Import QIAGEN Primers

To run the **Import QIAGEN Primers** tool, go to:

Import () | **Primers** () | **Import QIAGEN Primers** ()

The import wizard is shown in figure 5.1. The first step is to select the primers to import and a reference sequence.

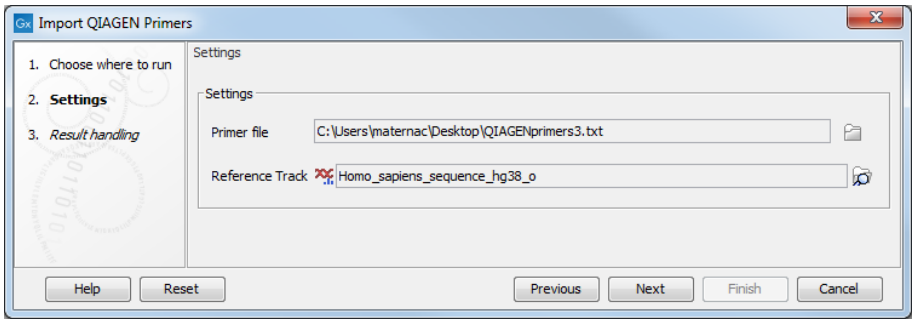


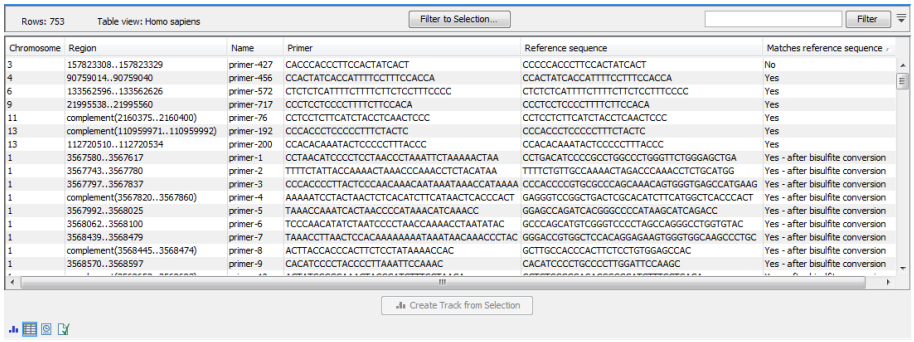
Figure 5.1: *Select the file to import.*

- **Primer File** Click on the folder icon to select the file you received upon purchase of a QIAseq Panel. The name of the file should include `primer3.txt` or `amplicons`.

- **Reference Track** Choose the hg19 or the hg38 reference sequence that are saved in the CLC Workbench after you have downloaded hg19 (or hg38) in the Reference Data Manager. The sequence can be found in the CLC_References folder in the Navigation Area tab, or using the QIAGEN Active sets folders in the Reference Data tab. Note that primers located on other chromosomes than those present in the selected reference track are skipped during import.

Click **Next** to go to the wizard step and choose to **Save** the imported primer location file.

Once the import completes, it is recommended to check the imported primers in the table view. If present, the "Matches reference sequence" column will have one of the following values: "Yes", "No", and in the case of QIAseq Targeted Methyl panels "Yes - after bisulfite conversion". A "No" indicates that the primer may have been designed against a more recent genome version, which has a corrected base compared to the reference. If there are many "No"s, it most likely indicates that an incorrect reference genome was supplied during import (figure 5.2).



Chromosome	Region	Name	Primer	Reference sequence	Matches reference sequence
3	157823308..157823329	primer-427	CACCCACCCTTCCACTATCACT	CCCCCACCCTTCCACTATCACT	No
4	90759014..90759040	primer-456	CCACTATCACCATTTTCCTTTCCACCA	CCACTATCACCATTTTCCTTTCCACCA	Yes
6	133562596..133562626	primer-572	CTCTCTCATTTTCTTTTCTTCTCCTTTCCCC	CTCTCTCATTTTCTTTTCTTCTCCTTTCCCC	Yes
9	21995538..21995560	primer-717	CCCTCCTCCCCTTTTCTTCCACA	CCCTCCTCCCCTTTTCTTCCACA	Yes
11	complement(2160375..2160400)	primer-76	CCTCCTCTTCATCTACCTCAACTCCC	CCTCCTCTTCATCTACCTCAACTCCC	Yes
13	complement(110959971..110959992)	primer-192	CCCACCCTCCCCCTTTCTACTC	CCCACCCTCCCCCTTTCTACTC	Yes
13	112720510..112720534	primer-200	CCACACAAATACTCCCCCTTTACCC	CCACACAAATACTCCCCCTTTACCC	Yes
1	3567580..3567617	primer-1	CCTAACATCCCCTCCTAACCCTAAATTCTAAAAACTAA	CCTGACATCCCCGCCTGGCCCTGGGTTCTGGGAGCTGA	Yes - after bisulfite conversion
1	3567743..3567780	primer-2	TTTTCTATTACCAAAACTAAACCCAAACCTCTACATAA	TTTTCTGTTGCCAAAACTAGACCCAAACCTCTGCATGG	Yes - after bisulfite conversion
1	3567797..3567837	primer-3	CCCAACCCCTTACTCCCAACAAACAATAAATAAACCATAAAA	CCCACCCCGTGCGCCCAGCAAACAGTGGGTGAGCCATGAAG	Yes - after bisulfite conversion
1	complement(3567820..3567860)	primer-4	AAAAATCCTACTAACTCTCACATCTTCATAACTCACCCACT	GAGGGTCCGGCTGACTCGCACATCTTCATGGCTCACCCACT	Yes - after bisulfite conversion
1	3567992..3568025	primer-5	TAAACCAAATCACTAACCCCATAAACATCAAACC	GGAGCCAGATCACGGGCCCCATAAGCATCAGACC	Yes - after bisulfite conversion
1	3568062..3568100	primer-6	TCCCAACATATCTAATCCCCTAACCAAAACCTAATATAC	GCCCAGCATGTCGGGTCCCCTAGCCAGGGCCTGGTGTAC	Yes - after bisulfite conversion
1	3568439..3568479	primer-7	TAAACCTTAACTCCACAAAAAAAATAAATAACAAACCCTAC	GGGACCGTGGCTCCACAGGAGAAGTGGGTGGCAAGCCCTGC	Yes - after bisulfite conversion
1	complement(3568445..3568474)	primer-8	ACTTACCACCCACTTCTCCTATAAAACCAC	GCTTGCCACCCACTTCTCCTGTGGAGCCAC	Yes - after bisulfite conversion
1	3568570..3568597	primer-9	CACATCCCCTACCCCTTAAATTCCAAAC	CACATCCCCTGCCCCTTGGATTCCAAGC	Yes - after bisulfite conversion

Figure 5.2: *The imported QIAGEN primers.*

QIAseq panel primer formats

The QIAseq panel primers are provided upon purchase of a kit, and the file can have the following formats.

The first file format is a **tab-separated file with 4 columns** defining:

1. Chromosome
2. Primer start/end position (0-indexed)
3. Whether the primer is on the plus strand indicated by an "L" or a "0", or on the minus strand indicated by an "R" or a "1".
4. The bases of the primer.

For example, the lines

```
chr1 1887011 L AGAATATTTTCTTGCTTAACCGTCACTTAACATCGA
chr1 1900114 R GGGACAAGACCTGGAACTACATTTCTGACT
```

define the primers: On chr1, from 1886977 to 1887012 (both are 0-indexed, inclusive), on the plus strand. On chr1, from 1900114 to 1900144 (both are 0-indexed, inclusive), on the minus strand.

The second file format is a **tab-separated file with 6 or 7 columns** defining:

1. Primer count (this value is ignored during import)
2. Chromosome
3. Start position (0-indexed) when on the "+" strand or End position (0-indexed) when on the "-" strand
4. End position (0-indexed) when on the "-" strand or Start position (0-indexed) when on the "+" strand
5. Strand ("+" or "-")
6. The bases of the primer
7. Target annotation (optional)

For example, the lines:

```
14 chr1 1886977 1887012 + AGAATATTTTCTTGCTTAACCGTCACTTAACATCGA COPA
2 chr1 1900114 1900144 – GGGACAAGACCTGGAACTACATTTCTGACT RCSD1
```

define the primers from the previous example, with their target annotation.

The third file format is a **tab-separated file with 11 or 14 columns** defining:

1. Gene identifier
2. Gene symbol
3. Chromosome
4. 5’ primer location (0-based)
5. 3’ primer location (0-based)
6. Genome strand (0 for binding to "-" but matching "+", and 1 for the opposite case)
7. The bases of the primer
8. Control primer flag (0 - not a control; 1 - reference gene expression control; 2-gDNA contamination control)
9. Genome blocks
10. Block sizes (comma-delimited)
11. Block starts (comma-delimited)
12. Whether the primer is designed for fusion calling (optional, 0 - no, 1 - yes)
13. Whether the primer is designed to target a SNP or indel (optional, 0 - no, 1 - yes)
14. Whether the primer is designed for use in gene expression (optional, 0 - no, 1 - yes)

For example, the lines:

```
ENSG00000000457 SCYL3 chr1 169859157 169859065 1 CTTCAATTCTGGATTCTTTACT
0 1 28 0
ENSG00000000457 SCYL3 chr1 169859969 169859079 1 GAGAACTTAGATCGATCGATTCCTG
0 1 30 0
```

define the primers from a RNAscan panel using the 11 column format.

The lines:

```
ENSG00000109685 WHSC1 chr4 1918581 1918613 0 GCATCCCAGTTTTTGGTCTTCTGTCAAAAACAC
33 0 0 0 0 1
ENSG00000109685 WHSC1 chr4 1959733 1961059 0 CAGAAAGGGAGAATTTGTTAACGAGTACGTTGG
7,26 0,1301 0 0 1 1
```

define the primers from a Fusion XP panel using the 14 column format.

All three file formats will be imported as paired primers when there is an even number of primers per chromosome and the lines are ordered with strands L, R, .. or R, L, .. (where L/R can also be +/- or 0/1 depending on the file format). Imported paired primers have an extra column "PrimerId" in the table view and work with **Trim Primers and their Dimers from Mapping**.

The fourth file format is a **tab-separated file with 12 columns** defining amplicons. Each amplicon implicitly defines two primers, and the same primer will be present multiple times if it amplifies multiple amplicons. The columns are:

1. Chromosome
2. 5’ primer location (0-based)
3. 3’ primer location (0-based)
4. Name of the left and right primers, separated by a "|"
5. This value is ignored during import
6. This value is ignored during import
7. 5’ primer location (0-based)
8. 3’ primer location (0-based)
9. This value is ignored during import
10. Block count - always "2" because each amplicon consists of two primers
11. Block sizes (comma-delimited)
12. Block starts (comma-delimited)

Primers from this fourth format will be paired primers with an extra column "PrimerId" in the table view. They will work with **Trim Primers and their Dimers from Mapping**.

The lines:

```
MN908947.3 3177 3412 QIAseq_23_LEFT|QIAseq_23_RIGHT 1 .  3177 3412 0 2 26,22
0,213
MN908947.3 3177 3425 QIAseq_23_LEFT|QIAseq_23-2_RIGHT 1 .  3177 3425 0 2
26,26 0,222
```

define primers from a QIAseq DIRECT SARS CoV-2 panel using this fourth format.

5.2 Quantify QIAseq RNA

The **Quantify QIAseq RNA** tool is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **QIAseq Tools** () | **Quantify QIAseq RNA** ()

The tool uses RNA-Seq reads as input, together with a reference sequence and target regions that are saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNA Panels hg38** reference data set.

We recommend using reads that have already been trimmed, as done for example in the Quantify QIAseq RNA Expression template workflow, see section 13.10. Trimming increases the results accuracy.

The tool performs multiple steps, as described below.

Mapping

The reads are mapped to the target regions using Map Reads to Reference (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Map_Reads_Reference.html) with default settings. "Auto-detect paired distances" is however disabled, as the automatic estimation of paired distances is not appropriate for targeted data.

The reference sequences of the target regions are created as follows:

- The sequence of the target regions is determined from the reference sequence.
- The exons of multi-exon targets are concatenated into one single target sequence.
- 12 ambiguous "N" nucleotides are prepended to the 5' region to account for UMIs.

Note that all other regions of the genome are ignored.

Filtering

Reads that successfully mapped to the target regions are removed if the UMI does not have the expected length of 12.

Merging

PCR and sequencing errors also happen in the UMIs. To account for this, UMIs can be merged as described in Peng et al., 2015. Briefly, UMIs that have a count of at most 2 or lower than 5% of the maximum observed UMI count, are considered for merging. For a UMI to be merged into another UMI, it must differ by at most one base pair and have a count that is at least 6 fold smaller.

5.3 QC for RNAscan Panels

The **QC for RNAscan Panels** is available from the Tools menu at:

Tools | **Biomedical Genomics Analysis** ( ) | **QIAseq Tools** ( ) | **QC for RNAscan Panels** ( )

Specify a RNA-Seq read mapping as input (figure 5.3).

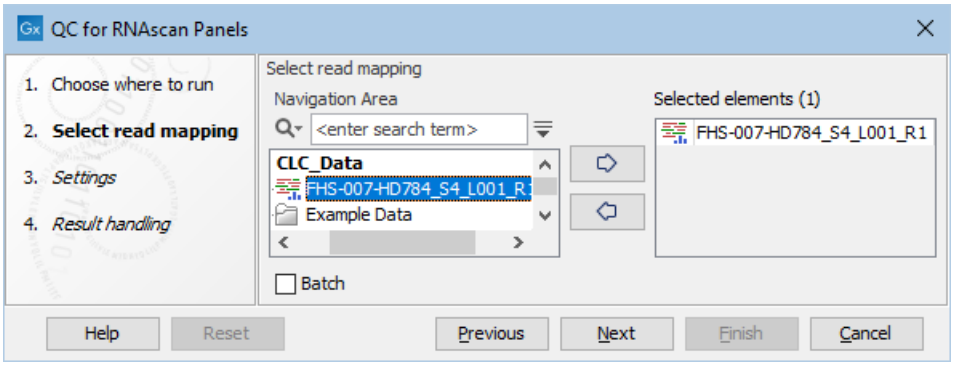


Figure 5.3: *Select a UMI read mapping.*

In the next dialog (figure 5.4), specify the mRNA track and the primer track that are saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNAscan Panels hg38** reference data set. You can also set a maximal distance between a read and a primer start for them to be considered matching. It is set by default to 0, which means that a read will not be considered as starting in the primer unless it maps exactly to the start of the primer.

The tool outputs a primer track with annotated read coverage and a report that recapitulates QC data (figure 5.5). The primer track gives information about each primer, as well as its read coverage, and whether it overlaps with target or housekeeping genes.

The QIAseq RNAscan Panels Report

A QC for RNAscan Panels Report contains the following information:

- Total number of mapped reads: The number of mapped reads, where each individual read is counted, i.e. paired reads are counted as two.
- Total number of mapped reads(-pair)s: The number of mapped reads, where each read pair is counted as one.
- Target genes: The names of the genes targeted by the primers.
- Number of target genes: The number of genes targeted by the primers.
- Primers in target genes: The number of primers within the target gene regions.
- Mean read coverage per target gene primer (start position match): Coverage by reads that start at the primer start site. The value is reported as "-" if the denominator in the calculation, "Primers in target genes", is 0.
- Mapped read(-pair)s matching primers (%). The percentage of read pairs mapping to any kind of primer, including primers in target genes, gDNA control primers, and reference gene control primers. A low value may indicate one or more of the following:
  - rRNA contamination

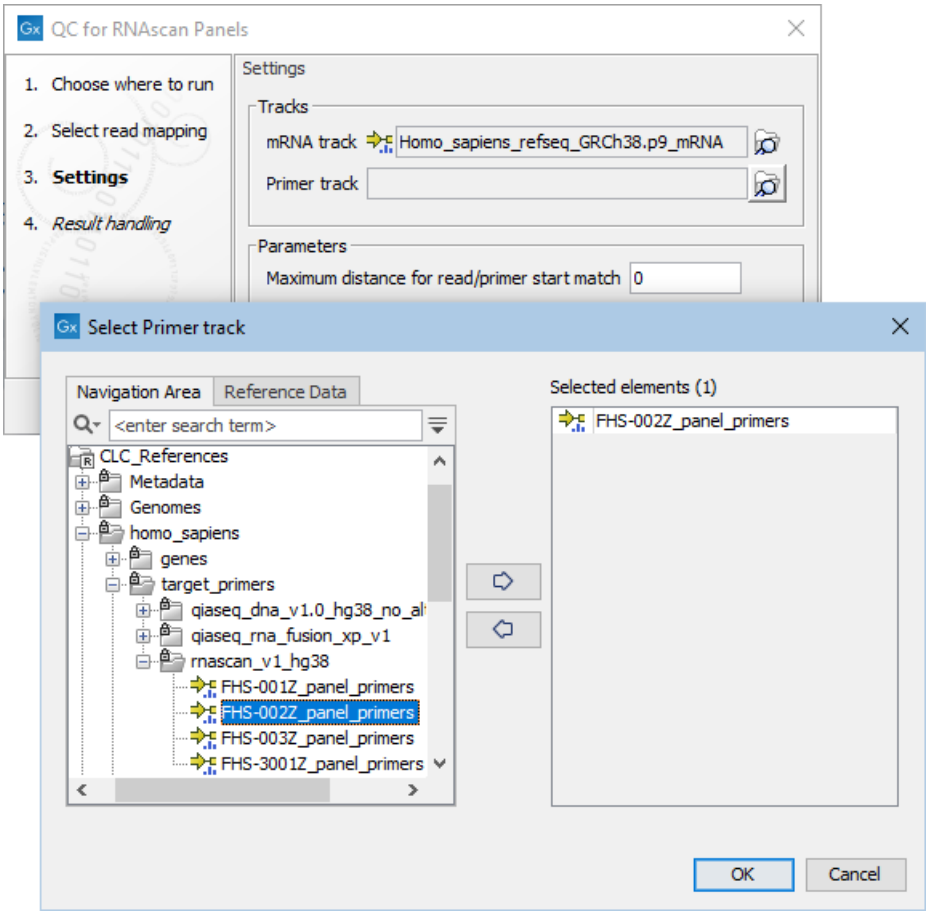


Figure 5.4: *Specify mRNA and primer tracks.*

- DNA contamination (estimated separately in this report when gDNA control primers are present)
- The presence of untrimmed adapters before the primer on the read, such that the read does not match the primer within the specified distance threshold
- Over-amplification. Reads from primers are more likely to share a UMI with other reads, and so are more likely to be merged into a single UMI read than off-target reads. Because of this, the percentage of read(-pair)s matching primers will typically decrease after the creation of UMI reads, especially when the average number of reads used to make a UMI read is very high.

This value is reported as "-" if the denominator in the calculation, "Total number of mapped read(-pair)s", is 0.

- Primers outside gene regions: The number of gDNA control primers outside target gene regions, used for detection of DNA contamination of samples.
- Mean read coverage per gDNA control primer: A value higher than 0 indicates DNA contamination of the sample. A mean coverage of around 50 reads or higher may increase the false positive signal level. This value is reported as "-" if the denominator in the calculation, "Primers outside gene regions", is 0.

1 QC For RNAscan Panels

Total number of mapped reads	3,279,860
Total number of mapped read(-pair)s	1,666,496
Target genes	ABL1, ALK, BCOR, BCR, CCDC88C, EML1, ETV6, FGFR1, FIP1L1, FLT3, INSL6, JAK2, KMT2A, LOC101928942, LOC107985554, MLLT1, MLLT10, MLLT3, NDE1, NPM1, NTRK3, NUMA1, PDGFRA, PDGFRB, PML, PRKG2, RANBP2, RARA, RCSD1, ZMYM2
Number of target genes	30
Primers in target genes	150
Mean read coverage per target gene primer (start position match)	137.65
Mapped read(-pair)s matching primers (%)	3.24
Primers outside gene regions	2
Mean read coverage per gDNA control primer	17.00
DNA contamination ratio	0.1235
Primers in reference genes (COPA, MRPS14, CIAO1, UBE3C)	4
Mean read coverage per reference gene control primer	8,338.25
Target gene vs reference gene coverage (%)	1.65

Figure 5.5: *QC for RNAscan Panels report.*

- DNA contamination ratio: Calculated as the ratio of "Mean read coverage per gDNA control primer" to "Mean read coverage per target gene primer (start position match)". This value is reported as "-" if any of the values "Primers outside gene regions", "Primers in target genes" or "Mean read coverage per target gene primer (start position match)" is 0. A ratio larger than 1 indicates that the sample contains more DNA than RNA.
- Primers in reference genes (COPA, MRPS14, CIAO1, UBE3C): QIAseq RNAscan panels typically include four reference gene primers.
- Mean read coverage per reference gene control primer: The reference genes should have a mean coverage of at least 300 reads, otherwise the effective input is too low, and false negatives are expected. This value is reported as "-" if the denominator in the calculation, "Primers in reference genes", is 0.
- Target gene versus reference gene coverage (%): Calculated as the ratio of the "Mean read coverage per target gene primer (start position match)" to "Mean read coverage per reference gene control primer", expressed as a percentage. This value is reported as "-" if "Primers in reference genes", "Primers in target genes" or "Mean read coverage per reference gene control primer" is 0.

5.4 Import Known Fusion Information Track

Import Known Fusion Information Track can import information about known fusions from a tab separated text file. It outputs an annotation track ( ) that can be used with **Annotate Fusions with Known Fusion Information**, see section 5.5.

The importer can be found here:

Import ( ) | **Import Known Fusion Information Track** ( )

The following options can be configured (figure 5.6):

- **File** A tab separated text file with the known fusion information. The first row in the file is a header. Each subsequent row lists the information for a known fusion.

  Information is extracted from the following columns in the file:

  - `5prime_gene` and `3prime_gene` (mandatory): the name of the 5'/3' gene.
  - `5prime_bp` and `3prime_bp` (optional): the position of the 5'/3' breakpoint. The position can be exact (e.g., 1364892), or uncertain (e.g., 1364887–1364895).
  - `5prime_chr` and `3prime_chr` (optional): the chromosome of the 5'/3' gene.
  - `5prime_strand` and `3prime_strand` (optional): the strand (+/-) of the 5'/3' gene.

  Further columns are added as additional attributes.

- **Gene track** An annotation track ( ) with the gene annotations. The 5' and 3' genes in the provided file are matched to this track by gene names, with support for synonyms. If chromosome and/or strand information is available in the file:

  - For genes that match the track:
    - If the chromosome in the file and track do not match, the fusion is not imported.
    - If the strand in the file and track do not match, the strand from the track is used.
    - If the breakpoint position is not available, the entire gene region is used.
  - For genes that do not match the track:
    - The fusion is imported with the chromosome/strand provided in the file.
    - Fusions with genes without a breakpoint position are not imported.

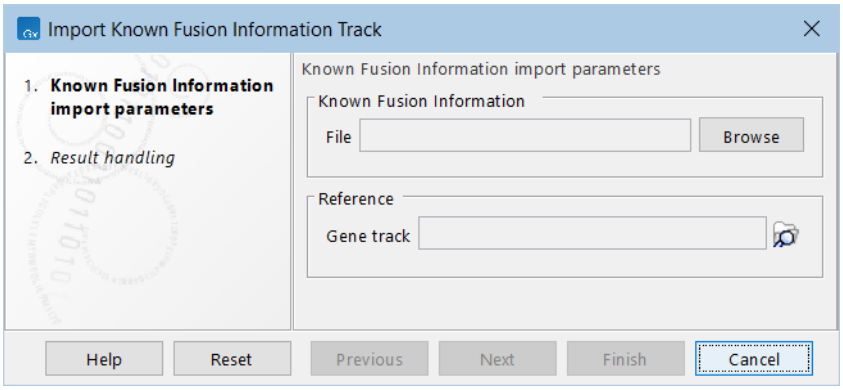


Figure 5.6: *The available options when importing known fusion information.*

5.5 Annotate Fusions with Known Fusion Information

The **Annotate Fusions with Known Fusion Information** tool annotates a fusion track (on wild type chromosomes only) with information from a fusion information file given as a feature track.

Annotate Fusions with Known Fusion Information is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **QIAseq Tools** ( ) | **Annotate Fusions with Known Fusion Information** ( )

Specify a fusion track as input (figure 5.7).

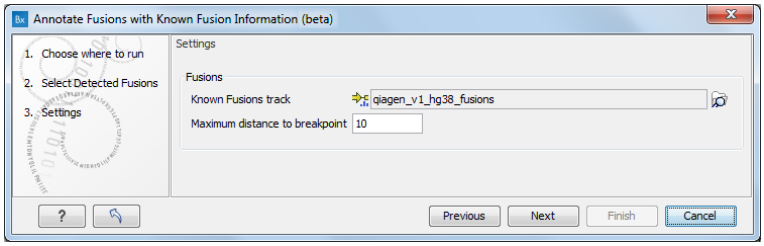


Figure 5.7: *Select a fusion track, and set up the maximum distance to breakpoint.*

In the next dialog, select the known fusion track you would like to use, for example the one that is saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNAscan Panels hg38** Reference Data Set for annotating with the wild type genome. You can also use a customized track imported using **Import Known Fusion Information Track**, see section 5.4.

In this dialog you can also set up the maximum distance to breakpoint: the tool will annotate only fusion breakpoints for which the distance between the detected breakpoint and the closest known fusion is smaller than the one specified.

The tool outputs a fusion track. It is similar to the one that was input, but includes a new Known Fusion column, indicating the Fusion ID Number of the matching fusion in the known fusion track, or -1. Additional information from the known fusion track if also added. For example, when using the known fusion track included in the **QIAseq RNAscan Panels hg38** Reference Data Set, the output track will have two more columns for Catalog Panels IDs and Cancer Types.

5.6 Validate QIAseq Read Structure (beta)

The **Validate QIAseq Read Structure (beta)** tool has the ability to analyze and validate the read structure for selected QIAseq panels and kits by predicting changes between UMI, common/adapter sequence and biological sequence based on nucleotide contributions for the four DNA nucleotides in the read. The tool takes a Graphical Report from **QC for Sequencing Reads** as input.

Validate QIAseq Read Structure (beta) predicts segments in the read by analyzing the changes in Standard Deviation (SD) of the nucleotide contributions in a sliding window throughout the read. If the nucleotide contribution for one of the four nucleotides in a position next to the window is significantly different from the nucleotide distribution inside the window, a potential change in read structure is found. The read structure prediction sensitivity parameter determines how many standard deviations the position next to the window should be from the mean of the window, in order to be significantly different.

When read segments have been predicted, the tool compares the predicted segments to the expected read structure. If all segments of the expected read structure are found in the sample, the read will pass the comparison. For paired end reads, R1 and R2 reads are analyzed and compared individually and the sample will pass if both of the reads pass.

The validation of read segments is based on three different criteria:

- Start and end positions: The start and end positions of each expected read segment must match the predicted read structure.
- Additional SD changes: If one of the expected read segments contains more than one predicted segment, the difference in nucleotide percentage SD for two adjacent predicted regions must be sufficiently small (by default the difference should be smaller than 10). See figure 5.9 for an example of additional SD changes.
- Nucleotide statistics: Average and SD values of the nucleotide percentage are calculated for each of the expected read segment. At least three of the four nucleotides, A, C, G and T, should have average and SD values that match the type of sequence in the segment. The acceptance criteria for average and SD values in the region depends on the type of sequence:
  - *UMI*: the nucleotides are evenly distributed and the standard deviation should be small (SD < 5). The average value should be between 8 and 50.
  - *Common sequence*: the nucleotides are unevenly distributed and the standard deviation should be high (SD > 8). No requirement for the average value.
  - *Staggered common sequence*: the nucleotides are unevenly distributed and the standard deviation should be high (SD > 6). No requirement for the average value.
  - *Biological sequence*: the nucleotides are evenly distributed however the standard deviation should be small or medium (SD < 20). The average value should be between 12 and 45.

  For the QIAseq Targeted Methyl Panel, the requirements on nucleotide statistics are different from the values above, since methylated panels have no cytosine in the biological sequence on R2, and no guanine in the biological sequence on R1.

Validate QIAseq Read Structure (beta) is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **QIAseq Tools** () | **Validate QIAseq Read Structure (beta)** ()

Double-click to run the tool.

Select one or more Graphical Report produced by **QC for Sequencing Reads** from the navigation area and add them to the list on the right hand side of the dialog using the arrows. Remember to enable Batch when analyzing multiple samples.

In the next wizard step select the QIAseq protocol used for generating the reads in the 'Expected QIAseq read structure' drop-down menu or select 'Unknown' if the kit is not supported. The tool will be preconfigured with optimal parameters for the different kits, however, sequencing is not an exact science and you might have to adjust the parameters to obtain the expected result. It is possible to adjust a number of parameters, see figure 5.8:

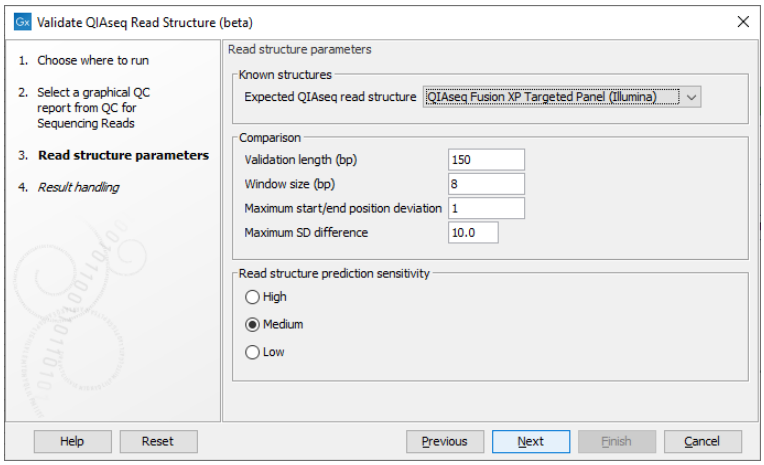


Figure 5.8: *Parameter wizard step of the Validate QIAseq Read Structure (beta)*

- **Comparison parameters:**
  - **Validation length (bp)** The number of residues in the read that are used for validation, counting from the 5' end of the read. Note that in some capture protocols a bias towards a single nucleotide, e.g., polyA-tails can dominate the end of the read, and in such cases only the stable part of the read should be included in the analysis.
  - **Window size (bp)** The number of residues considered in the sliding window for calculating standard deviations of the nucleotide percentages. Predicted read segments cannot be shorter than the windows size, so the value should not be smaller than the length of the UMI or common sequence, whichever is shortest. The default value is 8, which allows for more stable SD estimation compared to smaller values, the minimum allowed value is 5.
  - **Maximum start/end position deviation** A segment in the expected read structure fails if the difference between predicted and expected start or end position is larger than this value. Be careful to adjust this value because not all template workflows are build to tolerate a deviation.
  - **Maximum SD difference** A difference larger than this between nucleotide percentage standard deviations for two adjacent predicted regions will be highlighted. Large

differences between adjacent regions can indicate a departure from the expected read structure. The standard deviations are calculated as averages over the four DNA nucleotides, hence the SD difference for a single nucleotide could be larger than this value without making the segment fail the comparison.

Note that the different QIAseq protocols have different default values for this parameter, most protocols use 10. The QIAseq Multimodal RNA panel has a default of 30 since the capture protocol results in a polyT-rich region on R2 that will often lead to an unexpected SD change at position ∼30-32 which should be ignored.

- **Read structure prediction sensitivity:** The prediction sensitivity affects how sensitive the analysis is with respect to detecting changes in the read structure. The sensitivity might be chosen based on the quality of the reads and/or the sequencing technology. The options are:
  - **High** A high sensitivity results in more read segments being predicted. This setting should be used when sequencing has been performed on the more noisy platforms, e.g., Illuminas MiSeq.
  - **Medium** Default selection that fits most profiles.
  - **Low** A low sensitivity results in fewer read segments being predicted. With this setting, changes in the nucleotide contributions needs to be more significant for splitting the read into segments.

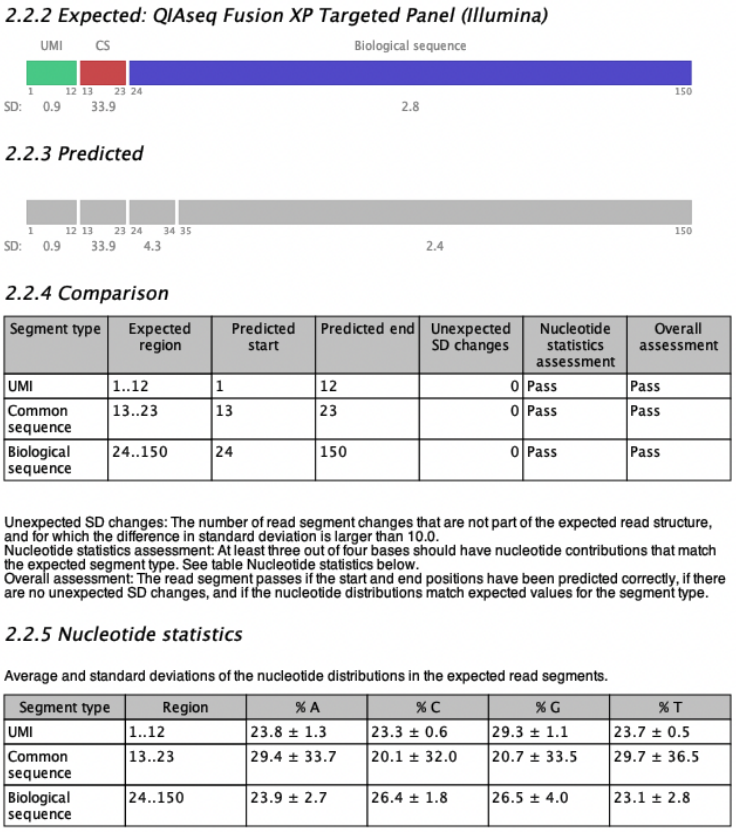
5.6.1 Output from Validate QIAseq Read Structure (beta)

Validate QIAseq Read Structure (beta) generates a report containing a summary section with information about whether the sample matches the expected read structure, and a read structure section showing plots of the read structure and a comparison of the expected and predicted structures. For paired end reads there are two read structure sections, since the two reads are handled separately.

An example of the read structure section is shown in figure 5.9 for a QIAseq Fusion XP Targeted Panel (Illumina) sample. The expected read structure of R2 is plotted in section 2.2.2 of the report; the QIAseq Fusion XP R2 read consists of a UMI, a common sequence and a biological sequence. Colored boxes represent read segments and the segment types are labeled above the segments (CS for common sequence). The start and end positions are shown underneath each segment together with the Standard Deviation (SD) of the segment as a measure of the nucleotide variation within the region. The segment SD is calculated as the SD of the nucleotide distribution in the region for each of the four DNA nucleotides, then averaged over the four nucleotides.

The predicted read structure is plotted in section 2.2.3 of the report. Grey boxes represent predicted read segments and similar to the expected read segments, they are labeled with SD values and start and end positions.

The expected and predicted read structures are compared in table 2.2.4. The table has a row for each expected segment. In figure 5.9, the start and end positions of all segments are predicted correctly, but the biological sequence is predicted as two segments with an additional break at position 34. The SD of the two biological regions are 4.3 and 2.4, respectively (shown under the plot in section 2.2.3 Predicted). Thus, the SD difference between the two regions is only 1.9,



2.2.2 Expected: QIAseq Fusion XP Targeted Panel (Illumina)

UMI CS Biological sequence

1 12 13 23 24 150

SD: 0.9 33.9 2.8

2.2.3 Predicted

1 12 13 23 24 34 35 150

SD: 0.9 33.9 4.3 2.4

2.2.4 Comparison

Segment type	Expected region	Predicted start	Predicted end	Unexpected SD changes	Nucleotide statistics assessment	Overall assessment
UMI	1..12	1	12	0	Pass	Pass
Common sequence	13..23	13	23	0	Pass	Pass
Biological sequence	24..150	24	150	0	Pass	Pass

Unexpected SD changes: The number of read segment changes that are not part of the expected read structure, and for which the difference in standard deviation is larger than 10.0.
Nucleotide statistics assessment: At least three out of four bases should have nucleotide contributions that match the expected segment type. See table Nucleotide statistics below.
Overall assessment: The read segment passes if the start and end positions have been predicted correctly, if there are no unexpected SD changes, and if the nucleotide distributions match expected values for the segment type.

2.2.5 Nucleotide statistics

Average and standard deviations of the nucleotide distributions in the expected read segments.

Segment type	Region	% A	% C	% G	% T
UMI	1..12	23.8 ± 1.3	23.3 ± 0.6	29.3 ± 1.1	23.7 ± 0.5
Common sequence	13..23	29.4 ± 33.7	20.1 ± 32.0	20.7 ± 33.5	29.7 ± 36.5
Biological sequence	24..150	23.9 ± 2.7	26.4 ± 1.8	26.5 ± 4.0	23.1 ± 2.8

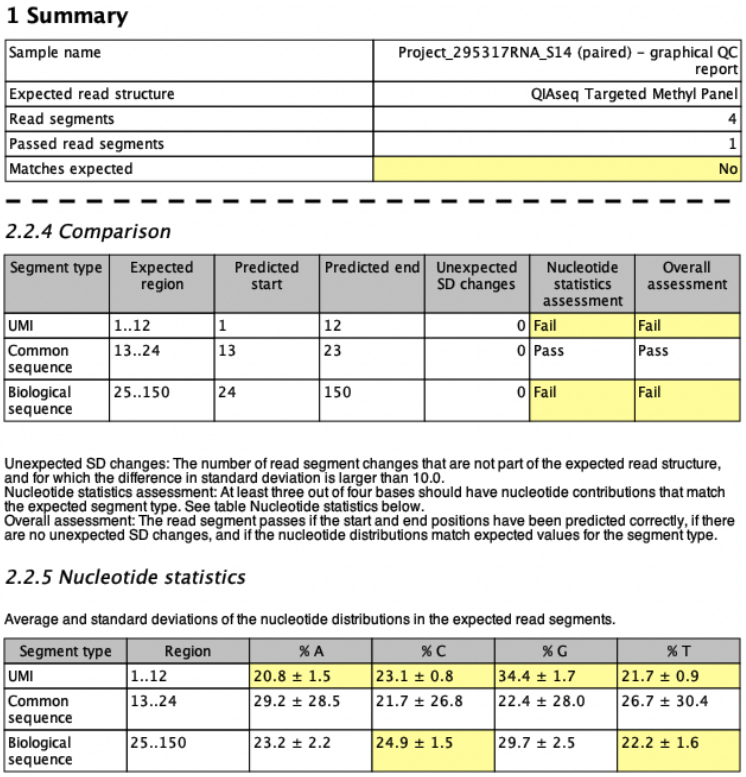
Figure 5.9: *An example of expected and predicted read structures from a read structure report when analyzing a QIAseq Fusion XP Targeted Panel Illumina sample.*

which is smaller than the default threshold of 10, and the difference is therefore not problematic and not included in the comparison table in the column 'Unexpected SD changes'.

Finally, table 2.2.5 Nucleotide statistics, shows the average and SD values of the four DNA nucleotide contributions in each of the expected read segments. Cells are highlighted in yellow if a read segment does not match the expected values (average and SD) for that type of segment. In this case all four nucleotide pass for all segments, and all cells are white.

If the sample does not match the expected read structure, it is highlighted in the Summary table of the report as shown in figure 5.10, where a QIAseq Fusion XP Targeted Panel Illumina sample is analyzed with the expected read structure QIAseq Targeted Methyl Panel. Yellow cells in the Comparison and Nucleotide statistics tables highlight where the comparison fails. In this case, the UMI and the biological sequence fail the comparison, while the common sequence pass (both QIAseq Targeted Methyl Panel and QIAseq Fusion XP Targeted Panel have a common sequence in positions 13-24 on R2).

For the UMI, the nucleotide distribution fails for all four bases. This is due to the fact that methylated UMIs have the form NNC (where N is random) and to pass the nucleotide statistics assessment, the sample should have SD values larger than 3 for T, larger than 8 for A and G, and larger than 20 for C. In addition, the biological sequence on R2 fails for C and T. C is supposed to have an average contribution smaller than 8% while T is supposed to have an average contribution larger than 35% in biological regions. As mentioned earlier, the segment comparison could also



1 Summary

Sample name	Project_295317RNA_S14 (paired) – graphical QC report
Expected read structure	QIAseq Targeted Methyl Panel
Read segments	4
Passed read segments	1
Matches expected	No

2.2.4 Comparison

Segment type	Expected region	Predicted start	Predicted end	Unexpected SD changes	Nucleotide statistics assessment	Overall assessment
UMI	1..12	1	12	0	Fail	Fail
Common sequence	13..24	13	23	0	Pass	Pass
Biological sequence	25..150	24	150	0	Fail	Fail

Unexpected SD changes: The number of read segment changes that are not part of the expected read structure, and for which the difference in standard deviation is larger than 10.0.
Nucleotide statistics assessment: At least three out of four bases should have nucleotide contributions that match the expected segment type. See table Nucleotide statistics below.
Overall assessment: The read segment passes if the start and end positions have been predicted correctly, if there are no unexpected SD changes, and if the nucleotide distributions match expected values for the segment type.

2.2.5 Nucleotide statistics

Average and standard deviations of the nucleotide distributions in the expected read segments.

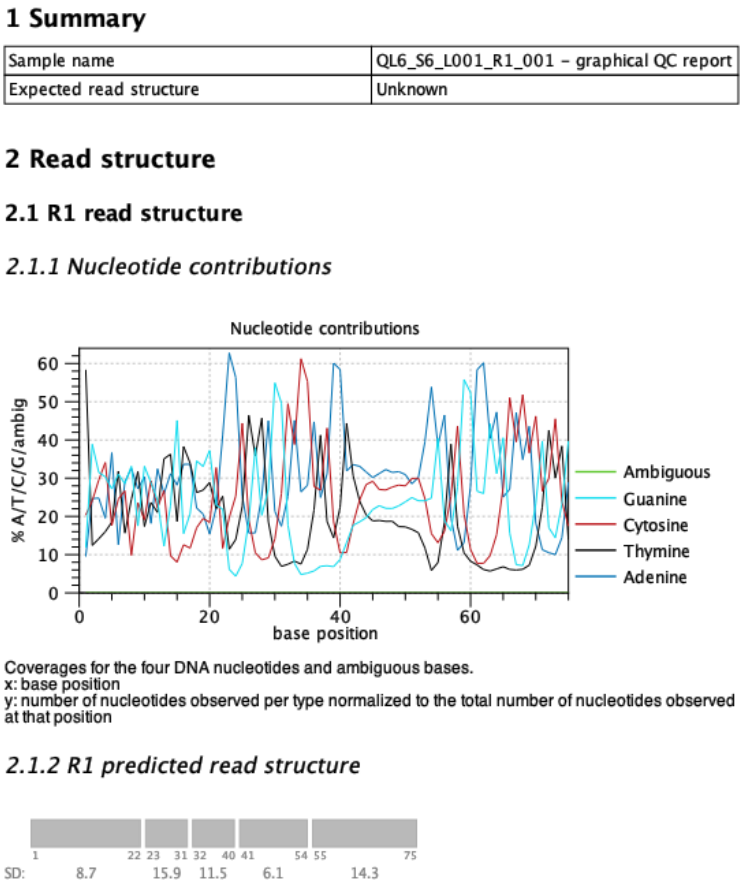
Segment type	Region	% A	% C	% G	% T
UMI	1..12	20.8 ± 1.5	23.1 ± 0.8	34.4 ± 1.7	21.7 ± 0.9
Common sequence	13..24	29.2 ± 28.5	21.7 ± 26.8	22.4 ± 28.0	26.7 ± 30.4
Biological sequence	25..150	23.2 ± 2.2	24.9 ± 1.5	29.7 ± 2.5	22.2 ± 1.6

Figure 5.10: *Snippets from a failed read structure report when analyzing a QIAseq Fusion XP Targeted Panel Illumina sample using the QIAseq Targeted Methyl Panel as expected read structure.*

fail if start or end positions of the expected read segments are not detected or if any additional segments are predicted with sufficiently large SD changes.

If the sequence protocol is unknown or not included in the list of supported protocols, the expected QIAseq read structure can be set to 'Unknown' to produce a report that only contains the predicted read structure and no comparison. An example is the QIAseq miRNA format, that is a bit tricky to analyze given the protocol, see figure 5.11. The structure consists of the miRNA segment, a common sequence, the UMI and an adapter/junk in that order. However, the miRNA sequence itself can vary in length making the read structure less predictable. It is possible to optimize the parameters in a way that the tool will predict the most likely structure of most reads in the sample.

The predicted read structure plot can be compared to the nucleotide contributions above to get an idea of the sequence type for each of the predicted segments. This can help decide if there are potential adapters that need to be trimmed. In the miRNA example from figure 5.11 the tool predicts the miRNA to be 22 bp long, which matches the length of most miRNAs. It can also identify the UMI (positions 41-54), but struggles a bit more with the common sequence which is staggered and does not align for the individual reads (the common sequence for this kit is AACTGTAGGCACCATCAAT). The staggered common sequence makes it hard to predict the exact start and end positions, which is the reason for the tool not supporting this protocol.



1 Summary

Sample name	QL6_S6_L001_R1_001 – graphical QC report
Expected read structure	Unknown

2 Read structure

2.1 R1 read structure

2.1.1 Nucleotide contributions

Coverages for the four DNA nucleotides and ambiguous bases.
x: base position
y: number of nucleotides observed per type normalized to the total number of nucleotides observed at that position

2.1.2 R1 predicted read structure

Figure 5.11: *The read structure report from Validate QIAseq Read Structure (beta) when analyzing a QIAseq miRNA sample. The parameters were tweaked to obtain the most likely structure by setting the sensitivity to high and using a smaller window size of 6.*

Chapter 6

Biomedical utility tools

Contents

6.1 Annotate Structural Variants

The main purpose of the Annotate Structural Variants tool is to calculate and add **Count**, **Coverage** and **Frequency** annotations to structural variants from InDel tracks detected by the InDels and Structural Variants tool.

The following tracks are taken as inputs:

- An InDel track. An output from the InDels and Structural Variants tool
- A Breakpoint track. An output from the InDels and Structural Variants tool
- A reads track. The read mapping used as input to the InDels and Structural Variants tool that led to the generation of the two tracks mentioned above.

These are used in the calculation and for further annotation of the resulting variant track. The Count, Coverage and Frequency are calculated as follows:

- **Count** The number of reads supporting the variant detected by the InDels and Structural Variants tool.
- **Coverage** The number of reads that overlap one or more relevant positions in the read mapping. "Relevant positions" here are the bases before and after the start and the end of the structural variant, and the bases before and after the breakpoints defining the structural variants.

- **Frequency**The number of reads in the mapping supporting the variant (Count) divided by the number of reads covering the relevant positions in the mapping, as defined above (Coverage).

Variants that are of the same type (SNV, MNV, insertions or deletions) are collapsed into a single variant if the breakpoints supporting the variant are within 20 bp and the calculation is performed on the consolidated variant.

Coverage for tandem duplications: Tandem duplications are insertions. For coverage of insertions, the reads overlapping the two positions before and after the insertion are considered. One breakpoint is the same as the insertion and so no new information is derived from this. The other breakpoint is at the other end of the detected duplicated interval, adding two relevant positions (before and after this breakpoint). So, for example, the coverage of a tandem duplication of length 40 at position 100 would be the number of mapped reads overlapping one or more of the following positions: 99, 100, 139, and 140.

Annotations from the Breakpoint track that are added to the variants are: p-value, Unaligned, Unaligned length, Perfect mapped, Not perfect mapped and Ignored mapped.

To run the Annotate Structural Variants tool, go to:

> **Tools** | **Biomedical Genomics Analysis** () | **Biomedical Utility Tools** () | **Annotate Structural Variants** ()

The Annotate Structural Variants tool generates a variant track containing the same variants as the original InDel track with **Count**, **Coverage**, **Frequency**, and additional annotations taken from the Breakpoint track.

6.2 Extract Reads Matching Primers

Contamination of next generation sequencing data is a common problem. Some of the problems when analyzing RNA sequencing data are caused by the presence of ribosomal RNA (rRNA) or genomic DNA (gDNA) in the sample. Likewise, RNA contamination can be a problem in DNA sequencing data.

Depletion of rRNA from RNA sequencing experiments is often performed using polyA enrichment for retaining RNA molecules with a polyA tail, a common feature of protein coding transcripts, and in this way eliminating rRNA sequences. Although the positive polyA selection usually is an efficient way to get rid of rRNA, polyA rich rRNA sequences do exist, and this type of rRNA may remain in sample.

A number of QIAseq RNA sequencing protocols can benefit from cleaning up the reads prior to mapping to remove contaminating rRNA reads. Using panel primer sequences as anchor, to only keep reads that match a primer sequence, can eliminate potential rRNA sequences. This is of particular importance when detecting gene fusions as polyA rich rRNA sequences can produce false positive gene fusions. Removal of these polyA rich rRNA sequences not only increases the quality of the fusion call but also decreases the run time as less fusion events are formed and analyzed.

In multi-modal applications, the presence of contaminating RNA in the DNA samples can also lead to false positive variant calls due to the different nature of the read composition where

especially InDels are problematic, but also RNA editing can introduce variants that are RNA editing artifacts.

To improve the sample purity and thereby potentially decreasing the number of false positive calls it can be useful to remove reads that do not match any primers. The tool "Extract Reads Matching Primers" extracts reads that match a primer and discards reads that do not match a primer. The tool takes unmapped DNA or RNA sequencing reads as input. We recommend using the "Extract Reads Matching Primers" tool on the raw sequencing reads before analyzing the data.

To run the Extract Reads Matching Primers tool, go to:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **Biomedical Utility Tools** ( ) | **Extract Reads Matching Primers** ( )

After you have specified whether you want to run the job locally or connected to a server, you are asked to select sequencing reads (figure 6.1).

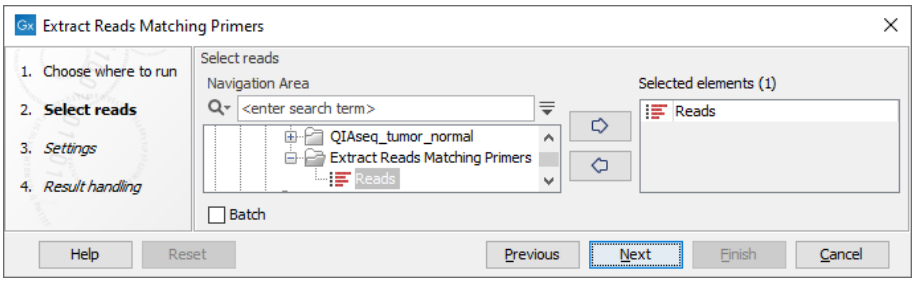


Figure 6.1: *Select unmapped sequencing reads.*

In the next dialog a number of different settings can be adjusted (figure 6.2).

The settings you can specify or adjust are:

- **Elements**
  - **Reference** A reference sequence track ( ) compatible with the selected primer track.
  - **Primers** A track containing the original primers and their intended primer locations ( ). A description of how to import pairs of primers can be found in the Import Primer Pairs section of the CLC Genomics Workbench manual: `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Import_Primer_Pairs.html`. A description of how to import QIAGEN primers can be found in section 5.1.
  - **Bisulfite** Check this box if you are analyzing methylation data using bisulfite converted primer sequences.
- **Primer matching**
  - **Primer location** Allows specification of where the primers are located on the input reads. One or more of the following options can be selected:
    - **Start of R1** The primers are found in the beginning of the R1 reads for paired-end reads or in the beginning of all reads for single-end reads.

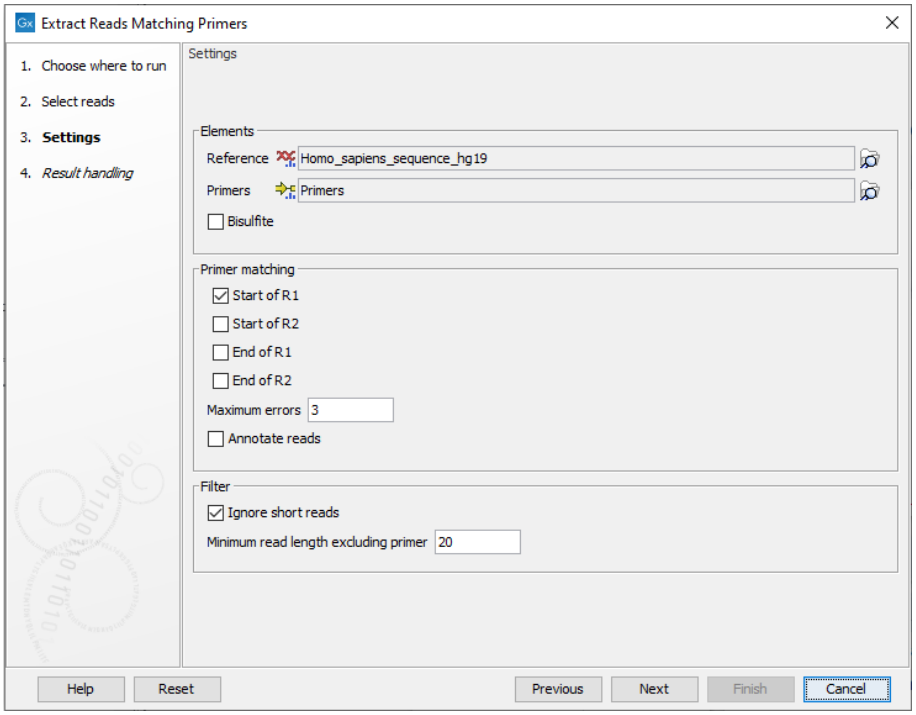


Figure 6.2: *In addition to selecting reference sequence and the relevant primer track, different settings can be adjusted in this dialog.*

  - **Start of R2** The primers are found in the beginning of the R2 reads for paired-end reads or in the beginning of all reads for single-end reads.
  - **End of R1** The primers are found at the end of R1 reads for paired-end reads or at the end of all reads for single-end reads.
  - **End of R2** The primers are found at the end of R2 reads for paired-end reads or at the end of all reads for single-end reads.
- **Maximum errors** The maximum number of mismatches allowed between the primer sequence and the read sequence. A read and a primer are not matched if the number of mismatches between primer and read exceeds the specified number of Maximum errors.
- **Annotate reads** Annotates each extracted read with information about whether it should be used for detection of fusions, detection of variants, and quantification of gene expression. This option only has an effect when this information is present on the primers.

- **Filter**
  - **Ignore short reads** Allows exclusion of short reads that match a primer but where the length of the read excluding the primer sequence is below a length that can be specified under "Minimum read length excluding primer".
  - **Minimum read length excluding primer**. When "Ignore short reads" is checked the "Minimum read length excluding primer" can be specified. Reads matching a primer but where the read excluding the primer sequence is shorter than this value will be discarded.

The output from the "Extract Reads Matching Primers" tool is a list of sequencing reads that match a primer and that have a length that is at least the length of what was specified under "Minimum read length excluding primer" if the option "Ignore short reads" was selected.

6.3 Identify Mispriming Events

Primers with high similarity to multiple genomic regions have the potential to be involved in mispriming, where reads are amplified from a region of the genome other than the intended target region. Reads resulting from such mispriming events are fused constructs: the primer part and the read part represent different regions of the genome.

A fraction of the population of a given primer may be involved in mispriming events, resulting in low frequency variants. How large this fraction is depends on the binding affinity and specificity of the primer, and the conditions that the lab work was performed under. Reads originating from mispriming events should be identified and removed from mappings to avoid calling false positive variants. If reads from mispriming events map to the region they originate from, and non-target regions of interest are known, then the primer can be unaligned, rather than removing the whole read.

Identify Mispriming Events generates a list of potential mispriming events for a set of panel primers and a specified reference genome. This list of mispriming events can then be supplied to **Trim Primers of Mapped Reads** to remove reads likely to represent a mispriming event, or to unalign primer parts of such reads, as relevant. This should precede variant detection, so as to minimize false positive variant calls due to artifacts generated from mispriming events.

Remove reads or unalign primer regions?

Trim Primers of Mapped Reads can handle misprimed reads when provided with a track of predicted mispriming events. Reads are either removed completely from the read mapping or having their primer region unaligned. This is done automatically during primer trimming to avoid calling false positiv variants, and the action needed depends on where reads resulting from a mispriming event are mapped:

- **To the original, intended target region** Reads amplified from a non-target region may still map to the original, intended target region if it has sufficient similarity to that region.
  - *Symptom:* No mismatches in the primer region in the mapping. Mismatches in the non-primer region if the sequence downstream of the intended primer site and downstream of the mispriming site differ.
  - *Action to take:* Such reads should be removed from the read mapping before calling variants, since the reads are mapped to a different region than the one they are amplified from.

  See figure 6.3 for an example.
- **To the non-target region it represents** The read maps best to the non-target region it was generated from.

- *Symptom:* Mismatches in the primer region of the read (unless the mispriming event has 100% identity). No mismatches in the non-primer region (besides mismatches due to true variants).
- *Action to take:* Unalign the primer part of such reads in regions relevant for calling variants. Even if the primer part is 100% identical to the non-target region sequence, unaligning the primer part of these reads is important, as this allows the correct frequencies of variants in that region to be determined. If unaligning primer regions in this circumstance is desired, a track containing the regions where variant calling is of interest must be provided when launching this tool.

See figure 6.4 for an example.

A given primer can be involved in mispriming events leading to the amplification of reads that map to the original target region and to the region they were amplified from.

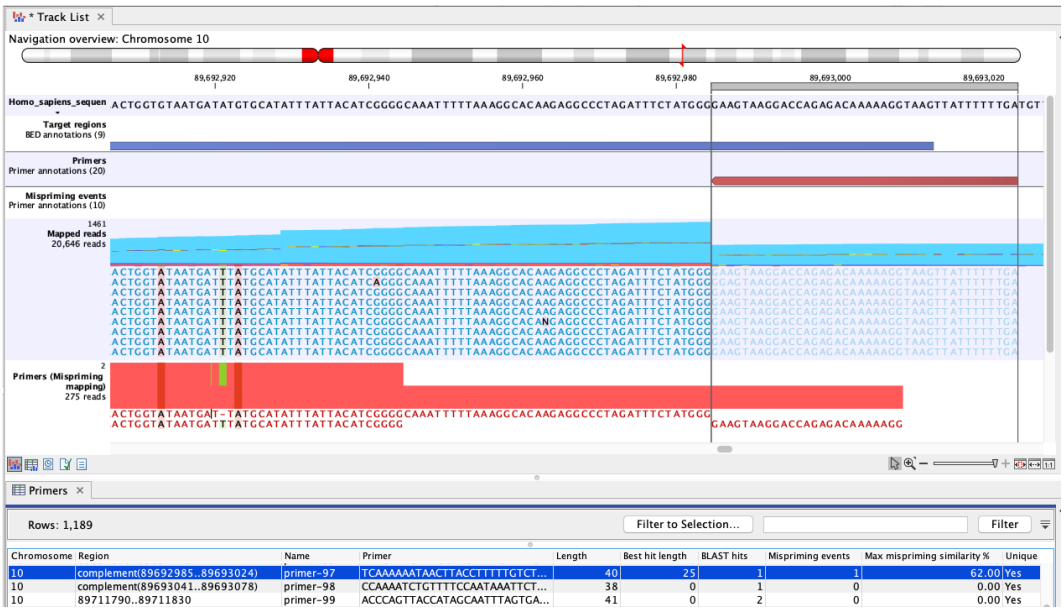


Figure 6.3: *An example of mispriming, where the reads map to the original intended target region. The two A variants and the single T variant, occuring in a non-primer part of the mapped reads, are consequences of mispriming. The reads supporting these variants should be removed from the mapping before variant detection is carried out. The reverse paired end reads (light blue) shown in the "Mapped reads" track were amplified from a mispriming binding site at chromosome 9 (not shown). While the primer had only 62% similarity with that site, the 3' primer end aligned perfectly, allowing it to anneal and for reads to be generated. Most of these reads mapped to the original target region, shown here, due to the low similarity of the primer region and high overall similarity with the intended target region.*

How the Identify Mispriming Events tool works

Identify Mispriming Events takes this approach:

1. A BLAST search is run using the primers as query sequences to search against a BLAST database of the relevant reference genome.

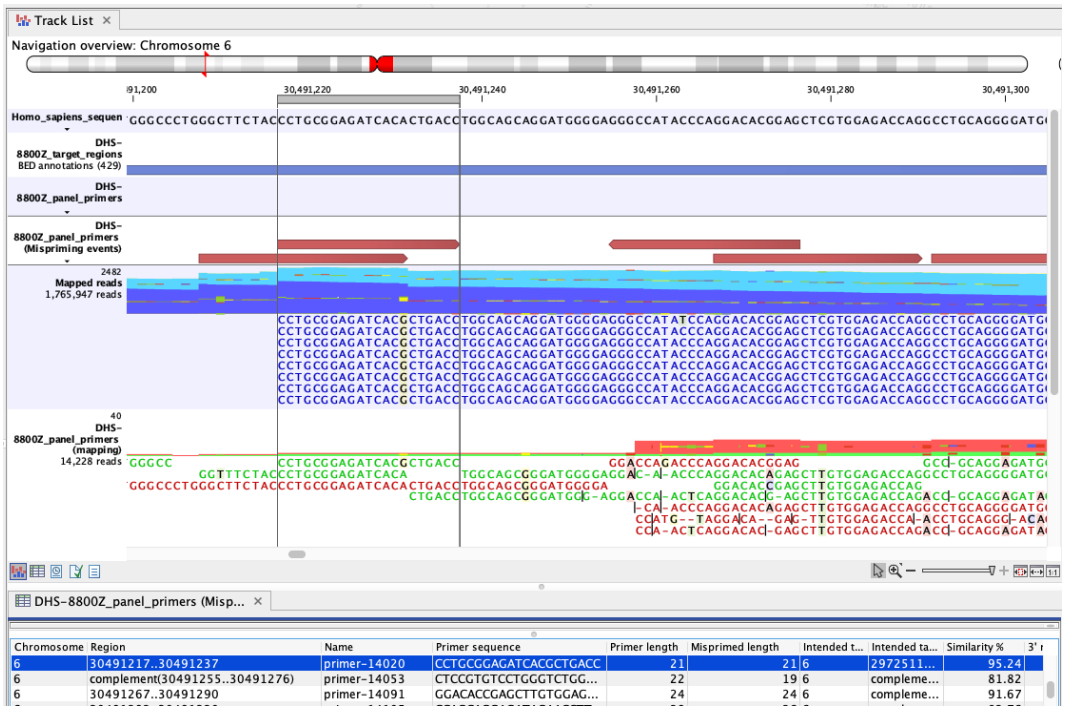


Figure 6.4: *An example of mispriming, where the reads map to the non-target region it represents. The A to G variant, found in the primer part of these forward, paired end reads (dark blue), is a consequence of mispriming. This primer was designed for a different region, but had 95.24% similarity with the region shown. Thus some copies of the primer annealed to this region and generate reads with a single mismatch, as shown in the "Mapped reads" track. The primer part of these reads should be unaligned, but the remaining part of the read can still be used for variant calling since it reflects the DNA fragments of the same genomic region.*

2. The BLAST hits returned are filtered. For each primer, hits are kept if that sequence has a high enough similarity to the intended target region and few mismatches at the 3' end.

3. The remaining BLAST hits for each primer are checked for their potential to cause mispriming artifacts of the two types mentioned above.

    - The sequence downstream of the intended target binding site is aligned to the sequence downstream of the mispriming site. If this pairwise alignment has a similarity fraction of at least 0.8, the BLAST hit is considered to be a mispriming event. The length of the sequence used of alignment can be changed using the parameter **Amplicon length (bp)**.
    - If a target region track is provided as input, and the mispriming region overlap a target region, the BLAST hit is considered a mispriming event.

    BLAST hits that are unlikely to represent sources of false positive variant calls are discarded.

Running the Identify Mispriming Events tool

To launch **Identify Mispriming Events**, go to:

Tools | **Biomedical Genomics Analysis** () | **Biomedical Utility Tools** () | **Identify Mispriming Events** ()

In the first dialog, select a primer track ( ) as input.

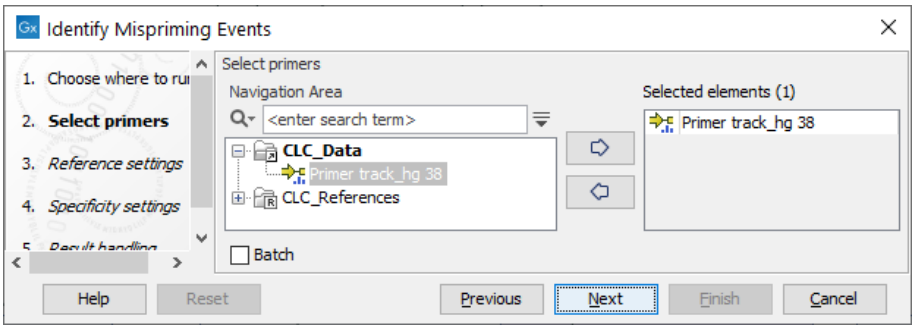


Figure 6.5: *Select a primer track.*

Settings related to the reference data are configured in the next wizard step (figure 6.6).

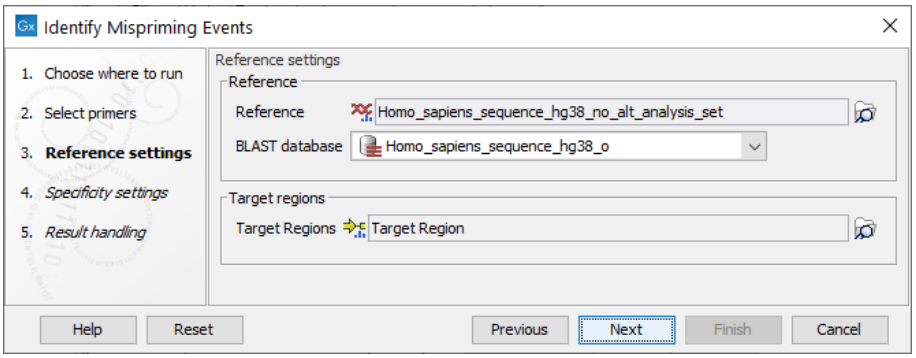


Figure 6.6: *Reference data settings for the Identify Mispriming Events tool.*

- **Reference**: A reference sequence track ( ) compatible with the selected primer track. Primer sequences are extracted from this reference. If a track is supplied in the "Target region track" field below, those regions are also extracted.
- **BLAST database**: A BLAST database of the selected reference genome.
- **Target regions**: An annotation track ( ) containing regions known to be of interest for variant calling. This track is required to obtain the type of mispriming events with mismatches in the primer region of the mapped read, where primer regions should be unaligned before variant calling.

Tip: Use **Create BLAST Database** if you do not already have a BLAST database of your reference genome (see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_local_BLAST_databases.html`). This tool takes a sequence list ( ) as input. If your reference genome is in a sequence track ( ), use **Convert from Tracks** to convert it to a sequence list ( ) before running **Create BLAST Database**.

Specificity settings are configured in the next wizard step (figure 6.7).

- **BLAST word size**: An exact match of at least this length between the primer and reference genome is required for BLAST to initiate an extension that might lead to a reported hit.

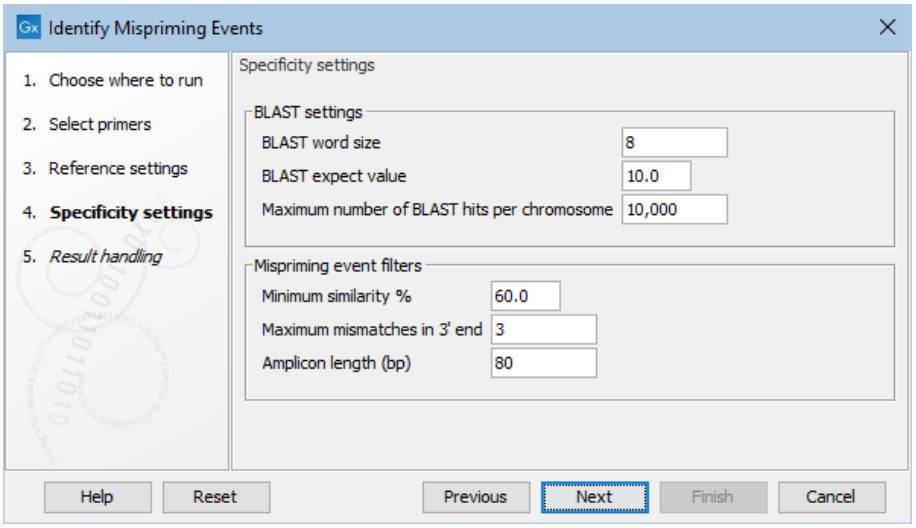


Figure 6.7: *The specificity settings of the Identify Mispriming Events tool. The default settings are a good starting point, but BLAST settings and/or Mispriming events filters can be adjusted to make the settings more relaxed or stringent.*

This value should be set based on the shortest primer in the primer set, and the word size should never be longer than half the length of the shortest primer. Increasing this value increases specificity, but too high a value could result in potential mispriming sites not being reported.

- **BLAST expect value**: Lower expect values are more stringent, leading to fewer chance matches being reported. However, virtually identical short alignments have relatively high values because of the way expect values are calculated. Raising this value can lead to more potential mispriming sites being identified, but can also increase the running time of the tool.

- **Maximum number of BLAST hits per chromosome**: The maximum number of BLAST hits to report per chromosome. Limiting the number of hits to return can decrease the time the running time of the tool and circumvent potential out-of-memory issues.

- **Minimum similarity %**: The minimum percentage similarity between a primer sequence and non-target regions in the reference genome for that region to be retained in the list of potential mispriming event sites.

- **Maximum mismatches in 3' end**: The number of mismatches allowed between a primer and a non-target region of the reference genome, counting from the 3' end of the primer. If this value is exceeded, the region is not retained in the list of potential mispriming event sites.

- **Amplicon length (bp)**: The length of sequence from downstream of the designed primer site and from downstream of the mispriming site that should be used to test if reads amplified from the mispriming site has the potential to map to the original, intended target region can cause false positive variant calls. If the pairwise alignment of these two regions has a similarity fraction of at least 0.8, the region is reported as a mispriming event.

6.3.1 Output from Identify Mispriming Events

Four outputs are produced from the **Identify Mispriming Events** tool:

- **Mispriming events**: An annotation track that can be used for **Trim Primers of Mapped Reads**. The track has a row for each mispriming event.
- **Primers**: An annotation track of primers annotated with different mispriming statistics.
- **Misprimed reads track**: A read mapping of reads representing mispriming events.
- **Report**: A report that summarizes the mispriming events identified by the tool.

Mispriming events

The mispriming events track includes the following annotations:

- **Primer sequence**: The sequence of the primer.
- **Primer length**: The length of the primer.
- **Misprimed length**: The length of the mispriming site where the primer sequence aligns.
- **Intended target chromosome**: The chromosome that the primer was designed for.
- **Intended target region**: The region that the primer was designed for.
- **Similarity %**: Similarity percentage between the primer sequence and the sequence of the mispriming site.
- **3' mismatches**: The number of nucleotide mismatches before the first match in the 3' primer end.
- **Primer part mismatch type**: Yes if the mispriming event potentially causes false positives in the primer part of the mapped read, otherwise No.
- **Non-primer part mismatch type**: Yes if the mispriming event potentially causes false positives in the non-primer part of the mapped read, otherwise No. Only evaluated if a target region track is provided.

Primers

The primer track includes the following annotations:

- **Length**: The length of the primer.
- **Best hit length**: The length of the BLAST hit with the highest similarity percentage.
- **BALST hits**: Number of filtered BLAST hits for this primer.
- **Mispriming events**: Number of mispriming events for this primer.
- **Mispriming events > 80%**: Number of mispriming events with a similarity percentage of at least 80% for this primer.

- **Mispriming events > 90%**: Number of mispriming events with a similarity percentage of at least 90% for this primer.
- **Mispriming events with non-primer part mismatches**: Number of mispriming events, originating from this primer, that potentially cause mismatches in the non-primer part of the mapped read.
- **Mispriming events with primer part mismatches**: Number of mispriming events, originating from this primer, that potentially causes mismatches in the primer part of the mapped read. Only evaluated if a target region track is provided.
- **Max mispriming similarity %**: Maximum similarity percentage among the identified mispriming events for this primer.
- **Unique primer**: Yes if the primer is unique in the reference genome, No if the primer has a 100% similarity match to another genomic region.

Misprimed reads track

Each mispriming event is represented by two reads in the read mapping: A read with the sequence of the primer aligned to the mispriming site, and a read with the sequence of the mispriming site aligned to the primer design region. Primers with multiple mispriming events will have a read for each mispriming event aligned at the primer design region.

For mispriming events that potentially cause false positives in the non-primer part of the read, two additional reads are included in the read mapping: A read with the downstream sequence of the primer aligned to the downstream of the mispriming site, and a read with the downstream sequence of the mispriming site aligned to the downstream of the primer design region. The mismatches in these reads show the potential false positive variants that can arise from mispriming.

Report

The mispriming event report includes the following information:

- **Summary**: A summary table showing the number of input primers and input target regions, as well as how many primers that have mispriming events and the types of potential false positives.
- **Primers with potential mispriming**: The section provides information about the primers for which one or more mispriming events have been found. The number of BLAST hits and the number of mispriming events for each primer are shown as distribution plots, as well as the maximum mispriming similarity percentage for these primers.
- **Mispriming events**: Different statistics about the BLAST hits and mispriming events identified by the tool.

6.4 Remove Ligation Artifacts

During the adapter ligation step of the library preparation, it can happen that two different DNA sequences also get ligated together. These ligation artifacts are more prone to occur between

short DNA fragments, such as the ones generated from FFPE samples. The tool **Remove Ligation Artifacts** removes reads which are likely the result of ligation artifacts. In addition, in cases of short fragments, a remnant of the common sequence can be found at the end of R1 reads. The tool will also remove these common sequence artifacts.

Remove Ligation Artifacts is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **Biomedical Utility Tools** ( ) | **Remove Ligation Artifacts** ( )

In the first dialog (figure 6.8), select a read mapping (it can also be an rna-seq read mapping).

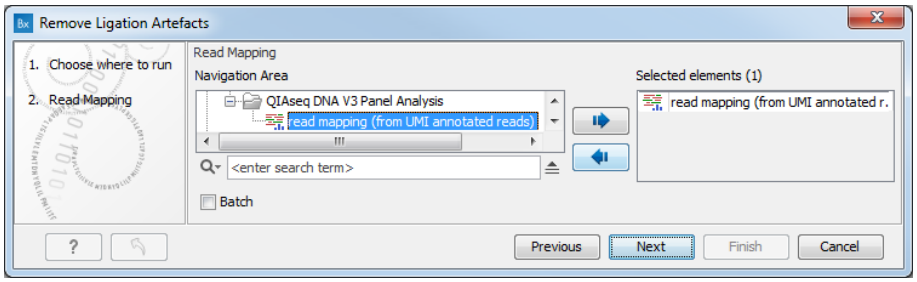


Figure 6.8: *Select a read mapping.*

In looking for **ligation artifact**, for each read:

- The tool looks at a window of a specific size (set by the option "Ligation artifact recognition length").
- The tool counts mismatches in the window. If there are less than 2 mismatches (value set by the "Minimum mismatches" parameter), the read is accepted. Note that any unaligned end counts as mismatches, i.e., if we have an unaligned end of size 3 that counts as 3 mismatches and the read will be subjected to the following steps.
- If there are at least 2 mismatches, the tool reverse complements the part of the read and tries to find a match within 250 bp on each side in the reference sequence.
- If a match is found, the read is deemed a ligation artifact and removed. It is possible to allow a single mismatch compared to the main sequence while still calling it a match ("Allow mismatch" under "Ligation artifacts").
- If the option "Remove entire Unique Molecular Index" is checked, all reads in a UMI group are removed if at least one ligation artifact read is found in the group.

In looking for **common sequence artifact**, for single reads and broken pairs:

- The tool looks at the 23 first and last bases (the window) in the read (defined by the "Full length common sequence search limit" parameter) and searches for the common sequence and the reverse complemented common sequence in the window. It is possible to allow a single mismatch between the common sequence and the read window and still call it a match (check "Allow mismatch" in the Common sequence artifacts section of the dialog).
- If no match is found, the tool searches for sub-strings down to a minimum of 4 bases ("Minimal partial common sequence length" parameter) on the read: When searching the

first bases of a read, the tool checks if suffixes of the common sequence match the start of the read. When searching for the last bases of a read, the tool checks if prefixes of the common sequence match the end of the read. It is here again possible to allow a single SNP in the common sequence sub-string and the read and still call them a match.

- The tool then counts mismatches in the window (the window is from the match (including the common sequence) out to the end of the read). If the percentage of mismatches in the window is less than 50% (defined by the "Minimum mismatch percentage" parameter), the read is accepted.
- If there are more than 50% mismatches in the window, the read is trimmed from the bases in the read in the window. Both unaligned and aligned bases will be removed. If there is just one, or if there are no aligned bases left after trimming, the read is removed.

In looking for **common sequence artifact** for paired reads, the tool will trim the overhang of the read that extends further than the beginning of the paired read carrying the UMI.

The setting options for the Remove Ligation Artifacts tools are as follow (figure 6.9).

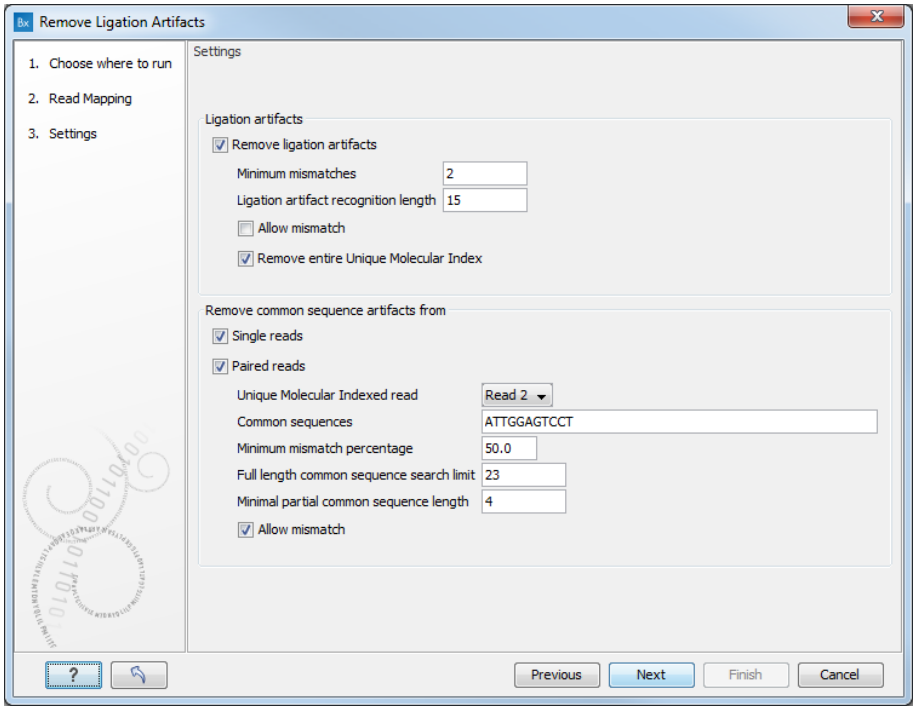


Figure 6.9: *Set the parameters for the remove Ligation Artifacts tool.*

- Ligation artifacts
  - **Remove ligation artifacts**: uncheck this option to keep ligation artifacts in your data.
  - **Minimum mismatches**: define the thresholds of mismatching characterizing a potential ligation artifact.
  - **Ligation artifact recognition length**: defines the size of the window being searched for mismatches.

  - **Allow mismatch**: checking this option will allow a single mismatch between the sequence window and the main sequence while still calling it a match.
  - **Remove entire Unique Molecular Index**: remove all reads in a UMI group if at least one ligation artifact read is found in the group.

- Remove common sequence artifacts from
  - **Single reads**
  - **Paired reads**
  - **Unique Molecular Indexed read**: can be set to Read 1 or Read 2 if you wish to restrict the removal of the common sequence artifacts to only one read in the pair.
  - **Common sequences** defined by the QIAseq DNA Panel kit. It can be one or several sequences separated by commas.
  - **Minimum mismatches percentage**: defines the thresholds of mismatching characterizing a potential common sequence artifact.
  - **Full length common sequence search limit**: size of the sequence window in which the tool will search for the common sequence.
  - **Minimal partial common sequence length**: size of a sub-string to look for matches between the beginning and the end of a read and the common sequence.
  - **Allow mismatch**: allows a single mismatch between the sequence window and the read sequence while still calling it a match.

Click **Next** to choose to **Open** or **Save** the tool output, i.e., the read mapping where the ligation and sequences artifacts have been removed. It is also possible to generate a read mapping containing the ligation artifacts, and a report.

6.5 Trim Primers of Mapped Reads

The tool **Trim Primers of Mapped Reads** removes the primer parts of mapped reads, as they reflect the primer that was added and not the actual sample. Note that the tool will also remove any insertion located immediately after the primer. The tool also removes artifacts due to mispriming events, happening when a primer binds an off-target location and thus amplifies an off-target sequence.

Trim Primers of Mapped Reads is available under the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Biomedical Utility Tools** () | **Trim Primers of Mapped Reads** ()

In the first dialog (figure 6.10), select a read mapping.

In the second dialog (figure 6.11), select the primer annotation track that was provided with the QIAseq Panel. This track contains the original primers and their intended primer locations.

In addition, set the following parameters:

- Primer location
  - **Default**: primers are at the end of single-end reads, or at the start of read 1 for paired-end reads.

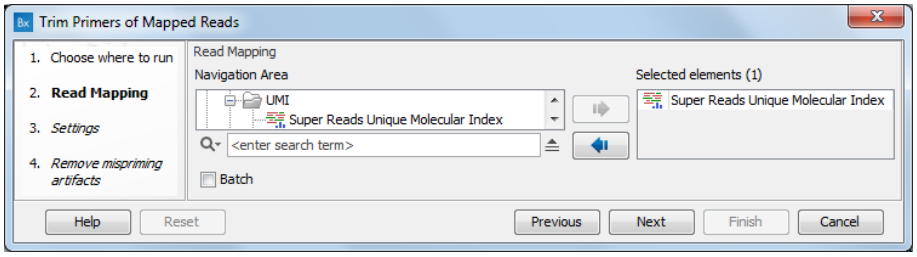


Figure 6.10: *Select a read mapping.*

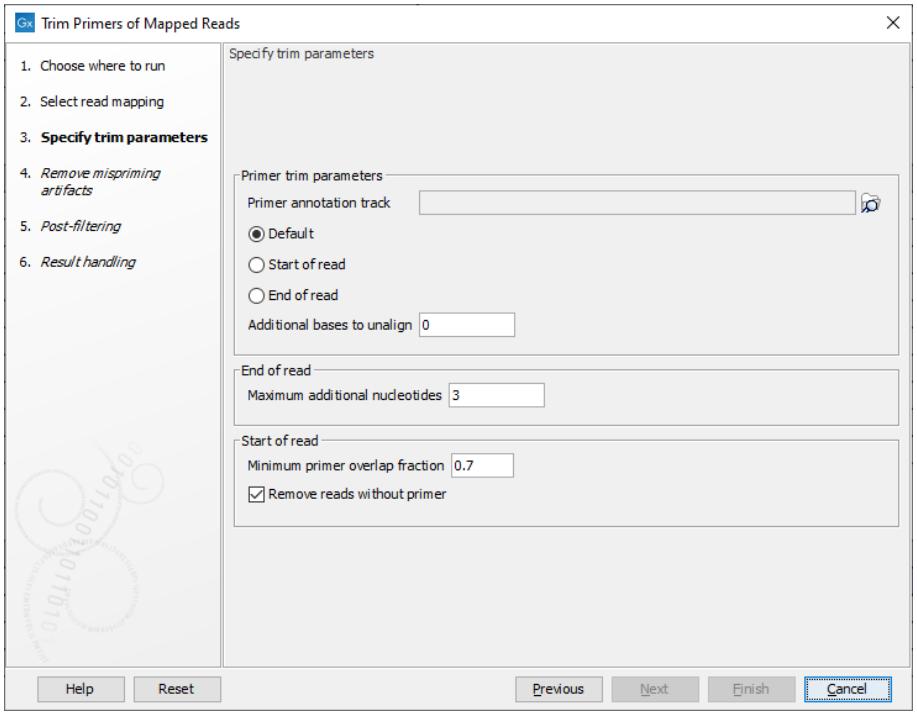


Figure 6.11: *Select the primer annotation track specific to the panel, and add the parameters needed to deal with the type of reads you are working with.*

  - **Start of read**: primers are at the start of single-end reads, or at the start of read 1 for paired-end reads.
  - **End of read**: primers are at the end of single-end reads. This option is not supported for paired-end reads.
  - **Start of read 2**: primers are at the start of read 2 of paired-end reads. This option is not supported for single-end reads.

- Decide on how many **Additional bases to unalign** immediately after the primer. This trimming is not done on reads for which dimer artifacts are identified.
- **Maximum additional nucleotides**: When trimming primers from the end of single-end reads, unalign reads that end up to this number of extra bases after the primer.
- **Minimum primer overlap fraction**: If an aligned read starts within the span of a primer, and if it overlaps at least this percentage of the primer, then it is said to "hit" the primer. For reads "hitting" a primer, the part of the read that overlaps the primer will be unaligned.

- **Remove reads without primer** When enabled, reads not "hitting" a primer and not coming from broken pairs will be removed from the output mapping. Broken pair reads are retained to help visualize genomic rearrangements. Note that it is possible to later remove broken pairs from the output mapping by running the tool Extract Reads with the option "Include reads from broken pairs" deselected.

If one read in a UMI group runs past the primer it overlaps, it means that all reads in that group were not created from that primer. If this happens, then the tool will not unalign any reads in this UMI group.

In the Remove mispriming artifacts dialog (figure 6.12), you can specify a Primer mispriming events track containing the predicted off-target priming locations of the original primers. Mispriming tracks can be generated by the tool **Identify Mispriming Events** (see section 6.3) and are available for each QIAseq panel from the Reference Data Manager.

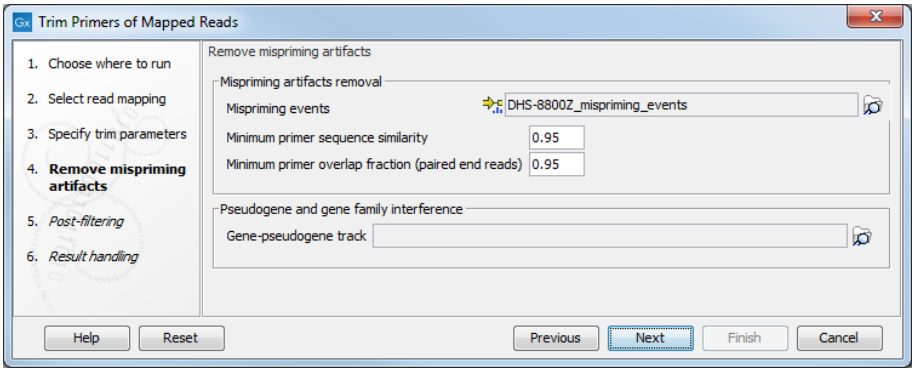


Figure 6.12: *Select a mispriming track specific to the panel, and configure the associated parameters if needed.*

In addition, set the following parameters:

- Mispriming artifacts removal. The tool will unalign primer part of misprimed reads only if the misprimer overlapping part of the read has at least the **Minimal primer sequence similarity** fraction to the original primer sequence. For paired end reads, the tool will unalign primer of misprimed paired end reads only if the primer part has at least the **Minimal primer overlap** fraction with the off-target primer.
- Pseudogene and gene family interference. It is possible to specify a gene-pseudogene track that contains gene and pseudogene links information for removing reads that map well to pseudogene locations.

In the Post-filtering dialog (figure 6.13), when the "Remove short reads" option is enabled, reads with an alignment length shorter than the value specified after primer trimming will be removed from the mapping.

The tool will output a trimmed read mapping. In addition, as seen in figure 6.14, you can choose to output off-target primers and/or regions tracks with counts.

If a mispriming track was specified in the earlier step, it is possible to output two additional tracks containing respectively region and primer mispriming statistics:

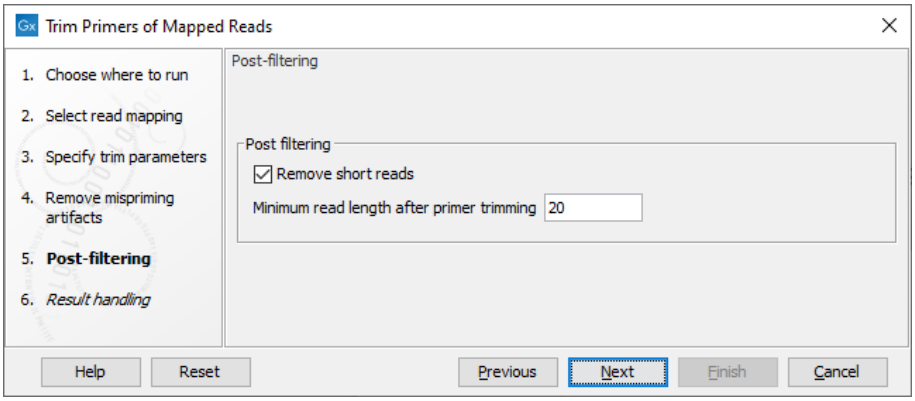


Figure 6.13: *Post-filtering parameters.*

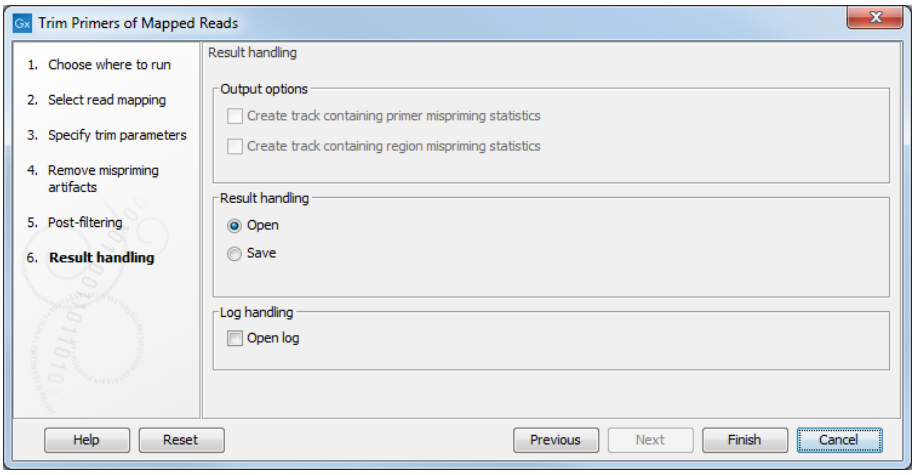


Figure 6.14: *Outputs options.*

- The primer mispriming statistics track (figure 6.15) includes mispriming events and the counts of read unalignment those events caused in a process of mispriming artifact removal.

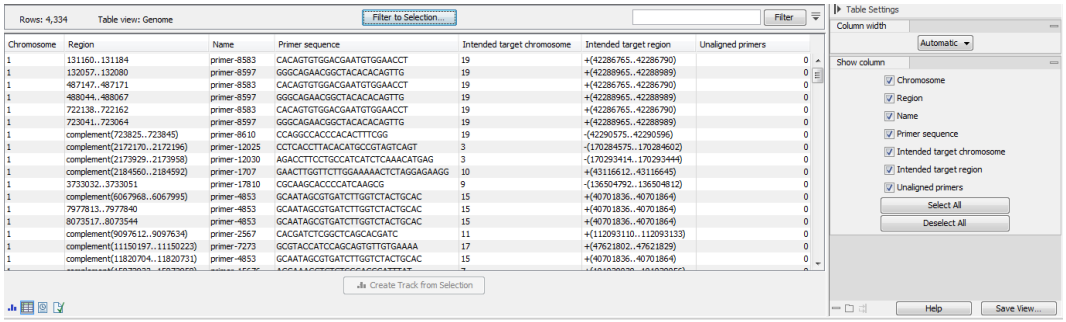


Figure 6.15: *A primer mispriming track.*

  - **Chromosome**, **Region** and **Name** referring to the identified mispriming locations.
  - **Intended target chromosome** and
  - **Intended region** referring to the true location of the primer.

  - **Unaligned primers** Total number of reads supporting the mispriming event. These reads are amplified from a mispriming event, and the primer part of the read has been unaligned because it matched the primer sequence and region. The primer part of the read might contain mismatches if the primer doesn't match the reference exactly. It is only unaligned if both the sequence and region overlap is sufficient, as determined by the two input parameters, "Minimum primer sequence similarity" and "Minimum primer overlap fraction (paired end reads)". For single end reads, only the first parameter is applicable.

- The regions mispriming statistics track (figure 6.16) includes regions that are considered as alternative mapping locations when attempting to remove mismapped reads. Primary target locations in this track are annotated with:

  - **Name** of the primer
  - **Primary target** True or false depending if the region describes the primer or a possible mispriming event.
  - **Reads checked** Total number of reads checked for better matches to a mispriming event region or to a position in the vicinity of the original mapping location.
  - **Reads matching mispriming location** Number of reads mapping with a better score to a mispriming region than to the original location after the primer sequence had been unaligned. These reads are removed from the read mapping.
  - **Reads matching target vicinity** Number of reads with a better match to a position in the vicinity of the original mapping location after the primer sequence had been unaligned. These reads are removed from the read mapping.

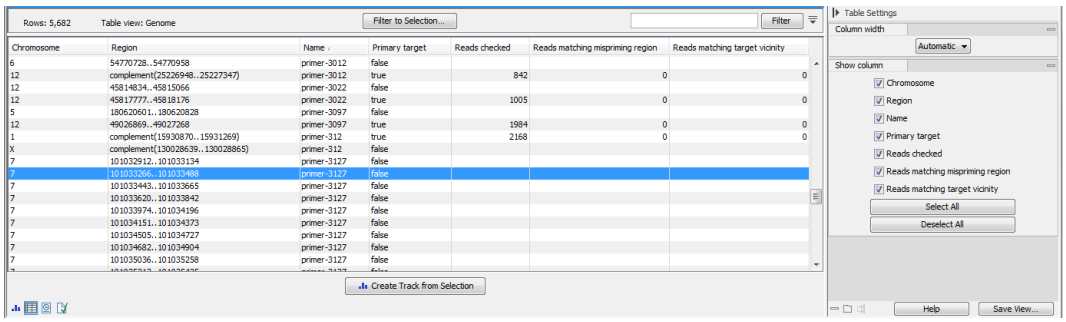


Figure 6.16: *A region mispriming track.*

6.6 Convert Annotation Track Coordinates

Convert Annotation Track Coordinates converts annotation coordinates, either from hg19 coordinates to hg38 coordinates, or vice versa. This remapping of coordinates, also referred to as 'liftover', makes use of the UCSC Lift Genome Annotations service (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`). Network access is therefore necessary to use this tool.

The assemblies used for remapping are hg19 (Feb. 2009 (GRCh37/hg19)) and hg38 (Dec. 2013 (GRCh38/hg38)).

Limitations:

- It is possible to remap annotation coordinates to and from the reference sequence hg38_no_alt_analysis_set, but annotations cannot be remapped to or from the alternative contigs that are included in hg38_no_alt_analysis_set.
- The tool can only convert files that are under 500Mb.

The UCSC Lift Genome Annotations service provides a list of options that can be adjusted. Of those, only **Minimum ratio of bases that must remap** is available in the **Convert Annotation Track Coordinates** tool. The remaining options are set to default values:

- BED 4 to BED 6 Options
  - Allow multiple output regions - Checked
  - Minimum hit size in query - 0
  - Minimum chain size in target - 0
- BED 12 Options
  - Minimum ratio of alignment blocks or exons that must map - 1
  - If thickStart/thickEnd is not mapped, use the closest mapped base - unchecked

For alternative settings, use the UCSC Lift Genome Annotations service directly. Note that CLC BED import/export, does not support thickStart/thickEnd.

To run **Convert Annotation Coordinates Track**, go to:

> **Tools** | **Biomedical Genomics Analysis** () | **Biomedical Utility Tools** () | **Convert Annotation Track Coordinates** ()

The tool takes annotation tracks, typically primer tracks or target region tracks, as input (see figure 6.17).

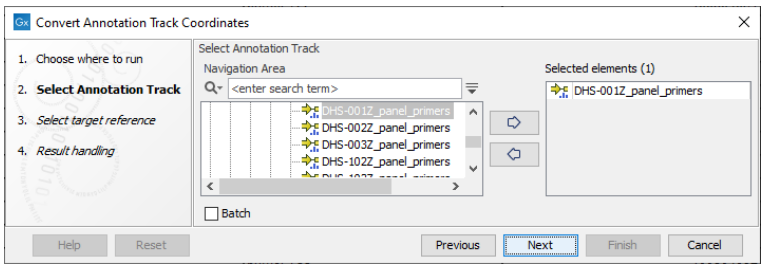


Figure 6.17: *Selection of a primer annotation track with hg19 coordinates for conversion to hg38 coordinates.*

In the next step, a target reference must be selected (see figure 6.18). If the annotation track used as input was based on an hg19 reference, select an hg38 reference genome. If the annotation track used as input was based on an hg38 reference, select an hg19 reference genome. If not already available, a copy of the relevant target reference can be downloaded using the Reference Data Manager.

Under settings you can adjust **Minimum ratio of bases that must remap**. This is the minimum fraction of bases in a region that must be directly aligned in gapless blocks from the first genome to the second.

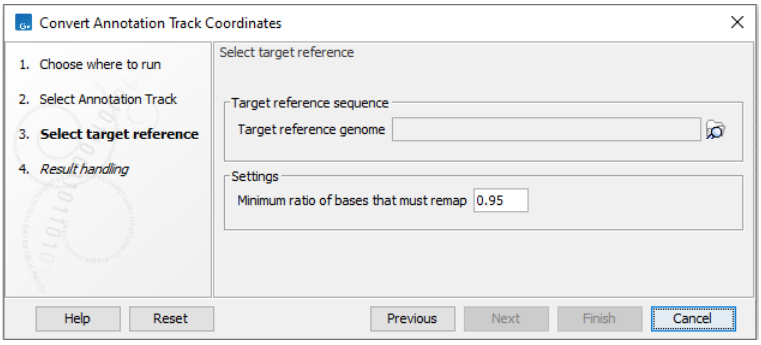


Figure 6.18: *This dialog allows selection of the reference sequence you would like to lift over to. Under settings you can choose to adjust the minimum ratio of bases that must remap.*

Convert Annotation Track Coordinates outputs an annotation track containing remapped annotations and a report. The report lists the skipped intervals, that is, the annotations that could not be remapped, together with a comment provided by UCSC (figure 6.19).

1 Coordinate Conversion

Converted coordinates from genome hg38 (Dec. 2013 (GRCh38/hg38)) to genome hg19 (Feb. 2009 (GRCh37/hg19)).

Converted coordinates for 191,994 out of 192,264 annotations.

Skipped intervals:

Chromosome	Name	Interval	Message
1	SEC22B	120160382..120160532	Partially deleted in new
1	SEC22B	120163208..120163372	Partially deleted in new
1	SEC22B	120168838..120168951	Partially deleted in new
1	PPIAL4A	120889819..120890317	Partially deleted in new
1	FCGR1B	121088652..121088910	Partially deleted in new

Figure 6.19: *The Convert Annotation Track Coordinates report lists skipped intervals.*

Chapter 7

Immune repertoire analysis

Contents

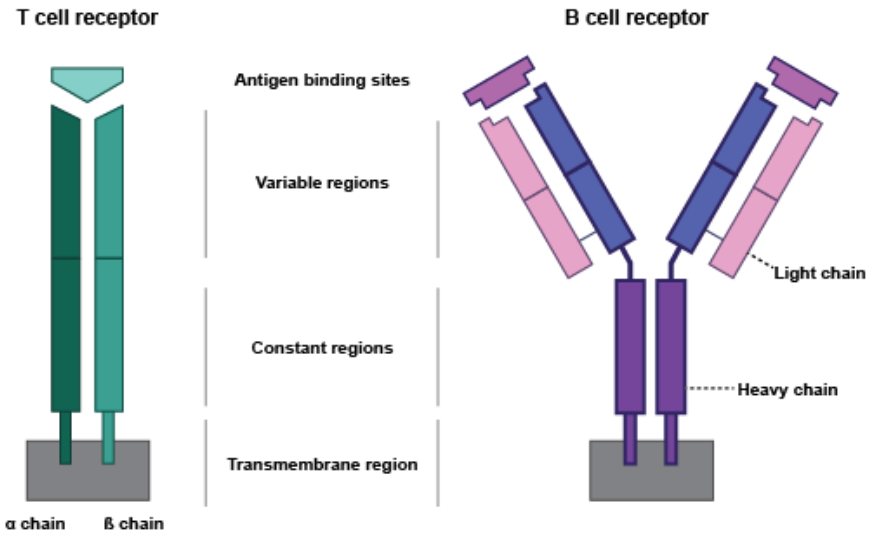


Figure 7.1: *T and B cell receptors structure. Each pair of chains forms an antigen binding site that binds to specific antigens.*

T and B cells form our acquired immune response. They both contain highly variable receptors (see figure 7.1) with binding sites that recognize antigens.

T and B cell receptors (TCR and BCR, respectively) are composed of multiple polypeptide chains: TCRs contain one pair, while BCRs contain two copies of a pair:

- TCR: α (TRA) and β (TRB), or γ (TRG) and δ (TRD).
- BCR: light and heavy. There are two types of light chains in humans: κ (IGK) and λ (IGL), while other animals also contain other types of light chains. Once set, the light chain class remains fixed for the life of the B cell. There are five types of heavy chains (IGH) for mammals: γ, δ, α, μ and η, defining the class of the receptor.

The chains are encoded by genes that undergo somatic recombination. During this process, gene segments are joined with random nucleotides at the junction sites. There are two types of recombination (see figure 7.2):

- VJ recombination, where one V (variable) gene segment is joined to a J (joining) gene segment;
- VDJ recombination, where a D (diversity) gene segment is added between the V and J gene segments.

For both types of recombination, a C (constant) gene segment is also added following the J segment.

The TRA and TRG chains are the result of VJ recombination, while the TRB and TRD chains are the result of VDJ recombination. The V, D, J and C gene segments are specific for each TCR chain type.

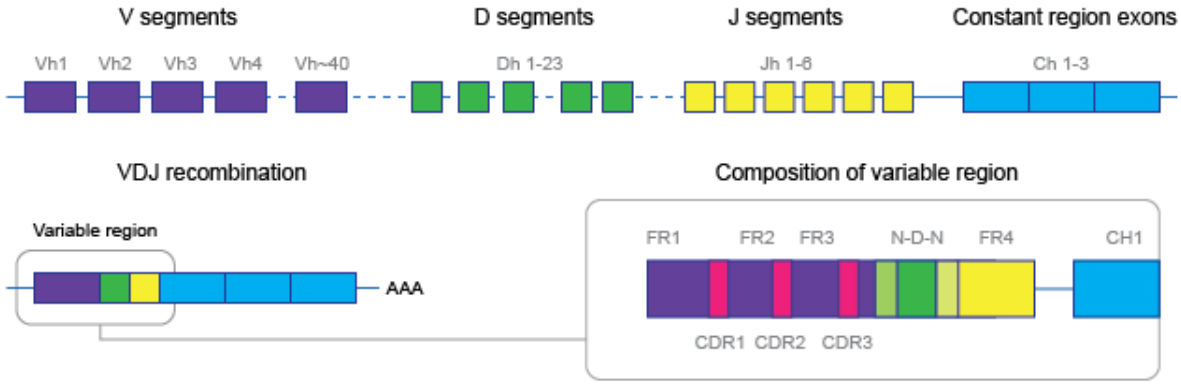


Figure 7.2: *VDJ recombination brings together a V, D, J and C gene segment.*

BCR light chains are the result of VJ recombination, while BCR heavy chains are the result of VDJ recombination. BCR heavy chains have three to four C gene segments. The V, D, J and C gene segments are specific for each BCR light chain type, while they are shared by the BCR heavy chains.

Each chain contains three "complementary-determining regions" (CDRs) (figure 7.2) which form loops in the antigen binding sites. The V(D)J recombination junction is located in the third CDR (CDR3). Due to inclusion of random nucleotides at the junctions between segments, the CDR3 is the most diverse among the three CDRs. Its beginning and end are marked by a conserved cysteine (C) and phenylalanine/tryptophan (F/W) amino-acid in the V and J segments, respectively.

Biomedical Genomics Analysis offers tools to clonotype reads and characterize the T or B cell receptor repertoire (section 7.3), filter the repertoires (section 7.5) and compare them (section 7.6). Here, clonotyping a read consists of identifying which V, D, J and C segments from the reference data (see section 7.3) are used, and extracting the CDR3 region found between the conserved amino acids.

7.1 Import/Export VDJtools Clonotypes

The `VDJtools format` is a simple clonotype text file format of the VDJtools software package [Shugay et al., 2015].

The file format is tab-separated. The first line of the file must start with a '#' and describe the contents of the columns. An example first line is:

```
#count freq cdr3nt cdr3aa v d j VEnd DStart DEnd JStart
```

Import and export from and to this format is supported for both **TCR clonotypes** ( ) and **BCR clonotypes** ( ) (see section 7.7).

Note that, if the license allows use within your organization, VDJtools can be used to convert several other clonotype formats into this format.

Import

To import this format, use the Standard Import tool:

Import () | Standard Import ()

Select `VDJtools Clonotypes (.txt)` in the **Force import as type** menu of the dialog.

The following columns are read from the file:

- "count", "freq", "cdr3nt", and "cdr3aa". These must be present in the file.
- "v", "d", "j", and "c". At least one of these must be present in the file.

All other columns are ignored during import.

For non-coding CDR3 sequences, the "cdr3aa" column may represent an incomplete codon using either "~" or "_".

The importer produces a TCR clonotypes element and / or BCR clonotypes element, depending on the data found in the file.

Export

To export clonotypes to a text file, launch the Export tool by clicking on the Export button in the toolbar, or going to the **File** menu and choosing **Export...**. To easily locate the "VDJ Tools" exporter, type "vdj" into the search field at the top of the tool.

7.2 Import Immune Reference Segments

The **Import Immune Reference Segments** tool can import reference sequences for V, D, J and C segments from a fasta file. The sequences are needed when running **Immune Repertoire Analysis** for either T or B cell receptor repertoires (TCR and BCR, respectively), see section 7.3.

The importer can be found here:

Import () | Import Immune Reference Segments () .

The importer can be used to import fasta files that are either in the IMSEQ [Kuchenbecker et al., 2015] or IMGT [Lefranc et al., 2009] format (see figure 7.3).

Both formats support allele numbering for the gene segments. If **Import only the first allele** is ticked, only segments without an allele or those with an allele defined as the number "1" (i.e "01" is also valid) will be imported. Otherwise, all segments are imported.

The two formats differ in how the sequence header is parsed for identifying the gene segment and related information, and how the conserved amino acids in the V and J segments are identified.

When saving the results, the reference data for either TCR, or BCR, or both, can be saved. The wizard will show an error message if an output option is ticked for which no relevant reference sequences are available.

The importer can only handle one fasta file at a time, but if two or more fasta files are imported, the resulting sequence lists can subsequently be combined to one list using the **Create Sequence List** tool.

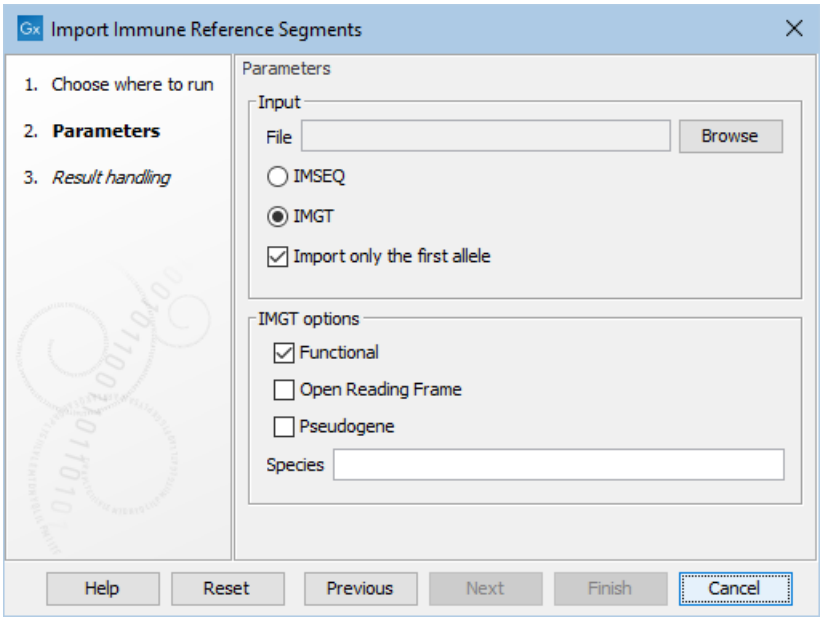


Figure 7.3: *The available options when importing immune reference segments.*

7.2.1 IMSEQ

For the IMSEQ format, the header contains the following elements, separated by "|":

- The chain: TRA, TRB, TRG or TRD for T cells, and IGH, IGK and IGL for B cells.
- The segment type: V, D, J, C.
- The segment ID. For B cells constant genes, the segment ID should also contain the letter corresponding to the encoded isotype.
- The segment allele.
- For J and V segments, the position of the first base of the conserved amino acid, counting from 0.

Currently only the heavy (IGH) and light κ and λ (IGK and IGL) chain types are supported for B cells.

Any segments with an unsupported chain or segment type are silently ignored.

7.2.2 IMGT

For the IMGT format, the header contains 15 elements, separated by "|". Only the following are read and used during import:

- (1) Accession number(s).
- (2) The segment name, including chain, segment type, ID and allele, in the format: `<chain><type><ID>*<allele>`, for example "TRAV1*01".

  Chain and segment type are the same as for IMSEQ. For B cells constant genes, the segment type contains instead the letter corresponding to the encoded isotype.

- (3) Species.
- (4) Allele functionality: F (functional), P (pseudogene) or ORF (open reading frame).
- (5) Extracted label(s): EX1, CH1 and C-REGION for C segments, and V-REGION, D-REGION, J-REGION for V, D, J segments, respectively.
- (8) The start of the codon, counting from 1, or "NR" for non coding labels.
- (9) The number of nucleotides added in 5' in the format `+n`.

The IMGT database contains chains, segment types and labels that are not listed above and are not supported. These are silently ignored.

While the IMSEQ format provides the position of the conserved amino acid, this needs to be calculated for the IMGT format. For this, the V region needs to be provided with gaps such that the conserved amino acid is found at approximately position 104 in the translated amino acid sequence. When downloading sequences from the IMGT database in fasta format, the "F+ORF+in-frame P nucleotide sequences with IMGT gaps" should be used. Alternatively, the corresponding "nt-WithGaps-F+ORF+inframeP" flat file can be downloaded from IMGT/GENE-DB.

If using custom reference data that is not downloaded from the IMGT database, it is recommended to use the IMSEQ format and specify the position of the conserved amino acid.

When importing files in the IMGT format, the following options are available (see figure 7.3):

- Which allele functionality(ies) should be imported. At least one must be chosen.
- Which species should be imported. After choosing the fasta file, the desired species can be chosen from the list of species identified in the file.

If element (9) in the header is not empty, the corresponding number of nucleotides are removed from the 5' end of the sequence.

Identification of the conserved amino acid

The nucleotide sequence (with IMGT gaps for the V segments), starting from position in element (8) in the header, is first translated to amino acids using the standard genetic code. The position of the conserved amino acid is calculated, and, if identified, translated to the position of the first nucleotide in the corresponding codon. Segments where the amino acid cannot be identified are silently ignored.

For the V segments, the amino acid position is calculated as follows:

- If the amino acid at position 104 is C, then position 104 is used.
- Otherwise, the position of the last occurrence of C after position 104 is used, if present.
- Otherwise, if the amino acid at position 104, 105 or 103 (in this order) is one base pair mutation away from C and not a stop codon (i.e. R, S, C, F, G, W, Y), then this position is used.

For the J segments, all 3 open reading frames (starting from nucleotide position 1, 2 or 3) are used. Note that "." below denotes any amino acid. The amino acid position is calculated as follows:

- The amino acid sequence "(F|W)G.G", if present, is identified.
  - The open reading frame that contains the amino acid sequence, no stop codon and has the lowest nucleotide starting position, if any, is used.
  - Otherwise, the open reading frame that contains the amino acid sequence and at least one stop codon, if any, is used. If multiple open reading frames match this criteria, none are used.
- Otherwise, the amino acid sequences "(F|W)X.G" and "(F|W)G.X", if present, are identified. Here, X denotes the amino acids that are one base pair mutation away from F/W and not a stop codon (i.e. A, R, S, C, D, E, V, W).
  - For each of the two amino acid sequences, the position is calculated as above.
  - If both amino acid sequences are present, the position that is closest to the end of the sequence is used.

V and J segments for which the amino acid position cannot be successfully identified are silently ignored.

7.2.3 Output from Import Immune Reference Segments

The importer outputs a sequence list that can be used for immune repertoire analysis. These can be added to a custom reference data set, to be used in workflows. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html for details.

The sequence list contains the reference sequences for the V, D, J and C segments, named in the format <chain>-<type>-<ID>*<allele>, for example "TRA-V-1*01". Note that for B cells constant genes, the letter corresponding to the encoded isotype will be used instead of the segment type.

If the gene segment does not have an allele or **Import only the first allele** is ticked, *<allele> is not added to the name.

By ticking "Show annotations" and "Region" in the Side Panel "Annotation layout" and "Annotation types" groups, respectively, the location of the conserved amino acid can be visualized (see figure 7.4).

The table view of the sequence list shows the chain and segment type of each sequence, and for the IMGT format, also the accession number(s) and species (see figure 7.5).

Figure 7.4: *Visualizing the location of the conserved amino acid.*

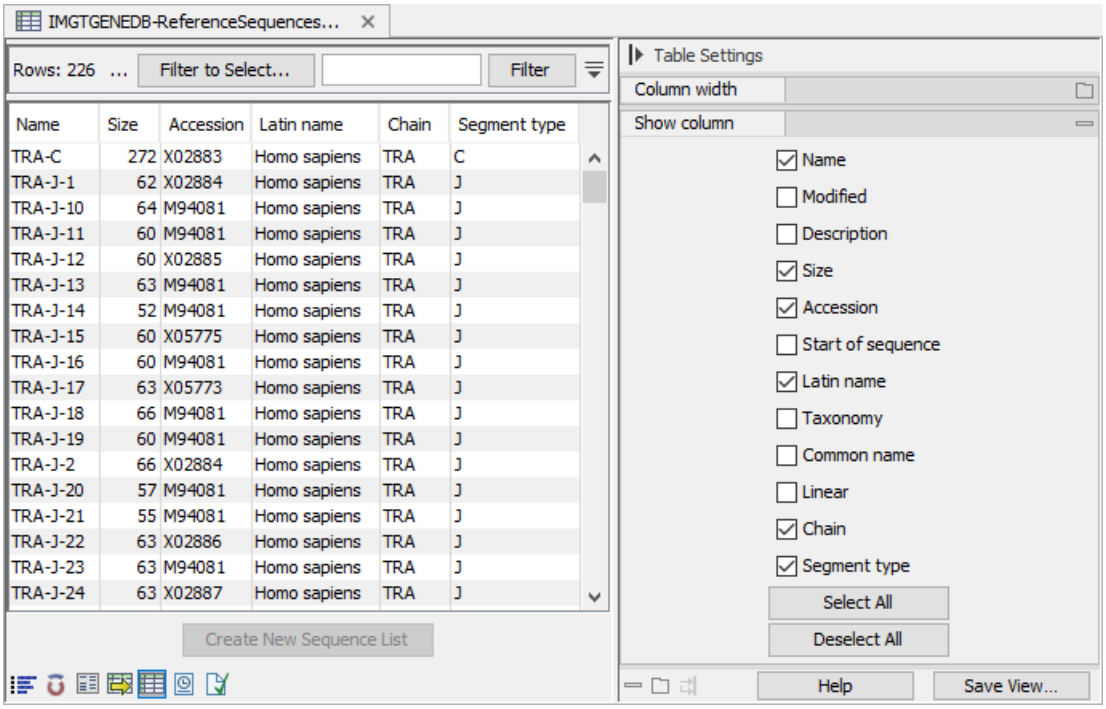


Figure 7.5: *Table view of imported sequence list showing the name, species and accession number when imported using the IMGT format.*

7.3 Immune Repertoire Analysis

Using RNA-Seq data as input, the **Immune Repertoire Analysis** tool can be used to characterize either the T or B cell receptor repertoire.

The tool requires a reference data sequence list ( ) containing reference sequences for the V, D, J and C segments.

Whether the tool identifies T or B cell receptors depends on the types of reference segments present in the provided sequence list. The tool does not accept sequence lists containing

reference sequences for both TCR and BCR.

The **Reference Data Manager** (see chapter 3) offers two QIAGEN sets for this tool. Each set contains a sequences list for **Immune Repertoire Analysis**:

- **QIAseq Immune Repertoire Analysis** for analysis of TCR human data.
- **QIAseq Immune Repertoire Analysis Mouse** for analysis of TCR mouse data.

If reference data is needed for BCR or for a different species than those above, **Import Immune Reference Segments** can be used to import reference data, see section 7.2.

The tool assumes that one read spans all segment types (V, D, J and C) in order to successfully report the clonotype. It is therefore recommended to collapse overlapping paired-end reads using **Merge Overlapping Pairs**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Merge_Overlapping_Pairs.html.

Identification of clonotypes

Clonotyping a read consists of identifying which V, D, J and C segments from the reference data are used and extracting the CDR3 sequence found between the conserved amino acids.

V and C segments are rather long (> 200 bp), whereas J segments are relatively short ($\approx 50 - 70$ bp) and D segments are even shorter ($\approx 10 - 30$ bp). The segments identification is therefore performed using different strategies.

First, the tool identifies the V and J segments. These segments are required for successfully clonotyping a read, because otherwise the CDR3 cannot be determined.

For V segments, the **Map Reads to Reference** tool is used internally.

For the J segment, a strategy similar to IMSEQ [Kuchenbecker et al., 2015] is used. First, a pairwise alignment with a 15 bp subsequence of the full segment called a Segment Core Fragment (SCF) is performed to find candidates for full pairwise alignments. If the pairwise alignment of an SCF to the read has a sufficiently small number of errors, it is nominated as a candidate. A full pairwise alignment is then made for all the segments corresponding to the candidate SCFs. If there is a sufficiently good match among the full alignments it will be assigned to the read.

Once both V and J segments are identified, only valid matches are kept:

- the V and J segments are for the same chain;
- the J segment is located on the read after the V segment.

The D and C segments are then identified for the reads with assigned V and J segments. These segments are optional.

For D segments, a local alignment is performed between the region of the read found between V and J, and the reference D segments for the same chain.

For C segments, the **Map Reads to Reference** tool is used internally. As the C segment is long and not variable, matches for the C segment for chains other than that identified for V and J indicate a false positive and the read is hence discarded.

Read length variability and sequencing errors can lead to one clonotype being reported as two separate clonotypes. See section 7.4 for details on how to merge such clonotypes into one single clonotype.

Running the tool

Immune Repertoire Analysis is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **Immune Repertoire Analysis** ( ) | **Immune Repertoire Analysis** ( )

This opens a dialog where the reads can be selected. The following options can be configured (figure 7.6):

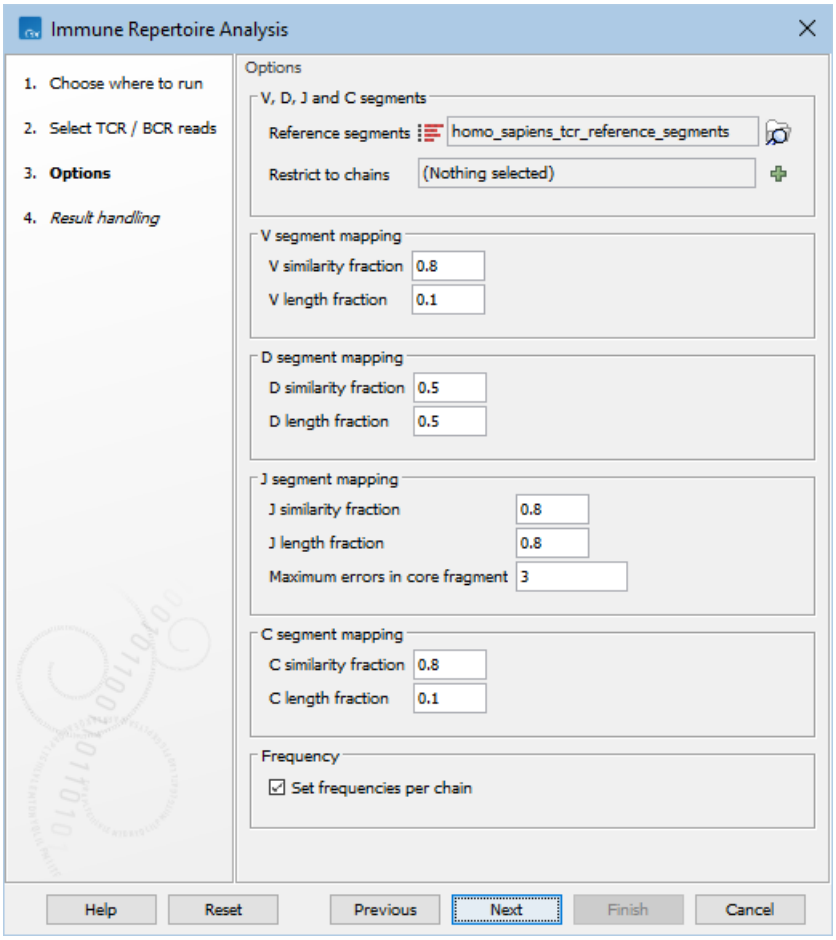


Figure 7.6: *Options for Immune Repertoire Analysis.*

- **Reference segments**. A sequence list containing the V, D, J and C segments, either from the reference data or imported using **Import Immune Reference Segments**, see section 7.2.
- **Restrict to chains**. A combination of 'TRA', 'TRB', 'TRG' and 'TRD' for TCR reference

segments, or 'IGH', 'IGK' and 'IGL' for BCR reference segments, can be chosen. Only clonotypes for the selected chains will be identified. If left empty, all chains will be used.

It can be useful to set this option when only specific chains have been sequenced, to remove false positives.

- **V / D / J / C similarity fraction**. Minimum identity fraction between the aligned region of the read and the segment.
- **V / D / J / C length fraction**. Minimum fraction of the segment that must match the read.
- **Maximum errors in core fragment**. Maximum number of errors allowed in the Segment Core Fragment (SCF) used for finding segment candidates for full pairwise alignment.
- **Set frequencies per chain**. Clonotype frequencies (see section 7.7.1) are calculated such that they add up to 100% across all chains. If **Set frequencies per chain** is ticked, the frequencies are instead calculated such that they add up to 100% for each individual chain.

The optimal values for the **Similarity fraction** and **Length fraction** are different for the different segment types.

As the V and C segments are at the ends of the read, they might not be covered entirely and the length fraction is expected to be considerably smaller than 1. On the other hand, the length fraction would typically be close to 1 for J. For V, J and C, the similarity fraction is usually close to 1 as not a lot of mutations are expected in these segments.

As the D segment is located in a region of high variability, both the similarity and length fractions would typically be lower to account for the high mutation rate.

7.3.1 Output from the Immune Repertoire Analysis tool

Immune Repertoire Analysis produces a **TCR clonotypes** () or **BCR clonotypes** () element, containing the identified clonotypes, see section 7.7 for more details.

The tool optionally outputs a **T cell receptor (TCR)** or **B cell receptor (BCR) report** () summarizing statistics of the detected clonotypes. The report includes the following information:

- **Summary** A summary table containing the number of
  - input reads;
  - input fragments, where each read pair is counted as one;
  - fragments for which a V / J / D / C segment was identified;
  - fragments that were successfully clonotyped;
  - clonotyped fragments per chain type;
  - unique clonotypes.

The identification of the segments is performed in the following order:

- Identification is performed first for the V segment.

- Identification is performed for the J segment for fragments where a V segment was identified.
- Identification is performed for the D and C segments for fragments where both a V and J segment was identified.

Therefore, the reported number of fragments for which a J / D / C segment was identified is bounded by the number of fragments from a previous identification step.

The remaining information in the report is given per chain type and only for those chain types for which clonotypes have been identified.

Note. All plots can be opened in table view by double-clicking on the plot and clicking on the table icon in the lower left corner.

- **Diversity indices**. Table containing diversity indices. As it is likely that some rare clonotypes are missing in the sequencing data, the extrapolated diversity indices give a projection of how many additional clonotypes there are and what the diversity would have been if the sample had been sequenced deep enough to capture all clonotypes.
  - Total number: The number of different clonotypes detected.
  - Extrapolated diversity (chaoE): The extrapolated number of detected clonotypes by the method described in [Chao, 1987].
  - Lorenz curve at 50% of total: The percent of clonotypes that account for 50% of the total fragment count. Also sometimes denoted as D50.
  - Inverse Simpson's index: Let c_i denote the fragment count for the ith clonotype and let $n = \sum_i c_i$. Then the inverse Simpon's index is defined as:

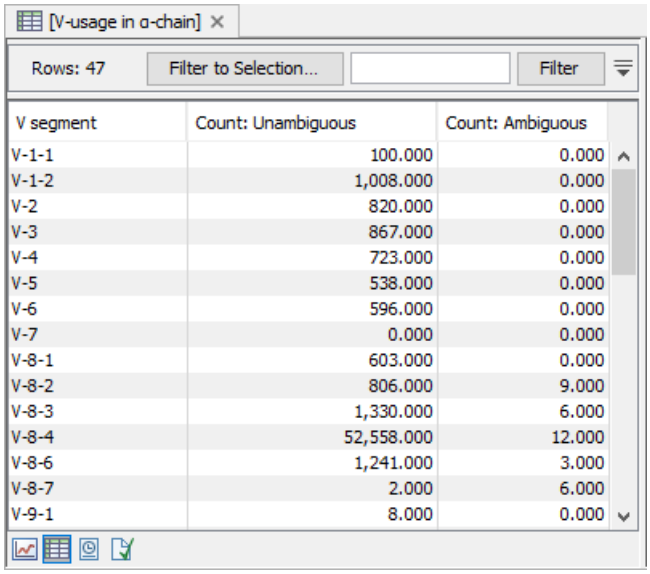
$$\sum_i \frac{1}{c_i/n}$$

  - Extrapolated Inverse Simpson's index (chaoE): The index by the method described in [Chao et al., 2014].
  - Shannon-Wiener index: With c_i and n defined as above, the Shannon-Wiener index is defined as:

$$\sum_i \frac{c_i}{n} \log\left(\frac{c_i}{n}\right)$$

    Note that the logarithm is the natural logarithm. To convert to base 2 logarithm the index can be multiplied by $\log_2(e) \approx 1.443$
  - Extrapolated Shannon-Wiener index (chaoE): The index by the method described in [Chao et al., 2013].
- **Rarefaction**. A plot for each chain with the rarefaction curve. See section 7.7.4 for more details.
- **CDR3 length**. A plot for each chain showing the length distribution of the CDR3 nucleotide sequences. See section 7.7.5 for more details.

- **V, D, J and C usage**. Histograms for each chain and segment type showing the frequency of each of the detected segments. See section 7.7.6 for more details. Double clicking the plot opens it in a new window. A table view can be selected from the bottom pane, showing counts for all segments, see figure 7.7.



V segment	Count: Unambiguous	Count: Ambiguous
V-1-1	100.000	0.000
V-1-2	1,008.000	0.000
V-2	820.000	0.000
V-3	867.000	0.000
V-4	723.000	0.000
V-5	538.000	0.000
V-6	596.000	0.000
V-7	0.000	0.000
V-8-1	603.000	0.000
V-8-2	806.000	9.000
V-8-3	1,330.000	6.000
V-8-4	52,558.000	12.000
V-8-6	1,241.000	3.000
V-8-7	2.000	6.000
V-9-1	8.000	0.000

Figure 7.7: *A full table showing segment usage can be obtained by double clicking the segment usage plot and selecting table view.*

- **Cumulative frequencies of clonotypes**. A plot for each chain showing the cumulative frequencies of the identified clonotypes. See section 7.7.7 for more details.
- **Productive summary**. The percentage distribution of CDR3 nucleotide sequences that are productive, out-of-frame or contain a premature stop codon

7.4 Merge Immune Repertoire

The **Merge Immune Repertoire** tool can be used for reducing false positives due to sequencing errors and variability in read quality and length:

- Read quality and length variability can impact the identified segments.
  - Reference segments may have a large degree of sequence identity due to recent duplication events [Glusman et al., 2001]. In order to uniquely identify a segment, reads need to be sufficiently long to cover the regions where paralogue segments differ. Shorter reads may lead to clonotypes containing multiple (ambiguous) segments.
  - Identification of the C segment requires reads that are sufficiently long. Shorter reads will be reported without an identified C segment.
- Sequencing errors in the CDR3 region can lead a highly expressed clonotype to be reported as multiple clonotypes.

Merge Immune Repertoire is available from the Tools menu at:

Tools | **Biomedical Genomics Analysis** ( ) | **Immune Repertoire Analysis** ( ) | **Merge Immune Repertoire** ( )

This opens a dialog where a **TCR clonotypes** ( ) or **BCR clonotypes** ( ) element can be selected. The following options can be configured (figure 7.8):

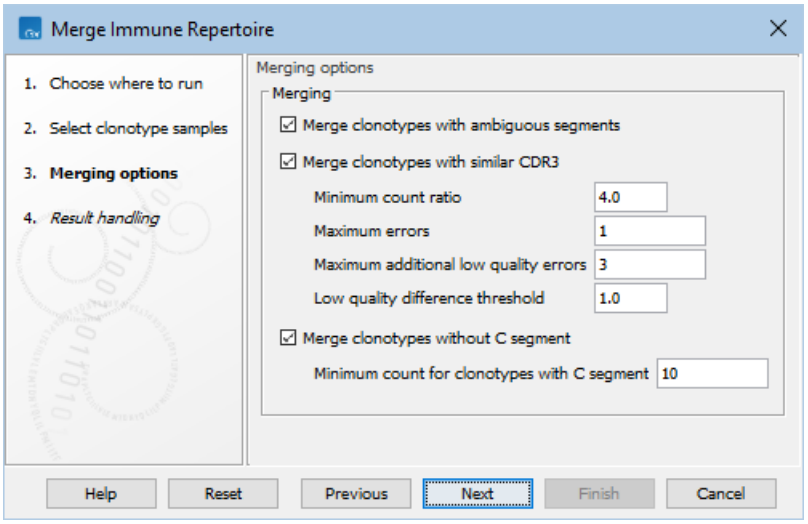


Figure 7.8: *Options for Merge Immune Repertoire.*

- **Merge clonotypes with ambiguous segments**. If selected, merge clonotypes with compatible V, J and C segments and the same CDR3 nucleotide sequence, where one clonotype has a unique segment and the other has ambiguous segments that include the former clonotype's segment.

  If two clonotypes are merged, the unique segment is preserved in the merged clonotype, regardless of the counts of the two clonotypes.

  Note that using this option can lead to some reads not being included in the alignments view. See section 7.7.2 for details.

- **Merge clonotypes with similar CDR3**. If selected, merge clonotypes with the same identified V, J, and C segments and with similar CDR3 nucleotide sequences. As the D segment is found within the CDR3, clonotypes are not required to have the same identified D segment.

  If two clonotypes are merged, the CDR3 sequence and identified D segment of the larger clonotype are preserved in the merged clonotype.

  - **Minimum count ratio**. A smaller clonotype is merged into a larger clonotype if the count of the larger clonotype is at least this number of times larger than the count of the smaller clonotype.

    E.g. if the minimum count ratio is 4 and a clonotype has count 8, only clonotypes with a count of at most 2 (8 / 4 = 2) will be considered for merging.
  - **Maximum errors**. Two clonotypes will be considered for merging if there are at most this many differences between their CDR3 sequences.
  - **Maximum additional low quality errors**. Two clonotypes where the number of differences between their CDR3 sequences exceeds **Maximum errors** can still be

considered for merging, if the number of additional errors at positions with low quality in the smaller clonotype does not exceed this number.

  - **Low quality difference threshold**. A position is considered of low quality if the average quality is more than this number of standard deviations lower than the average quality at each position in the CDR3 sequence.

    Note that **Maximum additional low quality errors** and **Low quality difference threshold** have no effect if the CDR3 quality scores are not available, see section 7.7.1.

- **Merge clonotypes without C segment**. If selected, merge clonotypes with the same identified V, J, and D segments and the same CDR3 nucleotide sequences, where one clonotype has an identified C segment and the other one does not.

  If two clonotypes are merged, the identified C segment is preserved in the merged clonotype.

  - **Minimum count for clonotypes with C segment**. A smaller clonotype with a C segment is merged with a larger clonotype without a C segment if the count of the smaller clonotype is at least this number.

    Note that a smaller clonotype without a C segment is always merged with a larger clonotype with a C segment.

7.4.1 Merging of clonotypes

The three merging options can be used together.

The following is an example of using both **Merge clonotypes with similar CDR3** and **Merge clonotypes without C segment**. Let us consider two clonotypes with the same identified V and J segments, sufficiently similar CDR3 nucleotide sequences, and where only one has an identified C segment:

- If both options are selected, the clonotypes will be merged.
- If only **Merge clonotypes with similar CDR3** is selected, the clonotypes will not be merged since they have different identified segments.
- If only **Merge clonotypes without C segment** is selected, the clonotypes will not be merged since they have different CDR3 sequences.

Multiple clonotypes with similar CDR3 sequences can be merged in different ways, depending on the values set for the different options. Consider the example from figure 7.9:

- The CDR3 sequences of clonotypes 1 and 2 differ by one nucleotide.
- The CDR3 sequences of clonotypes 2 and 3 differ by one nucleotide at a different position, such that the CDR3 sequences of clonotypes 1 and 3 differ by two nucleotides.
- The counts for clonotypes 1, 2 and 3 are 24, 12 and 3, respectively.

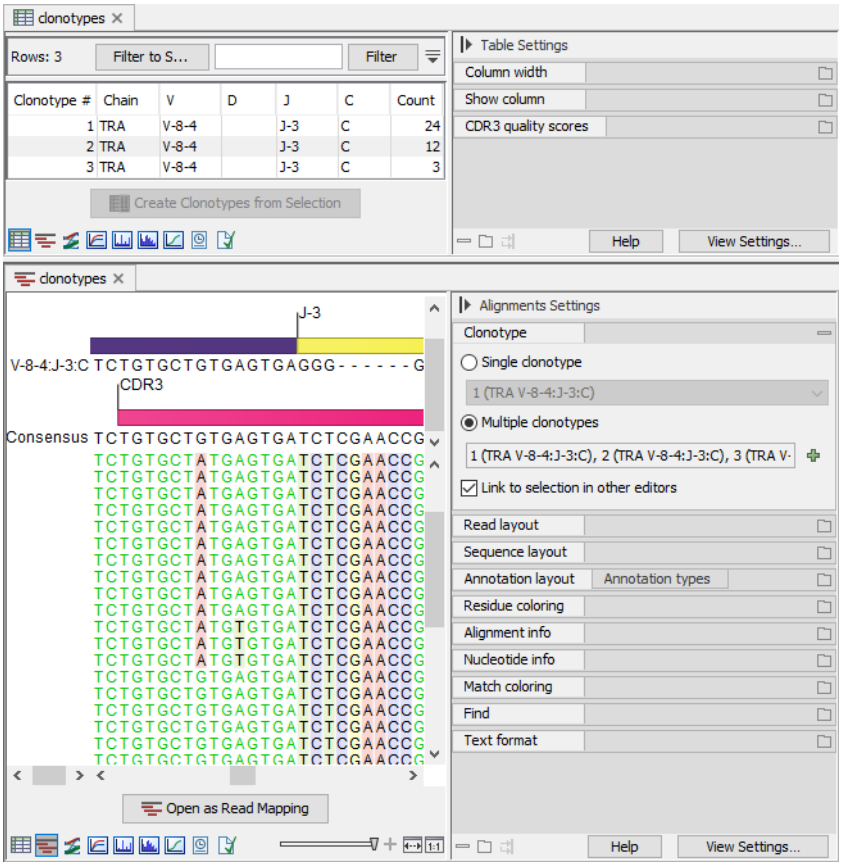


Figure 7.9: *Three clonotypes sharing the identified segments and with similar CDR3 sequences. Top: Clonotypes table view. Bottom: Clonotypes alignment view for all three clonotypes, highlighting the differences in the CDR3 sequence.*

Consecutive merges

Clonotypes merging is performed consecutively and hence multiple clonotypes may be merged into one. Setting **Minimum count ratio** to 1.5 and **Maximum errors** to 1:

- Clonotype 3 is merged into clonotype 2, leading to clonotype 2 having a count of 15.
- Clonotype 2 is merged into clonotype 1, because 24 / 15 > 1.5. Since clonotype 3 has already been merged into clonotype 2, this effectively also merges clonotype 3 into clonotype 1, even though the number of differences between their CDR3 sequences (2) exceeds **Maximum errors**.

Merges into multiple clonotypes

A clonotype may be merged into multiple clonotypes. Setting **Minimum count ratio** to 4 and **Maximum errors** to 2, clonotype 3 is merged into both clonotypes 1 and 2, and its count is distributed between clonotypes 1 and 2, proportional to their respective counts:

- Clonotype 2 receives 1 (3 x 12 / (12 + 24)), for a total count of 13.
- Clonogype 1 receives 2 (3 x 24 / (12 + 24)), for a total count of 26.

Clonotype 2 is not merged into clonotype 1 because 26 / 13 < 4.

Identical due to preceding merging

Two originally different clonotypes can end up being identical due to preceding merging, and then they will be merged into one. Consider the example from figure 7.10:

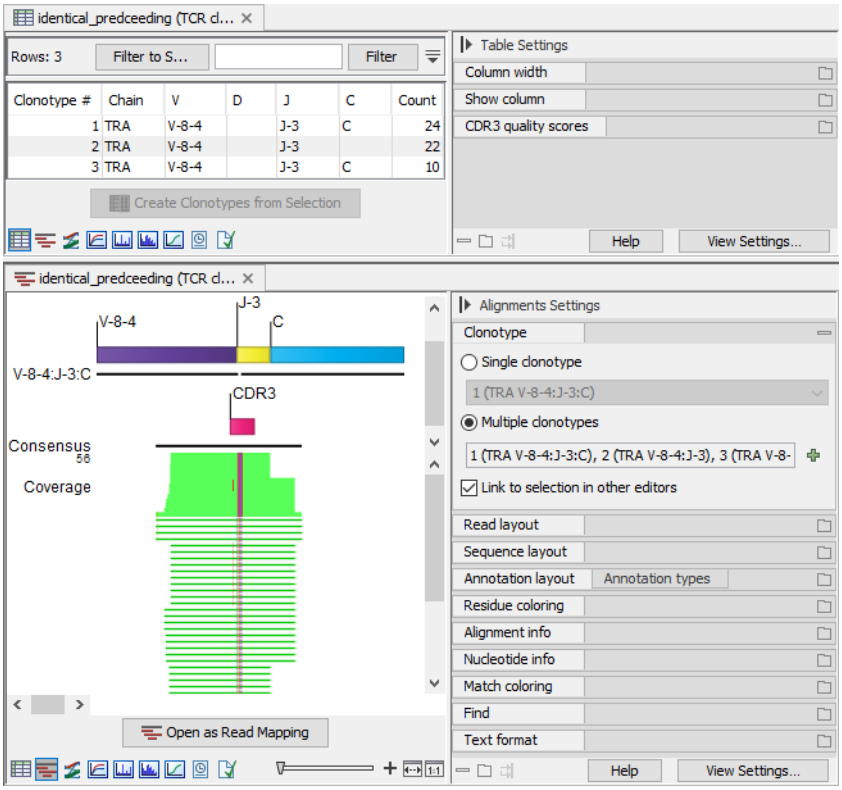


Figure 7.10: *Three clonotypes sharing the identified segments, with similar CDR3 sequences, and where one is missing the C segment. Top: Clonotypes table view. Bottom: Clonotypes alignment view for all three clonotypes, highlighting the differences in the CDR3 sequence and the shorter reads not covering the C segment.*

- Clonotypes 1 and 2 have the same CDR3 sequence, but only clonotype 1 has a C segment.
- Clonotype 3 has a CDR3 sequence that is sufficiently similar to that of clonotypes 1 and 2, and the same C segment as clonotype 1.
- The counts for clonotypes 1, 2 and 3 are 24, 22 and 10, respectively.

Running the tool with both **Merge clonotypes with similar CDR3** and **Merge clonotypes without C segment**, setting **Minimum count ratio** to 2, **Maximum errors** to 1 and **Minimum count for clonotypes with C segment** to 10:

- Clonotype 3 is merged into:
  - Clonotype 1 due to similar CDR3.

  - Clonotype 2 due to similar CDR3 and missing C segment.

- Clonotype 2 now has a C segment and is identical to clonotype 1 due to the preceding merging. The clonotypes are therefore merged.

7.4.2 Output from the Merge Immune Repertoire tool

Merge Immune Repertoire produces a **TCR clonotypes** () or **BCR clonotypes** () element, containing the merged clonotypes.

The tool optionally outputs a report () containing a summary of the performed merging, followed by statistics of the merged clonotypes, see section 7.3.1 for details.

The merging summary shows how many clonotypes were merged and the reason for merging:

- **Similar CDR3**. Clonotypes were merged according to **Merge clonotypes with similar CDR3**.
- **Without C segment**. Clonotypes were merged according to **Merge clonotypes without C segment**.
- **Similar CDR3 and without C segment**. Clonotypes were merged according to both **Merge clonotypes with similar CDR3** and **Merge clonotypes without C segment**.
- **Identical**. Clonotypes had the same identified segments and the same CDR3 nucleotide sequence.

  The Immune Repertoire Analysis tool (section 7.3) reports such clonotypes as only one clonotype. However, multiple but identical clonotypes can possibly be present in imported clonotypes (see section 7.1). Identical clonotypes are always merged by the tool, regardless of the chosen options.
- **Identical due to preceding merging**. Two originally different clonotypes became identical due to preceding merging steps. See section 7.4.1 for details.
- **Multiple merge events**. The same clonotype might be merged into several clonotypes for different reasons. For example, if a clonotype is merged into two different clonotypes because of similar CDR3, this clonotype will count towards **Similar CDR3**. If another clonotype is merged into two different clonotypes, one because of similar CDR3 and the other because of missing C, this clonotype will count towards both **Similar CDR3**, **Without C segment** and **Multiple merge events**.

7.5 Filter Immune Repertoire

The **Filter Immune Repertoire** tool can be used to restrict clonotypes to only a specific subset, for example, only productive clonotypes, or clonotypes with a specific chain. Alternatively, clonotypes can be filtered by creating a new element from a selection in the clonotypes table (section 7.7.1) or the clonotype sample comparison table (section 7.8.1).

Filter Immune Repertoire is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Immune Repertoire Analysis** () | **Filter Immune Repertoire** ()

This opens a dialog where a single **TCR clonotypes** (), **BCR clonotypes** (), or **Clonotype Sample Comparison** () element can be selected. The following filtering options can then be configured (see figures 7.11, 7.12, and 7.13). Note that the filters are applied independently.

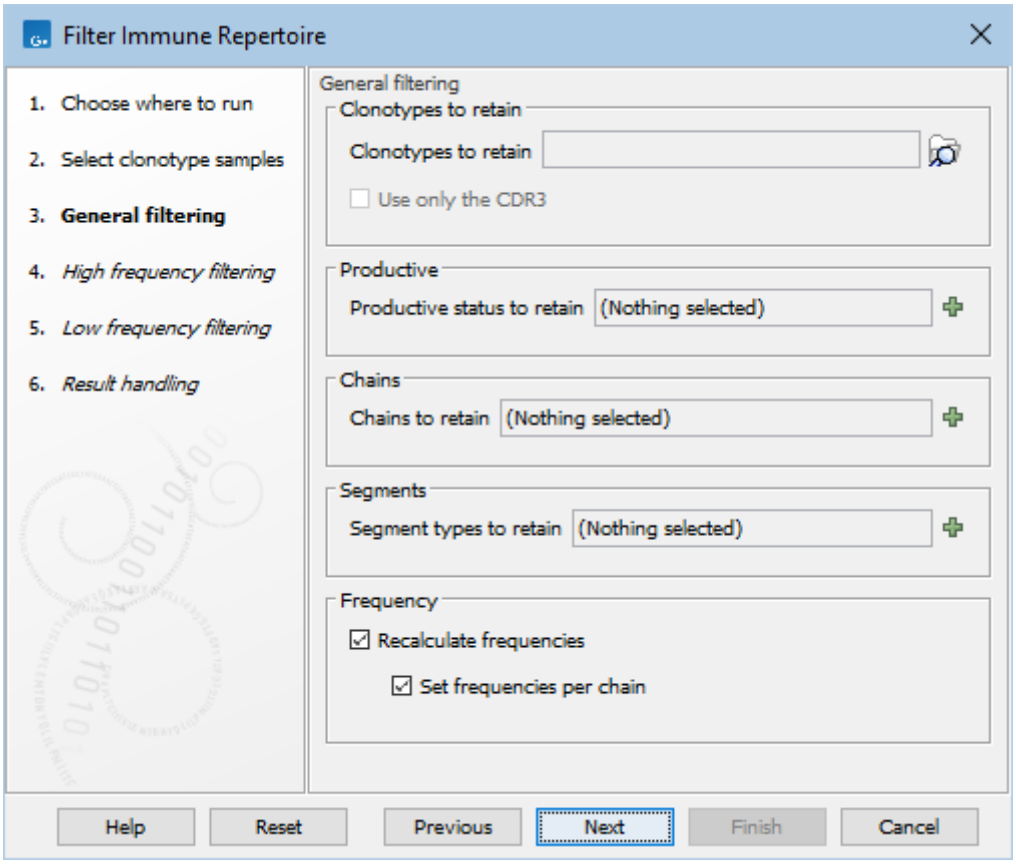


Figure 7.11: *General filtering options for Filter Immune Repertoire.*

- **Clonotypes to retain**. Retain clonotypes that are found in all provided **TCR clonotypes** (), **BCR clonotypes** (), or **Clonotype Sample Comparison** () elements. If left empty, no filter is applied.
- **Use only the CDR3**. When comparing the clonotypes in the input with those in the elements from **Clonotypes to retain**, only the CDR3 is used if this is ticked. Otherwise, the V and J segments together with the CDR3 are used to determine if two clonotypes are the same.
- **Productive status to retain**. A combination of 'Productive', 'Out of frame' and 'Premature stop codon' can be chosen and only the clonotypes with the respective productive status will be retained. If left empty, no filter is applied.
- **Chains to retain**. A combination of 'TRA', 'TRB', 'TRG' and 'TRD' for TCR data, or 'IGH', 'IGK' and 'IGL' for BCR data, can be chosen and only the clonotypes with the respective chains will be retained. If left empty, no filter is applied.
- **Segment types to retain**. A combination of 'V', 'D', 'J' and 'C' can be chosen and only the clonotypes that have identified segments for all respective segment types will be retained. This means that, for example, if 'D' is chosen, only chains for which the D segment is used

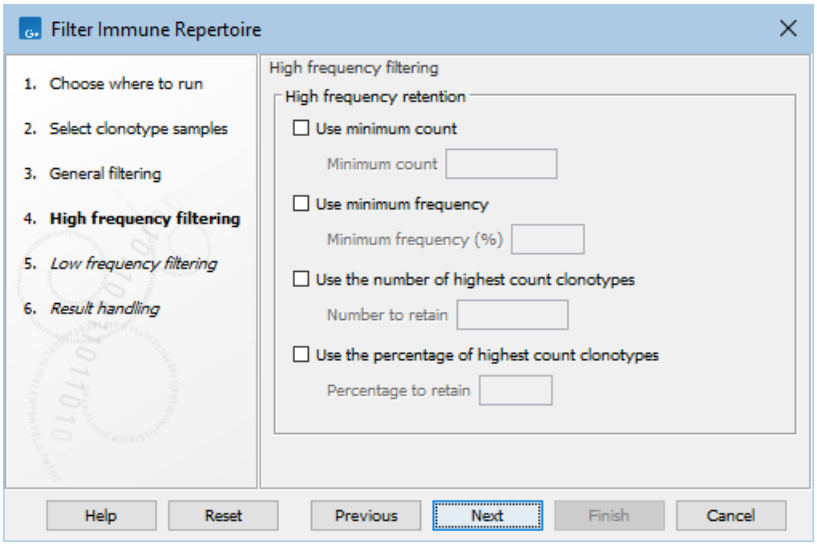


Figure 7.12: *High frequency filtering options for Filter Immune Repertoire.*

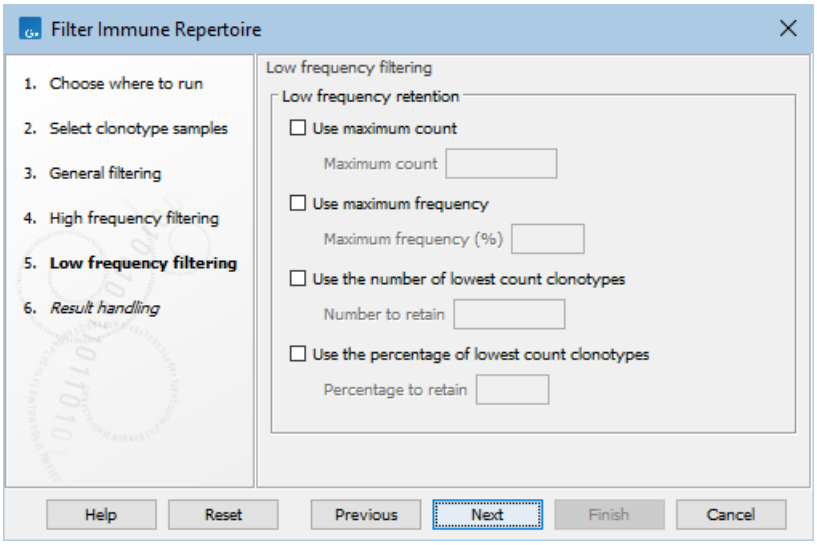


Figure 7.13: *Low frequency filtering options for Filter Immune Repertoire.*

will be retained, and for those chains, only the clonotypes for which the identification of the D segment was successful will be retained. If left empty, no filter is applied.

- **Recalculate frequencies**. If ticked, frequencies in the output clonotypes are recalculated such that they add up to 100% across all chains. Otherwise, the original frequencies found in the input are used.

  It can be useful to recalculate frequencies when removing noise (for example, removing clonotypes with a count of 1), but if a subset of clonotypes is created for the purpose of comparing clonotypes between samples, it might be more relevant to preserve the original frequencies. Note that the frequencies are always recalculated separately for each individual sample.

- **Set frequencies per chain**. If ticked, the frequencies are recalculated to add up to 100% for each individual chain. This option is enabled only when **Recalculate frequencies** is ticked.

- **High frequency retention**. The following filters for removing clonotypes with low frequencies can be enabled:
  - **Use minimum count**. Retain clonotypes with a count greater than or equal to **Minimum count**.
  - **Use minimum frequency**. Retain clonotypes with a frequency greater than or equal to **Minimum frequency (%)**.
  - **Use the number of highest frequency clonotypes**. Retain **Number to retain** clonotypes from each sample that have highest frequency.
  - **Use the percentage of highest frequency clonotypes**. Retain **Percentage to retain** percentage of clonotypes from each sample that have highest frequency.
- **Low frequency retention**. These filters can be used to remove clonotypes that have high frequencies:
  - **Use maximum count**. Retain clonotypes with a count less than or equal to **Minimum count**.
  - **Use maximum frequency**. Retain clonotypes with a frequency less than or equal to **Minimum frequency (%)**.
  - **Use the number of lowest frequency clonotypes**. Retain **Number to retain** clonotypes from each sample that have lowest frequency.
  - **Use the percentage of lowest frequency clonotypes**. Retain **Percentage to retain** percentage of clonotypes from each sample that have lowest frequency.

The tool outputs the filtered clonotypes and a report summarizing statistics of the filtered clonotypes. See section 7.3.1 for **TCR clonotypes** () and **BCR clonotypes** (), or section 7.6.2 for **Clonotype Sample Comparison** ().

7.6 Compare Immune Repertoires

The **Compare Immune Repertoires** tool contrasts properties of immune repertoires, such as diversity and similarity.

Compare Immune Repertoires is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Immune Repertoire Analysis** () | **Compare Immune Repertoires** ()

This opens a dialog where a combination of **TCR clonotypes** () or **BCR clonotypes** (), and **Clonotype Sample Comparison** () objects to be compared can be selected. Note that TCR and BCR clonotypes cannot be mixed and only one type should be used at a time.

In the next step, the following options are available:

- **Resolve clonotypes with ambiguous segments**. If checked, a clonotype in one sample with ambiguous V/D/J/C segments is resolved using a clonotype in another sample with unambiguous segments, where possible.
- **Resolve clonotypes without C segment**. If checked, a clonotype in one sample without a C segment is resolved using a clonotype in another sample with a C segment, where possible.

See section 7.6.1 for more details.

7.6.1 Resolving of clonotypes

In order for clonotypes to be considered the same, they need to have the same CDR3 sequence and V/D/J/C segments.

Sequencing errors and variability in read quality and length can lead to false positives, see section 7.3 and section 7.4. These can be reduced within one sample using the Merge Immune Repertoire tool, but such false positives can still remain when comparing samples.

If the same clonotype is present in two samples, up to differences due to ambiguous V/D/J/C segments and/or missing C segment, it will be considered as two separate clonotypes, leading to private clonotypes in each sample (see section 7.8.3). To consider them as one clonotype, the **Resolve clonotypes with ambiguous segments** and/or **Resolve clonotypes without C segment** can be used.

Resolve clonotypes with ambiguous segments: A clonotype with ambiguous segments may be corrected to have an unambiguous segment, if another sample has the same clonotype with unambiguous segment. Consider a clonotype with ambiguous V segments V-1/V-2 in sample A, and the same clonotype, but with V-1, in sample B. Then the clonotype from sample A is corrected to have V-1. If another sample C is also used, which has the same clonotype, but with V-2, it cannot be determined if V-1 or V-2 is the correct segment for sample A, so the clonotype is not changed.

Resolve clonotypes without C segment: A clonotype without a C segment may be corrected to have a C segment, if another sample has the same clonotype with a C segment. Consider a clonotype without a C segment in sample X, and the same clonotype, but with C-1, in sample Y. Then the clonotype from sample X is corrected to have C-1. If another sample Z is also used, which has the same clonotype, but with C-2, it cannot be determined if C-1 or C-2 is the correct segment for sample X, so the clonotype is not changed.

7.6.2 Output from Compare Immune Repertoires tool

Two outputs can be generated by the tool:

- **Clonotype Sample Comparison** (), containing the clonotypes from each sample, see section 7.8 for more details.
- **Compare Immune Repertoires (CIR) report** (). A report containing comparisons of repertoire properties.

CIR Report

The report starts with the following section:

- **Summary**. A summary table showing, for each input, the total number of clonotyped fragments, as well as the clonotyped fragments from each chain type. A fragment represents one single read or a pair of reads.

The remaining information in the report is given per chain type and only for those chain types for which clonotypes have been identified for at least two of the inputs.

- **Resolved clonotypes**. If **Resolve clonotypes with ambiguous segments** and/or **Resolve clonotypes without C segment** have been checked, this section will contain a table with information about how many clonotypes have been resolved:
  - The number of clonotypes, if any, for which segments have been disambiguated, for each segment type (V, D, J, and C).
  - The number of clonotypes, if any, for which a C segment has been added.
- **Diversity indices**. A table showing the diversity metrics 'Observed diversity', 'Extrapolated diversity (chaoE)' and 'Extrapolated Shannon-Wiener index (chaoE)', see section 7.3.1. Additionally, the table contains the diversity metric 'Interpolated to lowest sample diversity', showing an estimate of the diversity if all the inputs had the same number of clonotyped fragments as the input with the lowest number of clonotyped fragments.
- **Scatter plots**. If exactly two inputs are compared, this section will contain scatter plots with the clonotypes frequency in the two inputs. The **Clonotype Sample Comparison** ( ) can display the scatter plot for any two inputs at a time, see section 7.8.3 for more details.
- **Rarefaction**. A plot with rarefaction curves, also known as species accumulation curve. See section 7.8.4 for more details.
- **CDR3 length**. A table comparing CDR3 length summary statistics for the different inputs, across all chains. For each chain type, the CDR3 length distribution is also shown as a histogram, see section 7.8.5 for more details.
- **V, D, J and C usage**. Bar plots showing the V, D, J and C segment usage for each input. See section 7.8.6 for more details.

7.7 Clonotypes

TCR clonotypes ( ) and **BCR clonotypes** ( ) contain the clonotypes and have a number of views, displaying different properties / summaries of the clonotypes.

7.7.1 Table for Clonotypes

TCR clonotypes ( ) and **BCR clonotypes** ( ) show by default a table with the following columns:

- **Clonotype #**: A unique number identifying the clonotype.
- **Chain**: Which chain type the clonotype belongs to: TRA (α), TRB (β), TRG (γ) and TRD (δ) for TCR, or IGH (heavy), IGK (light κ), and IGL (light λ) for BCR. Note that other light BCR chain types are currently not supported.
- **V**, **D**, **J** and **C**: The identified V, D, J and C reference segment(s). If a single unambiguous V / D / J / C segment could not be identified, the segments will be listed separated by a comma.

- **CDR3 nucleotide sequence**: The nucleotide sequence for CDR3 including the V and J segment-encoded conserved amino acids.
- **CDR3 amino acid sequence**: The translated amino acid sequence for the CDR3 nucleotide sequence provided that it is in-frame.
- **CDR3 length**: The length of the CDR3 nucleotide sequence.
- **Count**: The number of fragments for which the specific clonotype was detected. A fragment represents one single read or one pair of reads.
- **Frequency (%)**: The count given as a percentage relative to the sum of all counts. Note that filtering can affect frequencies, see section 7.5.
- **Productive**: One of three categories are used to characterize the CDR3 nucleotide sequences:
  - Productive: sequences that are in frame and do not contain a premature stop codon;
  - Out-of-frame: sequences that have a length that is not a multiple of three;
  - Premature stop codon: sequences that are in-frame but contain a premature stop codon.

The clonotypes are sorted by frequency in decreasing order.

CDR3 quality sore

The CDR3 nucleotide sequences can be colored using the average quality score of each position, by using the "Color CDR3 by average quality sore" option in the Side Panel menu to the right. This option is disabled when there are no quality scores available.

Clonotypes from selection

At the bottom of the table, a button labeled **Create Clonotypes from Selection** is available. Select the relevant rows in the table and click the button to create new **TCR clonotypes** () or **BCR clonotypes** () that only include the selected clonotypes. When the button is clicked, a dialog with the following options is shown:

- **Recalculate frequencies**. If ticked, frequencies in the output clonotypes are recalculated such that they add up to 100% across all chains. Otherwise, the original frequencies found in the input are used.

  It can be useful to recalculate frequencies when removing noise (for example, removing clonotypes with a count of 1), but if a subset of clonotypes is created for the purpose of comparing clonotypes between samples, it might be more relevant to preserve the original frequencies.
- **Set frequencies per chain**. If ticked, the frequencies are recalculated to add up to 100% for each individual chain. This option is enabled only when **Recalculate frequencies** is ticked.

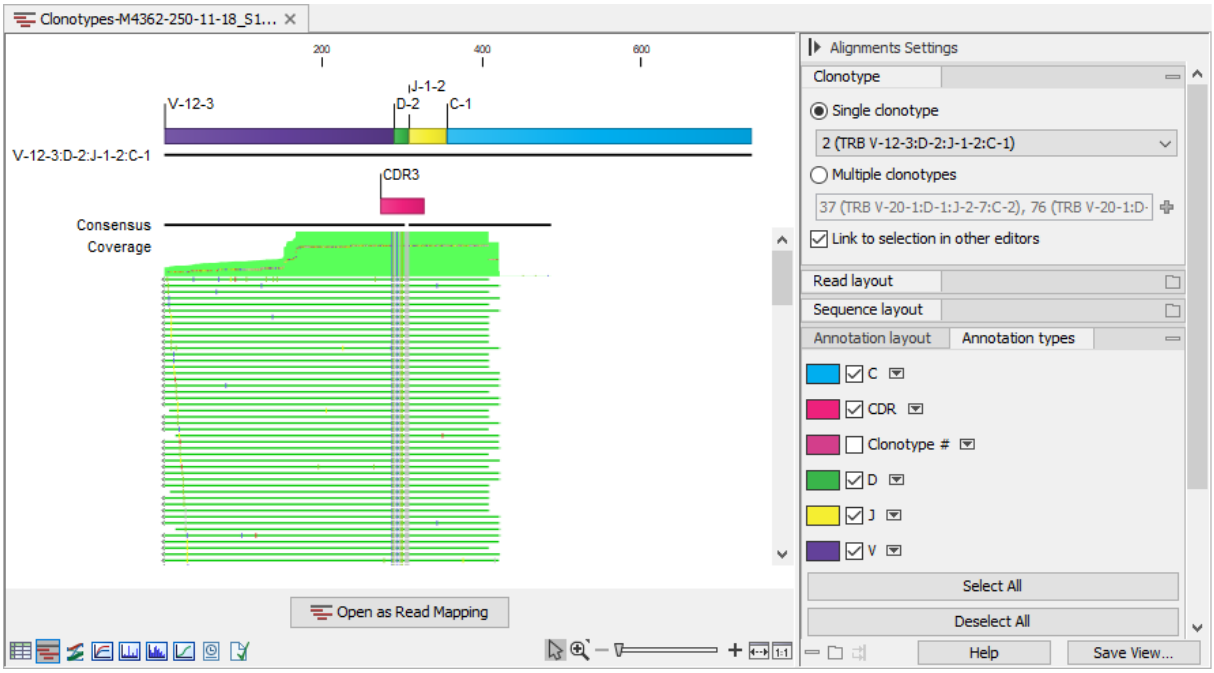


Figure 7.14: *Read mapping for a TRB clonotype. V, D, J and C segments are annotated on the reference sequence and the CDR3 is annotated on the consensus.*

7.7.2 Alignments for Clonotypes

The alignments view ( ) shows all reads mapping to a specific clonotype (figure 7.14).

The alignment contains:

- The reference sequence consisting of the identified V(D)JC segments. Annotations indicate the location of the different segment types. For clonotypes with ambiguous segments, the first segment from the list is used.
- The consensus sequence with an annotation indicating the CDR3 region.
- The aligned reads.

Various settings can be configured in the side panel, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html.

For further processing, the alignments can be opened and saved as a stand-alone read mapping by using the "Open as Read Mapping" button.

The clonotypes for which the alignment should be shown can be selected from the drop-down menus in the side panel, or from the clonotype table (section 7.7.1) while using a split view, see figure 7.15.

Alignments for multiple clonotypes can be shown together provided that they have the same V and J segments and the D / C segments are not contradictory: either the D / C segment is identified and the same, or it is missing (figure 7.15).

When viewing alignments for multiple clonotypes, it can be useful to change "Compactness" to "Not compact" and tick the "Clonotype #" annotation from the side panel. This way, it is easy to

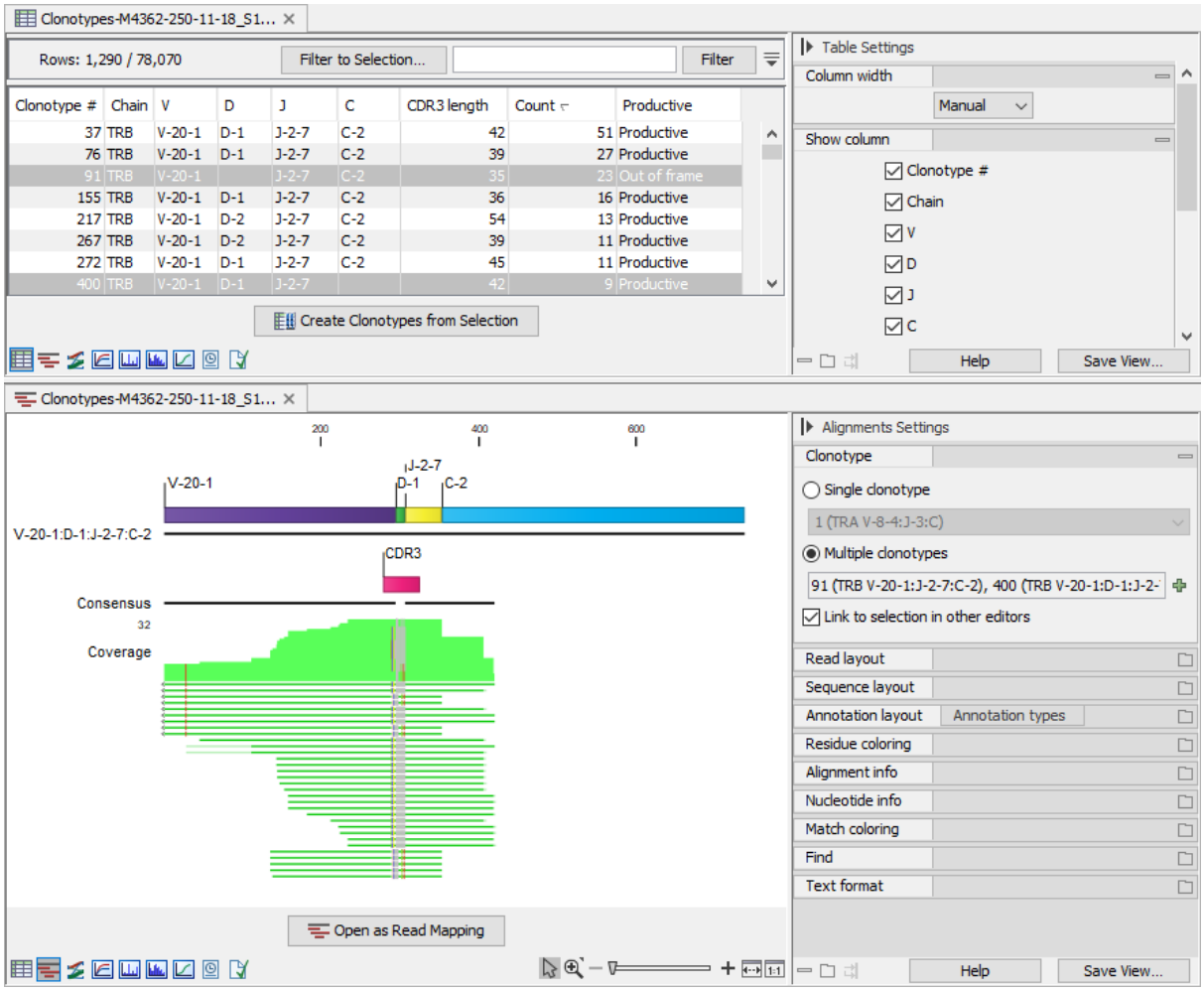


Figure 7.15: *Clonotypes split view. Top: Multiple clonotypes sharing the reference segments are selected in the table view. Bottom: Alignment view for the clonotypes selected in the table view.*

see the clonotype assigned to each read, figure 7.16.

Figure 7.15 shows an alignment for two clonotypes. Some of the reads do not span past the J segment and using the "Clonotype #" annotation, we can confirm that these reads belong to clonotype # 400, which is the clonotype without the C segment (figure 7.16). The two clonotypes do not share the D segment and have different CDR3 sequences. Using the alignment view (figure 7.15), it is straightforward to spot the differences between the two CDR3 sequences.

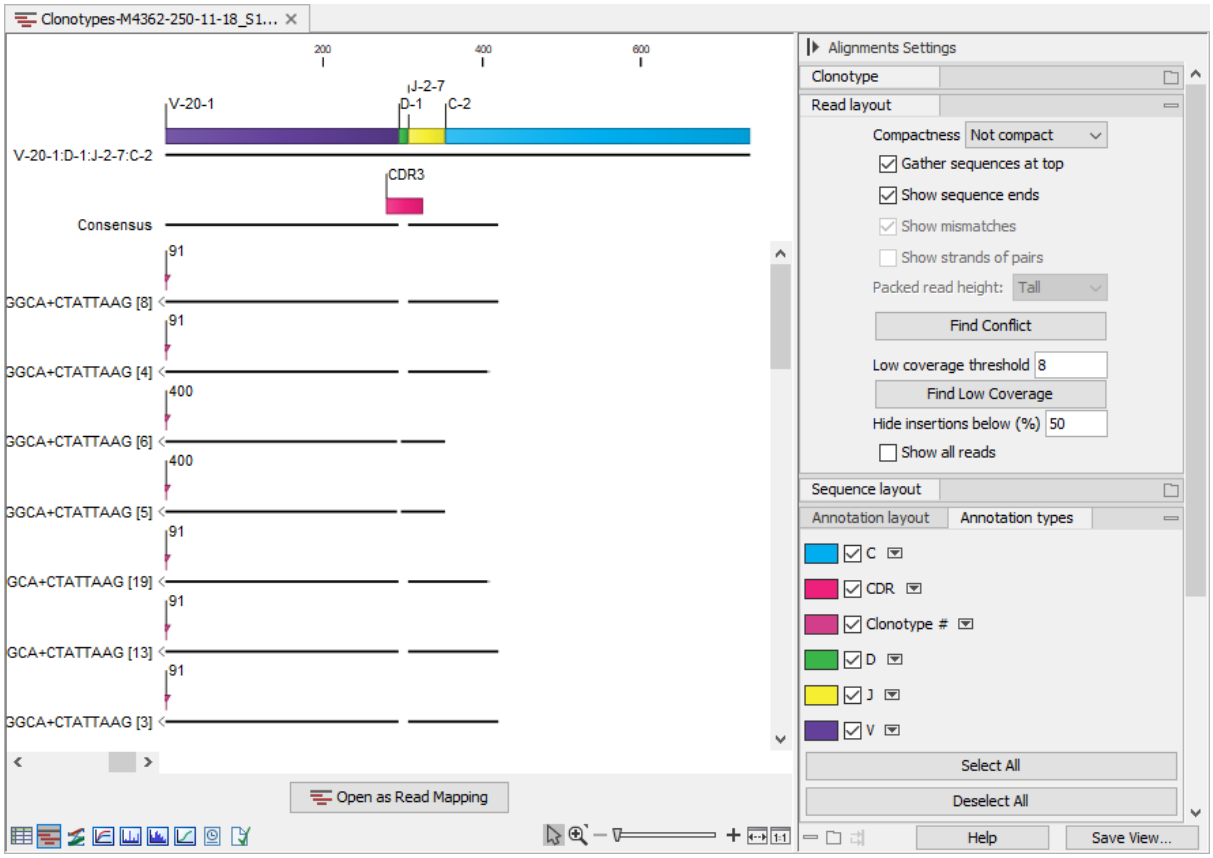


Figure 7.16: *Alignment view for multiple clonotypes where "Compactness" is set to "Not compact" and "Clonotype #" is ticked.*

Alignments after merging

Alignments for clonotypes with ambiguous segments are calculated using the first segment in the list. Thus when ambiguous segments have been resolved using **Merge Immune Repertoire**, see section 7.4, some reads may not be included in the alignments view. This happens when the unique segment that the ambiguous segments resolved to was not the first segment in the list.

Two situations can arise as illustrated in figure 7.17:

- The entire alignment cannot be determined. In this case, the alignment view shows an alignment of coverage 1, with one read that is identical to the reference (top right alignment in figure 7.17). This can happen when the most frequent clonotype had ambiguous segments, and ambiguity has been resolved using a less frequent clonotype. This alignment should not be used.
- The alignment for some reads cannot be determined. In this case, the alignment view does not include the reads for which the alignment could not be determined. This is apparent when the alignment coverage is lower than the clonotype count (bottom alignment in figure 7.17). This can happen when a less frequent clonotype had ambiguous segments, and ambiguity has been resolved using the most frequent clonotype. If the percentage of discarded reads is low, it will not impact the consensus sequence of the alignment.

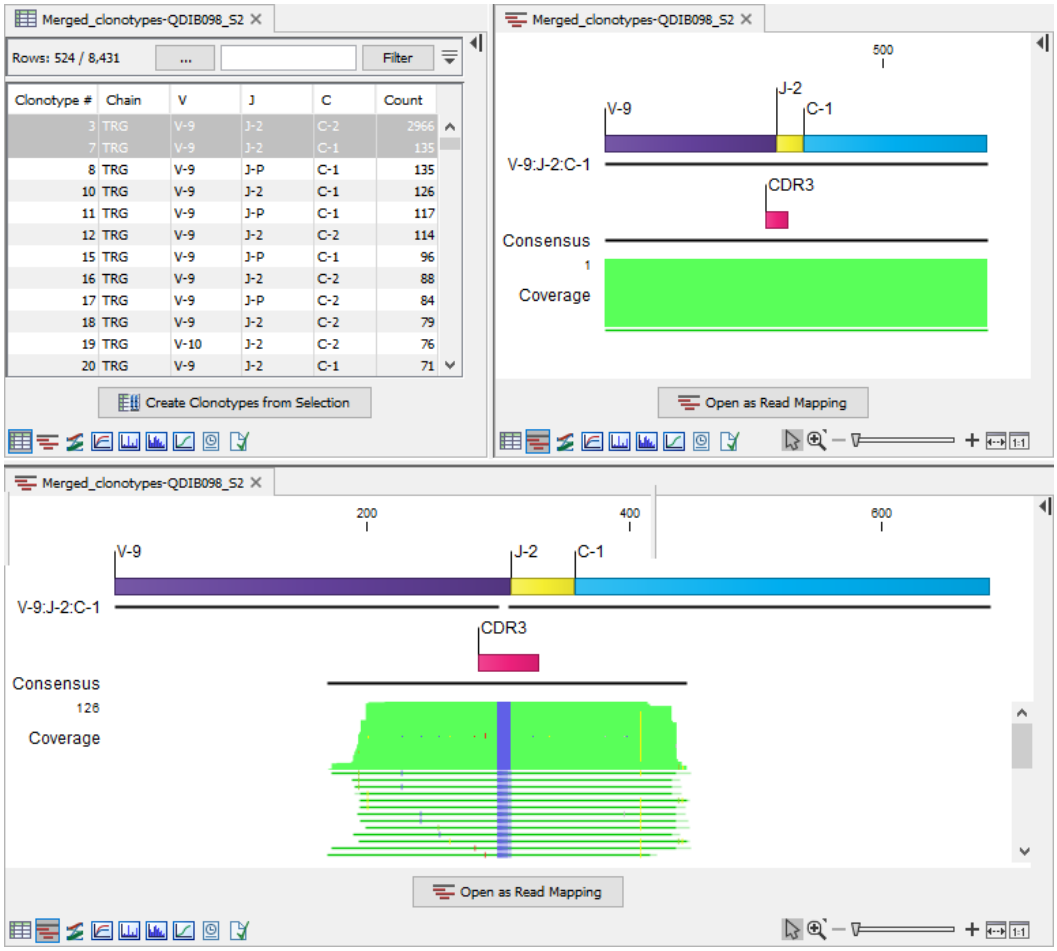


Figure 7.17: *Clonotypes split view. Top left: Table view of TRG clonotypes that have been merged. Top right: Alignment view for clonotype 3. The alignment could not be determined and the view shows an alignment of coverage 1. Bottom: Alignment view for clonotype 7. The clonotype count is 135, while the alignment coverage is only 126, i.e. the alignment for 9 reads could not be determined.*

7.7.3 Sankey plot for Clonotypes

The Sankey plot view ( ) shows how the segments of different types form the clonotypes for a given chain (figure 7.22). To keep the plot size manageable, it is recommended to first filter the clonotypes using **Filter Immune Repertoire** (see section 7.5).

For each selected segment type, the plot has a column that contains boxes for each segment. The box height reflects the total count for clonotypes with the given segment. The boxes are connected with flows. "Show continuous flows" controls the type of the flows:

- If not ticked, there are flows between boxes in consecutive columns for clonotypes having segments corresponding to the boxes. The height of the flow indicates the total count for these clonotypes.
- If ticked, the flows start from a fixed column. As before, flows between boxes in the first two columns reflect clonotypes with segments corresponding to the two boxes. Flows between boxes in the second and third column reflect clonotypes corresponding to the boxes in both

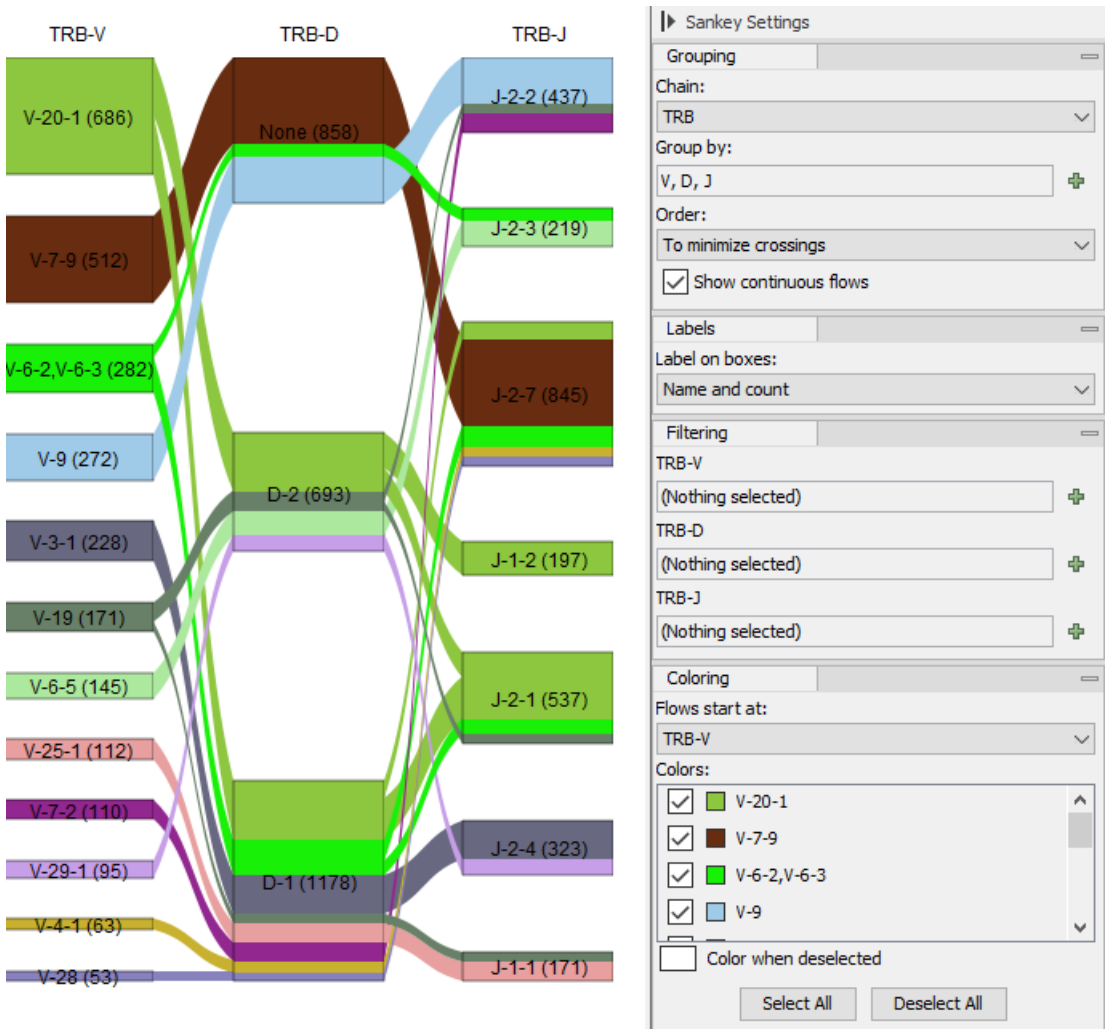


Figure 7.18: *Sankey plot for the TRB chain showing V, D and J segments. Numbers in brackets show the total count for the corresponding clonotypes. Flows show how many clonotypes with the specific segment combination are present. Clonotypes without D-segments are represented by the "None" box.*

the first, second and third column, and so on.

In both cases, the color of a flow indicates the element where the flow starts. "Flows start at" can be used to change the start column, by default the leftmost.

"Show continuous flows" has no effect when there are less than three columns.

Boxes can be removed from the plot by using the options under "Filtering" in the side panel (figure 7.18). The plot will show only boxes for the selected segments and the boxes to which the selected segments have a flow to (figure 7.19). If multiple filters are used, boxes are subject to all the restrictions (figure 7.20).

The columns and their order can be changed by using the options under "Group by" in the side panel (figure 7.18). The CDR3 amino acid sequences can also be shown, see figure 7.21.

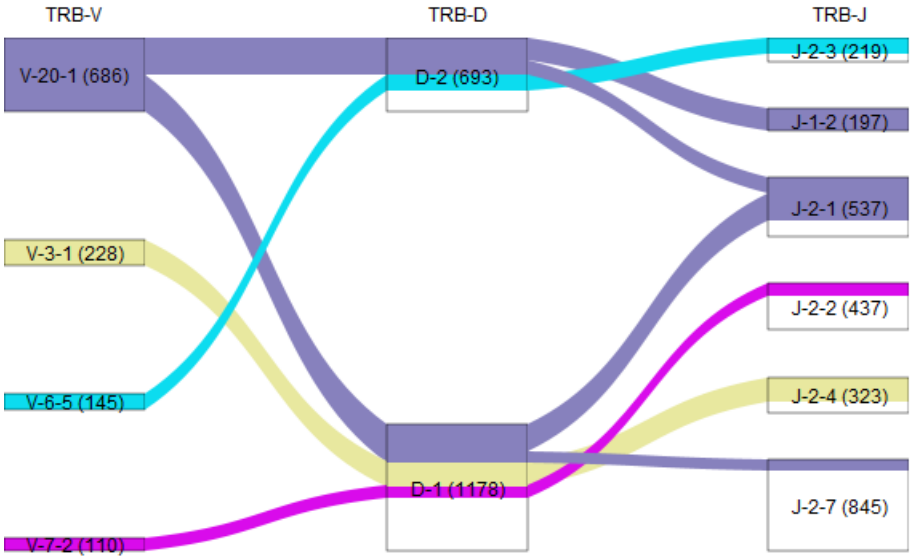


Figure 7.19: *Sankey plot for the TRB chain showing V segments restricted to V-20-1, V-3-1, V-6-5 and V-7-2. Boxes for D and J segments are only shown for segments present in clonotypes having one of the selected V-segments.*

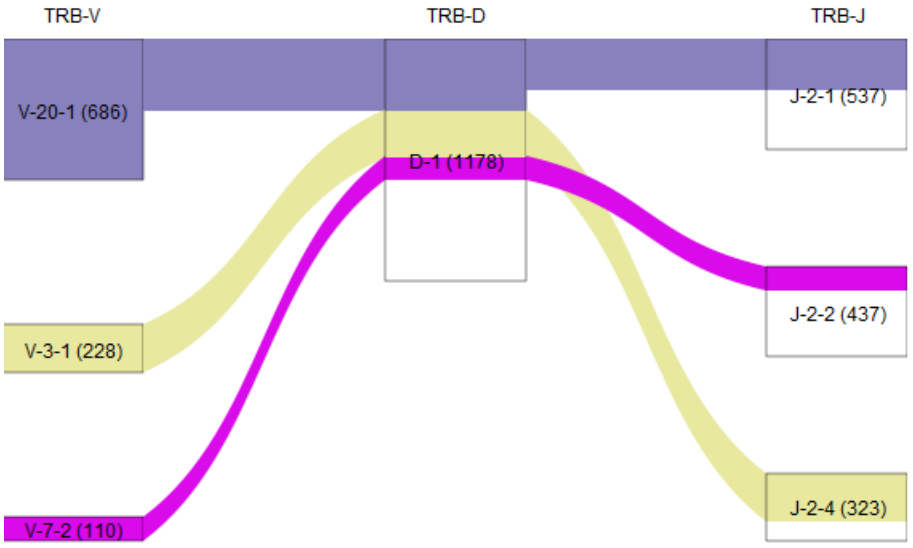


Figure 7.20: *Sankey plot for the TRB chain showing V segments restricted to V-20-1, V-3-1, V-6-5 and V-7-2; D segments restricted to D-1; and J segments restricted to J-2-3, J-2-1, J-2-2, J-2-3, and J-2-4. Note that V-6-5 is not in the plot because there are no clonotypes with V-6-5, D-1 and one of the selected J-segments.*

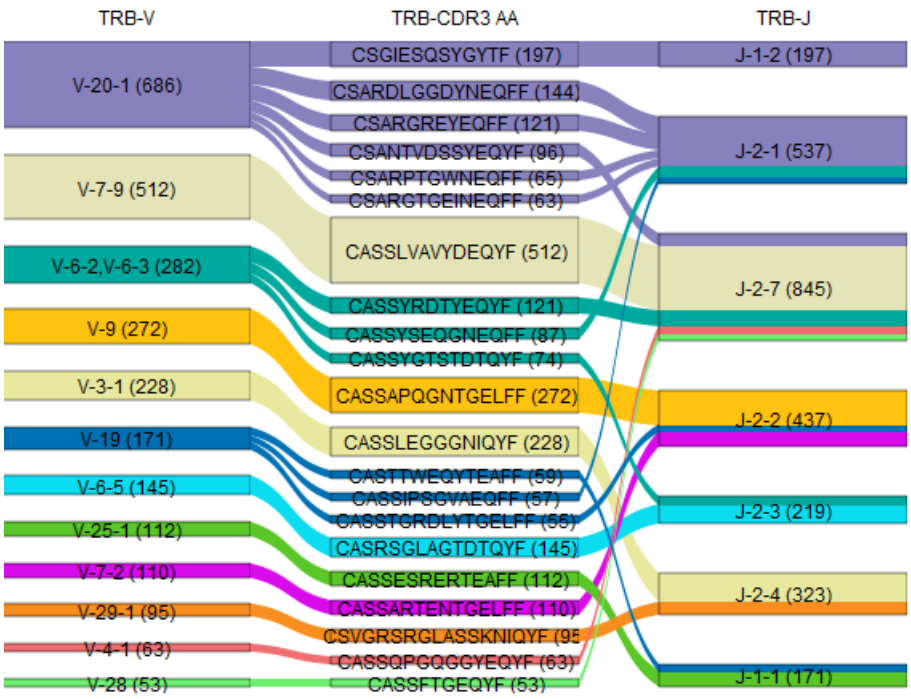


Figure 7.21: *Sankey plot for the TRB chain showing V segment, CDR3 and J segment.*

7.7.4 Rarefaction for Clonotypes

The rarefaction curve ( ) is also known as the species accumulation curve (figure 7.22). It shows the expected number of distinct clonotypes discovered as a function of the total number of detected clonotypes, together with the confidence interval (CI), obtained from a normal approximation. The curve is

- interpolated down to 0 clonotypes;
- extrapolated to twice the total number of detected clonotypes.

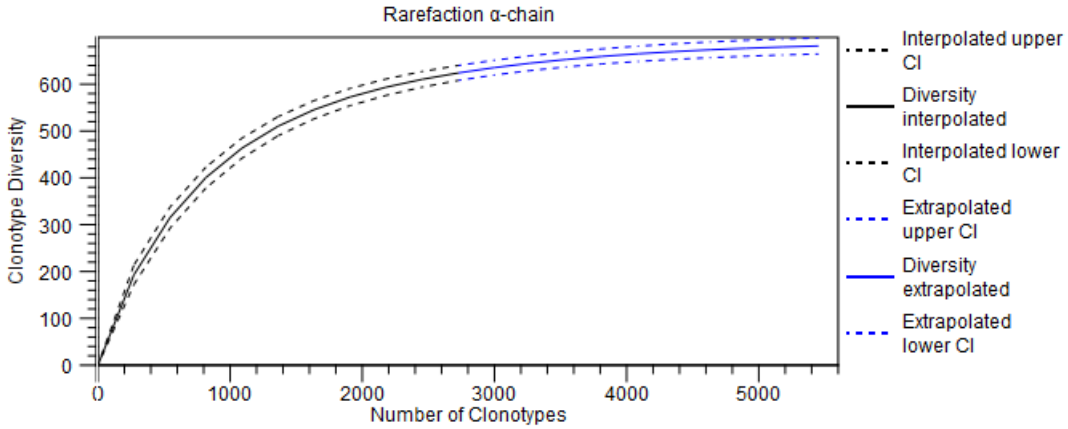


Figure 7.22: *Rarefaction plot for the TRA chain. The chain can be changed from the side panel.*

7.7.5 CDR3 length for Clonotypes

The CDR3 length plot ( ) shows the length distribution of the CDR3 nucleotide sequences (figure 7.23). Peaks are expected every 3 nt. due to repertoires consisting predominantly of in-frame CDR3 sequences.

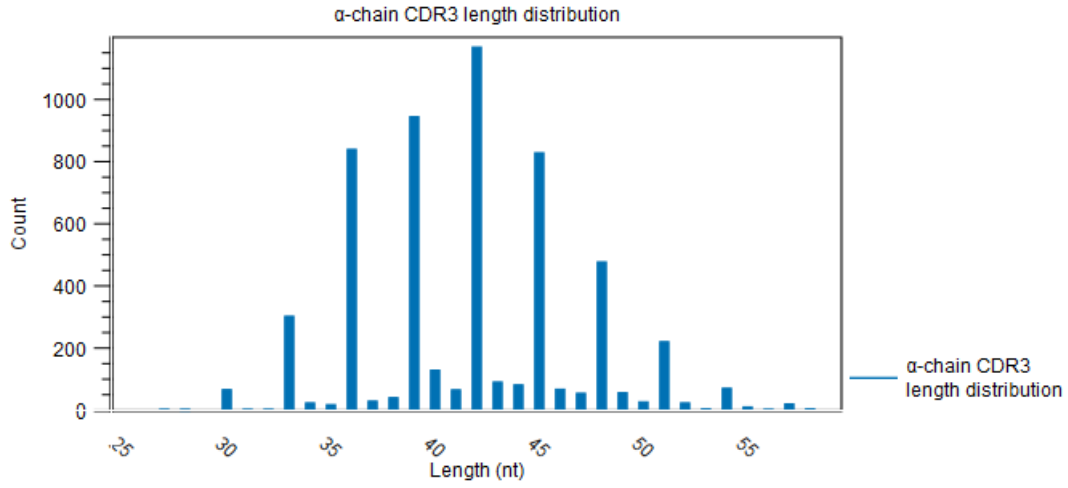


Figure 7.23: *CDR3 length distribution plot for the TRA chain with peaks every 3 nt. The chain can be changed from the side panel.*

7.7.6 Segment usage for Clonotypes

The segment usage histogram ( ) shows the clonotypes counts for the detected segments.

The chain and segment type can be changed from the side panel. It is also possible to restrict the histogram to selected segments.

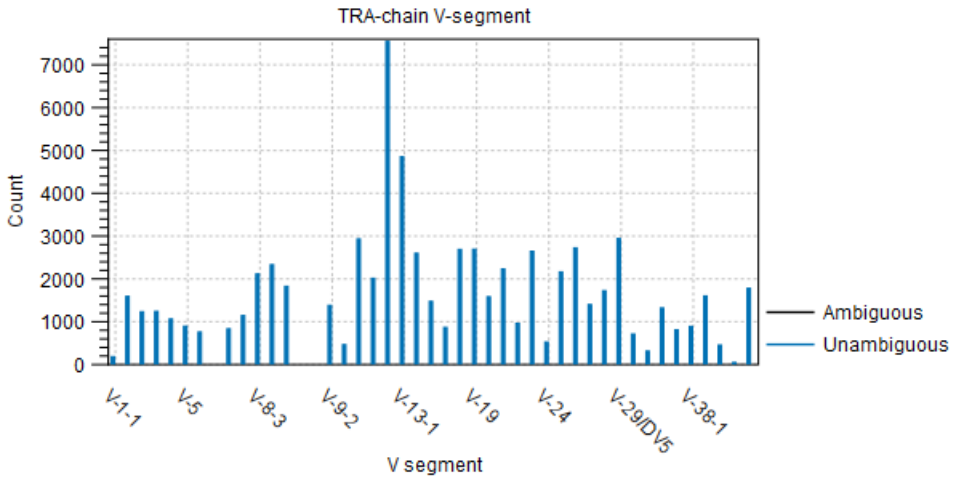


Figure 7.24: *Segment usage for TRA V segments.*

7.7.7 Cumulative frequencies for Clonotypes

The cumulative frequency curve ( ) shows the cumulative frequencies of the identified clonotypes, ordered by descending count (figure 7.25). If the curve is steep in the beginning and then flattens, it indicates that a few clonotypes account for most of the reads. On the other hand, if the curve is more linear, it indicates a more even distribution of reads among the clonotypes.

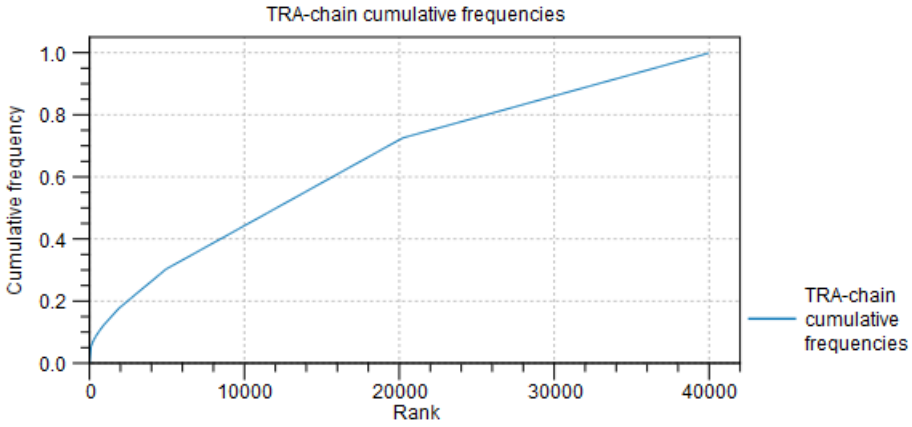


Figure 7.25: *The cumulative frequencies of the TRA chain. The chain can be changed from the side panel.*

7.8 Clonotype Sample Comparison

The **Clonotype Sample Comparison** ( ) contains clonotypes across multiple samples and has a number of views, displaying and comparing across samples different properties / summaries of the clonotypes.

7.8.1 Tables for Clonotype Sample Comparison

Clonotype Sample Comparisons have two table views:

- Clonotypes ( ). If the same clonotype is present in multiple samples, there will be just one row for the clonotype. The **Samples** column contains the number of samples in which the clonotype was identified.
- Clonotypes on Sample Level ( ). If the same clonotype is present in multiple samples, there will be one row for each sample. The **Sample** column contains the sample where the clonotype was identified.

Both tables have the following columns:

- **Chain**: Which chain type the clonotype belongs to: TRA (α), TRB (β), TRG (γ) and TRD (δ) for TCR, or IGH (heavy), IGK (light κ), and IGL (light λ) for BCR. Note that other light BCR chain types are currently not supported.
- **V**, **D**, **J** and **C**: The identified V, D, J and C reference segment(s). If a single unambiguous V / D / J / C segment could not be identified, the segments will be listed separated by a comma.
- **CDR3 nucleotide sequence**: The nucleotide sequence for CDR3 including the V and J segment-encoded conserved amino acids.
- **CDR3 amino acid sequence**: The translated amino acid sequence for the CDR3 nucleotide sequence provided that it is in-frame.
- **CDR3 length**: The length of the CDR3 nucleotide sequence.
- **Count**: The number of fragments for which the specific clonotype was detected. A fragment represents one single read or one pair of reads.
- **Frequency (%)**: The count given as a percentage relative to the sum of all counts. Note that filtering can affect frequencies, see section 7.5.
- **Productive**: One of three categories are used to characterize the CDR3 nucleotide sequences:
  - Productive: sequences that are in frame and do not contain a premature stop codon;
  - Out-of-frame: sequences that have a length that is not a multiple of three;
  - Premature stop codon: sequences that are in-frame but contain a premature stop codon.

The Clonotypes view ( ) has one **Count** and one **Frequency (%)** column for each sample.

Clonotype Sample Comparison from selection

At the bottom of each table, a button labeled **Create Comparison from Selection** is available. Select the relevant rows in the table and click the button to create a new **Clonotype Sample Comparison** () that only includes the selected clonotypes. When the button is clicked, a dialog with the following options is shown:

- **Recalculate frequencies**. If ticked, frequencies in the output clonotypes are recalculated such that they add up to 100% across all chains. Otherwise, the original frequencies found in the input are used.

  It can be useful to recalculate frequencies when removing noise (for example, removing clonotypes with a count of 1), but if a subset of clonotypes is created for the purpose of comparing clonotypes between samples, it might be more relevant to preserve the original frequencies. Note that the frequencies are always recalculated separately for each individual sample.

- **Set frequencies per chain**. If ticked, the frequencies are recalculated to add up to 100% for each individual chain. This option is enabled only when **Recalculate frequencies** is ticked.

Note, that the frequencies are always recalculated separately for each sample in the selection.

7.8.2 Sankey plot for Clonotype Sample Comparison

The Sankey plot () compares clonotypes frequencies across samples (figure 7.26).

For each selected sample, the plot has a column that contains boxes for each group of clonotypes (hereby referred to as simply clonotypes) with the selected properties. The properties, such as the segment type or the CDR3 amino acid sequence, are selected from the side panel under "Group by".

The height of a box indicates the frequencies of the clonotypes in the sample. The frequency is used instead of the count to make samples with a different number of reads comparable.

Note that the sum of the height of boxes may differ across samples. The frequency, across all chains, adds up to 100% when clonotypes are first constructed. It is unlikely that the sum of the frequencies for a specific chain adds up to the same total for two different samples. To achieve this, the clonotypes can be filtered to only contain the desired chain and the frequencies can be recalculated. Additionally, filtering can lead to frequencies adding to less than 100%. See section 7.5 for details on how to filter.

7.8.3 Scatter plot for Clonotype Sample Comparison

The scatter plot () shows the clonotypes frequency in two samples for a given chain (figure 7.27). The chain and the two samples to be compared can be changed from the side panel.

7.8.4 Rarefaction for Clonotype Sample Comparison

The rarefaction curve () is also known as the species accumulation curve (figure 7.22). It shows the expected number of distinct clonotypes discovered as a function of the total number of

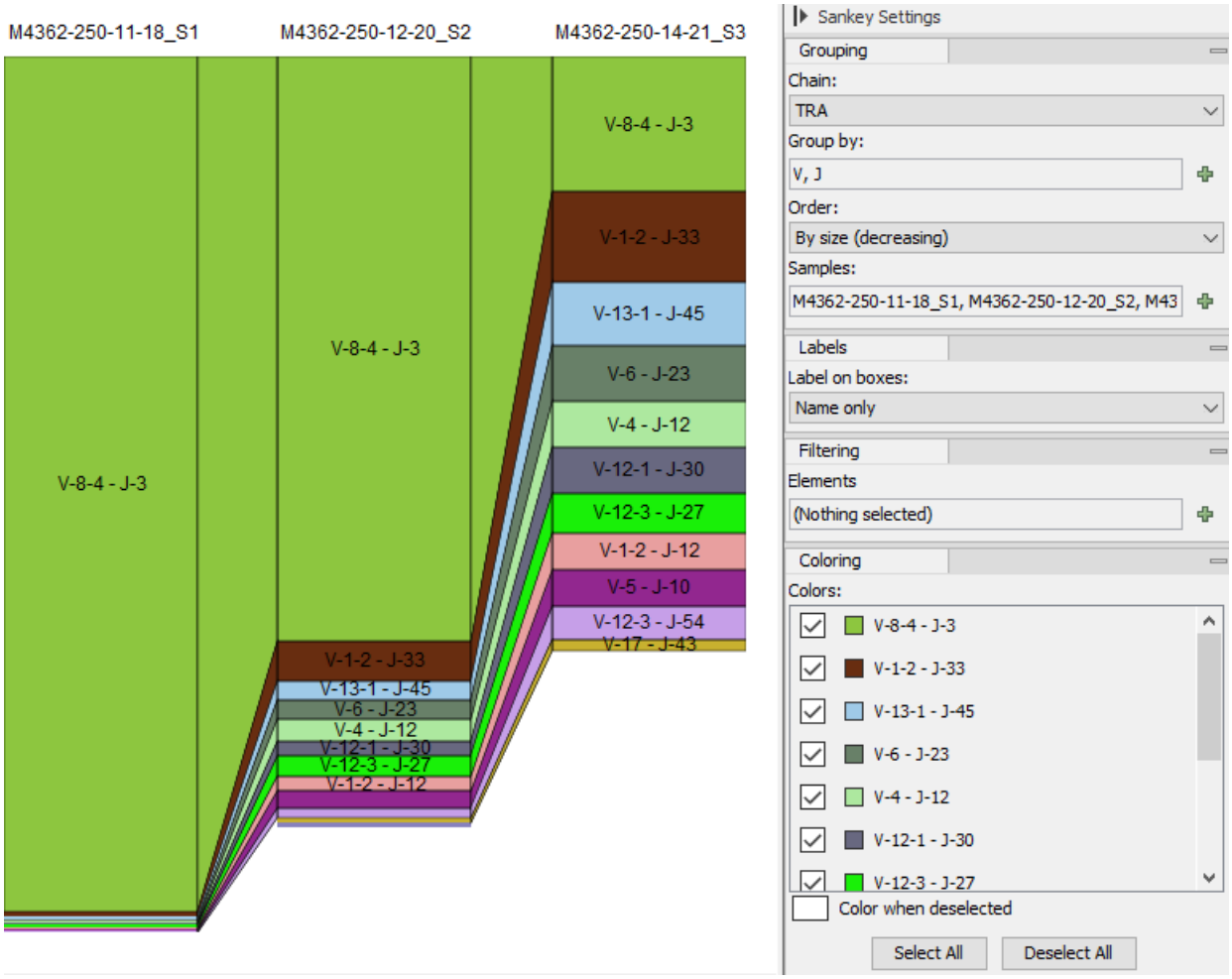


Figure 7.26: *Sankey plot for the TRA chain showing frequencies of clonotypes with specific V and J segments, compared across samples.*

detected clonotypes. The curve is extrapolated to twice the total number of detected clonotypes for the most abundant input.

7.8.5 CDR3 length for Clonotype Sample Comparison

The CDR3 length plot ( ) shows the length distribution of the CDR3 nucleotide sequences (figure 7.23). Peaks are expected every 3 nt. due to repertoires consisting predominantly of in-frame CDR3 sequences.

7.8.6 Segment usage for Clonotype Sample Comparison

The segment usage histogram ( ) shows the frequency of clonotypes with detected segments per sample for a given chain and segment type.

The samples, chain and segment type can be changed from the side panel. It is also possible to restrict the histogram to selected segments.

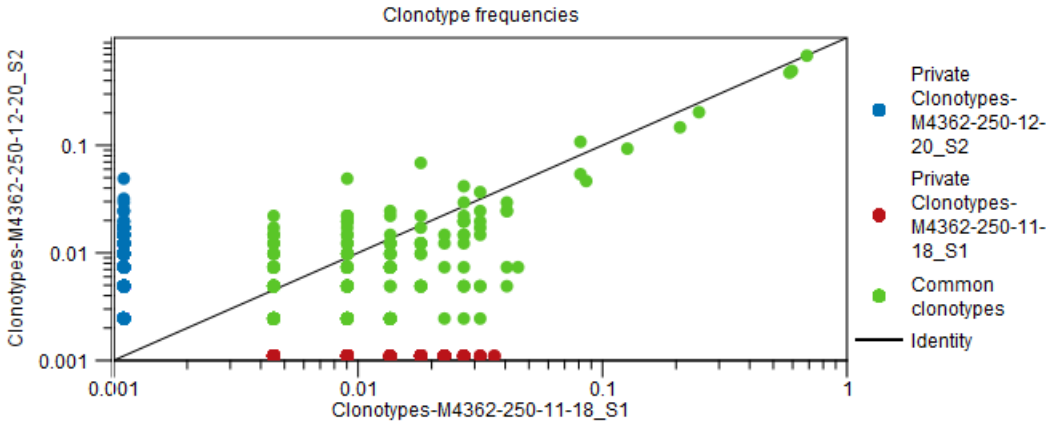


Figure 7.27: *Scatter plot with clonotype frequencies for a particular chain type. Note that private clonotypes have frequencies 0 in one of the samples. Due to the log scale, they cannot be plotted at frequency 0.*

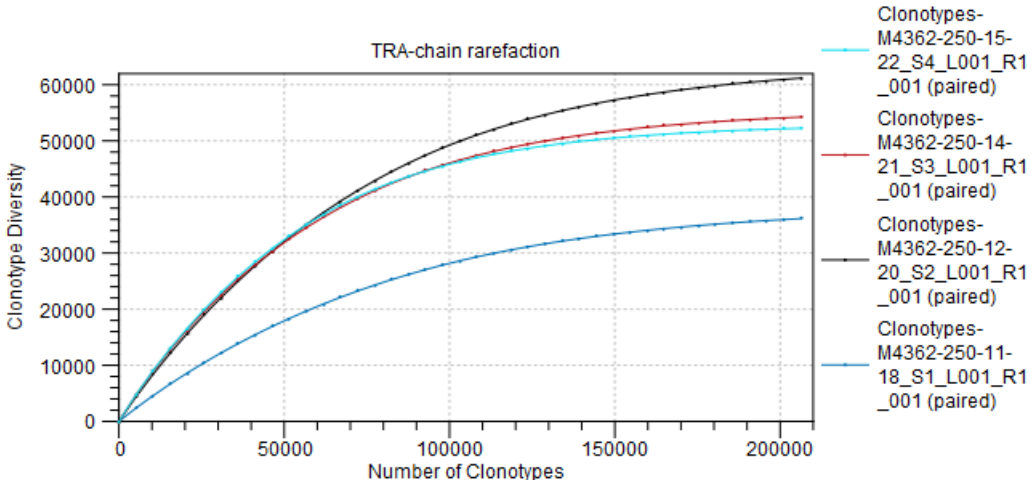


Figure 7.28: *Rarefaction plot for the TRA. The samples and chain can be changed from the side panel.*

7.8.7 Jaccard distance heat map for Clonotype Sample Comparison

The Jaccard distance heat map ( ) shows the Jaccard distance with samples clustered hierarchically.

For each pair of samples, the weighted Jaccard similarity between the two is computed. Let X_i, Y_i denote the relative frequencies of the i'th clonotype in the first and second sample respectively. The weighted Jaccard similarity is defined as

$$J(X,Y) = \frac{\sum_{i=1}^{n} \min(X_i, Y_i)}{\sum_{i=1}^{n} \max(X_i, Y_i)}\,.$$

The weighted Jaccard distance is defined as

$$D(X,Y) = 1 - J(X,Y)\,.$$

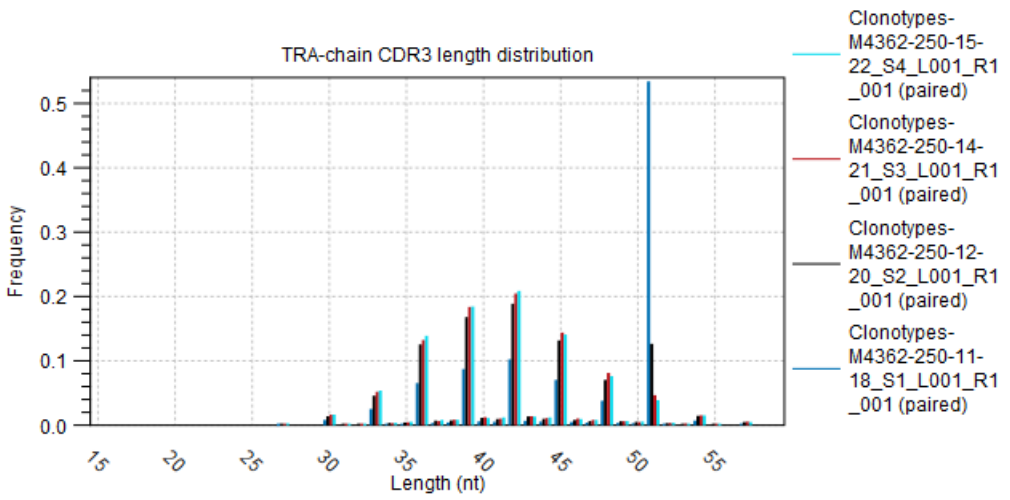


Figure 7.29: *CDR3 length distribution plot for the TRA chain. The samples and chain can be changed from the side panel.*

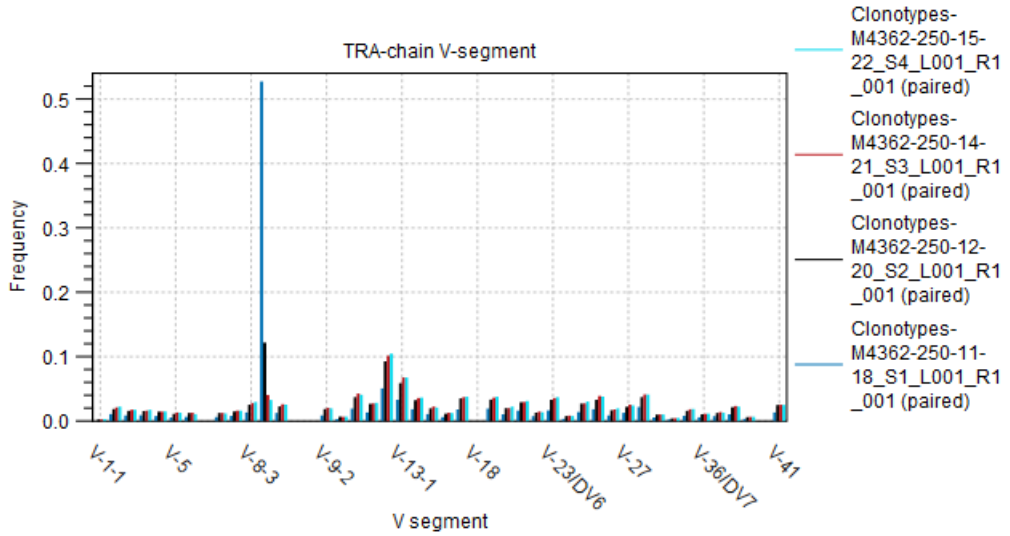


Figure 7.30: *Segment usage for TRA V segments.*

The color gradient used by the heat map can be changed from the side panel. It is also possible to restrict the heat map to selected samples.

The heat map can be opened and saved as a stand-alone heat map by using the "Open as New Distance Heat Map" button. Various settings can be configured in the side panel of the stand-alone heat map, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_heat_map_view.html. Note that the "Open as New Distance Heat Map" button opens a heat map with "Columns" and "Rows" in the side panel, instead of the "Samples" and "Features" described in the linked manual.

The "Open as New Similarity Table" button opens a stand-alone table containing the Jaccard similarity between each pair of samples.

Chapter 8

Oncology score estimation

Contents

The oncology score estimation folder contains tools related to TMB, MSI and HRD.

8.1 Calculate TMB Score

Calculate TMB Score takes a variant track and the set of regions to focus on, and calculates a TMB score, i.e. the number of variants per 1 million bases.

It is recommended that target regions with a coverage lower than 100X are discarded before running the tool. To do so, a workflow including the tools Create Mapping Graph and Identify Graph Threshold Area can be used to generate a target region file only containing target regions with at least 100X coverage (see figure 8.1).

The Calculate TMB Score tool currently considers only SNVs - and discards variants of any other type. First, it filters variants, keeping only variants that lie within exons within ROIs and outside the masking regions. It then applies successively various quality, germline and non-synonymous filters before calculating the TMB score as the a number of somatic variants multiplied by 1 million bases and divided by the length of the Region of Interest (ROI) in megabases (Mb) minus the length of masking regions in megabases (Mb).

Calculate TMB Score is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Oncology Score Estimation** () | **Calculate TMB Score** ()

The tool takes a variant track as input.

In the next dialog, tracks relevant to the analysis are specified (figure 8.2):

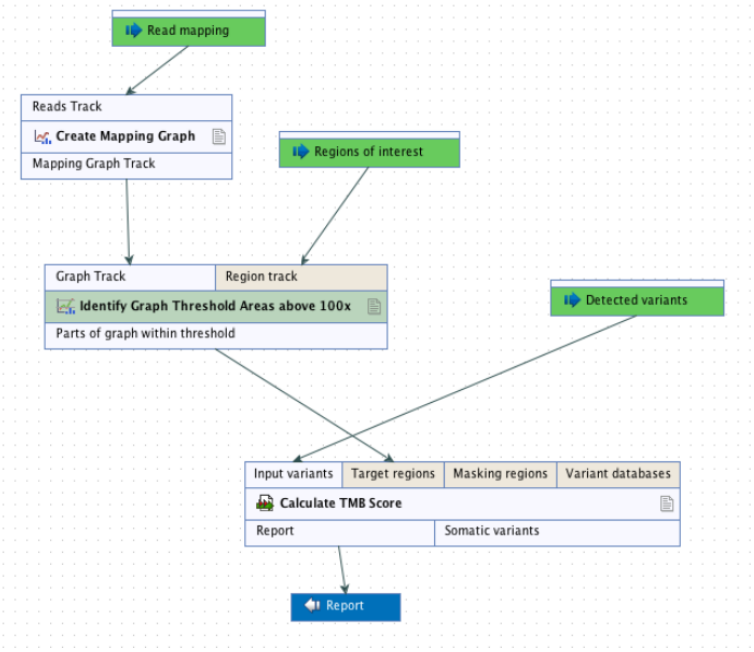


Figure 8.1: *Workflow to discard low coverage target regions.*

- **Target regions** A track containing the regions of interest.
- **Exon regions** An mRNA track containing the exons of interest
- **Masking regions** Regions that should not be considered

Only variants inside target regions and exons, and not within regions annotated on the masking track, are considered when calculating the TMB score.

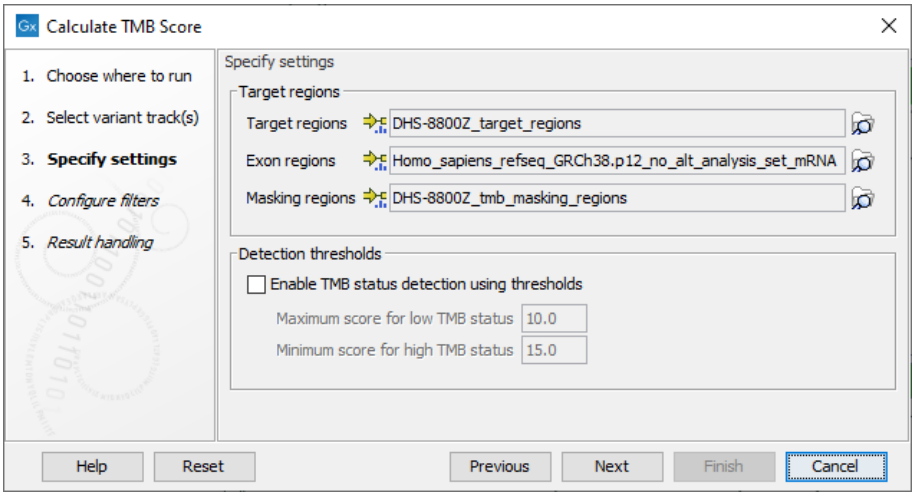


Figure 8.2: *Specifying tracks and parameters for calculating a TMB status.*

In addition, it is possible to enable the calculation of a TMB status based on a low and a high threshold, and which will appear as an additional item on the TMB report. The default values of 10 and 15 respectively have been chosen based on internal benchmark analyses of lung cancer

cell lines and different tissue cancer samples. Given the lack of standardization of methods and the heterogeneity of tumor mutation burden across many tumor types, it is difficult to establish cutoff values. Thresholds should be set according to the samples analyzed.

In the next dialog (figure 8.3), it is mandatory to provide a variant database of known germline variants as an input for filtering germline variants.

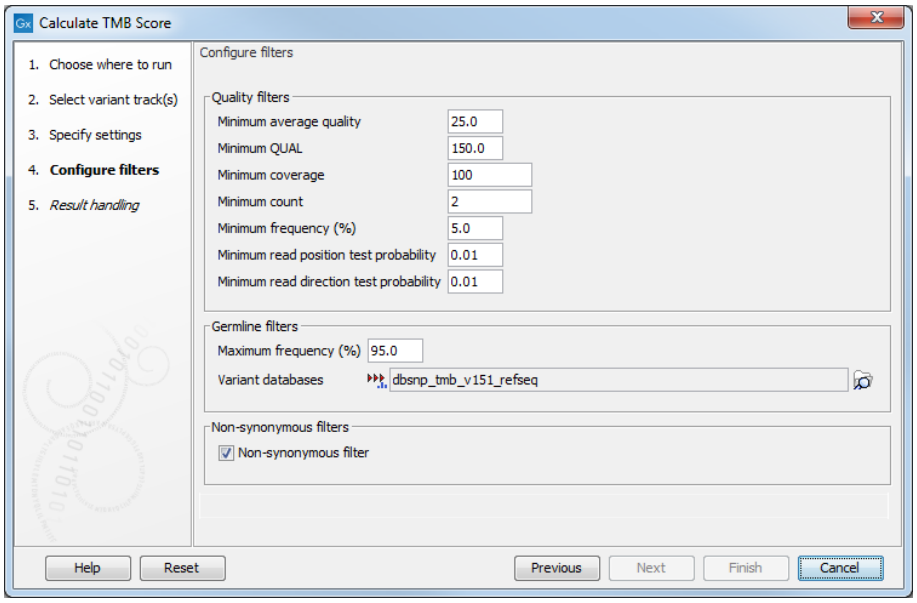


Figure 8.3: *Specifying tracks and parameters for calculating a TMB score.*

The parameters that can be configured are as follow:

- **Quality filters**
  - **Minimum average quality** The Avg Q of reads calculates the amount of sequences that feature individual PHRED-scores in 64 bins from 0 to 63. The quality score of a sequence is calculated as arithmetic mean of its base qualities. PHRED-scores of 30 and above are considered high quality.
  - **Minimum QUAL** Measure of the significance of a variant, i.e., a quantification of the evidence (read count) supporting the variant, relative to the coverage and what could be expected to be seen by chance, given the error rates in the data. The mathematical derivation depends on the set of probabilities of generating the nucleotide pattern observed at the variant site (1) by sequencing errors alone and (2) under the different allele models the variant caller allows. Qual is calculated as -10$\log_{10}$(1-p), p being the probability that a particular variant exists in the sample. Qual is capped at 200 for p=1, with 200: highly significant, 0: insignificant. In rare cases, the Qual value cannot be calculated for a specific variant and as a result the Qual field will be empty. This value is necessary for certain downstream analyses of the data after export in vcf format. A QUAL value of 10 indicates a 1 in 10 chance that the called variant is an error, while a QUAL of 100 indicates a 1 in 10^{10} chance that the called variant is an error.
  - **Minimum coverage** Only variants in regions covered by at least this many reads are called.

  - **Minimum count** Only variants that are present in at least this many reads are called.
  - **Minimum frequency (%)** The frequency is calculated as

$$count\ (reads\ having\ the\ variant)/coverage\ (reads\ covering\ the\ variant\ region) \tag{8.1}$$

    Only variants that are present at least at the specified frequency are called.
  - **Minimum read direction test probability** Tests whether the distribution among forward and reverse reads of the variant carrying reads is different from that of all the reads covering the variant position. This value reflects a balanced presence of the variant in forward and reverse reads (1: well-balanced, 0: un-balanced).
  - **Minimum read position test probability** Tests whether the distribution of the read positions in the variant carrying reads is different from that of all the reads covering the variant position.

- **Germline filters**
  - **Maximum frequency** Only variants whose frequency is equal to or lower than the specified value will be considered. Variants with a frequency above this value are considered germline.
  - **Variant databases** Specify a variant database such as dbSNP. Although dbSNP is thought to contain many erroneous calls, these may still be useful for removing variants that are not somatic, for example if they arise from common sequencing artifacts.
- **Non-synonymous filter** Only amino acids changing variants are kept and considered for the TMB score calculation.

Note that TMB filtering parameters are set conservatively. This is because for panels of 1MB size, a single false positive variant may increase the TMB score substantially.

The tool outputs a track of filtered somatic variants, i.e., the variants that remained after the filtering and that were included in the TMB score calculation. However, the main output is a report that includes filtering statistics and the calculated TMB score. It will also include a TMB status if the option was enabled (as shown in figure 8.4). By default, the TMB status is considered low if the TMB score is lower than 10; intermediate if the TMB score is between 10 and 15; and high if the TMB score is larger than 15. It is important to point out again that different cancer types have different somatic mutational load and thresholds should be set according to the samples analysed.

In addition, the report lists the length of the target regions, counts of various types of variants, and a value describing the tumor mutational burden calculated as the number of mutations per Mb. The quality filters statistics recapitulates how many variants were removed by the various filters applied by the tool, along with the frequency distributions of input and somatic variants.

The TMB status is assessed with a confidence level based on the size of the target regions included in the TMB score calculation, i.e., those with a coverage of at least 100X. This is illustrated by the color of the TMB status table cell in the report. If the analyzed target region size is below 900,000bp the cell will be colored in red, if it is between 900,000bp and 1,000,000bp it will be colored in yellow and if it is above 1,000,000bp it will not be colored. Note that if low coverage regions were not excluded from the target regions before TMB score calculation, the TMB status confidence level may wrongly be displayed as high.

1 Summary

TMB status	Low
Length of target regions (bp)	1,318,853
Variants inside target regions and after quality filters	84
Germline variants	72
Somatic variants	12
Non-coding somatic variants	0
Synonymous somatic variants	2
Non-synonymous somatic variants	10
Tumor mutational burden (mutations/Mb)	7.58

2 Quality filters statistics

Filter	Variants removed
Average quality filter	348
QUAL filter	7,931
Coverage filter	715
Count filter	0
Minimum frequency filter	2
Read position test probability filter	0
Read direction test probability filter	0

Figure 8.4: *A TMB report where the option to detect TMB status was enabled with default threshold values.*

8.2 Detect MSI Status

Detect MSI Status can be used to detect if a sample contains unstable microsatellites. It is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Oncology Score Estimation** () | **Detect MSI Status** ()

The tool detects whether a sample is stable or not by comparing it to a baseline composed of multiple microsatellite stable (MSS) samples. Baselines can be created using the **Generate MSI Baseline** tool (see section 8.3). This comparison is performed separately for each microsatellite locus and consists in evaluating whether the variations in the length distribution of the microsatellite observed in the sample are generally the same as the variations observed in the baseline samples.

We recommend that the MSI baseline is generated using samples that are sequenced under the same lab conditions as the sample for which the MSI status is calculated. The **Detect MSI Status** tool automatically inherits parameters from the selected MSI baseline and uses these parameters for generating a length distribution of the sample. This ensures that the length distributions are comparable between the baseline and the sample.

A microsatellite locus is said to be unstable when the length of the repeat region (e.g., tandem repeat of A nucleotides) is significantly different from the length in microsatellite stable (MSS) samples. To measure the locus lengths in a sample read mapping, the following steps are used:

1. For a given microsatellite locus, the flanking signature regions are identified in the reference genome on both sides of the locus. For example, if the flanking signature is 8 bp long and the sequence is GACTGCTGG**AAAAAAAAAA**TTTCGTAGC – where the sequence of repeated A's is the microsatellite – the left flanking signature is ACTGCTGG and the right flanking

signature is TTTCGTAG.

2. The tool searches for the left and right flanking signatures in all reads intersecting the locus.

3. The flanking signature might be present more than once in a read. This is increasingly likely with shorter flanking signatures. To account for this, we compare the nucleotide distribution for the microsattelite locus observed in the read to the reference sequence. This is done by determining the absolute difference in nucleotide fractions for each of the four nucleotides (A, C, G and T) between the reference sequence and the read sequence. Reads for which the sum for all four nucleotides is larger than 0.3 are removed. For example, for a 10 bp long homopolymer A region, the reference sequence has nucleotide fractions of 1.0 A's and 0.0 C's, G's, and T's. If a read has one A to C mismatch in the homopolymer region, its nucleotide fractions would be 0.9 A's, 0.1 C's, and 0.0 G's and T's. In this case the sum of the absolute differences would be 0.2, which is smaller than 0.3, hence the read is used in the MSI analysis.

4. For every read where both flanking signatures are identified and the nucleotide distribution of the locus is similar to the reference sequence, the length of the locus is used to update a frequency distribution of microsatellite locus lengths. A paired end read is only used if at least one of the reads in the read pair contains both flanking signatures. If, for example, read 1 contains only the left flanking signature (GACTGCTGGAAAAAA) and read 2 contains only the right flanking signature (AAAAAATTTCGTAGC), the read pair cannot be used for MSI detection since it is not possible to determine the length of the microsatellite. If both reads in the read pair contains both the left and right flanking signatures, the reads will count as one (and not two) in the frequency distribution, since the two reads originate from the same DNA fragment.

After counting the locus lengths in all reads, the statistical variation of the length distribution is calculated and compared to the baseline to determine if the locus is stable or unstable. If the proportion of unstable microsatellite loci is higher than a predefined threshold the sample is considered MSI-low or MSI-high depending on the settings.

When running **Detect MSI Status**, you will first need to select a read mapping. In the next dialog, specify an MSI baseline track (figure 8.5).

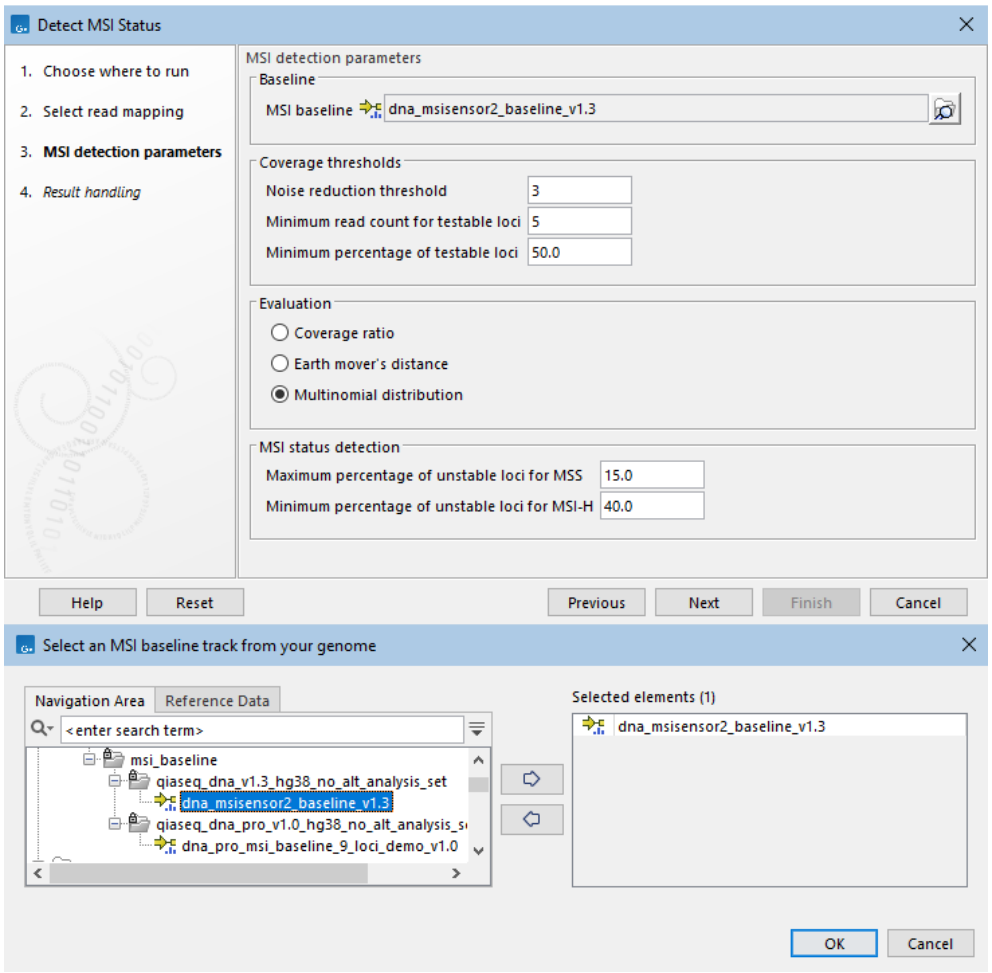


Figure 8.5: *Top: Parameters for Detect MSI Status. Bottom: MSI baselines from Reference Data Manager.*

The following baselines are available in the Reference Data Manager:

- dna_msisensor2_baseline_v1.3 is for QIAseq Targeted DNA panels, suitable for Human TMB and MSI Panel (DHS-8800Z) and Multimodal Pan-Cancer Panel (UHS-5000Z). The baseline is generated using 30 MSS samples that were mapped to hg38 (no alternative analysis set) and processed with the Generate MSI Baseline tool using default parameters and the msisensor2_loci_v1.0 loci track.
- dna_pro_msi_baseline_9_loci_demo_v1.0 is for QIAseq Targeted DNA Pro Panels (PHS-001Z, PHS-002Z, PHS-101Z, PHS-102Z, PHS-202Z, PHS-205Z, PHS-3000Z, PHS-3100Z, PHS-3200Z). The baseline is generated using 20 MSS samples that were mapped to hg38 (no alternative analysis set) and processed with the Generate MSI Baseline tool using default parameters and the qiaseq_msi_9_loci_v1.0 loci track. The baseline is only for demo use since it is generated with fewer than 30 samples.
- xHYB_CGP_msi_baseline_v1.0 is for QIAseq xHYB CGP DNA Panel. The baseline is generated using 30 MSS samples that were mapped to hg38 (no alternative analysis set) and processed with the Generate MSI Baseline tool using minimum read count of 1000 and the msisensor2_loci_v1.0 loci track.

The following parameters can be adjusted:

- **Noise reduction threshold** Locus lengths that are not supported by at least this number of reads are filtered away.
- **Minimum read count for testable loci** A locus is considered testable if the locus length can be determined in at least this number of reads after filtering away noisy locus lengths. If the read count is below this threshold, the locus will be evaluated as N/A.
- **Minimum percentage of testable loci** The MSI status for a sample is determined if at least this percentage of the loci is testable, i.e. has sufficient read count. If fewer loci are testable, the status is set to Undetermined.
- **Evaluation** The stability of the individual loci can be evaluated by three different methods. The coverage ratio method and the earth mover's distance method share the concept of a baseline length set. Each locus has its own baseline length set, containing locus lengths, that are found with a high frequency among the baseline samples. The baseline length set is created by determining all locus lengths, for which the frequency is at least 75% of the most frequent length in the distribution. This step is performed for all baseline samples individually, resulting in one or more locus lengths per sample, and the final baseline length set is created by combining the identified baseline lengths from all baseline samples.
  - **Coverage ratio** This method calculates the proportion of reads that have a microsatellite length, which is present in the baseline length set, relative to all reads. A Z-test is used to compare the sample to the baseline. The sample is evaluated as unstable if the coverage ratio of the sample is *smaller* than the average of baseline ratios minus three standard deviations.
  - **Earth mover's distance** This method measures the distance in the locus length distribution between the different bins (lengths) and the baseline length set. For each bin in the sample distribution, which is not in the baseline length set, we measure the distance to the closest baseline length bin and multiply it with the number of reads in the bin. Finally, the earth mover's distance is then calculated as the sum over all bins. A Z-test is used to compare the sample to the baseline. The sample is evaluated as unstable if the earth mover's distance of the sample is *larger* than the average of baseline distances plus three standard deviations.
  - **Multinomial distribution** This method estimates the probability for deletion from a multinomial distribution model. It compares the observed locus length to the length in the reference genome and calculates the probability of the locus being shorter than the reference length. A Z-test is used to compare the sample to the baseline. The sample is evaluated as unstable if the probability for deletion is *larger* than the average baseline probability plus three standard deviations.

  The coverage ratio and earth mover's distance methods can detect both microsatellite deletions and insertions, while the multinomial distribution is designed to only detect microsatellite deletions. Microsatellite instabilities are normally observed as deletions, however, if the aim is to detect insertions, the multinomial distribution method cannot be used.

The coverage ratio method is recommended for homogeneous baselines with relatively few baseline lengths forming a unimodal distribution. Earth mover's distance and multinomial distribution methods are recommended both for homogeneous and heterogeneous baselines, i.e. for unimodal or multimodal distribution of lengths.

By default, the multinomial distribution method is used, which gives good results across different QIAseq Targeted DNA and DNA Pro panels. We recommend validating the performance of a given MSI baseline using samples with known MSI status. If the results differ from the expectations, you can try another method and/or adjust the other parameters.

- **MSI status detection** Given the stability of the individual loci, the MSI status of a sample is determined based on whether the percentage of unstable loci is higher than a predefined threshold. There are two thresholds: one for low instability (MSI-L) and one for high instability (MSI-H). The percentages are calculated relative to the number of testable loci, as described above.
  - **Maximum percentage of unstable loci for MSS**, set by default at 15%
  - **Minimum percentage of unstable loci for MSI-H**, set by default at 40%

8.2.1 Output from Detect MSI Status

Three outputs are produced by the **Detect MSI Status** tool:

- **MSI loci track** An annotation track with MSI loci annotated with predicted stability.
- **MSI report** A report summarizing the overall status of the sample and showing length distribution plots of the individual loci.
- **Baseline cross-validation report** A report analyzing the quality of the MSI baseline.

The MSI report contains both combined and per loci information on stability and other descriptive statistics. The summary section contains information about the number of stable and unstable loci, as well as the MSI status of the sample (figure 8.6).

The loci overview section provides details about the analyzed loci and their stability. The table contains the following information:

- **Locus** Name of the locus. The name links to a plot showing the distribution of locus lengths for this locus.
- **Coverage** The number of reads intersecting the locus.
- **Read count** The number of reads that contains both flanking signatures and have been used for the calculation of the frequency distribution of microsatellite locus lengths as described above.
- **Baseline lengths** The set of baseline lengths used for determining the stability with coverage ratio and earth mover's distance methods. The column is not shown for the multinomial distribution method.

1 Summary

The MSI status was detected using Multinomial distribution method.
The MSI status of the sample can only be determined with at least 50.0% testable loci.

Sample name	L005_ILB06_16OCT18_SFS_S7
Number of loci	120
Number of unstable loci	107
Number of stable loci	11
Number of not testable loci	2
Percentage of unstable loci	90.68
MSI status	MSI-H
Clinical term	MSI-high

2 Loci Overview

Locus	Coverage	Read count	p-value	p-value threshold	Stability
1_7920926_10[A]	263	164	0.091	0.020	Unstable
1_11233311_11[A]	46	22	0.215	0.011	Unstable
1_15922234_11[T]	397	296	0.057	0.036	Unstable
1_77966823_14[A]	140	81	0.261	0.042	Unstable
1_77966964_11[A]	424	334	0.094	0.030	Unstable

3 Loci Length Distributions

3.1 1_7920926_10[A]

Stability: Unstable

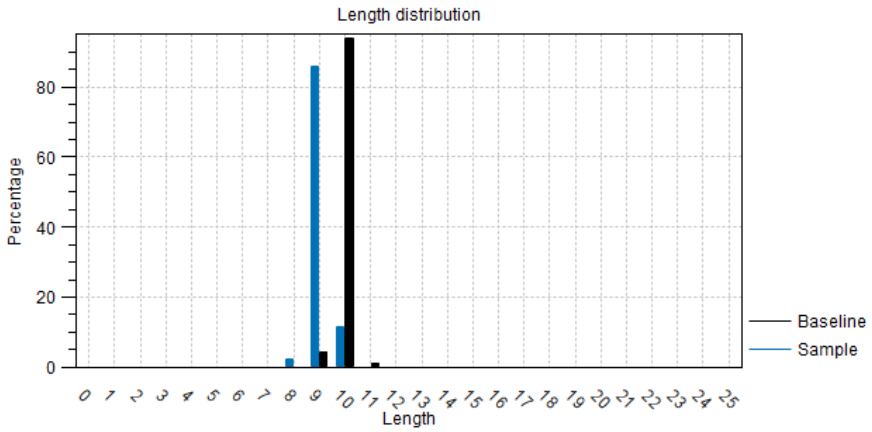


Figure 8.6: *Summary section from the MSI Report for an MSI-high sample analyzed using the coverage ratio method.*

- **Coverage ratio / Earth mover's distance / p-value** Stability value from the method of choice, see section 8.2 for details on how the metric is calculated.
- **Stability threshold / p-value threshold** Threshold calculated from the baseline used to assess the stability of the locus. A locus is unstable if the test sample has a stability value:
  - below this threshold for the coverage ratio method.
  - above this threshold for the earth mover's distance method.
  - above this threshold for the multinomial distribution method.

- **Stability** The stability of a locus can be stable, unstable or N/A. The stability is set to N/A if either the sample or the baseline has insufficient read count.

If the read count is low but the coverage is high, it could be an indication that the locus is highly unstable and only few reads are spanning the locus. Investigating the read mapping can help understanding the problems.

Figure 8.6 shows an example of an MSI report (the loci overview table is truncated by the dashed line), where a sample is compared to the dna_msisensor2_baseline_v1.3 baseline from the Reference Data. The baseline has 120 loci, where two of them are not testable due to too few reads. 107 of the remaining 118 loci are unstable, meaning that the overall assessment of the sample is MSI-high.

The length distribution plot compares the loci lengths observed for the sample (blue) and the baseline (black) for the locus 1_7920926_10[A]. The baseline distribution shows that >90% of the reads have a length of 10 bp, while 85% of reads in the sample have a length of 9 bp. The length distributions are significantly different between the sample and the baseline, and the locus is therefore evaluated as unstable.

The baseline cross-validation report (not shown) contains a table where the MSI status is presented for each sample in the baseline sample set. The cross-validation analysis verifies whether the baseline and selected parameters are suitable. For this, the MSI status of each sample from the baseline sample set is tested against a baseline created using all other samples of the set. Ideally, it is expected that all samples will be detected as stable (MSS) with a very low proportion of unstable loci. If this is not the case, the parameters might need to be adjusted and/or one or more samples should be removed from the baseline. Note that the cross-validation analysis is dependent on the parameters used for detection (exactly as for a test sample) and therefore each cross-validation is only valid for the selected set of parameter values used in the cross-validation run.

8.3 Generate MSI Baseline

The **Generate MSI Baseline** tool can be used to generate microsatellite instability (MSI) baseline tracks that are used when running **Detect MSI Status**. The tool is available from the Tools menu:

> **Tools** | **Biomedical Genomics Analysis** () | **Oncology Score Estimation** () | **Generate MSI Baseline** ()

The tool can generate a baseline by:

- Using an annotation track containing microsatellite loci. The annotation track can be:
  - A track containing specific loci targeted by the panel.
  - A baseline track generated by this tool.
- Scanning the reference genome for microsatellite loci. This increases the tool's runtime.

The tool requires at least five read mappings from microsatellite stable (MSS) samples as input. For a reliable baseline we recommend at least 30 samples, since the **Detect MSI Status** tool uses a Z-test to compare a test sample to the baseline.

The following options can be adjusted (figure 8.7):

- **MSI loci from track** Generate baseline for microsatellite loci in an annotation track.
  - **MSI loci track** An annotation track containing microsatellite loci. Three MSI loci tracks are available in the Reference Data Manager, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QIAGEN_Sets.html.
    - msisensor2_loci_v1.0 contains 2828 mono- and dinucleotide loci from msisensor2, see https://github.com/niu-lab/msisensor2/.
    - qiaseq_msi_9_loci_v1.0 contains 9 loci from the QIAseq MSI booster Panel (SDHS-10101-11981Z-48). These 9 loci are a subset of the 27 loci in qiaseq_msi_27_loci_v1.0.
    - qiaseq_msi_27_loci_v1.0 contains 27 loci from the QIAseq MSI booster Panel (SDHS-10101-11981Z-48).
- **Scan target regions or whole genome** Generate baseline for microsatellite loci that are automatically detected in the reference genome, by scanning the whole genome or just target regions. Scanning the whole genome increases the runtime.
  - **Target regions track** An optional annotation track containing non-overlapping target regions for scanning. Target regions can, for example, be regions in the genome that have coverage in the input MSS samples. Note that targeted DNA panels have their own specific target regions, typically covering hotspots and/or entire exons. As microsatellite loci are often intronic or intergenic, such panel target regions are generally unsuitable for scanning for microsatellite loci. Overlapping target regions can be collapsed using Collapse Overlapping Annotations, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Collapse_Overlapping_Annotations.html.
  - **Minimum locus length** Only loci with a length equal to or greater than this value are kept.
  - **Maximum repeat unit size** Only loci with the length of the repeat unit equal to or shorter than this value are kept. For example, 1 corresponds to only homopolymer loci.
  - **Minimum repeat times** Only loci where the repeat unit appears at least this number of times are kept.
- **Minimum read count** A locus is included in the baseline if its length can be determined from at least this many reads in the input read mappings. The length can be determined when the read spans the locus and includes both the left and right flanking signatures.
- **Flanking signature length** The length of the flanking signatures. The flanking signature should be long enough to be unique in the read, but short enough to be present in as many reads as possible.
- **Allow one mismatch in the flanking signature**: When checked, the flanking signature can contain one SNP compared to the reference genome.
- **Ignore broken pairs** When checked, broken pairs in paired end reads are not used for baseline generation.

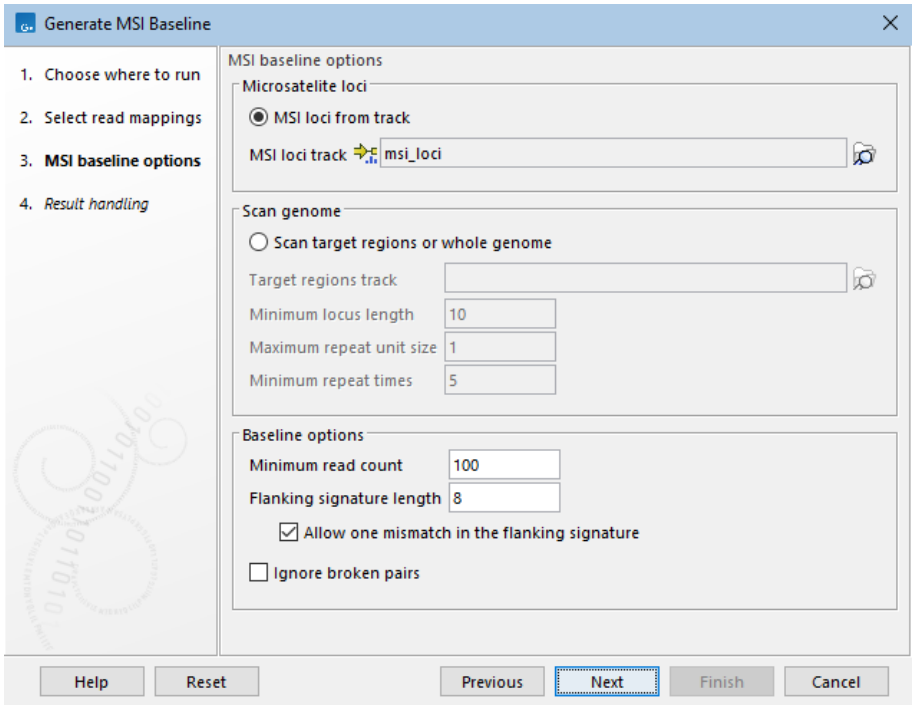


Figure 8.7: *Parameters for Generate MSI Baseline.*

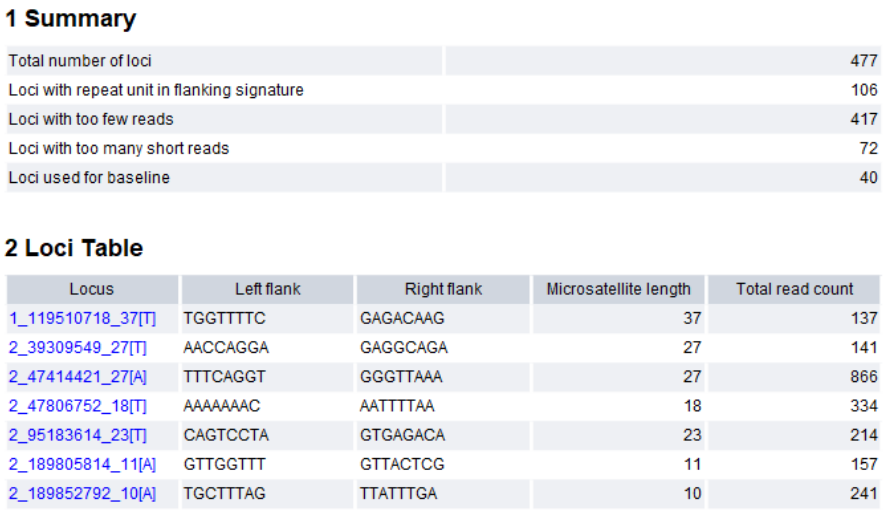
If **Scan target regions or whole genome** is selected, the tool first identifies a list of candidate microsatellite loci, which are otherwise provided in the **MSI loci track**. Subsequently, the tool extracts all reads overlapping the loci and analyzes the locus length by identifying the flanking signatures in the reads. See section 8.2 for more details about how the locus length is determined. Finally, the loci are filtered in three steps. A locus is removed if:

- It does not meet the **Minimum read count**. The report lists the number of loci filtered out due to this under **Loci with too few reads**.
- One of the flanking signatures is identical to the locus sequence except for one mismatch. Such flanking signatures can lead to an incorrect locus length. The report lists the number of loci filtered out due to this under **Loci with repeat unit in flanking signature**.
- At least 25% of reads have a locus length of 3 bp or less. A high percentage of short locus lengths suggests that the flanking signatures have been incorrectly determined. The report lists the number of loci filtered out due to this under **Loci with too many short reads**.

Generate MSI Baseline outputs an MSI baseline track and a report summarizing the loci in the baseline (figure 8.8).

Undesired loci can be manually removed from the baseline track by:

- Opening its table view ( ).
- Selecting the loci to be kept.
- Creating a new track using the **Create Track from Selection** button.



1 Summary

Total number of loci	477
Loci with repeat unit in flanking signature	106
Loci with too few reads	417
Loci with too many short reads	72
Loci used for baseline	40

2 Loci Table

Locus	Left flank	Right flank	Microsatellite length	Total read count
1_119510718_37[T]	TGGTTTTC	GAGACAAG	37	137
2_39309549_27[T]	AACCAGGA	GAGGCAGA	27	141
2_47414421_27[A]	TTTCAGGT	GGGTTAAA	27	866
2_47806752_18[T]	AAAAAAAC	AATTTTAA	18	334
2_95183614_23[T]	CAGTCCTA	GTGAGACA	23	214
2_189805814_11[A]	GTTGGTTT	GTTACTCG	11	157
2_189852792_10[A]	TGCTTTAG	TTATTTGA	10	241

Figure 8.8: *MSI baseline report obtained by scanning target regions for microsatellite sites.*

The report contains a summary section with the number of unfiltered and filtered loci. The **Total number of loci** is either the number of loci in the provided **MSI loci track**, or the number of loci initially identified by scanning.

The loci table contains all loci in the baseline, with the following columns:

- **Locus** Locus name. This links to the plot showing the distribution of locus lengths. The name is obtained either from the provided **MSI loci from track** or has the form {chromosome name}_{start position}_{repeat times}[repeat unit] when scanning.
- **Left flank** Left flanking signature used for identifying the locus length.
- **Right flank** Right flanking signature used for identifying the locus length.
- **Microsatellite length** Length of the microsatellite locus in the reference genome.
- **Total read count** Total number of reads in which the locus length could be determined.

8.4 Calculate HRD Score (beta)

The **Calculate HRD Score (beta)** tool is designed to calculate Homologous Recombination Deficiency (HRD) from targeted research resequencing experiments.

The tool takes a target-level ploidy track (from the Detect Regional Ploidy tool) and centromers as input.

Calculate HRD Score (beta) is available from the Tools menu at:

> **Tools** | **Biomedical Genomics Analysis** () | **Oncology Score Estimation** () | **Calculate HRD Score (beta)** ()

Select the target-level ploidy track generated by the Detect Regional Ploidy tool and click **Next**.

You are now presented with choices regarding HRD calculation.

- **Centromeres** Centromeres are used to define the chromosome arms and must be provided for correct calculation of the HRD score, as LOH, LST and TAI events are calculated per chromosome arm. Regions covered by provided centromeres are excluded from the HRD calculation.
- **LOH weight**, **LST weight**, **TAI weight** HRD is calculated as the sum of Loss of Heterozygosity (LOH), Large-scale State Transitions (LST), and Telomeric Allelic Imbalance (TAI) scores. These three scores can be weighted by the values given here before being summed to the final HRD score.
- **Minimum LOH region length (MB)** Loss of Heterozygosity (LOH) regions shorter than this are ignored in the LOH score calculation. The length is given in megabases.
- **Minimum LST region length (MB)** Large-scale State Transitions (LST) between regions shorter than this are ignored in the LST score calculation. The length is given in megabases.
- **Maximum LST region distance (MB)** Large-scale State Transitions (LST) with a distance larger than this are ignored in the LST score calculation. The distance is given in megabases.
- **Short LST region length (MB)** Regions shorter than this are filtered out before Large-scale State Transitions (LST) score calculation. The length is given in megabases.
- **Minimum merge size** Remove merged regions consisting of fewer than this number of targets.

When finished with the settings, click **Next** to start the algorithm.

HRD calculation

The HRD score is a count of chromosomal rearrangements that can be increased in tumors with HRD. It is calculated as the weighted sum of three different chromosomal rearrangements: The number of Telomeric Allelic Imbalances (TAI), Large-scale Transitions (LST), and long regions of Loss of Heterozygosity (LOH). The calculations are based on identified regions of copy number variations (CNV) as well as variant frequencies in a sample, which are identified beforehand.

The LOH score is defined as the number of regions with a minor allele count of zero. If all regions on a chromosome are affected by LOH, the whole chromosome is excluded from the LOH score calculation.

The TAI score is defined as the number of regions that:

- (1) Show an imbalance from the most prevalent copy number state of the whole chromosome. Whether a telomeric region is imbalance with the rest of the chromosome, is determined by comparing the copy number state of the region nearest the telomere with the most prevalent copy number state of the whole chromosome.
- (2) Do not cross the centromere.
- (3) Extend up to the telomeres. Note, the region located closest to the end of a chromosome is considered a proxy for the telomeric region. The CNV regions underlying TAI are filtered for a minimum number of probes (merge size) and a minimum region length.

Calculation of the three scores is inspired by:

- **Abkevich et al.** *Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer*, British Journal of Cancer. 2012, 107(10): 1776-1782. [Abkevich et al., 2012]
- **Birkbak et al.** *Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA damaging agents*, Cancer Discovery. 2012, 2(4): 366-375. [Birkbak et al., 2012]
- **de Luca et al.** *Using whole-genome sequencing data to derive the homologous recombination deficiency scores*. npj Breast Cancer. 2020, 6:33. [de Luca et al., 2020]

The LST score is the number of LST events. The LST score counts large rearrangements for each arm of a chromosome. The regions are merged and short regions removed iteratively. For each chromosome arm, as long as there are segments less than 3 MB, the segment at the first position, that is less than 3 MB is removed and adjacent segments across the whole chromosome arm with identical allele counts merged.

Calculate HRD score algorithm report

The report provides a table listing:

- **HRD score** The HRD score calculated from the LOH, LST and TAI scores.
- **HRD score LOH/LST/TAI weight** The weight that each of the underlying scores were given when calculating the HRD score.
- **Genomic coverage** The percentage of genomic positions covered by the panel used for HRD detection.
- **LOH score** The number of LOH regions.
- **LOH regions** The positions of the identified LOH regions. For each region, the chromosome and the start and end of the LOH region is included. As an example, the entry "2: 151M 169M" should be read as an LOH event on chromosome 2 occurring from position 51M to 169M.
- **LST score** The number of LST regions.
- **LST** The positions of the identified LST regions. As an example, in the entry "S1: 1-2 0M 13M -> 1-1 13M 248M" the parts before and after the arrow describe the chromosomal states on each side of the transition and should be read as: Start of chromosome 1, minor allele count 1, major allele count 2, positions 0M-13M changes to minor allele count 1, major allele count 1, positions 13M to 248M.
- **TAI score** The number of TAI regions.
- **TAI** The positions of the identified TAI regions. As an example "S1 TAI 2 1-2", should be read as start of chromosome 1, TAI event, most prevalent copy number state for the whole chromosome is 2, for the TAI event minor allele count is 1 and major allele count is 2. Correspondingly, "E1 CENT 125M 248M" should be read as end of chromosome 1, region extends from end of chromosome to centromere and is not counted as TAI,

positions 125M-248M and "E10 NO 2 1-1" should be read as end of chromosome 10, no TAI event, most prevalent copy number state for the whole chromosome is 2, and for the region closest to the end of the chromosome minor allele count is 1 and major allele count is 1. Hence, a TAI event is only counted when TAI is part of the annotation for a given chromosome arm.

Chapter 9

IPA and QCI Interpret Upload

Contents

The Biomedical Genomics Analysis plugin contains tools for uploading data from the CLC Workbench to QIAGEN Ingenuity Pathway Analysis (IPA) and QIAGEN Clinical Insight (QCI) Interpret.

9.1 Upload to IPA

QIAGEN Ingenuity Pathway Analysis (IPA) aids the interpretation of differential expression results by identifying enriched pathways and determining the activity of upstream regulators. It assesses potential impacts on downstream diseases, biological functions, and phenotypes. For more information, please visit https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/.

Upload to IPA takes statistical comparison tables ( ) or tracks ( ) as input and uploads them to IPA.

To run the tool, go to:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **IPA and QCI Interpret Upload** ( ) | **Upload to IPA** ( )

After selecting the inputs, use the **IPA login** wizard to log in to IPA. Select the relevant **IPA server** (US or China) and click the **Log in** button. This opens an IPA login page in an external browser. Once authenticated, the following options can be configured in the **Configuration** wizard (figure 9.1):

- **Project Name** The name of the project in IPA. The placeholder {1} will be substituted with the current date in the format YYYY-MM-DD.

- **Reference data** The reference data can be set to:
  - **Ingenuity Knowledge Base (Genes only)** The Ingenuity Knowledge Base is used as reference data. This is only suitable for inputs pertaining to gene expression. See section 9.1.1 for details.
  - **Uploaded dataset** The dataset itself is used as reference data.
- **Upload only** Only upload the dataset to IPA.
- **Upload and analyze** Upload to IPA and perform pathway analysis on features meeting the filter criteria specified using the following options:
  - **Minimum mean expression value** Only features with at least this "Max group mean" are analyzed.
  - **Maximum p-value** Only features with at most this p-value are analyzed. The **P-value type** can be set to **P-value**, **FDR p-value**, or **Bonferroni**.
  - Only features with an absolute fold change above a certain threshold are analyzed. The **Fold change type** can be set to **Fold change** or **Log$_2$ fold change**.
    - **Automatically calculate fold change threshold** If checked, the threshold is calculated for each input separately, such that the number of analyzed features is as close as possible to the **Target number of features**.
    - **Fold change threshold** The provided threshold is used.

Note that all features in the input are uploaded. To limit statistical comparisons to only the differentially expressed features, use **Filter on Custom Criteria**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html.

The summary at the bottom of the wizard (figure 9.1) provides an overview of the number of features to be uploaded, and, if applicable, the number of features to be analyzed and the automatically calculated fold change threshold.

9.1.1 Upload using the Ingenuity Knowledge Base

IPA recognizes gene ids for several species, see https://qiagen.my.salesforce-sites.com/KnowledgeBase/KnowledgeNavigatorPage?id=kA41i000000L6BTCA0&categoryName=IPA for a full list. Successive uploads to IPA are attempted, until the upload is successful:

- If the statistical comparison contains gene ids from a recognized database (Ensembl, Entrez, Hugo, or RefSeq), the ids are uploaded to IPA using the corresponding IPA identifier type. Otherwise, the 'Name' column is uploaded with IPA identifier types for Ensembl, Entrez, GenBank, miRBase (mature) and RefSeq.
- If the previous upload fails, it could be because the gene identifiers of the uploaded species are not supported by IPA. Gene names are often conserved across species, so uploads are attempted with gene names formatted according to the IPA human (upper case e.g. BRCA1) or mouse/rat (capitalized e.g. Brca1) gene names formats, using the corresponding IPA gene symbol identifier types. Uploads are performed in decreasing order of the number of unformatted gene names matching the human and mouse/rat formats.

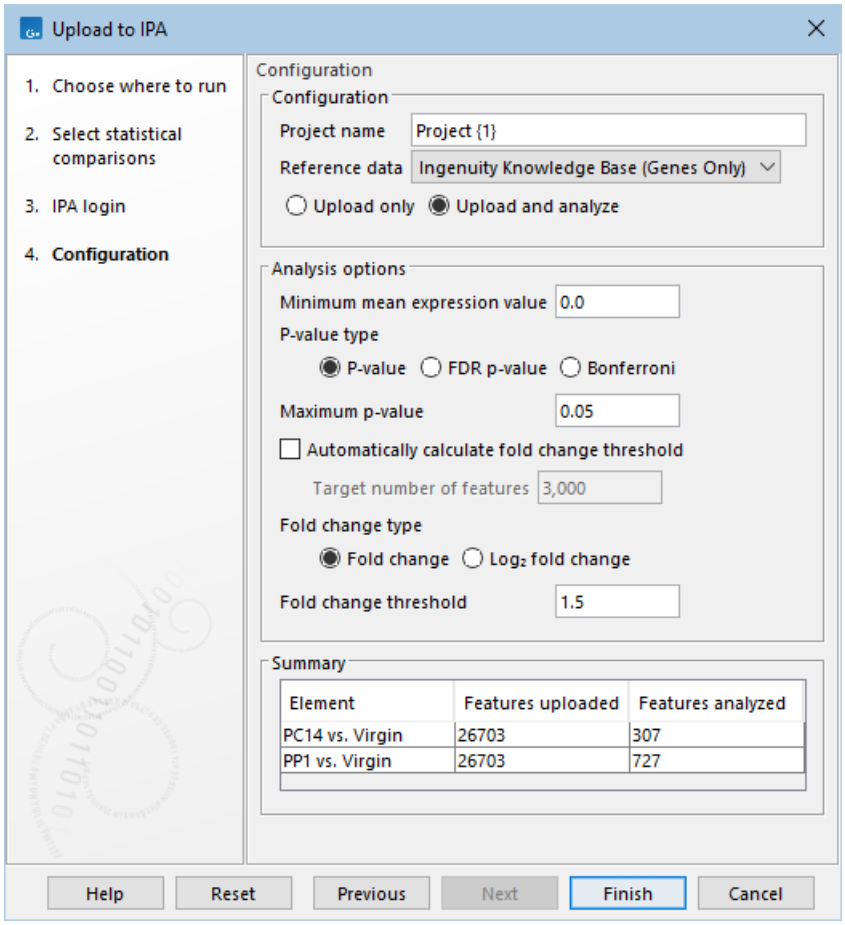


Figure 9.1: *Configuration options for Upload to IPA.*

Note that upload is successful even if just one gene has been successfully identified by IPA. All performed uploads and their error messages from IPA for failed uploads are written to the log.

If all upload attempts fail, the upload errors from IPA will be displayed. The errors might indicate that the species is not supported by IPA and the gene names did not match any of the human, mouse and rat genes.

9.1.2 Error handling

Concurrent IPA sessions

The IPA upload may fail if multiple sessions are active for the same username. This can occur if the IPA application is running while data is being uploaded via this tool or if the tool is part of a workflow with several simultaneous uploads.

To address upload failures caused by concurrent sessions, the tool automatically retries the upload. The waiting time between attempts is progressively increased with a random factor to prevent processes from blocking each other.

Upload multiple statistical comparisons

If the tool encounters an error during the upload of multiple statistical comparisons, it typically proceeds with the remaining uploads. However, if it receives one of the following errors from IPA, it immediately stops the upload, as these issue is unlikely to be resolves before the next upload attempt:

- Login error (invalid or expired login)
- User agreement not accepted
- License expired
- Upload limit exceeded
- Analysis limit exceeded

9.2 QCI Interpret Upload

QCI Interpret uses the QIAGEN Knowledge Base - a database that contains over 5 million variant findings as well as data from third party databases - and rules-based approaches to automatically compute pathogenicity classifications (Pathogenic to Benign) and actionability classifications (Tier 1 to 4) for each found variant. Pathogenicity and actionability classifications in QCI Interpret are accompanied by clear visibility into the criteria and evidence supporting the classifications. The final report is sample-specific and includes clinically-relevant variants, interpretations, approved or investigational therapies, and references specified throughout the assessment process.

The Upload to QCI Interpret tools support seamless upload of analysis results to QCI Interpret.

It is possible to upload all variants that can be exported to VCF from the CLC Workbench, as well as TMB and MSI scores. In addition, it is possible to upload CLC reports, which can then be accessed as PDFs in the QCI Interpret platform.

Use of these tools require an active QCI Interpret or QCI Interpret Translational account.

9.2.1 Prepare QCI Interpret Upload

The **Prepare QCI Interpret Upload** tool prepares a sample from several types of tracks and reports that can be uploaded to QCI Interpret or QCI Interpret Translational at a later time with **Upload Prepared QCI Interpret Report**.

This tool can be used in workflows to select all relevant tracks and reports. It sets name based on input or other metadata and outputs a single report that is ready to be uploaded to QCI Interpret. This is convenient because users won't have to choose multiple track inputs for every workflow output they want to upload to QCI Interpret.

To run the tool, go to:

> **Tools** | **Biomedical Genomics Analysis** () | **IPA and QCI Interpret Upload** () | **QCI Interpret Upload** () | **Prepare QCI Interpret Upload** ()

In the tool wizard, choose which elements to upload.

The first step defines inputs. Supported input elements are:

- **Variant track(s)** Add variant tracks to include variants. Choosing multiple variant tracks exports all variants in the variant tracks, for duplicates only the variant with the highest quality is exported.
- **CNV track** Add a CNV track to include copy number variations.
- **Fusion track** Add a fusion track to include fusions.
- **Inversion track** Add an inversion track to include inversions.

The next step is to define how the QCI Interpret sample will look in QCI:

- **Sample name** The name of the uploaded sample shown in QCI Interpret. Leaving it empty will use a name of a element. Use metadata if running as part of a workflow (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Configuring_Workflow_Output_Export_elements.html)
- **Subject ID** Set the subject ID, leave empty, or use metadata if running as part of a workflow.
- **Project** Set the project, leave empty or use metadata if running as part of a workflow.

The next step defines a reference and extra values that can be added to the sample:

- **Reference sequence track** To generate the VCF the tool needs a reference track. Hg19 and hg38 references are supported.
- **TMB report** Include tumor mutation burden value and status by including a report output from the **Calculate TMB Score** tool in the Biomedical Genomics Analysis plugin. The algo must have been set up to detect a TMB status.
- **MSI report** Include microsatellite instability status by including a report output from the **Detect MSI Status** tool in the Biomedical Genomics Analysis plugin.
- **Upload reports** Include these reports as PDF in the sample uploaded to QCI Interpret.

The next step defines how the uploaded VCF will be created. The Prefill VCF settings changes the VCF settings to suggested defaults depending on the data. The somatic and germline/hereditary workflows of QCI Interpret focus on different needs, with somatic being primarily focused on therapeutic, prognostic, and diagnostic actionability, while germline/hereditary is better suited for disease diagnosis/risk. Read more about VCF export settings in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Export_in_VCF_format.html.

Output

The output is a report that shows a summary of the data in the sample.

9.2.2 Upload Prepared QCI Interpret Report

The **Upload Prepared QCI Interpret Report** tool uploads a sample to QCI Interpret or QCI Interpret Translational from a prepared upload from the **Prepare QCI Interpret Upload** tool.

To run the tool, go to:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **IPA and QCI Interpret Upload** ( ) | **QCI Interpret Upload** ( ) | **Upload Prepared QCI Interpret Report** ( )

In the tool wizard, choose a QCI Interpret report from **Prepare QCI Interpret Upload**. You can also choose a already upload QCI Interpret report to upload it again.

Then choose your QCI Region and how to log in. To log in using a browser, choose the Browser option, then click the **Log in** button to open a new browser (or new tab) where you can log in, and give the CLC Workbench permission to upload samples to QCI on your behalf. To upload with API key you need an API user account, please send an email to bioinformaticslicense@qiagen.com and ask for confirmation if the QCI account is API-enabled. The license team can then provide the link to API Explorer page where you can log in to see the API key ID and key secret.

Note: Do not share workflows that contain your QCI Interpret login information. A workflow containing this tool contains your login information if the tool configuration has values set for API Key ID, or API Key Secret, or if you have logged in using a browser.

The uploaded sample can be shared with QCI users by adding a comma separated list of emails in the **Reviewers** setting.

Output

The output is a report that shows a summary of the data in the sample, information about the upload, and a link to the sample list in QCI Interpret.

9.2.3 Upload to QCI Interpret

The **Upload to QCI Interpret** tool can upload variants as well as TMB and MSI scores to QCI Interpret or QCI Interpret Translational. To facilitate the upload, variant tracks are first exported to VCF and then uploaded using provided login details.

To run the tool, go to:

> **Tools** | **Biomedical Genomics Analysis** ( ) | **IPA and QCI Interpret Upload** ( ) | **QCI Interpret Upload** ( ) | **Upload to QCI Interpret** ( )

In the first wizard step, select the elements containing the variants that should be uploaded (see figure 9.2). Supported input elements are:

- **Variant track(s)** Add variant tracks to include variants in the upload. If multiple variant tracks are selected, all variants in the variant tracks are uploaded except duplicate variants, where only the variant with the highest quality is included.
- **CNV track** Add a CNV track to include copy number variations.
- **Fusion track** Add a fusion track to include fusions.

- **Inversion track** Add an inversion track to include inversions.

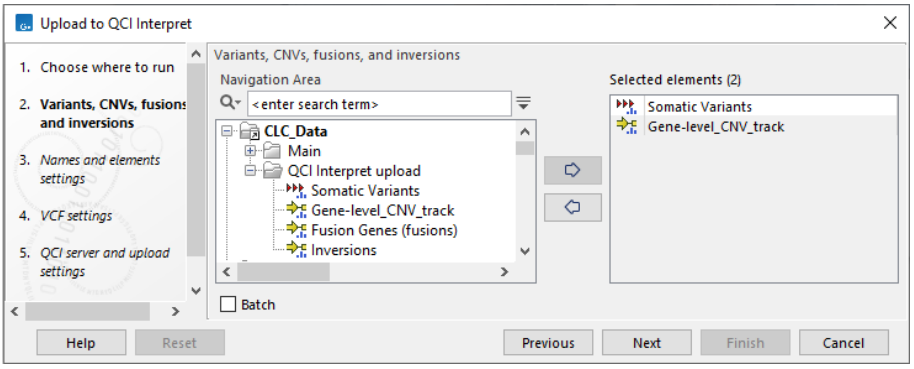


Figure 9.2: *Select the elements containing variants, CNVs, fusions and inversions that should be uploaded.*

In the next step, specify information that should be displayed in QCI Interpret (see figure 9.3). Under **Name**, if a field is left empty, the name of the first selected element is used as the sample name. It is possible to use metadata to define the sample name if the tool is run as part of a workflow (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Configuring_Workflow_Output_Export_elements.html).

- **Sample name** The name of the uploaded sample that will be shown in QCI Interpret.
- **Subject ID** The subject ID of the uploaded sample that will be shown in QCI Interpret.
- **Project** The project of the uploaded sample that will be shown in QCI Interpret.

Under **Elements**, specify the reference sequence, reports containing oncology scores that should be included in the upload and add any reports that should be available as PDFs in QCI Interpret:

- **Reference sequence track** To generate the VCF the tool needs a reference track. Hg19 and hg38 references are supported.
- **TMB report** Include tumor mutation burden value and status by including a report output from the **Calculate TMB Score** tool in the Biomedical Genomics Analysis plugin. The algo must have been set up to detect a TMB status.
- **MSI report** Include microsatellite instability status by including a report output from the **Detect MSI Status** tool in the Biomedical Genomics Analysis plugin.
- **Upload reports** Include any report. Specified reports will be shown as PDFs in QCI Interpret.

The next step defines how the uploaded VCF will be created. The Prefill VCF settings changes the VCF settings to suggested defaults depending on the data. The somatic and germline/hereditary workflows of QCI Interpret focus on different needs, with somatic being primarily focused on therapeutic, prognostic, and diagnostic actionability, while germline/hereditary is better suited for disease diagnosis/risk. Read more about VCF export settings in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Export_in_VCF_format.html.

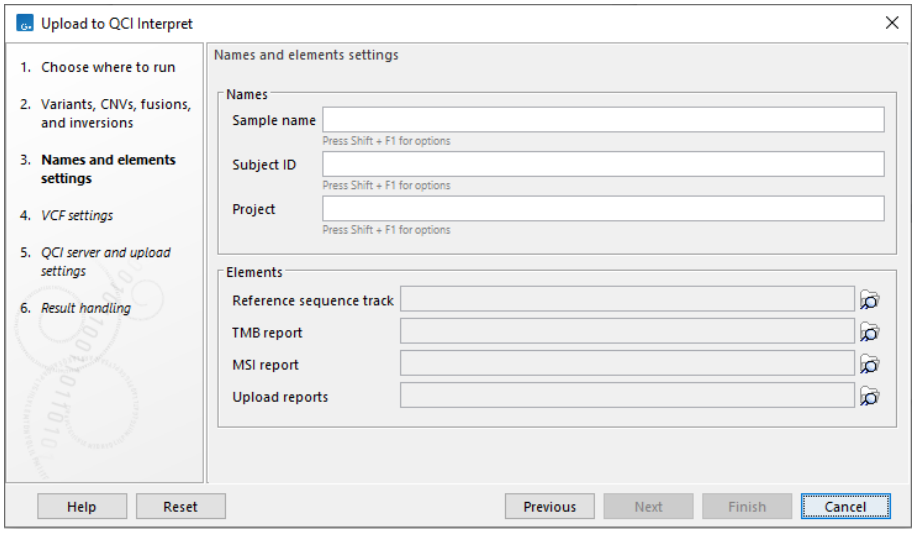


Figure 9.3: *Provide information that should be shown in QCI Interpret, the reference sequence for the uploaded variants, oncology scores and any reports that should be available as PDFs in QCI Interpret.*

In the final step choose your QCI Region and how to log in. See figure 9.4. To log in using a browser, choose the Browser option, then click the **Log in** button to open a new browser (or new tab) where you can log in, and give the CLC Workbench permission to upload samples to QCI on your behalf. To upload with API key you need an API user account. Please send an email to bioinformaticslicense@qiagen.com and ask for confirmation if the QCI account is API-enabled. The license team can then provide the link to API Explorer page where you can log in to see the API key ID and key secret.

Note: Do not share workflows that contain your QCI Interpret login information. A workflow containing this tool contains your login information if the tool configuration has values set for API Key ID, or API Key Secret, or if you have logged in using a browser.

The uploaded sample can be shared with QCI users by adding a comma separated list of emails in the **Reviewers** setting.

Output

The output is a report that shows a summary of the data in the sample, information about the upload, and a link to the sample list in QCI Interpret.

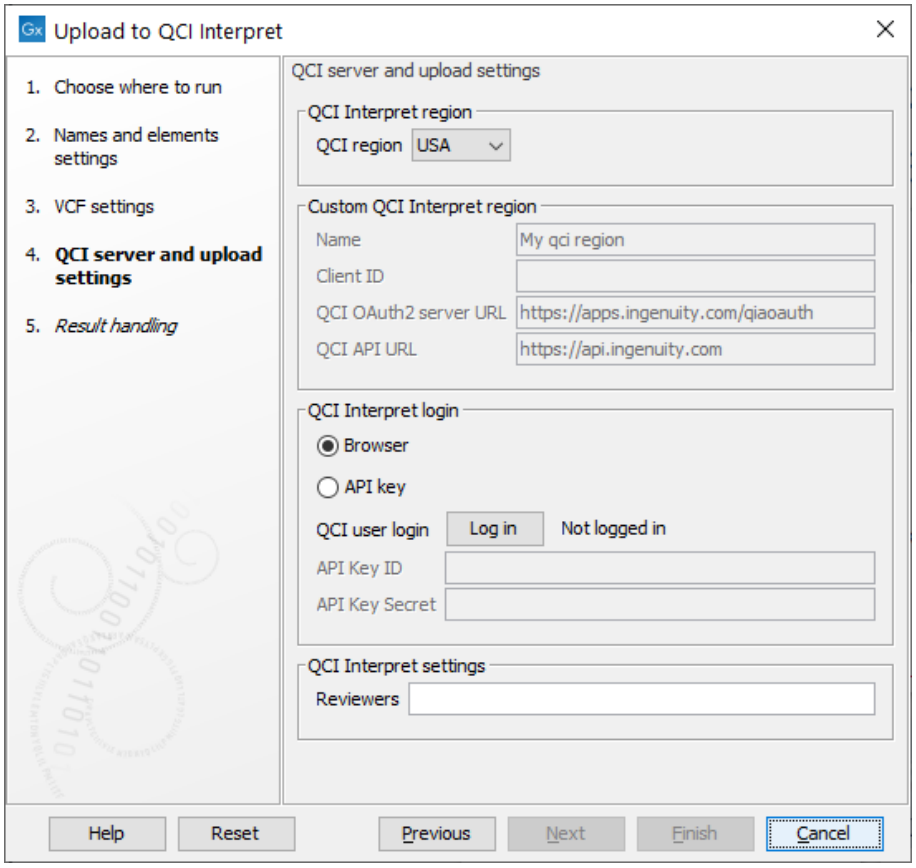


Figure 9.4: *Choosing QCI region and log in.*

Chapter 10

General tools

Contents

10.1 Filter Based on Name

Filter Based on Name takes as input an element that contains items with names, and outputs the subset of those items with names of interest. The names of interest can be provided in a separate element, or a text list.

To run the tool, go to:

> **Tools** | **Utility Tools** ( ) | **Tracks** ( ) | **Filter Tracks** ( ) | **Filter Based on Name** ( )

To see a list of element types that can be used as input, click on the ( ) icon in the top right corner of the input selection wizard (figure 10.1).

After selecting the input element, the names used for matching are defined (figure 10.2).

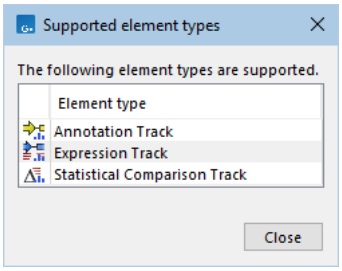


Figure 10.1: *Clicking on the info icon at the top right corner of the input selection wizard opens a window showing a list of element types that can be used as inputs.*

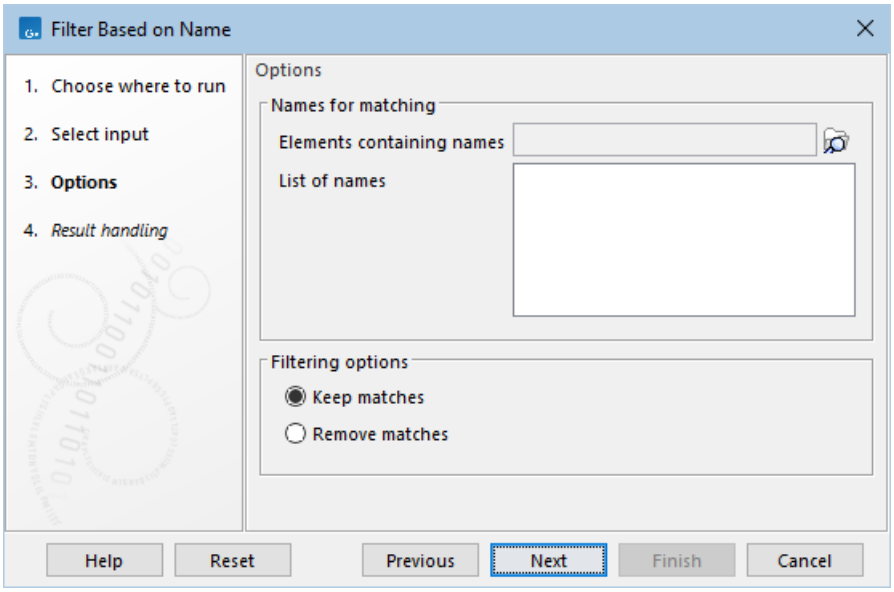


Figure 10.2: *Specify names for filtering.*

- **Elements containing names** The following can be chosen:
  - Elements containing named items. The same element types that can be used as input are supported.
  - Tables ( ) with names in the first column.
- **List of names** A list of names separated by space, newline, comma or semicolon. In cases where names contain one of the separators, we recommend importing the list of names as a table and using that table as the source of names. Tables can be imported using `Standard Import` from one of the supported `table formats`.

  All provided names, whether from elements containing names or the list of names, are used for filtering.

  For names provided using tables and list of names, the following wildcards are supported:
  - ? matches any single character.
  - * matches any sequence of characters, including none.

In the "Filtering" options section, the options **Keep matches** and **Remove matches** are available for specifying how matching items should be handled.

Special handling of certain input types

Some tools resolve duplicate names found in inputs or provided parameters by appending an underscore ("_") followed by a unique number. Examples of such tools include:

- `RNA-Seq Analysis`.
- `RNA-Seq Analysis for Long Reads`.

These modified names are then used by downstream tools, such as `differential expression tools`.

Filter Based on Name automatically accounts for this behavior in input types that may contain such de-duplicated names:

- Expression Tracks ( ).
- Statistical Comparison Tracks ( ).

For instance, if the list of names includes "ACE", any input items named "ACE_1" and "ACE_2" will be recognized as matching "ACE."

10.2 Annotate RNA Variants

The **Annotate RNA Variants** tool annotates variants likely to arise from alternative splicing rather than DNA changes. The annotations are useful for downstream filtering.

Three classes of annotations are added "known introns", "splice variants", and "closest exon". These are described briefly below.

Known introns annotation When RNA and DNA reads are sequenced together "index-hopping" can occur, which leads to small numbers of RNA reads being found in the DNA file. If these RNA reads splice across a short intron they can be interpreted as providing support for one or more deletions spanning the intron. An example of this is shown in figure 10.3. The Annotate RNA Variants tool slides each of these deletions independently to the left and right to see if they can be aligned with an exon boundary without changing the deleted sequence. If this is possible, the deletion is annotated with "Matches known intron = Yes".

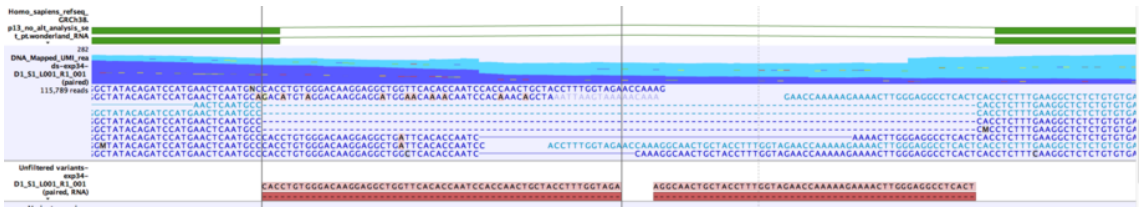


Figure 10.3: *Example of two variants that will be annotated as matching a known intron.*

Splice variants annotations Some variants likely to arise from alternative splicing rather than DNA changes. Although these are often due to tandem acceptor splice variants not present in the mRNA annotations, they can also be clinically relevant. A simple example unlikely to be of clinical relevance is shown in figure 10.4.

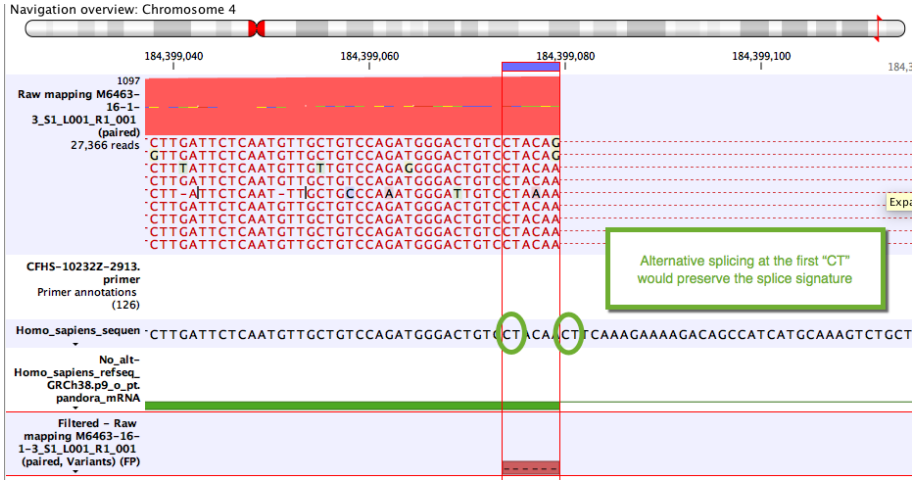


Figure 10.4: *Example of a variant that will be annotated as a splice variant conserving the splice signature.*

To detect such variants, the tool lists all variants within 30nt of each intron. All combinations of these variants are then generated such that for input variants v1, v2, v3, the output combinations would be: [v1], [v2], [v3], [v1,v2], [v1,v3], [v2,v3], [v1,v2,v3]. At most 31 combinations of variants are generated in this way as the number of possible combinations of variants quickly becomes unmanageable. Combinations of variants that are incompatible are not generated. For example if v1 and v2 are overlapping deletions, then combinations [v1,v2] or [v1,v2,v3] are not generated as these are not consistent with a single RNA molecule.

For each combination of variants, the DNA (extracted from the genome) is aligned to the variants + RNA (extracted from mRNA annotations). If the alignment can be made without mismatches, and with only a single gap, then the variants can be explained by a new intron with the coordinates of the gap. All such variants are then annotated with "Possible splice variant = Yes". Additional annotations are added based on the splice signature of the new intron:

- "Conserved splice signatures" with values "Yes" or "No" according to whether the splice signature matches that of the original intron.
- "Canonical splice signature" and zero or more comma-separated values from the list "GT-AG", "GC-AG", "AT-AC"
- "Possible splice signatures" and one or more comma-separated values from the list "AA-AA", "AA-AC" ... "TT-TT"

A variant may be investigated multiple times if it is close to multiple introns. If this happens, the annotations are updated such that "Conserved splice signature" may change from "No" to "Yes", and splice signatures may be added to the "Canonical splice signature" and "Possible splice signatures" lists.

Closest exon annotations A common error mode of RNA-Seq variant calling is when a variant lies within an intron close to an exon. In this case the frequency of the variant may be artificially inflated if reads without the variant are spliced and have no coverage in the intron.

An example is shown in figure 10.5. Here the coverage is 210 at the position where the SNV is called, but >2500 in the adjoining exon. The "real" variant frequency is closer to 16/2500 < 1% than 16/210 > 5%.

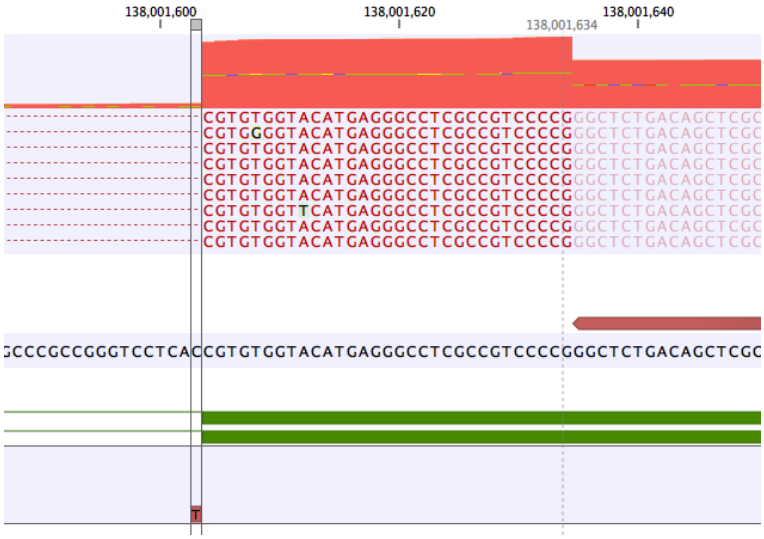


Figure 10.5: *Example of a variant that is called with frequency >5%, but whose frequency corrected to the coverage at the nearby exon is very low.*

The tool adds an annotation "Distance to nearest exon" to all non-reference variants within 100 nt of an exon. The nearest exon is that which returns the smallest positive distance among the following four options: (exonStart - variantStart, exonStart - variantEnd, variantStart-exonEnd, variantEnd-exonEnd).

If a read mapping is provided, the tool also annotates variants with the "Frequency compared to closest exon". This frequency is 100 * count / coverage at nearest exon. The coverage at the nearest exon is actually the coverage at the boundary of the exon. Coverage is calculated per match rather than per read i.e. broken pairs count as 2, pairs count as 1, and single reads count as 1. Coverage also includes ambiguously mapped reads.

In all cases where count > coverage at nearest exon, the tool reports "Frequency compared to closest exon = 100%" which in practice means that this annotation cannot be used to filter the variant. This can happen for a variety of reasons. For example:

- The nearest exon has no coverage - perhaps an exon with a boundary slightly further away from the variant has all the coverage
- Primers positioned in the intron generate more reads than primers in the adjacent exon

The **Annotate RNA Variants** tool is available under the Tools menu at:

Tools | **Resequencing Analysis** () | **Variant Annotation** () | **Annotate RNA Variants** ()

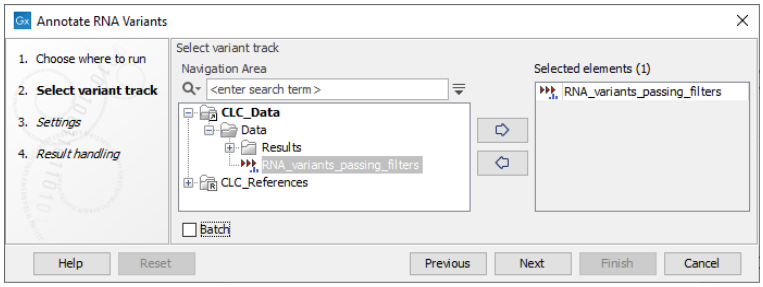


Figure 10.6: *Select a variant track.*

The tool takes a Variant Track as input (figure 10.6)

In the next dialog figure there are two parameters to set (figure 10.7)

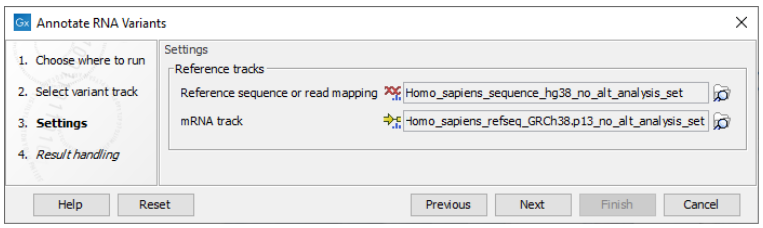


Figure 10.7: *Select a reference genome or read mapping, and an mRNA track.*

- Reference sequence or read mapping - variants will only be annotated with the coverage at the nearest exon if a read mapping is provided.
- mRNA track

The tool outputs an annotated Variant Track.

10.3 Detect Regional Ploidy

The **Detect Regional Ploidy** tool is designed to detect regional ploidy levels including loss-of-heterozygosity (LOH) from targeted research resequencing experiments.

The tool takes a target-level CNV events annotation track (from a CNV tool), somatic variants, and either germline variants or known segregating variants and optionally centromers.

Detect Regional Ploidy is available under the Tools menu at:

> **Tools** | **Resequencing Analysis** ( ) | **Variant Detection** ( ) | **Detect Regional Ploidy** ( )

Select the CNV target-level annotation track generated by Copy Number Variant Detection (Targeted), see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_Copy_Number_Variant_Detection_Targeted_tool.html`, and click **Next**.

You are now presented with choices regarding LOH detection.

- **Somatic variants** A track containing variants in the somatic sample. Their allele frequencies must be provided.

- **Type of variant track with known variants** Choose if the track with known variants is a variant database (Variant database) or a matching germline variant track (Germline variants). This will determine if LOH detection is performed in unpaired mode or in matched tumor normal mode.
- **Known variants** If "Variant database" is chosen above, provide a variant track of known SNPs in the population annotated with allele frequencies (e.g. dbSNP). The variant track can be restricted to the target region to improve computation time. This can be done using the **Filter Based on Overlap** tool, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_Based_on_Overlap.html. If "Germline variants" are chosen above, provide a variant track with matching germline variants. The variants are automatically filtered to heterozygous variants. For optimal performance, the variants should be high confidence.
- **Centromeres** If provided, targets overlapping the centromeric regions will be excluded.
- **Normalize coverage using allele frequencies** If enabled, allele frequencies will be used to find the correct coverage normalization. If a large fraction of targets are affected by say a deletion, the normalization factor used for the sample will be too low, resulting in underdetection. However, a deletion is both expected to affect the coverage and the allele frequencies and this information can be used to correct the normalization factor (see tables 10.2 and 10.3). As an example, if the control sample has copy number 2 for all targets, but the case sample has copy number 1 for all targets, the coverage after correcting for total library size should ideally be adjusted by a factor 0.5. Enabling this option is recommended for small panels where a large fraction of targets may be affected by CNV events.
- **Minimum normalization, Maximum normalization** If 'Normalize coverage using allele frequencies' is enabled this defines the limits to the amount of normalization done.
- **Minimum sample purity** The lowest sample purity the model can estimate. It is hard to distinguish a sample with only a few CNV and LOH events from a sample with very low purity. Set this parameter to the lowest purity that the model is allowed to use.
- **Transition factory** The transition factor controls the chance of switching state. A higher transition factor makes state switches less probable.
- **HMM decoding method** Method for optimizing and decoding Hidden Markov Model (VITERBI or POSTERIOR).
- **Minimum merge size** Remove merged regions consisting of fewer than this number of targets, when joining targets into regions.

Regional ploidy estimation

The algorithm implemented in the **Detect Regional Ploidy** tool is inspired by the following paper:

- **Beroukhim et al.** *Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays*, PLoS Computational Biology. 2006, 2(5): 323-332 [Beroukhim et al., 2006]

Based on coverage ratios and the allele ratios of putative heterozygous germline variants the tool detects targets and regions affected by Loss-of-heterozygosity events. The tool can handle both matched tumor normal data and unpaired tumor data. In both cases variants that are assumed to be heterozygous in normal tissue has to be identified.

Tumor-normal pairs: For matched tumor normal data, a track with somatic variants and a track with germline variants will be used. The variants used to detect LOH are simply the somatic variants overlapping heterozygous germline variants.

Tumor only: For unpaired tumor data, a somatic variant track and a database of known segregating variants are used (typically dbSNP common). The variants used in LOH calculation are the somatic variants overlapping the variants in the database.

The model operates with a number of ploidy states, which are characterized by their numbers of parental and maternal alleles (Table 10.1). The state together with the tumor purity (the percentage of cells in the sample originating from the tumor) determines the expected coverage ratio and the expected allele frequencies of the heterozygous variants. As an example, if a normal diploid sample would yield 200 reads, then a sample with purity 50% and copy-number 1 (deletion) would yield 150 reads (50%*200+50%*100). That means the coverage ratio is 150/200 = 75%. Table 10.2 shows the expected coverage ratios for different states and purities.

The state together with tumor purity also determines the expected allele frequencies of heterozygous variants. As an example, consider a sample with 60% purity where the cancer cells contain a deletion in a region with two alleles, A and B. If we take 100 cells:

- 60 cells (tumor) will contain one copy of allele A
- 40 cells (normal) will contain one copy of allele A and one copy of B

In total there will be 100 copies of allele A, and 40 copies of B. And the frequency of A will be 100 / (100 + 40) = 71.4%.

The tool estimates the purity using a hidden Markov model (HMM), that is then used to predict the most probable state for each target.

State	Allele-ratio	Copy-number	Loss-of-heterozygosity
Bi-allelic deletion	0:0	0	
Deletion	0:1	1	deletion LOH
Diploid	1:1	2	
Uniparental disomy	0:2	2	copy-neutral LOH
Duplication	1:2	3	
WGD	2:2	4	

Table 10.1: Ploidy states with their allele ratio, total copy number and whether the state is considered loss-of-heterozygosity.

Limitations

Detect Regional Ploidy is designed for ploidy estimation on autosomal chromosomes. The underlying model does not take into account that the normal state of sex chromosomes in male samples is haploid, and hence may mis-interpret detected allele frequencies and coverage

Purity	Bi-allelic deletion	Deletion	Diploid	Uniparental disomy	Duplication	WGD
10.0%	90.0%	95.0%	100.0%	100.0%	105.0%	110.0%
20.0%	80.0%	90.0%	100.0%	100.0%	110.0%	120.0%
30.0%	70.0%	85.0%	100.0%	100.0%	115.0%	130.0%
40.0%	60.0%	80.0%	100.0%	100.0%	120.0%	140.0%
50.0%	50.0%	75.0%	100.0%	100.0%	125.0%	150.0%
60.0%	40.0%	70.0%	100.0%	100.0%	130.0%	160.0%
70.0%	30.0%	65.0%	100.0%	100.0%	135.0%	170.0%
80.0%	20.0%	60.0%	100.0%	100.0%	140.0%	180.0%
90.0%	10.0%	55.0%	100.0%	100.0%	145.0%	190.0%
100.0%	0.0%	50.0%	100.0%	100.0%	150.0%	200.0%

Table 10.2: The expected coverage levels given tumor purity and the ploidy state.

Purity	Bi-allelic deletion	Deletion	Diploid	Uniparental disomy	Duplication	WGD
10.0%	50.0%	52.6%	50.0%	55.0%	52.4%	50.0%
20.0%	50.0%	55.6%	50.0%	60.0%	54.5%	50.0%
30.0%	50.0%	58.8%	50.0%	65.0%	56.5%	50.0%
40.0%	50.0%	62.5%	50.0%	70.0%	58.3%	50.0%
50.0%	50.0%	66.7%	50.0%	75.0%	60.0%	50.0%
60.0%	50.0%	71.4%	50.0%	80.0%	61.5%	50.0%
70.0%	50.0%	76.9%	50.0%	85.0%	63.0%	50.0%
80.0%	50.0%	83.3%	50.0%	90.0%	64.3%	50.0%
90.0%	50.0%	90.9%	50.0%	95.0%	65.5%	50.0%
100.0%		100.0%	50.0%	100.0%	66.7%	50.0%

Table 10.3: The expected frequencies of variants that are heterozygous in the normal tissue given tumor purity and the ploidy state.

ratios. If the tool is used to estimate ploidy for sex chromosomes, the results should be carefully assessed.

10.3.1 Output from Detect Regional Ploidy

The tool produces the following outputs:

- **Target-level Ploidy Track** A target region track where estimated ploidy is annotated to each of the original target regions in the target regions track provided to Detect Regional Ploidy. In addition, ploidy results are also provided for database/germline variants located in target regions.
- **Region-level Ploidy Track** A region-level track where adjacent target regions with the same estimated ploidies are collapsed to longer regions.
- **Report** A report providing the estimated purity, the normalization factor, an overview of identified target states as well as plots for log scaled fold changes and identified allele frequencies.

Target-level Ploidy Track

The target-level ploidy track contains the target regions and variants annotated with ploidy information. For target regions, original annotations from the tool Copy Number Variant Detection (Targeted) are also retained. The following columns have been modified or added by **Detect Regional Ploidy**:

- **Name** lists the type of information provided in the respective row:
  - **Coverage ratio** rows contain target level coverage information from the tool Copy Number Variant Detection (Targeted) and ploidy results from **Detect Regional Ploidy**.
  - **Variant Allele frequency** rows contain ploidy results as well as count and coverage.
  - **Reference allele** rows contain ploidy results as well as the population frequency if provided in the variant database track.
- **Adjusted RLR** is the relative log ratio adjusted by the normalization factor.
- **LOH** provides information about whether the estimated state can be considered LOH or not (Yes or No). LOH is characterized by loss of one chromosome, whereas the other chromosome is present in one or more copies.
- **State** is the predicted ploidy state.
- For columns **Normal diploid** to **(2,4)**, each column represents a target state. For example, Uniparental disomy represents the state where two identical chromosomes are present, and (1,3) represents the state where one chromosome is present in one copy and the other is present in three copies. For a given target region or variant, the predicted state is denoted with a 1 in the relevant column.
- **Count** is the variant count from the provided somatic variant track.
- **Coverage** is the variant coverage from the provided somatic variant track.
- **Population frequency** is the population frequency from the variant database. If germline variants are provided instead of a variant database this column is not included.

Region-level Ploidy Track

The region-level ploidy track contains predicted region level ploidy states where target regions with the same ploidy state have been collapsed to one region. In addition to chromosome and region, the track contains the following information:

- **Name** contains the predicted ploidy state of the region.
- **LOH** provides information about whether the estimated state can be considered LOH or not (Yes or No). LOH is characterized by loss of one chromosome, whereas the other chromosome is present in one or more copies.
- **Number of targets** is the number of original target regions included in the collapsed region.

Detect Regional Ploidy algorithm report

The purity and normalization factor table shows the estimated purity and normalization factor along with confidence intervals. Low purity or a wide confidence interval for purity is an indication that the regional ploidy predictions are uncertain. In the target regions table the number of targets predicted to be in each ploidy state is shown.

The log coverage ratio plots give a genome-wide overview over detected ploidy states and variant frequencies. There are plots for all chromosomes (Figure 10.8), and individual chromosomes with at least one target. The x-axis is the genomic position, the y-axis is the fold change. Vertical lines separate chromosomes, and the striped line shows average coverage. There are points for each input target with colors and shapes based on their calculated ploidy states. Plots come in pairs. The second plot shows variant frequencies. This plot is similar to the above one, but the y-axis is variant allele frequencies. There are points for each input somatic variant with colors and shapes based on their calculated ploidy states.

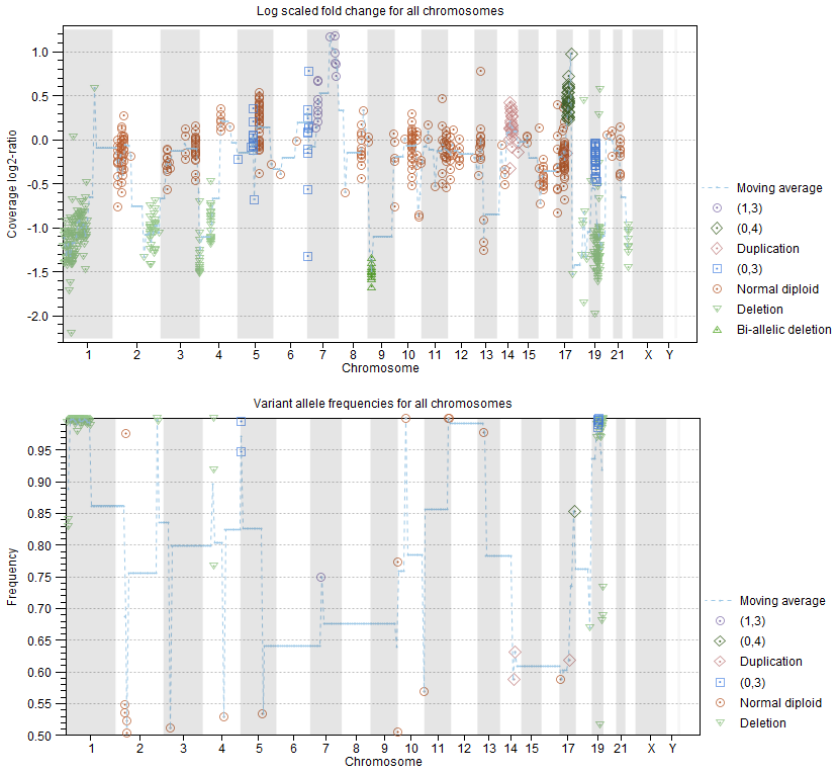


Figure 10.8: *Log-coverage ratios for CNV targets and variant allele frequencies for somatic variants.*

The next two subsections provide information useful for diagnosing potential problems with LOH detection. First, the expected coverage log-ratios for each ploidy state are shown along with the average coverage log-ratios for targets predicted to have this state. The expected coverage log-ratios are simply computed as in table 10.2 based on the estimated purity. Below the table is a plot with coverage log-ratios plotted against the base coverage. The points are colored by their predicted state and horizontal lines indicate the expected log-coverage ratio for each state (Figure 10.9).

Second, the expected allele frequencies for each ploidy state are shown along with the average allele frequency for variants predicted to have this state. Again the expected allele frequencies are computed as in table 10.3 based on the estimated purity. Below the table is a plot with allele

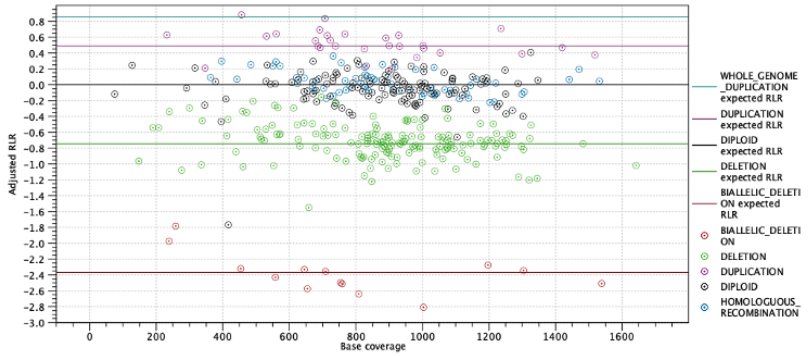


Figure 10.9: *Log-coverage ratios for each target with horizontal lines indicating the expected log-coverage ratio.*

frequencies plotted against their coverage. The points are colored by their predicted state and horizontal lines indicate the expected allele frequency for each state (Figure 10.10).

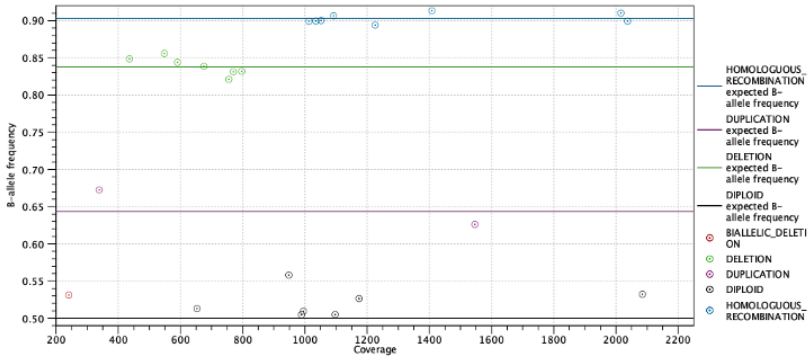


Figure 10.10: *Allele frequencies for each putative heterozygous variant with horizontal lines indicating the expected allele frequencies.*

10.4 Import Gene-Pseudogene Table

Pseudogenes originate from gene duplications that have taken place through evolution. The homology between gene-pseudogene pairs can be as high as 100% and this can be a challenge when analyzing sequencing data. In some cases, reads exist that map equally well to the gene and it's pseudogene and this can be taken into consideration when analyzing the sequencing data. More specifically, a built-in functionality in the tool **Trim Primers of Mapped Reads** allows removal of reads that originate from a pseudogene but have been mapped to the target gene, if a gene-pseudogene track has been provided as input to the primer trim tool. Gene-pseudogene tracks are provided for some of the QIAseq gene panels. If you would like to provide your own list of gene-pseudogene pairs, this can be imported using the **Import Gene-Pseudogene Table** tool, which can be found in the **Import** menu.

This importer takes as input a TXT or TSV file containing gene and pseudogene information. The input file has the following format: The first column contains a gene name, and the second column is a comma separated list of pseudogene names, so that each line contains information about one link between a gene and one or several pseudogenes.

```
LGALS9    LGALS9B,LGALS9DP,LGALS9C
LGMN    LGMNP1
```

When importing a gene-pseudogene table, a gene track must also be selected. Names of genes in the gene track will be used for matching names in the input file. If a name of a gene or a pseudogene from the input is not found in the gene track that was used during import, you will get a warning that will list the names of the genes that could not be matched and thus are absent from the imported track.

The output is a track of linked gene-pseudogene features that can be saved in the Navigation area. The gene-pseudogene track can be used as input to **Trim Primers of Mapped Reads**, but can only be used with the same reference genome as was used when importing the gene-pseudogene table.

10.5 Prepare Guidance Variant Track

The **Prepare Guidance Variant Track** tool generates a guidance variant track from the two outputs of the **InDels and Structural Variants** tool: a Structural variants feature track and an InDels variant track. This guidance track can be used for better realignment.

Prepare Guidance Variant Track is available under the Tools menu at:

Tools | **Resequencing Analysis** ( ) | **Prepare Guidance Variant Track** ( )

In the first dialog (figure 10.11), select an Indels variant track (it usually has the name of the read mapping it originates from, with (Indel) at the end of the full name).

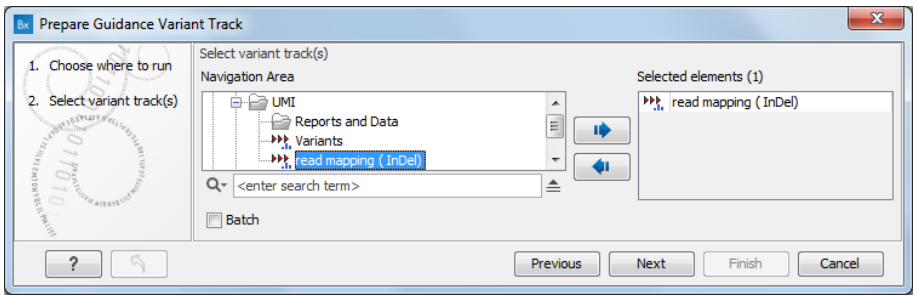


Figure 10.11: *Select a variant track.*

In the second dialog, select the Structural variant track (SV) that was output with the Indels variant (InDel) track from the same read mapping by the InDels and Structural Variants tool (as seen on figure 10.12). You can optionally specify a reference sequence. If set, it will be used to left-align indels to the extent possible. This will make a subsequent local alignment more in line with the recommendation/convention to left-align and that might in turn affect the downstream variant detection to also be left-aligned.

The output is a guidance track (Indel, guidance track), that combines valuable information from the indels and structural variants tracks (such as all the replacements that the tool detects), and that can be used as input for the Local Realignment tool. You can read more about the Indel and Structural variant tracks here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_Structural_Variants_InDels_output.

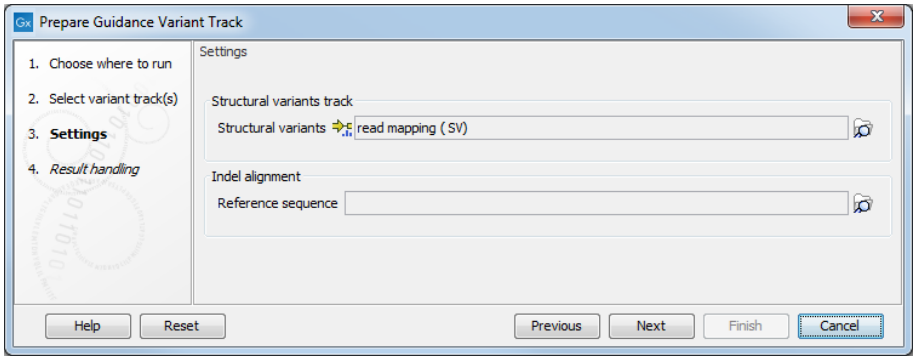


Figure 10.12: *Select a Structural variant track.*

10.6 Refine Read Mapping

Different sequencing technologies have different advantages and disadvantages and this is also the case when it comes to accuracy. The tool **Refine Read Mapping** can remove problematic mapped reads with many mismatches in close proximity and mapped reads with an unaligned end of a certain length. Removal of these types of mapped reads can minimize the number of potential false positive variants being reported. **Note:** The tool considers single end mapped reads. When analyzing paired end mapped reads the tool considers each mapped read in the pair separately.

Refine Read Mapping is available under the Tools menu at:

Tools | **Resequencing Analysis** ( ) | **Refine Read Mapping** ( )

In the first dialog (figure 10.13), select a read mapping.

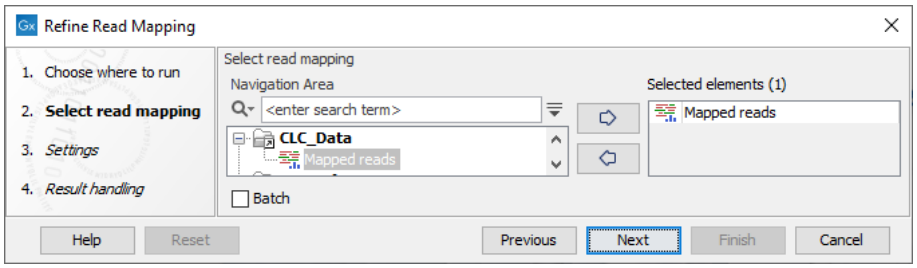


Figure 10.13: *Select a read mapping.*

In the second dialog (figure 10.14) the following parameters can be specified:

- Variants
  - **Remove mapped reads with variants**: When enabled, the Window size and Maximum variants can be specified.
  - **Window size**: The length of a region to be considered when counting the number of allowed variants specified under Maximum variants.
  - **Maximum number of variants allowed**: A mapped read with this number of variants found within a stretch of nucleotides of the length specified under Window size will be removed. Insertions, deletions, SNVs and MNVs each count as one variant. As the tool works on single end mapped reads, both mapped reads in a pair will be removed

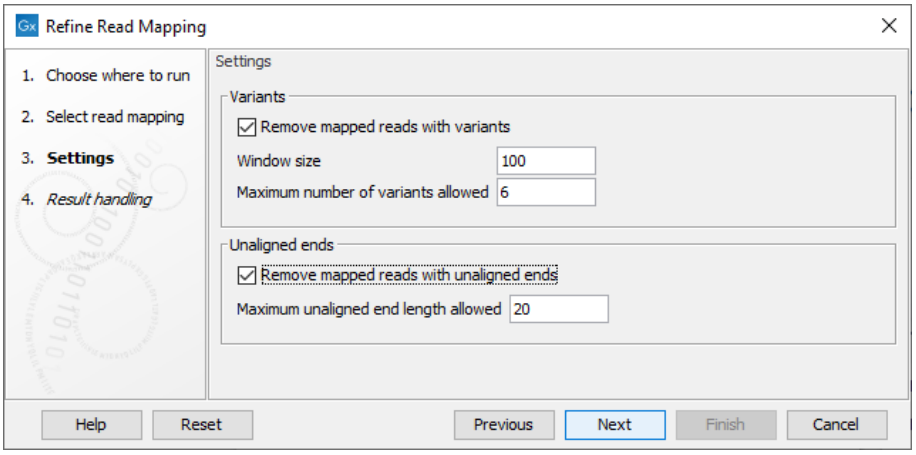


Figure 10.14: *Specify the settings for when to remove mapped reads with variants and when to remove mapped reads with unaligned ends.*

if this number of variants are found in any of the two single end mapped reads making up the pair.

- Unaligned ends
  - **Remove mapped reads with unaligned ends**: When enabled, the Maximum unaligned end length can be specified.
  - **Maximum unaligned end length allowed**: The maximum allowed number of nucleotides that are unaligned in a mapped read. If the number of unaligned nucleotides in a mapped read exceeds the specified Maximum unaligned end length, the mapped read is removed. A paired mapped read is removed if any of the two single end mapped reads making up the pair has an unaligned end longer than the specified maximum number of unaligned bases.

An example of a read mapping before and after using the **Refine Read Mapping** is shown in figure 10.15. In this example the mapped reads contain sequencing artifacts that can result in potential false positive calls.

The data used in this example are from a patient sample that were sequenced twice using two different sequencing technologies. Only one sequencing technology had problems with the G pattern shown in figure 10.15.

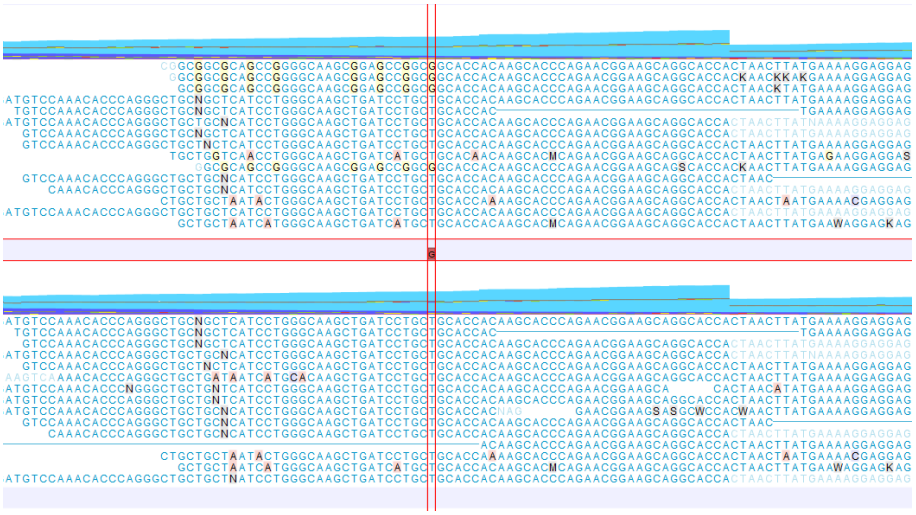


Figure 10.15: *A read mapping before (top) and after (bottom) using the tool Refine Read Mapping on the read mapping using default settings for Variants of Window size = 100 and Maximum variants = 6. A false positive variant was reported for one of the G positions before using the tool Refine Read Mapping. The variant is not reported after the mapped reads containing the G artifacts have been removed.*

10.7 Structural Variant Caller

The **Structural Variant Caller** tool identifies structural variants in read mappings based on evidence from unaligned read ends and coverage information. It builds on the same ideas around unaligned end read signatures as the existing **InDels and Structural Variants** tool, but to a larger extent relies on statistical reasoning and more refined components for consensus generation, mapping and alignment of the unaligned end sequences.

The tool:

- detects deletions, insertions (including tandem duplications), and inversions.
- is applicable to read mappings of Targeted, Exome, and WGS (Whole Genome Sequencing) NGS resequencing data.
- is developed for short read technologies (such as Illumina reads).
- detects germline as well as somatic variants.

The tool has the following limitations:

- Inter-chromosomal rearrangements are not supported.
- In read mappings of RNA-Seq data each part of a spliced read is treated independently.
- It can only process reads that are shorter than 5000 bp, reads that are longer are discarded.

The tool processes each chromosome in a genome individually, through several steps:

Breakpoint estimation: The tool looks for unaligned read ends at each chromosome position. Consensus sequences are constructed for the unaligned ends and aligned regions across the reads at a breakpoint (one consensus sequence for the unaligned end and one for the aligned region). The consensus sequence is based on a majority count of k-mers for the unaligned end, while the nucleotide count in each column is used for the aligned region. Breakpoints are labeled either as a 'left' or 'right' breakpoint. This labeling is from the perspective of a deletion, where a left breakpoint is on the left side of a deletion (which means there is a right unaligned end) and a right breakpoint is vice-versa on the right side of the deletion. For WGS applications, the tool makes a probabilistic assessment of how likely the breakpoint is to support a structural variant based on the coverage, the unaligned end read count, and the specified ploidy of the sample.

Coverage and complexity estimation: each chromosome is divided into bins. The tool then calculates the coverage and the complexity of the reference region in each bin.

- **Coverage** A bin size of 100bp is used when calculating the coverage, and uniquely mapped reads are then counted according to how much they cover a given bin (any non-specifically mapped reads are ignored). If for example a read covers half of the bin, then it will contribute with a value of 0.5 to the coverage. A structural variant's coverage is then calculated as the maximum coverage across the bins that it covers.
- **Complexity** The complexity is calculated using the Lempel-Ziv complexity measure and is used to avoid calling structural variants in low-complexity regions. A lower resolution is employed for this in comparison to the coverage, and the complexity therefore uses a bin size of 200bp.

Resolving structural variants: after breakpoints have been established, different combinations of left and right breakpoints are paired together. For each pair, the unaligned and aligned consensus sequences from one breakpoint are aligned to the other breakpoint. The alignment scores from each possible pairing are then stored in a matrix, and a dynamic programming algorithm is used to identify which breakpoints to pair together. Breakpoints that were not matched in this step are then each used as a single breakpoint to search for additional smaller insertions or deletions inferred from self-mapping evidence (where the unaligned consensus itself maps back to nearby its own location).

Running the Structural Variant Caller tool

To run the **Structural Variant Caller** tool, go to:

> **Tools** | **Resequencing Analysis** ( ) | **Variant Detection** ( ) | **Structural Variant Caller ( )**

Once the tool wizard has opened (figure 10.16), choose the read mapping you would like to analyze. The Structural Variant Caller tool accepts read mappings as either reads tracks or stand-alone read mappings.

In the next wizard step, specify the ploidy and application for the sample you are analyzing (figure 10.17). You can also specify to ignore broken pair reads. Ignoring broken pairs will typically reduce the computational time of the analysis. It may have a negative impact on sensitivity, but may also improve precision, depending on the source of the broken pair reads.

- **Ploidy** Specifies the ploidy of the sample. The value determines the maximum number of

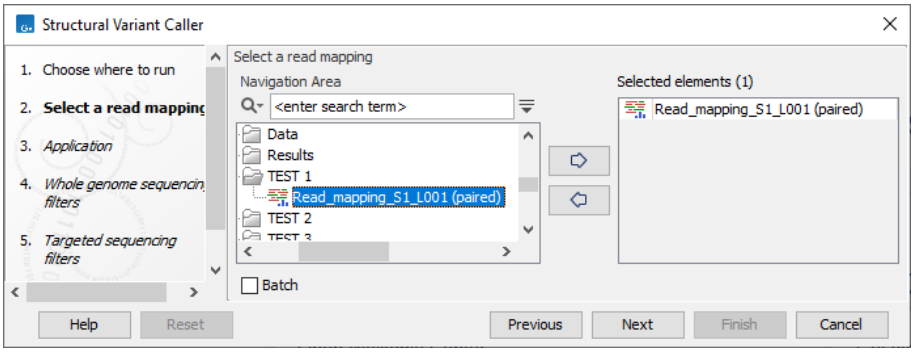


Figure 10.16: *Select one or several reads tracks or stand-alone read mappings.*

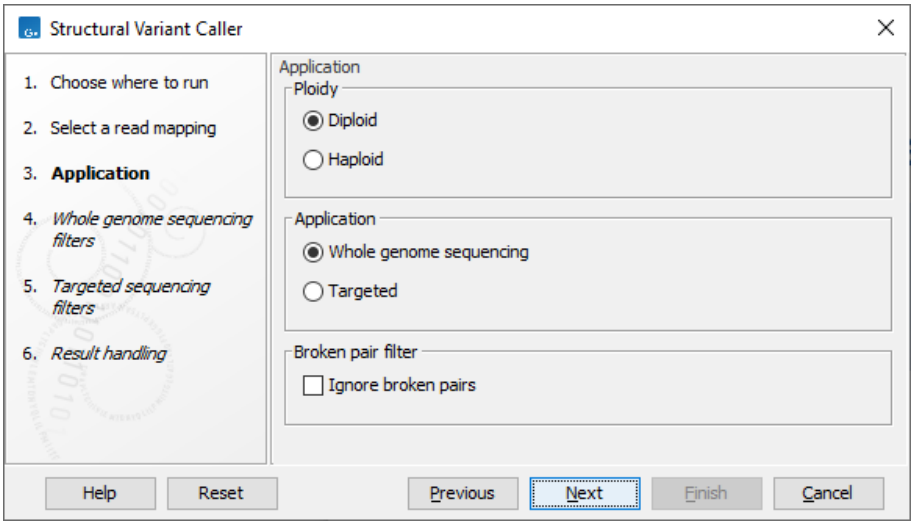


Figure 10.17: *Set the application parameters for the tool and specify if broken pair reads should be ignored.*

overlapping structural variants that can be detected, and, when Whole Genome Sequencing is specified as application, is also used for calculating breakpoint probabilities. Diploid should be chosen unless the data is from a haploid organism.

- **Application** Choose "Targeted" if running on a read mapping of targeted or whole exome sequencing data and otherwise choose "Whole Genome Sequencing". When Targeted is chosen the coverage and complexity analysis is not applied (it relies on model assumptions that are only appropriate for WGS applications) .

In the next steps you are asked to specify filter settings. The settings depend on whether you have specified the whole genome sequencing or the targeted application. The filter settings for the whole genome sequencing application (figure 10.18) are:

- **Minimum number of supporting reads** Minimum number of reads with unaligned ends required for a breakpoint to be detected. All of the detected breakpoints are used when searching for structural variants based on a pair of breakpoints.
- **Minimum breakpoint probability** Minimum required probability of a breakpoint to be considered (based on a statistical model and only applicable to Whole Genome Sequencing data).

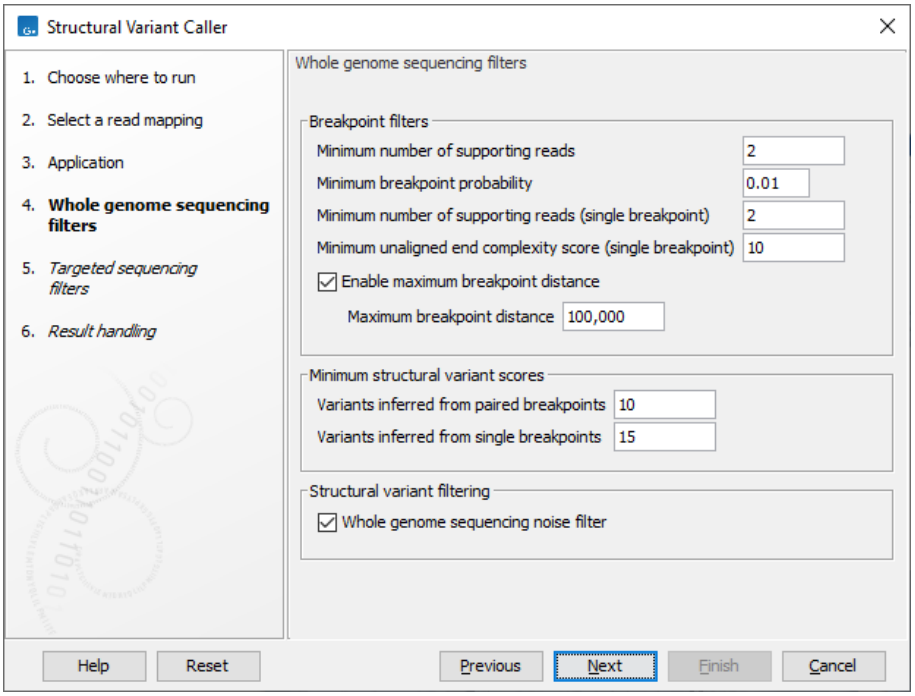


Figure 10.18: *Set filters for the whole genome sequencing applications.*

- **Minimum number of supporting reads (single breakpoint)** When searching for structural variants based on a single breakpoint, this is the minimum number of reads with unaligned ends required for a breakpoint to be considered.
- **Minimum unaligned end complexity score (single breakpoint)** When searching for structural variants based on a single breakpoint, a breakpoint must have an unaligned consensus sequence with this minimum complexity score to be considered.
- **Maximum breakpoint distance** If enabled, structural variants cannot be detected when a pair of breakpoints are further apart than this value. As most of the detected structural variants are found using breakpoint pairs, the maximum length of detected deletions, tandem repeats, and inversions will therefore typically be limited by this distance (note that re-alignment of a detected structural may occur, in which case the variant can extend beyond the breakpoint positions). Higher breakpoint distances will increase processing time, but will allow for the detection of longer deletions, tandem repeats, and inversions.
- **Variants inferred from paired breakpoints** Minimum score for structural variants based on a pair of breakpoints.
- **Variants inferred from single breakpoints** Minimum score for structural variants based on a single breakpoint. Structural variants that are based on a single breakpoint have less supporting evidence than variants based on a pair of breakpoints. It is therefore recommended to set this minimum score higher than the minimum for structural variants inferred from a pair of breakpoints.
- **Whole genome noise sequencing filter** This applies two steps of filtering. In the first step, breakpoints that appear unlikely are filtered out. This filtering is performed on the basis of breakpoint attributes such as unaligned end sequence complexity and local region coverage. In the next step, potential structural variants are filtered using a neural network

model that has been trained on the basis of whole genome sequencing data. The neural network filtering is applied to all structural variants except for insertions that are based on single breakpoints, as these are typically observed less frequently than other structural variant types.

For the targeted application the filters are (figure 10.19):

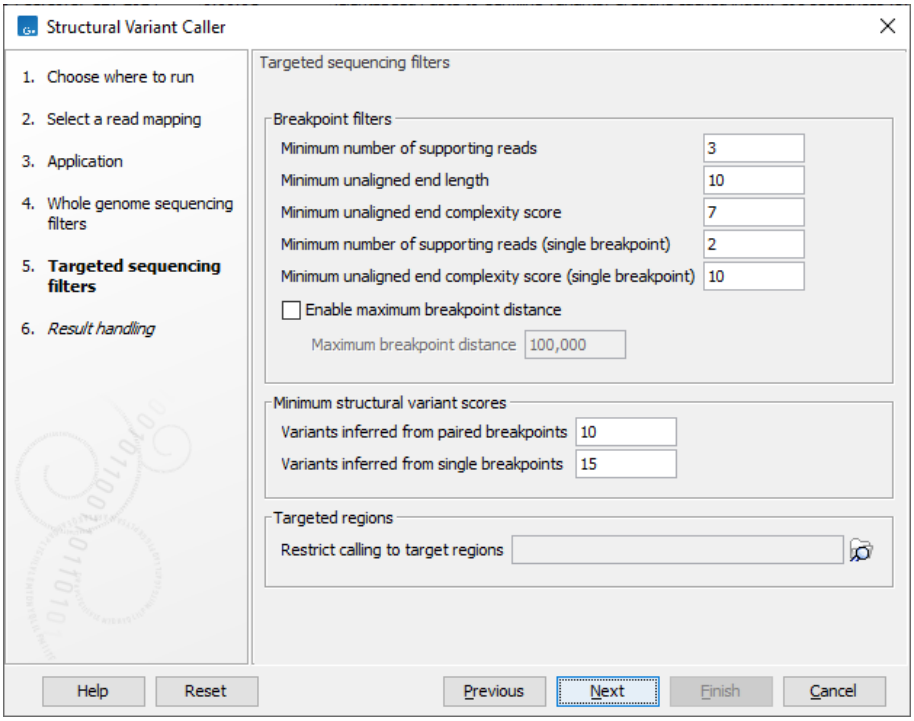


Figure 10.19: *Set filters for the targeted sequencing applications.*

- **Targeted regions** Allows you to specify an annotation track, to which the analysis will be restricted. When this is done, only breakpoints located within the specified regions (with a buffer of 15bp) will be considered and structural variant calls will be limited to those that are inferred from these breakpoints. This will typically decrease computational time, but may also cause variants with evidence located outside the specified regions to go undetected.

- **Minimum number of supporting reads** Minimum number of reads with unaligned ends required for a breakpoint to be considered.

- **Maximum breakpoint distance** If enabled, structural variants cannot be detected when a pair of breakpoints are further apart than this value. As most of the detected structural variants are found using breakpoint pairs, the maximum length of detected deletions, tandem repeats, and inversions will therefore typically be limited by this distance (note that re-alignment of a detected structural may occur, in which case the variant can extend beyond the breakpoint positions). Higher breakpoint distances will increase processing time and may make it more difficult to find smaller variants, but will allow for longer deletions, tandem repeats, and inversions to be detected.

- **Minimum unaligned end length** In the case of targeted data, this is the minimum length of the unaligned end required for a breakpoint to be detected.

- **Minimum unaligned end complexity score** In the case of targeted data, this is the minimum complexity of the unaligned end required for a breakpoint to be detected. The complexity is based on the Lempel-Ziv algorithm where each unique element in a sequence increases the complexity score by one. For example, if we process the sequence 'ACGGATTC' from left to right, then it has unique elements A, C, G, GA, T, and TC, resulting in a score of 6. Sequences are processed from left to right unless the resulting score is too low, in which case the sequence complexity from right to left is also calculated as this can yield a slightly different score.
- **Minimum structural variation score** Measure of the overall evidence supporting the structural variant detected. The value is based on the alignment scores of the unaligned ends or, in case of shorter indels, the length of the variation. This value may be increased to reduce the number of structural variants called.

10.7.1 Output from the Structural Variant Caller

The tool produces the following outputs:

- **Indels (Indels)** A variant track with indels (deletions and insertions (including tandem duplications)) that have lengths up to 100,000 bp
- **Long indels (Indels long))** An annotation track with long indels (those with lengths larger than 100,000 bp)
- **Inversions (Inv)** An annotation track with inversions
- **Breakpoints (BP)** An annotation track with a row for each breakpoint showing the unaligned ends used for the analysis
- **Report** A report giving an overview over analyzed references and found structural variants

The indels variant track, and the inversions and long indels annotation tracks can be exported to VCF.

The reason for putting the indels larger than 100,000 bp in a separate annotation track, is that the very long variants have very long either allele or reference entries in the variant track are challenging to work with in the track viewer.

Indels variant track The indels track contains all the standard variant annotations, except for the "Probability" and "QUAL" columns which are only preduced when the Whole genome Sequencing application is chosen. When produced, the content of the "Probability" column is the average of the probabilities of the breakpoints used to infer the feature, and the content of the "QUAL" column is the Phred score version of that probability.

As the indels are inferred indirectly from the unaligned ends, and hence are not necessarily directly visible within the aligned parts of the reads, the indel variant annotations are approximated from the breakpoints and the unaligned ends of reads in the read mapping. Figure 10.20 shows a read mapping of a 52bp deletion and the read mapping in which it was inferred by examination of the indirect evidence in the reads with unaligned ends, along with the approximated variant annotations (e.g count, coverage and frequency).

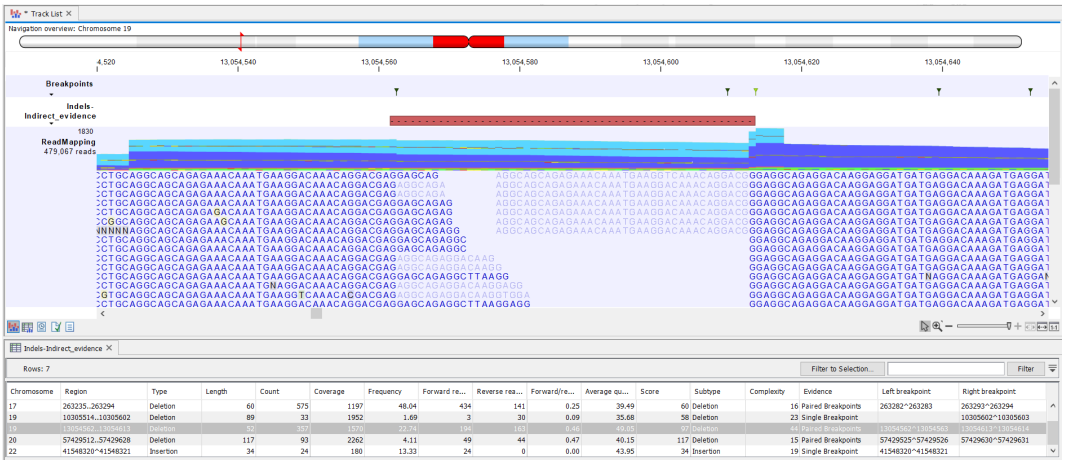


Figure 10.20: *A 52 bp deletion with approximated variant annotations and the read mapping in which it was inferred.*

In addition to the standard variant annotations, the indel track contains the following columns with characteristics of the inferred structural variant (figure 10.21):

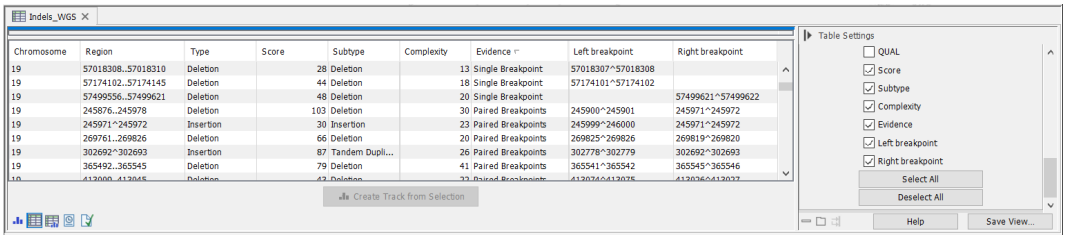


Figure 10.21: *The Structural Variant Caller indels track.*

- **Probability** Probability that a structural variant is correct. This value is estimated using the breakpoint probabilities and alignment scores that are associated with a structural variant. It is only available when the application has been set to whole genome sequencing.

- **Probability (NN)** Probability that a structural variant is correct. This value comes from the neural network model that is used in the whole genome sequence filtering, but it is also calculated when the filtering is not selected. It is only available when the application has been set to whole genome sequencing, and is not provided for insertions based on single breakpoints (see also whole genome noise sequencing filtering description). It should also be noted that there is a difference between the neural network model used for different structural variant types and their probabilities are therefore not directly comparable. These structural variant types are: short deletions, long deletions, short tandem duplications, long tandem duplications, short insertions, de novo based insertions, and deletions based on single breakpoints. A short structural variant in this case is defined as one that can be found within a read (for example this might be the case if a variant is less than 150bp in length), and anything else is considered to be a long variant.

- **Score** Measure of the overall evidence supporting the structural variant detected. The value is based on the alignment scores of the unaligned ends or, in case of shorter indels, the length of the variation.

- **Subtype** This is a more specific categorization of the structural variant type: either Insertion, Deletion, Tandem Duplication, Inversion, CNV Loss, or CNV Gain. Note that for Tandem

Duplications only one duplication is reported, even in cases where a sequence appears in more than two copies in the reads. The CNV loss and CNV gain annotations are CNVs inferred from a combination of coverage and breakpoint evidence, and are only inferred for Whole genome Sequencing applications.

- **Evidence** May be either Single Breakpoint, Paired Breakpoints, CNV + Breakpoint (i.e., based on coverage information and a single breakpoint), or Broken pairs. The broken pairs option is special since it is based on assembly of broken read pairs, where one of the reads in a pair maps at a different location in the genome. This allows for detection of insertions of Alu elements for example.
- **Complexity** Sum of the complexity of the left and right unaligned ends.
- **Left breakpoint** Position of the left breakpoint of the structural variant.
- **Right breakpoint** Position of the right breakpoint of the structural variant.

Long indels and inversions annotation tracks The long indels and inversions feature tracks contain the same columns as the indels track, except that the "Type", "Reference", "Allele" and "Reference allele" columns in the indels track are replaced by a single "Name" column in the feature track. The "Name" column specifies whether the feature is a deletion, insertion or inversion.

The report The report (figure 10.22) gives an overview of the numbers and types of structural variants found in the sample.

It contains:

- **A 'Summary' table** giving an overview of the numbers of breakpoints identified, and numbers of the different types and subtypes of the structural variants found
- **A 'References' table** with a row for each reference sequence, and information on the number of left and right unaligned breakpoint signatures and the resulting number of structural variants found on that reference sequence.
- **A 'Variants' table** with a row for each reference sequence, and information on the total number of variants, stratified into the different variant categories (Insertion, Deletion, Tandem Duplication, Inversion, CNV Loss, CNV Gain) found on that reference sequence.
- **A length distribution plot for short (<50 bp) structural variants**
- **A length distribution plot for long (>50 bp) structural variants**

Breakpoint track (BP) The breakpoint track (figure 10.23) contains a row for each called breakpoint with the following information

- **Chromosome** Chromosome on which the breakpoint is located.
- **Region** Location on the chromosome of the breakpoint.
- **Name** Type of the breakpoint ('left breakpoint' or 'right breakpoint').

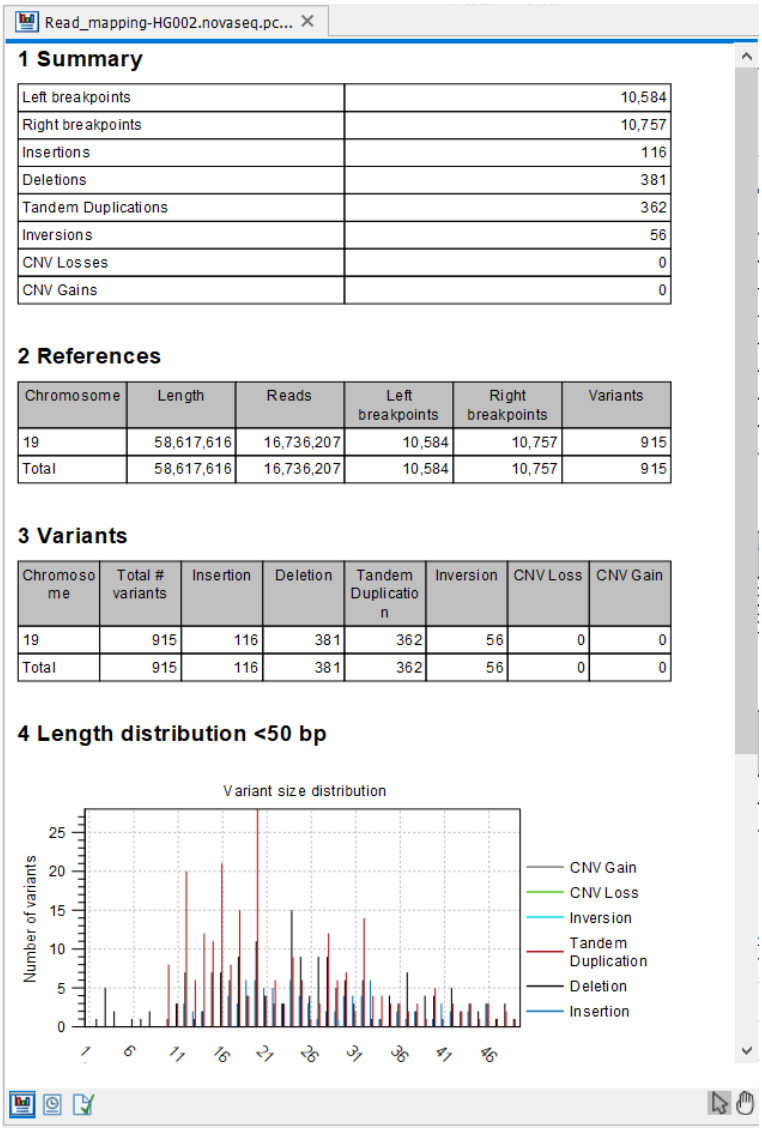


Figure 10.22: *The Structural Variant Caller report.*

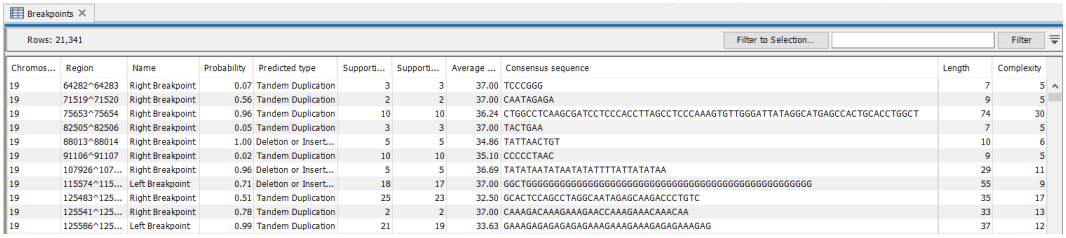


Figure 10.23: *The Structural Variant Caller breakpoints track.*

- **Probability** Estimate for how trustworthy the prediction is when running on WGS data
- **Predicted type** Whether the breakpoint appears to be part of a tandem duplication, a deletion or insertion. This is an initial estimate that does not include the possibility of inversions and is used with WGS data.
- **Supporting reads** Number of reads at the breakpoint position with an unaligned end.
- **Supporting reads (weighted)** Number of reads at the breakpoint position with an unaligned

end, but weighted according to an alignment probability that is assigned to each read.

- **Average quality** Average read quality of a single position in the supporting unaligned ends at a breakpoint. Only reads that have unaligned ends with the same direction as the breakpoint (i.e left or right) are included in the average.
- **Consensus sequence** The consensus sequence calculated across the unaligned ends of the reads that support the breakpoint.
- **Length** Length of the consensus sequence.
- **Complexity** Complexity of the consensus sequence.

10.8 Targeted Methyl associated tools

10.8.1 Finding differentially methylated regions

The Targeted Methyl application of the QIAseq Panel Analysis Assistant, see chapter 11 offers a "Detect Differentially Methylated Regions" tool to be run after the Detect QIAseq Methylation workflow. The tool requires at least two case and two control samples.

Click **Run** to open the tool's wizard. In the first dialog, select the case read mappings, and click **Next**.

In the following dialog you can optionally choose to provide target regions. If these are supplied, then each target is separately tested for differential methylation. It makes sense to provide targets if they are the unit of biological interest, or if they are short enough that methylation patterns within a target are expected to be constant i.e. most Cs are hypomethylated/hypermethylated/unchanged when compared to a control sample. If no target regions are provided, the tool will search for differential methylation by dividing the genome into 1kb regions.

In the final dialog, select the control read mappings. The output of the tool is a track of differentially methylated regions for CpG sites.

The above-described tool works by calling the Call Methylation Levels tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Call_Methylation_Levels.html) with parameters optimized for QIAseq Targeted Methyl data. To run the Call Methylation Levels tool directly with these optimized parameters:

- Uncheck Ignore duplicate matches: Duplicate matches have already been identified in the Detect QIAseq Methylation workflow using UMIs.
- Uncheck Confirm methylation contexts in reads: this option, when enabled, discards reads where the context is different in the reads and reference, for example due to a SNV. When this is disabled, SNVs that change the methylation context (for example from CpG to CHG) cannot be distinguished from changes in methylation. However this is unlikely to lead to incorrect interpretation of the data. The presence of a SNV can be confirmed by inspection in the read mapping.
- Set the Statistic mode to ANOVA

- Set the Minimum high-confidence site coverage to 10: Coverage should be high for most targets, so it may be acceptable to exclude low coverage sites which will typically be far from the primers, and which may add more noise than signal. However, first check that this is higher than the typical coverage of positions in each target using the Final_target_coverage output.
- Uncheck Create track of methylated cytosines: this would generate the same per-sample data already produced in the workflow.

10.8.2 Create Methylation Level Heat Map

The **Create Methylation Level Heat Map** tool generates a two dimensional heat map of methylation levels. Each column corresponds to a sample, and each row corresponds to a feature (a single CpG site or a larger target region including multiple CpG sites, e.g. promoter regions). A hierarchical clustering of the samples is performed. For up to 5000 features, a hierarchical clustering of features is also performed.

Calculation of the methylation levels is performed across all CpG sites in a given target. When the coverage of a CpG site is lower than a specified threshold, that site will be considered zero methylated, indicating that it is uninformative. For targets containing multiple CpG sites, only informative sites are considered and the methylation level is averaged across all the informative sites. For targets containing only a single CpG site, the methylation level is considered only for that site.

Clustering of features and samples

Features are clustered according to the similarity of their methylation level profiles over the set of samples. Samples are clustered according to the similarity of their methylation level patterns over the set of features.

The clustering has a tree structure that is generated by:

1. Letting each feature or sample be a cluster.
2. Calculating pairwise distances between all clusters.
3. Joining the two closest clusters into one new cluster.
4. Iterating 2 to 3 times, until a single cluster, containing all the features or samples, remains.

The tree is drawn such that the distances between clusters are reflected by the lengths of the branches in the tree.

Running the tool

Go to:

> **Tools** | **Epigenomics Analysis** () | **Bisulfite Sequencing** () | **Create Methylation Level Heat Map** ()

The tool takes as input methylation level tracks ( ) generated using the **Call Methylation Levels** tool with the "Report unmethylated cytosines" option selected, as shown in figure 10.24. This option is enabled by default when running the Detect QIAseq Methylation template workflow.

For valid comparisons to be made across samples, the inputs must have been generated using the same reference information, i.e. the same reference genome, target regions, etc.

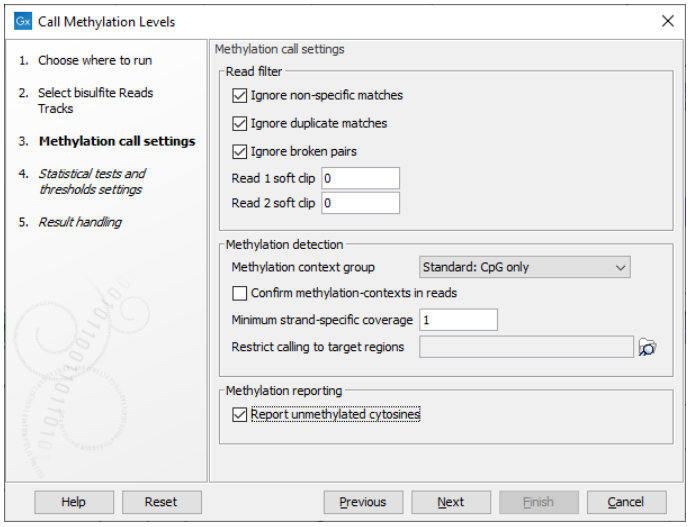


Figure 10.24: *The "Report unmethylated cytosines" option in Call Methylation Levels should be enabled when generating methylation level tracks for use with Create Methylation Level Heat Map.*

In the wizard step shown in figure 10.25, select the target region track containing the CpG sites. These may be single CpGs or larger targets (e.g. promoter regions). If no target region track is selected, single CpG sites from the methylation level tracks are used as features in the heat map.

At the bottom of this step, specify the minimum CpG site coverage value. CpG sites with coverage below this will be excluded from the analysis. By default, the value is 30. When only single CpG sites are analyzed, the methylation level of low coverage sites is set to 0.

A distance measure and a cluster linkage method for the hierarchical clustering is also specified here. The distance measure specifies how distances between two features or samples should be calculated. The cluster linkage method specifies how the distance between two clusters, each consisting of a number of features or samples, should be calculated.

There are three kinds of **distance measures**:

- **Euclidean distance**. The length of the segment connecting two points. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Euclidean distance between u and v is

$$|u - v| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}.$$

- **Manhattan distance**. The distance between two points measured along axes at right angles. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Manhattan distance

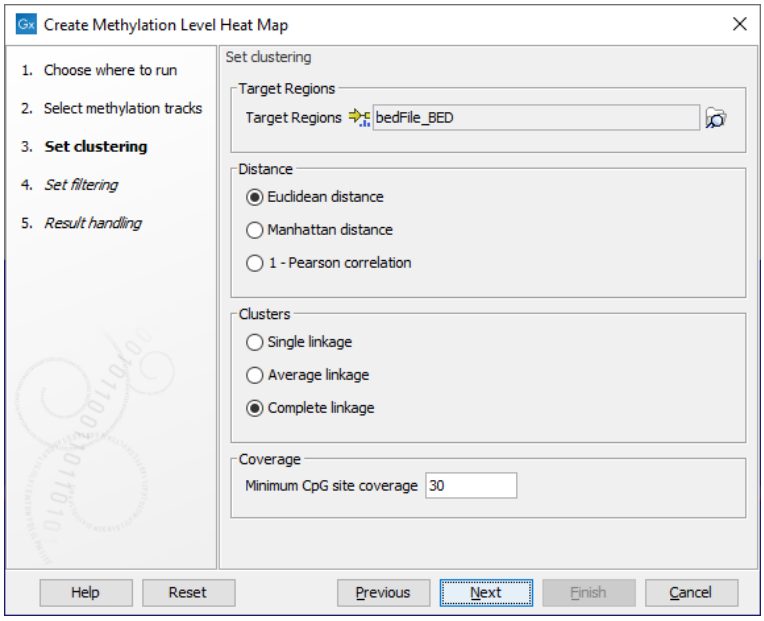


Figure 10.25: *The core options for Create Methylation Level Heat Map.*

between u and v is

$$|u - v| = \sum_{i=1}^{n} |u_i - v_i|.$$

- **1 - Pearson correlation**. The Pearson correlation coefficient between $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is defined as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \cdot \left(\frac{y_i - \overline{y}}{s_y} \right)$$

  where $\overline{x}/\overline{y}$ and s_x/s_y are the average and sample standard deviation, respectively, of the values in x/y values.

  The Pearson correlation coefficient ranges from -1 to 1, with high absolute values indicating strong correlation, and values near 0 suggesting little to no relationship between the elements.

  Using 1 - | Pearson correlation | as the distance measure ensures that highly correlated elements have a shorter distance, while elements with low correlation are farther apart.

The distance between two clusters is determined using one of the following linkage types:

- **Single linkage**. The distance between the two closest elements in the two clusters.
- **Average linkage**. The average distance between elements in the first cluster and elements in the second cluster.
- **Complete linkage**. The distance between the two farthest elements in the two clusters.

Filtering options are specified in the next step, as shown in figure 10.26.

The **Filter settings** options are described below. Some require additional information be provided in the sections underneath.

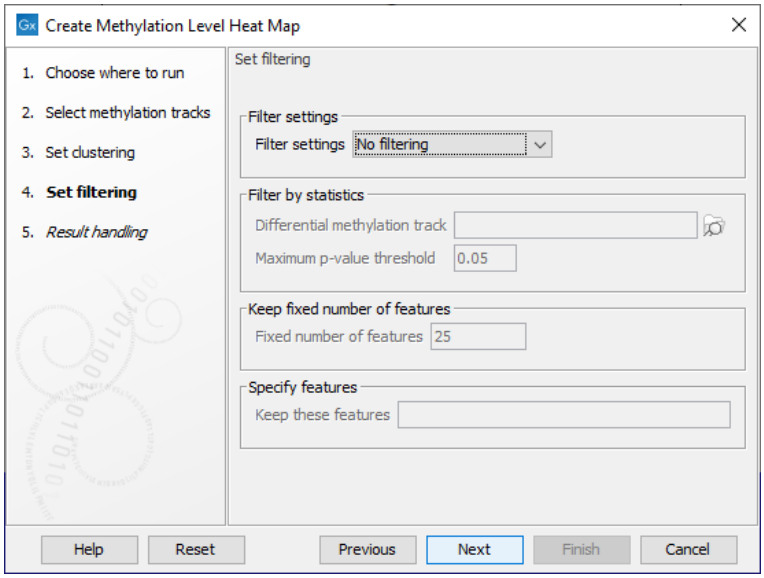


Figure 10.26: *The features to include in results can be customized using filtering options.*

- **No filtering** All features are reported in the outputs.
- **Filter by statistics** Only features meeting specified p-value and fold change thresholds in a differential methylation track you supply are reported in the outputs. Differential methylation tracks can be generated by running **Detect Differentially Methylated Regions** from the **Analyze QIAseq Panel** tool, as described in section 10.8.1.
- **Fixed number of features** Only the specified number of features with the highest index of dispersion (the ratio of the variance to the mean) are reported in the outputs.
- **Specify features** Only the features listed in the "Keep these features" field are included in the outputs. Enter the list of feature names, separated by white-space characters, commas or semi-colons. **Note**: This option can only be used if names have been defined for the target regions.

Create Methylation Level Heat Map generates two outputs: a heat map and a methylation expression track.

The methylation level heat map

Each row in the heat map corresponds to a feature (target region or single CpG site). Each column corresponds to a sample. The color in the i'th row and j'th column reflects the methylation level of feature i in sample j. The color scale can be set in the side panel settings. Heat map settings are described further at: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_heat_map_view.html.

The methylation expression track

The methylation expression track includes information from all the samples provided as input. It can be viewed as a graphical track ( ) or as a table ( ), as shown in figure 10.28.

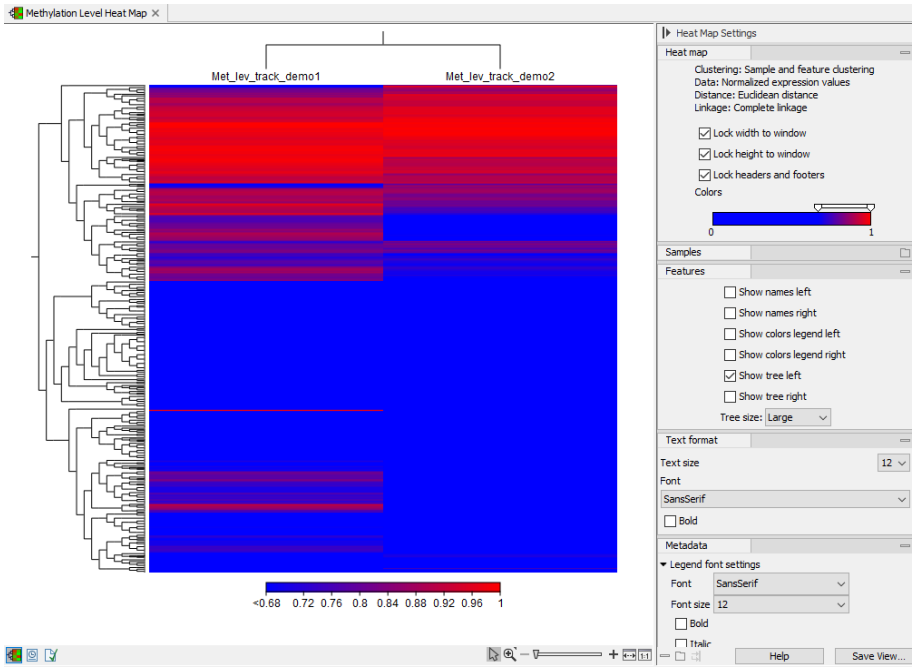


Figure 10.27: *A methylation level heat map.*

The following information is available for each feature:

- **Chromosome** The chromosome number
- **Region** The position of the target region on the chromosome
- **Expression value** Not relevant for this analysis type. All values in this column are reported as NaN.

The following four columns are provided for each sample, with the relevant sample name appended to the column name.

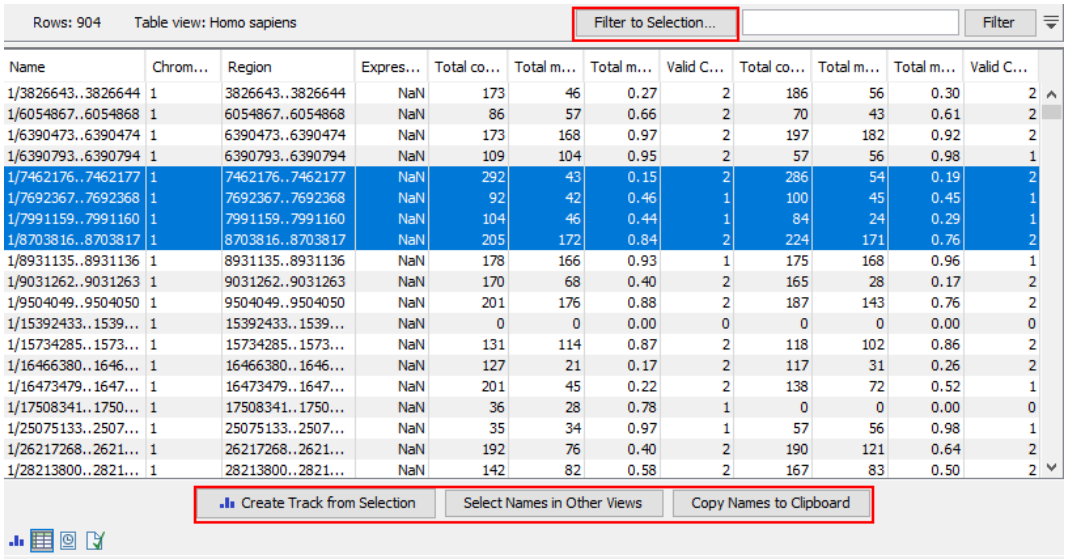
- **Total methylated coverage** Coverage of all informative methylated CpG sites
- **Total context coverage** Coverage of all informative CpG sites
- **Total methylation level** The coverage of informative methylated CpG sites divided by the coverage of all informative CpG sites
- **Valid CpG sites** The number of strand specific CpG sites included in the target region that meet the minimum CpG site coverage configured in the wizard step shown in figure 10.25

Viewing selected features

The heat map and methylation expression track created by Create Methylation Level Heat Map are linked. Selected elements in one of these outputs can be highlighted in the other. Open both outputs, preferably in a split view, and then:

- After selecting rows of interest in a heat map, right click and choose **Select Names in Other Views** from the menu, or
- After selecting rows of interest in the methylation expression track table, click on the **Select Names in Other Views** button.

The selections made in one of the outputs will now be selected in the other.



Name	Chrom...	Region	Expres...	Total co...	Total m...	Total m...	Valid C...	Total co...	Total m...	Total m...	Valid C...
1/3826643..3826644	1	3826643..3826644	NaN	173	46	0.27	2	186	56	0.30	2
1/6054867..6054868	1	6054867..6054868	NaN	86	57	0.66	2	70	43	0.61	2
1/6390473..6390474	1	6390473..6390474	NaN	173	168	0.97	2	197	182	0.92	2
1/6390793..6390794	1	6390793..6390794	NaN	109	104	0.95	2	57	56	0.98	1
1/7462176..7462177	1	7462176..7462177	NaN	292	43	0.15	2	286	54	0.19	2
1/7692367..7692368	1	7692367..7692368	NaN	92	42	0.46	1	100	45	0.45	1
1/7991159..7991160	1	7991159..7991160	NaN	104	46	0.44	1	84	24	0.29	1
1/8703816..8703817	1	8703816..8703817	NaN	205	172	0.84	2	224	171	0.76	2
1/8931135..8931136	1	8931135..8931136	NaN	178	166	0.93	1	175	168	0.96	1
1/9031262..9031263	1	9031262..9031263	NaN	170	68	0.40	2	165	28	0.17	2
1/9504049..9504050	1	9504049..9504050	NaN	201	176	0.88	2	187	143	0.76	2
1/15392433..1539...	1	15392433..1539...	NaN	0	0	0.00	0	0	0	0.00	0
1/15734285..1573...	1	15734285..1573...	NaN	131	114	0.87	2	118	102	0.86	2
1/16466380..1646...	1	16466380..1646...	NaN	127	21	0.17	2	117	31	0.26	2
1/16473479..1647...	1	16473479..1647...	NaN	201	45	0.22	2	138	72	0.52	1
1/17508341..1750...	1	17508341..1750...	NaN	36	28	0.78	1	0	0	0.00	0
1/25075133..2507...	1	25075133..2507...	NaN	35	34	0.97	1	57	56	0.98	1
1/26217268..2621...	1	26217268..2621...	NaN	192	76	0.40	2	190	121	0.64	2
1/28213800..2821...	1	28213800..2821...	NaN	142	82	0.58	2	167	83	0.50	2

Figure 10.28: *Methylation level results shown in a table view. After selecting rows in the table, the buttons highlighted can be used to work with the selection in various ways.*

Viewing results in context using a track list

Methylation expression tracks can be included in a track list with other relevant tracks, such as read mapping and annotation tracks, as shown in figure 10.29.

Further details about working with track lists can be found at: `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Track_lists.html`.

10.8.3 Predict Methylation Profile

The **Predict Methylation Profile** tool estimates the composition of an input sample. The tool is primarily designed for use with the QIAseq Targeted Methyl panel "T Cell Infiltration Panel (MHS-202Z)", though it can be used in other settings.

To start the tool, go to:

> **Tools** | **Epigenomics Analysis** ( ) | **Bisulfite Sequencing** ( )| **Predict Methylation Profile** ( )

In the first dialog, choose a methylation levels track produced by the tool **Call Methylation Levels**. It is recommended that the Call Methylation Levels tool has been run with the option to **Report unmethylated cytosines**. Click **Next** to configure the following parameters:

- **Reference**

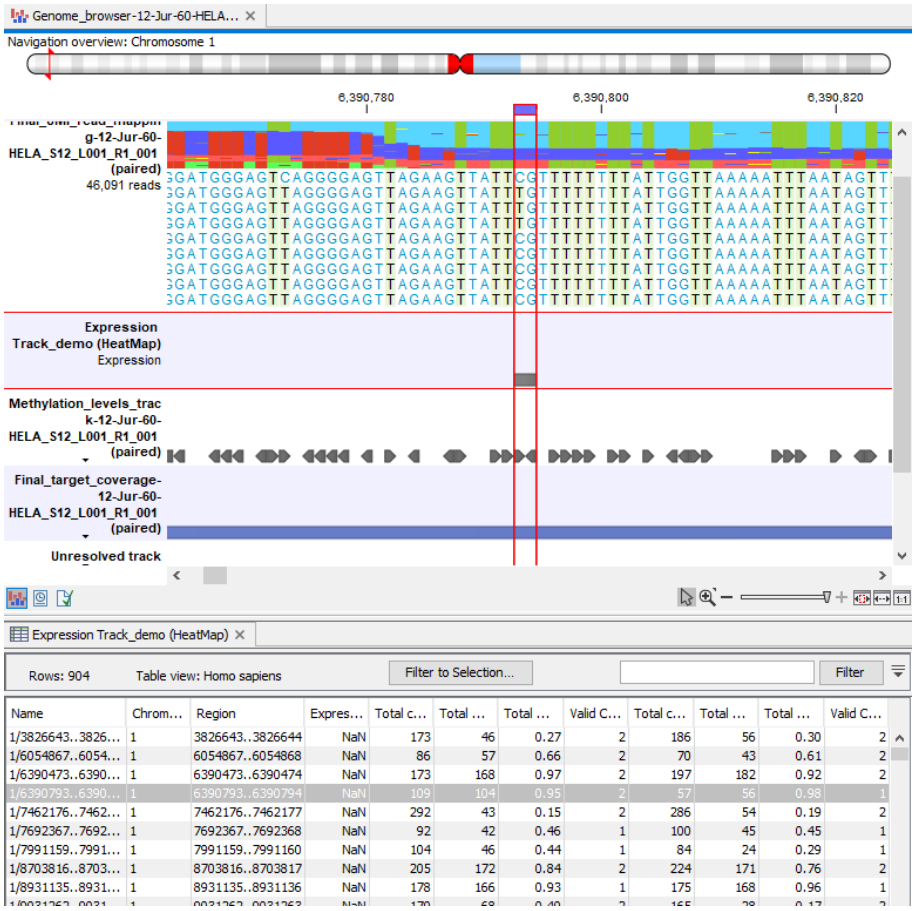


Figure 10.29: *Methylation results in the context of a track list, with the table view of the methylation expression track open in the bottom of the split view.*

- **Methylation database** An annotation track containing, for each region, known methylation levels in the range $[0, 1]$ for a 'pure' sample of a given type. There is no limit on the number of pure types that can be used, but prediction becomes more difficult as more types are added. Regions can be a single cytosine, a two nucleotide long CpG site, or a longer stretch of cytosines. When using longer stretches of cytosines, note that there are multiple ways of defining the known methylation level - for example, two ways might be as the arithmetic mean of the included cytosines, or the geometric mean. Best results will be obtained when the known methylation levels are defined as described for **Case methylation level** or **Control methylation level** in the **Call Methylation Levels** tool, for details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Call_Methylation_Levels.html.
- **Known methylation level columns** The names of columns within the Methylation database annotation track that refer to known methylation levels. The default values of "Epi", "Fib", and "IC", are chosen to match reference data distributed for the T Cell Infiltration Panel. These only need to be changed when a custom Methylation database is used.

- **Filtering**
  - **Minimum coverage** Regions are removed if their average coverage per cytosine present

in the input methylation levels track is lower than this.

– **Merge nearby regions** Methylation levels are often highly-correlated over short regions of the genome. Prediction works best when the input regions provide independent information about methylation. If two regions are closer than 1000 nt, and they appear to have compatible methylation levels, and this option is enabled, then they will be merged into a single region.

Constructing a custom methylation database

A database of two or three pure types can be created by using the **Create Methylation Database** tool described in section 10.8.4. Another option is to import a custom methylation database from an annotation file format such as GFF3, GTF, or BED, using **Import Tracks** (see `Import | Tracks`).

An example of a database with two pure types, "mcf" and "leuko", is provided in tab-delimited GFF3 format below. For more details of the format, refer to https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

```
##gff-version 3.1.26
1   .   sequence_feature   3567263 3567264 .   .   .   mcf=0.11;leuko=0.95
1   .   sequence_feature   3567272 3567273 .   .   .   mcf=0.16;leuko=1
1   .   sequence_feature   3567288 3567289 .   .   .   mcf=0.2;leuko=0.99
...
```

Note that every region must be annotated with every pure type - in the above example, the tool would skip regions that only had information for "mcf" and not "leuko". Also be sure to check that the imported coordinates match your expectations. This is easily done when the regions in the database are CpG sites, by making a new **Track List** of the reference track and the imported database, for details see: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Track_lists.html

When constructing a custom methylation database it is important to avoid bias. This can take different forms.

- **Bias in methylation level** Imagine a database where there are two pure types, "Epithelial cells" and "Blood", defined such that regions are always hypomethylated in epithelial cells and hypermethylated in blood. Then a sample with a generally high methylation will be classified as blood with high confidence. If instead the database had a mixture of hypermethylated regions for both pure types, then we might find that the same sample was predicted uncertainly with large confidence intervals.

- **Bias in the number of regions informative for each type** Imagine a database of three types "Epithelial cells", "Fibroblast cells", and "Immune cells", where regions are chosen for being able to distinguish one type from the other two. In this setup, there are three kinds of regions: those useful for distinguishing "Epithelial cells vs Fibroblast cells", "Epithelial cells vs Immune cells", or "Fibroblast cells vs Immune cells". If there are only few regions to distinguish "Fibroblast cells vs Immune cells" then the confidence intervals for these two types will likely be much larger than for epithelial cells.

The Predict Methylation Profile algorithm

The algorithm splits an input sample into a mixture of the types listed in the **Known methylation level columns** parameter. If there are n such types, then we define the fraction that is of type i to be f_i and impose the constraints $0 \leq f_i \leq 1$, $\sum_i^n f_i \leq 1$.

The regions of the methylation database are first filtered to remove regions whose coverage is lower than specified in the **Minimum coverage** option.

If the option **Merge nearby regions** is enabled, then the remaining regions are merged if they meet all the following conditions:

1. They are separated on the genome by < 1000 nt with no other region between them.
2. The known methylation levels, $k_{i,j}$ for pure type i and the two adjacent regions j and $j-1$ satisfy $|k_{i,j} - k_{i,j-1}| \leq 0.05$, $i = 1, 2 \ldots n$
3. The 95% confidence intervals of the methylation levels of the two regions overlap. The intervals are calculated using the Wilson score interval.

The n_r merged regions are then used for prediction using a maximum likelihood approach. In the underlying model, regions can be "good" regions or "bad" regions, such that $n_r = n_{good} + n_{bad}$. The predicted composition is only consistent with the methylation levels of the good regions.

The total number of bad regions is exponentially distributed to ensure that predictions requiring many bad regions are disfavored. Specifically, $n_{bad} \sim \exp(1)$. This means that the number of bad regions is a free parameter (though bound such that $n_{bad}/n_r < 0.4$) that must be determined at the same time as the sample compositions. This explicit modeling of bad regions reflects the observation that not all regions in the database will generalize to new samples as the methylation database is constructed from a limited number of samples.

For a given prediction, f_i, $i = 1, 2 \ldots n$, the expected methylation level at site j is $\beta_{expected} = \sum_i^n f_i * k_{i,j}$

The likelihood of a region with m methylated reads and coverage c being good is binomially distributed:

$$p_{good}(m|c, \beta_{expected}) \sim \mathcal{B}(c, \beta_{expected}) \tag{10.1}$$

All bad regions have the same likelihood, p_{bad}, which is a uniform distribution over the total number of regions, n_r:

$$p_{bad} = \frac{1}{1 + n_r} \tag{10.2}$$

The bad regions that maximize the log-likelihood for a given prediction f_i, $i = 1, 2 \ldots n$ are found by ordering regions by $\log p_{bad} - \log p_{good}$ in ascending order. Then up to the first n_{bad} regions with $\log p_{bad} - \log p_{good} < 0$ are bad regions.

The total log-likelihood for a given prediction is then:

$$\text{log-likelihood} = \sum_{bad} p_{bad} + \sum_{good} p_{good} - n_{bad} \tag{10.3}$$

The prediction maximizing the total log-likelihood is found by numerical optimization.

95% confidence intervals for f_i are estimated by taking the 5th percentile and 95th percentile values calculated from 1000 bootstrap iterations. In each iteration, the same number of regions as used for the initial prediction are randomly chosen with replacement from the list of regions used for that prediction, and the prediction is re-calculated on these.

The Predict Methylation Profile report

The predict methylation profile report includes the following information:

- **Summary**
  - **Total regions** The number of regions in the methylation database.
  - **Regions with too few counts** The number of regions removed by the **Low coverage** option.
  - **Regions not used** Of the remaining regions, the number that were classed as "bad" regions, and so were not used to estimate the composition of the input sample.
  - **Regions used** Regions that had sufficient coverage and were classed as "good" regions, and so were used to estimate the composition of the input sample.
  - **Merged regions used** The number of good regions after merging nearby regions. This row is only shown when **Merge nearby regions** is enabled.
- **Results** One row is shown for each type specified by the **Known methylation level columns** option.
  - **Cell type** The name of the type.
  - **Percentage** The percentage of the sample estimated to come from the given type.
  - **95% CI** A 95% confidence interval for the estimated percentage.

Note that the 95% confidence intervals only approximate the error due to the prediction being based on a selection of a small number of regions of the genome. Additional sources of error include:

- The uncertainty in each of the provided known methylation levels in the methylation database. This uncertainty can be minimized by using sites/regions with high coverage, and by validating them in many 'pure' samples.
- The presence of types in the sample that are not listed in the known methylation level columns.
- Bias in the construction of the methylation database.

As a rule-of-thumb, one should suspect the presence of these additional sources of error when $\sum_i^n f_i \ll 1$.

The Predict Methylation Profile region track

The columns of the predict methylation profile region track indicate:

- **Name** the name of the region in the Methylation database
- **Group** a number showing how regions are grouped by the **Merge nearby regions** option. If the option is disabled then each region will have a different number. Otherwise regions within 1000 nt of each other with compatible methylation levels will share a number, indicating that they were merged into one region during prediction.
- *<Known methylation level columns>* One column for each of the known methylations supplied to the tool.
- **Total methylated coverage** The sum of the coverage of all input methylated cytosines within the region.
- **Total context coverage** The sum of the coverage of all input cytosines within the region.
- **Number of cytosines** The number of input cytosines within the region for which methylation was reported. Note that the number of input cytosines is often smaller than the number of cytosines in the genome for the same region, for example because there is no read coverage at some positions.
- **Observed methylation level** The Total methylated coverage divided by the Total context coverage.
- **Predicted methylation level** The expected methylation level calculated by multiplying the predicted sample composition by the values in the *<Known methylation level columns>*. For example, if there are three such columns "Epi = 0.3", "Fib = 0.6" and "IC = 0.1", and the tool predicts that the sample is 50% Epi and 50% IC, then this value will be 0.5 x 0.3 + 0 x 0.6 + 0.5 x 0.1 = 0.2.
- **Low coverage** Whether the region was filtered away by the **Minimum coverage** option.
- **Used** Whether the region was predicted to be a "good" region i.e. not a "bad" region. Typically bad regions have large differences between the **Observed methylation level** and the **Predicted methylation level**.

10.8.4 Create Methylation Database

The **Create Methylation Database** tool can create databases for two or three conditions that can be used by the **Predict Methylation Profile**. The tool is primarily designed for use with the QIAseq Targeted Methyl panel, e.g., "T Cell Infiltration Panel (MHS-202Z)" where Fibroblast, Epithelial and Immune cells can be distinguished, further the tool is useful for creating Tumor/Normal databases for other QIAseq Targeted Methyl Cancer Panels. In addition, it can in principle be used to construct databases for any pure sample conditions where sufficient methylation differences exist.

To start the tool, go to:

Tools | **Epigenomics Analysis ()** | **Bisulfite Sequencing ()**| **Create Methylation Database ()**

In the first dialog, choose pure sample methylation level tracks produced by the tool **Call Methylation Levels** for two or three types. It is possible to use multiple tracks for each condition. It is recommended that the Call Methylation Levels tool has been run with the option to **Report unmethylated cytosines**. Specify the name of each condition matching the tracks selected, see figure 10.30.

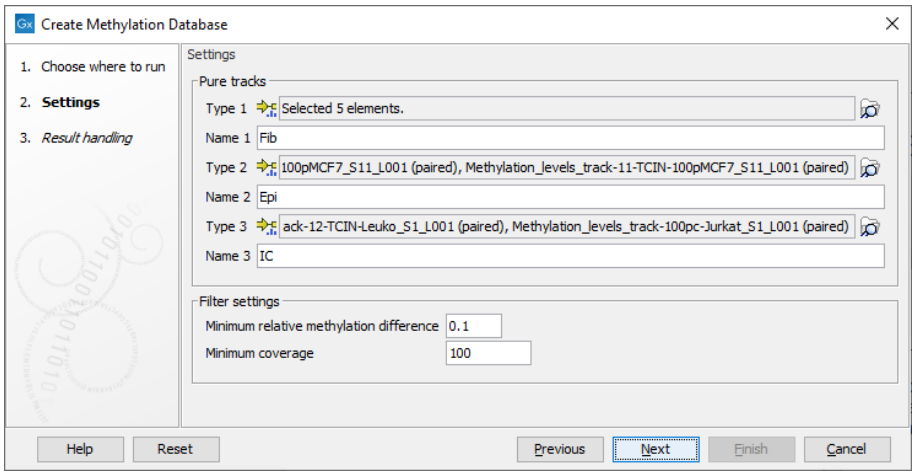


Figure 10.30: *Wizard step showing selection of pure tracks, naming and setting filters.*

Two filter options are available to specify how stringent the selection should be:

- **Minimum relative methylation difference** The minimum relative difference in methylation level allowed to include the CpG site in the database. Choose a value between 0.5 and >0.0 (default = 0.1). The choice will affect how many sites are selected and how much they differ. Different settings are shown in figure 10.31.
- **Minimum coverage** Minimum coverage in all samples for a CpG site to be considered. When the coverage is assessed across more pure samples of each type, both the main and complement strand coverage are calculated across the samples and the average is used.

We strongly recommend experimenting with the parameters to identify more optimal settings as these would differ between different experiments.

Click **Next**. The generated report will be valuable when assessing the constructed database.

The Create Methylation Database algorithm

The tool takes either two or three types of pure cells or conditions each represented by as many samples as wanted (one sample per track). It is recommended that the Call Methylation Levels tool has been run with the option to **Report unmethylated cytosines** when producing the tracks. The algorithm is constructed around two parameters, one for assessing coverage and one for specifying differences in methylation between the samples. A filtering cascade is used internally by the tool:

- **Step 1** Select all CpG sites.
- **Step 2** Keep only sites where a methylation level is assigned in all the tracks.

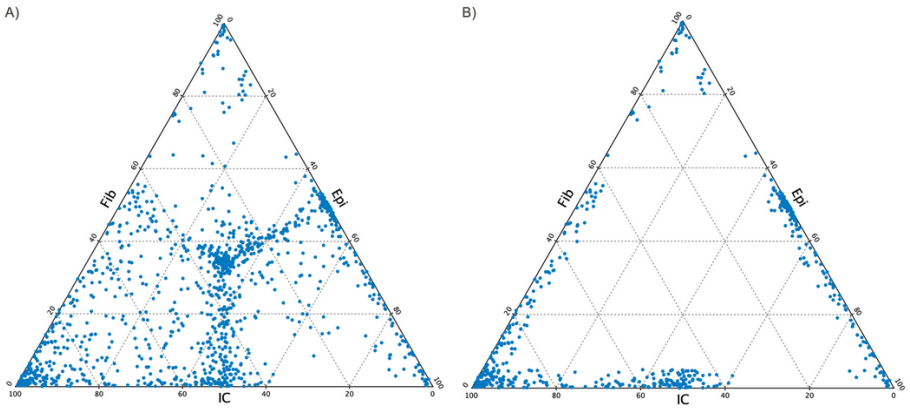


Figure 10.31: *Ternary plot created in the report when selecting three types of pure samples, Fib, Epi and IC for two different values of "minimum relative methylation difference". A) minimum relative methylation difference = 0.5, so all sites are selected. Note that the middle of the plot is populated and that these sites do not differ in methylation level between types, hence representing non-informative sites."B) A low value of the relative methylation difference illustrating a high difference between the types.*

- **Step 3** The population variance in methylation level within the set of input tracks for each pure type is less than 0.2". You can see sites filtered by this in the report, they will be reported as "Filtered due to inconsistent methylation level in pure samples
- **Step 4** Remove sites where the average coverage in the tracks for any type is below the **Minimum coverage** parameter.
- **Step 5** Remove sites where the average methylation across all samples is < 0.1. These are reported as "Filtered due to too low methylation level across samples"
- **Step 6** One of the cell type has a sufficiently different methylation level compared to the other cell types, defined by the minimum relative methylation difference parameter.

The Create Methylation Database report

The tool provides a report with a summary of selected sites. If no sites fulfill the criteria only the summary is available.

- **Summary**
  - **Included CpC Sites** The number of CpG sites in the database that is created.
  - **Filtered CpG sites** The total number of removed CpG sites.
  - **Filtered due to low coverage** The number of sites removed in step 4 of the algorithm.
  - **Filtered due to inconsistent methylation level in pure sample** The number of sites removed in step 3 of the algorithm.
  - **Filtered due to too low coverage across samples** The number of sites removed in step 5 of the algorithm.
  - **Filtered due to too small relative difference** The number of sites removed in step 6 of the algorithm.

- **Total CpG sites** The number of CpG sites in the data set.
- **Included cell types** a list of the included names.

In the next section the Average methylation levels are given per pure type category and for the individual tracks. The table is useful for assessing if the categories are evenly matched. The Average methylation level per sample should not differ too much.

In section 3 of the report the range of methylation values are shown for each of the pure input types. It is important that the hypo and hyper methylation levels is about the same within a category such that they can contribute evenly in the selected sites. This will provide the best estimates.

Finally in section 4, either a histogram or a ternary plot, depending on number of input types, illustrates the relative methylation levels across the selected CpG sites for the database. the ternary plot is illustrated in figure 10.31.

The Create Methylation Database database track

The output database track contains information on:

- **Chromosome** The chromosome number.
- **Region** The position of the CpG sites.
- **Name** A combination of chromosome and region prefixed by "Site_".
- **One column for each of the cell types in the database** The column contains the expected methylation level for the particular cell type. The header is the name of the type.

The produced track can be used as the input database for the **Predict Methylation Profile** tool. Validation can be done by creating mixtures with known amounts of each type.

10.9 Trim Primers and their Dimers from Mapping

Trim Primers and their Dimers from Mapping unaligns the primer parts of reads in read mappings, and also unaligns parts of reads identified as primer dimerization artifacts. Unaligned regions are not considered by downstream tools such as variant callers, where it would be undesirable to consider primer regions.

Trim Primers and their Dimers from Mapping was designed for use with data generated using GeneRead™ DNAseq Targeted Panels V2, where target specific primer pairs are used for multiplexed PCR-based target enrichment. We expect the tool to work with other targeted amplicon sequence data that employ target specific primer pairs, but we have not tested it for that purpose. This tool is included in the QIAGEN GeneRead Panel Analysis (legacy) template workflow, where the relevant workflow element has been named "Trim Primers and their Dimers of Mapped Reads".

Primer trimming

Target primer locations need to be imported before using this tool. Importing descriptions of primer locations from a generic text format file or from a QIAGEN gene panel primer file is described in the Import Primer Pairs section of the CLC Genomics Workbench manual: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Import_Primer_Pairs.html.

The fraction of the primer that must overlap with a read's aligned bases in order to record a primer hit is configurable.

By default, reads are only retained in the mapping if primer sequence is detected at the start of the read, but this behavior is optional.

For paired end data, if the primer found on R2 is not a member of the same target region primer pair as the primer found on R1, both members of the pair will be removed from the mapping.

Trim Primers and their Dimers from Mapping is strict regarding primer position. Primers are expected at the 5' end or the 3' end of the read, whether the end is aligned or unaligned. If there are any additional bases at the 5' or 3' end, the region will not be identified as primer sequence.

Primer dimer trimming

Two steps are involved in trimming primer dimerization artifacts:

1. All primers are compared against all others to look for pairs likely to dimerize. The minimum number of overlapping bases, used to identify primers that may dimerize with each other, is configurable. A list is compiled containing, for each primer, all primers with potential to dimerize with it.
2. If a primer *p* has been trimmed (unaligned), *and* the still-mapped section of the affected read starts with the sequence of a primer identified as dimerizing with *p*, it is assumed that the read contains primer-dimer artifact. This predicted dimerization artifact is then unaligned. If a read only contains primer-dimer artifact, the read is removed from the mapping and discarded.

The **Trim Primers and their Dimers from Mappings** tool also includes an option for trimming primers of amplicon fragments. This is particularly useful for trimming reads originating from short fragments. The **Trim primers of amplicon fragments** option allows trimming of both forward and reverse primers of reads in both directions. In cases where primers overlap, the innermost primer is used for trimming irrespective of the read orientation and expected primer pairing. In other words, the primer ID is ignored when selecting the **Trim primers of amplicon fragments** option. This allows trimming of reads that end in a region with multiple overlapping primers, where it cannot be determined which primer the read originated from and consequently how much of the read end is primer sequence.

Running the tool

To launch **Trim Primers and their Dimers from Mapping**, go to:

Tools | **Resequencing Analysis** () | **Trim Primers and their Dimers from Mapping** ()

In the first wizard step (figure 10.32), you are asked to select a read mapping. If you would like to analyze more than one read mapping, you can run the analysis in batch mode by ticking the "Batch" box in the lower left corner of the wizard. Running jobs in batch mode is described in the CLC Genomics Workbench manual: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batch_processing.html.

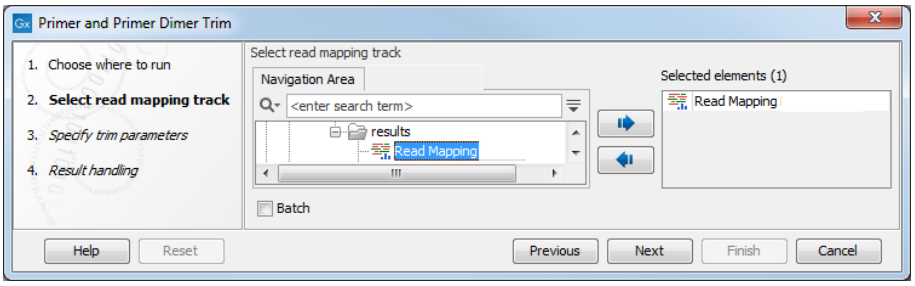


Figure 10.32: *Select a read mapping.*

In the next wizard step (figure 10.33), settings for the tool are configured.

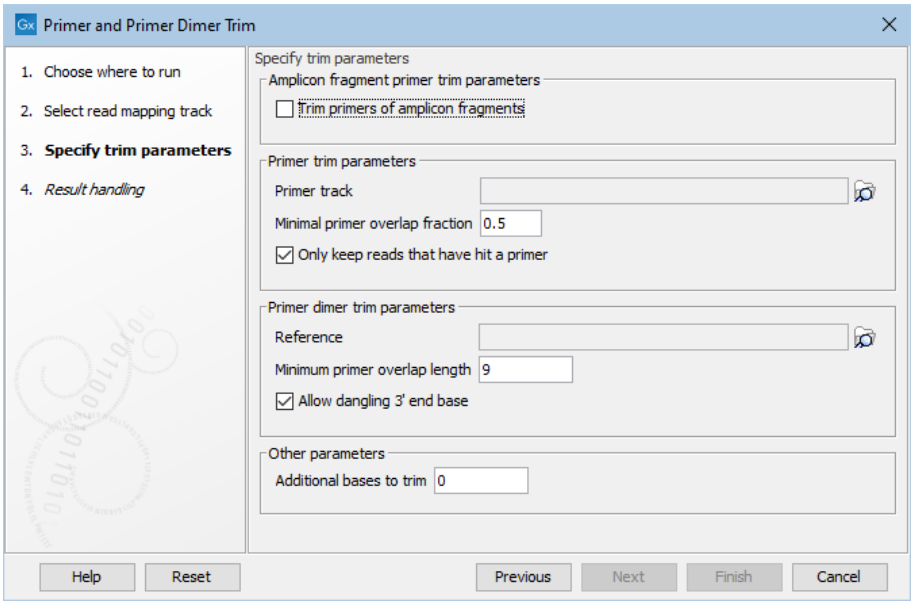


Figure 10.33: *Select your primer location file and choose whether you want to keep or discard reads with no matching primers. The option Trim primers of amplicon fragments is useful when working with reads that originate from short fragments.*

- **Amplicon fragment primer trim parameters**
  - **Trim primers of amplicon fragments** If you tick "Trim primers of amplicon fragments" all reads, regardless of orientation, can be trimmed with both forward and reverse primers.
    - **For read pairs mapping in the forward orientation (dark blue color) trim reads if:**
      - 5' end of Read 1 starts within a forward primer annotation
      - 5' end of Read 2 starts within a reverse primer annotation
      - 3' end of Read 2 ends within a forward primer annotation
      - 3' end of Read 1 ends within a reverse primer annotation
    - **For read pairs mapping in the reverse orientation (light blue color) trim reads if:**

- 5’ end of Read 2 starts within a forward primer annotation
- 5’ end of Read 1 starts within a reverse primer annotation
- 3’ end of Read 1 ends within a forward primer annotation
- 3’ end of Read 2 ends within a reverse primer annotation

– **Primer trim parameters**

* **Primer track** Click on the folder icon on the right-hand side of the wizard to select your primer location file.
* **Minimal primer overlap fraction** Specifies the fraction of the primer that must overlap with the read’s aligned bases in order to record a primer hit. Setting the fraction to 0.0 will disable this requirement.
* **Read handling configuration** If you tick "Only keep reads that have hit a primer", reads with no matching primers will be discarded.

– **Primer dimer trim parameters**

* **Reference** Click on the folder icon on the right-hand side of the wizard to select your reference location file.
* **Minimum primer overlap length** The minimum number of bases that need to bind for primers to dimerize and amplify.
* **Allow dangling 3’ end base** If you tick "Allow dangling 3’ end base", a mismatch is allowed in the primer dimerization at the 3’ end.

– **Other parameters**

* **Additional bases to trim** This number of nucleotides will be trimmed off a read right after the primer. This trimming is not done on reads for which primer-dimer artifacts were identified. This is set by default to 2 to avoid false positive calls and increase accuracy of the coverage calculation in the report.

In the last wizard step, choose the results to save and and click on **Finish**.

Output of Trim Primers and their Dimers from Mapping

The default output is a read mapping with primer and primer-dimer regions of the mapped reads unaligned. The name of the read mapping generated is based on the name of the mapping used as input, with "trimmed reads" appended. Optionally, a track containing the primer-dimer regions used for trimming the reads will also be generated. This track contains information about why each primer-dimer pair was predicted and the number of times it was used to partially trim a read or to remove a read from the mapping because it consisted only of primer-dimer sequence.

Part III

QIAseq Sample Analysis

Chapter 11

QIAseq Panel Analysis Assistant

Contents

The QIAseq Panel Analysis Assistant provides an easy entrance point for working with data generated with QIAseq panels and kits. Using the QIAseq Panel Analysis Assistant, information about the panels and kits can be accessed, available analyses can be viewed and run.

Most analyses offered via the QIAseq Panel Analysis Assistant are based on template workflows, which are available under the Workflows menu. Analyses launched using the QIAseq Panel Analysis Assistant have the appropriate reference data preselected. Additionally, some parameters are different to the template workflow, to account for the panel/kit design.

Validation of results should be carried out.

To start the QIAseq Panel Analysis Assistant, go to:

Workflows | **Template Workflows** | **QIAseq Panel Analysis Assistant (****)**

This opens a wizard listing different panels/kits categories on the left side, and analyses of panels/kits in the selected category on the right side (figure 11.1).

An analysis can be:

- A pre-configured template workflow, available from the Workflows menu.
- A pre-configured analysis tool, available from the Tools menu.
- A tool only available from within the QIAseq Panel Analysis Assistant.

Once an analysis has been selected, it can be started using **Run**. Additional actions for the selected analysis are available under **More**.

For detailed information on the QIAseq Panel Analysis Assistant, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QIAseq_Panel_Analysis_Assistant.html.

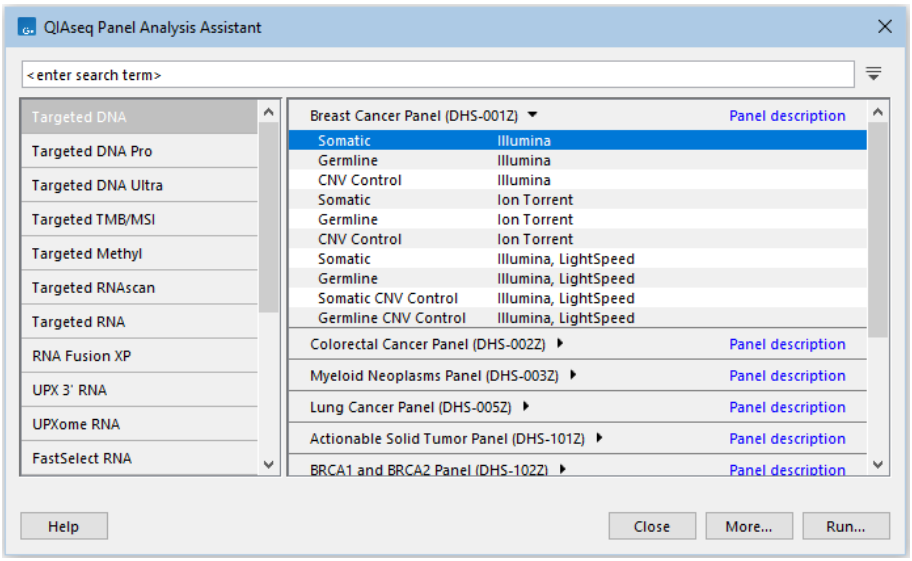


Figure 11.1: *The QIAseq Panel Analysis Assistant. Multiple analyses are available for the DHS-001Z panel. The "Panel description" links to more information about the panel.*

11.1 QIAseq custom panels

The QIAseq Panel Analysis Assistant only supports standard QIAseq panels/kits, for which reference data is already available in the *CLC Genomics Workbench*.

QIAseq custom panels are usually built from standard QIAseq panels and the analyses available for the standard panel are typically suitable for analyzing the data generated using the custom panel.

For custom panels, specific files are made available upon the purchase. These need to be imported into the *CLC Genomics Workbench* and used instead of the standard Reference Data Elements. Such a file could for example contain target primers, specified in txt files, see section 5.1 for how to import them.

Using an identified suitable analysis from the QIAseq Panel Analysis Assistant, data generated using QIAseq custom panels/kits can be analyzed by:

- **Using a Custom Set** containing the imported elements.
- **Using a workflow copy** configured to use the imported elements.

For more information on analyzing data generated using QIAseq custom panels/kits, see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QIAseq_custom_panels.html`.

11.2 Convert Legacy QIAseq Custom Analyses

Data produced using QIAseq custom panels should be analyzed as described in section 11.1.

Custom analyses created with the Analyze QIAseq Samples tool in versions 23.x and earlier[1] are

[1]Versions of the Biomedical Genomics Analysis plugin prior to 24.0 had support for adding custom analyses to the

no longer supported. The QIAseq Panel Analysis Assistant warns about existing legacy custom analyses (figure 11.2). Clicking on **Convert** starts the **Convert Legacy QIAseq Custom Analyses** tool. The tool can also be found here:

Tools | **Legacy Tools** ( ) | **Convert Legacy QIAseq Custom Analyses** ( )

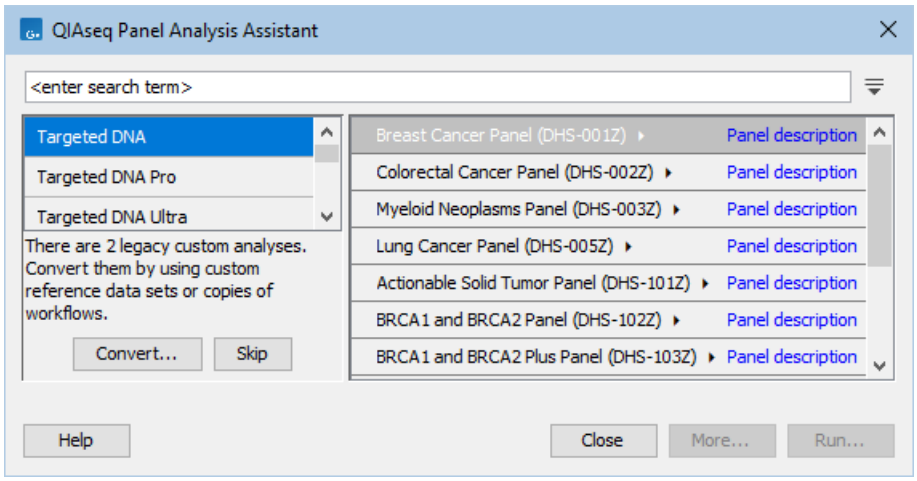


Figure 11.2: *The QIAseq Panel Analysis Assistant warns about existing legacy custom analyses.*

Convert Legacy QIAseq Custom Analyses helps converting custom analyses by creating custom sets and/or pre-configured workflow copies. Starting the tool opens a wizard listing the identified custom analyses on the left side, and details and available options for the selected custom analysis on the right side (figure 11.3).

The right side of the wizard contains the following:

Reference data set

- A custom analysis is based on a certain Reference Data Set, which is listed at the top of the wizard.
- All missing Reference Data Elements can be downloaded by clicking on the **Download to Workbench** button.
- If the *CLC Workbench* is logged into a *CLC Server*, there is also a **Download to Server** button.
- The download buttons are disabled when all Reference Data Elements have been downloaded.

Custom elements

- A custom analysis usually uses a few custom elements that are different than those in the Reference Data Set. These are listed below the Reference Data Set.
- Navigate to the different custom elements in the Navigation Area by clicking on the ( ) button.

Analyze QIAseq Samples tool. From version 24.0, the Analyze QIAseq Samples tool has been replaced by the QIAseq Panel Analysis Assistant, which does not support custom analyses.

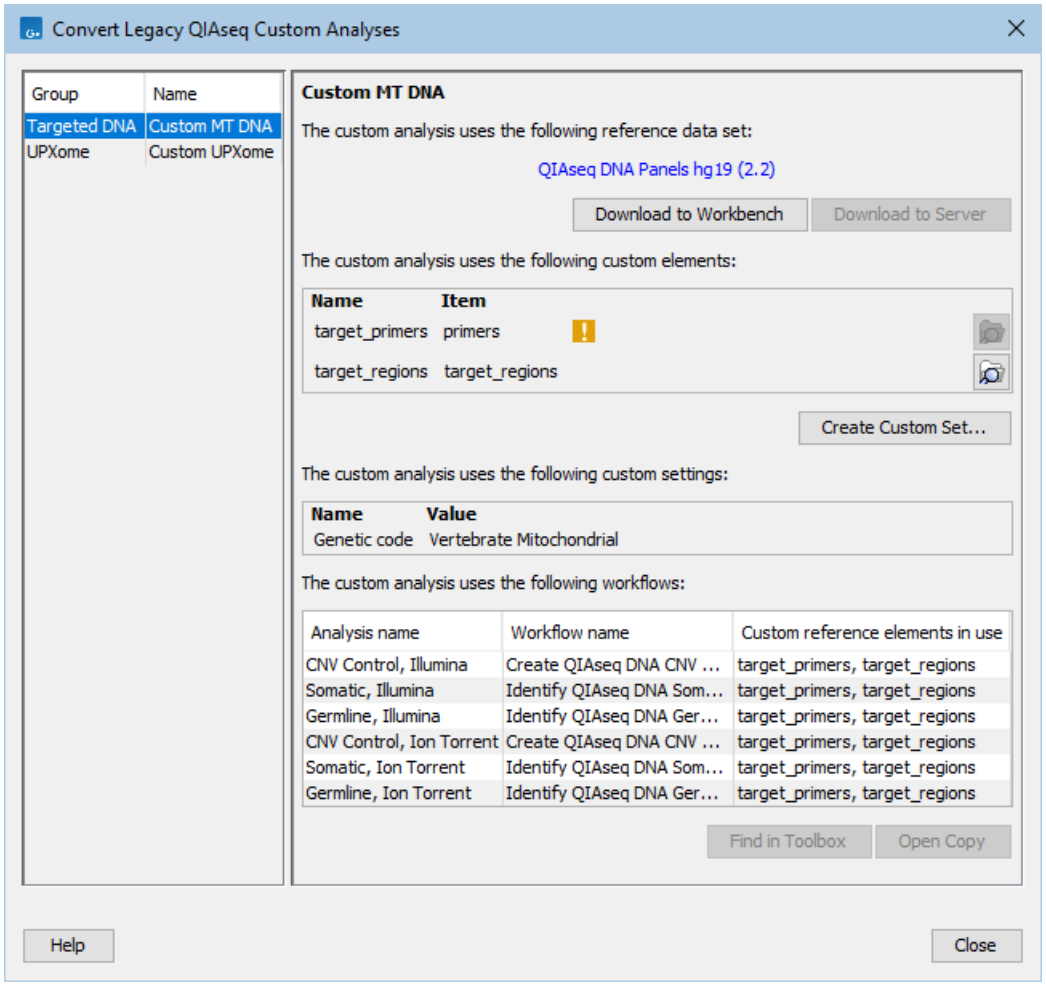


Figure 11.3: *A custom analysis for Targeted DNA that uses the "QIAseq DNA Panels hg19" Reference Data Set. There are two custom elements that are specific to this custom analysis, but the element used for the target primers could not be located. The custom analysis uses a different genetic code than the default in the template workflows.*

- Custom elements that cannot be located are marked with () and the () button is disabled.
- Create a custom set starting from the Reference Data Set and using the custom elements, by clicking on **Create Custom Set**.
  - If all needed Reference Data Elements have been previously downloaded to the *CLC Workbench*, or, if relevant, to the *CLC Server* that the *CLC Workbench* is logged into, the "Create Custom Data Set" wizard opens.
  - If there are missing Reference Data Elements, a warning wizard is shown (figure 11.4). Clicking **Yes** opens the "Create Custom Data Set" wizard and the missing Reference Data Elements are marked with () (figure 11.5).
  - The set can be further customized and saved from the "Create Custom Data Set" wizard (figure 11.5).

Custom settings

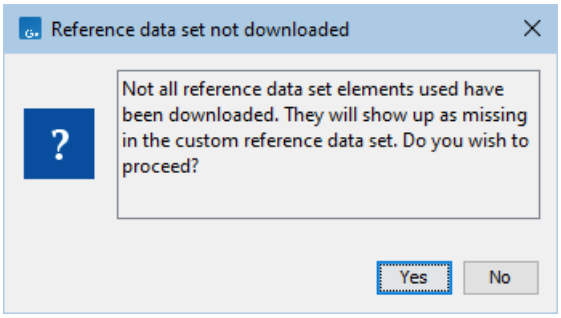


Figure 11.4: *Creating a Custom Set shows a warning message when there are missing Reference Data Elements.*

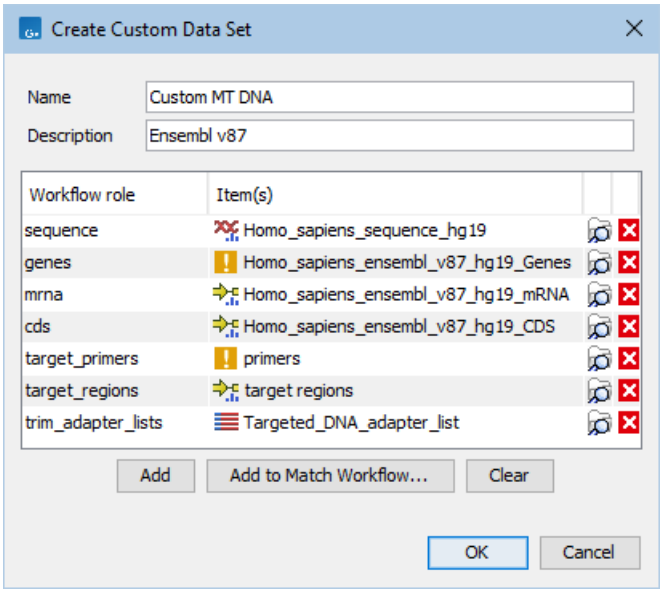


Figure 11.5: *Creating a Custom Set for the custom analysis. Both the genes Reference Data Element and the target primers custom element are missing.*

- Some custom analyses could use a different genetic code than default in the template workflows. If relevant, this is listed below the custom elements.

Workflows

- A custom analysis is based on an existing panel or kit, for which multiple workflows can be available. All relevant workflows are listed at the bottom of the wizard.
- Navigate to the selected template workflow by clicking on the **Find in Toolbox** button.
- Open a workflow copy by clicking on the **Open Copy** button.
  - The workflow copy is configured with the relevant Reference Data Set, including the custom elements.
  - If all needed Reference Data Elements have been previously downloaded to the *CLC Workbench*, or, if relevant, to the *CLC Server* that the *CLC Workbench* is logged into, the copy opens in the background.

- If there are missing Reference Data Elements, a "Reference data" wizard offers to download them. The download can be skipped by clicking **Finish**.
- When there are missing Reference Data Elements and/or custom elements, the workflow fails to validate. To run the workflow, either download the missing Reference Data Element and/or configure the workflow to use different data.
- If a different genetic code was used, the workflow is configured accordingly.
- The workflow can be further customized and once it is saved to the Navigation Area, it can be installed, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflow_installation.html for details.

Chapter 12

QIAseq DNA workflows

Contents

A series of template workflows supporting analysis of data generated using QIAseq panels and QIAseq library preparation kits are available. The workflows generally call variants, but workflows are also available that have additional functionality, such as calculating TMB or MSI.

All template workflows that include variant calling share a common structure:

- For unique molecular index (UMI) protocols, the UMI is removed
- Reads are trimmed for low quality nucleotides
- Reads are mapped
- For UMI protocols, UMI consensus reads are generated
- The read mapping is realigned
- Variants are called and filtered

The template workflows have been optimized using high quality data, settings may therefore not be appropriate for all protocols.

Template workflows are available for analysis of the following QIAseq panels:

- Single primer extension panels with UMIs:
  - QIAseq Targeted DNA Panels
  - QIAseq Targeted DNA Pro Panels
  - QIAseq Targeted DNA Ultra Panels
- Hybrid capture panels with or without UMIs:

  - QIAseq Human Exome
  - QIAseq xHYB Human Panels

In addition, template workflows designed to analyze whole genome sequencing (WGS) data can be used to analyze data from the following QIAseq library preparation kits:

- QIAseq FX DNA Library Kit
- QIAseq Multimodal DNA/RNA Library Kit
- QIAseq Ultralow Input Library Kit

12.1 Create QIAseq DNA CNV Control Mapping

Create QIAseq DNA CNV (Pro) Control Mapping template workflows generate mappings suitable for use as baseline, control mappings for the CNV detection step of the Identify QIAseq DNA (Pro) Variants template workflows (see section 12.4). We recommend a minimum of 3 control samples be used for creating control mappings.

The Create QIAseq DNA CNV Control Mapping workflows can be found at:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Create QIAseq DNA CNV Control Mapping (Illumina/Ion Torrent)** ()

The Create QIAseq DNA Pro CNV Control Mapping workflows can be found at:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Create QIAseq DNA Pro CNV Control Mapping (Illumina/Ion Torrent)** ()

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the sequencing reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details. For QIAseq Targeted DNA workflows, QIAseq DNA Panels hg19 will be pre-selected, whereas for QIAseq Targeted DNA Pro workflows, QIAseq DNA Pro Panels hg38 will be pre-selected.
- **Configure batching**. If running the workflow in **Batch** mode, you will be asked to define the batch units. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details.
- **Target primers**. Choose the relevant target primers from the drop down list.
- **Target regions**. Choose the relevant target regions from the drop down list.

- **Map Reads to Reference**. This step is available for QIAseq DNA Pro workflows only. Here, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Note that reads that span the origin of the MT chromosome are not trimmed by the Trim Primers of Mapped Reads tool when running the Identify QIAseq DNA Variants template workflows on data from the DHS-105Z panel.

Launching using the QIAseq Panel Analysis Assistant

The workflows are also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA and Targeted DNA Pro.

12.1.1 Output from the Create QIAseq DNA CNV Control Mapping workflow

The Create QIAseq DNA CNV Control Mapping template workflows produce the following outputs:

- A mapping of the UMI Reads () and a coverage table (), either can be used as a the baseline control mapping when running CNV detection.
- A per-region coverage track (), containing annotation for the coverage within the target region. The track can be used in a track list for visualization of the coverage.
- A coverage report () containing information about coverage within the target regions.
- A trim reads report () containing information about the trimming performed, including automatic adapter read-through trimming.
- A UMI Groups report () containing a breakdown of UMI groups with different number of reads, along with percentages of groups and reads.
- A Create UMI Report () containing information about the creation of UMI reads, including the numbers of reads ignored and the reason why these were not included in a UMI read.
- A sample report () containing summaries of the most important metrics from the above reports.

For more information about quality control see section 12.4.4.

12.2 Detect QIAseq MSI Status

The **Detect QIAseq MSI Status** template workflow has been designed to support the DHS-8800Z QIAseq Targeted DNA panels sequenced using Illumina technology. This particular workflow includes only the Detect MSI Status tool. The Detect MSI Status tool measures the statistical variation of the length distribution of each microsatellite locus and decides for each locus if it is stable or not by comparing the statistical variation of the test sample with the normal samples' baseline. If the proportion of unstable microsatellite loci is higher than a predefined threshold, then the sample is considered unstable.

To learn more about the workflow, read about the tool in section 8.2.

The **Detect QIAseq MSI Status** template workflow is available from the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA Workflows** ( ) | **Detect QIAseq MSI Status** ( )

The Detect QIAseq MSI Status workflow will take as input a read mapping previously generated by running the Identify TMB Status workflow.

Next, you need to select the appropriate Reference Data Set (figure 12.1).

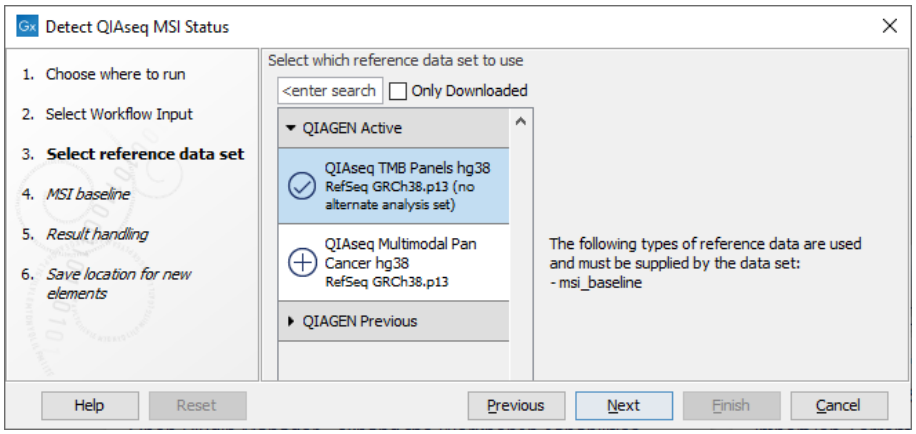


Figure 12.1: *Select the appropriate Reference Data Set.*

The Reference Data Manager includes the QIAseq TMB Panels hg38 (no alternative analysis set) set. This data set includes two template baseline tracks: one is based on a 27 microsatellite loci track containing all the microsatellite loci covered by the primers in the Human TMB and MSI Panel (DHS-8800Z). This baseline was created using 30 MSS samples that were mapped to the hg38 (no alternative analysis set) reference sequence and processed with the Generate MSI Baseline tool using the default parameters. The other is a subset of the first containing 9 loci. These loci were identified utilising lung FFPE MSS and MSI samples and were found to perform consistently well during benchmarking (figure 12.2). All 9 loci are mono nucleotide homopolymers.

The 27 loci baseline track should not be used for detecting MSI status. This baseline was included in the Reference Data Set to allow the selection of a subset of loci specific to a particular cancer type. To generate a cancer-specific baseline track, open a Track List of the 27 loci baseline track along with your samples read mappings to investigate quality of the loci of interest. Then select the loci that should be included in the baseline (in the table view of the 27 loci baseline track) and click on the **Create Track from Selection** button at the bottom of the

1 Summary

Locus	Chromosome/Region	Flanking Signature	Microsatellite Length	Total Coverage	Total Read Count
BAT40(T)37	Chromosome 1 - 119510718..119510754	TGGTTTTC/GAGACAAG	37	31971	1433
MONO-27(T)27	Chromosome 2 - 39309549..39309575	AACCAGGA/GAGGCAGA	27	29593	2491
BAT26(A)27	Chromosome 2 - 47414421..47414447	TTTCAGGT/GGGTTAAA	27	40705	14120
NR24(T)23	Chromosome 2 - 95183614..95183636	CAGTCCTA/GTGAGACA	23	23163	3070
BAT25(T)25	Chromosome 4 - 54732025..54732070	CTAAAGAG/GAGAACAG	46	53881	9711
NR22(T)21	Chromosome 11 - 125620871..125620891	GTTGAAGA/AATATGCA	21	40489	16117
HSP110-T17(T)17	Chromosome 13 - 31148484..31148500	GCACACTT/TCATGAGC	17	61654	28427
NR21(A)21	Chromosome 14 - 23183138..23183158	GTGTTGCT/GGCCAGGG	21	62027	16371
BAT34C4(A)18	Chromosome 17 - 7668827..7668865	CCAGTCTC/TTGACCCT	39	23970	9439

Figure 12.2: *Report generated when creating the 9 loci MSI baseline.*

table. Save the newly generated baseline in the Navigation Area, and create a custom Reference Data set by replacing the msi_loci and msi_baseline element with the track you just generated.

Also note that the samples were sequenced using NextSeq (Illumina) and will work best against samples sequenced on this sequencer. Similarly, when using Ion Torrent reads we recommend to create a custom baseline from at least 15 normal samples compatible with the samples of interest, i.e., samples from the same population and sequenced on same equipment and in the same lab to minimize sample preparation biases.

12.2.1 Output of the Detect QIAseq MSI Status workflow

The workflow will output a report indicating whether the sample was found to be stable (MSS) as seen in figure 12.3 or unstable (MSI) as in figure 12.4. Instability is reported as low instability (MSI-L) or high instability (MSI-H). Below the summary tables, are graphs representing length distributions for each locus (shown here to the right of the summary tables).

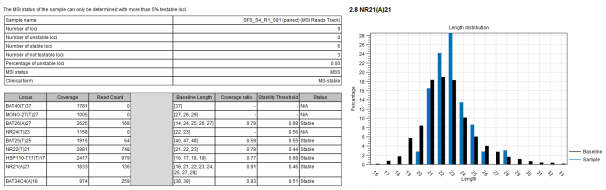


Figure 12.3: *Report from a stable sample, and graph for the locii NR21(A)21 loci.*

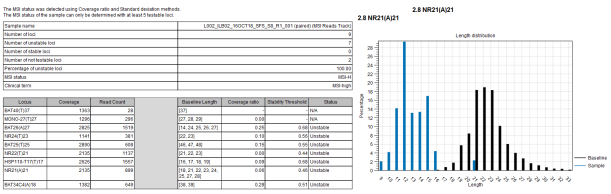


Figure 12.4: *Report from an unstable sample, and graph for the locii NR21(A)21 loci.*

Additional information about the MSI report is available in the Detect MSI Status tool description (section 8.2).

12.3 Detect MSI Status with Baseline Creation

The **Detect MSI Status with Baseline Creation** template workflow is designed to support the QIAseq MSI booster panel, SDHS-10101-11981Z. It can be used with either the hg19 or hg38 reference genomes.

This template workflow is available from the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA Workflows** ( ) | **Detect MSI Status with Baseline Creation** ( )

As input, select the reads to be analyzed (figure 12.5).

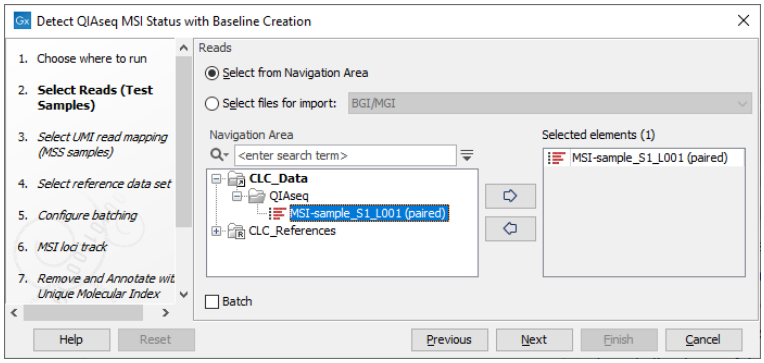


Figure 12.5: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

In the next dialog, select the UMI read mappings of the MSS samples forming the Loci Baseline (figure 12.6). Baseline mappings should be created from at least 15 MSS FFPE samples, preferably samples analyzed on the same sequencing instrument. Ideally, the samples would be from normal biopsies matched with the tumor samples being analyzed. By including the normal sample in the analysis, bias can be avoided.

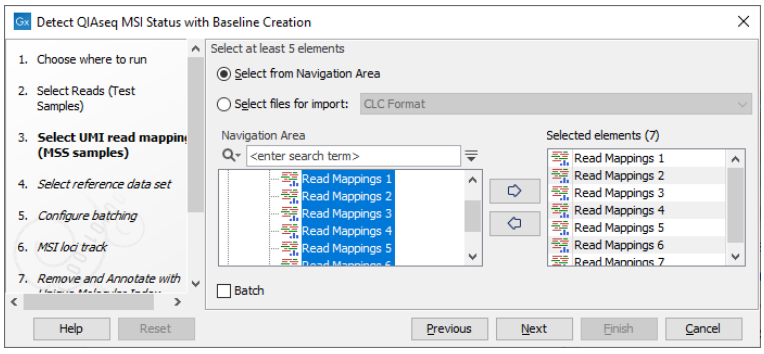


Figure 12.6: *Select at least 15 MSS samples to form a baseline.*

If you already have read mappings for the MSS samples, these can be used as input to this workflow. You may have such mappings if the MSS samples have been analyzed together with data from other QIAseq custom panels or from a QIAseq Targeted DNA panel. If you do not already have read mappings for the MSS samples then these can be generated as described at the end of this section.

In the next wizard step, choose either the QIAseq TMB Panels hg38 (no alternative analysis set) or a custom reference data set based on QIAseq DNA Panel hg19. Note that your reference data must match the reference sequence used for your MSS UMI read mappings.

To create a reference data set based on QIAseq DNA Panel hg19, first download the MSI Loci Track in hg19 coordinates. This is available as a Reference Data Element that can be found under the "QIAGEN Sets" tab of the Reference Data Manager. Then create a Custom Reference Data Set to match this workflow as described here http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

In the **Configure batching** and **Batch overview** wizard steps use the default settings. The **MSI Loci Track** step can be configured to use either the 9 loci or the 27 loci version of the loci (figure 12.6), or a custom selection if this has been created beforehand (for more on how to create a custom selection of loci see section 8.2).

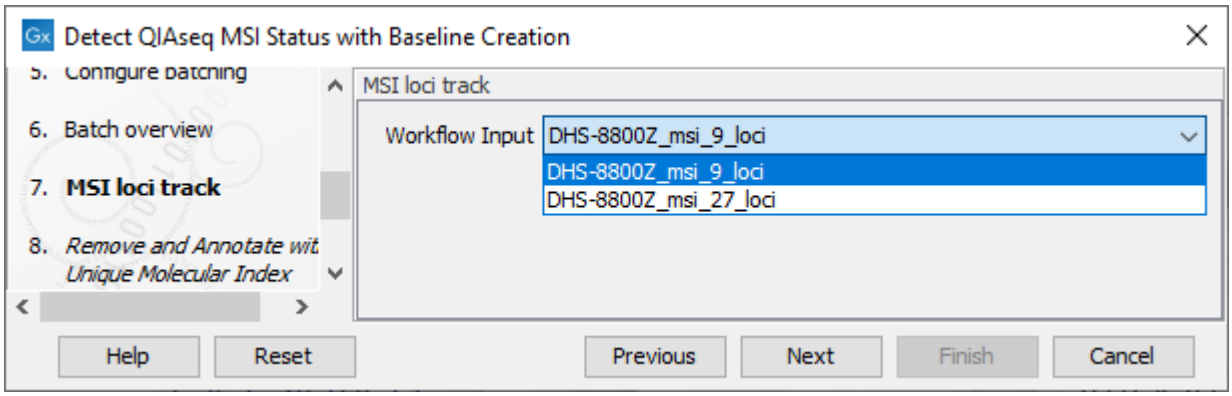


Figure 12.7: *Select an MSI loci track.*

In the Map Reads to Reference dialog, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.

Select settings for **Result handling** and a **Save location** for new elements and press **Finish** to start the analysis.

How to generate read mappings for the MSS samples

If you do not already have read mappings for the MSS samples then these can be generated by making and then running a modified version of **Detect MSI Status with Baseline Creation**, and then using the read mapping output from this as input to the original **Detect MSI Status with Baseline Creation** template workflow.

To do this:

- Open the Workflows tab in the Toolbox, in the bottom left side of the Workbench and navigate

to the location of the **Detect MSI Status with Baseline Creation** template workflow (given at the top of this section).

- Right-click on the workflow name and choose the menu option "Open Copy of Workflow".

  This opens a copy of the workflow in the Workflow Editor.

- Delete the following tools in the workflow: **Detect MSI Status** and **Generate MSI Baseline**. Then delete any orphaned Input and Output elements, i.e. elements originally linked to these tools, but now without links to other tools in the workflow.
- Save the workflow copy.
- Run the workflow copy, providing the MSS samples as input.
- Re-run the original **Detect MSI Status with Baseline Creation** template workflow using the read mapping output from the workflow copy as inputs.

12.4 Identify QIAseq DNA and QIAseq DNA Pro Variants

12.4.1 Introduction to the Identify QIAseq DNA Variants workflows

The Identify QIAseq DNA Variants template workflows are optimized to work with either somatic or germline applications from Illumina or Ion Torrent reads.

Two different types of panels are available for QIAseq Targeted DNA analysis, QIAseq Targeted DNA panels and QIAseq Targeted DNA Pro panels. The read structure is different between the two types of panels, and it is therefore important to choose the correct workflow to allow proper trimming and UMI grouping of the reads. Panel IDs for QIAseq Targeted DNA applications start with DHS or CDHS whereas panel IDs for QIAseq Targeted DNA Pro applications start with PHS or CPHS.

The workflows handling the two types of QIAseq panels are very similar, but default tool settings and the order of tools in the variant filtering cascades differ.

- General differences between QIAseq DNA and QIAseq DNA Pro analysis workflows:
  - A number of settings in the two tools Remove and Annotate with Unique Molecular Index and Trim reads differ, as they have been set up to handle reads from the relevant type of QIAseq panel appropriately.
  - In Pro workflows, an additional base after the primer is unaligned.
  - QIAseq DNA panels are designed against hg19, whereas QIAseq DNA Pro panels are designed against hg38. Consequently, using default settings, reads are mapped to hg19 or hg38, as relevant. In Pro workflows, it is possible to mask regions that are potentially false duplications using the GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set masking track during read mapping. Read about the masking track here: http://genomeref.blogspot.com/2021/07/one-of-these-things-doest-belong.html.
- Differences between Illumina QIAseq DNA and Illumina QIAseq DNA Pro analysis workflows:
  - In QIAseq DNA workflows, the minimum read length after trimming is set to 20. This has been increased to 40 in the QIAseq DNA Pro workflows.

  - The filtering cascades used for germline variant filtering varies widely between QIAseq DNA and QIAseq DNA Pro analysis workflows. Whereas the QIAseq DNA workflow has an extensive series of filtering steps, the QIAseq DNA Pro workflow has a relatively simple filtering cascade.

- Differences between Ion Torrent QIAseq DNA and Ion Torrent QIAseq DNA Pro analysis workflows:
  - In QIAseq DNA workflows, the mismatch cost and the insertion/deletion open and extend costs in Map Reads to References are 2, 6, 1, respectively. These have been increased to 6, 8, 2, respectively, in the QIAseq DNA Pro workflows.
  - In QIAseq DNA workflows, the Minimum supporting consensus fraction in Create UMI Reads from Grouped Reads is 0.0. This has been increased to 0.5 in the QIAseq DNA Pro workflows.
  - In workflows for somatic variant calling, the variant frequency in Remove False Positives is set to 0.5 in the QIAseq DNA workflow and 2 in the QIAseq DNA Pro workflow.

In the following, Identify QIAseq DNA and Identify QIAseq DNA Pro workflows are described together and are only mentioned specifically when there is a relevant difference.

To support QIAseq Targeted DNA analysis, the following workflows are available:

- Identify QIAseq DNA Somatic Variants (Illumina)
- Identify QIAseq DNA Somatic Variants (Ion Torrent)
- Identify QIAseq DNA Germline Variants (Illumina)
- Identify QIAseq DNA Germline Variants (Ion Torrent)
- Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina)

To support QIAseq Targeted DNA Pro analysis, the following workflows are available:

- Identify QIAseq DNA Pro Somatic Variants (Illumina)
- Identify QIAseq DNA Pro Somatic Variants (Ion Torrent)
- Identify QIAseq DNA Pro Germline Variants (Illumina)
- Identify QIAseq DNA Pro Germline Variants (Ion Torrent)

Note that the Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) differs from the other QIAseq DNA workflows by calling both somatic and germline variants in the same workflow and is described separately in section 12.4.5.

Somatic/germline specificity: For somatic variant detection, the template workflow uses the Low Frequency Variant Detection tool, a variant caller that does not base its statistical model on a bi-allelic assumption. This variant caller will thus declare a site heterozygous if it detects more than one allele at that site, even if one of the alleles is detected at very low frequency

and later filtered out. For germline applications, the workflows use the Fixed Ploidy Variant Detection tool. This variant caller has higher precision than the Low frequency Variant Detection tool, particularly at low to moderate levels of coverage (< 30x). At high levels of coverage (>100x) the Fixed Ploidy Variant Detection tool will exhibit low sensitivity for variants with allele frequencies far from what is expected for germline variants (that is 50 or 100%). For more information about the variant callers, please see: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html and https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Fixed_Ploidy_Variant_Detection.html.

Illumina/Ion Torrent specificity: Among various differences in the filtering strategy applied in the workflows aimed at analyzing data from a particular sequencing technology, the workflow for Ion Torrent data includes an extra step that removes non SNV type variants that are likely due to artifacts.

In each case, the parameter values applied as defaults have been optimized for high sensitivity and specificity when detecting variants.

The following description applies to the Identify QIAseq DNA (Pro) Variants template workflows optimized for calling either somatic or germline variants:

The QIAseq DNA workflows use the Reference Data set **QIAseq DNA Panels hg19** whereas the QIAseq DNA Pro workflows use **QIAseq DNA Pro Panels hg38**. Before starting one of the workflows for the first time, open the Reference Data Manager and select and download the relevant reference data set if you have not already done so.

12.4.2 Running the Identify QIAseq DNA Variants workflows

The Identify QIAseq DNA Variants template workflows are available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Identify QIAseq DNA Somatic/Germline Variants (Illumina/Ion Torrent)** ()

And the Identify QIAseq DNA Pro Variants template workflows are available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Identify QIAseq DNA Pro Somatic/Germline Variants (Illumina/Ion Torrent)** ()

Double-click on the relevant workflow to run the analysis.

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the sequencing reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details. For QIAseq Targeted DNA workflows, QIAseq DNA Panels hg19 will be pre-selected,

whereas for QIAseq Targeted DNA Pro workflows, QIAseq DNA Pro Panels hg38 will be pre-selected.

- **Configure batching**. If running the workflow in **Batch** mode, you will be asked to define the batch units. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details.
- **Target regions**. Choose the relevant target regions from the drop down list.
- **Target primers**. Choose the relevant target primers from the drop down list.
- **Map Reads to Reference**. This step is available for QIAseq DNA Pro workflows only. Here, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool. Note that the default value for this tool depends on the application chosen (somatic or germline). Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)**. Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 12.1. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Create Sample Report**. Specify QC items for assessment and subsequent flagging in the sample report generated by the workflow. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Variant filtering steps**. A series of dialogs outlines the filtering cascade in this workflow. Note that the dialog names can start with Identify candidate variants, Remove or Define. The filtering cascade has been tuned using samples of relatively high quality and coverage to provide the best possible sensitivity and precision. Note that selected default values vary by technology (Illumina / Ion Torrent), application (somatic or germline) and panel type (Targeted DNA/Targeted DNA Pro). Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.

- **Remove False Positives**. Set the minimum frequency for detected variants.
- **Add information about Amino Acid Changes**. Specify the genetic code to be used for amino acid translation.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Note that reads that span the origin of the MT chromosome are not trimmed by the Trim Primers of Mapped Reads tool when running the Identify QIAseq DNA Variants template workflows on data from the DHS-105Z panel.

Launching using the QIAseq Panel Analysis Assistant

The workflows are also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA and Targeted DNA Pro.

12.4.3 Output from the Identify QIAseq DNA Variants workflows

The Identify QIAseq DNA (Pro) Variants workflows produce a Genome Browser View () as well as the following files, available in a subfolder (as seen in figure 12.8):

- A Trim Reads report () where you can check that adapters were detected by the automatic detection option.
- A UMI Groups report () containing a breakdown of UMI groups with different number of reads, along with percentage of groups and reads.
- A Create UMI report () that indicates how many reads were ignored and the reason why they were not included in a UMI read.
- A Structural Variants report () giving an overview of the different types of structural variants inferred by the Structural Variant analysis.
- A Coverage report () and a coverage track () from the QC for Target Sequencing tool (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).
- A Sample report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The sample report also reports whether the QC thresholds specified in the Create Sample Report wizard dialog have been met.
- A read mapping of the UMI Reads ().
- Three variant tracks () : Two from the Variant Caller: the Unfiltered Variants is output before the filtering steps, the Variants passing filters is the one used in the Genome Browser View (for a definition of the variant table content, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html). The third is the Indels indirect evidence track produced by the Structural Variant Caller. This is also available in the Genome Browser View.

- An inversion and a long indels track ( ) containing identified inversions and indels longer than 100,000 bp.
- An Amino Acid track ( )
- If a read mapping was submitted in the Copy Number Variant Detection dialog, the workflow will also output three CNV tracks ( ) (Target-, Region- and Gene-level) and a CNV Results report.

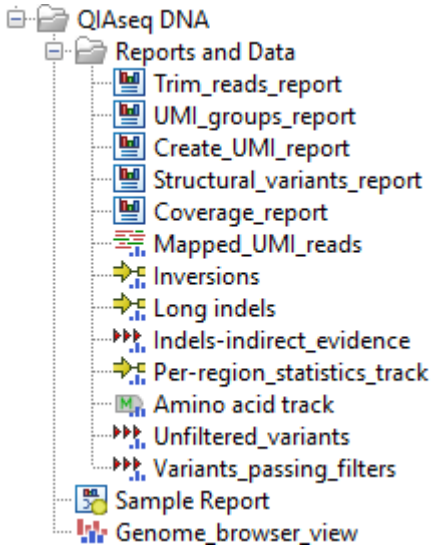


Figure 12.8: *Output from the Identify QIAseq DNA (Pro) Variants workflow without CNV detection.*

The Unfiltered variant track is included in the output so you can also review why a variant that was expected in the output would have been filtered out of the Variants passing filters track. The difference between the Unfiltered variant track and the Variants passing filters track depends on the following options available in the filtering steps:

- **Filter based on quality criteria**: Average Quality (quality of the sequenced bases that carry the variant), QUAL (significance of the variant), Read Position Test Probability (relative location of the variant in the reads that cover the variant position – only used in IonTorrent workflows) and Read Direction Test Probability (relative presence of the variant in the reads from different directions that cover the variant position).
- **Remove homopolymer error type variants**, i.e., errors of the indel type that occur in homopolymer regions. These regions are known to be harder to sequence than non-homopolymeric regions. Note that the definition of homopolymer regions differs between the pipelines due to differences in sequencing technology.
- **Remove false positive based on frequency** The variant's frequency needs to be above that threshold for the variant to be output by the workflow in the filtered variant track. Note that the unfiltered variant track is generated by the Low Frequency Variant Detection tool run with a frequency cut-off value of 0.5. This value can be considered a pre-filter, which is initially applied to each site in the alignment and determines which sites the variant caller should consider potential variant sites when it starts the error rate and site type/frequencies parameter estimation. In the case of this option, a frequency cut-off is applied on the final

candidate variant set (after variants that span across multiple alignment sites have been reconstructed). It is only meaningful to apply this post-filter at a value that is at least as high as the pre-filter value, and we actually recommend using a value that is as least twice as high (1.0). This allows for some wiggle-room when going from the single-site to the multiple site variant construction, in particular to avoid that long InDels are fragmented due to coverage difference throughout the considered region.

The workflow also produces a QC report for the target enrichment that offers statistics on the numbers of targets for which all positions are covered by the "Minimum coverage" threshold set in the QC for targeted sequencing dialog.

The read mapping of the merged UMI groups will let you verify the found variants, and examine why expected variants were not found. The UMI Groups Report gives information about the number of UMI groups found, and how many reads are in each. It includes the following information:

- How many reads were aligned to the reference (Reads in input).
- How many reads were mapped in multiple places and discarded if the Exclude ambiguously mapped reads has been selected, otherwise this value will be shown as zero.
- Groups merged: How many groups were created by merging singleton groups with other groups.
- Number of groups that were discarded for being too small (by default 0 but the option "Minimum group size" of the Calculate Unique Molecular Index Groups can be set up to discard small groups), and how many reads were thus discarded.
- How many groups were created, and of these how many were singletons groups (groups made with sequences sharing identical UMI).
- How many reads are in the largest group.
- How many different UMIs are in the most divergent group (different sequences with different UMIs can be in the same group, if they start on the same position and if they have UMIs that only differ with one character).
- Statistics about the number of reads in the groups.
- Statistics about groups size and reads not included in these groups (also available as graphs below the table).

12.4.4 Quality Control for the Identify QIAseq DNA Variants workflow

The Template QIAseq DNA Workflows for variant analysis have been configured by default to address general needs, provided that the data quality fulfills minimal standards of coverage, proper library preparation, and appropriate reads/UMI structure. We do recommend running the workflow from the QIAseq Panel Analysis Assistant a first time using the default configuration, and then inspect the QC reports: the UMI Groups Report and the Create UMI Report.

The following subsections describe how to perform proper quality controls and why they are important. For example, the relationship between the number of reads per UMI and the original coverage is not straightforward: if the original molecule has been amplified many

times, the resulting, seemingly deep, coverage will not add much information. So when the quality criteria are not fulfilled, we cannot guarantee the validity of the variant calls. After reviewing the quality controls described in this section, you can adjust workflows parameters and re-run the workflow in order to address specific experimental conditions. Parameters can be modified by opening a copy of the workflow and configuring workflow elements individually, as described at http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Basic_configuration_workflow_elements.html.

All figures in these subsections are based on three different samples that were analyzed with the QIAseq DHS-6600Z Human Tumor Mutational Burden Panel. The sample pictured to the left in each figure is a sample with few reads per UMI, the sample shown in the middle is an example of an ideal good quality sample and the one pictured to the right in each figure represents a sample with too many reads per UMI. All three samples are real samples and therefore also have different numbers of input reads which also needs to be taken into account when performing the quality control of the samples.

Proportion of reads per UMI group According to the QIAseq Targeted DNA panel Handbook and protocol, and depending on the input DNA for the library preparation, the ideal value for *Average reads per group* should be 2 to 4, with 4 being the best value for the highest DNA input (i.e., 40ng). This value can be found in the Groups table of the "UMI Groups report" (see red highlight in figure 12.9).

1 Summary

Key	Value
Reads in input	98,479,020
Reads mapped multiple places (discarded)	0
Groups merged	350,967
Groups not merged due to >1 candidate of equals size	6,456
Groups not merged due to parameter thresholds	306,402
Number of groups that were too small (discarded)	0
Number of reads in groups that were too small (discarded)	0
Output groups	32,969,355
Singleton groups	23,102,781
Reads in largest group	3,200
Number of Unique Molecular Indices in most divergent group	30
Average reads per group	1.52
Median reads per group	1.00
Standard deviation of reads per group	1.44
Reads by group size, 25th percentile	1
Reads by group size, 50th percentile	2
Reads by group size, 75th percentile	3
Groups with size = 0 (% of groups) (% of reads)	0 (0.00%) (0.00%)
Groups with size <= 1 (% of groups) (% of reads)	23,102,781 (70.07%) (46.08%)
Groups with size <= 2 (% of groups) (% of reads)	28,976,381 (87.89%) (69.51%)
Groups with size <= 3 (% of groups) (% of reads)	31,226,306 (94.71%) (82.97%)
Groups with size <= 4 (% of groups) (% of reads)	32,166,898 (97.57%) (90.47%)
Groups with size <= 5 (% of groups) (% of reads)	32,592,410 (98.86%) (94.71%)
Groups with size <= 7 (% of groups) (% of reads)	32,885,512 (99.75%) (98.41%)
Groups with size <= 10 (% of groups) (% of reads)	32,959,606 (99.97%) (99.67%)

1 Summary

Key	Value
Reads in input	57,081,901
Reads mapped multiple places (discarded)	0
Groups merged	573,287
Groups not merged due to >1 candidate of equals size	3,849
Groups not merged due to parameter thresholds	149,693
Number of groups that were too small (discarded)	0
Number of reads in groups that were too small (discarded)	0
Output groups	10,243,719
Singleton groups	4,712,264
Reads in largest group	51
Number of Unique Molecular Indices in most divergent group	9
Average reads per group	2.83
Median reads per group	2.00
Standard deviation of reads per group	2.66
Reads by group size, 25th percentile	2
Reads by group size, 50th percentile	5
Reads by group size, 75th percentile	8
Groups with size = 0 (% of groups) (% of reads)	0 (0.00%) (0.00%)
Groups with size <= 1 (% of groups) (% of reads)	4,712,264 (46.00%) (16.27%)
Groups with size <= 2 (% of groups) (% of reads)	6,459,008 (63.05%) (28.33%)
Groups with size <= 3 (% of groups) (% of reads)	7,539,584 (73.60%) (39.52%)
Groups with size <= 4 (% of groups) (% of reads)	8,276,298 (80.79%) (49.70%)
Groups with size <= 5 (% of groups) (% of reads)	8,813,285 (86.04%) (58.96%)
Groups with size <= 7 (% of groups) (% of reads)	9,510,815 (92.85%) (74.44%)
Groups with size <= 10 (% of groups) (% of reads)	10,005,385 (97.67%) (89.45%)

1 Summary

Key	Value
Reads in input	72,156,326
Reads mapped multiple places (discarded)	0
Groups merged	797,088
Groups not merged due to >1 candidate of equals size	10,913
Groups not merged due to parameter thresholds	117,838
Number of groups that were too small (discarded)	0
Number of reads in groups that were too small (discarded)	0
Output groups	3,842,755
Singleton groups	1,457,437
Reads in largest group	463
Number of Unique Molecular Indices in most divergent group	19
Average reads per group	9.59
Median reads per group	3.00
Standard deviation of reads per group	20.08
Reads by group size, 25th percentile	11
Reads by group size, 50th percentile	28
Reads by group size, 75th percentile	73
Groups with size = 0 (% of groups) (% of reads)	0 (0.00%) (0.00%)
Groups with size <= 1 (% of groups) (% of reads)	1,457,437 (37.93%) (3.95%)
Groups with size <= 2 (% of groups) (% of reads)	1,768,238 (46.01%) (5.64%)
Groups with size <= 3 (% of groups) (% of reads)	2,017,125 (52.49%) (7.67%)
Groups with size <= 4 (% of groups) (% of reads)	2,227,156 (57.96%) (9.95%)
Groups with size <= 5 (% of groups) (% of reads)	2,407,414 (62.65%) (12.39%)
Groups with size <= 7 (% of groups) (% of reads)	2,696,299 (70.17%) (17.46%)
Groups with size <= 10 (% of groups) (% of reads)	2,994,734 (77.93%) (24.67%)

Figure 12.9: *Summary section of the UMI group report.*

Average reads per group values smaller than 2 will make it impossible to create a consensus read, and therefore difficult to improve the Q scores and achieve higher precision (the advantage of using UMIs in the first place). In this case, Average quality should be adjusted for the Identify candidates variants (Low average quality variants).

Similarly, if the value is more than 6 to 8, it indicates that not enough DNA input in the library preparation resulted in an excessive PCR amplification of the same fragments. This in turn leads to decreased efficiency of the library and a lower UMI coverage as a result of creating consensus

across many reads per group. As the filtering of the variants depends on the counts (number of times a variant is observed at a UMI read), the Count filters might need further adjustments (available for Illumina workflows when configuring the Identify Candidate Variants tools for Low, Medium and High counts).

Reads by group size, 25th, 50th and 75th percentile shows the distribution of reads by group size. Examples from the figure where the 50th percentile is 5 meaning that half the reads will be in UMI groups containing up to 5 reads. If the 75th percentile is 73 it mean that 75 percent of all reads are in UMI groups containing up to 73 reads. The histogram shows the distribution of all reads by groups.

Proportion of singletons The proportion of singletons indicates how many UMI barcodes are shared across different reads. If the number is too high, this corresponds to a very low number of reads per UMI on average. This value can be found in the Groups table (see red highlight in figure 12.9) and assessed with the Read by group size plot (figure 12.10) of the "UMI groups report". When the proportion of singletons is too high, you can modify the Average quality and Count filters as described above.

Figure 12.10: *Distribution of reads by group size also showing the proportion of singletons as seen in the UMI Groups report*

Median Q scores The median Q score is correlated to the generation of a good consensus among reads with the same UMIs. When there are not enough "big UMI" reads (i.e., reads created out of consensus of 2 or more reads carrying the same UMI barcode), the Q scores will be similar to the original Q scores calculated for the raw sequencing. In this case, you cannot benefit of stringent filtering based on average Q scores that are expected to improve by the creation of UMI reads.

The improvements of the Q score values can be inspected in the report called "Create UMI Report" Summary table (figure 12.11). When there is no difference in Q score between reads and UMI reads, adjust the Average quality parameter as described above. Note that paired UMI reads are considered two reads when counting UMI reads, hence when inspecting the table in figure 12.11, there are almost twice as many UMI reads as there are UMI groups.

Composition of the barcodes used in the library prep. The base composition of the barcodes used for UMI tagging is not dependent on the user (the barcodes are random) and can also be independent of the quality of the library preparation (input, reads per UMIs etc), but it will affect the quality of the consensus creation.

If the barcodes are not unique (i.e., their base composition is in percentage of the total barcodes very small), there is a risk that the UMI consensus step will group barcodes together that are meant to tag different fragments, and therefore reducing the quality of the consensus. This can

1 Summary

Input reads	98,479,020
Ignored UMI reads (too low average quality score)	41,080
Ignored UMI reads (no consensus of any base)	0
Ignored UMI reads (too many mismatches)	776
Ignored reads due to group size smaller than minimum	0
UMI groups	32,969,355
UMI reads	64,227,884
Average Q scores for UMI reads	41.99
Median Q scores for UMI reads	36.00
Average Q scores for input reads	34.87
Median Q scores for input reads	36.00
Average Q scores for UMI read nucleotides	42.12
Median Q scores for UMI read nucleotides	36.00
Average Q scores for input read nucleotides	34.72
Median Q scores for input read nucleotides	36.00
UMI reads longer than groups reads	257,559

1 Summary

Input reads	57,081,901
Ignored UMI reads (too low average quality score)	19,500
Ignored UMI reads (no consensus of any base)	0
Ignored UMI reads (too many mismatches)	1,084
Ignored reads due to group size smaller than minimum	0
UMI groups	10,243,719
UMI reads	19,906,270
Average Q scores for UMI reads	46.95
Median Q scores for UMI reads	55.00
Average Q scores for input reads	33.54
Median Q scores for input reads	34.00
Average Q scores for UMI read nucleotides	47.86
Median Q scores for UMI read nucleotides	60.00
Average Q scores for input read nucleotides	33.53
Median Q scores for input read nucleotides	36.00
UMI reads longer than groups reads	671,698

1 Summary

Input reads	72,156,326
Ignored UMI reads (too low average quality score)	69,601
Ignored UMI reads (no consensus of any base)	0
Ignored UMI reads (too many mismatches)	699
Ignored reads due to group size smaller than minimum	0
UMI groups	3,842,755
UMI reads	6,774,670
Average Q scores for UMI reads	48.22
Median Q scores for UMI reads	58.00
Average Q scores for input reads	33.46
Median Q scores for input reads	34.00
Average Q scores for UMI read nucleotides	49.62
Median Q scores for UMI read nucleotides	60.00
Average Q scores for input read nucleotides	33.36
Median Q scores for input read nucleotides	36.00
UMI reads longer than groups reads	36,407

Figure 12.11: *Summary section of the Create UMI report highlighting average and median Q scores*

be inspected in the plot "Nucleotide percentages of the unique molecular barcode symbols" of the report "Create UMI report" (figure 12.12). For good quality, we expect the distribution to be skewed to the left as much as possible.

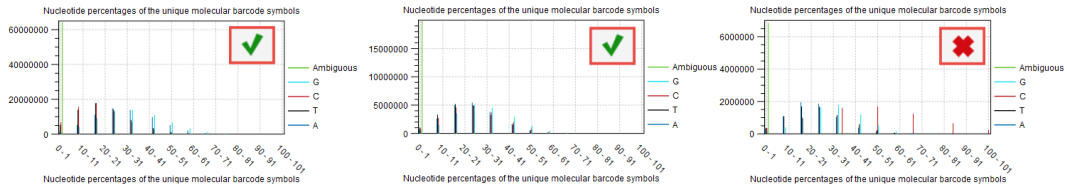


Figure 12.12: *Item no. 18 from the Create UMI report showing the nucleotide composition of the barcode.*

Coverage and accuracy when using UMIs Appropriate coverage is important to achieve the best results when using UMIs, particularly when you wish to call variants with an allele-fraction lower than 5%. In our analyses, we observe that a UMI-coverage of at least 2000X is recommended to achieve the best results.

We analyzed only variants with variant allele-fraction (VAF) of 1%, thus measuring the performance of the workflow only on variants considered more difficult to detect. In these scenarios, UMIs are expected to reduce significantly the number of false-positives, and therefore show the highest impact on precision. The plots in figure 12.13 illustrate the relationship between sensitivity and coverage, precision and overage, and accuracy and coverage, with sensitivity, precision and accuracy defined as:

Sensitivity = $TP/(TP + FN)$

Precision = $TP/(TP + FP)$

Accuracy calculated at F1 = 2 * ((Precision * Sensitivity) / (Precision + Sensitivity))

and TP: True Positives, FN: False Negatives, FP: False Positives.

A/ For sensitivity, the total number of reads are known to be an important factor, influencing the number of variants called, and hence sensitivity. Comparing our results at 2000X UMI-coverage with the results at the same level of raw coverage, the sensitivity of UMI analysis is 90.13%, while a traditional pipeline achieves 77.28%. However, comparing the results on the same subsampled

data, the effect of UMI on low VAF loci is more evident only at higher levels of coverage, because the creation of consensus reads reduces the total number of reads, although increasing their quality. The sensitivity analysis (A) also highlights an important message: when sequencing at very high coverage, hard filters may not be appropriate in all situations, as they tend to over-filter, reducing sensitivity. In these cases, adaptive filters may be necessary.

B/ The precision results show the clear advantage of UMI in reducing common errors, and hence the number of false- positive variants. Our results show that precision in detecting variants is higher for the UMI-aware analysis even at the lower coverage levels: 69.79% for UMI versus 63.77% without UMI. While at higher coverage levels, UMI-aware workflow reaches 95.22% in the present analysis, versus 78.48% of the non-UMI workflow. The results of this analysis on VAF 1% variants show that overall the precision of a UMI-aware workflow is higher than the analysis ignoring UMI information, at all coverage levels, and accuracy shows its best performance at higher levels, where sensitivity is >90%.

C/ In the following figure, we observed an accuracy of 92% for UMI-aware at this coverage, compared to 78.4% achieved by the non-UMI workflow.

The relationship between the number of reads per UMI and the original coverage is not straightforward: if the original molecule has been amplified many times, the resulting, seemingly deep, coverage will not add much information.

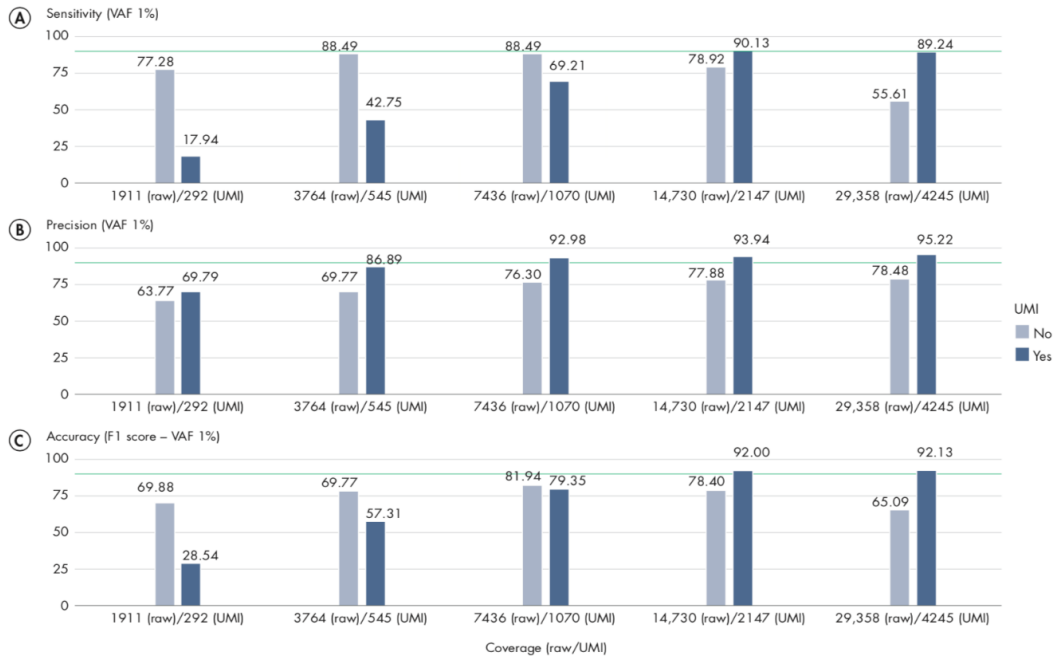


Figure 12.13: *Sensitivity, Precision and Accuracy of calling variant allele-fraction of 1% at different raw and UMI coverage values.*

12.4.5 Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina)

The ability to rapidly sequence matched tumor and normal samples enables integrated analysis of germline and somatic variants. This in turn allows for increased understanding of how each genome contributes to the disease.

Using matched tumor and normal samples as input, the Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) template workflow detects and reports somatic variants and germline variants. To ensure that no somatic variants are filtered out due to field effects in apparently normal tissue, or tumor cells being present in matched blood specimens used as germline DNA, this workflow also reports somatic variants found at low frequency in germline sample in a separate output.

The workflow can be used with all QIAseq DNA panels and before running the workflow, you must first open the Reference Data Manager, select **QIAseq DNA Panels hg19**, download the set, if you have not done so already, and close the References management window. Please note that two reference data sets are available for this workflow; an ENSEMBL set and a RefSeq set, both matching the hg19 genome assembly.

The **Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina)** template workflow is available under the Workflows menu at:

> **Workflows |Template Workflows | Biomedical Workflows** () **| QIAseq Sample Analysis** () **| QIAseq DNA workflows** () **| Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina)**

Note: If you have more than one matched tumor-normal pair that you would like to analyze, you can consider analyzing them in one workflow run. To be able to do this, the matched samples must be paired correctly. This is done using metadata, which, in essence, is a table containing information that can match the samples based on sample IDs and information about whether a sample is of tumor or normal origin. A description of how to import a metadata table or create a metadata table in the CLC Workbench can be found by following this link `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Creating_metadata_tables.html`.

Double-click on the Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) to run the analysis.

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

The next two dialogs allow you to select the tumor sequencing reads (figure 12.14) and matching normal sequencing reads (figure 12.15) that you wish to analyze.

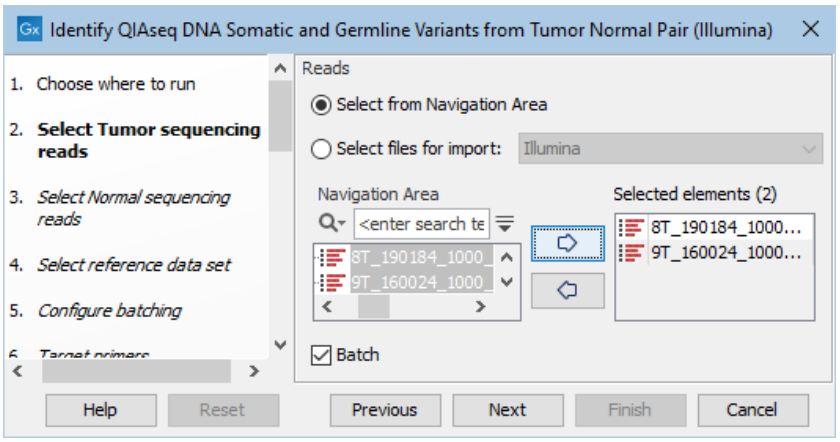


Figure 12.14: *Select the tumor sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

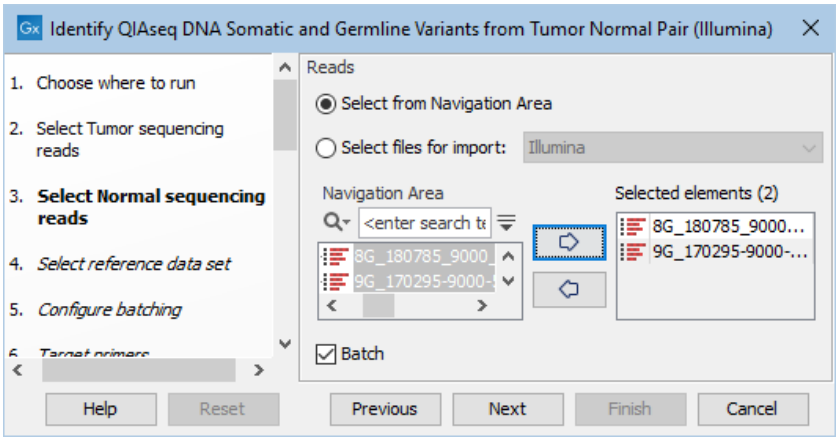


Figure 12.15: *Select the normal sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side. If you want to analyze multiple samples in one analysis run, the Batch option should be checked in the lower left corner of the dialog.*

If you would like to analyze multiple matched tumor-normal samples in one analysis run, you must use the batch option. This is described in the CLC Workbench user manual `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html`.

The following dialog helps you set up the relevant Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. This is shown in figure 12.16.

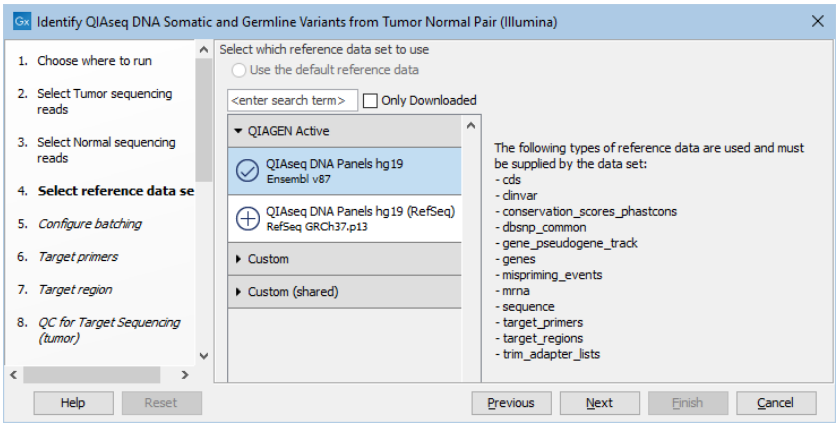


Figure 12.16: *In the central part of the dialog, the relevant Reference Data Set is highlighted. In the right-hand side, the types of references needed by the workflow are listed.*

Note that if you wish to Cancel or Resume the Download, you can close the template workflow and open the Reference Data Manager where the Cancel, Pause and Resume buttons are available.

If the Reference Data Set was previously downloaded, the option "Use the default reference data" is available and will ensure the relevant data set is used. You can always check the "Select a reference set to use" option to be able to specify another Reference Data Set than the one suggested.

If you have chosen to run more than one matched tumor-normal pair and have checked the batch box when selecting tumor and normal sequencing reads, you will now be asked to configure the batching as shown in figure 12.17.

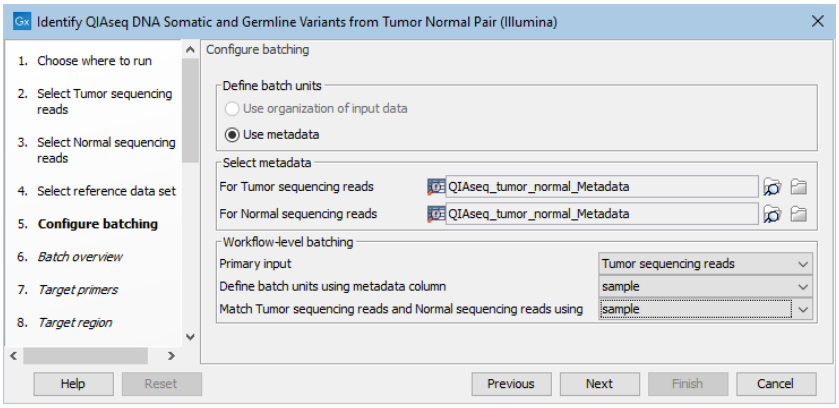


Figure 12.17: *Define the batch units using metadata if you are analyzing more than one matched tumor-normal pair in a workflow run.*

A metadata table must be provided for both the tumor and normal sequencing reads. The metadata table can be one table that holds information about both the tumor and normal samples, or it can be two individual metadata tables, one with information about the tumor samples and one with information about the normal samples. In this step you must also define the **Workflow-level batching**:

- **Primary input** For this workflow it makes no difference whether the primary input is tumor or normal as the primary input determines the number of times the workflow should be run and the number of tumor and normal samples is the same when working with matched samples.
- **Define batch units using metadata column** The column in the metadata table specifying the group the data belongs to. Each group makes up a single batch unit.
- **Match Tumor sequencing reads and Normal sequencing reads using** The column in the metadata file(s) that is used to ensure that the correct data from each workflow input are included together in a given batch run. A column with this name must be present in the metadata file(s) or table(s).

For more information about running the workflow using metadata, please see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batching_workflows_with_more_than_one_input_changing_per_run.html`.

The next dialog allows you to check how the tumor and normal samples are being matched based on the batch configuration done in the previous step (figure 12.18).

In the next dialog, specify the relevant **target primers** from the drop down list (figure 12.19).

Next, specify the relevant **target regions** from the drop down list (figure 12.20).

In the dialog called QC for Target Sequencing (tumor), you can modify the Minimum coverage for the tumor sample, which is the minimum coverage needed on all positions in a target for this target to be considered covered. This is shown in figure 12.21.

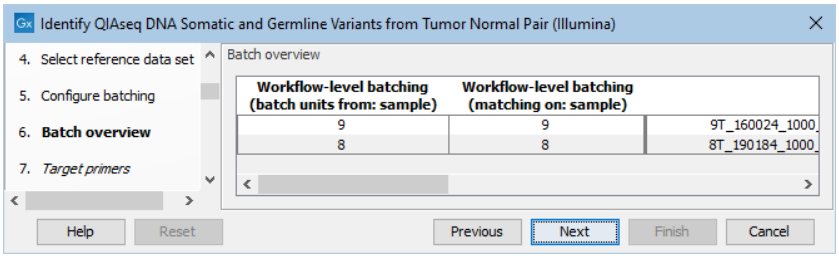


Figure 12.18: *Check how the tumor and normal samples are being matched in batch units and proceed to the next step if everything looks as expected.*

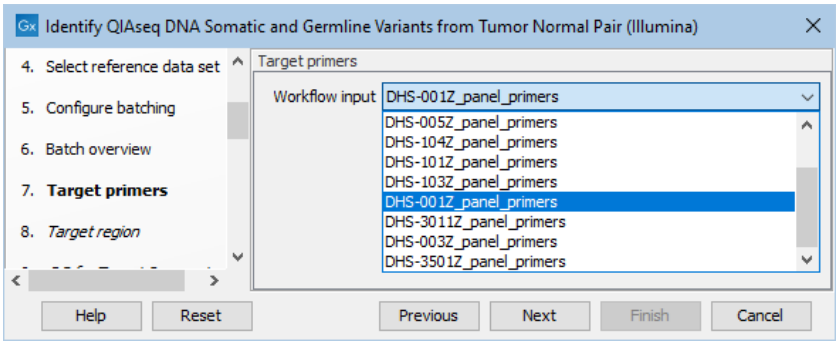


Figure 12.19: *Select the target primers file specific to the panel used.*

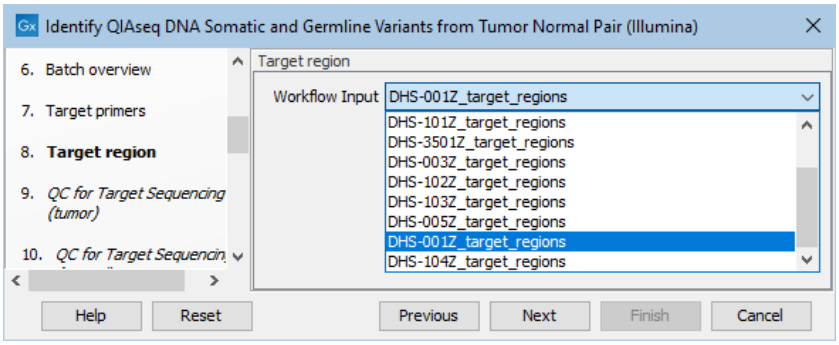


Figure 12.20: *Select the target regions file specific to the panel used.*

In the dialog called QC for Target Sequencing (normal), you can modify the Minimum coverage for the normal sample, which is the minimum coverage needed on all positions in a target for this target to be considered covered. This is shown in figure 12.22.

The Identify candidate variants (Low average quality variants) (germline) dialog allows adjustment of the cutoff for the average base quality score of the bases supporting a variant for germline variants (figure 12.23). Variants below the specified cutoff will be filtered out.

Likewise, the Identify candidate variants (Low average quality variants) (somatic) dialog allows adjustment of the cutoff for the average base quality score of the bases supporting a variant for somatic variants (figure 12.24). Variants below the specified cutoff will be filtered out.

The Identify candidate variants (Low count variants likely due to artifacts) (somatic) dialog allows adjustment of the cutoff for the average quality for somatic variants as well as the read direction test probability (figure 12.25). Variants below either of the specified cutoffs will be filtered out.

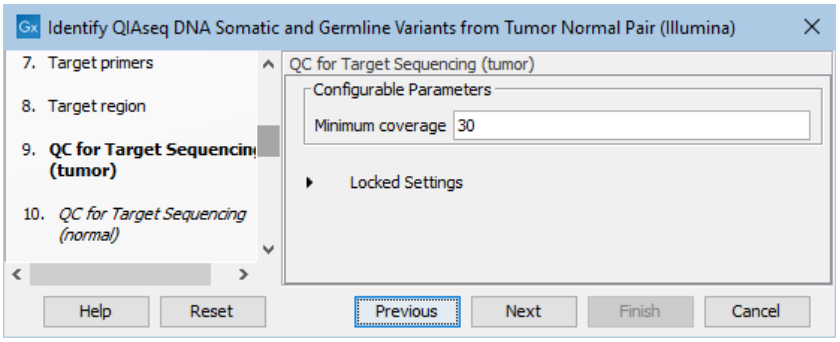


Figure 12.21: *Setting the Minimum coverage parameter of the QC for Target Sequencing for the tumor sample.*

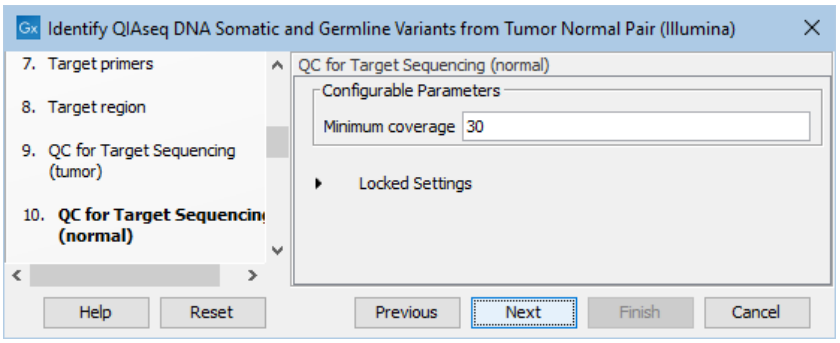


Figure 12.22: *Setting the Minimum coverage parameter of the QC for Target Sequencing for the normal sample.*

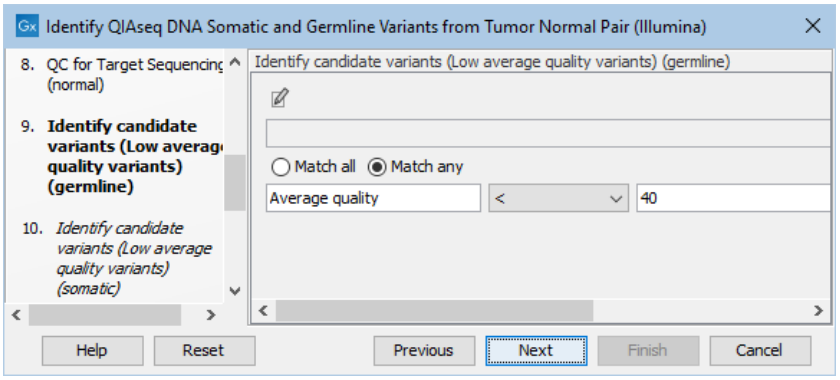


Figure 12.23: *Setting the cutoff for the minimum average quality for germline variants.*

The Remove False Positives (Filter on allele frequency) (germline) dialog allows adjustment of the cutoff for the minimum frequency required for reporting a germline variant (figure 12.26). Variants below the specified frequency cutoff will be filtered out.

Finally, you can specify where to save the data. If you are analyzing more than one matched tumor-normal pair in one run, make sure that you check the box **Create subfolders per batch unit** in the Result handling step (figure 12.27) to get a separate folder for the results from each individual matched tumor-normal pair.

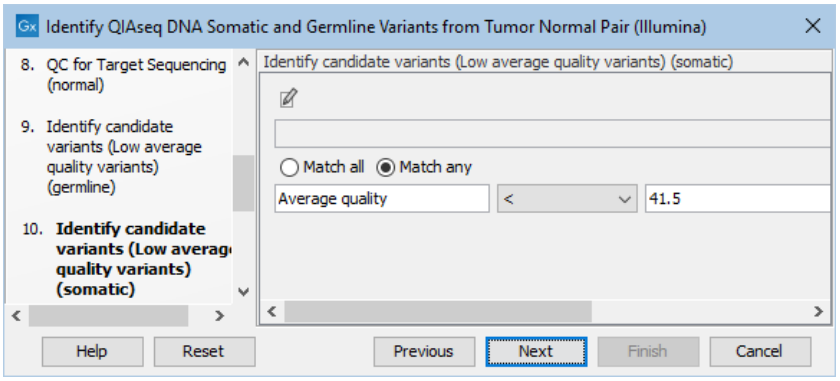


Figure 12.24: *Setting the cutoff for the minimum average quality for somatic variants.*

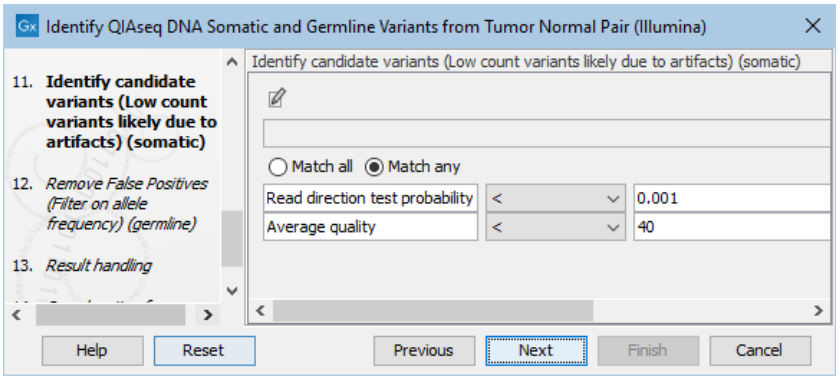


Figure 12.25: *Setting the cutoff for the minimum average quality and read direction test probability for low count variants.*

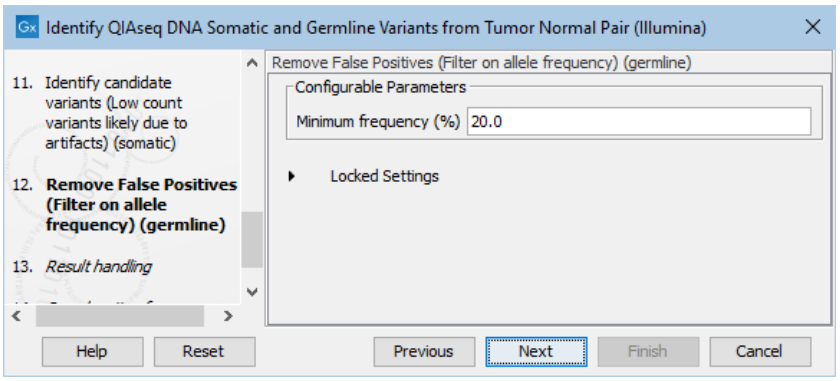


Figure 12.26: *Setting the cutoff for the minimum average quality and read direction test probability for low count variants.*

12.4.6 Output from the Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) workflow

The Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) workflow produces a number of different outputs, some of which are available in a subfolder called Reports and Data. Figure 12.28 shows the structure of the output when looking at the analysis output for one matched tumor-normal pair.

The outputs generated from the Identify QIAseq DNA Somatic and Germline Variants from Tumor

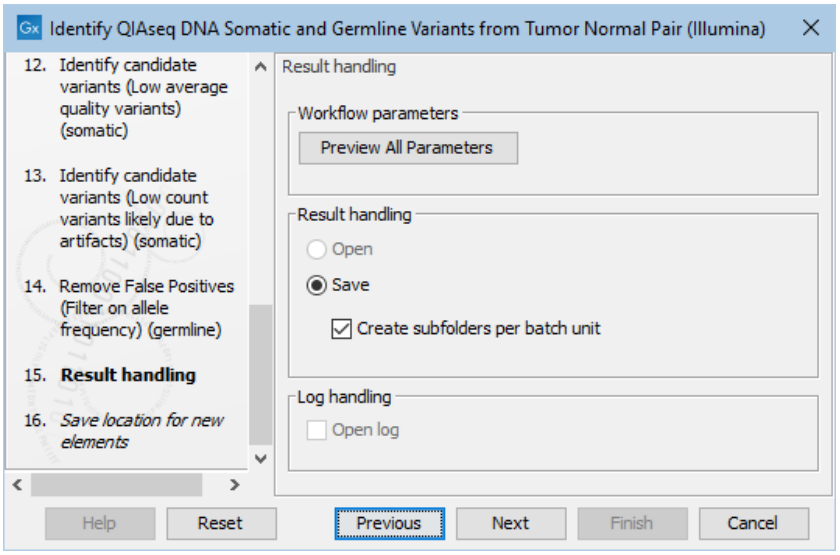


Figure 12.27: *In the Results handling step you can choose to save the analysis results in subfolders if you check the Create subfolders per batch unit box.*

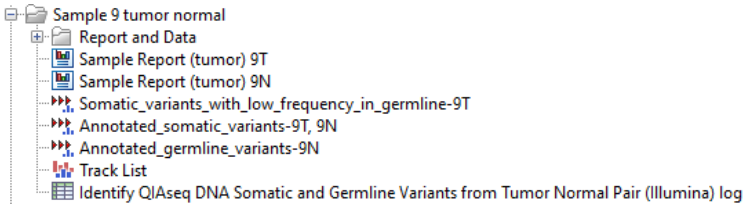


Figure 12.28: *The outputs produced by the Identify QIAseq DNA Somatic and Germline Variants from Tumor Normal Pair (Illumina) workflow. A Track List, three main variant tracks and sample reports for the tumor and the normal sample, respectively are directly accessible, whereas the remaining outputs are placed in a subfolder called Reports and Data.*

Normal Pair (Illumina) workflow are:

- **Sample Report Tumor** ( ): A sample report for the tumor sample that summarizes information about the most important metrics for the tumor sequencing reads.
- **Sample Report Normal** ( ): A sample report for the normal sample that summarizes information about the most important metrics for the normal sequencing reads.
- **Somatic_variants_with_low_frequency_in_germline** ( ): This variant track holds somatic variants that also are found at a low frequency in the normal sample. More specifically, somatic variants that also are found in more than five normal UMI reads but with a frequency below 20% in the normal sample are reported in this track. For a definition of the variant table content please see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html.
- **Annotated_somatic_variants** ( ): This variant track holds somatic variants detected in the tumor sample with a variant allele frequency cutoff at 4% and observed in at most five normal reads.

- **Annotated_germline_variants** (): This variant track holds germline variants detected with a frequency of at least 20% in the normal reads.
- **Track List** (): A graphic representation that allows for visual inspection of the results. The Track list contains the following collection of tracks:
  - Homo sapiens reference sequence
  - Reference gene track
  - Reference mRNA track
  - Per-region_statistics_track_normal
  - Per-region_statistics_track_tumor
  - Mapped_UMI_reads_normal
  - Mapped_UMI_reads_tumor
  - Annotated_somatic_variants
  - ClinVar variants
  - Amino_acid_changes_somatic_variants
  - Somatic_variants_with_low_frequency_in_germline
  - Annotated_germline_variants
  - dbSNP Common
  - Amino_acid_changes_germline_variants
  - Low_coverage_in_normal
  - Low_coverage_in_tumor
  - Primer track
  - Unfiltered_somatic_variants
- **Reports and Data** (): A subfolder holding the outputs listed below.
  - **Mapped_UMI_reads_tumor** (): A read mapping of the UMI Reads for the tumor sample.
  - **Per-region_statistics_track_tumor** (): A coverage statistics track for the tumor sample with coverage information for each targeted region.
  - **Coverage_report_tumor** (): A coverage report for the tumor sample, generated by the QC for Target Sequencing tool (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).
  - **Low_coverage_in_tumor** (): A coverage track reporting regions with low coverage in tumor. Regions with a coverage below 40x is reported as low-coverage regions.
  - **Mapped_UMI_reads_normal** (): A read mapping of the UMI Reads for the normal sample.
  - **Low_coverage_in_normal** (): A coverage track reporting regions with low coverage in the normal sample. Regions with a coverage below 40x is reported as low-coverage regions.
  - **Unfiltered_low_frequency_germline_variants** (): All detected low-frequency germline variants before applying any filters.

- **Unfiltered_somatic_variants** (): All detected somatic variants before applying any filters.
- **Low_frequency_germline_variants** (): Germline variants found with a frequency below 20%.
- **Unfiltered_germline_variants** (): All detected germline variants before applying any filters.
- **Amino_acid_changes_somatic_variants** (): A track showing amino acid changes introduced by somatic variants.
- **Amino_acid_changes_germline_variants** (): A track showing amino acid changes introduced by germline variants.

Many of the generated reports hold information about different quality control metrics. For further information about what specifically to be aware of regarding quality control, please see section 12.4.4.

The unfiltered variant tracks are provided to allow you to review the raw unfiltered variants. This can be relevant in cases where expected variants are missing from the filtered variants output and potentially have been filtered out due to low quality. The difference between the unfiltered variant track and the variants passing filters track is described in section 12.4.3.

12.5 Identify QIAseq DNA Pro Somatic Variants with LOH Detection

The **Identify QIAseq DNA Pro Somatic Variants with LOH Detection (Illumina)** template workflow has been designed to detect LOH events on chromosomes 1p and 19q using these panels:

- PHS-004Z Brain Cancer Research Panel
- PHS-104Z Brain Cancer Focus Panel
- PHS-3000Z Comprehensive Cancer Research Panel
- PHS-3100 Comprehensive Cancer Focus Panel

Reference data for the panels is available in the reference data set **QIAseq DNA Pro Panels hg38**. Hence, when running the workflow, all the Targeted DNA Pro panels are available for selection, however, only the ones listed above have been designed to support detection of LOH.

To run the workflow go to:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA Workflows** () | **Identify QIAseq DNA Pro Somatic Variants with LOH Detection (Illumina)** ()

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

A detailed description of the individual steps of the workflow is available in section 12.4.2. Note that to run LOH detection the workflow needs CNV control mappings or control coverage tables which can be provided in the Copy Number Variant Detection (Targeted) dialog.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA Pro.

12.5.1 Output from the Calculate LOH workflow

This workflow creates three elements in addition to the normal DNA Pro workflows output elements: Target-level_ploidy_track, Region-level_ploidy_track, and Regional_ploidy_results_report. These elements are described in the manual for Detect Regional Ploidy (see section 10.3).

12.6 Identify QIAseq DNA Pro Somatic Variants with MSI (Illumina)

The **Identify QIAseq DNA Pro Somatic Variants with MSI (Illumina)** template workflow has been designed to identify somatic variants and detect microsattelite instability (MSI) status). It is set up to analyze data from one of the following panels:

- PHS-001Z Breast Cancer Research Panel
- PHS-002Z Colorectal Cancer Research Panel
- PHS-101Z Breast Cancer Focus Panel
- PHS-102Z Colorectal Cancer Focus Panel
- PHS-202Z Hereditary Colorectal Cancer Panel
- PHS-205Z Hereditary Pancreatic Cancer Panel
- PHS-3000Z Comprehensive Cancer Research Panel
- PHS-3100Z Comprehensive Cancer Focus Panel
- PHS-3200Z Comprehensive Hereditary Cancer Research Panel

Reference data for the panels is available in the reference data set **QIAseq DNA Pro MSI Panels hg38**.

To run the workflow go to:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA Workflows** () | **Identify QIAseq DNA Pro Somatic Variants with MSI (Illumina)** ()

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

A detailed description of the individual steps of the workflow is available in section 12.4.2. Note that to detect the MSI status of a sample, the workflow requires an MSI baseline which is provided in the MSI baseline dialog. Per default, the workflow will use a demo baseline based on 20 microsattelite stable (MSS) samples. We recommend creating a new baseline with at least 30 MSS samples processed under the same lab conditions as test samples, see section 8.3 for additional details.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA Pro.

12.6.1 Output from the Identify QIAseq DNA Pro Somatic Variants with MSI (Illumina) workflow

This workflow creates two elements in addition to the standard QIAseq DNA Pro workflow output elements:

- **MSI report** (): Summarizes the performed MSI detection and contains the MSI status of the sample. See **Output from Detect MSI Status**, see section 8.2.1 for details.
- **Loci track**: MSI loci annotated with predicted stability.

12.7 Identify QIAseq DNA Somatic Variants with HRD Score (beta)

The Identify QIAseq DNA Somatic Variants with HRD Score (Illumina) (beta) template workflow has been designed to calculate a homologous recombination deficiency (HRD) score using custom QIAseq panels with enrichment of SNPs throughout the genome. The HRD score is based on detected copy number variants (CNVs) and shifts in observed allele frequencies of variants that are expected to be heterozygous.

The workflow is built on **Identify QIAseq DNA Somatic Variants** and only a few changes have been made compared to the original workflow. Therefore, to see a general description of this workflow and information about how to run it, go to the manual page for the Identify QIAseq DNA Variants template workflows (section 12.4). The changes implemented with this workflow include addition of the Detect Regional Ploidy and Calculate HRD Score (beta) tools with relevant inputs and that an input to Control mappings in the Copy Number Variant Detection (Targeted) tool is now mandatory. Note that only variants and target-level CNVs that overlap benchmark regions defined by the Genome in a Bottle Consortium are used as input for the Detect Regional Ploidy tool.

The Detect Regional Ploidy tool takes a target-level annotation track produced by the CNV detection tool as input. Therefore, to run this workflow, control mappings or coverage tables for establishing a CNV baseline must be available. These can be generated with the workflow **Create QIAseq DNA CNV Control Mapping (Illumina)**, see section 12.1.

In order to facilitate easy inspection of variants in 15 selected homologous recombination repair genes an additional variant output listing variants in target regions overlapping these genes is included.

Based on internal testing and validation of a limited number of samples, we have identified 50 as a potential HRD score cutoff for distinguishing between normal samples (below 50) and samples with HRD (equal to or above 50). Note that 50 may not be an optimal cutoff for all protocols, and should only be considered as guidance.

The reference data necessary to run this template workflow is available in the reference data set QIAseq DNA Panels hg19. However, primers and target regions for the relevant QIAseq custom panel must be configured independently. To import the target regions/roi file for a QIAseq

custom panel, use **Import Tracks** (see `Import | Tracks`). To import the QIAseq custom panel primers, use **Import QIAGEN Primers** (see section 5.1).

12.7.1 Output from the Identify HRD Score workflow

All of the original outputs from the underlying workflow have been retained. To view information about those, go to the Identify QIAseq DNA Variants template workflows section (section 12.4).

There are four outputs that are specific to the Identify HRD Score workflow:

- The Detect Regional Ploidy tool outputs a **Target-level_ploidy_track** and a **Regional_ploidy_results_report**. The ploidy track contains the called target-level ploidy states and the report contains graphical overviews of the called ploidy regions and the underlying CNVs and SNP allele frequencies, allowing for easy inspection of the results. See the Detect Regional Ploidy tool for a detailed description (section 10.3).
- The **HRD report** contains the HRD score and the individual LOH, LST and TAI scores. It also details the regions/events called for each type of chromosomal aberation. See the Calculate HRD Score (beta) tool for a detailed description (section 8.4).
- A variant track called **Variants_15_HRR_genes_passing_filters** lists identified variants in target regions of the 15 HRR genes ATM, BARD1, BRCA1, BRCA2, BRIP1, CDK12, CHEK1, CHEK2, FANCA, FANCL, PALB2, RAD51B, RAD51C, RAD51D, RAD54L. The variants are also comprised in the full list of filtered variants. The additional output has only been included to facilitate easy inspection of variants in HRR genes.

12.8 Identify QIAseq DNA Somatic Variants with TMB Score

The Identify QIAseq DNA Somatic Variants with TMB Score (Illumina) or (Ion Torrent) template workflows have been designed to support the DHS-8800Z and DHS-6600Z QIAseq Targeted DNA panels. These panels cover a significantly larger region of the genome than classic Targeted DNA panels, which increases the difficulty of variant calling especially with regards to specificity. Through a series of tools and filters, the Identify TMB Status template workflow has the ability to accurately call variants and to compute a TMB score and score confidence that can be classified as low, intermediate or high.

To run the workflows go to:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA Workflows** () | **Identify QIAseq DNA Somatic Variants with TMB Score (Illumina/Ion Torrent)** ()

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the sequencing reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.

- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details. QIAseq TMB Panels hg38 will be pre-selected and is recommended for running this workflow.
- **Configure batching**. If running the workflow in **Batch** mode, you will be asked to define the batch units. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details.
- **Target regions**. Choose the relevant target regions from the drop down list.
- **Target primers**. Choose the relevant target primers from the drop down list.
- **Mispriming events**. Choose the relevant track for correction of mispriming events from the drop down list.
- **Gene-pseudogene track**. Choose the relevant track for pseudogene and gene family interference correction from the drop down list.
- **Map Reads to Reference**. Here, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)**. Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 12.1. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Identify candidate variants**. The filtering cascade has been tuned using samples of relatively high quality and coverage to provide the best possible sensitivity and precision. Note that selected default values vary by technology (Illumina / Ion Torrent). Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.
- **Remove False Positives**. Set the minimum frequency for detected variants.

- **Add information about Amino Acid Changes**. Specify the genetic code to be used for amino acid translation.
- **Calculate TMB Score**. Specify whether the calculation of a TMB status should be performed and if so, set appropriate thresholds. The Calculate TMB Score tool performs an additional series of filtering including removal of known germlines variants and variants that are likely to be germline based on the observed variant allele frequency.
- **Create Sample Report**. Specify QC items for assessment and subsequent flagging in the sample report generated by the workflow. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted TMB/MSI.

12.8.1 Output from the Identify TMB Status workflow

The Identify TMB Status workflow produces a **TMB report** ( ) that contains the TMB score (calculated as the number of mutations per Mb) and TMB confidence values (figure 12.29). The TMB confidence is based on the size of the target regions included in the TMB score calculation, i.e., those with a coverage at least 100X: TMB confidence is low if fewer than 900,000bp of target regions have sufficient coverage, high if more than 1,000,000 bp of target regions have been included in the calculation, and intermediate in between these 2 values.

1 Summary

TMB status	High
Length of target regions (bp)	1,318,853
Variants inside target regions and after quality filters	442
Germline variants	410
Somatic variants	32
Non-coding somatic variants	2
Synonymous somatic variants	10
Non-synonymous somatic variants	20
Tumor mutational burden (mutations/Mb)	15.16
Tumor mutational burden confidence	High(*)

Figure 12.29: *An example of a TMB report.*

Note that it is possible to configure the Calculate TMB Score tool to include TMB status information in the report (see section 8.1).

In addition, the report lists the adjusted length of the target regions (after removal of the regions whose coverage was below 100X). To estimate how the removal of region of low coverage impacted the original target regions, see Section 1.5 of the "QC for Targeted Sequencing - Coverage Report" output, which offers statistics on the numbers of targets for which all positions are covered by the "Minimum coverage" threshold set in the QC for targeted sequencing dialog (100 by default).

The quality filters statistics of the TMB report recapitulates how many variants are removed by the various filters applied by the tool, and the frequency distributions of input and somatic variants.

Here are two high TMB score reports based on different samples: the first sample (to the left) is a cell line cancer (pure) and the second (to the right) is from tissue (mixed with normal tissue). It can be seen from the distribution of the variants at the different frequencies that the pure sample contains somatic variants at higher frequency compared to the tissue sample where the low frequency variants contribute to the TMB score instead (figure 12.30).

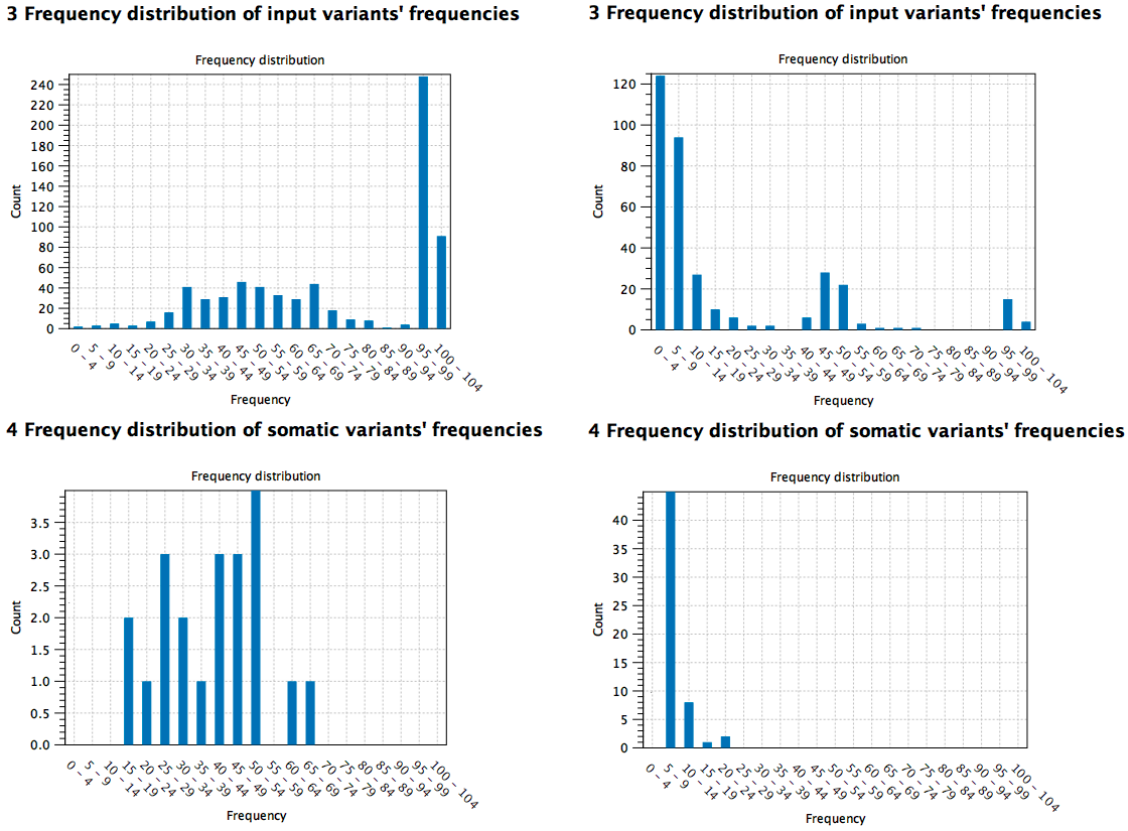


Figure 12.30: *Comparison of high TMB score reports based on different samples (pure cell cancer sample to the left versus mixed tissue sample to the right).*

The workflow will also generate a Genome Browser View ( ) as well as the following files:

- A **Trim Reads report** ( ) where you can check that adapters were detected by the automatic detection option.
- A **UMI Groups report** ( ) containing a breakdown of UMI groups with different numbers of reads, along with percentages of groups and reads (see section 4.2).
- A **Create UMI report** ( ) that indicates how many reads were ignored and the reason why they were not included in a UMI read (see section 4.3).
- A **Structural Variants report** ( ) giving an overview of the different types of structural variants inferred by the Structural Variant analysis.

- A **Sample report** () that contains compiled QC metrics from other reports and provides an overview of a given sample. The sample report also reports whether the QC thresholds specified in the Create Sample Report wizard dialog have been met.
- A **Read Mapping** of the UMI reads ().
- A **Coverage report** () and a **Coverage track** () from the QC for Target Sequencing tool (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).
- Four variant tracks: The **Unfiltered variant** track () allows you to review why a variant that was expected in the output would have been filtered out of the Variants passing filters track. The **Filtered variant** track () includes all variants that remained after the first series of filtering. These filtered variants are input in the Calculate TMB Score 8.1 tool where they underwent further filtering. The **TMB Somatic Variants** track () includes variants that remained after the stringent filtering from the tool. These are the variants that were used to calculate the TMB score. The fourth variant track is the **Indels indirect evidence** track produced by the Structural Variant Caller. This is also available in the Genome Browser View. See http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html for a definition of the variant table content.
- An inversion and a long indels track () containing any inversions and indels longer than 100,000 bp respectively, detected.
- An **amino acid** track () that displays a graphical representation of the amino acid changes. The track is based on the CDS track and in addition to the amino acid sequence of the coding sequence, all amino acids that have been affected by variants are shown as individual amino acids below the amino acid track. Changes causing a frameshift are symbolized with two arrow heads, and variants causing premature stop are marked with an asterisk.
- If a read mapping was submitted in the Copy Number Variant Detection dialog, the workflow also ouptputs three **CNV** tracks () (**Target**-, **Region**- and **Gene**-level) and a **CNV Results** report ().

The difference between the Unfiltered variant track and the Variants passing filters track depends on the following options available in the filtering steps:

- **Filter based on quality criteria**: Average Quality (quality of the sequenced bases that carry the variant), QUAL (significance of the variant), Read Position Test Probability (relative location of the variant in the reads that cover the variant position) and Read Direction Test Probability (relative presence of the variant in the reads from different directions that cover the variant position).
- **Remove homopolymer error type variants**, i.e., errors of the indel type that occur in homopolymer regions. These regions are known to be harder to sequence than non-homopolymeric regions. Note that the definition of homopolymer regions differs between the pipelines due to differences in sequencing technology.
- **Remove false positive based on frequency** The variant's frequency needs to be above this threshold (2.5% for the TMB application) for the variant to be output by the workflow in the filtered variant track.

The difference between the Variants passing filters track and the TMB Somatic Variant track is that even more stringent filters are applied to exclude variants before calculating the TMB score. For example, only variants with a frequency equal to or higher than 5% will be included in the TMB score. Germline variants, synonymous variants and variants outside of coding regions are also excluded.

The read mapping of the merged UMI groups will let you verify the found variants, and examine why expected variants were not found. The UMI Groups Report gives information about the number of UMI groups found, and how many reads are in each. It includes the following information:

- How many reads were aligned to the reference (Reads in input).
- How many reads were mapped in multiple places and thus discarded.
- Groups merged: How many groups were created by merging singleton groups with other groups.
- Number of groups that were discarded for being too small (by default 0 but the option "Minimum group size" of the Calculate Unique Molecular Index Groups can be set up to discard small groups), and how many reads were thus discarded.
- How many groups were created, and of these how many were singletons groups (groups made with sequences sharing identical UMI).
- How many reads are in the largest group.
- How many different UMIs are in the most divergent group (different sequences with different UMIs can be in the same group, if they start on the same position and if they have UMIs that only differ with one character).
- Statistics about the number of reads in the groups.
- Statistics about groups size and reads not included in these groups (also available as graphs below the table).

12.9 Identify QIAseq DNA Ultra Somatic Variants

The Identify QIAseq DNA Ultra Somatic Variants template workflow supports analysis of Illumina QIAseq Ultra panel data. The ultra panels are designed to provide high coverage in targeted regions to allow identification of low frequency variants in cfDNA. As the read structure is different from standard QIAseq panels, the Identify QIAseq DNA Ultra Somatic Variants template workflow should only be used to process data from from QIAseq Ultra panels.

The Identify QIAseq DNA Ultra Somatic Variants workflow is set up to detect very low frequency variants. Please note that to call low frequency variants, coverage must be high. In low coverage samples or regions, very low frequency variants are unlikely to be represented in the reads.

The primers and target regions for the Ultra panels are available in the reference data set QIAseq DNA Ultra Panels hg38.

To run the workflow go to:

Template Workflows | **Biomedical Workflows** () | **QIAseq Panel Analysis** () | **QIAseq DNA Workflows** () | **Identify QIAseq DNA Ultra Somatic Variants (Illumina)** ()

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the sequencing reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details. QIAseq DNA Ultra Panels hg38 will be pre-selected and is recommended for running this workflow.
- **Configure batching**. If running the workflow in **Batch** mode, you will be asked to define the batch units. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details.
- **Target regions**. Choose the relevant target regions from the drop down list.
- **Target primers**. Choose the relevant target primers from the drop down list.
- **Map Reads to Reference**. Here, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.
- **Create UMI Reads from Grouped Reads**. Specify settings for UMI grouping. The QIAseq DNA Ultra data is expected to contain very large UMI groups and more PCR or sequencing errors may consequently be present in the UMIs compared to other sequencing protocols. Therefore, settings for grouping reads into UMI groups should be more relaxed than settings for standard panels. This is reflected in the default settings for Create UMI Reads from Grouped Reads in this workflow. See section 4.3 for details about UMI grouping using the tool Create UMI Reads from Grouped Reads.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)**. Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 12.1. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the

control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

- **Create Sample Report**. Specify QC items for assessment and subsequent flagging in the sample report generated by the workflow. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Variant filtering steps**. A series of dialogs outlines the filtering cascade in this workflow. Note that the dialog names can start with Identify candidate variants, Remove or Define. The filtering cascade has been tuned using samples of relatively high quality and coverage to provide the best possible sensitivity and precision. Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.
- **Remove False Positives**. Set the minimum frequency for detected variants.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA Ultra.

12.9.1 Output from the Identify QIAseq DNA Ultra Somatic Variants template workflow

The Identify QIAseq DNA Ultra Somatic Variants template workflow produces a Genome Browser View () as well as the following files, available in a subfolder as seen in figure 12.31):

- A Trim Reads Report () where you can check that adapters were detected by the automatic detection option.
- A UMI Groups Report () containing a breakdown of UMI groups with different number of reads, along with percentage of groups and reads.
- A Create UMI Report () that indicates how many reads were ignored and the reason why they were not included in a UMI read.
- A Structural Variants Report () giving an overview of the different types of structural variants inferred by the Structural Variant analysis.
- A mapping of the UMI Reads ().

- A coverage report ( ) and a Per-region statistics track ( ) from the QC for Target Sequencing tool (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).
- Two variant tracks ( ): Unfiltered variants contains a list of variants that is output before filtering and Variants passing filters contains filtered variants expected to be high confidence and is also the one used in the Genome Browser View. The difference between the Unfiltered variants track and the Variants passing filters track depends on the settings in the variant filtering steps of the workflow. The Unfiltered variants track is included to allow review of why a variant that was expected in the output would have been filtered out of the Variants passing filters track. For a definition of the variant table content, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html.
- An Amino Acid track ( ).
- If a read mapping was submitted in the Copy Number Variant Detection dialog, the workflow will also output three CNV tracks ( ) (Target-, Region- and Gene-level) and a CNV Results report.
- A Sample Report ( ) that contains compiled QC metrics from other reports and provides an overview of a given sample. The sample report has been set up to report if two QC thresholds have been met. First, a sample must have UMI coverage >3000x at >90 percent of positions in the target regions. This ensures that samples have sufficient coverage for detection of very low frequency variants. Second, the average number of UMIs per read must be >=5. Having many reads per UMI increases the confidence of identified variants. This is particularly relevant when detecting low frequency variants that are only supported by few reads. Both of the QC thresholds can be adjusted in the tool Create Sample Report.

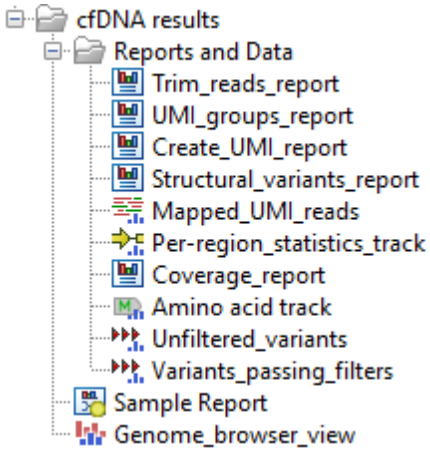


Figure 12.31: *Output from the Identify QIAseq DNA Ultra Somatic Variants template workflow.*

12.9.2 Create QIAseq DNA Ultra CNV Control Mapping

The Create QIAseq DNA Ultra CNV Control Mapping template workflow generates mappings suitable for use as control mappings for the CNV detection step of the Identify QIAseq DNA Ultra

Somatic Variants template workflow. We recommend a minimum of 3 control samples be used for creating control mappings.

The Create QIAseq DNA Ultra CNV Control Mapping template workflow can be found at:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq Analysis Workflows** () | **Create QIAseq DNA Ultra CNV Control Mapping (Illumina)** ()

This workflow includes the same initial processing steps as the Identify QIAseq DNA Ultra Somatic Variants template workflow, which is described in section 12.9. Thus, the initial steps to launch these workflows are similar.

We recommend that the default QIAseq reference set is selected for use, both when generating control mappings and when analyzing cases using the Identify QIAseq DNA Ultra Somatic Variants template workflow.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted DNA Ultra.

12.9.3 Output from the Create QIAseq DNA Ultra CNV Control Mapping template workflow

The Create QIAseq DNA Ultra CNV Control Mapping template workflow produces the following outputs:

- A mapping of the UMI Reads () and a coverage table (), either can be used as the control mapping when running the Identify QIAseq DNA Ultra Somatic Variants template workflow for CNV detection.
- A per-region coverage track (), containing annotation for the coverage within the target region. The track can be used in a track list for visualization of the coverage.
- A coverage report () containing information about coverage within the target regions.
- A trim reads report () containing information about the trimming performed, including automatic adapter read-through trimming.
- A UMI Groups report () containing a breakdown of UMI groups with different number of reads, along with percentages of groups and reads.
- A Create UMI report () containing information about the creation of UMI reads, including the numbers of reads ignored and the reason why these were not included in a UMI read.
- A Structural Variants report () giving an overview of the different types of structural variants inferred by the Structural Variant analysis.
- A sample report () containing summaries of the most important metrics from the above reports.

12.10 Create QIAseq Hybrid Capture CNV Control Mapping (Illumina)

The **Create QIAseq Hybrid Capture CNV Control Mapping (Illumina)** template workflow is designed to support the analysis of data generated with the hybrid capture QIAseq Human Exome and QIAseq xHYB Human kits. The workflow generates control mappings suitable for the CNV detection step of the Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina) (see section 12.12) and Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina) (see section 12.13) template workflows. We recommend a minimum of 3 control samples be used for creating control mappings.

The Create QIAseq Hybrid Capture CNV Control Mapping (Illumina) workflow can be found at:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Create QIAseq Hybrid Capture CNV Control Mapping (Illumina)** ()

This workflow includes the same initial processing steps as the Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina) and Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina) template workflows. Thus, the initial steps to launch these workflows are similar.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Human Exome and xHYB Human.

12.10.1 Output from the Create QIAseq Hybrid Capture CNV Control Mapping (Illumina)

The **Create QIAseq Hybrid Capture CNV Control Mapping (Illumina)** template workflow produces the following outputs:

- A coverage table () and a read mapping () that can be used interchangeably as controls when running the Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina) and Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina) template workflows with CNV detection.
- A target region coverage report () containing information about coverage within the target regions.
- A sample report () containing summaries of trimming results, read length distribution, mapping information, variants etc.

12.11 Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio

The **Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio** template workflow is designed to support the analysis of data generated with hybrid capture QIAseq Human Exome and QIAseq xHYB Human kits. The workflow identifies putative disease causing, inherited variants in a family of three, where there is an affected parent, unaffected parent and a proband.

The first steps of the workflow involve trimming off any remaining PCR adapters. This is followed by mapping the trimmed reads to the human reference sequence. The Structural Variant Caller

then generates a guidance track that is used in the Local Realignment tool to improve the mapping. The improved mapping is then input to the Fixed Ploidy Variant Detection tool. The resulting variants are filtered to remove those located outside defined target regions. Remaining variants are then annotated with information such as the relation to repeat/homopolymer regions or gene elements. Finally, a series of filtering steps removes variants likely to be artifacts.

The putative disease causing variants are identified by creating a list of variants present in both affected individuals and subtracting all variants in the unaffected individual. Additional checks are then carried out, allowing variants to be classified, for example as de novo variants, recessive variants, etc. The variants are output, along with reports and other associated results.

The Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio template workflow can be found at:

> **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA workflows** ( ) | **Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio** ( )

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

Separate dialog steps are presented for providing the sample data for the proband, the affected parent and the unaffected parent. The names of the steps in the left hand side of the wizard indicate the data that should be entered in that step. For example, sequencing reads for the proband would be selected in the step shown in figure 12.32.

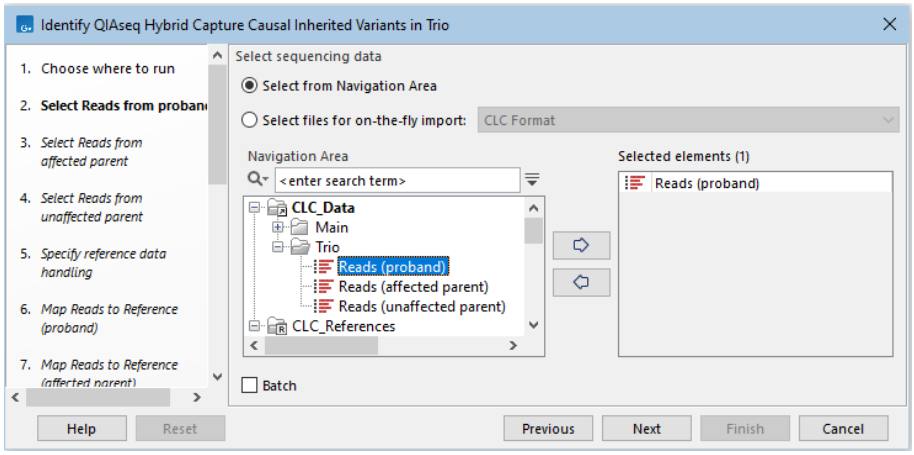


Figure 12.32: *The sequence reads for the proband are specified at the "Select reads from proband" wizard step.*

The following dialog helps you set up the relevant Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button.

Note that if you wish to Cancel or Resume the Download, you can close the template workflow and open the Reference Data Manager where the Cancel, Pause and Resume buttons are available.

If the Reference Data Set was previously downloaded, the option "Use the default reference data" is available and will ensure the relevant data set is used. You can always check the "Select a reference set to use" option to be able to specify another Reference Data Set than the one

suggested.

Both Map Reads to Reference and Fixed Ploidy Variant Detection are configured in separate dialog steps for the proband, the affected parent and the unaffected parent samples. The names of the steps in the left hand side of the wizard, and near the top of each dialog, indicate which sample the parameters apply to. For example, the first wizard steps for Fixed Ploidy Variant Detection, displayed in figure 12.33, has the word "proband" in the title near the top of the dialog.

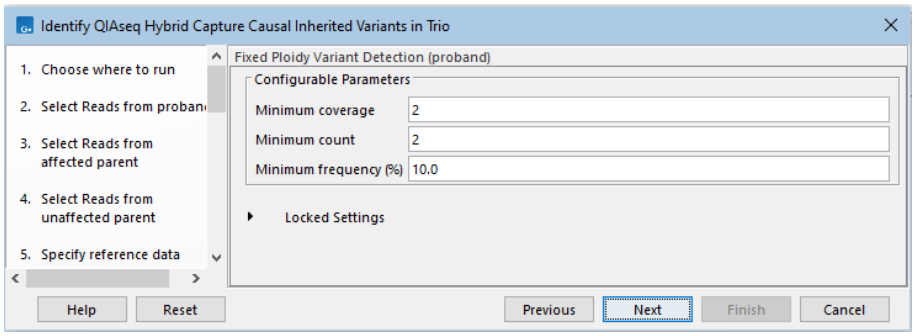


Figure 12.33: *Configuration of Fixed Ploidy Variant Detection tool for the proband sample analysis. This tool is configured separately for the anlaysis of each family member.*

In the Map Reads to Reference dialog, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.

The Fixed Ploidy Variant Detection settings:

- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

In the next two wizard steps, individual filtering settings can be specified for SNVs and Indels for the proband.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

Launching using the QIAseq Panel Analysis Assistant

The workflow is not available in the QIAseq Panel Analysis Assistant, see chapter 11.

12.11.1 Output from the Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio

The **Identify QIAseq Hybrid Capture Causal Inherited Variants in Trio** template workflow produces the following outputs:

- A track list called Genome Browser View ( ).
- A sample report ( ) summarizing the results for the entire trio analysis.
- A read mapping ( ) for each family member.
- A mapping report ( ) for each family member.
- A target region coverage report ( ) and a target region coverage track ( ) for each family member.
- A variant track ( ) for each family member containing annotated and filtered variants.
- A variant track ( ) showing de novo variants in the proband.
- A variant track ( ) showing recessive variants in the proband.
- A gene list ( ) with variants containing the identified putative compound heterozygous variants in the proband.
- A gene list ( ) with variants containing the identified recessive variants in the proband.
- A gene list ( ) with variants containing the identified de novo variants in the proband.
- An amino acid track ( ) showing de novo mutations.
- An amino acid track ( ) showing the recessive variants.

12.12 Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina)

The **Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina)** template workflow is designed to call germline variants from data generated with e.g. QIAseq Multimodal DNA Library kit without UMIs or QIAseq FX DNA Library kit followed by hybrid capture-based target enrichment without addition of mitochondrial spike-in probes, such as QIAseq Exome, QIAseq xHYB Human, or panels from a third party provider. For panels from a third party provider, the same approach as described in section 11.1 is recommended.

If mitochondrial spike-in probes have been added, the Identify QIAseq Hybrid Capture DNA Germline Variants including Mitochondrial (Illumina) should be used (see section 12.12.2).

The first steps of the workflow involve trimming off any remaining PCR adapters. This is followed by mapping the trimmed reads to the human reference sequence. The Structural Variant Caller then generates a guidance track that is used in the Local Realignment tool to improve the mapping. The improved mapping is then input to the Fixed Ploidy Variant Detection tool. The resulting variants are filtered to remove those located outside defined target regions. Remaining variants are then annotated with information such as the relation to repeat/homopolymer regions or gene elements. Finally, a series of filtering steps removes variants likely to be artifacts. The retained variants are output, along with reports and other associated results.

The workflow can be found at:

> **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA workflows** ( ) | **Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina)** ( )

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the sequencing reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details. QIAseq DNA Hybrid Capture and WGS hg38 will be pre-selected and is recommended for running this workflow.
- **Configure batching**. If running the workflow in **Batch** mode, you will be asked to define the batch units. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details.
- **Target regions**. Choose the relevant target regions from the drop down list.
- **Map Reads to Reference**. Here, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.
- **Fixed Ploidy Variant Detection**. Configure the variant detection options.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)**. Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 12.10. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Create Sample Report**. Specify QC items for assessment and subsequent flagging in the sample report generated by the workflow. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Variant filtering steps**. A series of dialogs outlines the filtering cascade in this workflow. Note that the dialog names can start with Identify candidate variants, Remove or Define. The filtering cascade has been tuned using samples of relatively high quality

and coverage to provide the best possible sensitivity and precision. Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.

- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Human Exome and xHYB Human.

12.12.1 Output from the Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina) workflow

The **Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina)** workflow produces the following outputs:

- A **QC graphical report** ( ) summarizing the QC for the reads in the sample analyzed.
- A **Mapping report** ( ) with information related to the mapping of the reads.
- A **Duplicates report** ( ) with details on the results of the removal of duplicate mapped reads.
- A **Structural Variants Report** ( ) giving an overview of the different types of inferred structural variants.
- A **Read mapping** ( ).
- A **Coverage report** ( ) and a **Per region** coverage track ( ) from the QC for Target Sequencing tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html.
- Three variant tracks ( ): Two from the Variant Caller: the **Unfiltered Variants** is output before the filtering steps, the **Annotated variants** contains the variants after filtering and is the one used in the Genome Browser View (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html for a definition of the variant table content). The third is the **Indels indirect evidence** track produced by the Structural Variant Caller. This is also available in the Genome Browser View.
- An **Amino acid** track ( ).

- If a read mapping was submitted in the Copy Number Variant Detection (Targeted) dialog, three CNV tracks ( ) (Target-, Region- and Gene-level) and a CNV results report ( ) will be produced.
- A **Genome Browser View** ( ).
- A **Sample report** ( ) that contains compiled QC metrics from other reports and provides an overview of a given sample. The sample report also reports whether the QC thresholds specified in the Create Sample Report wizard dialog have been met.

12.12.2 Identify QIAseq Hybrid Capture DNA Germline Variants including Mitochondrial (Illumina)

The **Identify QIAseq Hybrid Capture DNA Germline Variants including Mitochondrial (Illumina)** template workflow is designed to call germline variants from data generated with e.g. QIAseq Multimodal DNA Library kit without UMIs or QIAseq FX DNA Library kit followed by hybrid capture-based target enrichment with addition of mitochondrial spike-in probes, such as QIAseq Exome, QIAseq xHYB Human, or panels from a third party provider. For panels from a third party provider, the same approach as described in section 11.1 is recommended.

The workflow calls germline variants in the panel's target regions, as described in section 12.12, and in target regions associated with a mitochondrial spike-in.

It can be found at:

> **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA workflows** ( ) | **Identify QIAseq Hybrid Capture DNA Germline Variants including Mitochondrial (Illumina)** ( )

Mitochondrial variant calling is performed using the Low Frequency Variant Detection tool. The parameters for this can be set in the relevant wizard step when launching the workflow (figure 12.34). For a description of the different parameters that can be adjusted, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html. If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the template workflow. Resulting variants are filtered using a different filtering cascade compared to variants obtained from the Fixed Ploidy Variant Detection tool. The settings of the filtering cascade are locked by default.

Please note that copy number variation analysis is restricted to the panel's target regions and will therefore not be performed for mitochondrial target regions.

The following output files are created separately for mitochondrial data:

- A **Coverage report** ( ) and a **Per region** coverage track ( ) from the QC for Target Sequencing tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html.
- Three variant tracks ( ): Two from the Low Frequency Variant Detection tool: the **Unfiltered Variants** is output before the filtering steps, the **Annotated variants** contains the variants after filtering and is the one used in the Genome Browser View (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html for a definition of the variant table content). The third is the **Indels indirect**

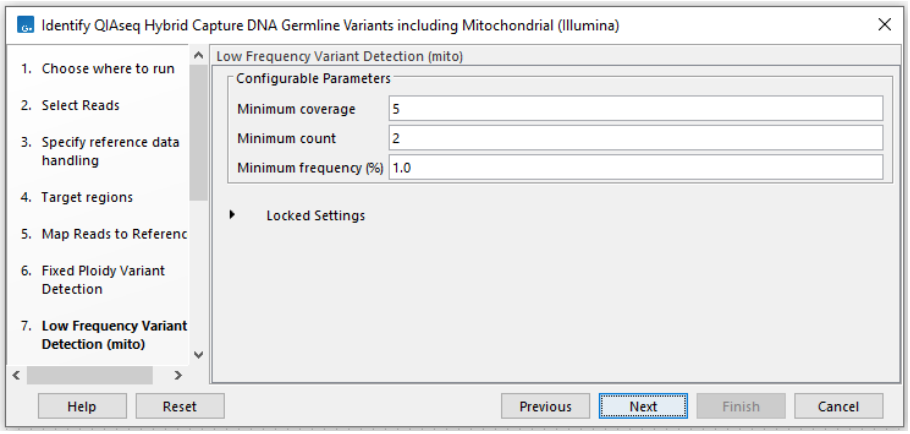


Figure 12.34: *Specify the parameters for the Low Frequency Variant Detection tool.*

evidence track produced by the Structural Variant Caller. This is also available in the Genome Browser View.

- An **Amino acid** track ( ) generated using the mitochondrial genetic code.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Human Exome and xHYB Human.

12.13 Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina)

The **Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina)** template workflow is designed to call somatic variants from data generated with e.g. QIAseq Multimodal DNA Library kit without UMIs or QIAseq FX DNA Library kit followed by hybrid capture-based target enrichment without addition of mitochondrial spike-in probes, such as QIAseq Exome, QIAseq xHYB Human, or panels from a third party provider. For panels from a third party provider, the same approach as described in section 11.1 is recommended.

If mitochondrial spike-in probes have been added, the Identify QIAseq Hybrid Capture DNA Somatic Variants including Mitochondrial (Illumina) should be used (see section 12.13.1).

For data generated using QIAseq Multimodal DNA Library kit with UMIs, see section 12.14.

The first steps of the workflow involve trimming off any remaining PCR adapters. This is followed by mapping the trimmed reads to the human reference sequence. The Structural Variant Caller then generates a guidance track that is used in the Local Realignment tool to improve the mapping. The improved mapping is then input to the Low Frequency Variant Detection tool. The resulting variants are filtered to remove those located outside defined target regions. Remaining variants are then annotated with information such as the relation to repeat/homopolymer regions or gene elements. Finally, a series of filtering steps removes variants likely to be artifacts. The retained variants are output, along with reports and other associated results.

The workflow can be found at:

Template Workflows | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina)** ()

Running the workflow, and the outputs generated, is similar to the Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina) workflow, which is described in detail in section 12.12. However, the Low Frequency Variant Detection tool is used for variant detection instead of the Fixed Ploidy Variant Detection tool and a different filtering step has been introduced as outlined below:

- **Low Frequency Variant Detection**. Configure the variant detection options. For additional details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html.
- **Remove Marginal Variants**. Set the minimum frequency for detected variants.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Human Exome.

12.13.1 Identify QIAseq Hybrid Capture DNA Somatic Variants including Mitochondrial (Illumina)

The **Identify QIAseq Hybrid Capture DNA Somatic Variants including Mitochondrial (Illumina)** template workflow is suitable for data generated with e.g. QIAseq Multimodal DNA Library kit without UMIs or QIAseq FX DNA Library kit followed by hybrid capture-based target enrichment, such as QIAseq Exome, QIAseq xHYB Human, or panels from a third party provider. It calls somatic variants in the panel's target regions, as described in section 12.13, and in target regions associated with a mitochondrial spike-in.

It can be found at:

Template Workflows | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Identify QIAseq Hybrid Capture DNA Somatic Variants including Mitochondrial (Illumina)** ()

For additional details, please see section 12.12.2.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Human Exome.

12.14 Identify QIAseq Multimodal DNA Library Kit Variants

DNA data **without UMIs** produced with the **QIAseq Multimodal DNA/RNA Library Kit**, can be analyzed using the following template workflows, described elsewhere in the manual:

- Identify QIAseq Hybrid Capture DNA Germline Variants (Illumina), see section 12.12.
- Identify QIAseq Hybrid Capture DNA Somatic Variants (Illumina), see section 12.13.

- Identify QIAseq Somatic Variants (WGS) (Illumina) template workflow, see section 12.15.

DNA data **with UMIs** produced with the **QIAseq Multimodal DNA/RNA Library Kit** can be analyzed using the following two template workflows, which are designed to call somatic variants:

- **Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina)** for WGS data.
- **Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina)** for data that has been subjected to hybrid capture-based target enrichment without addition of mitochondrial spike-in probes such as QIAseq Exome, QIAseq xHYB Human, QIAseq xHYB CGP DNA, or panels from a third-party provider.

  The workflow optionally detects CNVs, calculates a TMB score, and detects MSI status.

The workflows include all necessary steps for processing and analyzing the DNA reads:

- Various statistics summarizing and visualizing the input reads are produced using `QC for Sequencing Reads`.
- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1.
- Reads are trimmed using `Trim Reads`.
- Reads are mapped using `Map Reads to Reference`.
- UMI reads are created using **Calculate Unique Molecular Index Groups**, see section 4.2, and **Create UMI Reads from Grouped Reads**, see section 4.3.
- A guidance track is generated from the mapped (UMI) reads using **Structural Variant Caller**, see section 10.7.
- An improved mapping is obtained by realigning the mapped (UMI) reads using the guidance track and `Local Realignment`.
- Variants are called from the improved mapping using `Low Frequency Variant Detection`. For panel data, variant calling is restricted to the relevant target regions.
- The variants are annotated with various information, such as the relation to repeat/homopolymer regions or gene elements, and are subsequently filtered to remove those that are likely to be artifacts through a filtering cascade using `Filter on Custom Criteria`.
- CNVs are optionally detected from the improved mapping using `Copy Number Variant Detection (Targeted)`.
- A TMB score is optionally calculated using **Calculate TMB Score**, see section 8.1.
- MSI status is optionally detected using **Detect MSI Status**, see section 8.2.
- A summary report is created using `Create Sample Report`.

Launching the workflows

To run these workflows, go to

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA Workflows** ()

and select:

> **Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina)** ()
>
> **Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina)** ()

For general information about launching workflows, see Launching workflows individually and in batches.

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**.

  For the Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina) workflow:

  - Select the **QIAseq Multimodal Library Kit and Hybrid Capture hg38** Reference Data Set.

  For the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) workflow:

  - Select the **QIAseq Multimodal Library Kit and Hybrid Capture hg38** Reference Data Set if the data was hybrid captured using the QIAseq Exome or QIAseq xHYB Human panels.
  - Select the **QIAseq DNA xHYB CGP hg38** Reference Data Set if the data was hybrid captured using the QIAseq xHYB CGP DNA panel.

  See chapter 3 for details.
- **Configure batching**. If running the workflow in Batch mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Identify candidate variants**. The filtering cascade has been tuned using samples of relatively high quality and coverage to provide the best possible sensitivity and precision. Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See Filter on Custom Criteria for details on how to adjust the options.

- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Options in the following dialogs can additionally be configured for the Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina) workflow:

- **Map Reads to Reference**. In the Map Reads to Reference dialog, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.

Options in the following dialogs can additionally be configured for the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) workflow:

- **Specify workflow path**. Select whether you want to:
  - **Detect CNVs**. Note that CNV detection requires CNV controls (i.e. normal samples).
    If you select "Yes", you must further specify whether the CNV controls contain targets on X and Y:
    * Select **Yes (X and Y not in controls)** if there are no target regions on X and Y in the CNV controls.
    * Select **Yes (X and Y in controls)** if there are target regions on X and Y in the CNV controls and you are analyzing a sample from the same sex as the CNV controls. This will enable detection of CNVs on the X and Y chromosomes.

    For the QIAseq xHYB CGP DNA panel, a CNV coverage table based on 13 different normal samples without X and Y regions is available in the Reference Data Manager.
  - **Detect MSI.** Note that MSI detection requires an MSI baseline.
    For the QIAseq xHYB CGP DNA panel, an MSI baseline based on 30 normal samples and 176 loci from msisensor2 is available in the Reference Data Manager.
  - **Calculate TMB.** Note that TMB calculation is only recommended for panels covering at least 667 kb [Vega et al., 2021].
- **Target regions**. Choose the relevant target regions. If the data is produced using a custom panel or a panel from a third party provider, see section 11.1.
- **QC for Target Sequencing**. Set the Minimum coverage parameter of the QC for Target Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Calculate TMB Score**. Configure the options for TMB score calculation. See section 8.1 for details.
- **Create Sample Report**. Specify QC items for assessment and subsequent flagging in the sample report generated by the workflow. For additional information, see `Create Sample Report`.

Launching using the QIAseq Panel Analysis Assistant

The two workflows are also available in the QIAseq Panel Analysis Assistant, see chapter 11. The **Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina)** template workflow is available under Multimodal Library Kit, and the **Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina)** template workflow is available under xHYB CGP.

12.14.1 Output from the Identify QIAseq Multimodal DNA Library Kit Variants workflows

The following outputs are generated:

- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the `browser`.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- **Variants passing filters** (): The filtered variants.
- **Gene-level CNV track** (), if running the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) workflow with CNV detection: Genes affected by CNVs. See `Gene-level annotation track` for details.
- **QC & Reports** folder:
  - **Coverage report** (): Summarizes the coverage.

    For the Identify QIAseq Multimodal DNA Library Kit with UMI Somatic Variants (WGS) (Illumina) workflow:

    * Coverage is summarized across the entire genome. See `QC for Read Mapping` for details.

    For the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) workflow:

    * Coverage is summarized within the target regions. See `Coverage summary report` for details.
  - **Mapping report** (): Summarizes the performed read mapping. See `Summary mapping report` for details.
  - **QC report** (): Summarizes and visualizes various statistics of the input reads. See `QC for Sequencing Reads` for details.
  - **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
  - **Remove ligation artifacts report** (): Summarizes ligation artifacts found in and removed from the read mapping. See section 6.4 for details.
  - **Structural variants report** (): Summarizes the number and different types of identified structural variants. See section 10.7.1 for details.

  - **Trim reads report** (): Summarizes the performed read trimming. See `Trim output` for details.
  - **UMI group report** (): Summarizes the identified UMI groups. See section 4.2 for details.
  - **UMI reads report** (): Summarizes the UMI reads. See section 4.3 for details.

  Additionally, if running CNV detection, TMB score calculation, and MSI detection:

  - **CNV results report** (): Summarizes the identified CNVs. See `CNV results report` for details.
  - **MSI report** (): Summarizes the performed MSI detection and contains the MSI status of the sample. See section 8.2.1 for details.
  - **TMB report** (): Contains the TMB score and confidence values. See section 8.1 for details.

- **Tracks** folder:

  - **Indels indirect evidence** (): The filtered indels generated by the Structural Variant Caller. See section 10.7.1 for details.
  - **Mapped UMI reads** (): The UMI reads mapped to the reference genome.
  - **Unfiltered variants** (): The variants identified before filtering.

  Additionally, for the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) workflow:

  - **Amino acid track** (): The amino acid changes introduced by variants in the filtered variant track. See `Amino Acid Changes` for details.
  - **MSI loci track** (), if running MSI detection: MSI loci annotated with predicted stability. See section 8.2.1 for details.
  - **Per-region statistics track** (): Coverage statistics for each target region. See `Per-region statistics` for details.
  - **Region-level CNV track** (), if running CNV detection: Regions affected by CNVs. See `Region-level annotation track` for details.
  - **TMB somatic variants** (), if running TMB score calculation: Filtered variants that are included in the TMB score calculation. See section 8.1 for details.
  - **Target regions without XY** (): Target regions that can be used for CNV detection, where regions on X and Y chromosomes are removed.

12.14.2 Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina)

The **Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina)** template workflow is designed to support the analysis of UMI data generated with the hybrid capture QIAseq Human Exome and QIAseq xHYB Human kits in combination with the QIASeq Multimodal DNA/RNA Library Kit.

The workflow generates controls suitable for the CNV detection step of the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) template workflow, see section 12.14. We recommend a minimum of 3 control samples be used for creating control mappings.

The Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina) workflow can be found at:

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA workflows** () | **Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina)** ()

This workflow includes the same initial processing steps as the Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) template workflow. Thus, the initial steps to launch these workflows are similar.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11, under Human Exome.

Output from the Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina)

The **Create QIAseq Hybrid Capture CNV Control Mapping (with UMI) (Illumina)** template workflow produces the following outputs:

- **Coverage table** () and **Read mapping** (): Can be used interchangeably as controls when running the Identify QIAseq Hybrid Capture DNA Germline Variants (with UMI) (Illumina) and Identify QIAseq Hybrid Capture DNA Somatic Variants (with UMI) (Illumina) template workflows with CNV detection.
- **Coverage report** (): Summarizes coverage statistics in target regions. See `Coverage summary report` for details.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.

12.15 Identify QIAseq Somatic Variants (WGS) (Illumina)

Identify QIAseq Somatic Variants (WGS) (Illumina) is designed to call somatic variants from DNA data produced with the library kits **QIAseq Multimodal DNA/RNA Library Kit**, **QIAseq FX DNA Library Kit** and **QIAseq Ultralow Input Library Kit**.

The workflow includes all necessary steps for processing and analyzing the DNA reads:

- Various statistics summarizing and visualizing the input reads are produced using `QC for Sequencing Reads`.
- Reads are trimmed using `Trim Reads`.
- Reads are mapped using **Map Reads to Reference**, see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Map_Reads_Reference.html`
- A guidance track is generated from the mapped (UMI) reads using **Structural Variant Caller**, see section 10.7

- An improved mapping is obtained by realigning the mapped (UMI) reads using the guidance track and **Local Realignment**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Local_Realignment.html
- Variants are called from the improved mapping using **Low Frequency Variant Detection**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html. For panel data, variant calling is restricted to the relevant target regions.
- The variants are annotated with various information, such as the relation to repeat/homopolymer regions or gene elements, and are subsequently filtered to remove those that are likely to be artifacts through a filtering cascade using **Filter on Custom Criteria**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html
- A summary report is created using `Create Sample Report`.

Launching the workflow

To run the workflow, go to

> **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq DNA Workflows** () | **Identify QIAseq Somatic Variants (WGS) (Illumina)** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options in the following dialogs can be configured:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the DNA reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq Multimodal Library Kit and Hybrid Capture hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Identify candidate variants**. The filtering cascade has been tuned using samples of relatively high quality and coverage to provide the best possible sensitivity and precision. Additional filtering may be needed, or filtering values may need to be adjusted, when working with low quality/coverage samples or when seeking a different balance between sensitivity and precision. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.

- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

12.15.1 Output from the Identify QIAseq Somatic Variants (WGS) (Illumina) template workflow

The following outputs are generated:

- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the `browser`.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- **Variants passing filters**: The filtered variants.
- **QC & Reports** folder:
  - **QC report** (): Summarizes and visualizes various statistics of the input DNA reads. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Sequencing_Reads.html` for details.
  - **Trim reads report** (): Summarizes the performed read trimming. See `Trim output` for details.
  - **Mapping report** (): Summarizes the performed read mapping. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Summary_mapping_report.html` for details.
  - **Remove ligation artifact report** (): Summarizes ligation artifacts found in and removed from the read mapping. See **Remove Ligation Artifacts**, see section 6.4 for details.
  - **Coverage report** (): Summarizes the coverage. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html` for details.
- **Tracks** folder:
  - **Read mapping**: The reads mapped to the reference genome.
  - **Unfiltered variants**: The variants identified before filtering.
  - **Indels indirect evidence**: The filtered indels.

Chapter 13

QIAseq RNA workflows

Contents

13.1 Detect QIAseq RNAscan Fusions

The **Detect QIAseq RNAscan Fusions** workflow can be used for analyzing data produced with the **QIAseq Targeted RNAscan Panels**.

The workflow includes all necessary steps for processing and analyzing the reads:

- Various statistics summarizing and visualizing the input reads are produced using `QC for Sequencing Reads`.
- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- Reads are trimmed using `Trim Reads`.
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4
- Expression levels are quantified using **RNA-Seq Analysis**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_Analysis.html
- Fusions are detected using `Detect and Refine Fusion Genes`.
- A QC report and primer coverage statistics for the reads are created using **QC for RNAscan Panels**, see section 5.3
- A summary report is created using `Create Sample Report`.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** () | **Detect QIAseq RNAscan Fusions** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq RNAscan Panels hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Primers**. Select the primers corresponding to the panel used to generate the reads.
- **Remove and Annotate with Unique Molecular Index**. Specify read structure.
- **Create UMI Reads from Reads**. Specify read structure.
- **Detect and Refine Fusion Genes**. Configure the following options as needed:

  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action
  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see Exclude lists.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.

- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted RNAscan.

13.1.1 Output from the Detect QIAseq RNAscan Fusions workflow

The following outputs are generated:

- **Gene expression** (): A track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.
- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the browser. Two browsers are generated, one for wild type (WT) and one for the fusion chromosomes.
- **PASS fusion genes (WT)** (): The breakpoints on the reference genome of detected fusions that have passed all relevant filters. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See Create Sample Report output for details.
- **QC & Reports** folder:

  - **Create UMI reads report** (): Summarizes the identified UMI groups. See section 4.4 for details.
  - **Fusion report** (): Summarizes the identified fusions. See `Output from Detect and Refine Fusion Genes` for details.
  - **QC for RNAscan Panels report** (): Summarizes various statistics of the mapped reads. Two reports are generated, one for the mapping of all reads ('WT') and one for the mapping of 'Fusion' reads only. See section 5.3 for details.
  - **QC report** (): Summarizes and visualizes various statistics of the input reads. See `QC for Sequencing Reads` for details.
  - **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
  - **Remove ligation artifacts report** (): Summarizes ligation artifacts found in and removed from the read mapping. See section 6.4 for details.
  - **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. See `RNA-Seq report` for details.
  - **Trim adapters**, **Trim homopolymers**, and **Trim on quality** reports (): Summarize the performed trimming. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html` for details. The order of these three trimming steps can be seen in the Sample report.

- **Tracks** folder:
  - **WT** subfolder containing:
    * **Fusion genes (WT)** (): The breakpoints on the reference genome of all detected fusions. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html` for details.
    * **Read mapping** (): The reads mapped to the reference genome. See `RNA-Seq result handling` for details.
    * **Read mapping refined (WT)** (): The reads that mapped best to the reference genome. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html` for details.
    * **Unaligned ends (WT)** (): The unaligned ends mapped to the reference genome. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html` for details.
  - **Fusions** subfolder containing:
    * **Fusion genes (fusions)** (): The breakpoints on the artificial fusion chromosomes of all detected fusions. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html` for details.
    * **Panel primers (fusions)** (): Information about each primer and its read coverage. See section 5.3 for details.
    * **Read mapping (fusions)** (): The reads that mapped best to the artificial fusion chromosomes. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html` for details.

* **Reference sequence (fusions)** ( ) , **Genes (fusions)** ( ), **mRNA (fusions)** ( ), **CDS (fusions)** ( ), and **Primers (fusions)** ( ): The reference sequence, gene regions, mRNA transcripts, CDS regions, and panel primer regions corresponding to the detected fusions on the artificial fusion chromosomes. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

Reviewing the results

The easiest way to review the results is to open the Genome Browser Views, as shown in figure 13.1.

Figure 13.1: *An example of a fusion in a Genome Browser View.*

Double-click on the fusion track name (to the left of the Genome Browser View). The fusion track will open as a table in split view, below the Genome Browser View. Clicking on a fusion event in the table will zoom in to its location in the read mapping, allowing you to review the reads supporting the detected fusion.

Each line in the table corresponds to a fusion breakpoint, such that a fusion event is represented by two lines in the table. The two lines are linked by sharing the same 'Fusion number', which identifies the fused genes, and the same 'Fusion pair', which identifies the event for the gene. For more details on the table, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html.

We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

13.2 Perform QIAseq RNA Fusion XP Analysis

The **Perform QIAseq RNA Fusion XP Analysis (Illumina)** and the **Perform QIAseq RNA Fusion XP Analysis (Ion Torrent)** workflows can be used for analyzing data produced with the **QIAseq RNA Fusion XP panels** (JHS-001Z, JHS-002Z, JHS-003Z, JHS-004Z, JHS-005Z, JHS-3001Z, and JHS3002Z).

The panels allow fusion detection, variant calls, and expression values to be generated from the same RNA sample, and therefore improve on the capabilities of the existing QIAseq RNAscan panels. These extended capabilities are supported by primers which have been annotated with the purpose for which they are intended: 'Variant', 'Fusion', and 'GEX'.

In the Perform QIAseq RNA Fusion XP Analysis workflows, variants are called based on all reads while fusions are detected using reads matching 'Fusion'-annotated primers and gene expressions are calculated based on reads matching 'GEX'-annotated primers.

The workflows include all necessary steps for processing and analyzing the reads:

- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- Reads are trimmed using `Trim Reads`.
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4
- The reads are mapped using **RNA-Seq Analysis**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_Analysis.html
- Various statistics summarizing the mapped reads are produced using **QC for RNAscan Panels**, see section 5.3
- Variants are called from the mapped reads using **Low Frequency Variant Detection**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html
- The variants are annotated with various information, such as the relation to repeat/homopolymer regions or gene elements, and are subsequently filtered to remove those that are likely to be artifacts through a filtering cascade using **Filter on Custom Criteria**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html
- A summary report is created using `Create Sample Report`.
- Specifically for 'Fusion' reads:
  - 'Fusion reads are extracted using **Filter on Custom Criteria**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html
  - The reads are mapped using **RNA-Seq Analysis**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_Analysis.html
  - Various statistics summarizing the mapped reads are produced using **QC for RNAscan Panels**, see section 5.3
  - Fusions are detected using `Detect and Refine Fusion Genes`.
- Specifically for 'GEX' reads:
  - 'GEX' reads are extracted using **Filter on Custom Criteria**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html
  - Expression is quantified using `RNA-Seq Analysis`.

Launching the workflows

To run these workflows, go to

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** ()

and select:

> **Perform QIAseq RNA Fusion XP Analysis (Illumina)** ()
>
> **Perform QIAseq RNA Fusion XP Analysis (Ion Torrent)** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq RNA Fusion XP Panels hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **RNA primers**. Select the primers that were used to produce the data.

  The workflow requires the panel primers to be specified, either by selecting them from the Reference Data Set (QIAseq RNA Fusion XP Panels hg38) for QIAGEN panels or, for custom panels, by importing them as described in section 5.1.
- **Identify candidate variants**. The filtering cascade has been configured to provide the best sensitivity and precision in the output variants. The cascade has been tuned using samples of relatively high quality and coverage. Therefore, additional filtering might be needed, or filtering values adjusted when working with low quality/coverage samples. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.
- **Remove False Positives (filter on allele frequency)**. Optionally adjust the minimum frequency of detected variants.
- **Detect and Refine Fusion Genes**. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes

  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action
  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see Exclude lists.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.

- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflows are also available in the QIAseq Panel Analysis Assistant, see chapter 11 under RNA Fusion XP.

13.2.1 Output from the Perform QIAseq RNA Fusion XP Analysis workflows

The following outputs are generated:

- **Gene expression** (): A track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.
- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the browser.Two browsers are generated, one for wild type (WT) and one for the fusion chromosomes.
- **PASS fusion genes (WT)** (): The breakpoints on the reference genome of detected fusions that have passed all relevant filters. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.
- **RNA variants passing filters** (): The RNA variants that have passed all relevant filters. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html for details.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See Create Sample Report output for details.
- **QC & Reports** folder:

- **Fusion report** (): A graphical report of the fusions found in the sample. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.
- **QC report** (): Summarizes various statistics of the mapped reads. Two reports are generated, one for the mapping of all reads and one for the mapping of 'Fusion' reads only. See section 5.3 for details.
- **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
- **Remove ligation artifacts report** (): Summarizes ligation artifacts found in and removed from the read mapping. Two reports are generated, one for the mapping of all reads and one for the mapping of 'Fusion' reads only. See section 6.4 for details.
- **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. Three reports are generated, one for the RNA-Seq analysis of all reads, one for the RNA-Seq analysis of 'Fusion' reads, and one for the RNA-Seq analysis of 'Expression' reads. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html for details.
- **Trim adapters**, **Trim homopolymers**, and **Trim on quality** reports (): Summarize the performed trimming. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details. The order of these three trimming steps can be seen in the Sample report.
- umiFromReadsReportOutput. See section 4.4 for details.

- **Tracks** folder:
  - **Amino acid track** (): A graphical representation of the amino acid changes.
    The track is based on the CDS track and in addition to the amino acid sequence of the coding sequence, all amino acids that have been affected by variants are shown as individual amino acids below the amino acid track. Changes causing a frameshift are symbolized with two arrow heads, and variants causing premature stop are marked with an asterisk. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html for details.
  - **Read mapping (Variants)** (): Mapping of all reads ('Variant', 'Fusion', and 'GEX') to the reference genome. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_result_handling.html for details.
  - **RNA unfiltered variants** (): The variants identified before filtering. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html for details.

  Furthermore, the output tracks from Detect and Refine Fusion Genes are organized in two subfolders:
  - **WT**:
    - **Fusion genes (WT)** (): The breakpoints on the reference genome of all detected fusions.
    - **Primers (WT)** (): The panel primer regions on the reference genome.
    - **Read mapping refined (WT)** (): The reads that mapped best to the reference genome. See also Read mapping (fusions) below.

    - **Unaligned ends (WT)** (): The unaligned ends mapped to the reference genome.
  - **Fusions**:
    - **Reference sequence (fusions)** () , **Genes (fusions)** (), **mRNA (fusions)** (), **CDS (fusions)** (), and **Primers (fusions)** (): The reference sequence, gene regions, mRNA transcripts, CDS regions, and panel primer regions corresponding to the detected fusions on the artificial fusion chromosomes.
    - **Fusion genes (fusions)** (): The breakpoints on the artificial fusion chromosomes of all detected fusions.
    - **Read mapping (fusions)** (): The reads that mapped best to the artificial fusion chromosomes.

    See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

13.3 Perform QIAseq FastSelect RNA Analysis

The **Perform QIAseq FastSelect RNA Analysis** template workflow can be used for analyzing data produced with the **QIAseq FastSelect RNA Library Kits**.

QIAseq FastSelect RNA Library Kits can be used with the N6-T RT primer (random hexamer), the ODT-T RT primer (oligo-dT), or a combination of both. The ODT-T RT primer uses a TTTV tag to capture RNA fragments. The short polyT tag efficiently captures both exons and polyA tails, making it particularly advantageous for low-input samples or exosomes, where exon usage is of interest. This does not introduce a 3' bias, and approximately 50% of reads may map to exons, depending on the sample. While well-suited for differential expression analysis among samples with similar RNA input levels, it is not designed for RNA abundance analysis, as genes with more polyT regions tend to show higher expression levels.

The workflow includes all necessary steps for processing and analyzing the reads:

- Various statistics summarizing and visualizing the input reads are produced using QC for Sequencing Reads.
- Reads are trimmed using Trim Reads.
- Expression is quantified using RNA-Seq Analysis.
- A summary report is created using Create Sample Report.
- Additionally, if using the N6-T RT primer:
  - N6-T RT primer reads are extracted using **Trim Reads** and a trim adapter list to remove reads containing the ODT-T RT primer, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html
  - Fusions are optionally detected using **Detect and Refine Fusion Genes**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** () | **Perform QIAseq FastSelect RNA Analysis** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select the primers used to produce the data.

  Select whether you want to detect fusions (for the N6-T RT primer only). Skipping fusion detection may speed up workflow execution time.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq UPXome and FastSelect RNA hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Detect and Refine Fusion Genes**, if running fusion detection. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action
  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see `Exclude lists`.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.
- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.

- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under FastSelect RNA.

13.3.1 Output from the Perform QIAseq FastSelect RNA Analysis workflow

The following outputs are generated:

- **Gene expression** (): A track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.

  If using both ODT-T RT and N6-T RT primers, two tracks are generated, one for gene expressions from all reads and one for gene expressions from the N6-T RT primer reads only.

- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the browser.

  If running fusion detection, two browsers are generated, one for wild type and one for the fusion chromosomes.

- **PASS fusion genes (WT)** (), if running fusion detection: The breakpoints on the reference genome of detected fusions that have passed all relevant filters. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

- **Sample report** (): A report containing essential information from all reports produced by the workflow. See Create Sample Report output for details.

- **QC & Reports** folder:

  - **QC report** (): Summarizes and visualizes various statistics of the input reads. See QC for Sequencing Reads for details.
  - **R2 trimming** and **Trimmed reads** reports (): Summarize the performed trimming. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details.

    If using both ODT-T RT and N6-T RT primers, two trimmed reads reports are generated, one for all reads and one for the N6-T RT primer reads only.
  - **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. See RNA-Seq report for details.

    If using both ODT-T RT and N6-T RT primers, two reports are generated, one for statistics from all reads and one for statistics from the N6-T RT primer reads only.

    Additionally, if using the N6-T RT primer:

  - **Extract N6-T RT primer reads report** ( ): Summarizes the trimming performed to extract N6-T RT primer reads. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details.
  - **Fusion report** ( ), if running fusion detection: Summarizes the identified fusions. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

- **Tracks** folder:
  - **Read mapping** ( ): The reads mapped to the reference genome. See RNA-Seq result handling for details.

    If using both ODT-T RT and N6-T RT primers, two mappings are generated, one for the mapping of all reads and one for the mapping of N6-T RT primer reads only.

    If running fusion detection, additional output tracks are organized in two subfolders:
  - **WT**:
    * **Fusion genes (WT)** ( ): The breakpoints on the reference genome of all detected fusions.
    * **Read mapping refined (WT)** ( ): The reads that mapped best to the reference genome during fusion detection. See also read mapping (fusions) below.
    * **Unaligned ends (WT)** ( ): The unaligned ends mapped to the reference genome.
  - **Fusions**:
    * **Fusion genes (fusions)** ( ): The breakpoints on the artificial fusion chromosomes of all detected fusions.
    * **Read mapping (fusions)** ( ): The reads that mapped best to the artificial fusion chromosomes.
    * **Reference sequence (fusions)** ( ) , **Genes (fusions)** ( ), and **mRNA (fusions)** ( ): The reference sequence, gene regions, and mRNA transcripts corresponding to the detected fusions on the artificial fusion chromosomes.

    See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

  We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

13.4 Detect Wells for UPXome

In the following sections we describe the different options available for analyzing samples generated by the **QIAseq UPXome RNA Library Kits**. It is possible to view which plate wells contain reads and compare this to the wells that were intended to be used. Samples can be analyzed in two ways, either by running the workflow through the QIAseq Panel Analysis Assistant, see chapter 11 or by using one of the QIAseq UPXome workflows. Common to both are that they require Barcodes for demultiplexing. These Barcodes have been built into the workflow when running through the QIAseq Panel Analysis Assistant, but can also be found as a reference element in the QIAseq UPXome and FastSelect RNA hg38 data set.

The UPXome RNA application of the QIAseq Panel Analysis Assistant offers a Detect Wells tool which can be run to verify the presence of reads in the expected plate well positions. The preview

shows how reads are distributed in wells and produces a list of user selected wells and their barcodes. The list can be used as an input to the QIAseq UPXome workflows. The Detect Wells tool is only available through the QIAseq Panel Analysis Assistant.

Click **Run** to open the wizard. In the first dialog, select the input reads, and click **Next**.

In the following dialog (figure 13.2)

1. Specify whether a mismatch per barcode should be allowed.
2. Select all the wells used in the experiment in the diagram in the Select wells area (details below).

Selecting the wells used

A diagram of the 96 well plate, is shown in the Select wells area. Wells identified automatically as being used in the experiment are shaded in blue. A well is identified automatically if has at least 0.5% as many reads as the well with the maximum number of reads.

All wells used in the experiment must now be selected. Select wells using functionality associated with the plate diagram, and buttons below it:

- Select a single well by clicking on it in the plate diagram.
- Select individual rows and columns using the checkboxes located to the right and below the diagram.
- Select all wells with the **Select All** button.
- Deselect all selected wells with the **Deselect All** button.
- Select only the wells that were automatically detected with the **Select Detected** button.
- Invert the current selection with the **Invert** button. Using this button, all selected wells are deselected, and vice versa.

Selected wells are indicated by a dark blue circle around the well. When you click on **Next**, it is the barcodes and plate location for these wells that will be output. The generated table can now be used directly in the QIAseq UPXome workflows and will only contain wells with content.

Output from Detect Wells A list of the selected wells and their barcode and location on the plate is output, see figure 13.3. The list can be used when running the Quantify UPXome workflows from the assistant. It can also be exported to a file for running the same template workflow directly, from under the Workflows menu. The list is also compatible with the **Demultiplex Reads** tool that can be run to create the demultiplexed samples without further processing.

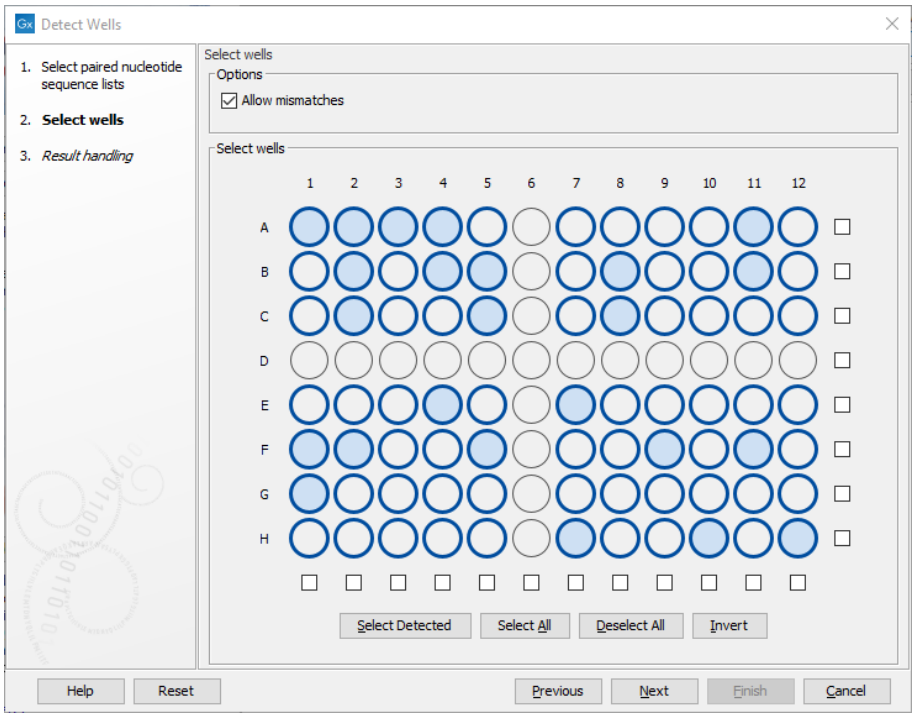
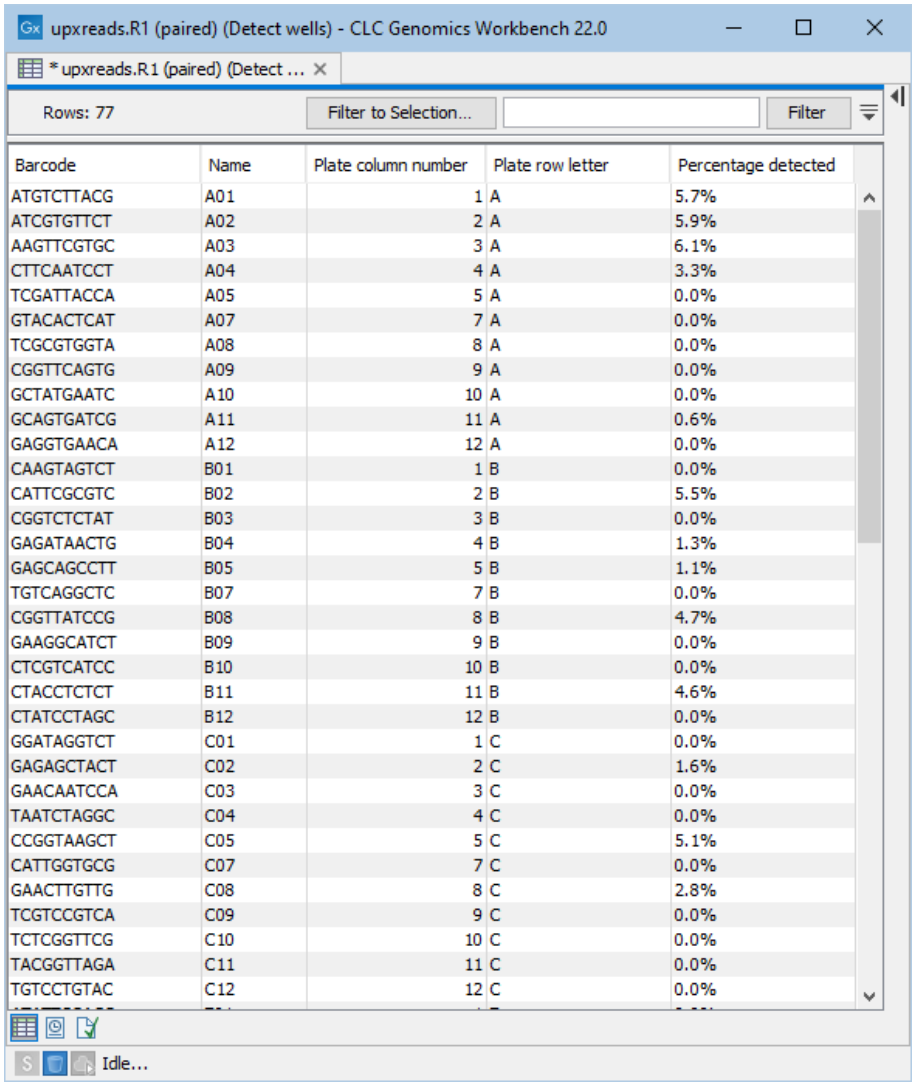


Figure 13.2: *Wells identified automatically as being used in the experiment are shaded in light blue. Here, all wells except row D and column 6 have been selected for use.*



Barcode	Name	Plate column number	Plate row letter	Percentage detected
ATGTCTTACG	A01	1	A	5.7%
ATCGTGTTCT	A02	2	A	5.9%
AAGTTCGTGC	A03	3	A	6.1%
CTTCAATCCT	A04	4	A	3.3%
TCGATTACCA	A05	5	A	0.0%
GTACACTCAT	A07	7	A	0.0%
TCGCGTGGTA	A08	8	A	0.0%
CGGTTCAGTG	A09	9	A	0.0%
GCTATGAATC	A10	10	A	0.0%
GCAGTGATCG	A11	11	A	0.6%
GAGGTGAACA	A12	12	A	0.0%
CAAGTAGTCT	B01	1	B	0.0%
CATTCGCGTC	B02	2	B	5.5%
CGGTCTCTAT	B03	3	B	0.0%
GAGATAACTG	B04	4	B	1.3%
GAGCAGCCTT	B05	5	B	1.1%
TGTCAGGCTC	B07	7	B	0.0%
CGGTTATCCG	B08	8	B	4.7%
GAAGGCATCT	B09	9	B	0.0%
CTCGTCATCC	B10	10	B	0.0%
CTACCTCTCT	B11	11	B	4.6%
CTATCCTAGC	B12	12	B	0.0%
GGATAGGTCT	C01	1	C	0.0%
GAGAGCTACT	C02	2	C	1.6%
GAACAATCCA	C03	3	C	0.0%
TAATCTAGGC	C04	4	C	0.0%
CCGGTAAGCT	C05	5	C	5.1%
CATTGGTGCG	C07	7	C	0.0%
GAACTTGTTG	C08	8	C	2.8%
TCGTCCGTCA	C09	9	C	0.0%
TCTCGGTTCG	C10	10	C	0.0%
TACGGTTAGA	C11	11	C	0.0%
TGTCCTGTAC	C12	12	C	0.0%

Figure 13.3: *Table listing the selected wells to be used when running the Quantify QIAseq UPXome workflows.*

13.5 Demultiplex QIAseq UPXome Reads

The **Demultiplex QIAseq UPXome Reads** workflow can be used to demultiplex reads generated by the **QIAseq UPXome RNA Library Kits**.

The workflow demultiplexes reads into individual samples using **Demultiplex Reads**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Demultiplex_Reads.html for details.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** () | **Demultiplex QIAseq UPXome Reads** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the input reads.
- **Demultiplex Reads**. Specify the barcodes, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Demultiplex_Reads.html for details. This can for example be done by loading a barcode table generated by **Detect Wells for UPXome**, see section 13.4.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under UPXome RNA.

Output from the Demultiplex QIAseq UPXome Reads workflow

The following outputs are generated:

- **Demultiplex reads report** (): Summarizes the number of reads found for each barcode, i.e. for each sample. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Demultiplex_Reads.html for details.
- **Reads** folder:

- **Reads** ( ): A sequence list containing the demultiplexed reads for each sample. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Demultiplex_Reads.html for details.

13.6 Perform QIAseq UPXome RNA Analysis

The **Perform QIAseq UPXome RNA Analysis** template workflow can be used for analyzing data produced with the **QIAseq UPXome RNA Library Kits**.

QIAseq UPXome RNA Library Kits can be used with the N6-T RT primer (random hexamer), the ODT-T RT primer (oligo-dT), or a combination of both. The ODT-T RT primer uses a TTTV tag to capture RNA fragments. The short polyT tag efficiently captures both exons and polyA tails, making it particularly advantageous for low-input samples or exosomes, where exon usage is of interest. This does not introduce a 3' bias, and approximately 50% of reads may map to exons, depending on the sample. While well-suited for differential expression analysis among samples with similar RNA input levels, it is not designed for RNA abundance analysis, as genes with more polyT regions tend to show higher expression levels.

The workflow contains a Demultiplex Reads element, but otherwise resembles the **Perform QIAseq FastSelect RNA Analysis**, see section 13.3 template workflow. This allows for simultaneous analysis of all samples in a multiplexed experiment. The workflow quantifies expression for each individual samples and subsequently performs different analyses across all samples. The ability to run parts of the workflow on a per-sample basis and other parts based on all samples, is possible due to the **Iterate** and **Collect and Distribute** elements, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Control_flow_elements.html

The workflow includes all necessary steps for processing and analyzing the reads:

- For each individual sample:
  - The input reads are demultiplexed into individual samples using **Demultiplex Reads**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Demultiplex_Reads.html
  - Various statistics summarizing and visualizing the input reads are produced using `QC for Sequencing Reads`.
  - Reads are trimmed using `Trim Reads`.
  - Expression is quantified using `RNA-Seq Analysis`.
  - A summary report is created using `Create Sample Report`.
  - Additionally, if using the N6-T RT primer:
    - N6-T RT primer reads are extracted using **Trim Reads** and a trim adapter list to remove reads containing the ODT-T RT primer, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html
    - Fusions are optionally detected using **Detect and Refine Fusion Genes**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html
- Across all samples:

- Principal Component Analysis (PCA) plots are created using **PCA for RNA-Seq**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PCA_RNA_Seq.html
- Expression browsers are created using **Create Expression Browser**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Expression_Browser.html
- A combined report is created using **Combine Reports**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Combine_Reports.html

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** () | **Perform QIAseq UPXome RNA Analysis** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select the primers used to produce the data.

  Select whether you want to detect fusions (for the N6-T RT primer only). Skipping fusion detection may speed up workflow execution time.
- **Select Reads**. Select the input reads.
- **Specify reference data handling**. Select the **QIAseq UPXome and FastSelect RNA hg38** Reference Data Set, see chapter 3 for details.
- **Demultiplex Reads**. Specify the barcodes, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Demultiplex_Reads.html for details. This can for example be done by loading a barcode table generated by **Detect Wells for UPXome**, see section 13.4.
- **Detect and Refine Fusion Genes**, if running fusion detection. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action

  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see Exclude lists.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.

- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under UPXome RNA.

13.6.1 Output from the Perform QIAseq UPXome RNA Analysis workflow

The following outputs are generated:

- **Combined report** (): A report containing essential information from all reports produced by the workflow for all samples. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Combine_Reports_output.html for details.
- **Gene expression browser** (): A browser for inspecting feature expressions, annotations, and statistics for all samples. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_expression_browser.html for details.

  If running fusion detection, two browsers are generated, one for gene expressions from all reads and one for gene expressions from the N6-T RT primer reads only.
- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the browser.

  If running fusion detection, two browsers are generated, one for wild type and one for the fusion chromosomes.
- **PCA plot** (): A Principal Component Analysis (PCA) plot. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PCA_RNA_Seq.html for details.

  If using both ODT-T RT and N6-T RT primers, two plots are generated, one for gene expressions from all reads and one for gene expressions from N6-T RT primer reads only.
- **QC & Reports** folder:

  - **Demultiplex report** (): Summarizes the number of reads found for each barcode, i.e. for each sample. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Demultiplex_Reads.html for details.
  - A subfolder for each sample containing:
    * **QC report** (): Summarizes and visualizes various statistics of the input reads. See QC for Sequencing Reads for details.
    * **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. See RNA-Seq report for details.
      If using both ODT-T RT and N6-T RT primers, two reports are generated, one for statistics from all reads and one for statistics from the N6-T RT primer reads only.
    * **Sample report** (): A report containing essential information from all reports produced by the workflow. See Create Sample Report output for details.
    * **Trim reads report** (): Summarizes the performed read trimming. See Trim output for details.
      If using both ODT-T RT and N6-T RT primers, two trimmed reads reports are generated, one for all reads and one for the N6-T RT primer reads only.

      Additionally, if using the N6-T RT primer:
    * **Extract N6-T RT primer reads report** (): Summarizes the trimming performed to extract N6-T RT primer reads. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details.
    * **Fusion report** (), if running fusion detection: Summarizes the identified fusions. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

- **Reads** folder:
  - **Reads** (): A sequence list containing the demultiplexed reads for each sample. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Demultiplex_Reads.html for details.

- **Tracks** folder:
  - A subfolder for each sample containing:
    * **Gene expression** (): A track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.
      If using both ODT-T RT and N6-T RT primers, two tracks are generated, one for expression values from all reads and one for expression values from the N6-T RT primer reads only.
    * **Read mapping** (): The reads mapped to the reference genome.
      If using both ODT-T RT and N6-T RT primers, two read mappings are generated, one with all reads and one with the N6-T RT primer reads only. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_result_handling.html for details.

      If running fusion detection, additional output tracks are organized in two subfolders:

* **WT**:
  - **Fusion genes (WT)** (): The breakpoints on the reference genome of all detected fusions.
  - **PASS fusion genes (WT)** (), if running fusion detection: The breakpoints on the reference genome of detected fusions that have passed all relevant filters.
  - **Read mapping refined (WT)** (): The reads that mapped best to the reference genome during fusion detection. See also read mapping (fusions) below.
  - **Unaligned ends (WT)** (): The unaligned ends mapped to the reference genome.
* **Fusions**:
  - **Fusion genes (fusions)** (): The breakpoints on the artificial fusion chromosomes of all detected fusions.
  - **Read mapping (fusions)** (): The reads that mapped best to the artificial fusion chromosomes.
  - **Reference sequence (fusions)** () , **Genes (fusions)** (), and **mRNA (fusions)** (): The reference sequence, gene regions, and mRNA transcripts corresponding to the detected fusions on the artificial fusion chromosomes.

  See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html for details.

We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

13.7 Perform QIAseq Multimodal RNA Library Kit Analysis

RNA data produced with the **QIAseq Multimodal DNA/RNA Library Kit** can be analyzed using the following two template workflows:

- **Perform QIAseq Multimodal RNA Library Kit Analysis (Illumina)** for whole transcriptome sequencing data.
- **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** for data that has been subjected to hybrid capture-based target enrichment with the **QIAseq xHYB CGP RNA Panel**.

The workflows include all necessary steps for processing and analyzing the RNA reads:

- Various statistics summarizing and visualizing the input reads are produced using QC for Sequencing Reads.
- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1.
- Reads are trimmed using Trim Reads.
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4.
- Expression is quantified using RNA-Seq Analysis.

- Fusions are optionally detected using `Detect and Refine Fusion Genes`.
- A summary report is created using `Create Sample Report`.

The sensitivity of fusion detection is improved when RNA libraries are prepared using hybrid capture with Nextera adapter blockers.

Launching the workflows

To run these workflows, go to

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** ()

and select:

> **Perform QIAseq Multimodal RNA Library Kit Analysis (Illumina)** ()
>
> **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** ()

For general information about launching workflows, see `Launching workflows individually and in batches`.

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select whether you want to detect fusions. Skipping fusion detection may speed up workflow execution time.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq Multimodal RNA Library Kit and Hybrid Capture hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Detect and Refine Fusion Genes**, if running fusion detection. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - For the **Perform QIAseq Multimodal RNA Library Kit Analysis (Illumina)** workflow:
    - Gene filter action
    - Genes for filtering (tracks)
    - Fusion filter action

    - Fusions for filtering (tables)
  - For the **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** workflow:
    - Fusion filter action
    - Fusion for filtering (tables)
      By default, the gene filtering options are set such that the workflow detects only fusions involving the genes targeted by the **QIAseq xHYB CGP RNA Panel**.

    For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see `Exclude lists`. For general details about fusion detection, see `Detect and Refine Fusion Genes`.

- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The two workflows are also available in the QIAseq Panel Analysis Assistant, see chapter 11. The **Perform QIAseq Multimodal RNA Library Kit Analysis (Illumina)** workflow is available under Multimodal Library Kit, and the **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** workflow is available under xHYB CGP.

13.7.1 Output from the Perform QIAseq Multimodal RNA Library Kit Analysis workflows

The following outputs are generated:

- **Gene expression**: A track with gene expression annotations, including counts and expression values for each gene. See `RNA-Seq result handling` for details.

  The track produced by the **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** workflow contains only the genes targeted by the **QIAseq xHYB CGP RNA Panel**.

- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the `browser`.

  If running fusion detection, two browsers are generated, one for wild type (WT) and one for the artificial fusion chromosomes.

- **PASS fusion genes (WT)** (), if running fusion detection: The breakpoints on the reference genome of detected fusions that have passed all relevant filters. See `Output from Detect and Refine Fusion Genes` for details.

- **Sample report** (): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.

- **QC & Reports** folder:
  - **Fusion report** (): Summarizes the identified fusions. See Output from Detect and Refine Fusion Genes for details.
  - **QC report** (): Summarizes and visualizes various statistics of the input reads. See QC for Sequencing Reads for details.
  - **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
  - **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. See RNA-Seq report for details.
  - **Trim adapters**, **Trim homopolymers**, and **Trim on quality** reports (): Summarize the performed trimming. See Trim output for details. The order of these three trimming steps can be seen in the sample report.
  - **UMI reads report** (): Summarizes the identified UMI groups. See section 4.4 for details.
- **Tracks** folder:
  - **Read mapping** (): The reads mapped to the reference genome. See RNA-Seq result handling for details.
  - **Gene expression unfiltered** (), if running the **Perform QIAseq Hybrid Capture RNA Analysis (Illumina)** workflow: An unfiltered track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.

  If running fusion detection, additional output tracks are organized in two subfolders:
  - **WT**:
    - **Fusion genes (WT)** (): The breakpoints on the reference genome of all detected fusions.
    - **Read mapping refined (WT)** (): The reads that mapped best to the reference genome. See also Read mapping (fusions) below.
    - **Unaligned ends (WT)** (): The unaligned ends mapped to the reference genome.
  - **Fusions**:
    - **Reference sequence (fusions)** () , **Genes (fusions)** (), **mRNA (fusions)** (), and **CDS (fusions)** (): The reference sequence, gene regions, mRNA transcripts, and CDS regions corresponding to the detected fusions on the artificial fusion chromosomes.
    - **Fusion genes (fusions)** (): The breakpoints on the artificial fusion chromosomes of all detected fusions.
    - **Read mapping (fusions)** (): The reads that mapped best to the artificial fusion chromosomes.

    See Output from Detect and Refine Fusion Genes for details.

  We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

13.8 QIAseq miRNA Differential Expression

The workflow calculates differential expressions for expression tables with associated metadata using multi-factorial statistics based on a negative binomial Generalized Linear Model (GLM). Both Grouped on Mature and Grouped on Seed expression tables can be used.

The expression tables are sent to:

- **Create Heat Map for miRNA** The tool creates a two dimensional heat map of expression values. Each column corresponds to one sample, and each row corresponds to a feature (a miRNA or seed sequence). The samples and features are both hierarchically clustered.
- **Differential Expression for miRNA** The tool performs a statistical differential expression test and outputs Statistical Comparison Tables that are used as input for the following tools:
  - **Gene Set Test** Note that this will not yield any results when the workflow is run with Seeds, as no Gene Ontology annotations are known for seeds.
  - **Create Expression Browser** This creates a table where each row includes the expression values of all samples and the contents of all the statistical comparison tables.
  - **Create Venn Diagram for RNA-Seq** The Venn diagram comparison visualizes the overlap between the differentially expressed miRNAs in the selected statistical comparison tables. The miRNA considered to be differentially expressed can be controlled by setting appropriate p-value and fold change thresholds.

To run the workflow, go to:

Template Workflows | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq RNA Workflows** ( ) | **QIAseq miRNA Differential Expression** ( )

Choose the expression data to be analyzed (figure 13.4). The workflow takes Grouped on mature, Grouped on custom database and Grouped on seed expression tables, but note that only samples generated using the same parameters should be analyzed together.

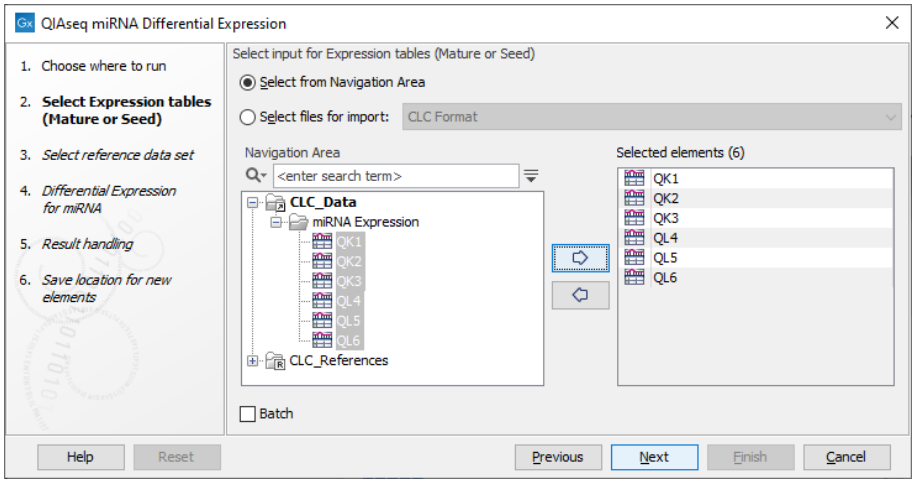


Figure 13.4: *Select the expression data to be compared.*

Then select the QIAseq Small RNA Reference Data Set as in figure 13.5. You can download the Data Set if you have not done so before.

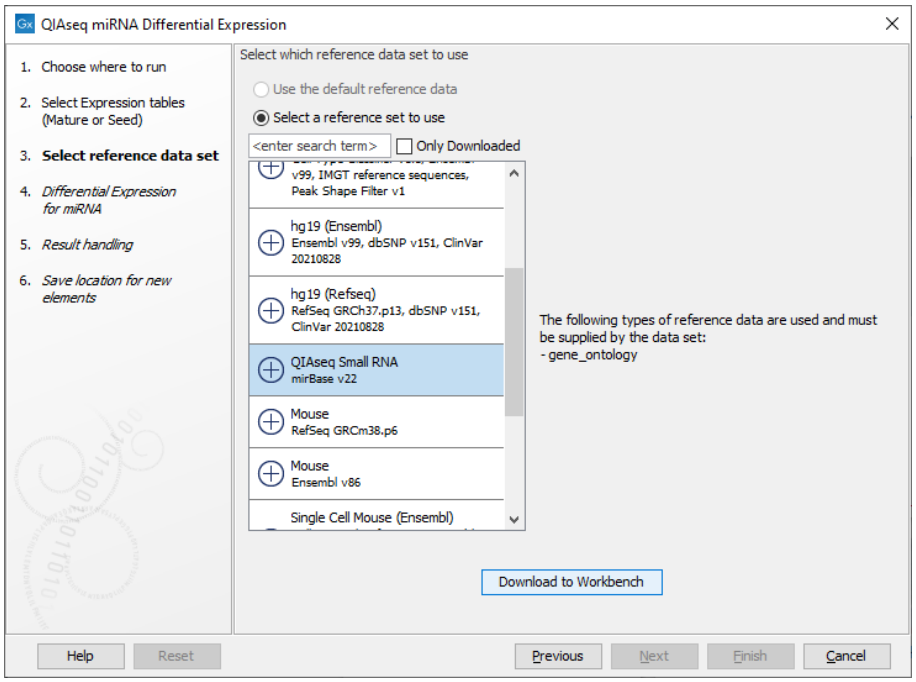


Figure 13.5: *Select and download if necessary the relevant Reference Data Set.*

Following this, the parameters for the QIAseq miRNA Differential Expression need to be specified (figure 13.6).

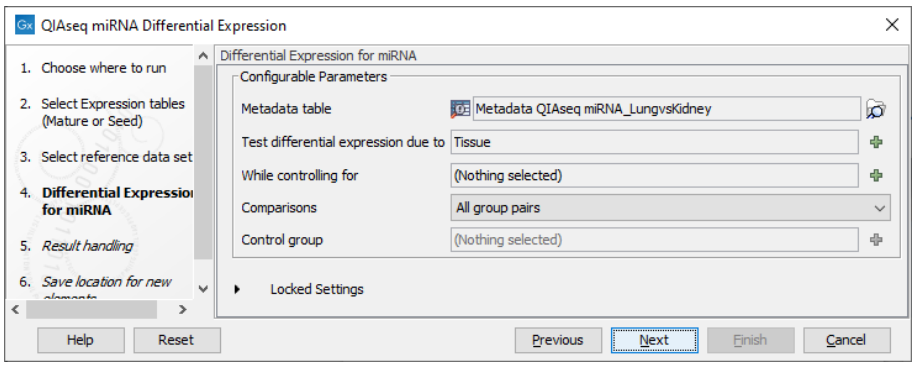


Figure 13.6: *Selecting parameters for QIAseq miRNA Differential Expression.*

Metadata table Select a metadata object that associates the selected input objects to metadata used by the RNA-Seq analysis.

Test differential expression due to Select the factor to be tested for differential expression.

Comparisons Select groups to be compared. It is possible to choose between "Across groups", "All group pairs", and "Against control group".

Control group If "Against control group" was selected in "Comparisons", a control group must be selected.

An example of a metadata table is shown in figure 13.7.

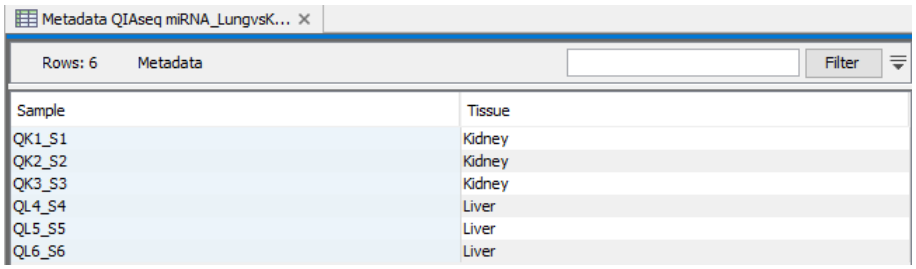


Figure 13.7: *Select the expression data to be compared.*

Metadata is required when defining the experimental design in the Differential Expression for miRNA tool, and can be used to add extra layers of insight in the Create Heat Map for miRNA tool. To learn more about how to create a metadata table, how to import a metadata table, or how to associate data elements with metadata, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Metadata.html.

In the final step, standard result handling is performed: the selected parameters can be previewed, and an output location must be chosen.

The workflow will output the following files:

- a Gene Set Test table, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Gene_Set_Test.html.
- a Venn diagram, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Venn_Diagram_RNA_Seq.html.
- a Statistical comparison table, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_Differential_Expression_tools.html. This table can be uploaded to Ingenuity Pathway Analysis.
- an Expression browser, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Expression_Browser.html.
- a Heat Map, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Heat_Map_RNA_Seq.html.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under miRNA.

13.9 QIAseq miRNA Quantification

The QIAGEN miRNA Quantification workflow quantifies the expression in a sample of the miRNAs found in miRBase. The workflow includes a Trim Reads step, but note that this step only affects Ion Torrent reads as Illumina reads do not have the 5' adapter.

To run the workflow, go to:

Template Workflows | **Biomedical Workflows** | **QIAseq Sample Analysis** | **QIAseq RNA Workflows** | **QIAseq miRNA Quantification**

In the first dialog (figure 13.8), specify the reads to analyze. If the reads come from different samples, remember to check the Batch option. The next dialog will allow you to review the batch units.

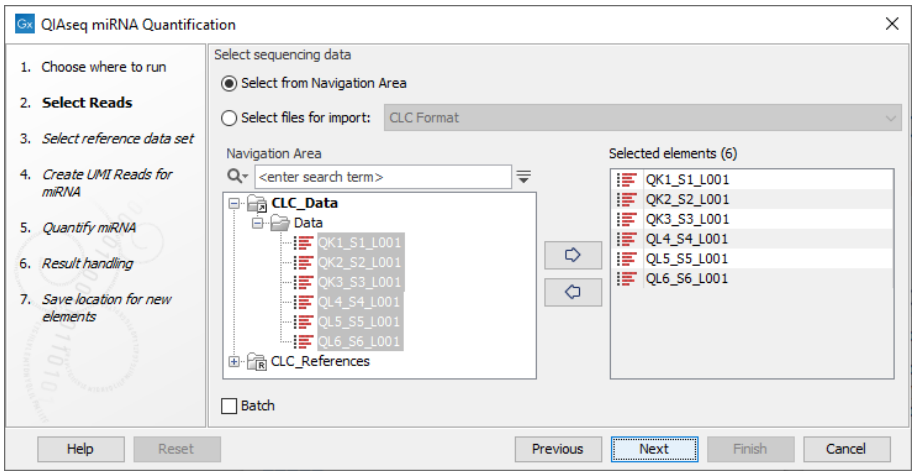


Figure 13.8: *Select the reads.*

In the dialog shown in figure 13.9, specify the Reference Data Set to be used, for example QIAseq Small RNA when using QIAseq data. This set includes GO annotations, a mapping from miRNA identifiers to GO entries, the miRBase database, spike-ins data and a trim adapter list. Note that alternative data sets can be created as explained in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

In the Create UMI Reads for miRNA dialog (figure 13.10), you can set the parameters needed to run the tool (see section 4.5). Note that in this workflow all options have been preconfigured by default to work with Illumina data. How to change the settings when working with Ion Torrent reads is described below.

- **Allow indels in common sequence** This option is unchecked by default, but should be enabled when working with Ion Torrent data.
- **Allow indels in UMI** This option is unchecked by default, but should be enabled when working with Ion Torrent data.
- **Maximum differences in small RNA sequence** Number of allowed differences in the miRNA when merging UMI groups. This is set to 1 difference for Illumina reads, but 2 mismatches should be allowed when working with Ion Torrent data.
- **Allow indels in small RNA sequence** This option is unchecked by default, but should be enabled when working with Ion Torrent data.

In the Quantify miRNA dialog (figure 13.11), specify the parameters for the tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Quantify_miRNA.html).

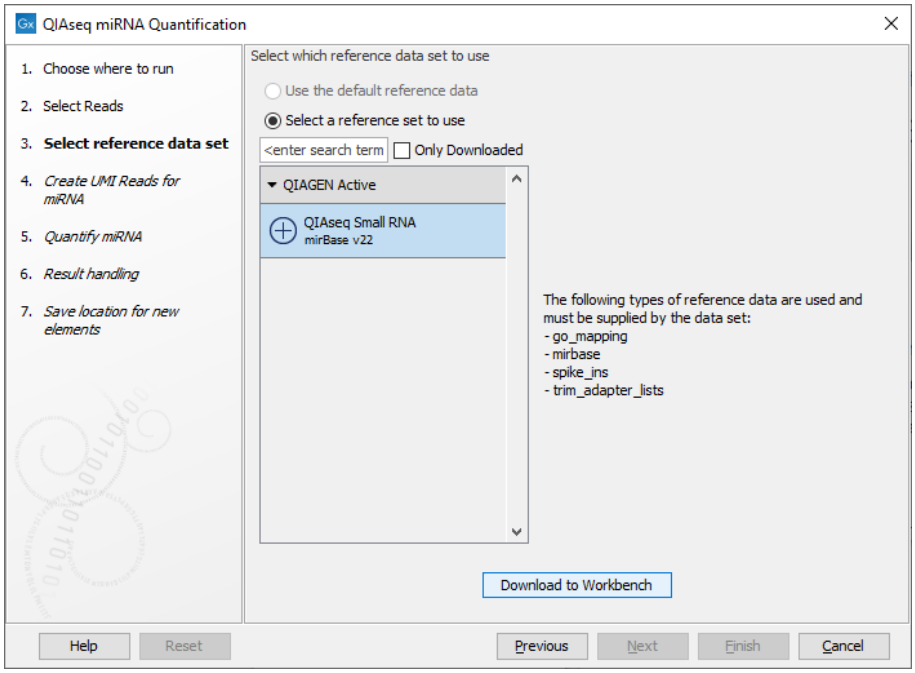


Figure 13.9: *Select the relevant Reference Data Set.*

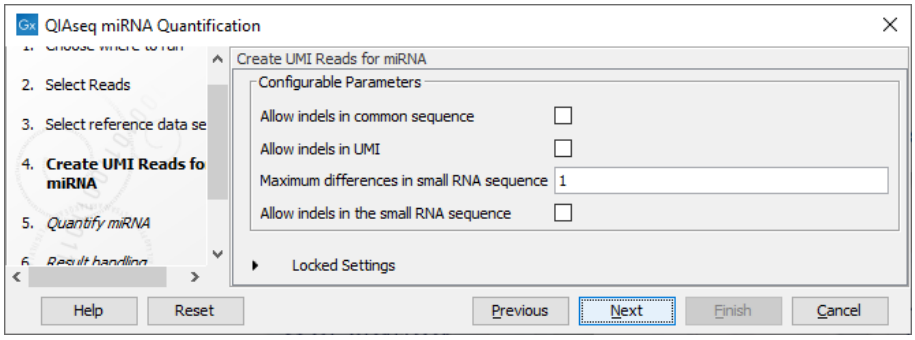


Figure 13.10: *Set up the parameters for the Create UMI Reads for miRNA tool.*

You can choose whether or not to enable the spike-ins analysis and which miRBase database to use. You can also prioritize a list of species used for annotations, with the actual species sequenced always as the first prioritized. The prioritization is important in cases where a read matches two miRNAs equally well, because only the highest priority match is used when linking the miRNA with GO annotations during gene set testing.

Known limitations of prioritization: when two miRNAs have identical sequences, the prioritization determines which is reported. However, if two miRNAs differ in sequence, prioritization will have no effect. For example, in figure 13.12, species are prioritized in the order 'human', 'chimpanzee', 'mouse'. The sample is a human sample. Nevertheless some reads are assigned to the chimpanzee sequence ptr-miR-143 because they map equally well or better to this miRNA than to human hsa-miR-143-3p.

In the last step, choose whether to **Open** or **Save** your results.

Launching using the QIAseq Panel Analysis Assistant

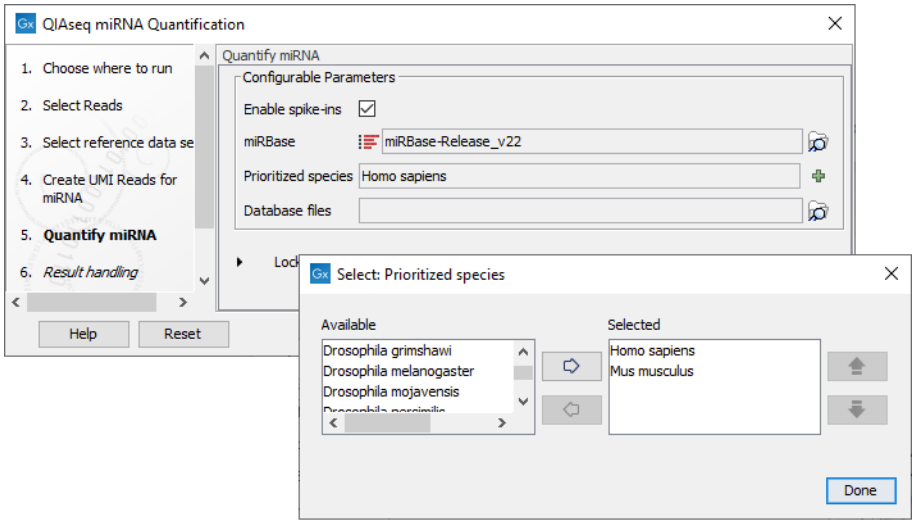
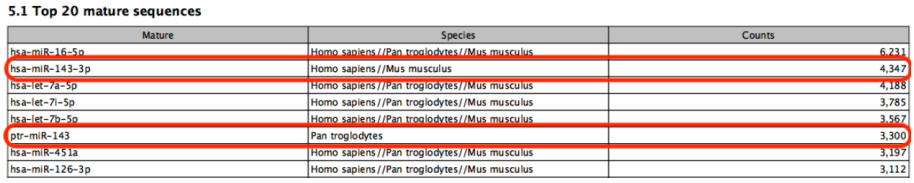


Figure 13.11: *Specify the parameters for the Quantify miRNA tool, including the prioritized list of species that should be used to annotate the miRNA.*



5.1 Top 20 mature sequences

Mature	Species	Counts
hsa-miR-16-5p	Homo sapiens//Pan troglodytes//Mus musculus	6,231
hsa-miR-143-3p	Homo sapiens//Mus musculus	4,347
hsa-let-7a-5p	Homo sapiens//Pan troglodytes//Mus musculus	4,188
hsa-let-7i-5p	Homo sapiens//Pan troglodytes//Mus musculus	3,785
hsa-let-7b-5p	Homo sapiens//Pan troglodytes//Mus musculus	3,567
ptr-miR-143	Pan troglodytes	3,300
hsa-miR-451a	Homo sapiens//Pan troglodytes//Mus musculus	3,197
hsa-miR-126-3p	Homo sapiens//Pan troglodytes//Mus musculus	3,112

Figure 13.12: *A miRNA will not be prioritized above another when they have different sequences.*

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under miRNA.

13.9.1 QIAseq miRNA Quantification outputs

The tool will output expression tables. The "Grouped on mature" table has a row for each mature miRNA. The same mature miRNA may be produced from different precursor miRNAs. The "Grouped on seed" table has a row for each seed sequence. The same seed sequence may be found in different mature miRNAs. If a custom database was provided, a "Grouped on custom database" will be added to the output folder.

Grouped on seed ()

In this expression table, there is a row for each seed sequence (figure 13.13).

This table contains the following information:

- **Name** An example of an expressed mature miRNA that has this seed sequence.
- **Feature ID** The sequence of the miRNA seed
- **Expression value** Counts

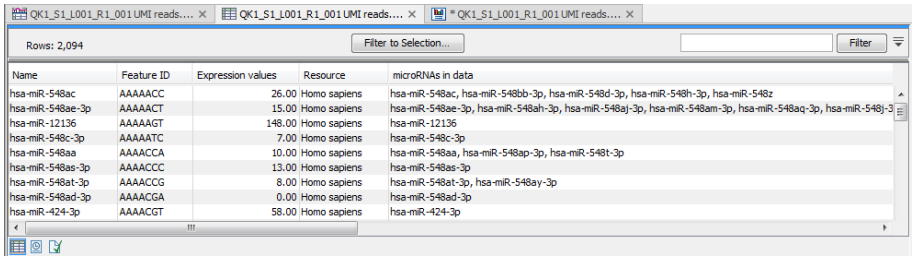


Figure 13.13: *Expression table grouped on seed.*

- **Resource** The database used for identifying miRNAs. For miRBase the species name will be shown.
- **microRNAs in data** A complete list of expressed mature miRNAs with this seed sequence

Grouped on mature ()

In this table, there is a row for each mature miRNA in the database, including those for which the expression is zero (figure 13.14). Double click on a row to open a unique reads alignment (seen at the bottom of figure 13.14). Unique reads result from collapsing identical reads into one. The number of reads that are collapsed into a unique read is indicated in parentheses to the right of the miR name of the unique mature read. The alignment shows all possible unique reads that have aligned to a specific miRNA from the database. Mismatches to the mature reference are highlighted in the alignment and recapitulated in their name as explained in section 13.9.1.

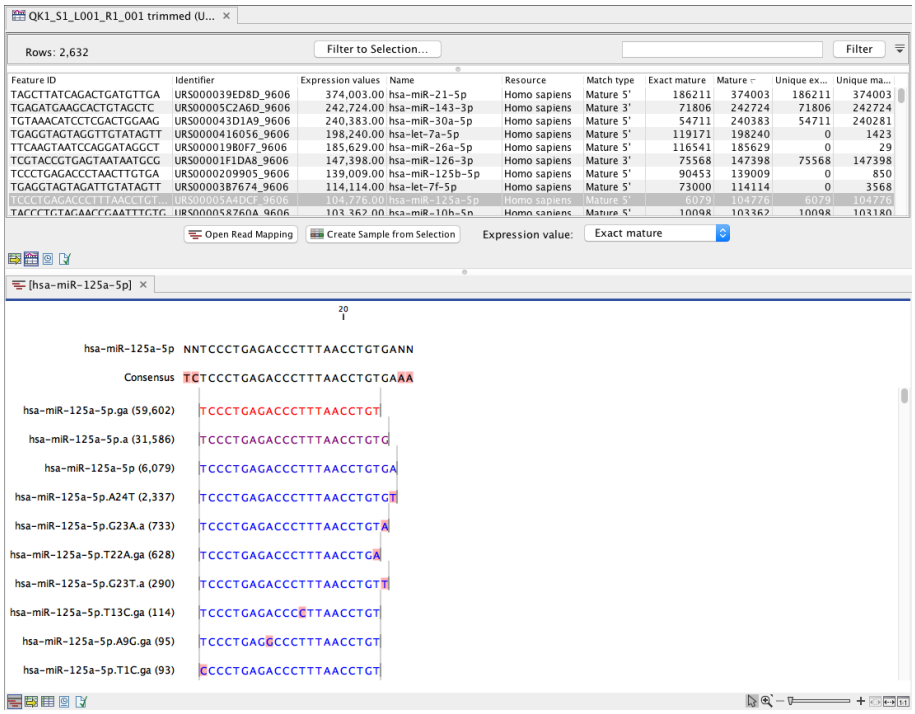


Figure 13.14: *Expression table grouped on mature, with a view of a unique reads alignment.*

This table contains the following information:

- **Feature ID** Sequence of the mature miRNA
- **Identifier** The RNAcentral Accession of the mature miRNA
- **Expression value** Counts in the Mature column
- **Name** Name of the annotation sequence
- **Resource** This is the source of the annotation. For miRBase the species name will be shown.
- **Match type** Mature 5' or Mature 3'
- **Exact mature** Number of mature reads that exactly match the miRBase sequence.
- **Mature** Number of all mature reads, i.e., exact and variants
- **Unique exact mature** In cases where one read has several hits, the counts are distributed evenly across the references. The difference between Exact mature and Unique exact mature is that the latter only includes reads that are unique to this reference.
- **Unique mature** Same as above but for all mature, including variants

Grouped on custom database 

In this table, there is a row for each mature smallRNA in the database, including those for which the expression is zero (figure 13.15). Double click on a row to open a unique reads alignment (seen at the bottom of figure 13.15). Unique reads result from collapsing identical reads into one. The number of reads that are collapsed into a unique read is indicated in parentheses to the right of the miR name of the unique mature read. The alignment shows all possible unique reads that have aligned to a specific miRNA from the database. As with the table *Grouped on mature*, mismatches to the reference are highlighted in the alignment and recapitulated in their name as explained in section 13.9.1.

This table contains the following information:

- **Feature ID** Sequence of the mature miRNA
- **Expression values** Counts in the Mature column
- **Name** Name of the annotation sequence
- **Resource** This is the source of the annotation, usually the name of the custom database input.
- **Exact mature** Number of reads matching to a subsequence of the custom database reference.
- **Mature** Number of reads matching to a subsequence of the custom database reference where mismatches are allowed (within the limit of what was specified in the wizard during configuration)
- **Unique exact mature** In cases where one read has several hits, the counts are distributed evenly across the references. The difference between Exact mature and Unique exact mature is that the latter only includes reads that are unique to this reference.

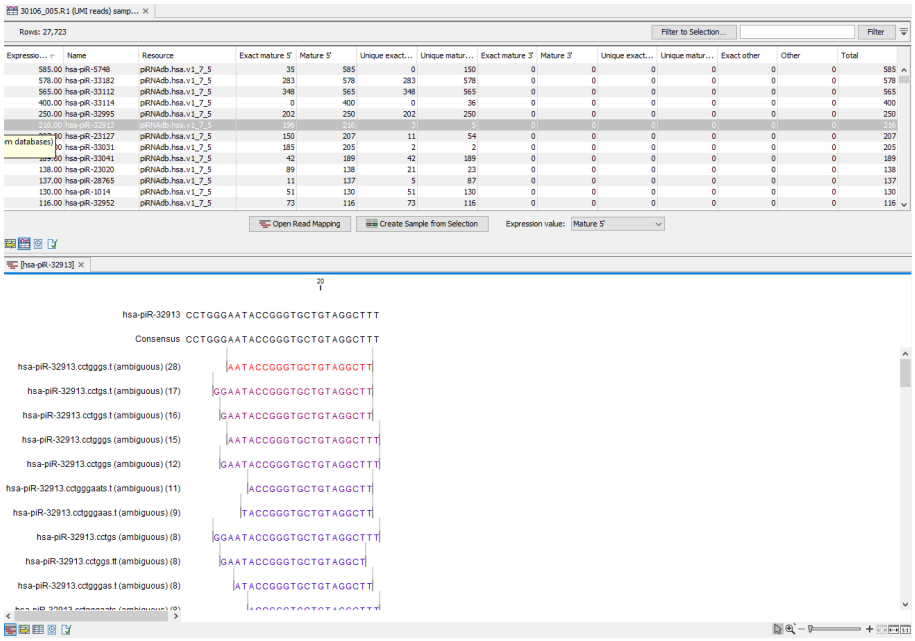


Figure 13.15: *Expression table grouped on custom database, with a view of a unique reads alignment.*

- **Unique mature** Same as above but for all mature, including variants
- **Exact other** Always 0
- **Other** Always 0
- **Total**

Reports and discarded reads The workflow also outputs reports:

- A trimming report, only relevant when working with Ion Torrent reads
- A UMI Report that indicates how many reads were ignored and the reason why they were not included in a UMI read.
- A sequence list containing discarded reads for review.
- A Quantification report

The quantification report contains the following main sections:

- Quantification summary, with information of the number of features that were annotated in the sample.
- Spike-ins, a statistical summary of the reads mapping to the spike-ins (only when spike-ins were enabled).
- Unique search sequences counts, a small RNA reads count distribution.

- Map and Annotate, with Summary, Resources, Unique search sequences, Reads, Read count proportions and Annotations (miRBase).
- Reference sequences, a table with the Top 20 mature sequences, and a table with the Top custom databases sequences when one was provided.
- Seeds report, with tables listing the Top 20 seeds (reference) and Top 20 novel seeds.

It is later possible to combine all reports issued for one sample using the Create Combined miRNA Report tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Combined_miRNA_Report.html).

Naming isomiRs

The names of aligned sequences in mature groups adhere to a naming convention that generates unique names for all isomiRs. This convention is inspired by the discussion available here: https://github.com/miRTop/incubator/blob/master/isomirs/isomir_naming.md

Deletions are in lowercase and there is a suffix s for 5' deletions (figure 13.16):

```
ACGT (ref)
 CGT .as
ACG  .t
  GT .acs
```

Figure 13.16: *Naming of deletions.*

Insertions are in uppercase and there is a suffix s for 5' insertions (figure 13.17):

```
 ACGT  (ref)
 ACGTT .T
 ACGTG .G
AACGT  .As
 ACGTTT .TT
 ACGTGT .GT
```

Figure 13.17: *Naming of insertions.*

Note that indels within miRNAs are not supported.

Mutations (SNVs) are indicated with reference symbol, position and new symbol. Consecutive mutations will not be merged into MNVs. The position is relative to the reference, so preceding (5') indels will not offset it (figure 13.18):

Deletions followed by insertions will be annotated as shown in figure 13.19:

If a sequence maps to multiple miRBase entries or to multiple entries in a custom database, we will add the suffix 'ambiguous' to its name. This can happen when multiple species are selected, as they will often share the same miRBase (or other reference) sequence, or when a read does not map perfectly to any miRBase entry, but is close to two or more entries, distinguished by just one SNV, for example.

```
ACGT (ref)
ATGT .C2T
ATAT .C2T.G3A
AATGT .As.C2T (and not C3T)
  TGT .as.C2T (and not C1T)
```

Figure 13.18: *Naming of mutations.*

```
ACGT (ref)
TCGT .A1T (and not .as.Tes)
ACGA .T4A (and not .t.Ae)
```

Figure 13.19: *Naming of deletions followed by insertions.*

13.10 Quantify QIAseq RNA Expression

The **Quantify QIAseq RNA Expression** template workflow can be found here:

> **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq RNA Workflows** ( ) | **Quantify QIAseq RNA Expression** ( )

Before running the workflow, make sure proper adapter trimming is performed. Run QC for Sequencing Reads to assess the reads and remove any unexpected adapters before running the workflow. Reads with adapters are likely not to map to the reference which can lead to loss of read count.

Double-click on the Quantify QIAseq RNA Expression template workflow to run the analysis.

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

The workflow can and should be run in batch mode, allowing the analysis of several samples at once. Once you have checked the **Batch** option, you can select the **folder** holding the samples that should be analyzed (figure 13.20).

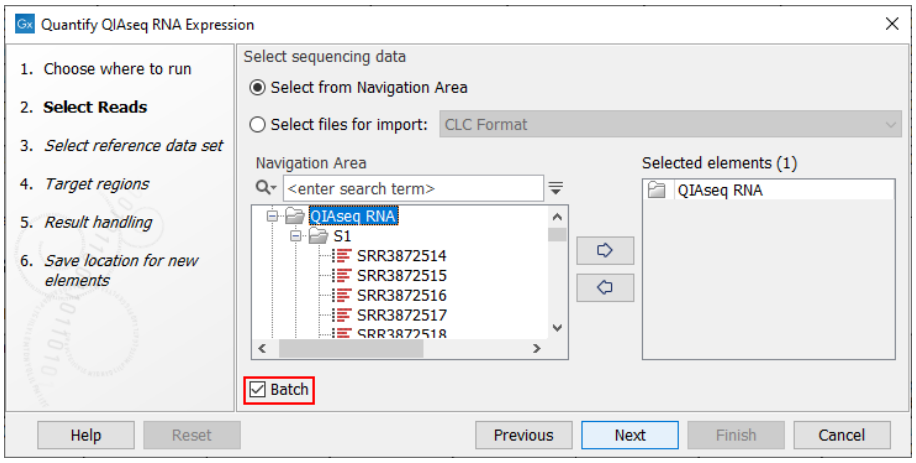


Figure 13.20: *Select the samples to analyze by working in batch mode and choosing the top folder holding all samples.*

When working in batch mode, it is important to select the folder containing the samples, and not

the subfolders containing the sequence lists, nor the reads themselves. In the two latter cases, each sequence list would be considered as an independent sample, when in fact, individual samples are usually made of several sequence lists.

The following dialog helps you set up the relevant Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. Select **QIAseq RNA Panels hg38** and **Download** the set if you have not done so before (figure 13.21).

Note that if you wish to Cancel or Resume the Download, you can close the template workflow and open the Reference Data Manager where the Cancel, Pause and Resume buttons are available.

If the Reference Data Set was previously downloaded, the option "Use the default reference data" is available and will ensure the relevant data set is used. You can always check the "Select a reference set to use" option to be able to specify another Reference Data Set than the one suggested.

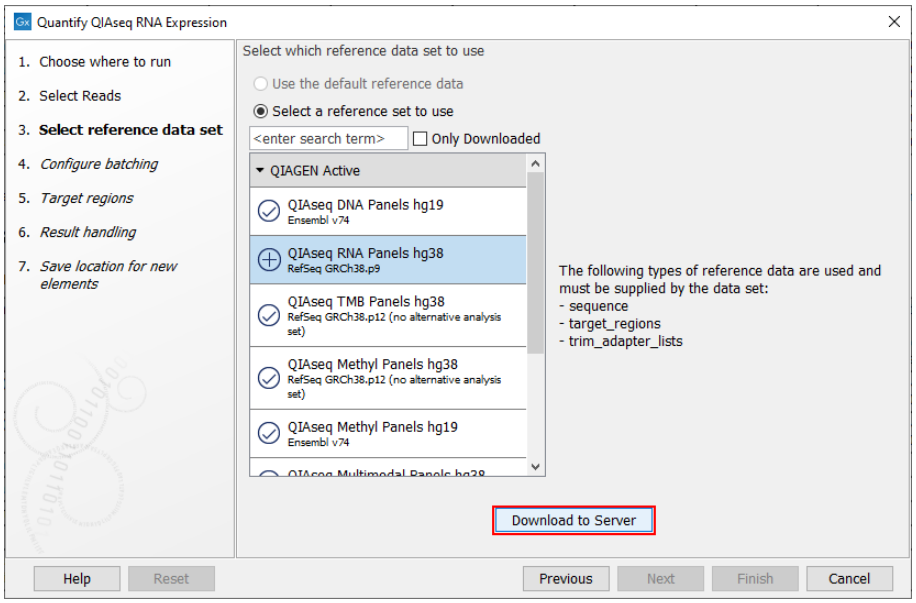


Figure 13.21: *The relevant Reference Data Set is highlighted. In the text to the right, the types of reference needed by the workflow are listed.*

In the next dialog you can can choose to use metadata to configure the batching if you do not want to use the default organization of input.

The batch overview dialog that comes next in the wizard allows you to check that the batch unit is the sample, as opposed to independent sequence lists. You can take advantage of this dialog to exclude some samples from your analysis (figure 13.22).

In the Target regions dialog, specify the Target regions file that correspond to the panel used from the drop down list (figure 13.23).

Finally, in the last wizard step, choose to **Save** the results of the workflow before clicking **Finish**.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under

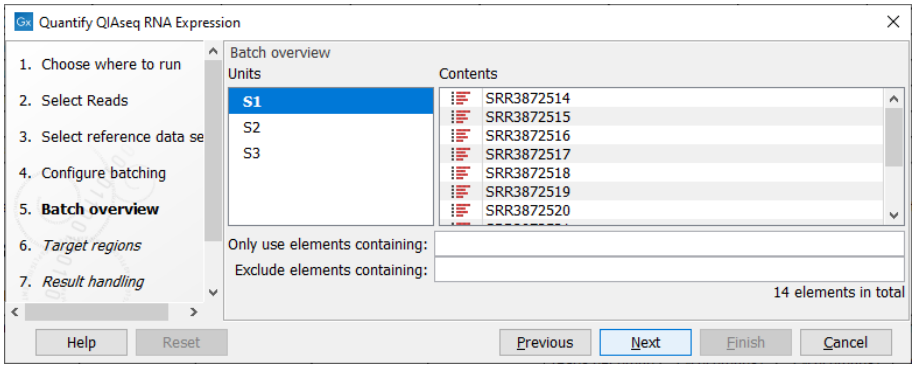


Figure 13.22: *Check in this dialog that the batch unit is the sample and not the reads.*

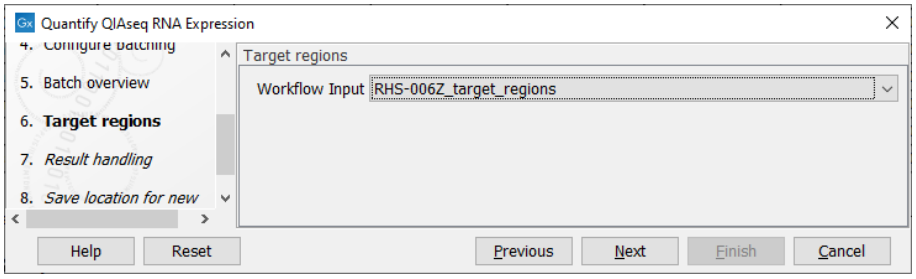


Figure 13.23: *Select the Target regions track that correspond to the panel used.*

Targeted RNA.

13.10.1 Output from the Quantify QIAseq RNA Expression workflow

The **Quantify QIAseq RNA Expression** template workflow produces the following outputs:

- Expression tracks ( ) for each sample.
- **Sample report** ( ): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- QC & Reports folder:
  - Trim Reads report. A report summarizing the performed adapter trimming. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html` for details.
  - Quantify QIAseq RNA report.

Expression tracks are displayed as tables listing genes included in the panel and several measures of their expression levels.

- **Expression value** The number of distinct UMIs seen for this gene.
- **TPM** (Transcripts per Million) The number of transcripts per million that come from this gene. This is computed as the relative abundance per million $(X_i / \sum_j X_j)$.

- **RPKM** (Reads Per Kilobase per Million reads) There is no good definition of RPKM for targeted amplicon data. We therefore define RPKM to be equal to TPM, which preserves the expected property that RPKM is proportional to TPM.
- **Total gene reads** The number of distinct UMIs seen for this gene.
- **Read counts** The total number of reads mapping to this gene. Several reads may have the same UMI.
- **UMI counts** The number of distinct UMIs seen for this gene.
- **Mean reads per UMI** The mean number of reads for each UMI seen for this target.

Expression tracks can be further analyzed using statistical tools from the RNA-Seq Analysis folder of the CLC Workbench, such as PCA for RNA-Seq and Create Heat Map for RNA-Seq.

QC metrics can be found in the sample report. They indicate how many reads were ignored and the reason they were not included in a UMI read. The "Accepted reads" column contains the number of reads that passed the filtering carried out by **Quantify QIAseq RNA**, see section 5.2.

When you are designing a customized panel online, you will see that there is an option to add a set of 6 "GDC controls". These values are your negative controls of genomic DNA. You can also see these in the target regions where they have names like "GDC_CONTROL_06". The number of control targets with more than 10 UMI reads are listed in the "Expressed GDC controls" table cell. This cell will be colored pink and a warning will be shown if any controls are expressed.

Why does the workflow not produce a read mapping? The very short target regions in a QIAseq Targeted RNA Panel are not suited to downstream analyses that require a read mapping, such as variant calling. If a read mapping is desired, for example to investigate suspected off-target effects, we recommend using the Map Reads to Reference tool.

13.11 Demultiplex QIAseq UPX 3' Reads

The UPX 3' RNA application of the QIAseq Panel Analysis Assistant offers a Demultiplex tool to be run before starting the **Quantify QIAseq UPX 3'** workflow. The workflow requires the reads to be sequenced with a QIAseq UPX 3' Transcriptome Kit or a QIAseq UPX 3' Targeted RNA Panel.

Click **Run** to open the wizard. In the first dialog, select the reads you want to demultiplex, and click **Next**.

In the following dialog (figure 13.24):

1. Select the sequencing platform and the plate size. The initial selections have been automatically inferred. Those selections can be adjusted as needed. (Note: For data from NovaSeq, select the MiSeq/HiSeq as the sequencing platform.)
2. Specify whether a mismatch per barcode should be allowed.
3. Select all the wells used in the experiment in the diagram in the Select wells area (details below).

Selecting the wells used

A diagram of the plate, based on the selected well size, is shown in the Select wells area. Wells identified automatically as used in the experiment are shaded in blue.

All wells used in the experiment must now be selected. Select wells using functionality associated with the plate diagram, and buttons below it:

- Select a single well by clicking on it in the plate diagram.
- Select individual rows and columns using the checkboxes located to the right and below the diagram.
- Select all wells with the **Select All** button.
- Deselect all selected wells with the **Deselect All** button.
- Select only the wells that were automatically detected with the **Select Detected** button.
- Invert the current selection with the **Invert** button. Using this button, all selected wells are deselected, and vice versa.

Selected wells are indicated by a dark blue circle around the well. When you click on **Next**, it is the barcodes for these wells that will be used for demultiplexing.

Outputs of demultiplexing

For each sample analyzed, a sequence list is generated *for each* selected well for which reads have been detected. There are also optional outputs: a sequence list containing reads that did not contain any of the expected barcodes, and a report containing information about the number of reads found for each barcode.

Requirements when running demultiplexing in batch mode

When running demultiplexing in batch mode, all samples being analyzed should have been sequenced on the same platform, and the same wells should have been used for each plate. This is because:

- The wells selected when setting up the analysis will be used for each batch unit.
- The same adapter trimming settings are employed for each batch unit.

Automatic selection of plate size and wells

The automatic selection of plate size and the indication of which wells were used is done by sampling the reads used as input. When using a single file as input, the first 10,000 reads are scanned for barcodes, and this information is used. When multiple files are used as input for a sample, a selection of 10,000 reads is also used, but the selection is taken from across the files.

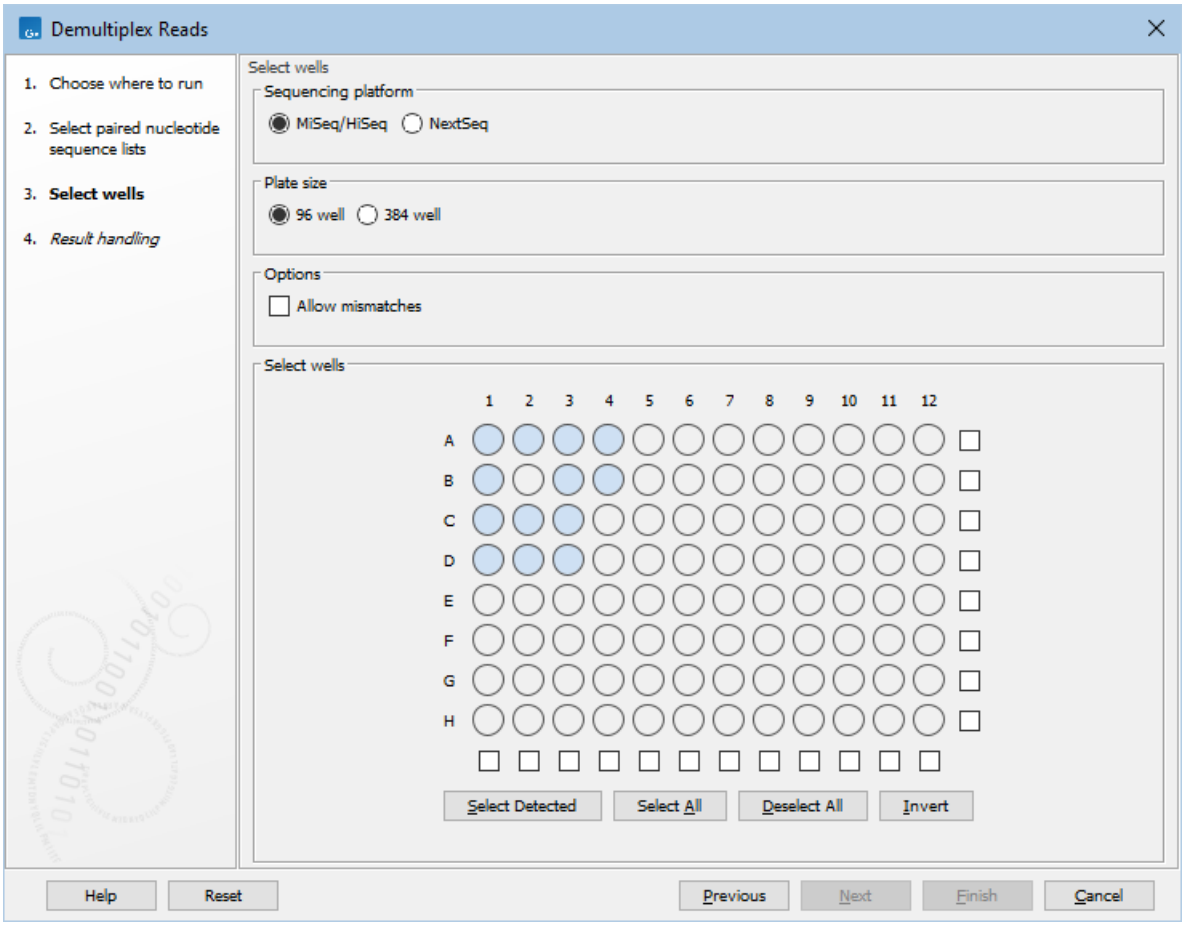


Figure 13.24: *Wells identified automatically as being used in the experiment are shaded in light blue. Here, none of the wells have yet been selected for use.*

13.12 Quantify QIAseq UPX 3'

The **Quantify QIAseq UPX 3'** template workflow can be used for analyzing data produced with a **QIAseq UPX 3' Transcriptome Kit** or a **QIAseq UPX 3' Targeted RNA Panel**.

The workflow includes all necessary steps for processing and analyzing reads that have been demultiplexed with **Demultiplex QIAseq UPX 3' Reads**, see section 13.11:

- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- Reads are trimmed using `Trim Reads`.
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4
- Expression is quantified using `RNA-Seq Analysis`.
- A summary report is created using `Create Sample Report`.

The workflow always reports expression across the whole genome, which allows you to see when reads map off-target.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **QIAseq RNA Workflows** () | **Quantify QIAseq UPX 3'** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq UPX 3' Panels hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **RNA-Seq Analysis**. Specify whether spike-in controls were used when producing the data.

  Be aware that the workflow is configured to expect reads that are strand specific. This is because the reaction design ensures that reads from 3' fragments will appear forward-stranded (read 1 mapping sense to the transcript). It is not unusual for 5-10% of reads to be dropped in the workflow for having an unexpected strandedness.
- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The **Quantify QIAseq UPX 3'** workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under UPX 3' RNA.

13.12.1 Output from the Quantify QIAseq UPX 3' workflow

The following outputs are generated:

- **Gene expression** (): A track with gene expression annotations, including counts and expression values for each gene. See RNA-Seq result handling for details.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See Create Sample Report output for details.
- **QC & Reports** folder:
  - **Merge UMI reads report** (): Summarizes the identified UMI groups. See section 4.4 for details.
  - **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
  - **RNA-Seq report** (): Summarizes various mapping statistics and biotypes distributions. See RNA-Seq report for details.
  - **Trimming first round**, **Trimming second round**, and **Trimming third round** reports (): Summarize the performed trimming. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details.

Chapter 14

Other QIAseq workflows

Contents

14.1 Detect QIAseq Methylation

The **Detect QIAseq Methylation** workflow can be used to call methylation levels on data generated with **QIAseq Targeted Methyl Panels**. Note that the workflow only supports Illumina paired-end sequencing data.

The workflow includes all necessary steps for processing and analyzing the reads:

- Various statistics summarizing and visualizing the input reads are produced using `QC for Sequencing Reads`.
- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- Reads are trimmed using `Trim Reads`.
- Reads are mapped using **Map Bisulfite Reads to Reference**, see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Map_Bisulfite_Reads_Reference.html`

- UMI reads are created using **Calculate Unique Molecular Index Groups**, see section 4.2 and 4.3
- Target coverage statistics are generated using **QC for Targeted Sequencing**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html
- Differential methylation is detected using **Call Methylation Levels**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Call_Methylation_Levels.html
- Sample composition is optionally predicted using **Predict Methylation Profile**, see section 10.8.3
- A summary report is created using `Create Sample Report`.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **Other QIAseq workflows** () | **Detect QIAseq Methylation** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select whether methylation profile prediction should be performed. Only select 'Yes' for data produced with T Cell Infiltration panels such as the 'T Cell Infiltration (MHS-202Z)' panel.
- **Select Reads**. Select the input reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq Methyl Panels hg19** Reference Data Set, see chapter 3 for details.

  Note that the QIAseq Targeted Methyl Panels are designed to use hg19 as reference, but custom panels can be designed against either hg19 or hg38. When analyzing data from a custom panel designed against hg38, a custom Reference Data Set must first be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html and section 5.1 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.

- **Target primers**. Select the primers corresponding to the panel used to generate the reads.

  Specify the primers, either by selecting them from the Reference Data Set (QIAseq Methyl Panels hg19) for QIAGEN panels or, for custom panels, by importing them as described in section 5.1

- **Target regions**. Select the target regions that match the selected primers.
- **Map Bisulfite Reads to Reference**. Specify whether a 'Directional' (reads from both strands) or 'Non-directional' protocol was used to generate the data.
- **QC for Target Sequencing**. Specify the minimum coverage needed on all positions in a target for it to be considered covered. For somatic calling, we recommend setting this to at least 100x.
- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Predict Methylation Profile**. If you have selected to predict methylation profile, you will be asked to specify the minimum coverage a CpG site must have to be used in the prediction.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Targeted Methyl.

14.1.1 Output from the Detect QIAseq Methylation workflow

The following outputs are generated:

- **Genome Browser View** ( ): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the `browser`.
- **Sample report** ( ): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- **QC & Reports** folder:
  - **Initial mapping report** ( ): Summarizes and visualizes various mapping statistics from Map Bisulfite Reads to Reference. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Summary_mapping_report.html` for details.
  - **Initial sequence QC report** ( ): Summarizes and visualizes various statistics of the input reads. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Sequencing_Reads.html` for details.

- **Methylation profile report** (), if running methylation profile prediction: Shows the percentage of the sample estimated to come from epithelial cells (Epi), fibroblast cells (Fib), and immune cells (IC). See section 10.8.3 for details.
- **Methylation report** (): Summarizes various statistics for the methylation level calls. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Call_Methylation_Levels.html for details.
- **Remove and annotate UMI report** (): Summarizes the identified UMIs. See section 4.1 for details.
- **Target coverage report** (): Summarizes various coverage statistics for the target regions. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coverage_summary_report.html for details.
- **Trim reads report** (): Summarizes the performed read trimming. See `Trim output` for details.
- **UMI group creation report** (): Summarizes the UMI reads. See section 4.3 for details.

  Note that data generated with QIAseq Targeted Methyl Panels uses 'NNCNNCNNCNN' UMIs, which results in a higher percentage of the C nucleotide in the 'Nucleotide percentages of the unique molecular barcode symbols' plot.
- **UMI group report** (): Summarizes the identified UMI groups. See section 4.2 for details.

- **Tracks** folder:
  - **Final UMI read mapping** (). Read mapping from Map Bisulfite Reads to Reference after additional UMI collapsing. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Map_Bisulfite_Reads_Reference.html, section 4.2, and section 4.3 for details.
  - **Methylation levels** (). Information about methylated and unmethylated cytosines. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Call_Methylation_Levels.html for details.
  - **Methylation profile** (), if running methylation profile prediction. See section 10.8.3 for details.
  - **Target coverage** (). Information about coverage for each target region. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Per_region_statistics.html for details.

The provided methylation reference database (Homo_sapiens_CpG_hg19_EpiFibIC) in Predict Methylation Profile is constructed from a limited set of epithelial, fibroblast, and immune cell types. It is possible that the epithelial, fibroblast, and immune cell types in any given sample differ from these, and have different methylation profiles. If this is the case, then prediction performance may suffer.

We advise testing the database on pure samples and mixtures with known proportions to see if it is suitable for the analysis.

If it is necessary to create a database, use the tool described in section 10.8.4.

14.2 Perform QIAseq Immune Repertoire Analysis

The **Perform QIAseq Immune Repertoire Analysis** template workflow can be used to characterize the T cell receptor (TCR) immune repertoire for RNA-Seq data produced with the **QIAseq Immune Repertoire RNA Library Kit**.

The workflow includes all necessary steps for processing the RNA-Seq reads and characterizing the repertoire:

- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4
- Overlapping paired UMI reads are merged using **Merge Overlapping Pairs**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Merge_Overlapping_Pairs.html
- Both merged and not merged UMI reads are trimmed using **Trim Reads**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html
- Clonotypes are identified using **Immune Repertoire Analysis**, see section 7.3
- Identified clonotypes are merged using **Merge Immune Repertoire**, see section 7.4
- Merged clonotypes are filtered to remove false positives using **Filter Immune Repertoire**, see section 7.5
- A summary report is created using `Create Sample Report`.

Launching the workflow

The **Perform QIAseq Immune Repertoire Analysis** template workflow is available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **Other QIAseq Workflows** () | **Perform QIAseq Immune Repertoire Analysis** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the RNA-Seq reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.

- **Specify reference data handling**. Select the relevant Reference Data Set. The **Reference Data Manager** (see chapter 3) offers two QIAGEN sets:
  - **QIAseq Immune Repertoire Analysis** for analysis of TCR human data.
  - **QIAseq Immune Repertoire Analysis Mouse** for analysis of TCR mouse data.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Create UMI Reads from Reads**. Adjust **Minimum UMI group size** if needed. UMI groups supported by fewer reads than this number will be discarded. See section 4.4 for details.
- **Filter Immune Repertoire**. Uncheck **Use minimum count** if the identified clonotypes should not be filtered. Otherwise, adjust **Minimum count** if needed. Clonotypes supported by fewer UMI reads than this number will be discarded. See section 7.5 for details.
- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Immune.

14.2.1 Output from the Perform QIAseq Immune Repertoire Analysis workflow

The following outputs are generated:

- **Clonotypes** (): The TCR clonotypes, after merging and filtering, identified in the sample. See section 7.7 for details.
- **Sample report** (): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- **QC & Reports** folder:
  - **Remove and annotate with UMI report** (): summarizes the identified UMIs. See section 4.1 for details.
  - **UMI read report** (): summarizes the identified UMI groups. See section 4.4 for details.
  - **Merge overlaps report** (): summarizes how overlapping paired reads have been merged. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Merge_Overlapping_Pairs.html` for details.

  - **Trim merged** and **not merged reads reports** (): summarize the performed trimming. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html for details.
  - **Immune repertoire analysis report** (): summarizes all detected clonotypes. See section 7.3.1 for details.
  - **Merge immune repertoire report** (): summarizes clonotypes merging. See section 7.4.2 for details.
  - **Filter immune repertoire report** (): summarizes clonotypes filtering. See section 7.5 for details.

- **Supplemental** folder:
  - **Trimmed UMI merged** and **not merged reads** (): The processed reads that were used for the clonotype identification.
  - **Raw clonotypes** (): TCR clonotypes, before merging and filtering, identified in the sample. See section 7.7 for details.

14.3 Perform QIAseq Targeted TCR Analysis

The **Perform QIAseq Targeted TCR Analysis** workflow can be used to characterize the T cell receptor (TCR) immune repertoire for RNA-Seq data produced with the **QIAseq Targeted RNA-seq Panel for T-cell Receptor**.

The workflow includes all necessary steps for processing the RNA-Seq reads and characterizing the repertoire:

- UMIs are removed using **Remove and Annotate with Unique Molecular Index**, see section 4.1
- Common sequence is removed using **Trim Reads**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html
- UMI reads are created using **Create UMI Reads from Reads**, see section 4.4
- Overlapping paired UMI reads are merged using **Merge Overlapping Pairs**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Merge_Overlapping_Pairs.html
- Both merged and not merged UMI reads are trimmed using **Trim Reads**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html
- Clonotypes are identified using **Immune Repertoire Analysis**, see section 7.3
- Identified clonotypes are merged using **Merge Immune Repertoire**, see section 7.4
- Merged clonotypes are filtered to remove false positives using **Filter Immune Repertoire**, see section 7.5
- A summary report is created using `Create Sample Report`.

Launching the workflow

The **Perform QIAseq Targeted TCR Analysis** template workflow is available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **Other QIAseq Workflows** () | **Perform QIAseq Targeted TCR Analysis** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Select Reads**. Select the RNA-Seq reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the relevant Reference Data Set. The **Reference Data Manager** (see chapter 3) offers two QIAGEN sets:
  - **QIAseq Immune Repertoire Analysis** for analysis of TCR human data.
  - **QIAseq Immune Repertoire Analysis Mouse** for analysis of TCR mouse data.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **Create UMI Reads from Reads**. Adjust **Minimum UMI group size** if needed. UMI groups supported by fewer reads than this number will be discarded. See section 4.4 for details.
- **Immune Repertoire Analysis**. Set **Restrict to chains** if only specific chains have been sequenced. See section 7.3 for more details.
- **Filter Immune Repertoire**. Uncheck **Use minimum count** if the identified clonotypes should not be filtered. Otherwise, adjust **Minimum count** if needed. Clonotypes supported by fewer UMI reads than this number will be discarded. See section 7.5 for details.
- **Create Sample Report**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Immune.

When launching from the assistant, the analysis can be restricted to the sequenced chains (figure 14.1).

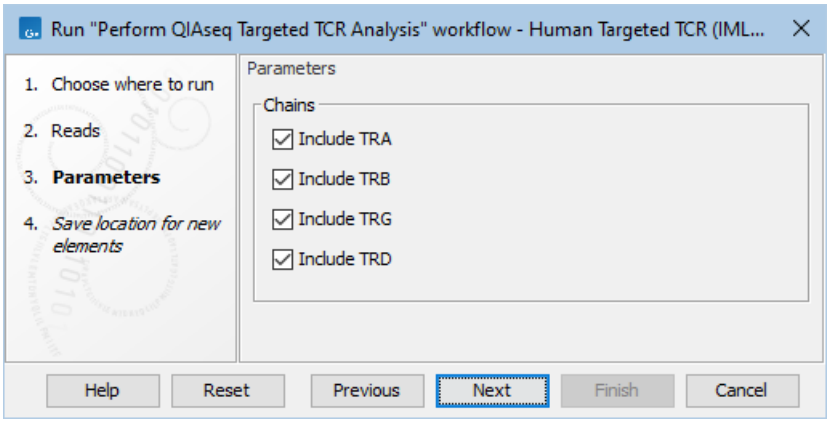


Figure 14.1: *Selecting the chains to be used for the analysis when executing the workflow from the QIAseq Panel Analysis Assistant.*

14.3.1 Output from the Perform QIAseq Targeted TCR Analysis workflow

The following outputs are generated:

- **Clonotypes** ( ): The TCR clonotypes, after merging and filtering, identified in the sample. See section 7.7 for details.
- **Sample report** ( ): A report containing essential information from all reports produced by the workflow. See `Create Sample Report output` for details.
- **QC & Reports** folder:
  - **Remove and annotate with UMI report** ( ): summarizes the identified UMIs. See section 4.1 for details.
  - **Trim common sequence report** ( ): summarizes trimming of the common sequence. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html` for details.
  - **UMI read report** ( ): summarizes the identified UMI groups. See section 4.4 for details.
  - **Merge overlaps report** ( ): summarizes how overlapping paired reads have been merged. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Merge_Overlapping_Pairs.html` for details.
  - **Trim merged** and **not merged reads reports** ( ): summarize the performed trimming. See `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html` for details.

  - **Immune repertoire analysis report** (): summarizes all detected clonotypes. See section 7.3.1 for details.
  - **Merge immune repertoire report** (): summarizes merging of clonotypes. See section 7.4.2 for details.
  - **Filter immune repertoire report** (): summarizes clonotypes filtering. See section 7.5 for details.

- **Supplemental** folder:
  - **Trimmed UMI merged** and **not merged reads** (): The processed reads that were used for the clonotype identification.
  - **Raw clonotypes** (): TCR clonotypes, before merging and filtering, identified in the sample. See section 7.7 for details.

14.4 Perform QIAseq Multimodal Panel Analysis

The **Perform QIAseq Multimodal Panel Analysis (Illumina)** template workflow can be used for analyzing DNA and/or RNA reads generated using QIAseq Multimodal Panels.

Note that the QIAseq Multimodal Panels are designed against genome build hg19 for the DNA panel and hg38 for the RNA panel. BED files are provided in the respective genome build. However, the template workflow requires that reference data for both DNA and RNA is for the same genome build. The two QIAseq Multimodal Reference Data Sets provided by the Reference Data Manager are for genome build hg38, where the reference data for the DNA panel has been converted to hg38 as described below.

For custom panels, the DNA panel BED file needs to be imported against hg19, after which it should be converted to hg38 using the tool **Convert Annotation Track Coordinates**, see section 6.6. If many regions are lost during conversion, it can cause reads to be discarded that would have otherwise mapped to the lost target regions. To avoid such issues, a copy of the template workflow can be used, containing only the analysis of the DNA reads, and the workflow should be run using the imported BED file against hg19.

The workflow is built by combining variant calling from the Identify QIAseq DNA Somatic Variants (Illumina) workflow, see section 12.4, and fusion detection from the Perform QIAseq RNAscan Fusion XP workflow, see section 13.2, with some minor adjustments. Specifically, two tools to further annotate variants have been added:

- Annotate RNA Variants, see section 10.2
- Annotate with Repeat and Homopolymer Information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Annotate_with_Repeat_Homopolymer_Information.html

The annotations added by these tools are used to filter away variant calls that most likely originate from RNA contamination and variants appearing within repeat or homopolymer regions.

The workflow can be run with the Reference Data Set **QIAseq Multimodal Panels hg38**. This set contains Catalog Panel Primers and Target Regions that have been lifted to the hg38 reference sequence. You can either download the reference data set before starting the analysis or download the default data set during execution of the workflow.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **Other QIAseq workflows** () | **Perform QIAseq Multimodal Panel Analysis (Illumina)** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select whether you want to analyze DNA and/or RNA reads.
- **Select DNA/RNA Reads**. Select the reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq Multimodal Panels hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **DNA/RNA primers**. Select the primers that were used to produce the data.
- **Target regions**. Select the target regions that match the selected primers.
- **Map Reads to Reference**. Configure masking.

  By default, the GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set masking track is selected, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1. Changing the masking mode from 'No masking' to 'Exclude annotated' excludes these regions.
- **QC for Target Sequencing**. Specify the minimum coverage needed on all positions in a target for it to be considered covered. For somatic calling, we recommend setting this to at least 100x
- **Copy Number Variant Detection (Targeted)**. Specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify

a control mapping, or if the target regions files contains fewer than 50 regions, the Copy Number Variation analysis will not be carried out.

- **Create Sample Report (DNA/RNA)**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Identify candidate variants**. The filtering cascade has been configured to provide the best sensitivity and precision in the output variants. The cascade has been tuned using samples of relatively high quality and coverage. Therefore, additional filtering might be needed, or filtering values adjusted when working with low quality/coverage samples. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details on how to adjust the options.
- **Remove False Positives (Filter on allele frequency)**. Specify the minimum frequency of detected variants.
- **Add Information about Amino Acid Changes**. Leave the genetic code set to 1 Standard.
- **Detect and Refine Fusion Genes**. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action
  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see Exclude lists.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Multimodal Panels.

14.4.1 Output from the Perform QIAseq Multimodal Panel Analysis (Illumina)

The Perform QIAseq Multimodal Panel Analysis workflow produces a large number of files organized into a number of subfolders as well as single elements. All the files described here are generated when both DNA and RNA reads are analyzed. If only DNA reads are analyzed, RNA-specific output files will not be produced, and vice versa.

The root folder contains four subfolders (QC & Reports, Tracks (WT), Tracks (Fusions) and VCF exportable tracks) in addition to the following output elements:

- A Workflow Result Metadata table keeping track of all generated output.
- A Gene Expression Track () with gene expression counts.
- A DNA Sample Report () summarizing important QC values for the DNA run.
- An RNA Sample Report () summarizing important QC values for the RNA run.
- A Fusion Report () (described in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html) with graphical representations of the fusions found. Double-clicking on a fusion plot in the report will open the plot in a view that allows it to be exported as a high resolution image.
- A Genome Browser View (WT) () containing the WT part of the analysis including fusions, DNA and RNA read mappings and variant callings.
- A Genome Browser View (Fusions) () containing only the fusion chromosomes, which are used for refining the fusions. Tracks for this view can be found in the Tracks (Fusion) folder.

The subfolder QC & Reports contain Reports for both the DNA and RNA part of the workflow. Each report has the prefix DNA or RNA as appropriate. The folder includes, among others, the following report types:

- Remove and annotate UMI reports containing statistics on UMI barcodes.
- Trim reads reports for adapter, homopolymer, and quality trimming.
- A DNA UMI group report containing a breakdown of UMI groups with different number of reads, along with percentage of groups and reads (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_output.html).
- Two UMI read reports for RNA and DNA reads, respectively, showing how many reads were ignored and the reason why the ignored reads were not included in a UMI read. Please note that the reports are generated by different tools and have different content.
- A DNA coverage report from the QC for Target Sequencing tool (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).
- An RNA-Seq report with statistics on the mapping of RNA reads (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html).

- A CNV results report, if CNV detection has been run.

The subfolder called Tracks (WT) includes:

- Four UMI Read Mappings (). Three of these relate to the RNA reads; the 'RNA read mapping' is the original RNA-Seq mapping, 'Fusion genes unaligned ends' is the re-mapping of the unaligned ends of the RNA read mapping, and 'Read mapping refined' is the reads that map to the original chromosomes when all reads are mapped to both wild-type and fusion chromosomes. The final read mapping is the DNA mapping.
- An unfiltered variant track () holds all detected variants before filters have been applied.
- A Per-Region Statistical Report track () holds information about coverage.
- An amino acid track () that displays a graphical representation of the amino acid changes. The track is based on the CDS track and in addition to the amino acid sequence of the coding sequence, all amino acids that have been affected by variants are shown as individual amino acids below the amino acid track. Changes causing a frameshift are symbolized with two arrow heads, and variants causing premature stop are marked with an asterisk.
- A Fusion Gene track () from the Detect and Refine Fusion Genes tool.
- Region- and Gene-level CNV tracks () if CNV detection has been run.
- Indels indirect evidence () produced by the Structural Variant Caller.
- An inversion and a long indels track () containing any inversions and indels longer than 100,000 bp respectively, produced by the Structural Variant Caller.

The folder Tracks (fusion) contains data related to the fusion chromosomes

- The Reference Elements for the Fusion genome (Reference sequence, Genes, mRNA, CDS and Primers).
- A Read Mapping () of the RNA UMI reads against the fusion chromosomes.

The final folder, VCF Exportable Tracks, contains outputs that can be exported together as a single VCF file using the **VCF** exporter. This folder contains a variant track of variants passing filters, a track of fusions, and, if CNV detection has been run, a CNV target-level track.

The difference between the Unfiltered variant track in the Tracks (WT) folder and the Variants passing filters track depends on the following options available in the filtering steps:

- **Filter based on quality criteria**: Average Quality (quality of the sequenced bases that carry the variant), QUAL (significance of the variant), and Read Direction Test Probability (relative presence of the variant in the reads from different directions that cover the variant position).
- **Remove homopolymer error type variants**, i.e., errors of the indel type that occur in homopolymer regions. These regions are known to be harder to sequence than non-homopolymeric regions.

- **Remove false positive based on frequency** Variants with a frequency above the specified threshold will be included in the filtered variant track. Note that the unfiltered variant track is generated by the Low Frequency Variant Detection tool run with a frequency cut-off value of 0.5. This value can be considered a pre-filter, which is initially applied to each site in the alignment and determines which sites the variant caller should consider potential variant sites when it starts the error rate and site type/frequencies parameter estimation. In the case of this option, a frequency cut-off is applied on the final candidate variant set (after variants that span across multiple alignment sites have been reconstructed). It is only meaningful to apply this post-filter at a value that is at least as high as the pre-filter value, and we actually recommend using a value that is as least twice as high (1.0). This allows for some wiggle-room when going from the single-site to the multiple site variant construction, in particular to avoid that long indels are fragmented due to coverage difference throughout the considered region.

We recommend evaluating the support for identified fusions, see Interpretation of fusion results for details.

14.4.2 Running multimodal workflows in batch using metadata

In order to run the workflow in batch, metadata must be provided to describe which DNA and RNA reads belong together.

See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html for details on batch analysis, https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Metadata.html for general information about metadata, and https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Importing_metadata.html for how to import a metadata table.

Metadata can easily be supplied from an Excel spreadsheet or CSV file. A minimal spreadsheet should contain the start of the file names and a column containing the sample information:

Fastq	Sample
DNA-S1	S1
RNA-S1	S1
DNA-S2	S2
RNA-S2	S2

Start the workflow normally, but remember to tick batch twice - once when selecting DNA reads (figure 14.2), and again when selecting RNA reads (figure 14.3).

Metadata can be chosen in the 'Configure batching' dialog (figure 14.4).

The next dialog will show how the batching will be performed (figure 14.5).

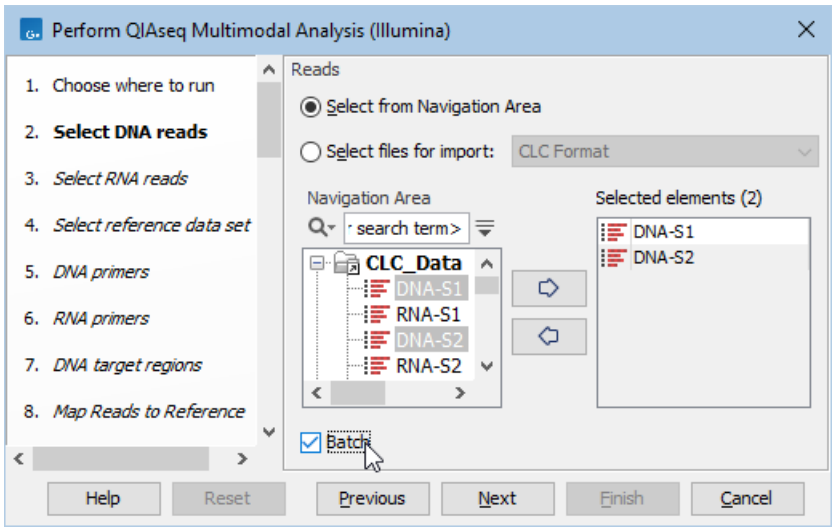


Figure 14.2: *Both DNA samples are selected at the same time. They will be grouped into batches in a later step.*

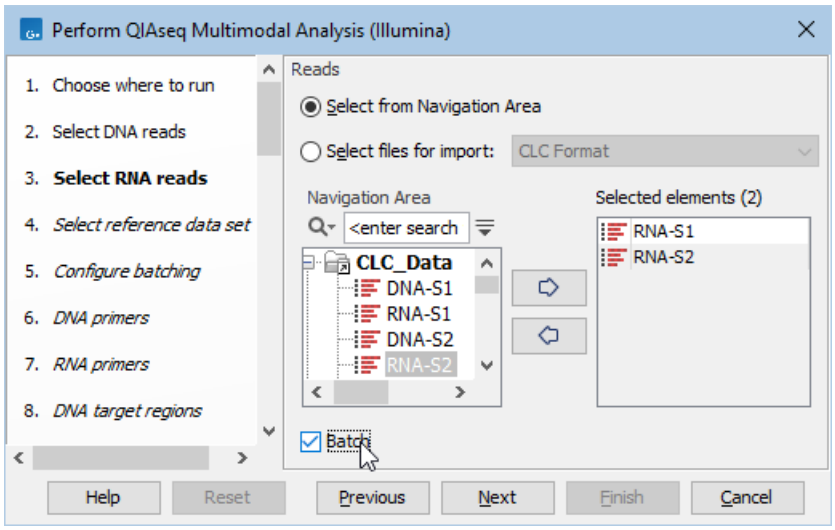


Figure 14.3: *Both RNA samples are selected at the same time. They will be grouped into batches in a later step.*

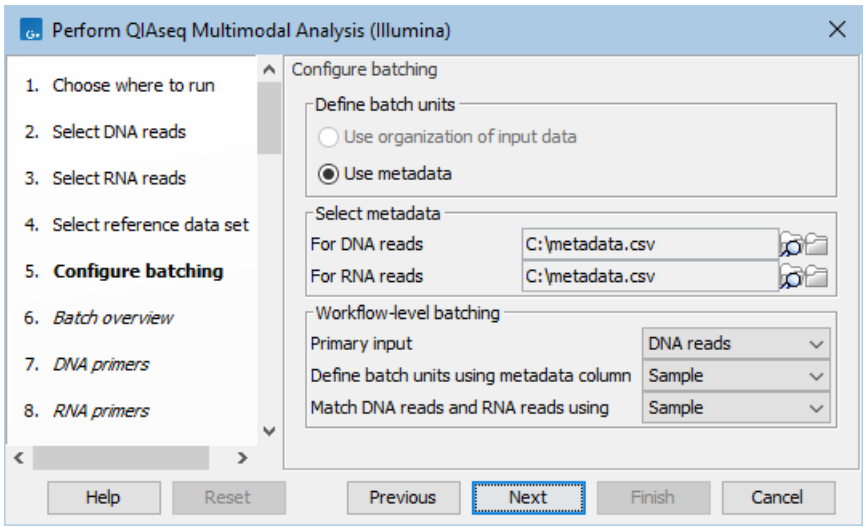


Figure 14.4: *Configuration of batch units based on metadata. Each batch unit is named after the DNA file name. DNA and RNA reads are grouped together if they share a value in the 'Sample' column.*

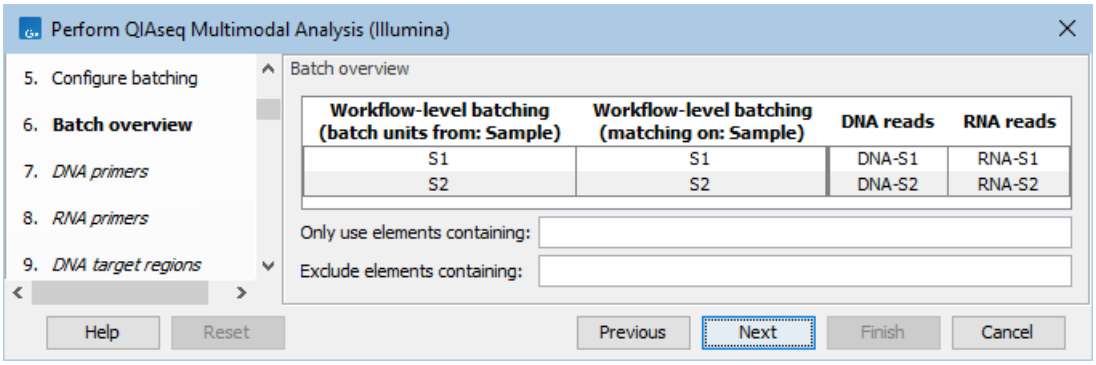


Figure 14.5: *Overview of the batch units. Each batch unit is named after the DNA file name. DNA and RNA reads are grouped together if they share a value in the 'Sample' column.*

14.5 Perform QIAseq Multimodal Panel Analysis with TMB and MSI

The **Perform QIAseq Multimodal Panel Analysis TMB and MSI (Illumina)** template workflow can be used for analyzing DNA and/or RNA reads generated using the QIAseq Multimodal Pan Cancer panel (UHS-5000Z or QHS-5000Z). The workflow extends the Perform QIAseq Multimodal Panel Analysis (Illumina) template workflow, see section 14.4, to calculate a TMB score as well as assess MSI status.

Note that the QIAseq Multimodal Panels are designed against genome build hg19 for the DNA panel and hg38 for the RNA panel. BED files are provided in the respective genome build. However, the template workflow requires that reference data for both DNA and RNA is for the same genome build. The two QIAseq Multimodal Reference Data Sets provided by the Reference Data Manager are for genome build hg38, where the reference data for the DNA panel has been converted to hg38 as described below.

For custom panels, the DNA panel BED file needs to be imported against hg19, after which it should be converted to hg38 using the tool **Convert Annotation Track Coordinates**, see section 6.6. If many regions are lost during conversion, it can cause reads to be discarded that would have otherwise mapped to the lost target regions. To avoid such issues, a copy of the template workflow can be used, containing only the analysis of the DNA reads, and the workflow should be run using the imported BED file against hg19.

Launching the workflow

To run this workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **QIAseq Sample Analysis** () | **Other QIAseq workflows** () | **Perform QIAseq Multimodal Panel Analysis TMB and MSI (Illumina)** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

Options can be configured in the following dialogs:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select whether you want to analyze DNA and/or RNA reads.
- **Select DNA/RNA Reads**. Select the reads. When analyzing more than one sample at a time, check the **Batch** checkbox in the lower left corner of the dialog.
- **Specify reference data handling**. Select the **QIAseq Multimodal Pan Cancer hg38** Reference Data Set, see chapter 3 for details.
- **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.

- **Batch overview**, if running in batch mode. Verify that the batching is as intended.
- **MSI baseline**. A default MSI baseline from the Reference Data Set is provided for this workflow, but this is for demo purpose only and will not give the true microsatellite instability status. We recommend that the MSI baseline is generated using samples that are sequenced under the same lab conditions as the multimodal samples (see section 8.3).
- **Map Reads to Reference**. Configure masking.

  By default, the GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set masking track is selected, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1. Changing the masking mode from 'No masking' to 'Exclude annotated' excludes these regions.
- **QC for Target Sequencing**. Specify the minimum coverage needed on all positions in a target for it to be considered covered. For somatic calling, we recommend setting this to at least 100x
- **Copy Number Variant Detection (Targeted)**. Specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping, or if the target regions files contains fewer than 50 regions, the Copy Number Variation analysis will not be carried out.
- **Remove False Positives (filter on allele frequency)**. Specify the minimum frequency of detected variants.

  The frequency cutoff is the only open parameter in this workflow and the workflow can detect down to 1% variant frequency. Even when setting the frequency lower it will not output lower frequencies as the variant calling is initially done down to 1% by the variant caller in this workflow. Further adjustments needs to be done by opening a copy of the workflow.
- **Calculate TMB Score**. Optionally calculate TMB status based on a low and a high threshold.

  Note that the default values of 10 and 15 have been chosen based on internal benchmark analyses and should be set according to the samples analyzed.
- **Create Sample Report (DNA/RNA)**. Select relevant summary items and specify thresholds for quality control. Summary items, thresholds, and an indication of whether specified thresholds were met, will be shown in the quality control section of the `sample report`. The default summary items are appropriate for many data sets, but may need to be adjusted.
- **Detect and Refine Fusion Genes**. Configure the following options as needed:
  - Detect exon skippings
  - Detect novel exon boundaries
  - Detect novel exon boundaries in both genes
  - Gene filter action
  - Genes for filtering (tracks)
  - Fusion filter action

  - Fusions for filtering (tables)

  For details about the elements used by default in 'Genes for filtering (tracks)' and 'Fusions for filtering (tables)', see Exclude lists.

  For general details about fusion detection, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Detect_Refine_Fusion_Genes.html.

- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

Launching using the QIAseq Panel Analysis Assistant

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under Multimodal Panels.

14.5.1 Output from the Perform QIAseq Multimodal Panel Analysis with TMB and MSI (Illumina)

The Perform QIAseq Multimodal Panel Analysis workflow produces a large number of files organized into a number of subfolders as well as single elements.

Besides two more reports in the subfolder QC & Reports called TMB_report and MSI_report, respectively and two additional tracks containing TMB_somatic_variants and Loci_track, the output is identical to the output produced from the Perform QIAseq Multimodal Panel Analysis (Illumina) template workflow described in section 14.4.1.

The TMB and MSI reports are also included in the DNA_sample_report that is provided for quality control. The TMB report includes a distribution plot of variant frequencies for the somatic variants that can be quite informative. The MSI report includes the status and performance for the different loci. The two tracks are included in the Genome Browser View (WT) and is useful for visualizing the variants and loci.

Part IV

Biomedical Template Workflows

Chapter 15

SARS-CoV-2 workflows

Contents

The SARS-CoV-2 workflows

Three workflows are available for analyzing SARS-CoV-2 data (figure 15.1), one workflow is a generic workflow for use with ARTIC V3 SARS-CoV-2 primers designs, one workflow is customized for use with Ion AmpliSeq SARS-CoV-2 Research Panel data and the last workflow is customized for use with QIAseq DIRECT SARS-CoV-2 Panel data. All workflows can take one or multiple samples as input, which allows for analysis of a single sample or comparison of multiple samples based on a single workflow run.

The general approach of both workflows is mapping the reads to a reference, generating a consensus sequence from the mapping, calling variants, and generating outputs that allow for efficient review of results, including cross-sample comparison.

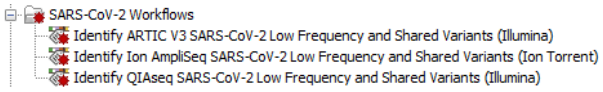


Figure 15.1: *The available SARS-CoV-2 template workflows*

Two variant tracks are produced by each workflow, one containing variants likely to be true variants, those with frequencies between 50% and 100%, and another containing all potential variants, called low frequency variants with defaults down to between 10% and 20% depending on sequencing technology. Potential low frequency variants are likely to need further validation, as they may represent new mutations in the sample, but may be due to other factors, for example reverse transcriptase or sequencing errors.

In more detail, each workflow takes this general approach:

- Reads are trimmed, as needed for the sequencing protocol used.
- Reads are mapped to a reference.
- Structural variants are identified in the mapping. Ploidy set to diploid to support mixed viral populations.
- The mapping is locally re-aligned using the structural variants identified above as guidance track.
- Marginal reads, i.e. reads that contain large unaligned ends, as well as primers (when relevant), are removed from the mapping.
- A consensus sequence is generated from the mapping using Create Consensus Sequence from Variants, where consensus calls are made by substitution of identified variants found in a sample. Areas with coverage below 30x will be represented by ambiguous nucleotides (N).
- The Low Frequency Variant Detection tool is used to call variants in the mapping and variants are further filtered based on different quality metrics. Two variant tracks are generated, one with variants of frequencies above 50% and another with frequencies down to the low frequency cut-off (Default between 10% and 20%).
- Reports are generated by various tools in the workflow, and summaries of these reports are collected together and output as a combined report, which can be used for quality control.
- Track lists are generated, allowing for detailed, visual review of results.

Part of each workflow runs on each sample individually, with the per-sample results then being combined to aid inter-sample comparison. Thus, it is assumed that data for multiple samples will be provided when the workflow is launched. If data for only one sample is provided, the workflow will still run, and the results for the individual sample are still valid.

The workflow outputs can be used with the tools in CLC Microbial Genomics Module. Examples include:

- Understanding sample contamination through taxonomic profiling of unmapped reads.
- Functional analysis with BLAST and DIAMOND.
- Tree construction from consensus sequences or variant calls to trace the evolution of the virus.

Please see the sections on Taxonomic Analysis, Functional Analysis, Phylogenetic trees using SNPs and k-mers and MLST Scheme Tools in https://resources.qiagenbioinformatics.com/manuals/clcmgm/current/index.php?manual=Introduction.html for further details.

15.1 Identify ARTIC V3 SARS-CoV-2 Low Frequency and Shared Variants (Illumina)

The **Identify ARTIC V3 SARS-CoV-2 Low Frequency and Shared Variants (Illumina)** workflow includes all necessary steps for processing paired-end reads from SARS-CoV-2 samples, such as sample QC, trimming of adapters (note that it might be necessary to adjust trim adapters depending on propocol) and primers, variant calling relative to reference MN908947.3 and extraction of a consensus sequence. Default the workflow is configured to trim ARTIC V3 primers and adapters of the reads, however if these steps are unnecessary they can be removed or adjusted.

This template workflow can be found under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **SARS-CoV-2 Workflows** () | **Identify ARTIC V3 SARS-CoV-2 Low Frequency and Shared Variants (Illumina)** ()

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

In the next step, select the data to be analyzed. These can be sequence lists containing paired-end reads selected from the Navigation Area, or by using the "Select files for on-the-fly import" option, where files containing paired-end read data can be selected from disk. These will be imported as part of the workflow run. When importing Illumina paired-end data, the "Paired reads" option needs to be enabled.

The workflow contains an Iterate element, allowing each sample to be analyzed individually, before the results are combined for comparison. The "Batch" check box, at the bottom of the dialog, should normally remain *unchecked* when launching this workflow.

In the next step the relevant Reference Data Set is selected. The workflow uses SARS-CoV-2 reference MN908947.3 by default. Note that alternative reference data sets can be created, as described in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html. Depending on protocol trim adapters adjustments might be necessary. Either extend the current data element or create a new element containing the correct trim adapter sequences and add it to a Custom Sets as described in the link above.

In the next step, you specify how the batch units are defined, that is, which data files come from each individual sample and thus should be analyzed together. Batch units can be defined through the organization of the input data or by using metadata. Further information can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html. When the metadata option is chosen, selecting an Excel file that describes the data will often be the most convenient method, and it is the only option available when input data will be imported as part of the workflow run. When using already-imported data as input, existing metadata tables, where associations from the input data are already in place, can also be selected.

In the next step, a preview of the batch units is shown. If this looks as expected, you can proceed to configure the analysis settings.

In the next step the Remove False Positives (high frequency) quality filter can be adjusted. High frequency variants (>=50%) with an average base quality lower than the specified value will be

discarded (figure 15.2).

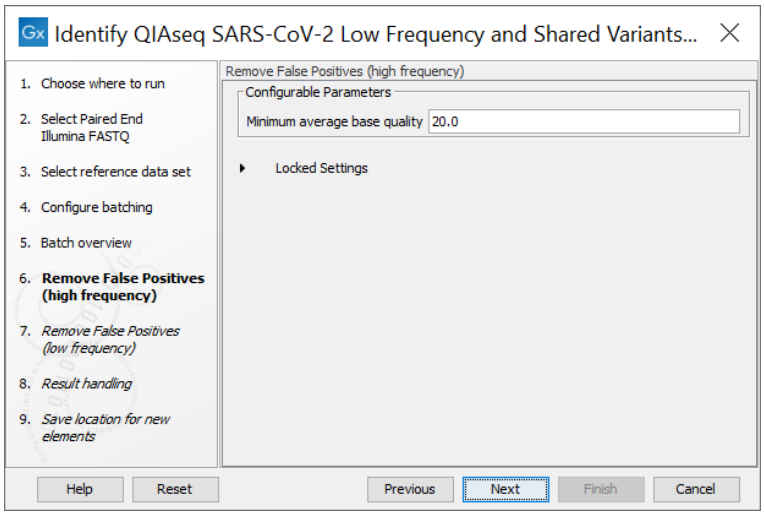


Figure 15.2: *The minimum quality for variants to be included in the output track*

In the next step, the Remove False Positives (low frequency) quality filter applying filters for both quality and frequency. Lower frequency variants (Default >=10%) with an average base quality lower than the specified value will be discarded by default.

Finally, choose where to save the results.

The outputs generated are described in section 15.4.

15.2 Identify QIAseq SARS-CoV-2 Low Frequency and Shared Variants (Illumina)

The **Identify QIAseq SARS-CoV-2 Low Frequency and Shared Variants (Illumina)** workflow includes all necessary steps for processing paired-end reads from SARS-CoV-2 samples, such as sample QC, trimming of adapters and primers, variant calling relative to reference MN908947.3 and extraction of a consensus sequence. Default this workflow is configured to run with the QIAseq DIRECT primers.

This template workflow can be found under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **SARS-CoV-2 Workflows** ( ) | **Identify QIAseq SARS-CoV-2 Low Frequency and Shared Variants (Illumina)** ( )

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

In the next step, select the data to be analyzed. These can be sequence lists containing paired-end reads selected from the Navigation Area, or by using the "Select files for on-the-fly import" option, where files containing paired-end read data can be selected from disk. These will be imported as part of the workflow run. When importing QIAseq paired-end data, the "Paired reads" option needs to be enabled.

The workflow contains an Iterate element, allowing each sample to be analyzed individually,

before the results are combined for comparison. The "Batch" check box, at the bottom of the dialog, should normally remain *unchecked* when launching this workflow.

In the next step the relevant Reference Data Set is selected. The workflow uses SARS-CoV-2 reference MN908947.3 by default. Note that alternative reference data sets can be created, as described in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

In the next step, you specify how the batch units are defined, that is, which data files come from each individual sample and thus should be analyzed together. Batch units can be defined through the organization of the input data or by using metadata. Further information can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html. When the metadata option is chosen, selecting an Excel file that describes the data will often be the most convenient method, and it is the only option available when input data will be imported as part of the workflow run. When using already-imported data as input, existing metadata tables, where associations from the input data are already in place, can also be selected.

In the next step, a preview of the batch units is shown. If this looks as expected, you can proceed to configure the analysis settings.

In the next step the Remove False Positives (high frequency) quality filter can be adjusted. High frequency variants (>=50%) with an average base quality lower than the specified value will be discarded (figure 15.3).

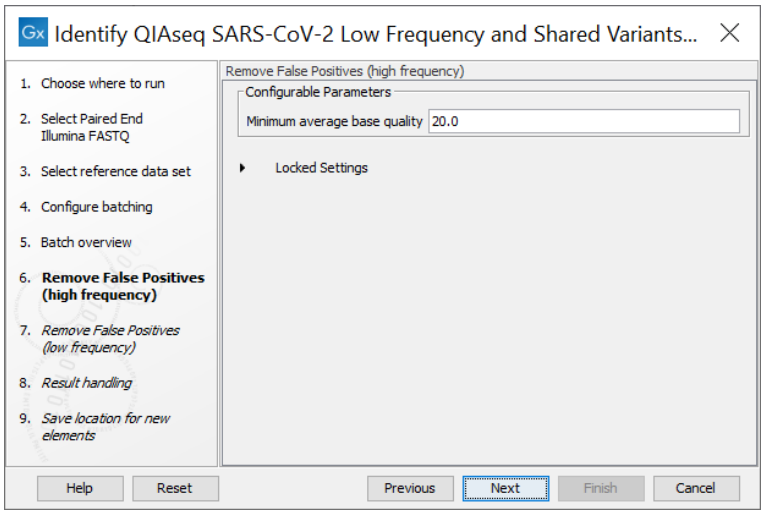


Figure 15.3: *The minimum quality for variants to be included in the output track*

In the next step, the Remove False Positives (low frequency) quality filter applying filters for both quality and frequency. Lower frequency variants (Default >=10%) with an average base quality lower than the specified value will be discarded by default.

Finally, choose where to save the results.

The outputs generated are described in section 15.4.

The workflow is also available in the QIAseq Panel Analysis Assistant, see chapter 11 under SARS-CoV-2.

15.3 Identify Ion AmpliSeq SARS-CoV-2 Low Frequency and Shared Variants (Ion Torrent)

The **Identify Ion AmpliSeq SARS-CoV-2 Low Frequency and Shared Variants (Ion Torrent)** workflow includes all necessary steps for processing the SARS-CoV-2 reads generated with the IonTorrent AmpliSeq pipeline, such as sample QC, filtering of human control reads, variant calling relative to reference MN908947.3 and extraction of a consensus sequence.

This template workflow can be found under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** () | **SARS-CoV-2 Workflows** () | **Identify Ion AmpliSeq SARS-CoV-2 Low Frequency and Shared Variants (Ion Torrent)** ()

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

In the next step, select the data to be analyzed. These can be sequence lists containing single-end reads selected from the Navigation Area, or by using the "Select files for on-the-fly import" option, where files containing single-end read data can be selected from disk. These will be imported as part of the workflow run. Importing Ion AmpliSeq reads should be done using the Ion Torrent importer.

The workflow contains an Iterate element, allowing each sample to be analyzed individually, before the results are combined for comparison. The "Batch" check box, at the bottom of the dialog, should normally remain *unchecked* when launching this workflow.

In the next step the relevant Reference Data Set is selected. The workflow uses SARS-CoV-2 reference MN908947.3 by default. Note that alternative reference data sets can be created, as described in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

In the next step, you specify how the batch units are defined, that is, which data files come from each individual sample and thus should be analyzed together. Batch units can be defined through the organization of the input data or by using metadata. Further information can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html. When the metadata option is chosen, selecting an Excel file that describes the data will often be the most convenient method, and it is the only option available when input data will be imported as part of the workflow run. When using already-imported data as input, existing metadata tables, where associations from the input data are already in place, can also be selected.

In the next step, a preview of the batch units is shown. If this looks as expected, you can proceed to configure the analysis settings.

In the next step the Remove False Positives (high frequency) quality filter can be adjusted. High frequency variants (>=50%) with an average base quality lower than the specified value will be discarded (figure 15.4).

In the next step, the Remove False Positives (low frequency) quality filter applying filters for both quality and frequency. Lower frequency variants (Default >=20%) with an average base quality lower than the specified value will be discarded by default.

Finally, choose where to save the results.

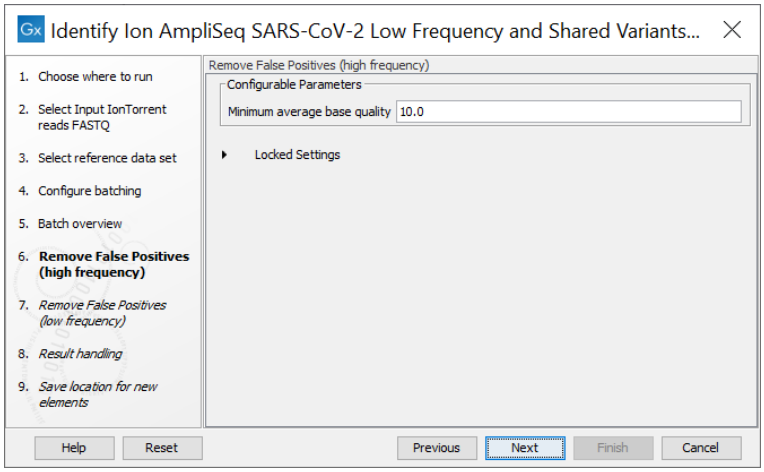


Figure 15.4: *The minimum quality for variants to be included in the output track*

The outputs generated are described in section 15.4.

15.4 SARS-CoV-2 workflow output

The outputs from the SARS-CoV-2 Low Frequency and Shared Variants workflows include results for each sample, cross-sample summaries, and a track list for efficient visual investigation and comparison of results. An example of the outputs generated from an analysis involving two samples is shown in figure 15.5.

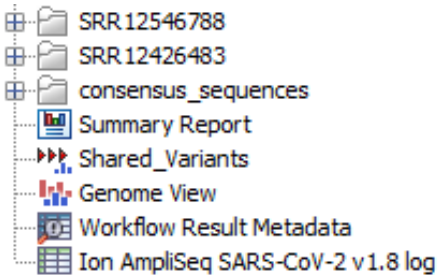


Figure 15.5: *The SARS-CoV-2 Low Frequency and Shared Variants workflow output*

15.4.1 Summary outputs

1. **Summary Report**: A cross-sample, combined report. This report summarizes key QC statistics such as read mapping summary and coverage statistics. See section 15.4.2 for a detailed description.
2. **Shared_Variants**: A track containing variants with >=50% frequency found in one or more samples.
3. **Consensus sequences:** A folder containing the consensus sequence for each sample.
4. **Genome View**: A track list containing:
   - The reference SARS-CoV-2 genome MN908947.3
   - The reference SARS-CoV-2 CDS regions
   - Sample coverage graphs
   - The reference SARS-CoV-2 gene regions
   - The Shared_Variants track

   An example is shown in figure 15.6.

15.4.2 Sample specific outputs

A separate folder of results is created for each sample and contains sample specific reports, tracks and additional supplemental outputs, (figure 15.7).

Reports

The following reports are produced:

- **trim_report**: Reports read lengths before and after trimming, as well as the number of reads discarded because they did not pass the minimum length threshold of 50bp after being trimmed.

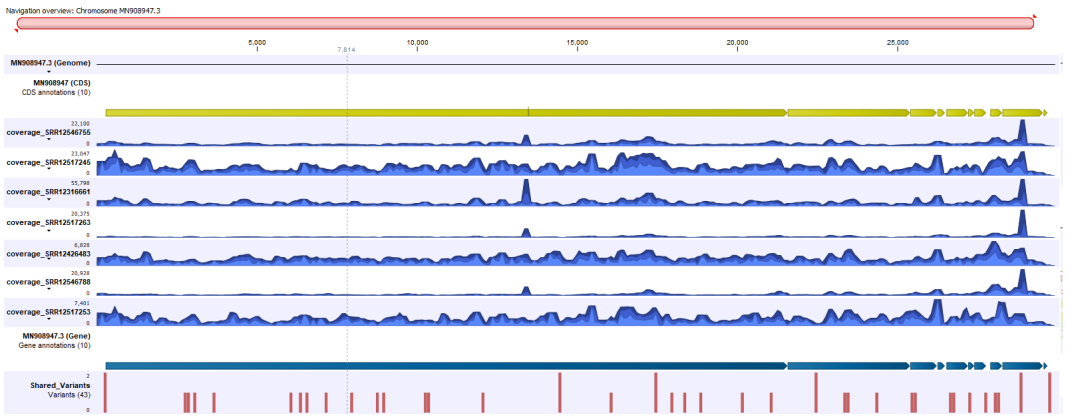


Figure 15.6: *An example of the Genome View track list created from analysis of seven SARS-CoV-2 samples*

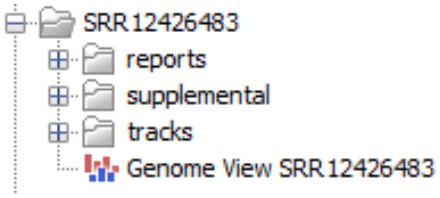


Figure 15.7: *The per sample outputs generated by the SARS-CoV-2 Low Frequency and Shared Variants workflow*

- **ligation_artifacts_report**: Summarizes any ligation artifacts found in and removed from the read mapping.
- **QC_report**: Contains quality information such as sequence length, GC content and quality, to help detect any sequencing bias in the sample reads.
- **structural_variants_report**: Provides an overview of any potential structural variants found in the sample.
- **mapping_report**: Summarizes the number of mapped and unmapped reads, as well as providing coverage statistics for the SARS-CoV-2 reference genome.
- **coverage_report**: Gives an overview of coverage in the targeted regions e.g. minimum coverage in % of target regions and number of targets passing the coverage threshold of 30x. For the Ion AmpliSeq protocol, "target regions" refers to the targeted amplicons. For the other workflows Target region is defined by the full length genome. The coverage report is helpful for identifying regions where coverage is insufficient to reliably call variants. Variants present in regions with insufficient coverage is not called although the read mapping shows evidence of the variant being present in the reads. The report can help understand such results and be used for adjusting coverage requirements if deemed necessary.
- **variant_report**: Contains information about the estimated error model for the variant calls.
- **variant_track_statistics_report**: Contains information about statistics on variants classified by types.

- **human_control_genes_mapping_report** (Ion AmpliSeq SARS-CoV-2 Low Frequency and Shared Variants workflow only): Contains mapping information for the 5 human expression gene controls.

Tracks

The following tracks are produced:

- **breakpoints**: Contains a row for each potential breakpoint with information on region, p-value, mapping information and number of reads supporting breakpoint.
- **InDels_track**: The InDels used as guidance variants for the local realignment. Note, in some cases it might be necessary to create the consensus sequence using the coordinates of the full insertion from the InDels track. This can happen when no reads span the entire region.
- **realigned_regions**: Contains a row for each region in which the mapping was improved following local realignment.
- **read_mapping**: The read mapping track after local realignment, trimming and removal of marginal reads.
- **coverage_below_30**: Shows regions that failed to meet the coverage threshold of 30x and were therefore not used when calling variants nor used towards building the consensus sequence.
- **amino_acid_track**: Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is displayed in red.
- **coverage**: Coverage per position across the SARS-CoV-2 reference.
- **variants_above_50_frequency**: A list of variants that passed the >=50% frequency quality filter. There is strong evidence these variants are present in the sample.

Genome View

The Genome View track list contains the following collection of tracks:

- The reference SARS-CoV-2 genome MN908947.3
- The reference SARS-CoV-2 CDS regions
- Amino acid changes track produced using unfiltered variant calls
- Coverage graph
- Read mapping after local realignment, trimming and removal of marginal reads
- The reference SARS-CoV-2 gene regions
- Low coverage regions (<30x coverage)
- Variants with >=50% frequency that passed the quality filter

An example is shown in figure 15.8.

Figure 15.8: *An example of the Genome View track list created for each sample*

Supplemental

The following results are placed in a folder called Supplemental:

- **unmapped reads**: Reads that did not map to the SARS-CoV-2 reference or the human control genes when using the Ion AmpliSeq SARS-CoV-2 Low Frequency and Shared Variants workflow.
- **unfiltered_variant_track**: The variant track generated by Low Frequency Variant Detection before any further filtering was carried out.
- **low_frequency_variants**: A track containing all variants passing the quality filter with a frequency (Optional value, Default >=10% or 20%).

Chapter 16

TruSight Oncology 500

Contents

TruSight Oncology panels

The TruSight Oncology 500 (TSO500) bundle enables analysis of both DNA and RNA in cancer samples and cover a wide range of cancer genes known to host driver mutations as well as fusion genes. The panel has a size that enables analysis of tumor mutational burden (TMB) scores.

Two template workflows are available for analyzing TruSight Oncology 500 panels, **Perform TSO500 DNA Analysis (Illumina)** for analyzing DNA data, and **Perform TSO500 RNA Analysis (Illumina)** for analyzing RNA data.

The **Perform TSO500 DNA Analysis (Illumina)** workflow includes somatic variant calling, CNV detection and reporting of a TMB score. The RNA workflow, **Perform TSO500 RNA Analysis (Illumina)**, includes an expression matrix and reporting of detected fusions. Both workflows are configured with hg38 reference data using RefSeq annotations and use a traditional re-sequencing strategy for DNA and RNA sample analysis, respectively.

In both the DNA and the RNA workflows, the reads are initially subjected to the following steps:

- Annotation with UMI information and trimming off the UMI barcode
- Quality trimming to remove low quality bases, adapter read-through and homoploymer regions at the read ends
- Grouping reads into consensus reads based on UMI and sequence similarity
- Mapping the UMI reads to a reference

The DNA analysis is followed by local re-alignment of the reads in the read mapping using a guidance track inferred by the Stuctural Variant Caller and variant detection using the Fixed Ploidy

Variant Detection tool. The variants are filtered on quality parameters such as QUAL, Average Quality, BaseQRankSum and read direction test probability in different filters focusing on High UMI count, low coverage, BaseQRankSum bias and homopolymer evidence. The read mapping is analyzed for a coverage profile and the somatic variants are further filtered in the Calculate TMB Score tool to evaluate the tumor mutational burden with suitable cutoffs. When a matched normal sample is available or a baseline encompassing between 3 to 5 samples are supplied as controls, CNVs are reported, if present in the sample. Finally, reports are generated by different tools in the workflow, and summaries of these reports are available in a combined report that can be used for quality control. The generated Track list allows for detailed, visual review of the results.

In the RNA workflow, a gene expression track, summarizing the expression at gene-level, and a read mapping is produced by the RNA-Seq Analysis tool. The mapped RNA reads are used for fusion gene detection. Detection of fusion genes involves creation of potential fusion chromosomes that are evaluated by a refinement approach where the RNA reads are remapped to a reference that now includes the new potential fusion genes. Finally, analysis reports are generated by different tools in the workflow, and summaries of these reports are available in a combined report that can be used for quality control. Two Track lists are generated, allowing for detailed, visual review of the results, one for the wildtype chromosomes and one for the fusion gene chromosomes.

NOTE: The current workflow is not including tools for MSI detection, as a proper baseline needs to be created for this analysis to report the true identity of MSI status. This is work in progress and the expectation is that this will be provided in a later version.

16.1 Perform TSO500 DNA Analysis

The **Perform TSO500 DNA Analysis (Illumina)** template workflow includes all necessary steps for processing paired-end reads from TSO500 DNA samples, such as sample QC, adapter trimming, somatic variant calling, TMB score calculation and CNV analysis when control samples are provided.

This template workflow is available under the Workflows menu at:

> **Workflows** | **Biomedical Workflows** () | **TSO Panel Analysis** () | **Perform TSO500 DNA Analysis (Illumina)** ()

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

In the next step, select the DNA sequencing reads to analyze. The input can be sequence lists containing paired-end reads selected from the Navigation Area, or samples can be imported using the "Select files for on-the-fly import" option, where files containing paired-end read data can be selected from disk. If choosing the "Select files for on-the-fly import" option, the "Paired reads" option needs to be enabled.

If you would like to analyze more than one sample in one workflow run, check the "Batch" box in the lower left corner of the dialog. When running multiple imported samples, metadata needs to be provided. Further information can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html.

After selecting reference data as described below you can configure the batch unit and see the

batch overview. When using metadata, selecting an Excel file that describes the data will often be the most convenient method. Providing metadata directly from an Excel file is the only option available when input data is imported as part of the workflow run.

The following dialog helps you set up the relevant Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. The dialog for selection of reference data is shown in figure 16.1.

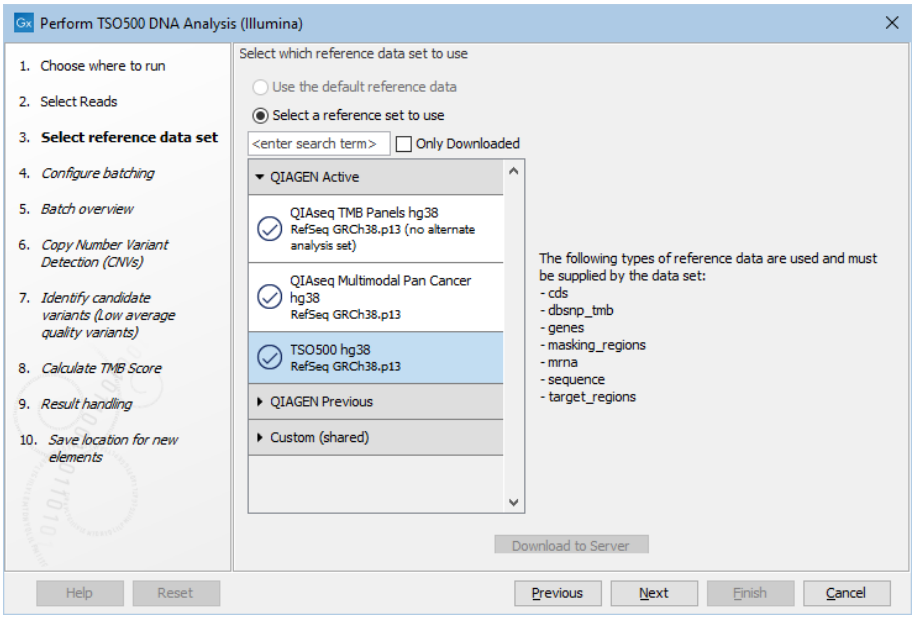


Figure 16.1: *The relevant Reference Data Set is highlighted. The text to the right lists the types of references needed by the workflow.*

Note that if you wish to Cancel or Resume the Download, you can close the template workflow and open the Reference Data Manager where the Cancel, Pause and Resume buttons are available.

If the Reference Data Set was previously downloaded, the option "Use the default reference data" is available and will ensure the relevant data set is used. You can always check the "Select a reference set to use" option to be able to specify another Reference Data Set than the one suggested.

In the Map Reads to Reference dialog, it is possible to configure masking. A custom masking track can be used, but by default, the masking track is set to GenomeReferenceConsortium_masking_hg38_no_alt_analysis_set, containing the regions defined by the Genome Reference Consortium, which serve primarily to remove false duplications, including one affecting the gene U2AF1.

In the next step, the **Average quality** cutoff for high UMI evidence SNP variants can be adjusted. This might need adjustment depending on the UMI grouping, where 2-4 UMI on average is good. A higher number of reads per UMI could indicate fragmented or bad input DNA and hence lead to lower average quality scores even for high UMI count samples. Besides, this can affect coverage and will show up as untrustworthy TMB score estimates due to low coverage (<100x coverage) in the 1 MB region required for TMB estimation.

The next step, **Calculate TMB score**, has the following Configurable Parameters (figure 16.2):

- **Enable TMB status detection using thresholds**: When enabled, the TMB status can be calculated using thresholds by customizing the minimum and maximum scores to consider when calculating the TMB status. When enabled, the TMB report will include a TMB status calculated using the specified Maximum score for low TMB status and the Minimum score for high TMB status.
- **Maximum score for low TMB status**: A score of maximum this value will be considered having low TMB status. Scores above this value will be considered intermediate or high MB status, depending on what is specified for Minimum score for high TMB status.
- **Minimum score for high TMB status**: A score of minimum this value will be considered having high TMB status. Scores below this value will be considered intermediate or low TMB status, depending on what is specified for Maximum score for low TMB status.

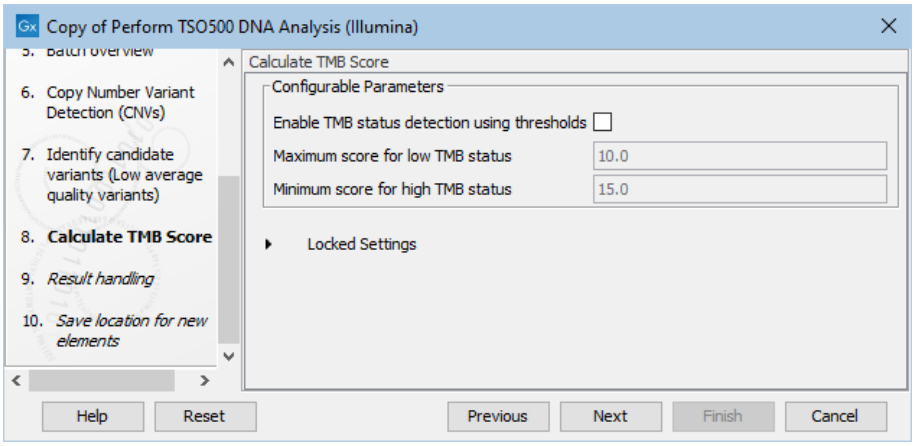


Figure 16.2: *In this dialog you can Enable TMB status detection using thresholds if you would like to base the calculation of TMB status on specific minimum and maximum scores for high and low TMB status, respectively.*

Finally, choose where to save the results.

16.1.1 Output from Perform TSO500 DNA Analysis

The **Perform TSO500 DNA Analysis (Illumina)** workflow output has been ordered into a number of single key elements and folders with different purpose.

1. **QC & Reports**: A folder that contains the following reports.
   - **DNA_remove_and_annotate_with_UMI_report**: Contains information about the number of reads being processed and the number and fraction of reads with UMIs.
   - **DNA_trim_reads_report**: Contains detailed trimming results and information about automatic adapter read-through trimming.
   - **DNA_UMI_read_report**: A comprehensive report containing statistical metrics and graphical representations on UMI reads. Focus is on duplex, read grouping, read quality and read length distribution.
   - **DNA_mapping_report**: Contains information on mapping statistics.

- **DNA_remove_ligation_artifacts_report**: Summarizes any ligation and common sequence artifacts found in and removed from the read mapping.
- **DNA_structural_variant_caller_report**: Gives an overview of the numbers and types of structural variants found in the sample.
- **DNA_variant_report**: Summarizes variants detected in the read mapping.
- **Coverage_report**: Contains statistical metrics of the covered targets.
- **TMB_report**: Contains information on variant distribution and coverage for detecting the TMB score as well as the score itself.
- **CNV_results_report**: Contains statistical metric on the CNV analysis. This output only exists when control sample(s) are provided.

These reports are summarized in the output called **DNA_combined_report**, which is useful for quick QC assessment.

2. **Tracks**: A folder that contains the following tracks.

   - **Mapped_UMI_reads**: The read mapping produced by mapping the UMI reads.
   - **Unfiltered_variants**: All variants called by the Low Frequency Variant Detection tool.
   - **Indels-Indirect_evidance**: Indels inferred from indirect evidence by the Structural Variant Caller and not detected by the Low Frequency Variant Detection tool.
   - **Amino acid track**: Track of amino acids sequences and changes introduced by the detected variants.
   - **TMB_somatic_variants**: Variants included in the TMB score estimation.
   - **Per-region_statistics_track**: Coverage report of the covered target region.
   - **Gene_level_CNV_track**: Contains CNV evidence per gene. This output only exists when control sample(s) are provided.

   These elements are compatible with the **Track List** that in addition contains VCF Exportable Tracks element described in item 3 and reference elements from the **CLC_References** location (Reference sequence, Genes, CDS and mRNA). The Track List is helpful for visual inspection of variants and CNV regions.

3. **VCF Exportable Tracks**: The VCF Exportable Tracks folder contains outputs that can be exported together as a single VCF file using the **VCF** exporter. The tracks found in the VCF Exportable Tracks folder can also be exported as a single VCF together with the **Final_fusion_genes (WT)** from the RNA part of the pipeline.

   - **Variants_passing_filters**: All variants of high enough quality to be kept. As this track contains the possible driver mutations, it is the track selected for VCF export rather than the TMB somatic variant track.
   - **Region-level_CNV_track**: CNV track with information on regions rather then genes. The region-level annotation track provides CNV breakpoints that are useful for export. This output only exists when control sample(s) are provided.

16.2 Perform TSO500 RNA Analysis

The **Perform TSO500 RNA Analysis (Illumina)** workflow includes all necessary steps for processing paired-end reads from TSO500 RNA samples, such as sample QC, adapter trimming, mapping resulting in gene expression counts and fusion gene identification.

This template workflow is available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **TSO Panel Analysis** ( ) | **Perform TSO500 RNA Analysis (Illumina)** ( )

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.

In the next step, select the RNA sequencing reads to analyze. The input can be sequence lists containing paired-end reads selected from the Navigation Area, or samples can be imported using the "Select files for on-the-fly import" option, where files containing paired-end read data can be selected from disk. If choosing the "Select files for on-the-fly import" option, the "Paired reads" option needs to be enabled.

If you would like to analyze more than one sample in one workflow run, check the "Batch" box in the lower left corner of the dialog. When analyzing multiple imported samples, metadata needs to be provided. Further information can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Running_workflows_in_batch_mode.html.

After selecting reference data as described below you can configure the batch unit and see the batch overview. When using metadata, selecting an Excel file that describes the data will often be the most convenient method. Providing metadata directly from an Excel file is the only option available when input data is imported as part of the workflow run.

The following dialog helps you set up the relevant Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. The dialog for selection of reference data is shown in figure 16.1.

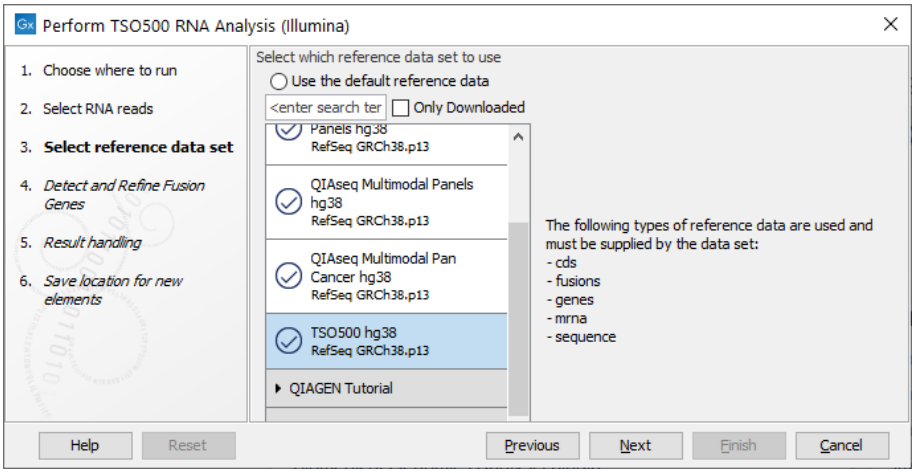


Figure 16.3: *The relevant Reference Data Set is highlighted. The text to the right lists the types of references needed by the workflow.*

Note that if you wish to Cancel or Resume the Download, you can close the template workflow and

open the Reference Data Manager where the Cancel, Pause and Resume buttons are available.

If the Reference Data Set was previously downloaded, the option "Use the default reference data" is available and will ensure the relevant data set is used. You can always check the "Select a reference set to use" option to be able to specify another Reference Data Set than the one suggested.

In the next step, you can adjust the parameters for specifying the fusion gene detection (figure 16.4).

The **Configurable Parameters** for Detect and Refine Fusion Genes are:

- **Promiscuity threshold**: The number of genes a fusion gene is allowed to fuse with. We recommend to keep this value under 10. The higher this value is, the more potential false positives can be produced. The likelihood for true fusion genes to be able to fuse with multiple genes is not that high. Genes with a high promiscuity rate are often ribosomal genes which are more likely to have a long polyA tail that can fuse in different places and thereby indicate potential gene fusions. This is especially true when the "Detect fusion with novel exon boundaries" is enabled, as this option allows fusion to intronic regions where polyA stretches are more prominent. However, some genes do tend to fuse in intronic regions creating novel exons such as PLM-RARA where this phenomenon is often seen.
- **Detect exon skippings**: Allows for detection of novel transcripts that origin from exon skipping events.
- **Detect fusions with novel exon boundaries**: When enabled fusions with breakpoints that are within a specified distance to the known exon boundary, but not at canonical exon boundaries, are also reported. The default setting for "Maximum distance to known exon boundary" is 8.

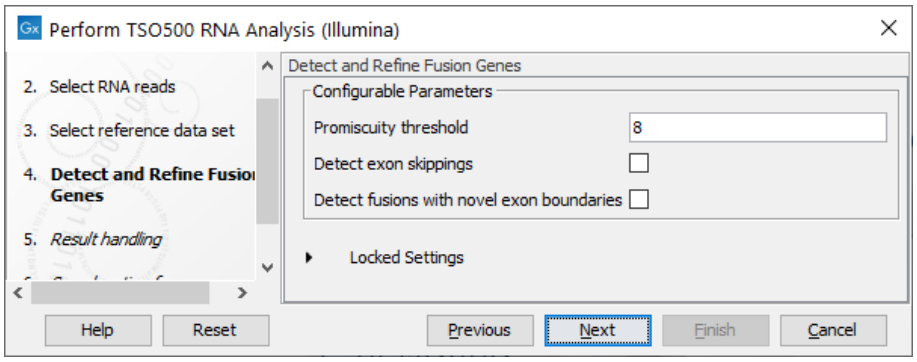


Figure 16.4: *Adjustable parameters for the Detect and Refine Fusion Genes wizard step. Options include detection of exon skipping and fusions with novel exon boundaries.*

Finally, choose where to save the results.

16.2.1 Output from Perform TSO500 RNA Analysis

The **Perform TSO500 RNA Analysis (Illumina)** workflow produces a relatively large amount of outputs. The outputs have been ordered in a number of folders and single elements as described below.

The two main outputs are:

- **Gene_expression**: A track with gene expression counts.
- **Fusion_report**: This report includes summary statistics and a graphical representation of the detected fusions. For a detailed description of the Fusion report, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_from_Detect_Refine_Fusion_Genes.html.

1. **QC & Reports**: A folder that contains the following reports.
   - **RNA_remove_and_annotate_with_UMI_report**: Contains information about the number of reads being processed and the number and fraction of reads with UMIs.
   - **RNA_adapter_trim_report**: Summarizes the trimming results such as automatic adapter read-through trimming.
   - **RNA_UMI_read_report**: A comprehensive report containing statistical metrics and graphical representations on UMI reads. Focus is on duplex, read grouping, read quality and read length distribution.
   - **RNA_quality_trim_report**: Contains information about quality trimming.
   - **RNA-Seq_report**: Contains information on mapping statistics.
   - **RNA_remove_ligation_artifacts_report**: Summarizes ligation and common sequence artifacts found in and removed from the read mapping.

   These reports are summarized in **RNA_combined_QC_report**, which is useful for quick QC assessment.
2. **Tracks (WT)**: A folder that contains the following tracks.
   - **RNA_read_mapping (WT)**: The mapping of UMI reads produced by the RNA-Seq Analysis tool
   - **Fusion_genes_unaligned_ends (WT)**: The unaligned ends produced by the RNA-Seq Analysis tool
   - **Final_fusion_genes (WT)**: The breakpoints on the reference genome of all fusions detected by Detect and Refine Fusion Genes
   - **Read_mapping_refined (WT)**: Read mapping from the Detect and Refine Fusion Genes tool, where Fusion chromosomes are removed

   These elements are presented in the **Genome Browser View (WT)** together with the VCF Exportable Tracks element described in item 4 and reference elements from the **CLC_References** location (Reference sequence, Genes, CDS and mRNA). Only the refined read mapping is loaded by default, but the remaining elements can be added for investigating missed fusions.
3. **Tracks (fusion)**: A folder that contains the fusion chromosome reference tracks and fusion evidence:
   - **CDS (fusions)** Artificial CDS track matching the artificial reference fusion chromosome
   - **Genes (fusions)** Artificial genes track matching the artificial reference fusion chromosome

- **mRNA (fusions)** Artificial mRNA track matching the artificial reference fusion chromosome
- **Reference_sequence (fusions)** Artificial sequence track matching the artificial reference fusion chromosome
- **RNA_read_mapping (fusions)**: UMI reads mapped to the artificial fusion chromosomes
- **Fusion_genes (fusion)**: The breakpoints on the artificial fusion chromosomes of all detected fusions

These elements are used in the **Genome Browser View (Fusions)** that facilitates visualization of the individual fusions.

4. **VCF Exportable Tracks**: This folder contains the VCF exportable fusion element called **PASS_fusion_genes (WT)**, listing the breakpoints on the reference genome of detected fusions that have passed all relevant filters. The track can be exported to VCF format alone or together in a single VCF output together with the **VCF Exportable Tracks** from the DNA part of the pipeline.

Chapter 17

WGS, WES, TAS and WTS template workflow descriptions

Contents

CLC Workbench contains several template workflows that support analysis of cancer data, but also analysis of hereditary diseases and other conditions that are best studied using family analysis.

Before running an application workflow, it is important to prepare the sequencing reads as explained in the following diagram (figure 17.1).

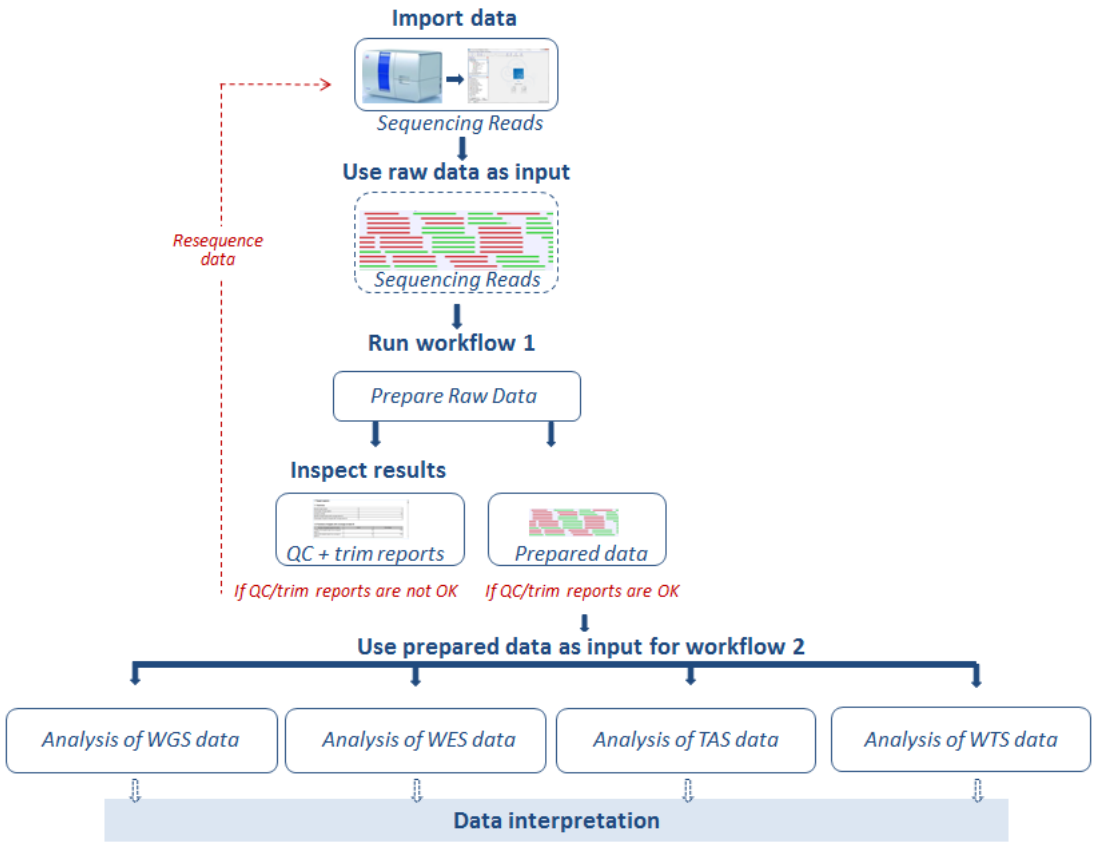


Figure 17.1: *From sequencing reads to data interpretation.*

The workflows are specific to the type of data used as input: Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), Targeted Amplicon Sequencing (TAS) and Whole Transcriptome Sequencing (WTS). For each of the first three categories, WGS, WES, and TAS, **General Analysis** workflows can be used for general identification and annotation of variants irrespective of disease. In **Somatic Cancer** you can find workflows designed specifically for cancer research. Finally, use **Hereditary Disease** workflows to study variants that cause rare diseases or hereditary diseases (HD).

The template workflows found under each of the first three applications have similar names (with the only difference that "WGS", "WES", or "TAS" have been added after the name). However, each have been tailored to the individual applications with parameter settings that have been adjusted to fit the expected differences in coverage between the different application types. We therefore recommend that you use the template workflow that is found under the relevant application heading.

17.1 General workflows

The general template workflows are universal workflows in the sense that they can be used independently of the disease that is being studied. Two workflows exist in this category:

- **Annotate Variants** annotates variants with gene names, conservation scores, amino acid

changes, and information from relevant databases.

- **Identify Known Variants in One Sample** maps sequencing reads and looks for the presence or absence of user-specified variants in the mapping.

17.2 Somatic cancer

The somatic cancer template workflows are workflows that have been tailored to cancer research. In this category it is possible to find workflows that can compare variants in matched tumor-normal pairs. The workflows found in the Somatic Cancer category use the Low Frequency Variant Detection for variant calling. The advantages of using this variant caller when analyzing cancer data are that 1) it does not take ploidy into consideration, and 2) it is particularly good at picking up low frequency variants in contrast to the other variant callers.

The workflows that are available in this category are:

- **Filter Somatic Variants** removes variants outside the target region (only targeted experiments) and common variants present in publicly available databases. Annotates with gene names, conservation scores, and information from relevant databases.
- **Identify Somatic Variants from Tumor Normal Pair** removes germline variants by referring to the control sample read mapping, removes variants outside the target region (in case of a targeted experiment), and annotates with gene names, conservation scores, amino acid changes, and information from relevant databases.
- **Identify Variants** calls variants in the mapped and locally realigned reads, removes false positives and, in case of a targeted experiment, removes variants outside the targeted region. Variant calling is performed with the Low Frequency Variant Detection tool.

17.3 Hereditary disease

Hereditary disease template workflows have been developed to support identification of disease-causing mutations in families.

Hereditary diseases can be non-cancer related diseases, such as inherited heart diseases or familial hypercholesterolemia, or it can be inherited cancers such as hereditary colorectal cancer or hereditary breast cancer. In addition to the hereditary diseases, family analysis can help researchers identify rare disease causing mutations that can be:

- a new mutation, also known as a de novo mutation, that is only present in a child and not in any of the parents
- a combination of events that occur in the same gene but at different positions in each of the parents, which is not disease causing by itself in either of the parents, but when both variants are found in a child, it becomes disease causing; this type of variant is known as a compound heterozygous variant.

We offer workflows tailored to two family sizes, 1) a classical "Trio", consisting of a mother, father, and an affected child (the proband), and 2) a "Family of Four", which is mother, father,

affected child, and either a sibling in the workflows that detects rare diseases or, in the workflows that detect inherited diseases, an other affected family member, e.g., a sibling or a grand-parent.

These workflows use the "Fixed Ploidy Variant Detection" tool, designed to call variants in samples with known ploidy from read mapping data.

Workflows designed to detect rare variants can both pick up de novo variants as well as compound heterozygous variants.

In addition to the Trio and Family of Four workflows, additional workflows exist that have been designed to pick up variants that are inherited from either the mother or the father.

The available workflows in this category are:

- **Filter Causal Variants** removes variants outside the target region (only targeted experiments) and common variants present in publicly available databases. Annotates with gene names, conservation scores, and information from relevant databases.
- **Identify Causal Inherited Variants in a Family of Four** identifies putative disease causing inherited variants by creating a list of variants present in all three affected individuals and subtracting all variants in the unaffected individual. The workflow includes a back-check for all family members.
- **Identify Causal Inherited Variants in a Trio** identifies putative disease causing inherited variants by creating a list of variants present in both affected individuals and subtracting all variants in the unaffected individual. The workflow includes a back-check for all family members
- **Identify Rare Disease Causing Mutations in a Family of Four** identifies de novo and compound heterozygous variants from an extended family of four, where the fourth individual is not affected.
- **Identify Rare Disease Causing Mutations in a Trio** identifies de novo and compound heterozygous variants from a Trio. The workflow includes a back-check for all family members.
- **Identify Variants (HD)** calls variants in the mapped and locally realigned reads, removes false positives and, in case of a targeted experiment, removes variants outside the targeted region. Variant calling is performed with the Fixed Ploidy Variant Detection tool.

Chapter 18

Whole genome sequencing (WGS)

Contents

The most comprehensive sequencing method is whole genome sequencing that allows for identification of genetic variations and somatic mutations across the entire human genome. This type of sequencing encompasses both chromosomal and mitochondrial DNA. The advantage of sequencing the entire genome is that not only the protein-coding regions are sequenced, but information is also provided for regulatory and non-protein-coding regions.

A number of template workflows are available for analysis of whole genome sequencing data (figure 18.1). The concept of the pre-installed template workflows is that read data are used as input in one end of the workflow, and after running it, the workflow will output a Track List and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual template workflows can be used for and go through step by step how to run the workflows.

Remember you will have to prepare data with the **Prepare Raw Data** workflow described in section 2 before you proceed to running any of these workflows.

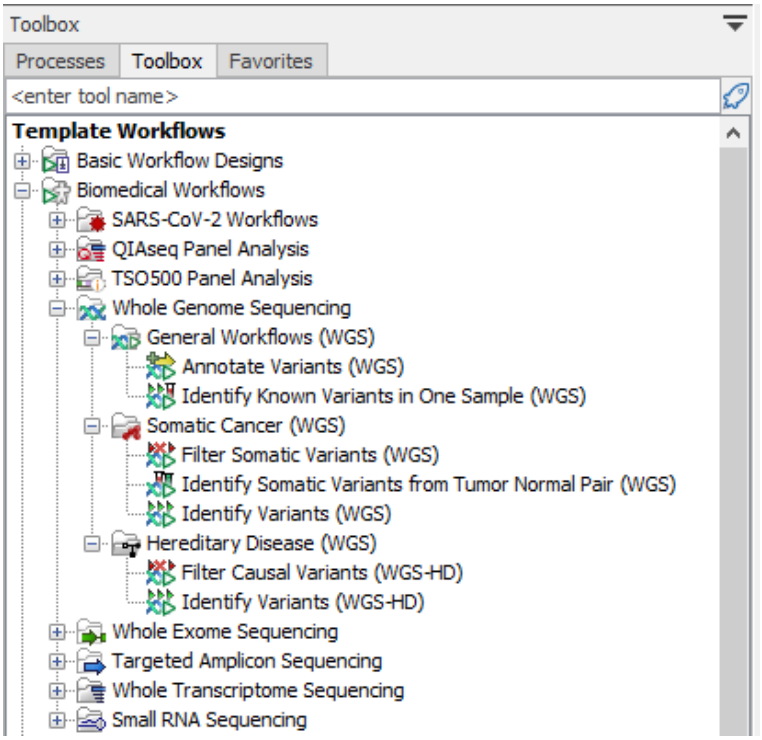


Figure 18.1: *Workflows available for analyzing whole genome sequencing data.*

18.1 General Workflows (WGS)

18.1.1 Annotate Variants (WGS)

The **Annotate Variants (WGS)** template workflow can add the following annotation types to a variant track, annotation track, expression track or statistical comparison track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.
- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP Common** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (indels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

Run the Annotate Variants (WGS) workflow

1. To run the **Annotate Variants (WGS)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** () | **Whole Genome Sequencing** () | **General Workflows (WGS)** () | **Annotate Variants (WGS)** ()

2. In the first wizard step, select the input variant track, annotation track, expression track or statistical comparison track to annotate (figure 18.2).

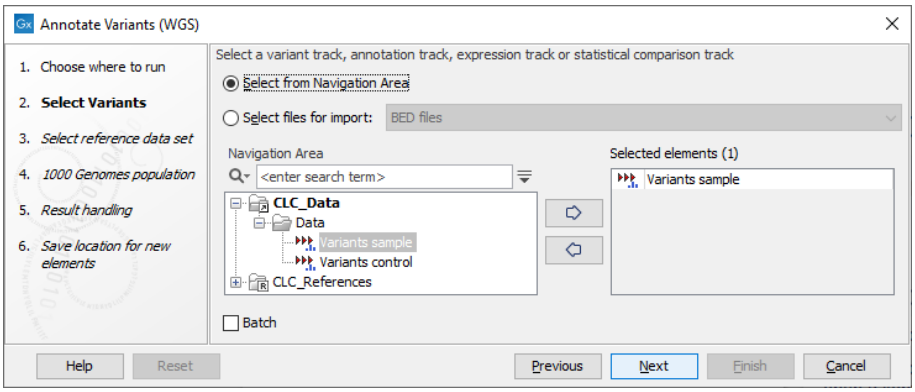


Figure 18.2: *Select the relevant track to annotate.*

3. In the next dialog, you have to select which data set should be used to annotate variants (figure 18.3).

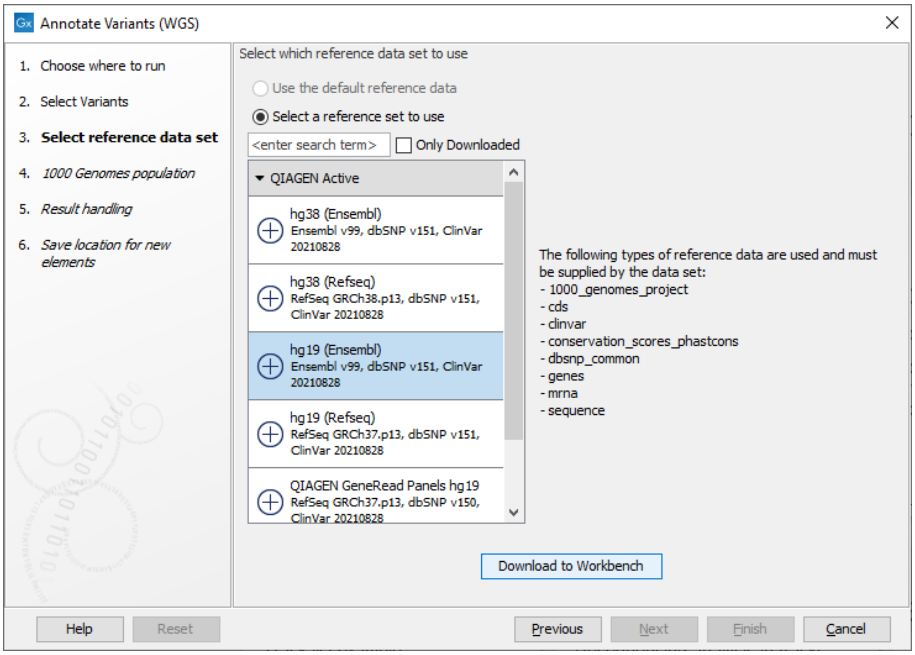


Figure 18.3: *Choose the relevant reference Data Set to annotate.*

4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 18.4).

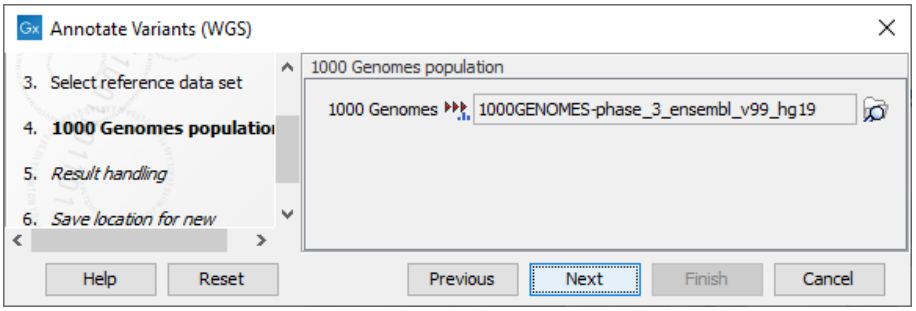


Figure 18.4: *Use the preselected 1000 Genomes population(s) or select another variant track.*

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Annotate Variants (WGS) workflow The outputs generated are:

1. **Filtered Annotated Variant Track** Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

3. **Track List Annotated Variants** A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP Common, ClinVar, 1000 Genomes, and PhastCons conservation scores (see figure 18.5).

It is possible to add tracks to the Track List by dragging the track directly from the **Navigation Area** to the Track List view. On the other hand, if you delete the annotated variant track, this track will also disappear from the Track List.

Open the annotated track as a table (see figure 18.6). The table and the Track List are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Track List view.

You may be met with a warning as shown in figure 18.7. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in ClinVar or dbSNP Common, the corresponding annotation column headers are missing from the result.

Adding information from other sources may help you identify interesting candidate variants for further research. For example, common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Furthermore, variants not found in the ClinVar database can

Figure 18.5: *The output from the Annotate Variants template workflow is a track list containing individual tracks for all added annotations.*

be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals in the region containing the variant can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) are prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

18.1.2 Identify Known Variants in One Sample (WGS)

The **Identify Known Variants in One Sample (WGS)** template workflow combines data analysis and interpretation. It should be used to identify known variants as specified by the user (e.g., known breast cancer associated variants) for their presence or absence in a sample. This workflow will not identify new variants.

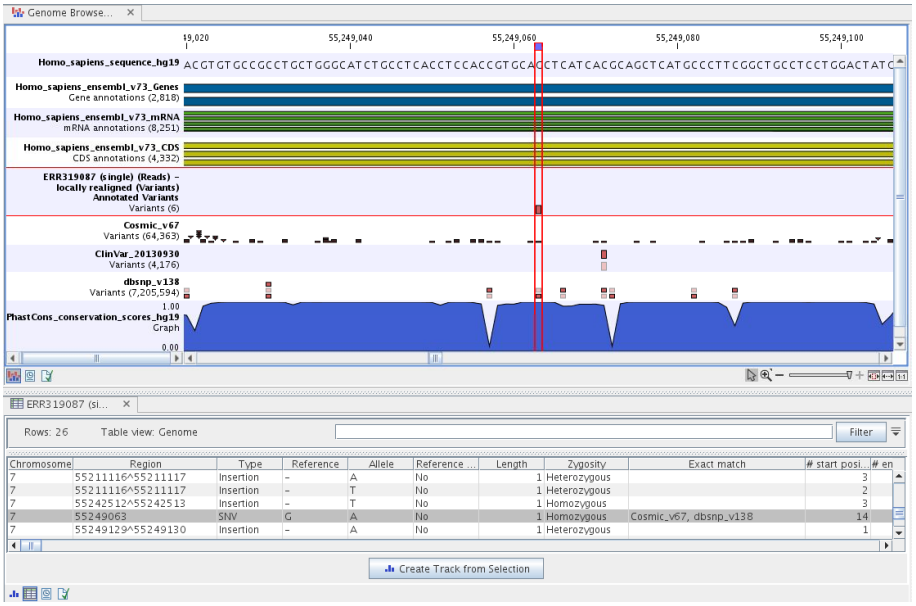


Figure 18.6: *The output from the Annotate Variants template workflow is a track list linked with the variant table view.*

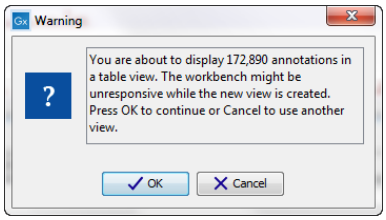


Figure 18.7: *Warning that appears when you work with tracks containing many annotations.*

The workflow maps the sequencing reads to a human genome sequence and does a local realignment of the mapped reads to improve the subsequent variant detection. In the next step, only variants specified by the user are identified and annotated in the newly generated read mapping.

Before starting the workflow, you may need to import your known variants in GVF or VCF format with the **Import | Tracks** tool (see `Import | Tracks`).

Run the Identify Known Variants in One Sample (WGS) workflow

1. To run the **Identify Known Variants in One Sample (WGS)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** () | **Whole Genome Sequencing** () | **General Workflows (WGS)** () | **Identify Known Variants from One Sample (WGS)** ()

2. First select the trimmed sequencing reads of the sample that should be tested for presence or absence of your known variants (figure 18.8).

   If several samples from different folders should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" and specifying the folders that hold the data you wish to analyse.

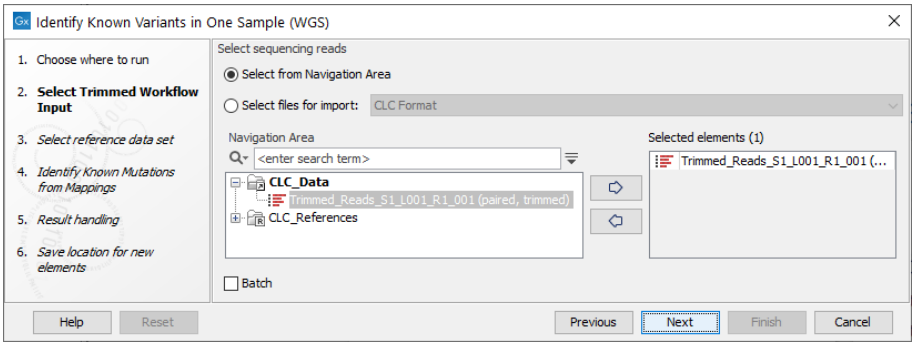


Figure 18.8: *Select the trimmed sequencing reads from the sample you would like to test for your known variants.*

3. In the next wizard step, select the reference data set should be used to identify the known variants (figure 18.9).

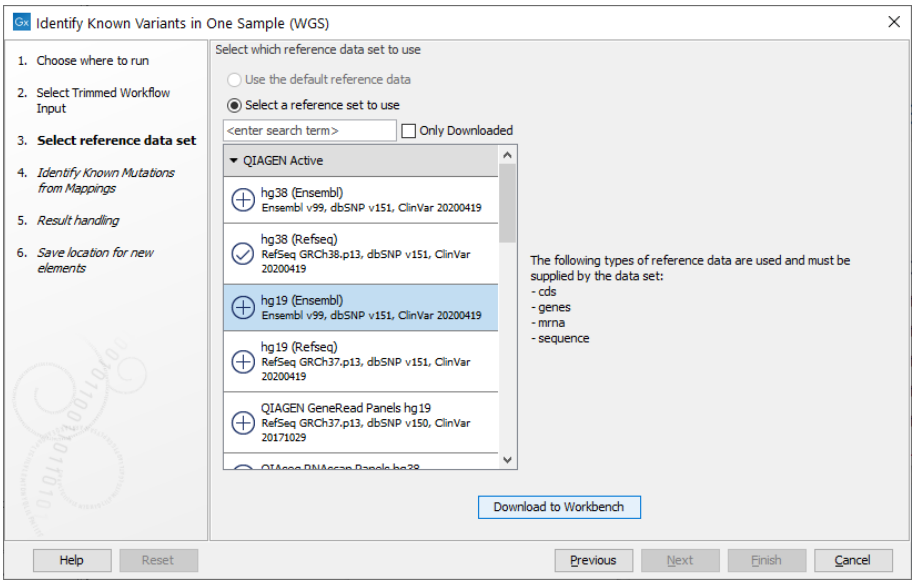


Figure 18.9: *Choose the relevant reference Data Set to identify the known variants.*

4. In the Identify Known Mutations from Mappings, select a variant track containing the known variants you want to identify in the sample (figure 18.10).

   The parameters that can be set are:

   - **Minimum coverage** The minimum number of reads that covers the position of the variant, which is required to set "Sufficient Coverage" to YES.
   - **Detection frequency** The minimum allele frequency that is required to annotate a variant as being present in the sample. The same threshold will also be used to determine if a variant is homozygous or heterozygous. In case the most frequent alternative allele at the position of the considered variant has a frequency of less than this value, the zygosity of the considered variant will be reported as being homozygous.

   The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or

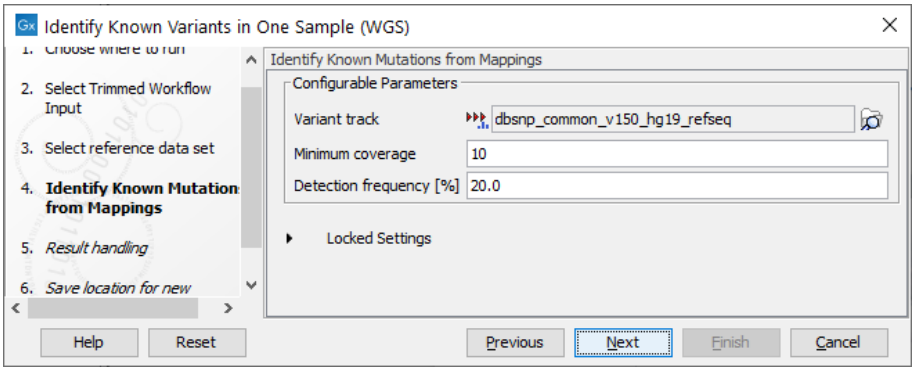


Figure 18.10: *Specify the track with the known variants that should be identified.*

not (observed frequency <= specified frequency). Moreover, it will determine if a variant should be labeled as heterozygous (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygous (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click **Finish**.

Output from the Identify Known Variants in One Sample (WGS) workflow The **Identify Known Variants in One Sample (WGS)** tool produces four different output types.

1. **Read Mapping Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads.

2. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

3. **Variants Detected in Detail** ( ) Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads, as well as detailed information about the Most Frequent Alternative Allele (labeled MFAA).

4. **Track List Identify Known Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

It is a good idea to start looking at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90% of the reads are mapped to the human reference sequence.

When this has been done you can open the Track List file (see 18.11).

The Track List includes the overview track of known variants and the detailed result track in the context to the human reference sequence, genes, transcripts, coding regions, targeted regions, and mapped sequencing reads.

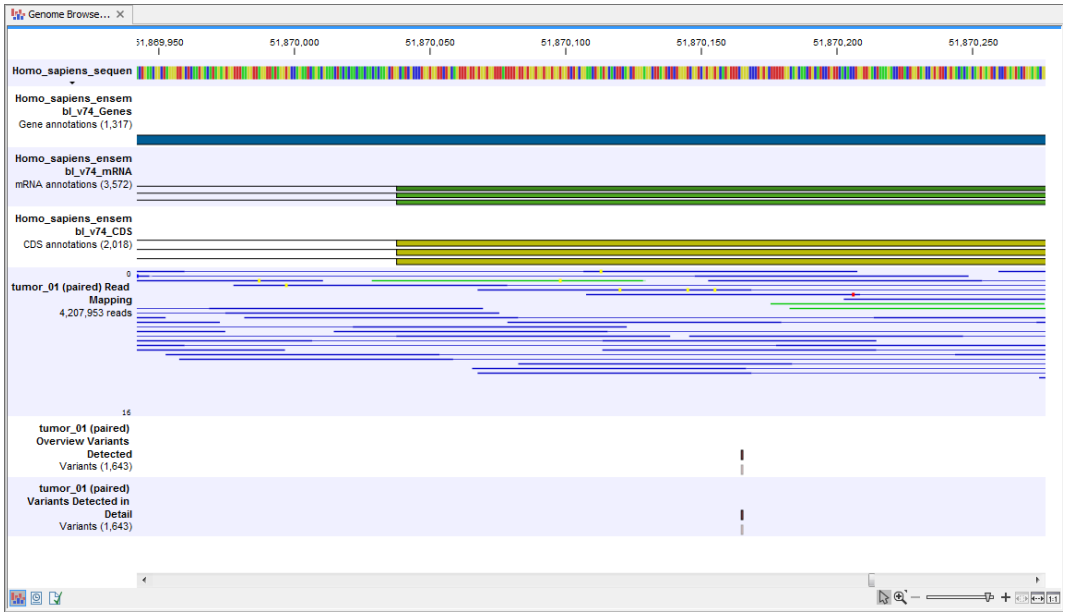
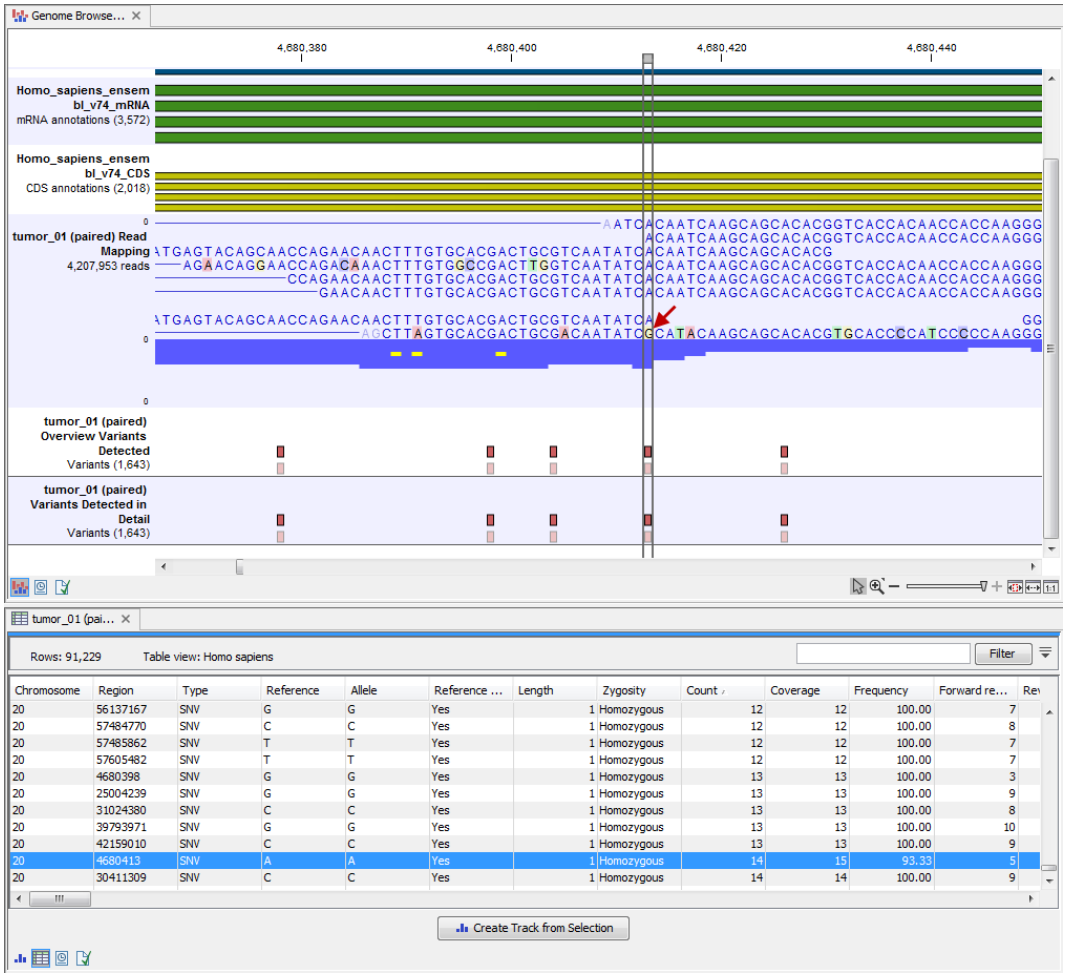


Figure 18.11: *Track List that allows inspection of the identified variants in the context of the human genome and external databases.*

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

Open the annotated variant as a table showing all variants and the added information/annotations (see 18.12).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Forward re...
20	56137167	SNV	G	G	Yes	1	Homozygous	12	12	100.00	7
20	57484770	SNV	C	C	Yes	1	Homozygous	12	12	100.00	8
20	57485862	SNV	T	T	Yes	1	Homozygous	12	12	100.00	7
20	57605482	SNV	T	T	Yes	1	Homozygous	12	12	100.00	7
20	4680398	SNV	G	G	Yes	1	Homozygous	13	13	100.00	3
20	25004239	SNV	G	G	Yes	1	Homozygous	13	13	100.00	9
20	31024380	SNV	C	C	Yes	1	Homozygous	13	13	100.00	8
20	39793971	SNV	G	G	Yes	1	Homozygous	13	13	100.00	10
20	42159010	SNV	C	C	Yes	1	Homozygous	13	13	100.00	9
20	4680413	SNV	A	A	Yes	1	Homozygous	14	15	93.33	5
20	30411309	SNV	C	C	Yes	1	Homozygous	14	14	100.00	9

Figure 18.12: *Track List with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

18.2 Somatic Cancer (WGS)

18.2.1 Filter Somatic Variants (WGS)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same subject, you can use the **Filter Somatic Variants (WGS)** template workflow to identify potential somatic variants. The purpose of this template workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same subject is available.

This workflow accepts variant tracks ( ) (e.g. the output from the Identify Variants template workflow) as input. Variants that are identical to the human reference sequence are first filtered away and then variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

Please note that this tool will likely also remove inherited cancer variants that are present at a

low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes, conservation scores and information from ClinVar (known variants with medical impact) and dbSNP (all known variants).

To run the **Filter Somatic Variants (WGS)** workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Genome Sequencing** ( ) | **Somatic Cancer** ( ) | **Filter Somatic Variants** ( )

1. Double-click on the **Filter Somatic Variants (WGS)** to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.

2. Next, you will be asked to select the variant track you would like to use for filtering somatic variants (figure 18.13).

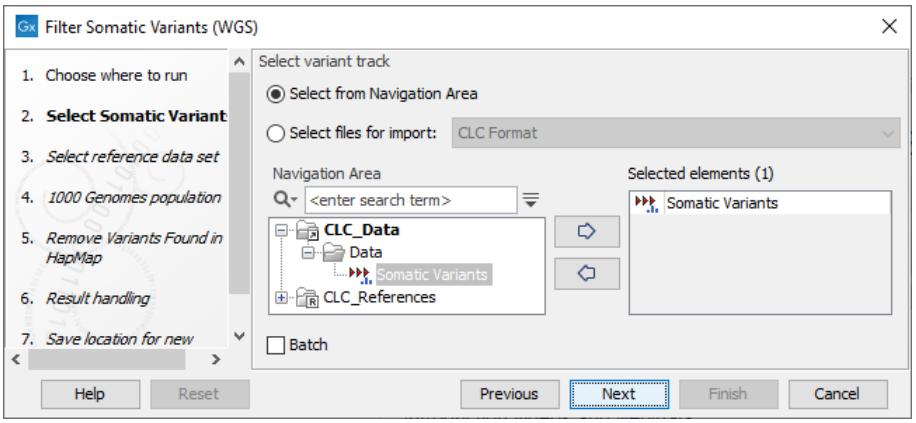


Figure 18.13: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter somatic variants (figure 18.14).

4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 18.15).

5. For databases that provide data from more than one population as HapMap does, the populations relevant to the data set can be specified. Click on the plus symbol ( ) and choose the population that matches the population your samples are derived from (figure 18.16). Please note that different populations are available and can be downloaded via the Reference Data Manager found in the top right corner of the CLC Workbench.

6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

7. Choose to **Save** your results and click **Finish**.

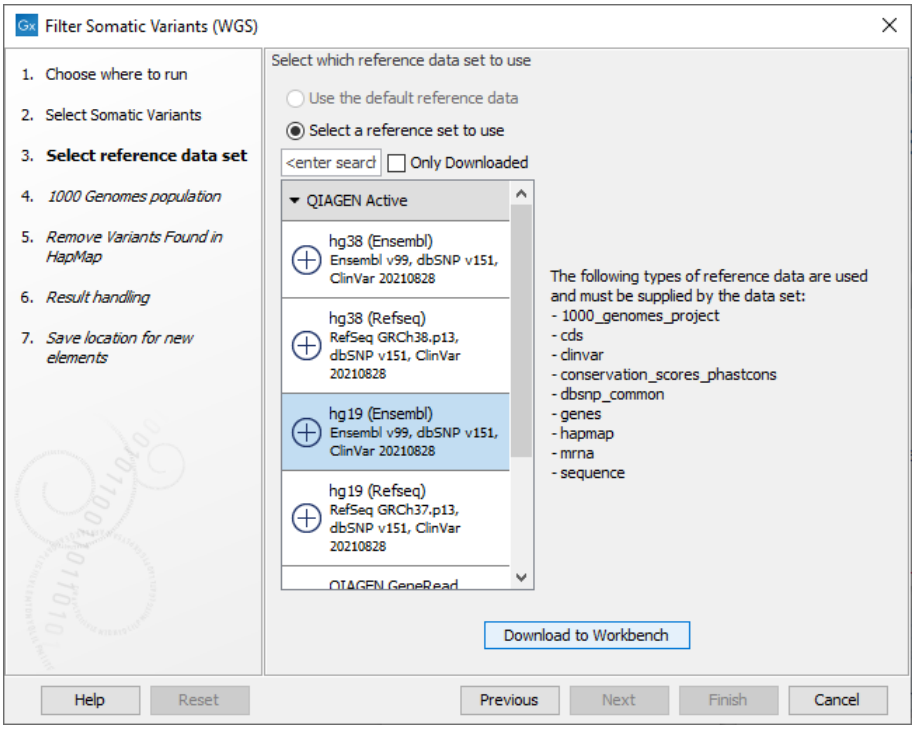


Figure 18.14: *Choose the relevant reference Data Set to annotate.*

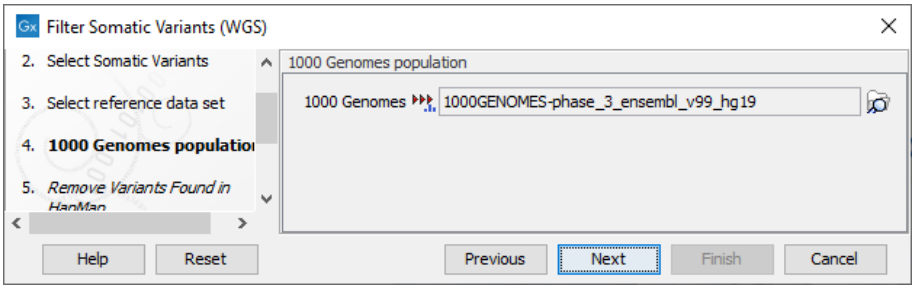


Figure 18.15: *Use the preselected 1000 Genomes population(s) or select another variant track.*

Output from the Filter Somatic Variants (WGS) workflow Two types of output are generated:

1. **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Track List. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.

2. **Track List Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, 1000 Genomes, and the PhastCons conservation scores (see figure 18.17).

The track with the conservation scores allows you to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant. Mapped sequencing reads as well as other tracks can be easily added to the Track List.

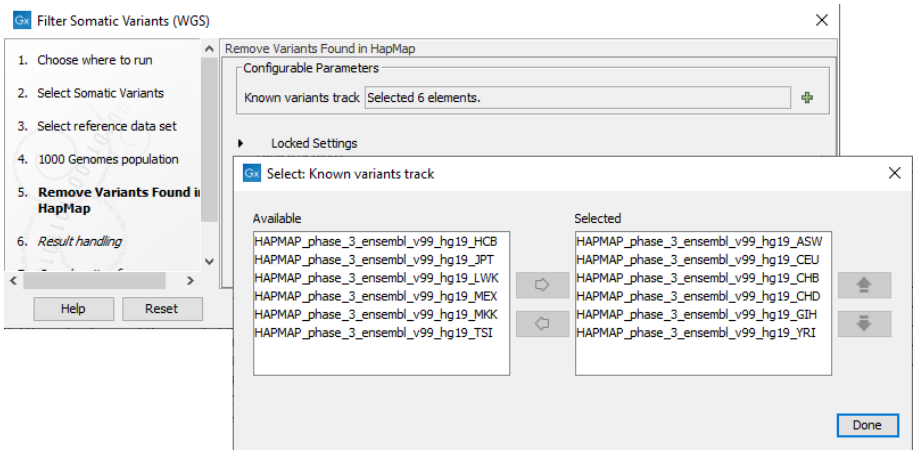


Figure 18.16: *Specify which HapMap population to use for filtering out known variants.*

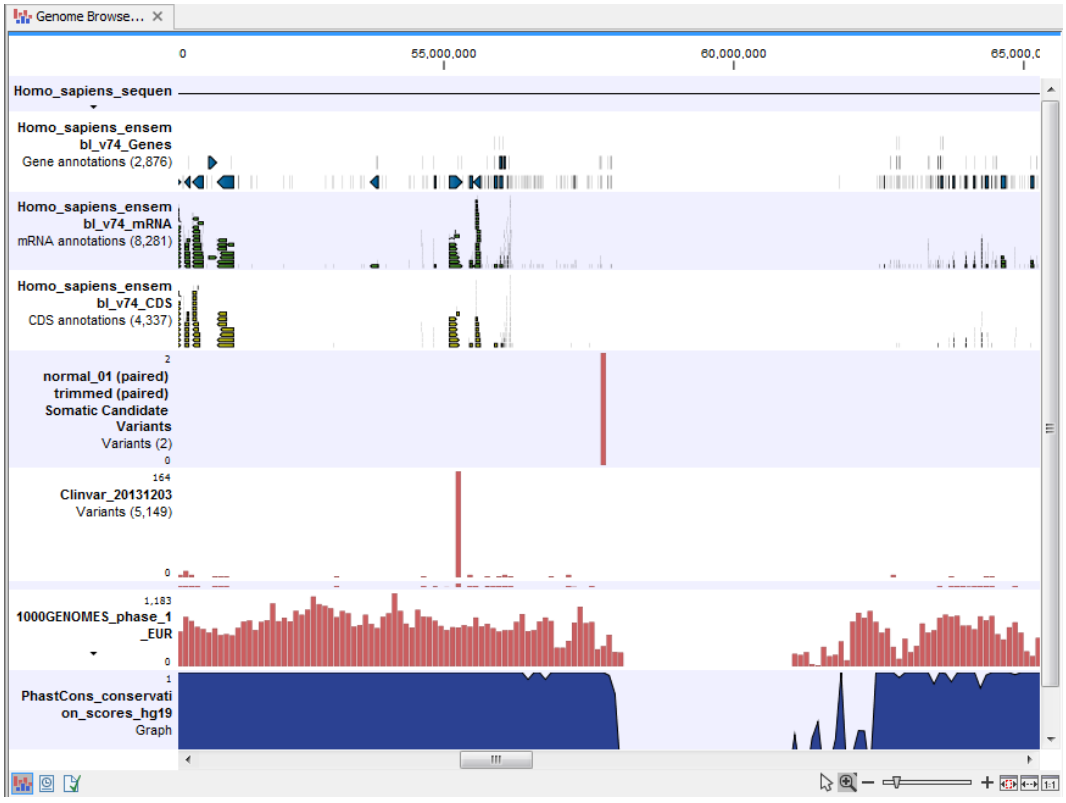


Figure 18.17: *The Track List showing the annotated somatic variants together with a range of other tracks.*

Open the variant track as a table showing all variants and the added information/annotations (see figure 18.18.

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid

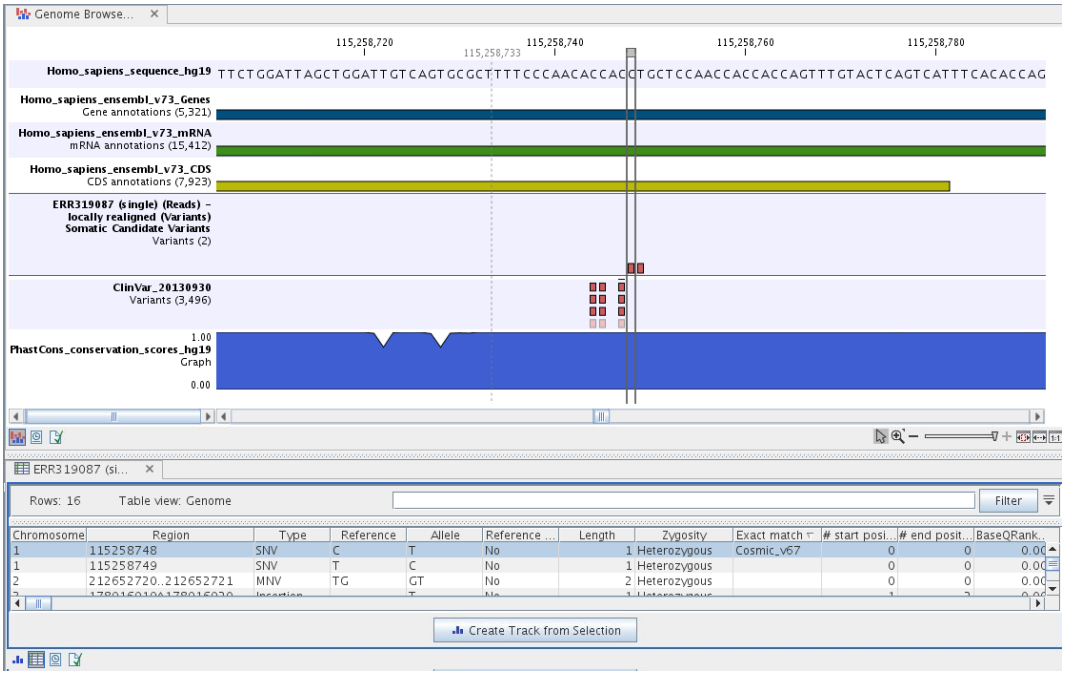


Figure 18.18: *The Track List showing the annotated somatic variants together with a range of other tracks.*

level.

A high conservation level, between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

18.2.2 Identify Somatic Variants from Tumor Normal Pair (WGS)

The **Identify Somatic Variants from Tumor Normal Pair (WGS)** template workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same subject.

When running this workflow the trimmed reads are mapped and variants identified. Germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads. Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from relevant databases like ClinVar (variants with clinically relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

Run the Identify Somatic Variants from Tumor Normal Pair (WGS) workflow

1. To run the **Identify Somatic Variants from Tumor Normal Pair (WGS)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** () | **Whole Genome Sequencing** () | **Somatic Cancer** () | **Identify Somatic Variants from Tumor Normal Pair (WGS)** ()

2. First (figure 18.19), select the trimmed tumor sample reads.

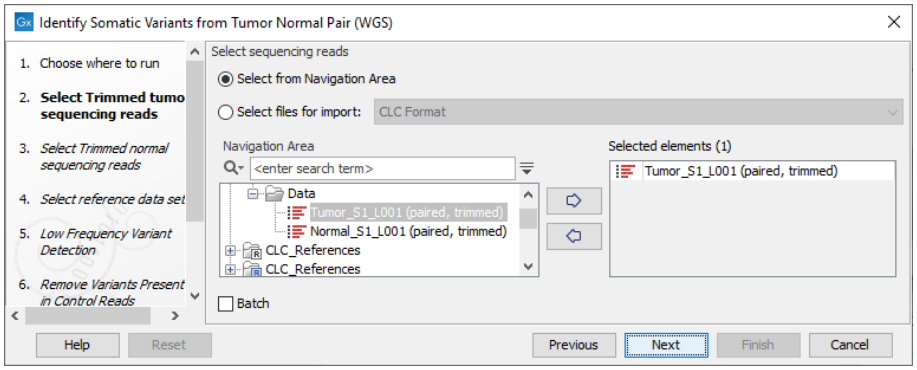


Figure 18.19: *Select the trimmed tumor sample reads.*

3. In the next wizard step, specify the trimmed normal sample reads.

4. In the next dialog, select which reference Data Set should be used to identify variants (figure 18.20).

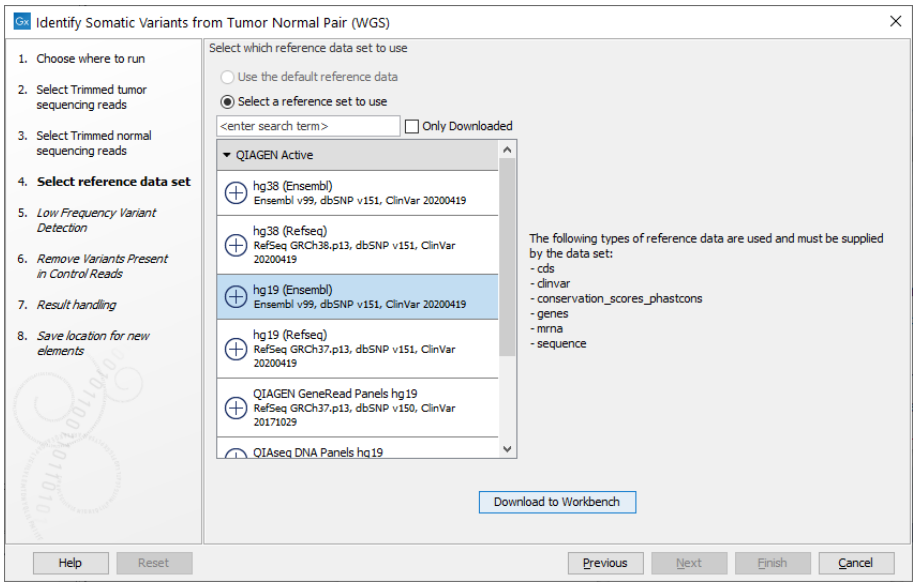


Figure 18.20: *Choose the relevant reference Data Set to identify variants.*

5. In the next wizard step you can adjust the settings used for variant detection (figure 18.21).

   For a description of the different parameters that can be adjusted, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_

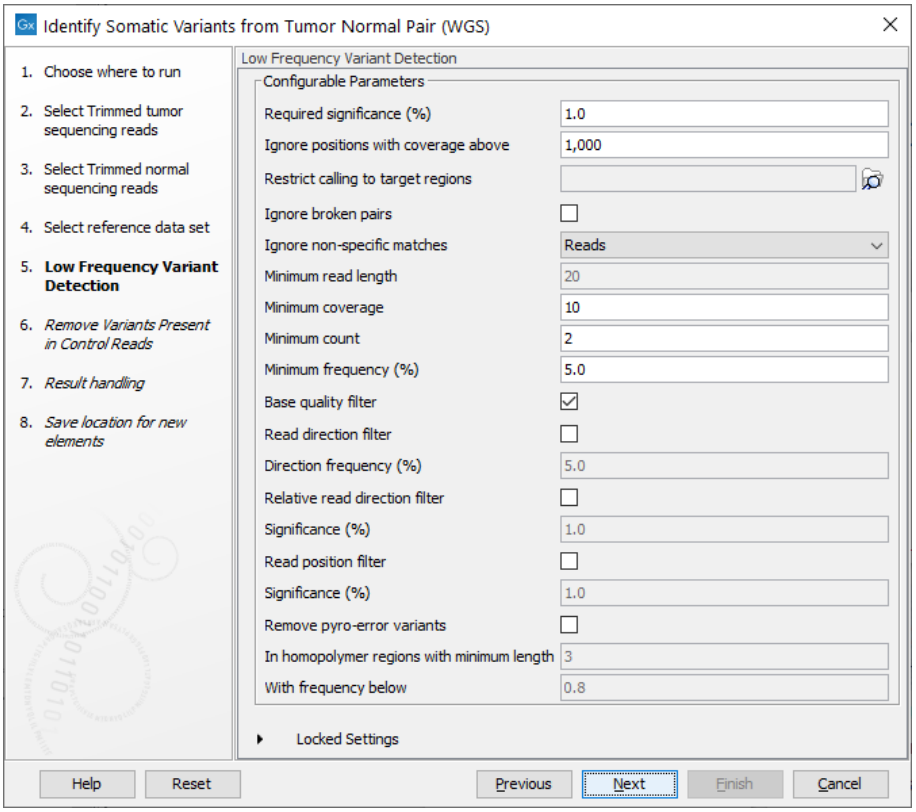


Figure 18.21: *Specify the settings for the variant detection.*

`Variant_Detection.html`. If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the template workflow.

6. In the Remove Variants Present in Control Reads step, you can adjust the settings for removal of germline variants (figure 18.22).

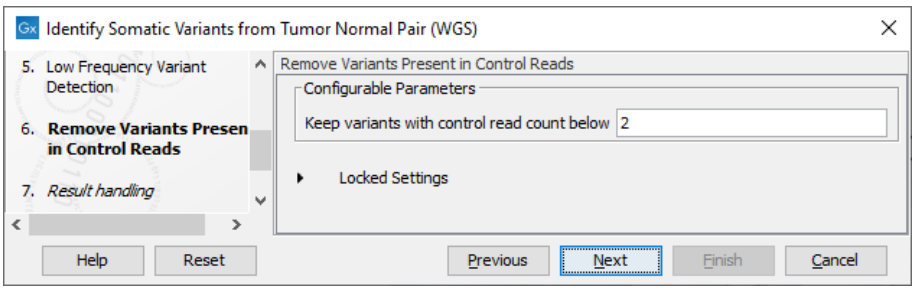


Figure 18.22: *Specify setting for removal of germline variants.*

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click **Finish**.

Output from the Identify Somatic Variants from Tumor Normal Pair (WGS) workflow The following outputs are generated:

1. **Read Mapping Tumor** ( ) The mapped sequencing reads for the tumor sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).
2. **Read Mapping Normal** ( ) The mapped sequencing reads for the normal sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously.
3. **Mapping Report Tumor** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.
4. **Mapping Report Normal** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the normal sample.
5. **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
6. **Annotated Somatic Variants** ( ) A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
7. **Track List Tumor Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar database, and finally a track showing the conservation score (see figure 18.23).

18.2.3 Identify Variants (WGS)

The **Identify Variants (WGS)** template workflow takes trimmed sequencing reads as input and returns identified variants in a Track List.

Sequencing reads provided as input are initially mapped to the human reference sequence. The resulting read mapping is analyzed by the Structural Variant Caller to infer indels and other structural variants from unaligned end read patterns. Subsequently, the mapping is realigned, guided by the indels detected by the Structural Variant Caller. The locally realigned read mapping is analyzed by the Low Frequency Variant Detection tool. The Low Frequency Variant Detection tool produces a track of unfiltered variants; these are post-filtered to remove variants that are likely due to artifacts or noise. The variants called by the Low Frequency Variant Detection tool that pass the post filtering criteria can be found in the Identified variants track. Variants inferred by the Structural Variant Caller, and not detected by the Low Frequency Variant Detection tool, are also subjected to a number of post filters; those that pass the post filter criteria can be found in the Indels indirect evidence track.

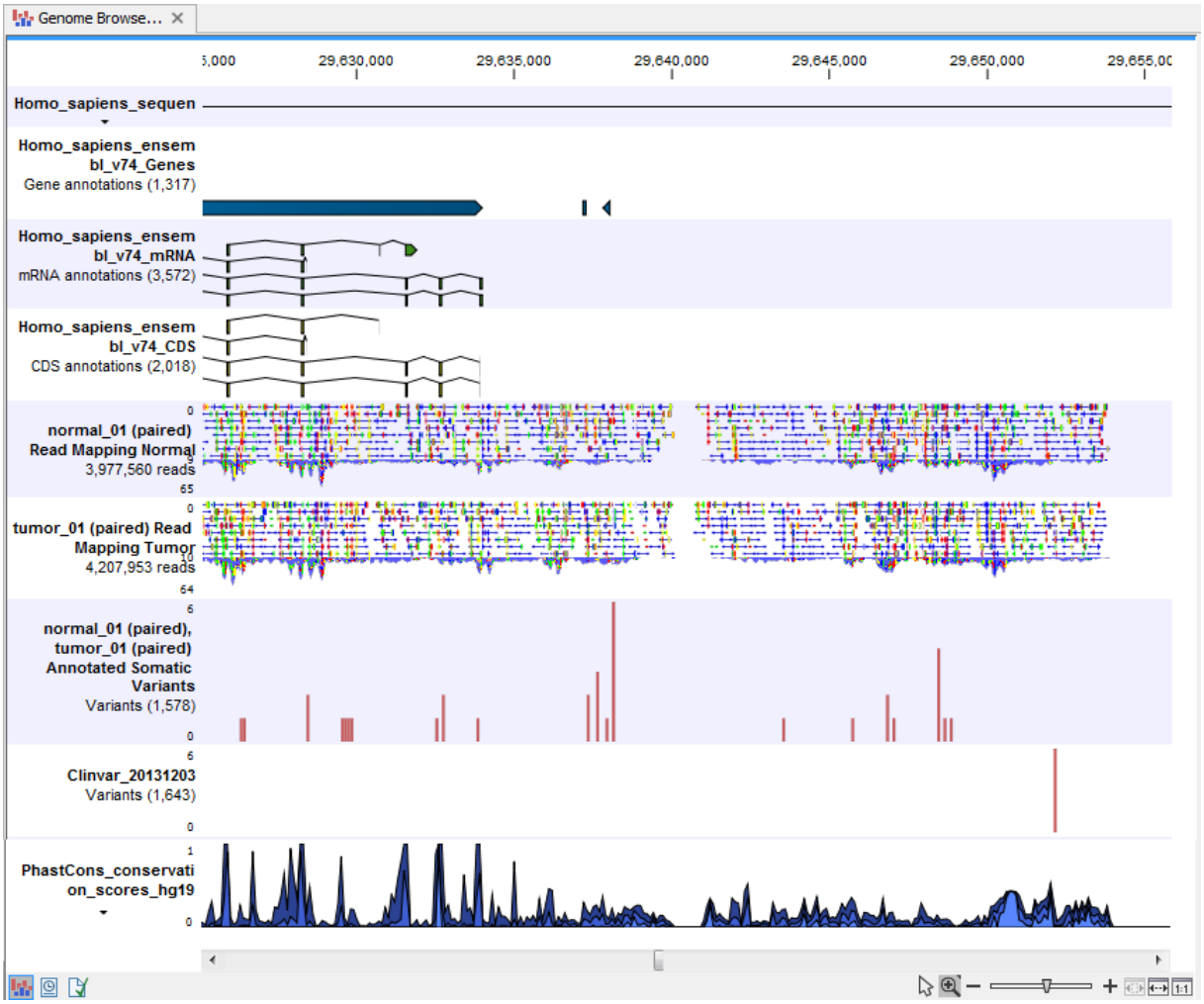


Figure 18.23: *The Track List presents all the different data tracks together and makes it easy to compare different tracks.*

A detailed mapping report is created with summaries on the mapping and coverage.

Run the Identify Variants (WGS) workflow

1. To run the **Identify Variants (WGS)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Genome Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Variants (WGS)** ( )

2. Select the trimmed sequencing reads from the sample that should be analyzed (figure 18.24).

   If several samples should be analyzed, the tool has to be run in batch mode. This is done by checking "Batch" and selecting the **folder** that holds the data you wish to analyze.

3. In the next dialog, you have to select which reference data set should be used to identify variants (figure 18.25).

4. In the Low Frequency Variant Detection dialog (figure 18.26), you can specify the parameters

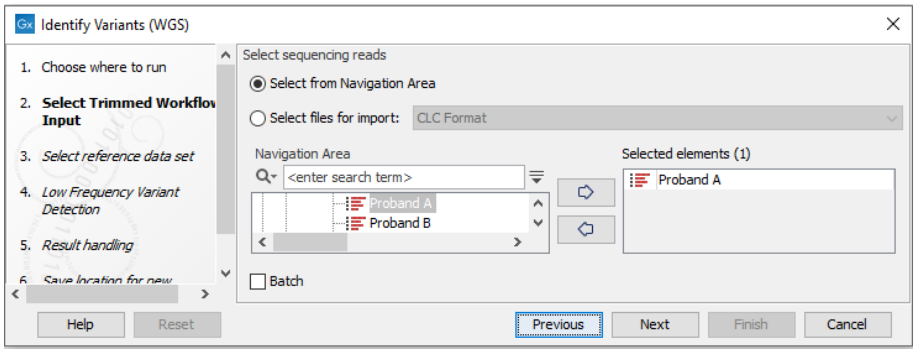


Figure 18.24: *Please select trimmed sequencing reads from the sample to be analyzed.*

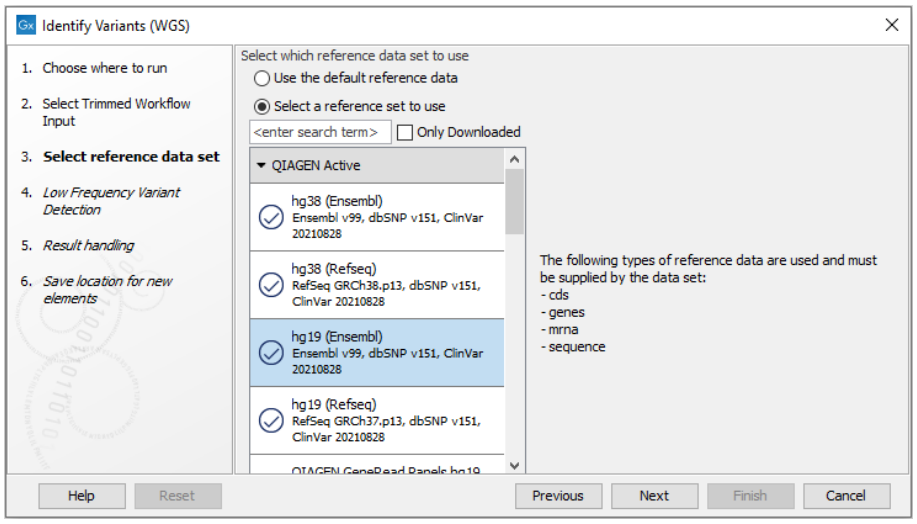


Figure 18.25: *Choose the relevant reference Data Set to identify variants in your sample.*

for variant detection.

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click **Finish**.

Output from the Identify Variants (WGS) workflow

The **Identify Variants (WGS)** tool produces the following outputs:

1. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

2. **Read Mapping Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads.

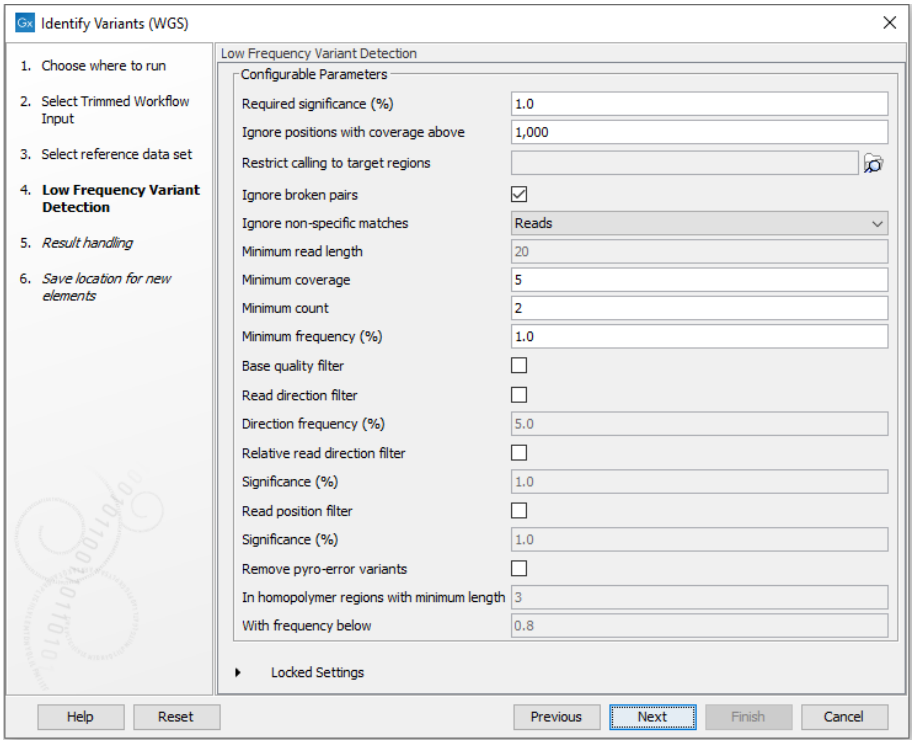


Figure 18.26: *Specify the parameters that should be used to detect variants.*

3. Two variant tracks ( ): The **Identified Variants** track containing the variants identified by the Low Frequency Variant Detection tool after the post-filtering has been applied, and the **Indels indirect evidence** track which contains the indels inferred by the Structural Variant Caller. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

4. **Genome Browser View** ( ) A Genome Browser view containing the collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the Indels indirect evidence variants (see figure 18.5).

Before looking at the identified variants, we recommend that you first take a look at the mapping report to see the performance of the mapping. E.g., check that at least 90% of the reads map to the human reference sequence.

Next, open the Genome Browser View (see figure 18.27). It lists the track of the identified variants in context to the human reference sequence, genes, transcripts, coding regions, and mapped sequencing reads.

By double-clicking on the Indels indirect evidence variant track in the Genome Browser View, a table will be shown that lists all inferred larger insertions and deletions (see figure 18.28).

Figure 18.27: *The Genome Browser View allows easy inspection of the identified variants in the context of the human genome.*

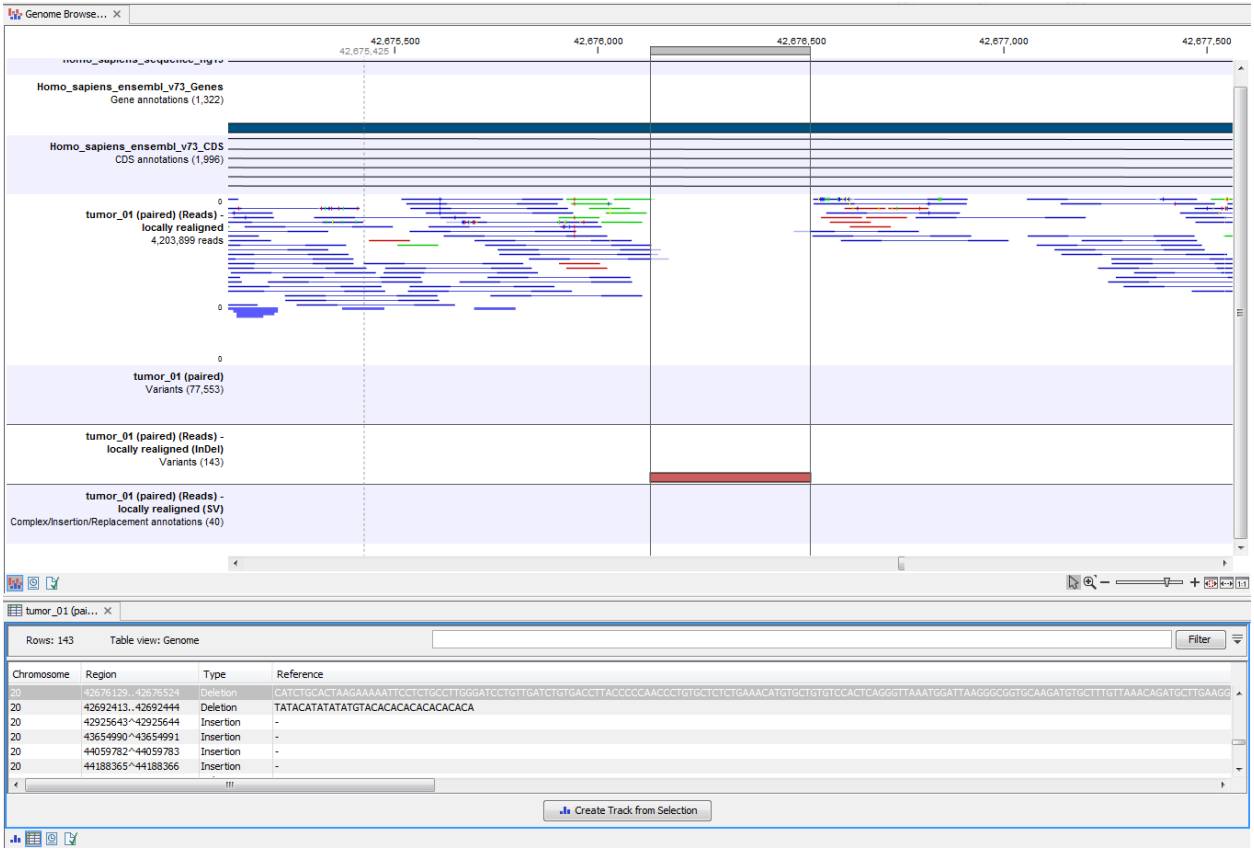


Figure 18.28: *This figure shows a Genome Browser View with an open track table. The table allows deeper inspection of the identified variants.*

18.3 Hereditary Disease (WGS)

18.3.1 Filter Causal Variants (WGS-HD)

If you are analyzing a list of variants, you can use the **Filter Causal Variants (WGS-HD)** template workflow to remove variants that are outside the target region, as well as common variants present in publicly available databases. The workflow will annotate the remaining variants with gene names, conservation scores, and information from relevant databases.

Run the Filter Causal Variants (WGS-HD) workflow

To run the **Filter Causal Variants (WGS-HD)** workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Genome Sequencing** ( ) | **Hereditary Disease** ( ) | **Filter Causal Variants (WGS-HD)** ( )

1. Double-click on the workflow name to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Select the **variant track** you want to use for filtering causal variants (figure 18.29).

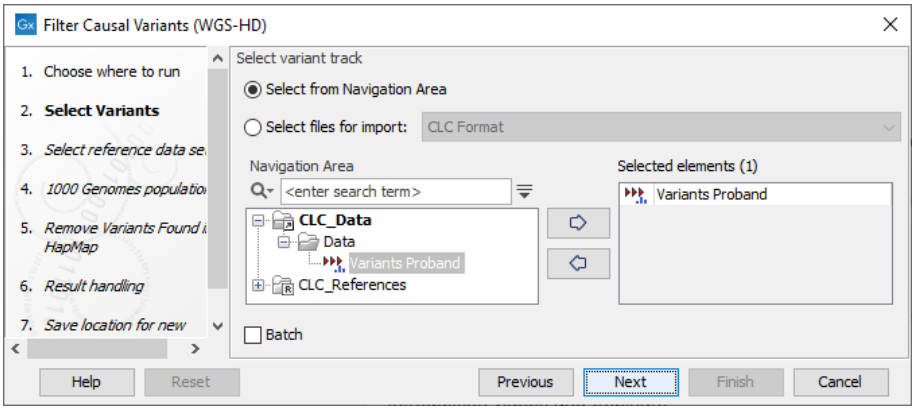


Figure 18.29: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter causal variants (figure 18.30).
4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track (figure 18.31).
5. Specify the **Hapmap populations** that should be used for **filtering out** variants found in Hapmap (figure 18.32).
6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.
7. Choose to **Save** your results and click on the button labeled **Finish**.

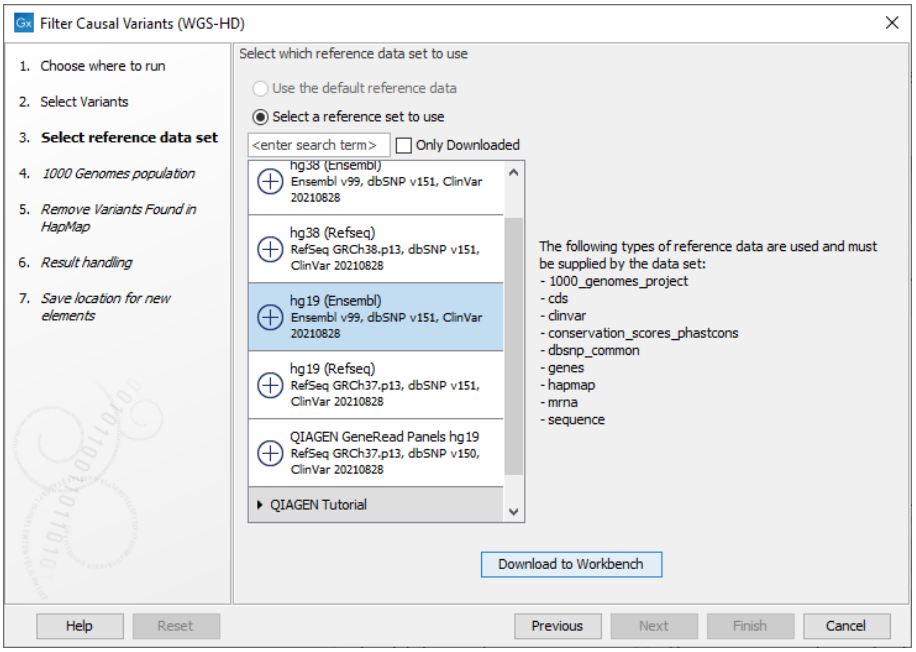


Figure 18.30: *Choose the relevant reference Data Set to annotate.*

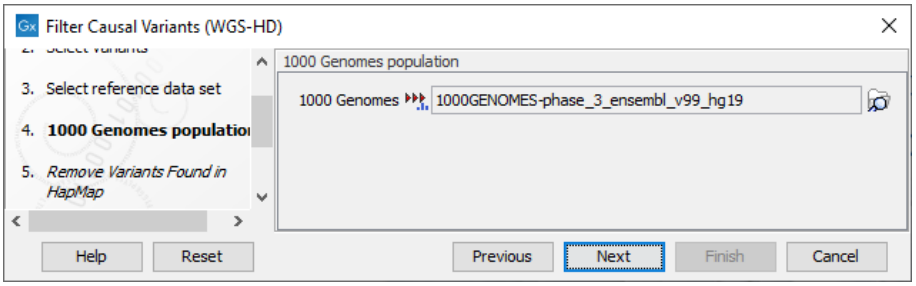


Figure 18.31: *Use the preselected 1000 Genomes population(s) or select another variant track.*

Output from the Filter Causal Variants (WGS-HD) workflow

The following outputs are generated:

- An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
- A **Track List** A collection of the tracks presented together.
- A **Filtered Variant Track** Shows all reported variants for this sample.

18.3.2 Identify Variants (WGS-HD)

The **Identify Variants (WGS-HD)** template workflow takes trimmed sequencing reads as input and returns identified variants in a Track List.

Sequencing reads provided as input are initially mapped to the human reference sequence. The resulting read mapping is analyzed by the Structural Variant Caller to infer indels and other

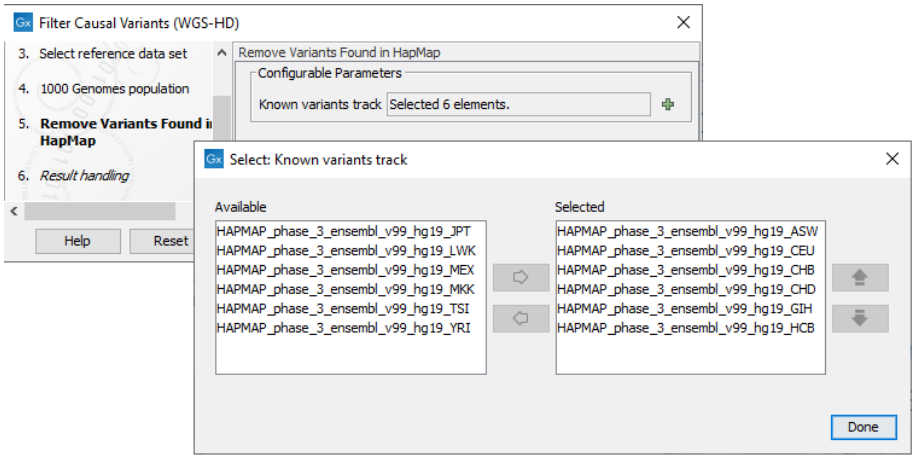


Figure 18.32: *Select the relevant Hapmap population(s).*

structural variants from unaligned end read patterns. Subsequently, the mapping is realigned, guided by the indels detected by the Structural Variant Caller. The locally realigned read mapping is analyzed by the Fixed Ploidy Variant Detection tool. The Fixed Ploidy Variant Detection tool produces a track of unfiltered variants; these are post-filtered to remove variants that are likely due to artifacts or noise. The variants called by the Fixed Ploidy Variant Detection tool that pass the post filtering criteria can be found in the Identified variants track. Variants inferred by the Structural Variant Caller, and not detected by the Fixed Ploidy Variant Detection tool, are also subjected to a number of post filters; those that pass the post filter criteria can be found in the Indels indirect evidence track.

A detailed mapping report is created with summaries on the mapping and coverage.

Run the Identify Variants (WGS-HD) workflow

1. To run the **Identify Variants (WGS-HD)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Genome Sequencing** ( ) | **Hereditary Disease** ( ) | **Identify Variants (WGS-HD)** ( )

2. Select the trimmed sequencing reads you want to analyze (figure 18.33).

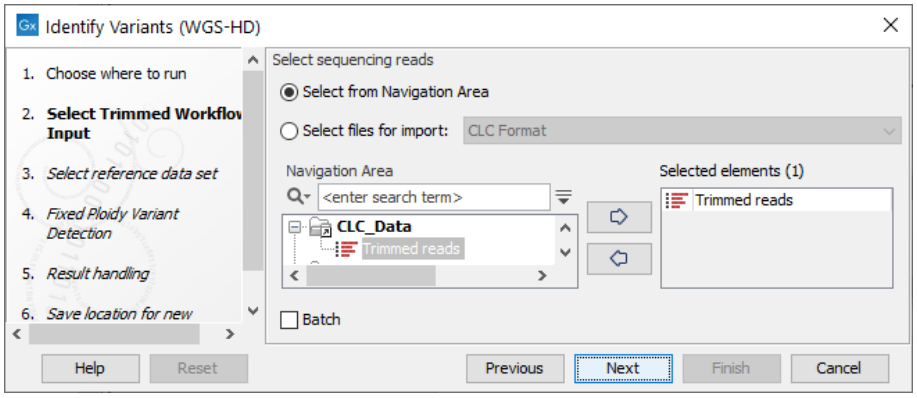


Figure 18.33: *Specify the trimmed sequencing reads for the sample.*

3. In the next dialog, you have to select which reference data set should be used for the analysis (figure 18.34).

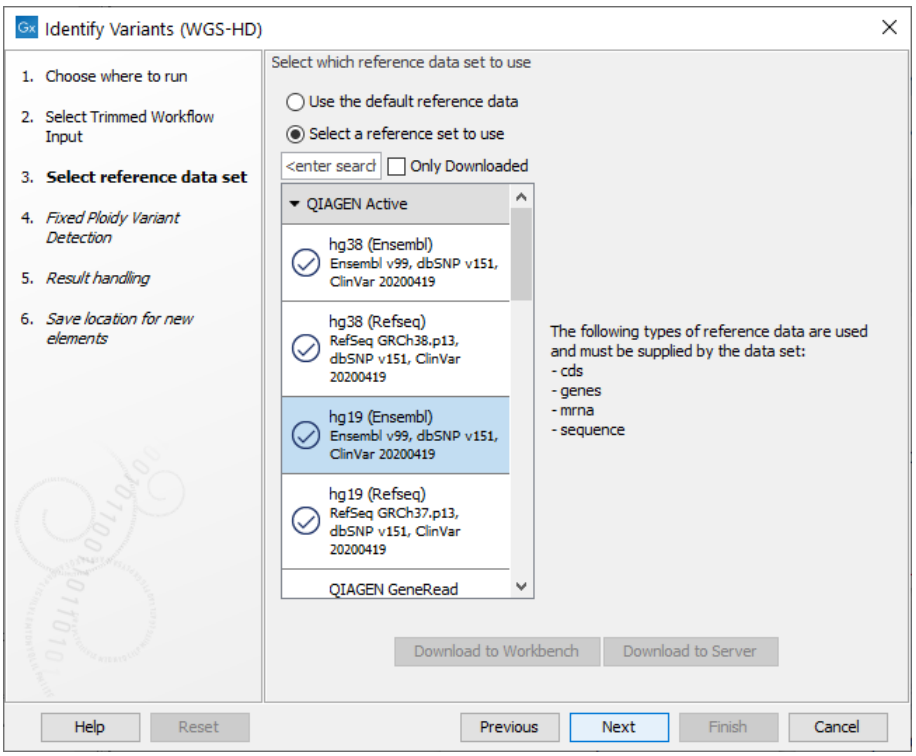


Figure 18.34: *Choose the relevant reference Data Set to identify variants.*

4. Specify the parameters for the Fixed Ploidy Variant Detection tool (figure 18.35).

   The parameters used by the Fixed Ploidy Variant Detection tool can be adjusted. We have optimized the parameters to the individual analyses, but you may want to tweak some of the parameters to fit your particular sequencing data. A good starting point could be to run an analysis with the default settings.

   The parameters that can be set are:

   - **Required variant probability** is the minimum probability value of the 'variant site' required for the variant to be called. Note that it is not the minimum value of the probability of the individual variant. For the Fixed Ploidy Variant detector, if a variant site - and not the variant itself - passes the variant probability threshold, then the variant with the highest probability at that site will be reported even if the probability of that particular variant might be less than the threshold. For example if the required variant probability is set to 0.9 then the individual probability of the variant called might be less than 0.9 as long as the probability of the entire variant site is greater than 0.9.
   - **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored

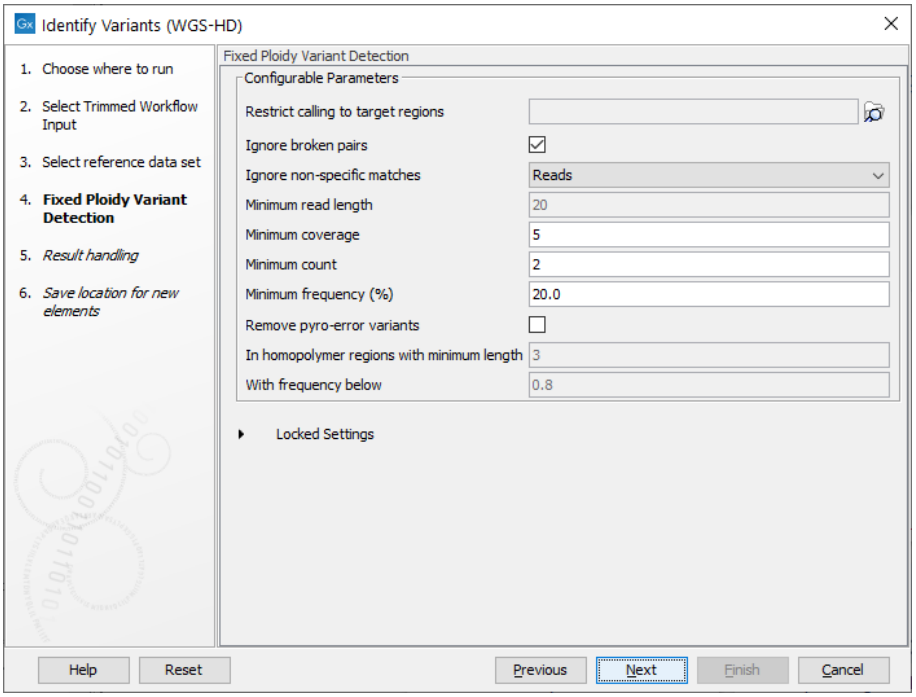


Figure 18.35: *Specify the parameters for the Fixed Ploidy Variant Detection tool.*

some true variants may go undetected. Please note that ignored broken pair reads will not be considered for any non-specific match filters.

- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Identify Variants (WGS-HD) workflow

The following outputs are generated:

- **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

- **Read Mapping Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads.

- Two variant tracks ( ): The **Identified Variants** track containing the variants identified by the Fixed Ploidy Variant Detection tool after the post-filtering has been applied, and the **Indels indirect evidence** track which contains the indels inferred by the Structural Variant Caller. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

- **Genome Browser View** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 18.5).

Chapter 19

Whole exome sequencing (WES)

Contents

The protein coding part of the human genome accounts for around 1 % of the genome and consists of around 180,000 exons covering an area of ~30 megabases (Mb) [Ng et al., 2009]. By targeting sequencing to only the protein coding parts of the genome, exome sequencing is a cost efficient way of generating sequencing data that is believed to harbor the vast majority of the disease-causing mutations [Choi et al., 2009].

A number of template workflows are available for analysis of whole genome sequencing data (figure 19.1). The concept of the pre-installed template workflows is that read data are used as input in one end of the workflow and in the other end of the workflow you get a track list and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual template workflows can be used for and go through step by step how to run the workflows.

Remember you will have to prepare data with the **Prepare Raw Data** workflow described in section 2 before you proceed to running any of these workflows.

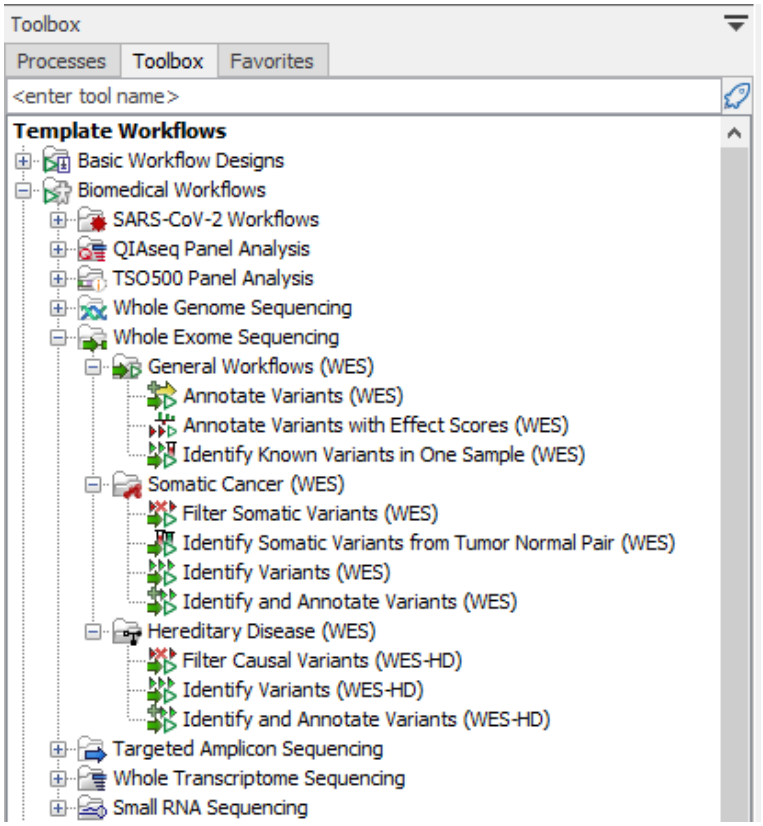


Figure 19.1: *The workflows available for analyzing whole exome sequencing data.*

19.1 General Workflows (WES)

19.1.1 Annotate Variants (WES)

The **Annotate Variants (WES)** template workflow can add the following annotation types to a variant track, annotation track, expression track or statistical comparison track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.
- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP Common** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (indels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

Run the Annotate Variants (WES) workflow

1. To run the **Annotate Variants (WES)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **General Workflows (WES)** ( ) | **Annotate Variants (WES)** ( )

2. In the first wizard step, select the input variant track, annotation track, expression track or statistical comparison track to annotate (figure 19.2).

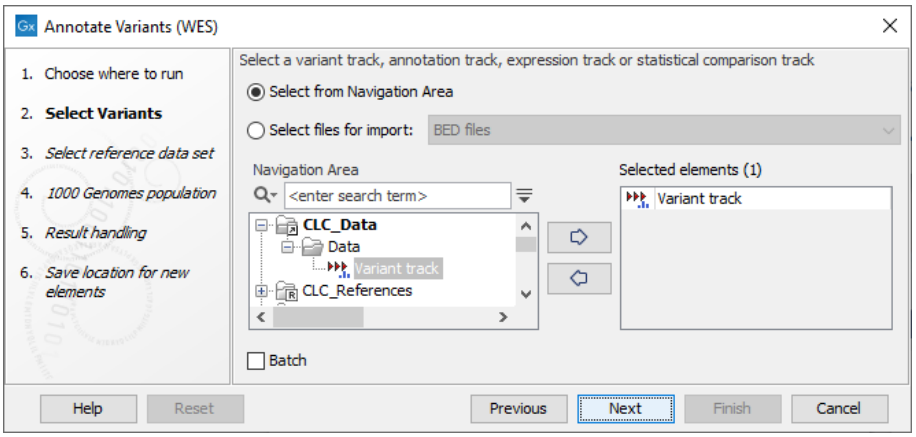


Figure 19.2: *Select the relevant track to annotate.*

3. In the next dialog, you have to select which data set should be used to annotate variants (figure 19.3).

4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 19.4).

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Annotate Variants (WES) workflow

The output generated are:

1. **Filtered Annotated Variant Track** ( ) Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

3. **Genome Browser View Annotated Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes,

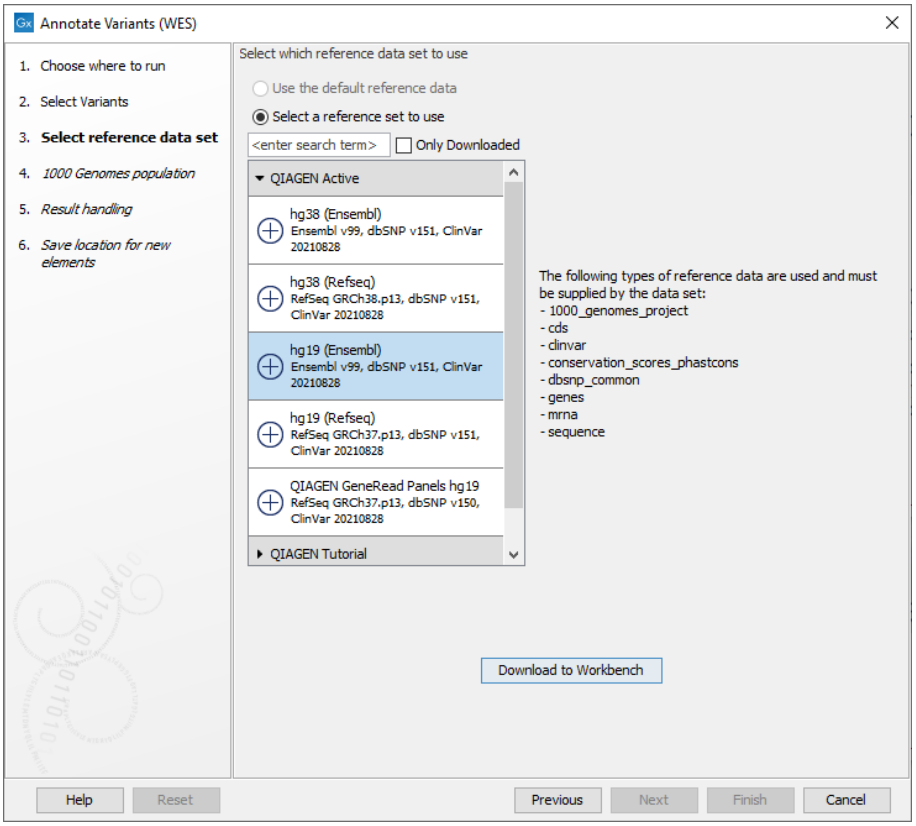


Figure 19.3: *Choose the relevant reference Data Set to annotate.*

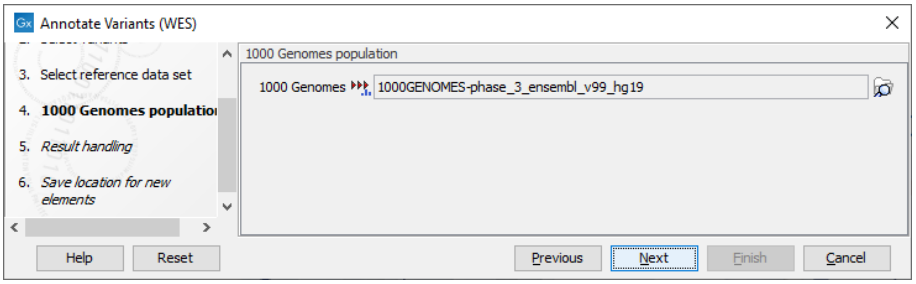


Figure 19.4: *Use the preselected 1000 Genomes population(s) or select another variant track.*

transcripts, coding regions, and variants detected in dbSNP Common, ClinVar, 1000 Genomes, and PhastCons conservation scores (see figure 19.5).

It is possible to add tracks to the Genome Browser View by dragging the track directly from the **Navigation Area** to the Genome Browser View. On the other hand, if you delete the annotated variant track, this track will also disappear from the Genome Browser View.

Open the annotated track as a table (see figure 19.6). The table and the Genome Browser View are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Genome Browser View.

You may be met with a warning as shown in figure 19.7. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in ClinVar or dbSNP

Figure 19.5: *The output from the Annotate Variants template workflow is a track list containing individual tracks for all added annotations.*

Common, the corresponding annotation column headers are missing from the result.

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar database, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments

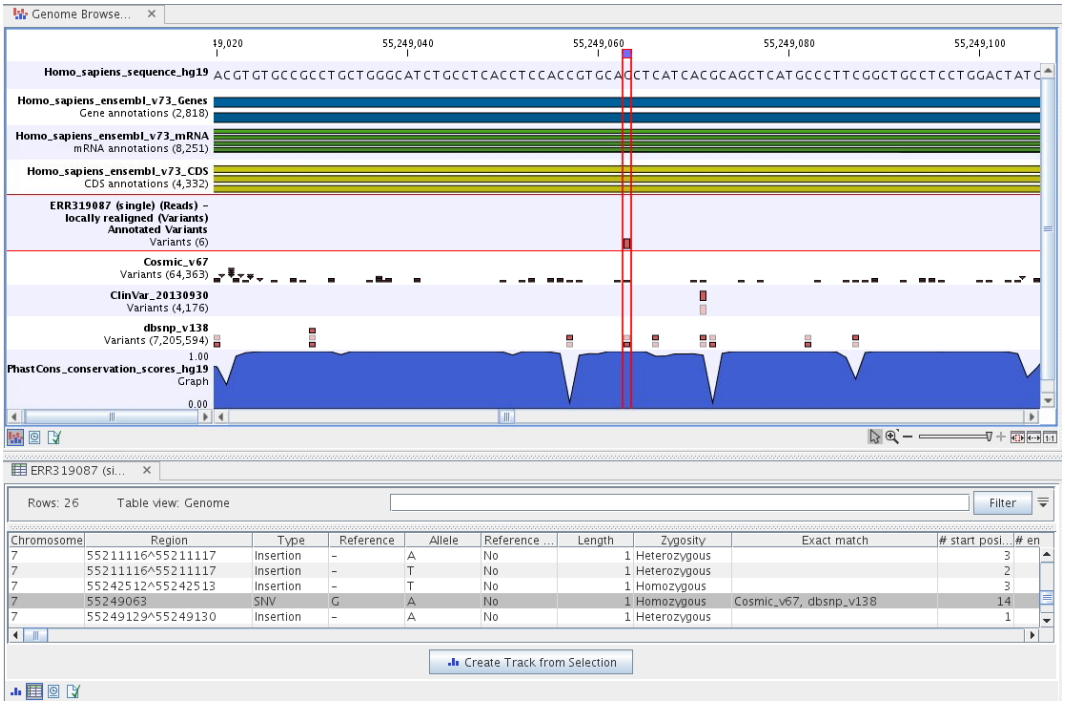


Figure 19.6: *The output from the Annotate Variants template workflow is a Genome Browser View linked with the variant table view.*

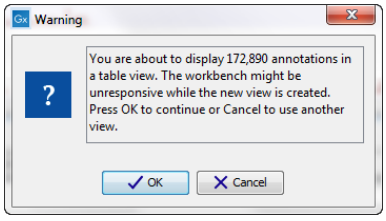


Figure 19.7: *Warning that appears when you work with tracks containing many annotations.*

where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

19.1.2 Annotate Variants with Effect Scores (WES)

The **Annotate Variants with Effect Scores (WES)** template workflow takes a variant track as input, and outputs a variant track with effect scores annotations added to non-reference SNVs.

See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Annotate_with_Effect_Scores.html for information about the underlying tool used to add these annotations.

Run the Annotate Variants with Effect Scores (WES) workflow

1. To run the **Annotate Variants with Effect Scores (WES)** template workflow, go to:

   Template Workflows | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **General Workflows (WES)** ( ) | **Annotate Variants with Effect Scores (WES)** ( )

2. In the first wizard step (figure 19.8), select the input variant track.

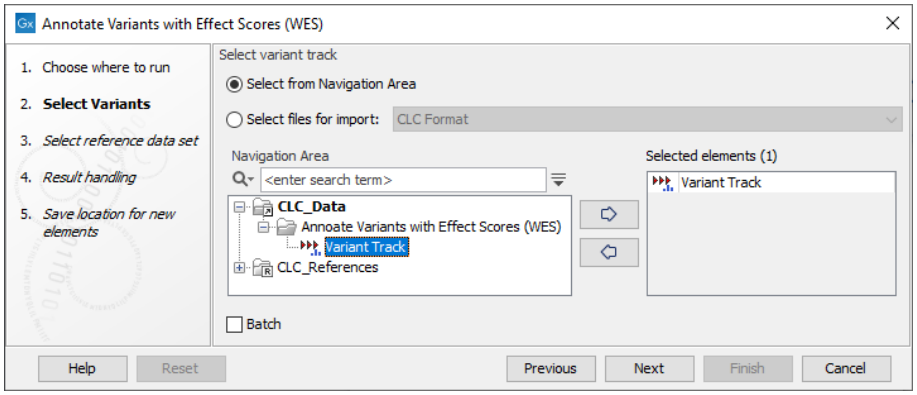


Figure 19.8: *Select the relevant variant track to annotate.*

3. In the next wizard step (figure 19.9), select the effect score data.

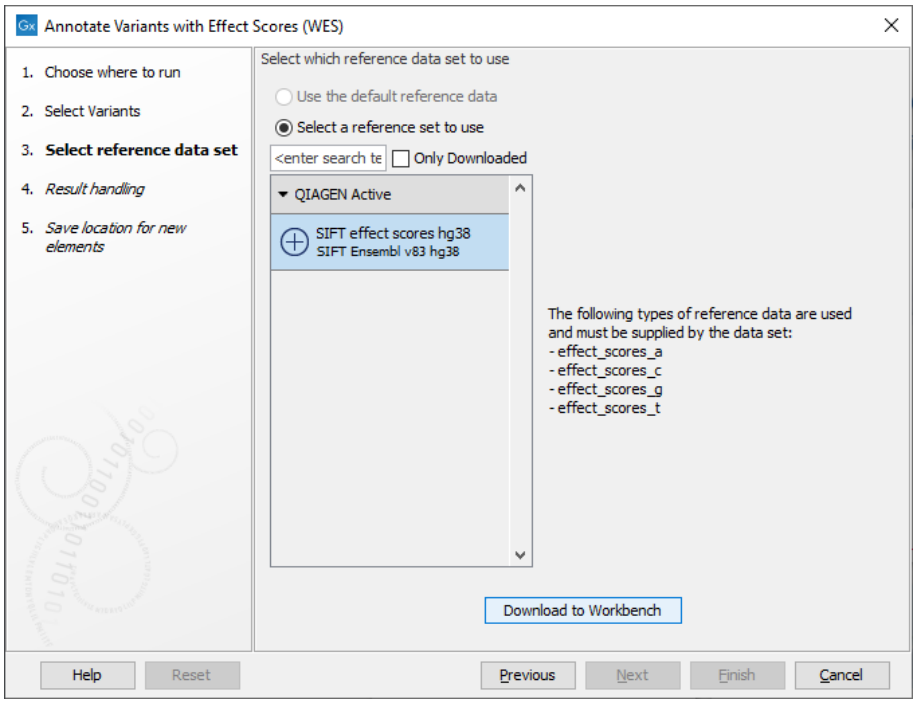


Figure 19.9: *Choose the relevant reference data set.*

4. In the last wizard step you can check all the workflow settings by clicking on the button labeled **Preview All Parameters**. If you wish to make changes to unlocked parameters, click on the **Previous** button until you get back to the relevant wizard step.

5. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Annotate Variants with Effect Scores (WES) workflow

The **Annotate Variants with Effect Scores (WES)** outputs a variant track () with the original annotations, plus the effect score annotations on non-reference SNVs. Effect scores may not

be available for particular positions, since some effect scores can only be computed for coding regions.

When hovering the mouse cursor over an individual variant, a tooltip appears containing detailed information about that variant, include the effect score, where available.

The effect score values can also be reviewed by opening the table view of the variant track, as shown in figure 19.10.

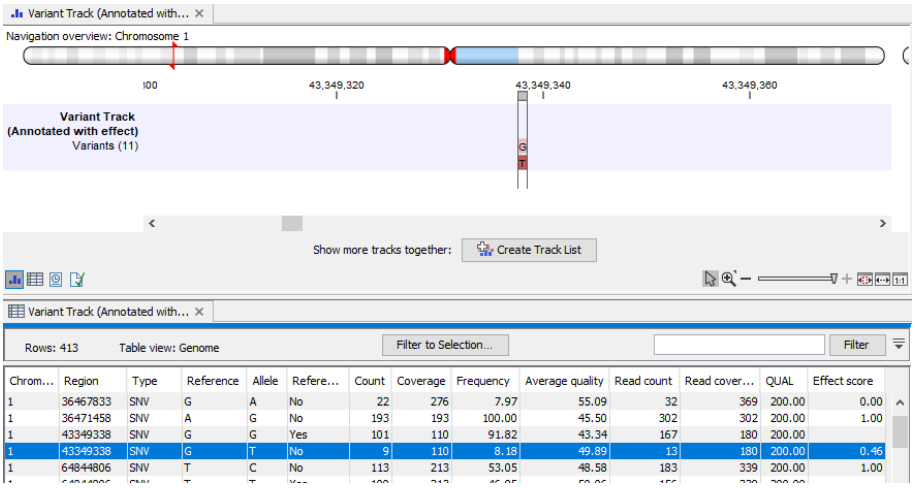


Figure 19.10: *Variants annotated with effect scores, here shown in a split view with the track view at the top and the table view at the bottom.*

19.1.3 Identify Known Variants in One Sample (WES)

The **Identify Known Variants in One Sample (WES)** template workflow combines data analysis and interpretation. It should be used to identify known variants as specified by the user (e.g., known breast cancer associated variants) for their presence or absence in a sample. This workflow will not identify new variants.

The workflow maps the sequencing reads to a human genome sequence and does a local realignment of the mapped reads to improve the subsequent variant detection. In the next step, only variants specified by the user are identified and annotated in the newly generated read mapping.

Before starting the workflow, you may need to import your reads with the **Import | Tracks** tool (see `Import | Tracks`):

- **Import your known variants**. Variants can be imported in GVF or VCF format.
- **Import your targeted regions**. A file with the genomic regions targeted by the amplicon or hybridization kit can usually be provided by the vendor, either BED or GFF format.

Run the Identify Known Variants in One Sample (WES) workflow

1. To run the **Identify Known Variants in One Sample (WES)** template workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **General Workflows (WES)** ( ) | **Identify Known Variants from One Sample (WES)** ( )

2. First select the reads of the sample that should be tested for presence or absence of your known variants (figure 19.11).

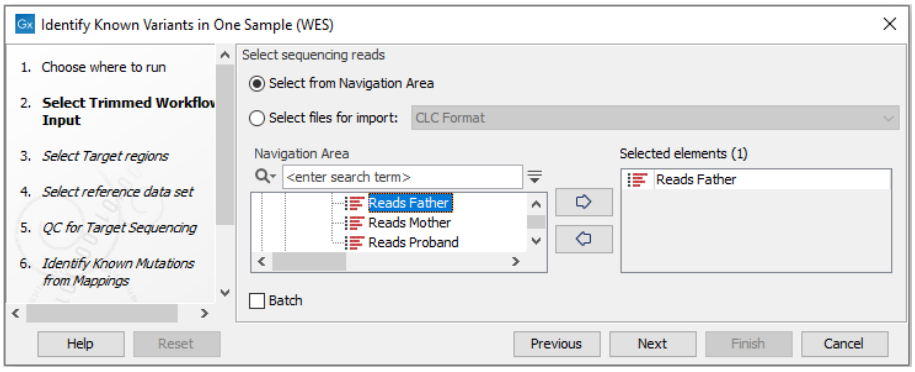


Figure 19.11: *Select the sequencing reads from the sample you would like to test for your known variants.*

If several samples from different folders should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" and specifying the folders that hold the data you wish to analyse.

3. Next, specify the target regions (figure 19.12). The target regions are used by the tools Structural Variant Caller and QC for Targeted Sequencing. Providing the Structural Variant Caller with target regions, reduces the processing time of the workflow as the tool only analyzes defined target regions.

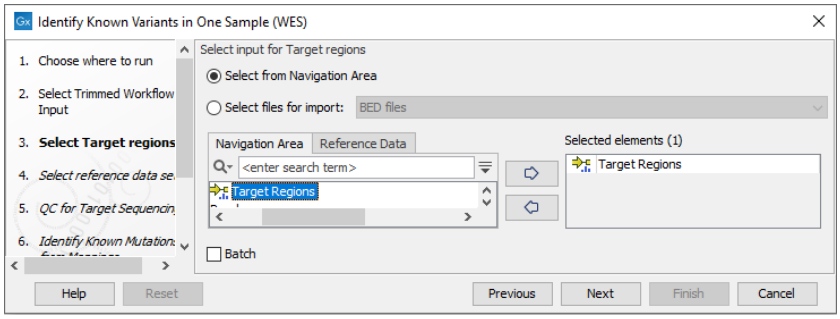


Figure 19.12: *Specify the target regions file.*

4. In the next wizard step, select the reference data set that should be used to identify the known variants (figure 19.13).

5. Specify the parameters for the QC for Target Sequencing tool (figure 19.14).

When working with targeted data (WES or TAS data), quality checks for the targeted sequencing are included in the workflows. This step is not optional, and you need to specify the targeted regions file adapted to the sequencing technology you used. Choose to use the default settings or to adjust the parameters.

The parameters that can be set are:

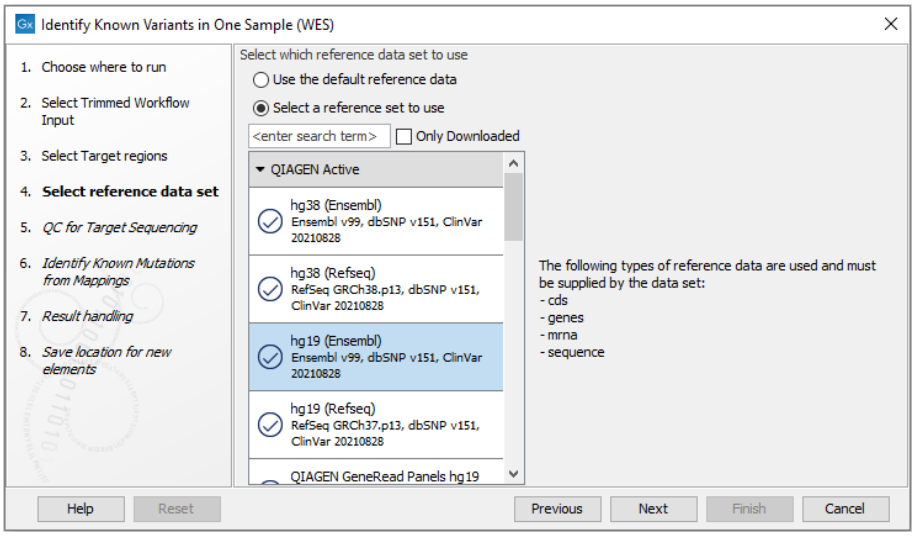


Figure 19.13: *Choose the relevant reference Data Set to identify the known variants.*

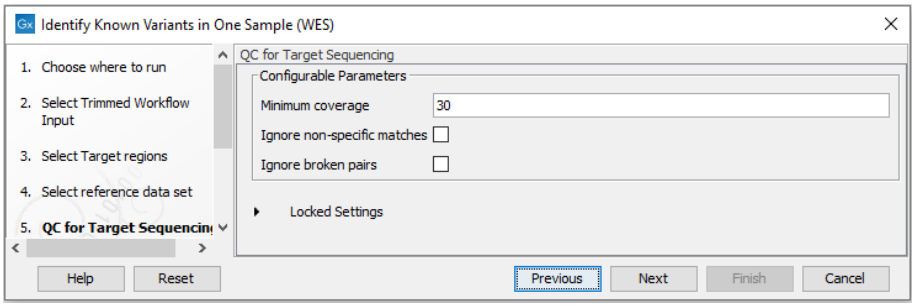


Figure 19.14: *Specify the parameters for the QC for Target Sequencing tool.*

- **Minimum coverage** provides the length of each target region that has at least this coverage.
- **Ignore non-specific matches** reads that are non-specifically mapped will be ignored.
- **Ignore broken pairs** reads that belong to broken pairs will be ignored.

6. In the Identify Known Mutations from Mappings, select a variant track containing the known variants you want to identify in the sample (figure 19.15).

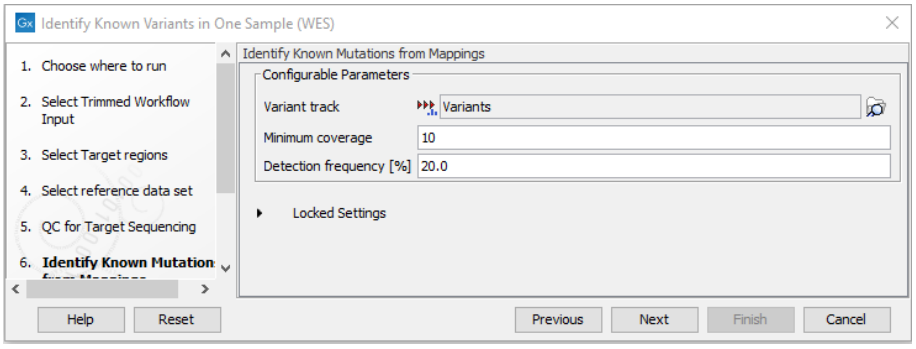


Figure 19.15: *Specify the track with the known variants that should be identified.*

The parameters that can be set are:

- **Minimum coverage** The minimum number of reads that covers the position of the variant, which is required to set "Sufficient Coverage" to YES.
- **Detection frequency** The minimum allele frequency that is required to annotate a variant as being present in the sample. The same threshold will also be used to determine if a variant is homozygous or heterozygous. In case the most frequent alternative allele at the position of the considered variant has a frequency of less than this value, the zygosity of the considered variant will be reported as being homozygous.

The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or not (observed frequency <= specified frequency). Moreover, it will determine if a variant should be labeled as heterozygous (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygous (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click **Finish**.

Output from the Identify Known Variants in One Sample (WES)

The **Identify Known Variants in One Sample (WES)** tool produces five different output types:

- **Read Mapping** () The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).
- **Target Regions Coverage** () A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.
- **Target Regions Coverage Report** () The report consists of a number of tables and graphs that in different ways show e.g. the number, length, and coverage of the target regions and provides information about the read count per GC%.
- **Variants Detected in Detail** () Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads, as well as detailed information about the Most Frequent Alternative Allele (labeled MFAA).
- **Track List Identify Known Variants** () A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

It is a good idea to start looking at the Target Regions Coverage Report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, we also recommend that you check that the majority of the reads are mapping to the targeted region.

When you have inspected the target regions coverage report you can open the Track List (see 19.16).

The Track List includes an overview track of the known variants and a detailed result track presented in the context of the human reference sequence, genes, transcripts, coding regions, targeted regions, and mapped sequencing reads.

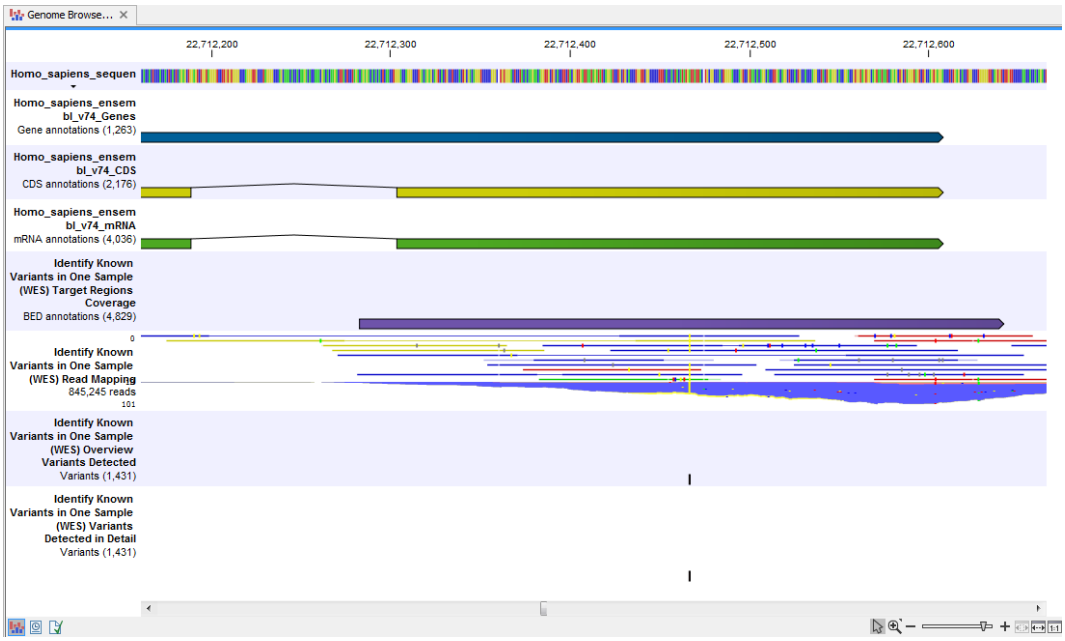


Figure 19.16: *Track List that allows inspection of the identified variants in the context of the human genome and external databases.*

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

Open the annotated variant as a table showing all variants and the added information/annotations (see 19.17).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.

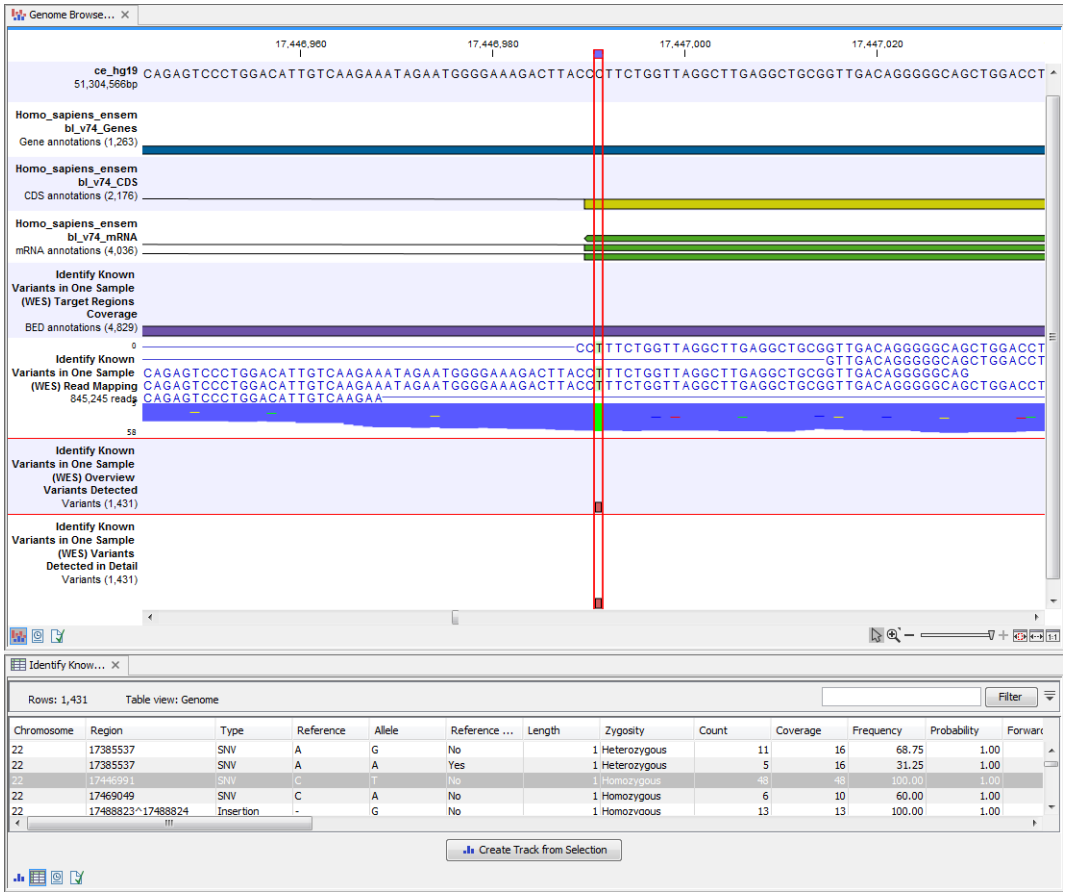


Figure 19.17: *Track List with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

19.2 Somatic Cancer (WES)

19.2.1 Filter Somatic Variants (WES)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same subject, you can use the **Filter Somatic Variants (WES)** template workflow to identify potential somatic variants. The purpose of this template workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same subject is available.

This workflow accepts variant tracks ( ) (e.g. the output from the Identify Variants template workflow) as input. In cases with heterozygous variants, the reference allele is first filtered away, then variants outside the targeted region are removed, and lastly, variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

Please note that this tool will likely also remove inherited cancer variants that are present at a low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes,

conservation scores and information from ClinVar (known variants with medical impact) and dbSNP (all known variants).

Run the Filter Somatic Variants (WES) workflow

To run the **Filter Somatic Variants (WES)**, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Somatic Cancer** ( ) | **Filter Somatic Variants** ( )

1. Double-click on the **Filter Somatic Variants** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Next, you will be asked to select the variant track you would like to use for filtering somatic variants (figure 19.18).

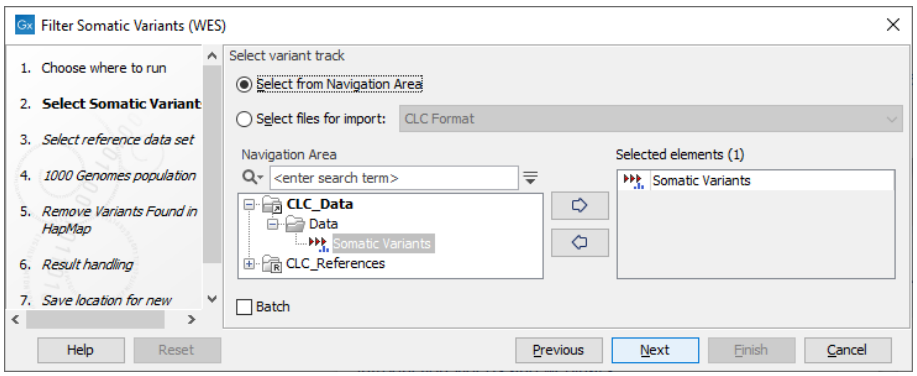


Figure 19.18: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter somatic variants (figure 19.19).
4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 19.20).
5. For databases that provide data from more than one population as HapMap does, the populations relevant to the data set can be specified. Click on the plus symbol ( ) and choose the population that matches the population your samples are derived from (figure 19.21). Please note that different populations are available and can be downloaded via the Reference Data Manager found in the top right corner of the CLC Workbench.
6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.
7. Choose to **Save** your results and click **Finish**.

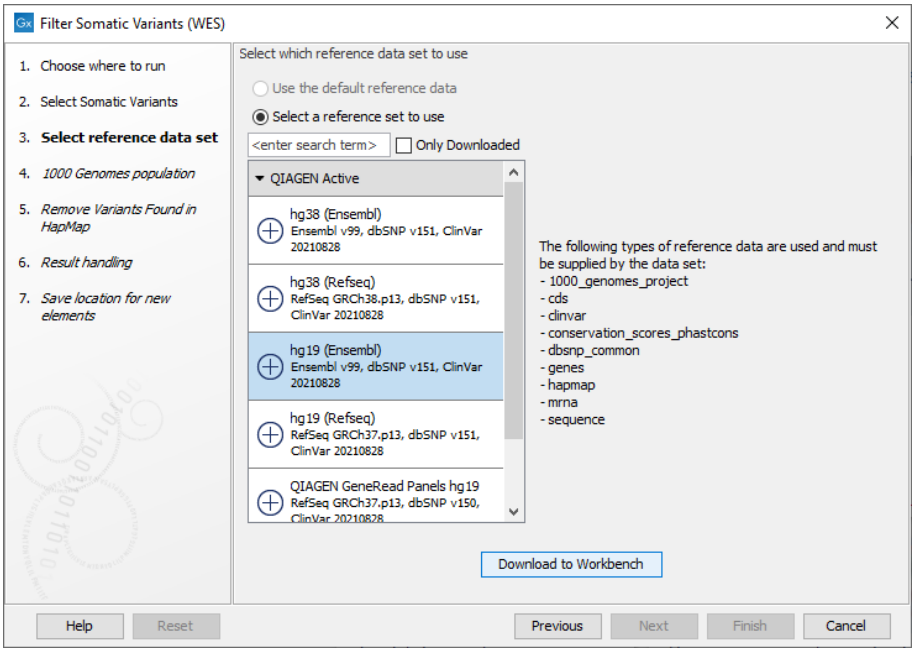


Figure 19.19: *Choose the relevant reference Data Set to annotate.*

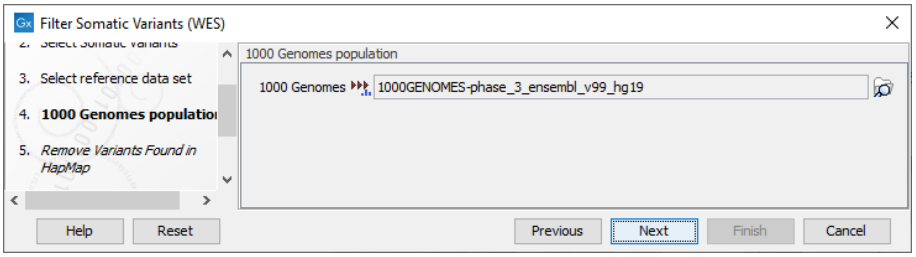


Figure 19.20: *Use the preselected 1000 Genomes population(s) or select another variant track.*

Output from the Filter Somatic Variants (WES) workflow Two types of output are generated:

- **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
- **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Track List. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.
- **Track List Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, 1000 Genomes, and the PhastCons conservation scores (see figure 19.22).

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well. Mapped

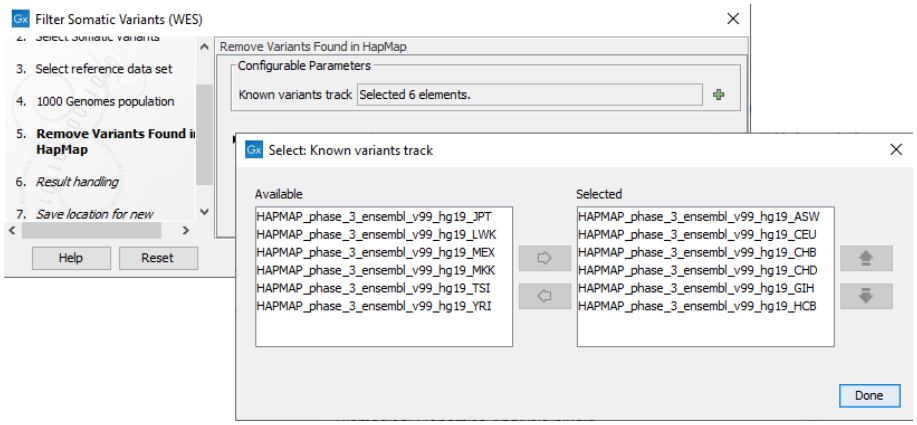


Figure 19.21: *Specify which HapMap population to use for filtering out known variants.*

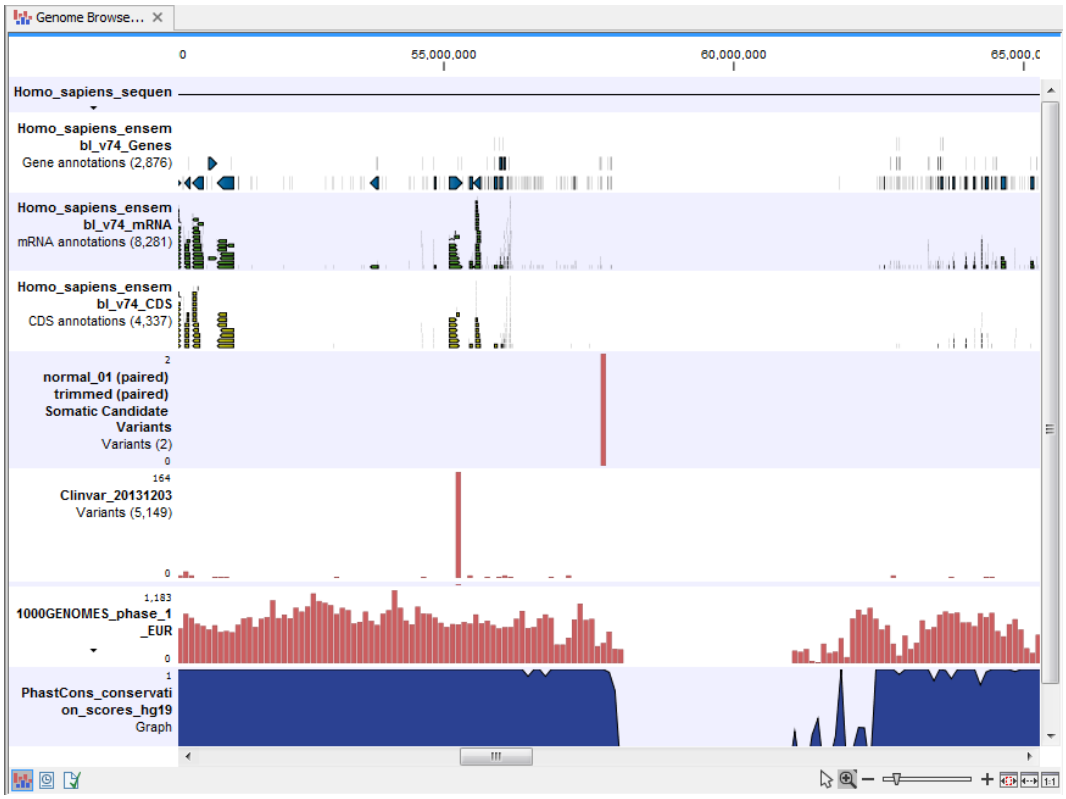


Figure 19.22: *The Track List showing the annotated somatic variants together with a range of other tracks.*

sequencing reads as well as other tracks can be easily added to this Track List. Open the variant track as a table showing all variants and the added information/annotations (see figure 19.23).

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar database, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

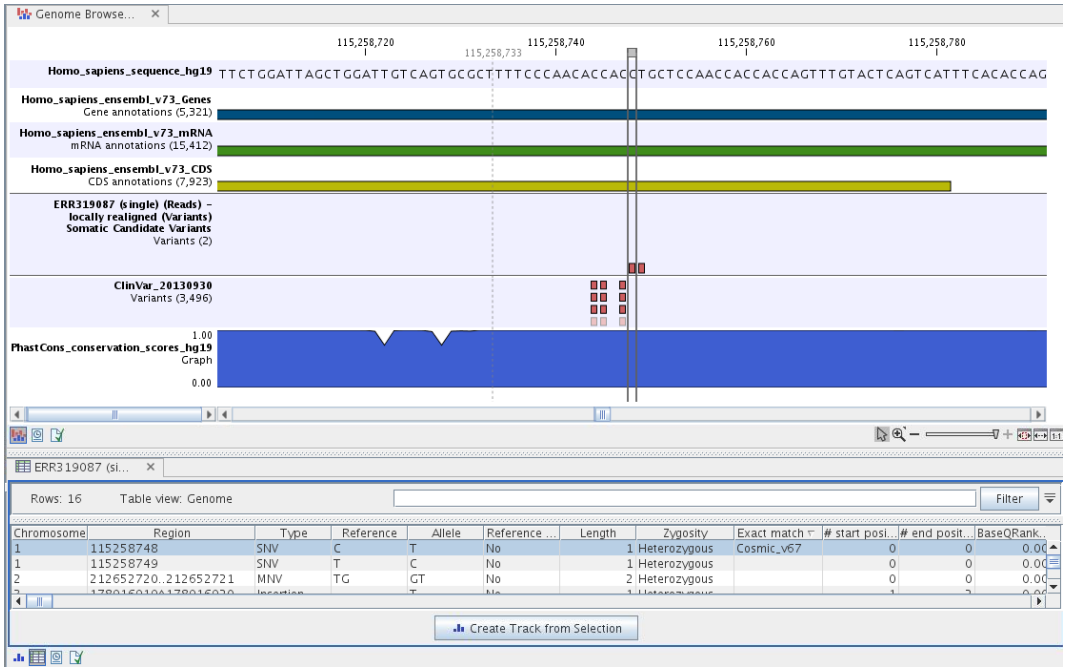


Figure 19.23: *The Track List showing the annotated somatic variants together with a range of other tracks.*

A high conservation level, between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

19.2.2 Identify Somatic Variants from Tumor Normal Pair (WES)

The **Identify Somatic Variants from Tumor Normal Pair (WES)** template workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same individual.

When running this workflow the reads are mapped and the variants identified. Germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads. Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from relevant databases like ClinVar (variants with clinically relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually

available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

Run the Identify Somatic Variants from Tumor Normal Pair (WES) workflow

1. To run the **Identify Somatic Variants from Tumor Normal Pair (WES)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Somatic Variants from Tumor Normal Pair (WES)** ( )

2. First (figure 19.24), select the tumor sample reads.

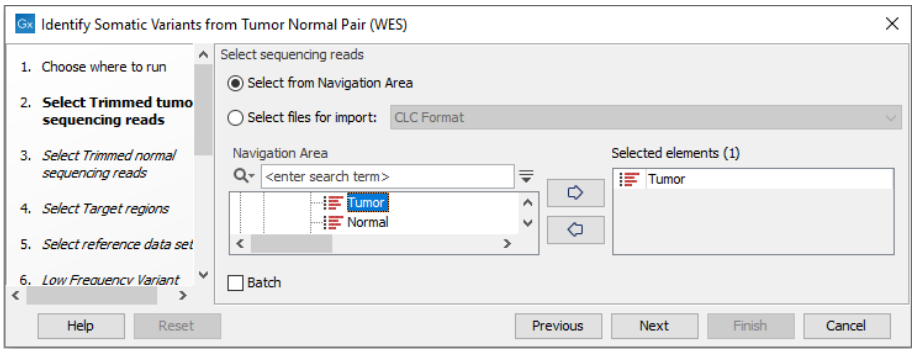


Figure 19.24: *Select the tumor sample reads.*

3. In the next wizard step, specify the normal sample reads.

4. The following step allows you to restrict variant calling to target regions, both for tumor and normal reads (figure 19.25). Variants found outside the targeted regions will not be included in the output that is generated with the template workflow.

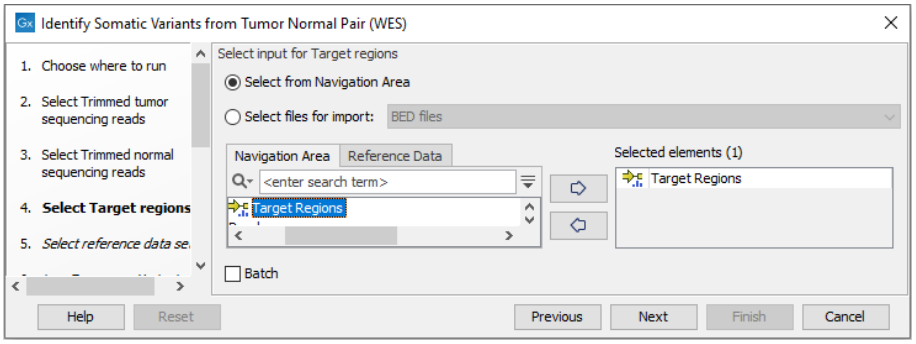


Figure 19.25: *Specify the target regions track.*

5. In the next dialog, select which reference data set should be used to identify variants (figure 19.26).

6. Set the parameters for the Low Frequency Variant Detection step (figure 19.27).

   For a description of the different parameters that can be adjusted, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_

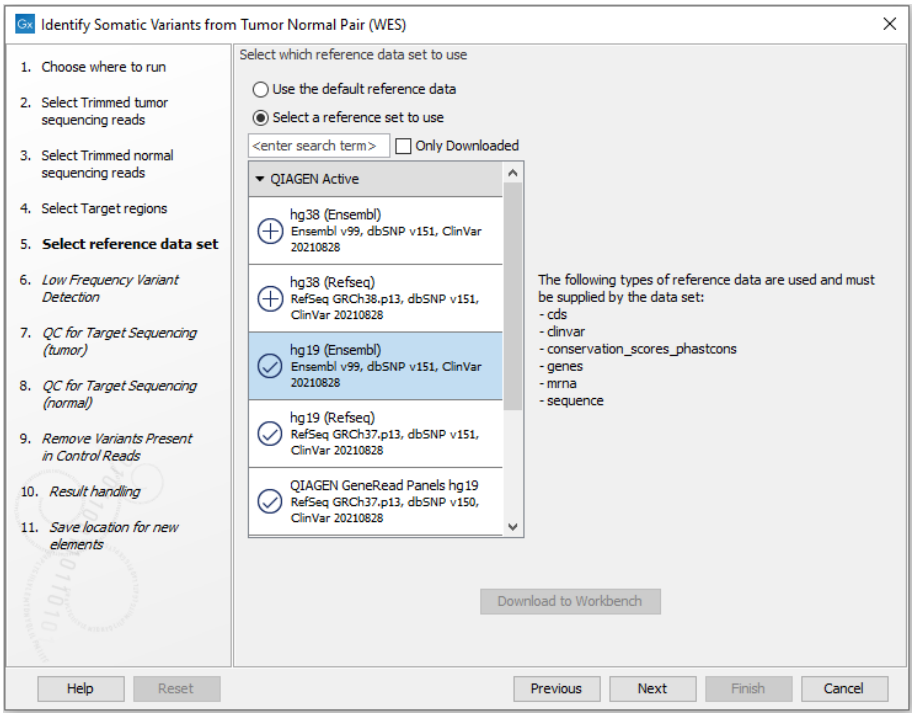


Figure 19.26: *Choose the relevant reference Data Set to identify variants.*

`Variant_Detection.html`. If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the template workflow.

7. In the following 2 wizard steps, you can specify the settings for QC for Target Sequencing which provides quality metrics for the performance of the targeted re-sequencing experiment for both tumor and normal samples (figure 19.28).

   For a description of the different parameters that can be adjusted, see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html`. If you click on "Locked Settings", you will be able to see all parameters used for the QC for Targeted Sequencing tool in the template workflow.

8. In the Remove Variants Present in Control Reads step, you can adjust the settings for removal of germline variants (figure 19.29).

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

10. Choose to **Save** your results and click **Finish**.

Output from the Identify Somatic Variants from Tumor Normal Pair (WES) workflow The following outputs are generated:

1. **Read Mapping Normal** ( ) The mapped sequencing reads for the normal sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see `http://resources.`

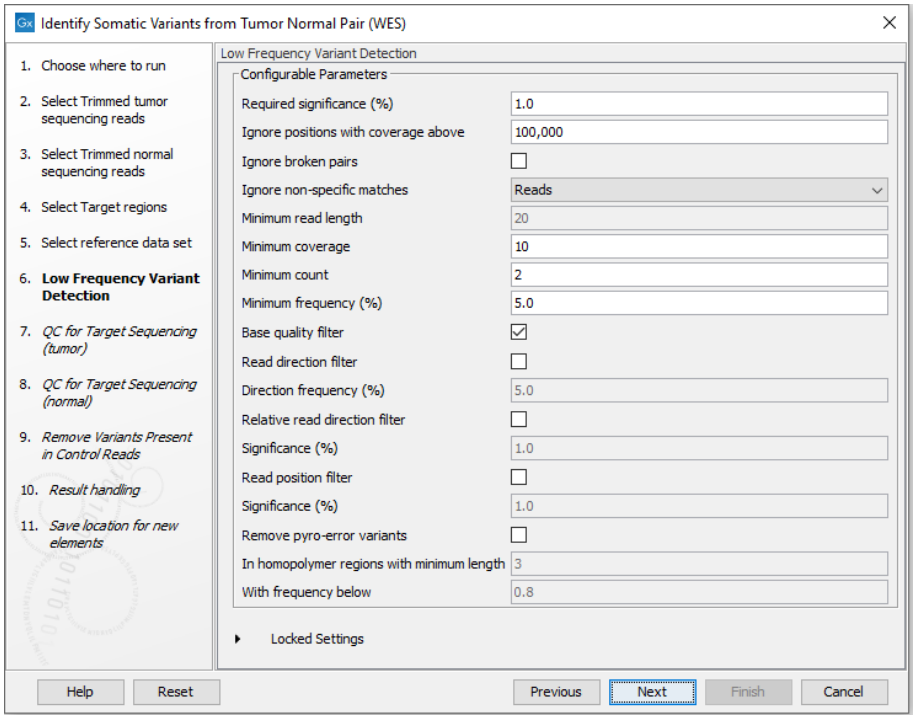


Figure 19.27: *Specify the settings for the variant detection.*

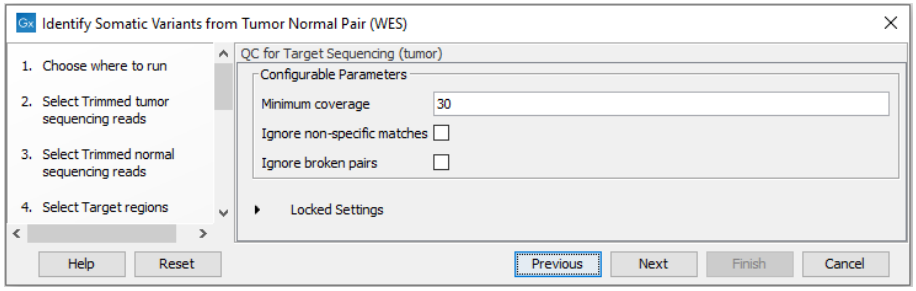


Figure 19.28: *Set the parameters for the QC for targeted regions.*

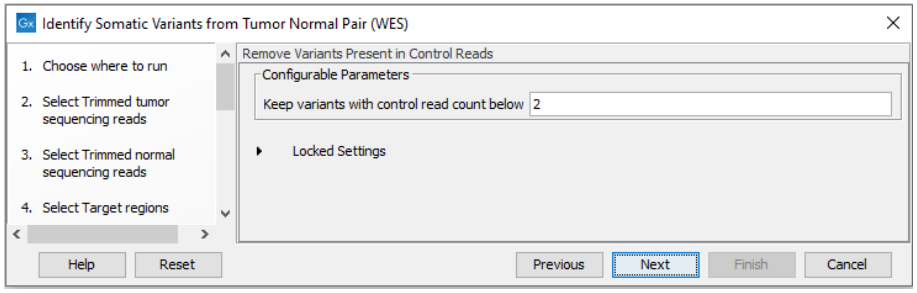


Figure 19.29: *Specify setting for removal of germline variants.*

qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

2. **Read Mapping Tumor** ( ) The mapped sequencing reads for the tumor sample.
3. **Target Region Coverage Report Normal** ( ) The report consists of a number of tables and

graphs that in different ways provide information about the mapped reads from the normal sample.

4. **Target Region Coverage Tumor** ( ) A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.

5. **Target Region Coverage Report Tumor** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.

6. **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

7. **Annotated Somatic Variants** ( ) A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

8. **Track List Tumor Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar database, and finally a track showing the conservation score (see figure 19.30).

19.2.3 Identify Variants (WES)

The **Identify Variants (WES)** template workflow takes sequencing reads as input and returns identified variants as part of a Track List.

Sequencing reads provided as input are initially mapped to the human reference sequence. This is followed by the removal of duplicate mapped reads (to reduce biases introduced by target enrichment). The resulting read mapping is analyzed by the Structural Variant Caller to infer indels and other structural variants from unaligned end read patterns. Subsequently, the mapping is realigned, guided by the indels detected by the Structural Variant Caller. The locally realigned read mapping is analyzed by the Low Frequency Variant Detection tool. The Low Frequency Variant Detection tool produces a track of unfiltered variants; these are subjected to a number of post filters to remove variants that are likely due to artifacts or noise. The variants called by the Low Frequency Variant Detection tool that pass the post filtering criteria can be found in the Identified variants track. Variants inferred by the Structural Variant Caller, and not detected by the Low Frequency Variant Detection tool, are also subjected to a number of post filters; those that pass the post filter criteria can be found in the Indels indirect evidence track.

In addition, a targeted region report is created to inspect the overall coverage and mapping specificity in the targeted regions.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

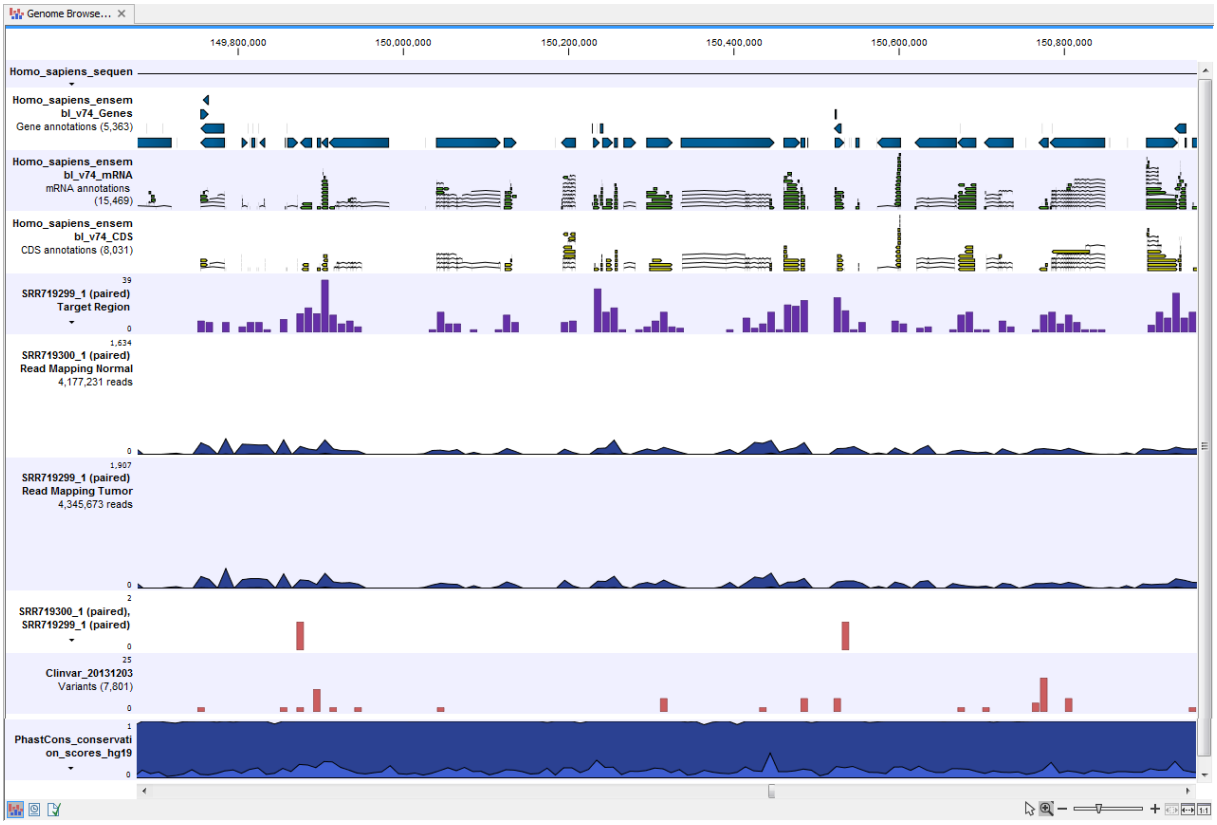


Figure 19.30: *The Track List presents all the different data tracks together and makes it easy to compare different tracks.*

Run the Identify Variants (WES) workflow

1. To run the **Identify Variants (WES)** template workflow, go to:

   Workflows |**Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Variants (WES)** ( )

2. Select the sequencing reads from the sample that should be analyzed (figure 19.31).

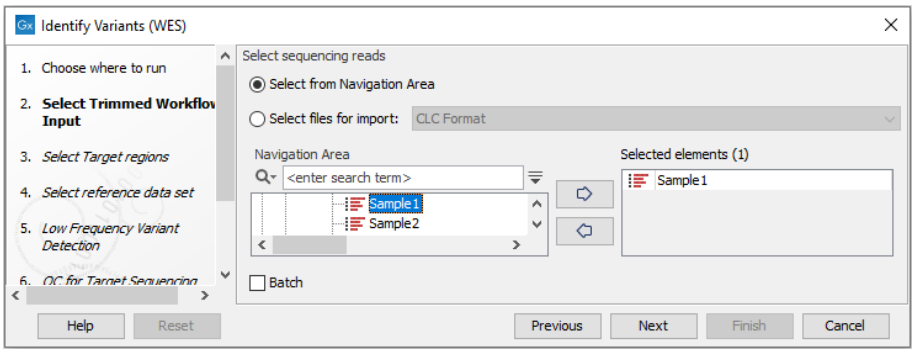


Figure 19.31: *Please select all sequencing reads from the sample to be analyzed.*

If several samples should be analyzed, the tool has to be run in batch mode. This is done by checking "Batch" and selecting the **folder** that holds the data you wish to analyze.

3. Next, in the Target regions dialog you need to specify the target regions for your application. The variant calling will be restricted to these regions (figure 19.32).

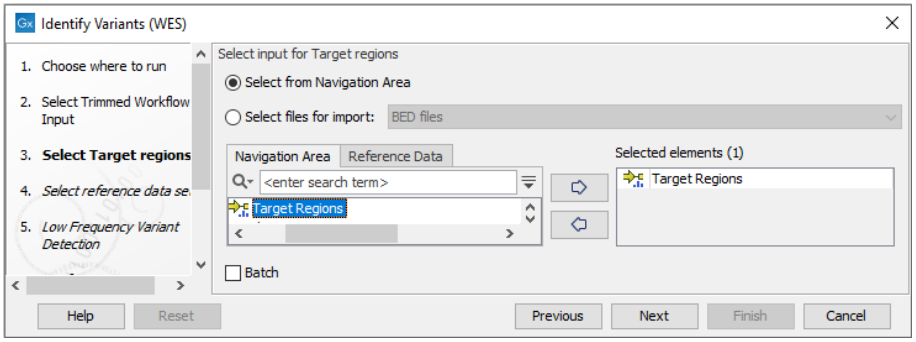


Figure 19.32: *Select the track with the targeted regions from your experiment.*

4. In the next dialog, you have to select which reference data set should be used to identify variants (figure 19.33).

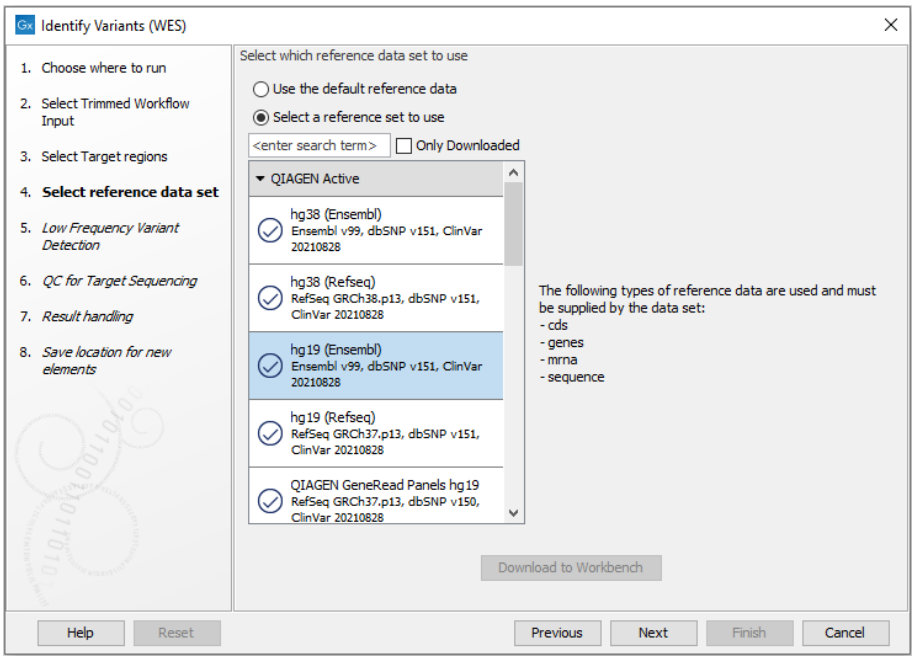


Figure 19.33: *Choose the relevant reference Data Set to identify variants in your sample.*

5. In the next wizard step (figure 19.34), you can specify the parameters for variant detection.

6. In the QC for Target Sequencing step (figure 19.35) you can specify the minimum read coverage, which should be present in the targeted regions.

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click **Finish**.

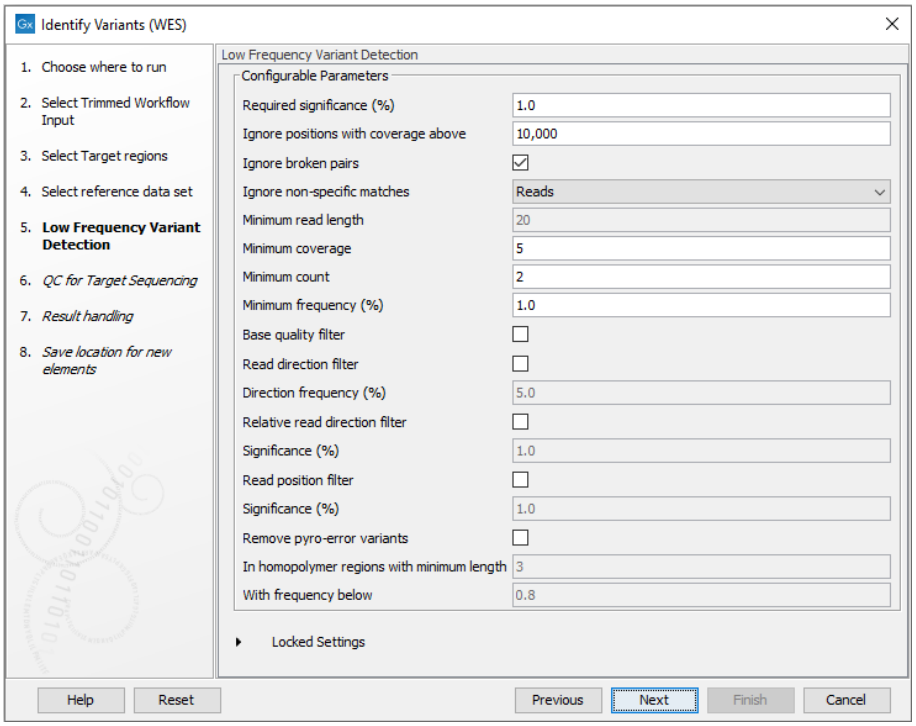


Figure 19.34: *Specify the parameters for variant detection.*

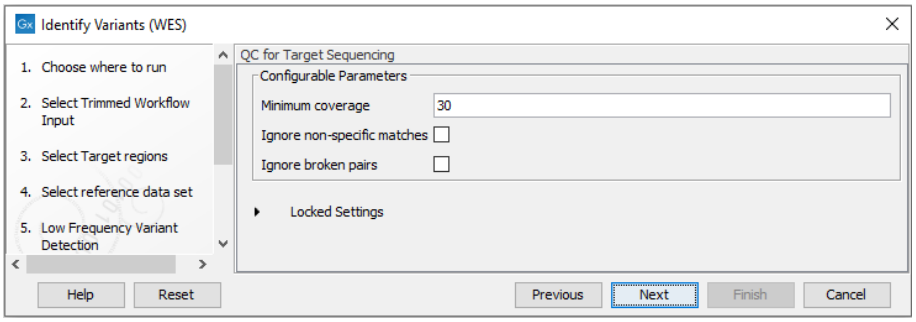


Figure 19.35: *Specify the minimum coverage for the QC for Targeted sequencing.*

Output from the Identify Variants (WES) workflow

The **Identify Variants (WES)** tool produces the following outputs:

1. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html`).

2. **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

3. **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

4. Three variant tracks ( ): Two from the Variant Caller: the **Unfiltered Variants** is output before the filtering steps, the **Variants passing filters** is the one used in the Genome Browser View (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@_annotated_variant_table.html for a definition of the variant table content). The third is the **Indels indirect evidence** track produced by the Structural Variant Caller. This is also available in the Genome Browser View. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

5. **Genome Browser View** ( ) A collection of tracks presented together. Shows the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the indels indirect evidence variants (see figure 19.5).

It is important that you do not delete any of the produced files individually as some of the outputs are linked to other outputs. If you would like to delete the outputs, please always delete all of them at the same time.

We recommend that you first inspect the target region coverage report to check that the majority of reads are mapping to the targeted region, and to see if the coverage is sufficient in regions of interest. Furthermore, check that at least 90% of reads are mapped to the human reference sequence.

Afterwards please open the Track List file (see 19.36).

The Genome Browser View includes the track of identified variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions and mapped sequencing reads.

Open the variant track as a table to see information about all identified variants (see 19.37).

19.2.4 Identify and Annotate Variants (WES)

The **Identify and Annotate Variants (WES)** template workflow should be used to identify and annotate variants in one sample. The workflow is a combination of the **Identify Variants** and the **Annotate Variants** workflows.

The workflow starts with mapping the sequencing reads to the human reference sequence, followed by a local realignment to improve the variant detection that is run afterwards. After the variants have been detected, they are annotated with gene names, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and information from common variants present in the common dbSNP Common, HapMap, and 1000 Genomes database. Furthermore, a detailed targeted region coverage report is created to inspect the overall coverage and mapping specificity.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

Figure 19.36: *The Genome Browser View allows you to inspect the identified variants in the context of the human genome.*

Run the Identify and Annotate Variants (WES) workflow

To run the **Identify and Annotate Variants (WES)** workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify and Annotate Variants (WES)** ( )

1. Double-click on the workflow name to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. First select the sequencing reads from the sample that should be analyzed (figure 19.38).

   If several samples should be analyzed, the tool has to be run in batch mode. This is done by checking "Batch" and selecting the **folder** that holds the data you wish to analyze.
3. In the Target regions dialog (figure 19.39), you can specify the target regions track. Only variants within the specified regions will be detected.
4. In the next dialog, you have to select which reference data set should be used to identify and annotate variants (figure 19.40).
5. In the next wizard step (figure 19.41) you can select the population from the 1000 Genomes project that you would like to use for annotation.
6. In the next dialog (figure 19.42), you have to specify the parameters for the variant detection.

   For a description of the different parameters that can be adjusted, see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html`. If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the template workflow.

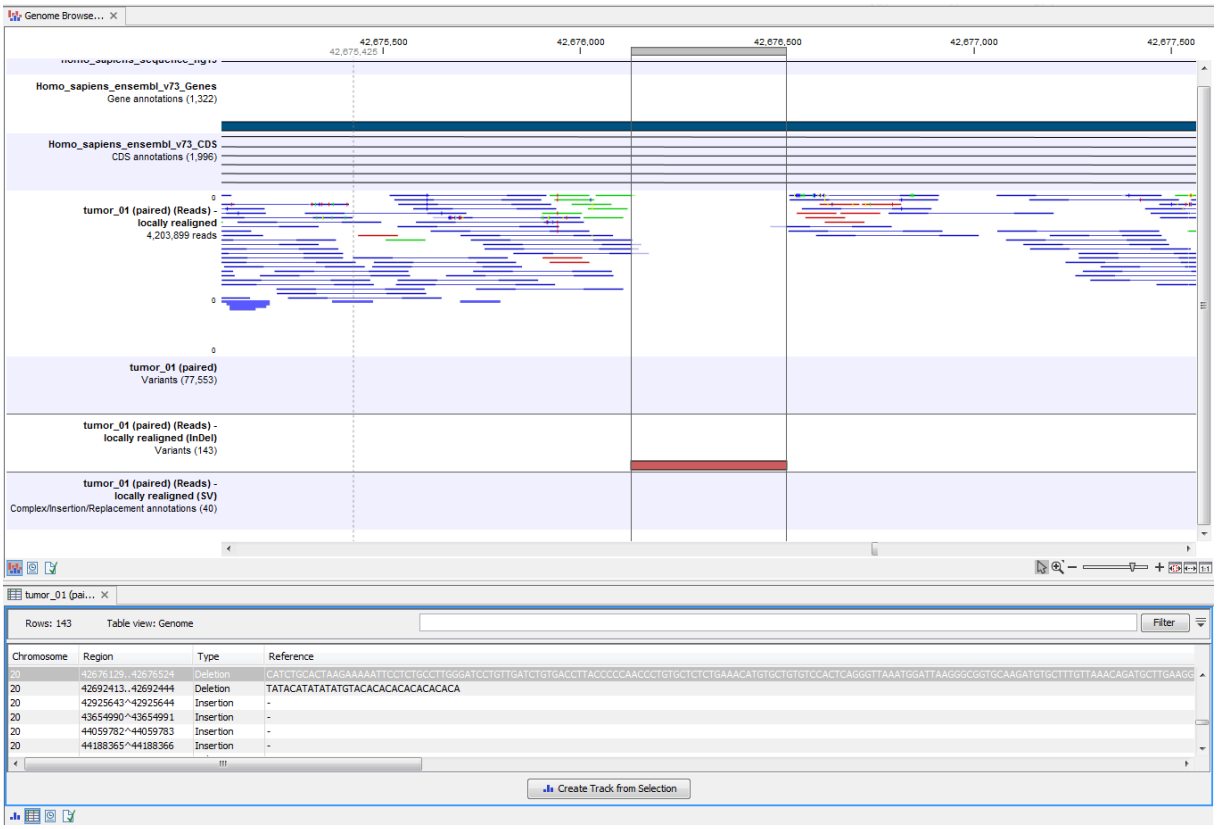


Figure 19.37: *Genome Browser View with an open track table to inspect identified variants more closely in the context of the human genome.*

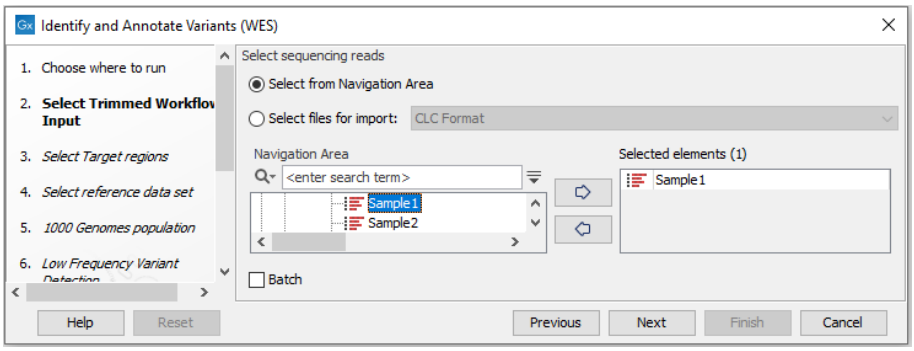


Figure 19.38: *Select all sequencing reads from the sample to be analyzed.*

7. In the QC for Target Sequencing step (figure 19.43) you can specify the minimum read coverage that should be present in the targeted regions.

8. Finally, select a population from the HapMap database (figure 19.44). This will add information from the Hapmap database to your variants.

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

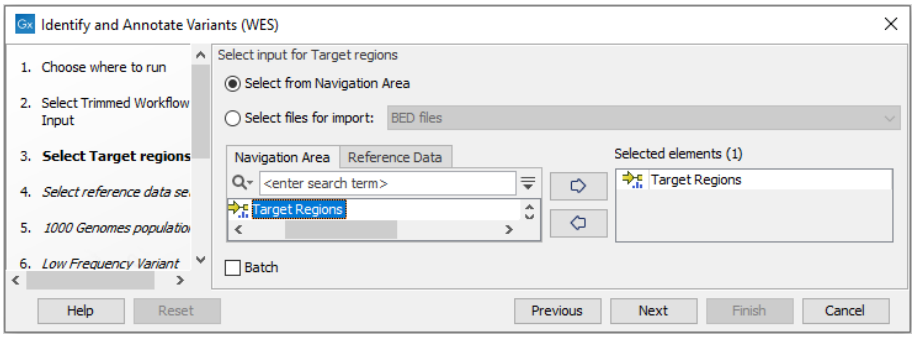


Figure 19.39: *Specify the target regions.*

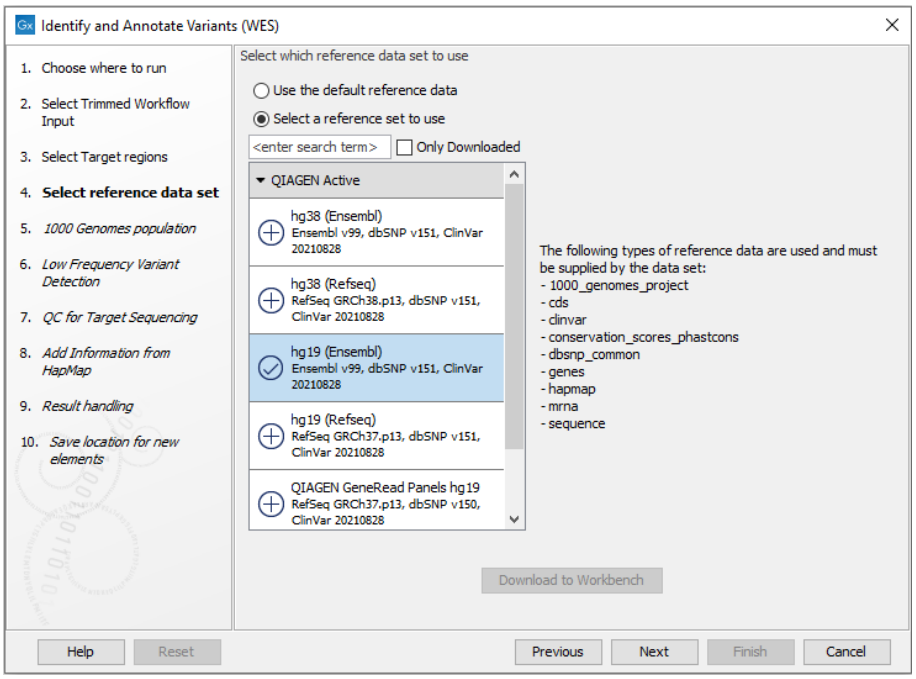


Figure 19.40: *Choose the relevant reference Data Set to identify and annotate.*

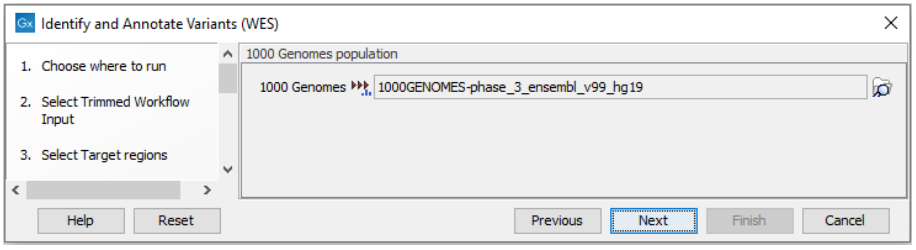


Figure 19.41: *Select the population from the 1000 Genomes project that you would like to use for annotation.*

10. Choose to **Save** your results and click **Finish**.

Output from the Identify and Annotate Variants (WES) workflow

The **Identify and Annotate Variants (WES)** workflow produces several outputs.

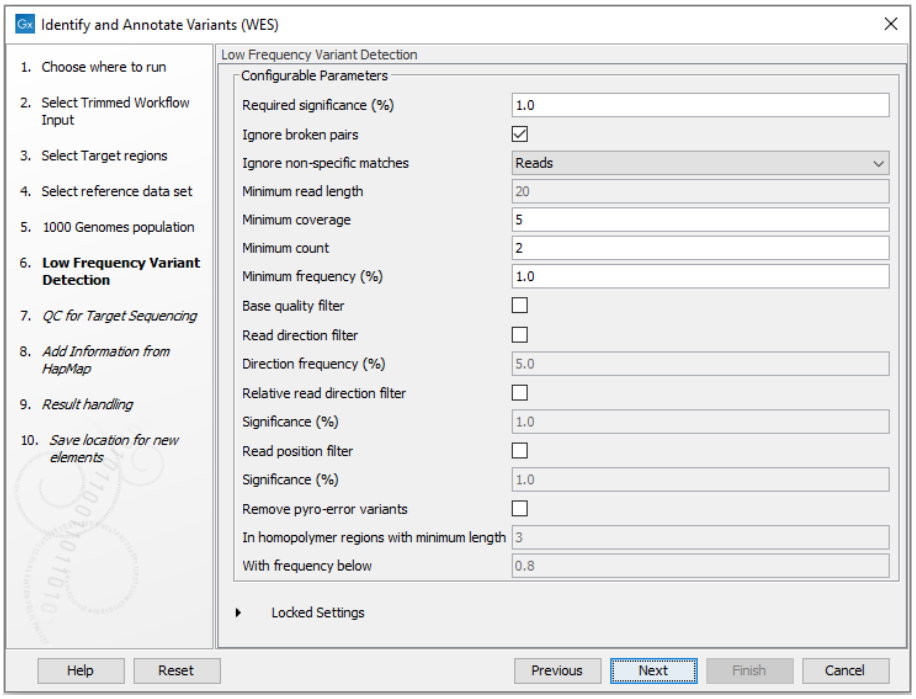


Figure 19.42: *Specify the parameters for variant calling.*

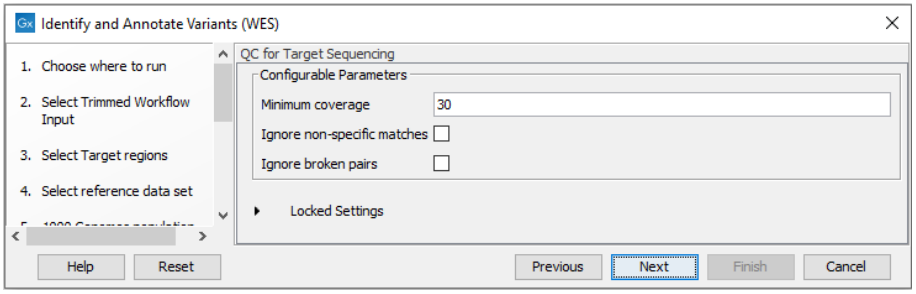


Figure 19.43: *Set the coverage threshold for the QC of the Targeted sequencing.*

1. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

2. **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

3. **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

4. Three variant tracks ( ): Two from the Variant Caller: the **Unfiltered Variants** is output before the filtering steps, the **Variants passing filters** is the one used in the Genome Browser View (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html http://resources.qiagenbioinformatics.

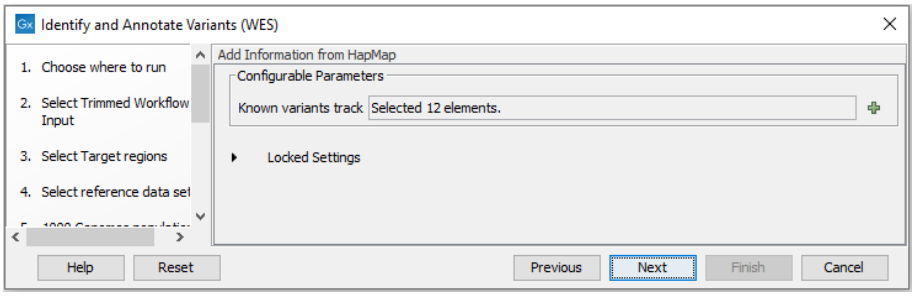


Figure 19.44: *Select a population from the HapMap database to add information from the Hapmap database to your variants.*

com/manuals/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@_annotated_variant_table.html for a definition of the variant table content). The third is the **Indels indirect evidence** track produced by the Structural Variant Caller. This is also available in the Genome Browser View. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

5. **Amino acid changes** Adds information about amino acid changes caused by the variants.

6. **Genome Browser View** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, amino acid changes, the mapped reads, the identified variants, and the indels indirect evidence variants (see figure 19.5).

Please do not delete any of the produced files alone as some of them are linked to other outputs. Please always delete all of them at the same time.

A good place to start is to take a look at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Furthermore, please check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of the reads are mapping to the targeted region.

Next, open the Genome Browser View (see figure 19.45).

The Genome Browser View includes a track of the identified annotated variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, relevant variants in the ClinVar database as well as common variants in common dbSNP Common, HapMap, and 1000 Genomes databases.

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well.

By double-clicking on the annotated variant track in the Genome Browser View, a table will be shown that includes all variants and the added information/annotations (see figure 19.46).

The added information will help you to identify candidate variants for further research. For example can common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) easily be seen.

Figure 19.45: *Genome Browser View to inspect identified variants in the context of the human genome and external databases.*

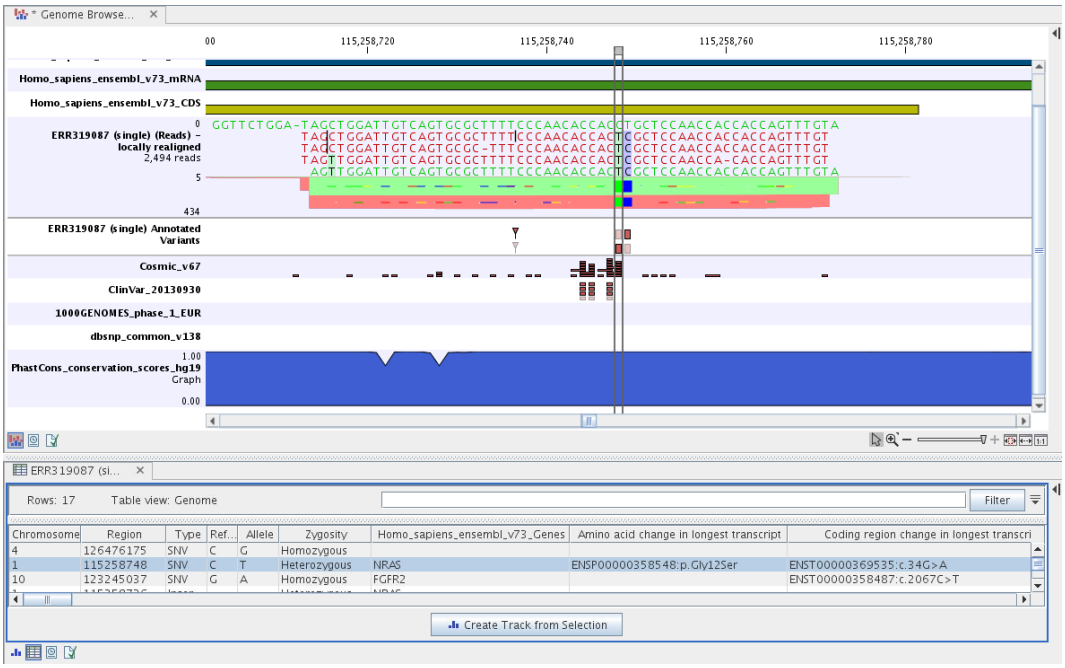


Figure 19.46: *Genome Browser View with an open track table to inspect identified somatic variants more closely in the context of the human genome and external databases.*

Not identified variants in ClinVar, can for example be prioritized based on amino acid changes (do they cause any changes on the amino acid level?). A high conservation level on the position of the variant between many vertebrates or mammals can also be a hint that this region could have an important functional role and variants with a conservation score of more than 0.9 (PhastCons score) should be prioritized higher. A further filtering of the variants based on their annotations can be facilitated using the table filter on top of the table.

If you wish to always apply the same filter criteria, the **Create new Filter Criteria** tool should be used to specify this filter and the **Identify and Annotate Variants (WES)** workflow should be extended by the **Identify Candidate Tool** (configured with the Filter Criterion). See the reference manual for more information on how preinstalled workflows can be edited.

Please note that in case none of the variants are present in ClinVar or dbSNP Common, the corresponding annotation column headers are missing from the result.

In case you like to change the databases as well as the used database version, please use the Reference Data Manager.

19.3 Hereditary Disease (WES)

19.3.1 Filter Causal Variants (WES-HD)

If you are analyzing a list of variants, you can use the **Filter Causal Variants (WES-HD)** template workflow to remove variants that are outside the target region, as well as common variants present in publicly available databases. The workflow will annotate the remaining variants with gene names, conservation scores, and information from relevant databases.

Run the Filter Causal Variants (WES-HD) workflow

To run the **Filter Causal Variants (WES-HD)** workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Hereditary Disease** ( ) | **Filter Causal Variants (WES -HD)** ( )

1. Double-click on the workflow name to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Select the **variant track** you want to use for filtering causal variants (figure 19.47).

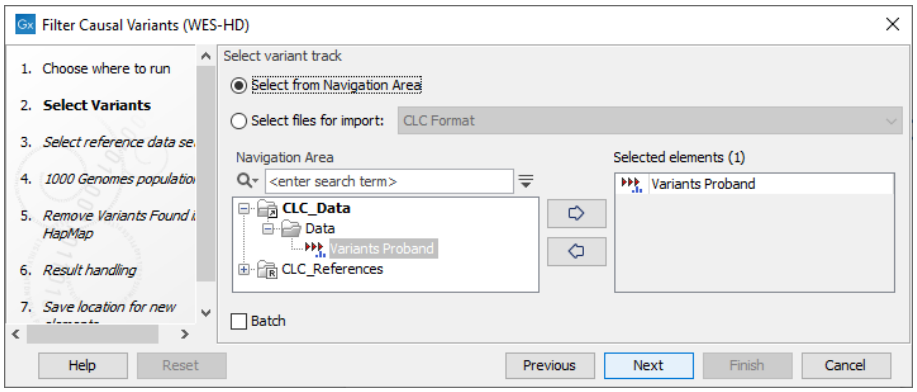


Figure 19.47: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter causal variants (figure 19.48).
4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track (figure 19.49).

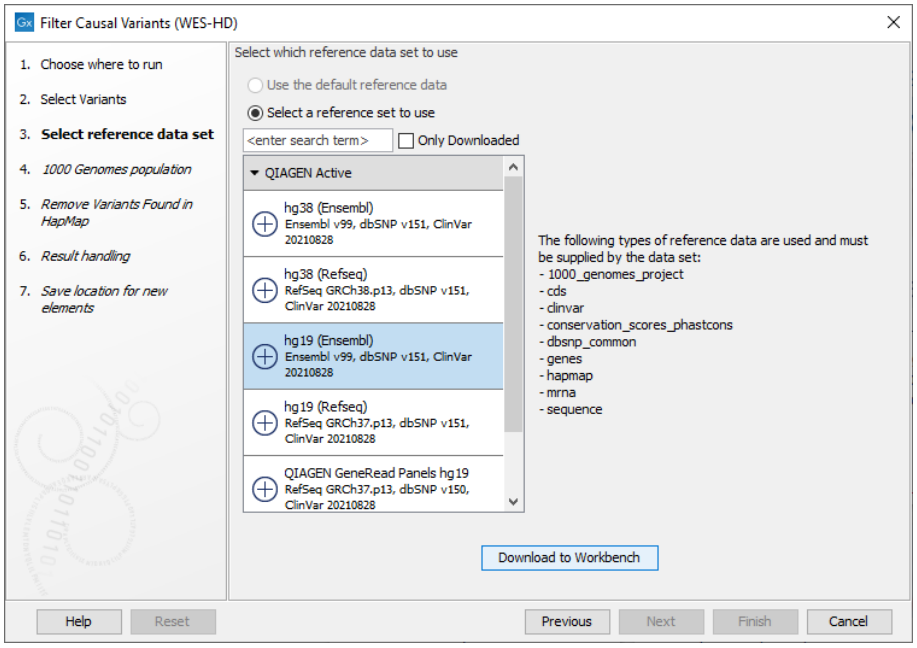


Figure 19.48: *Choose the relevant reference Data Set to annotate.*

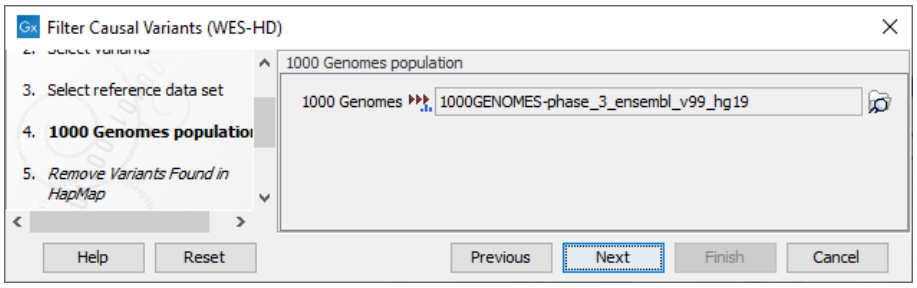


Figure 19.49: *Use the preselected 1000 Genomes population(s) or select another variant track.*

5. Specify the **Hapmap populations** that should be used for **filtering out** variants found in Hapmap (figure 19.50).

6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

7. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Filter Causal Variants (WES-HD) workflow

The following outputs are generated:

- An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

- A **Track List**

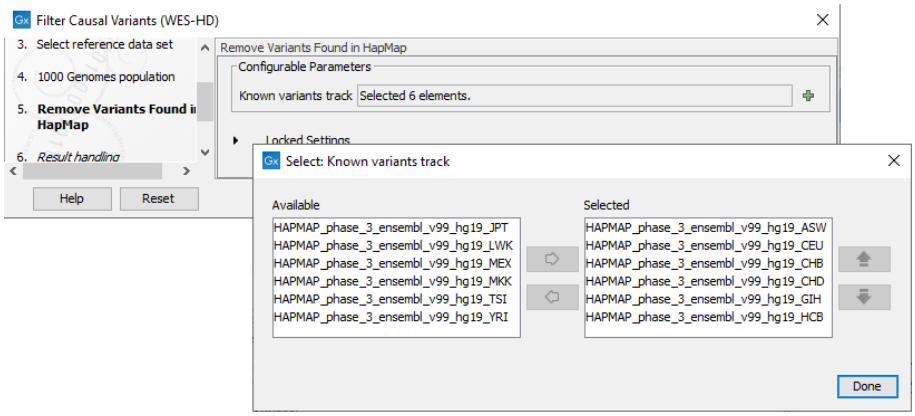


Figure 19.50: *Select the relevant Hapmap population(s).*

- A **Filtered Variant Track**

19.3.2 Identify Variants (WES-HD)

You can use the **Identify Variants (WES-HD)** template workflow to call variants in the mapped and locally realigned reads. The workflow removes false positives and, in case of a targeted experiment, removes variants outside the targeted region. Variant calling is performed with the Fixed Ploidy Variant Detection tool and the Structural Variant Caller.

The **Identify Variants (WES-HD)** template workflow accepts sequencing reads as input.

Run the Identify Variants (WES-HD) workflow

1. To run the **Identify Variants (WES-HD)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Hereditary Disease** ( ) | **Identify Variants (WES-HD)** ( )

2. Select the sequencing reads you want to analyze (figure 19.51).

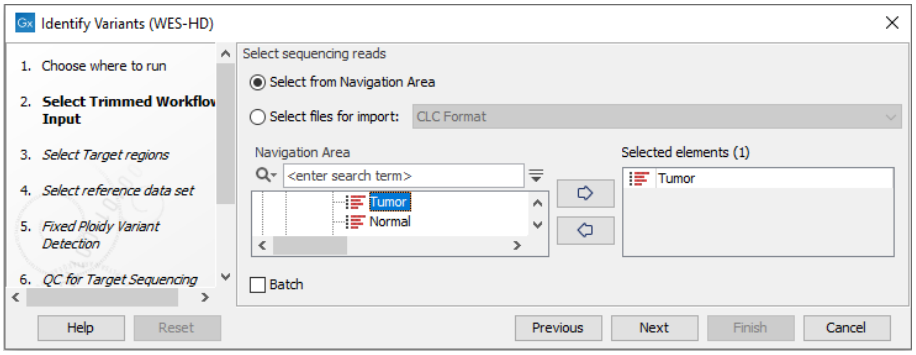


Figure 19.51: *Specify the sequencing reads for the sample.*

3. Specify a target region file (figure 19.52).

   The targeted region file is a file that specifies which regions have been sequenced, when working with whole exome sequencing or targeted amplicon sequencing data. This file is

something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents.

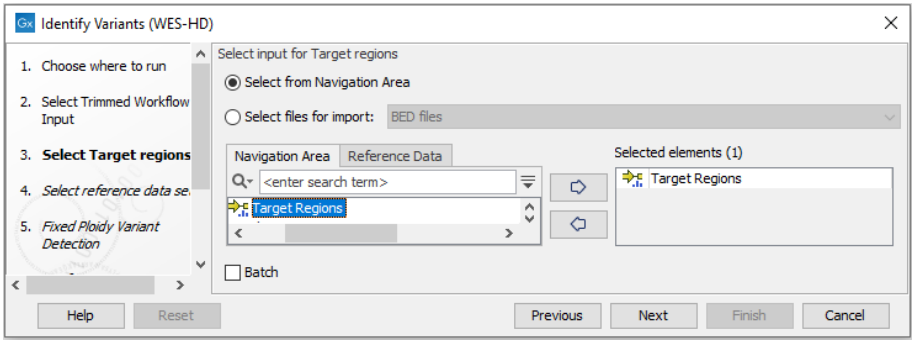


Figure 19.52: *Specify the target regions.*

4. In the next dialog, you have to select which reference data set should be used in the analysis (figure 19.53).

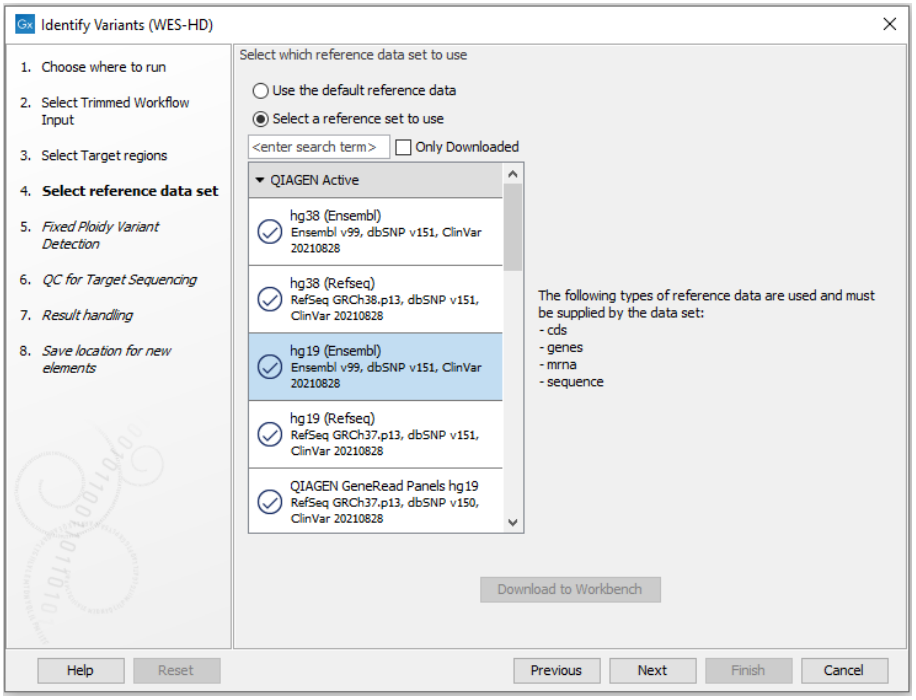


Figure 19.53: *Choose the relevant reference Data Set to identify variants.*

5. Specify the parameters for the Fixed Ploidy Variant Detection tool (figure 19.54).

   The parameters used by the Fixed Ploidy Variant Detection tool can be adjusted. We have optimized the parameters to the individual analyses, but you may want to tweak some of the parameters to fit your particular sequencing data. A good starting point could be to run an analysis with the default settings.

   The parameters that can be set are:

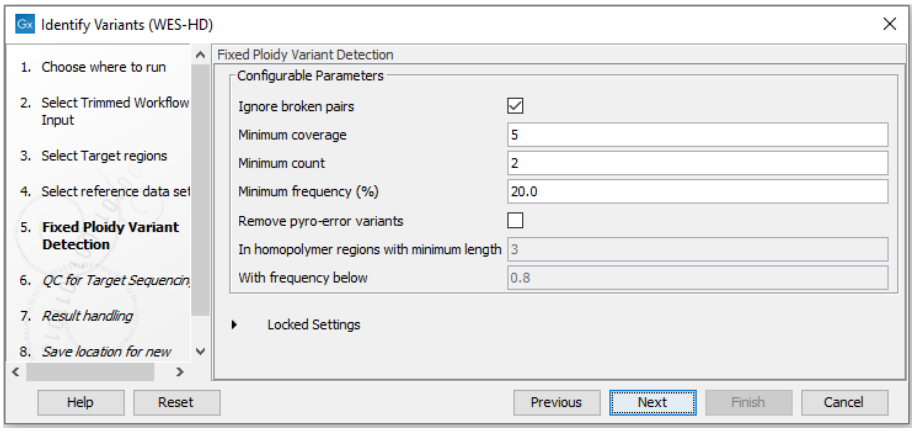


Figure 19.54: *Specify the parameters for the Fixed Ploidy Variant Detection tool.*

- **Required variant probability** is the minimum probability value of the 'variant site' required for the variant to be called. Note that it is not the minimum value of the probability of the individual variant. For the Fixed Ploidy Variant detector, if a variant site - and not the variant itself - passes the variant probability threshold, then the variant with the highest probability at that site will be reported even if the probability of that particular variant might be less than the threshold. For example if the required variant probability is set to 0.9 then the individual probability of the variant called might be less than 0.9 as long as the probability of the entire variant site is greater than 0.9.
- **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected. Please note that ignored broken pair reads will not be considered for any non-specific match filters.
- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

6. Specify the parameters for the QC for Targeted Sequencing tool (figure 19.55).

   When working with targeted data (WES or TAS data), quality checks for the targeted sequencing is included in the workflows. Again, you can choose to use the default settings, or you can choose to adjust the parameters.

   The parameters that can be set are:

   - **Minimum coverage** provides the length of each target region that has at least this coverage.
   - **Ignore non-specific matches**: reads that are non-specifically mapped will be ignored.
   - **Ignore broken pairs**: reads that belong to broken pairs will be ignored.

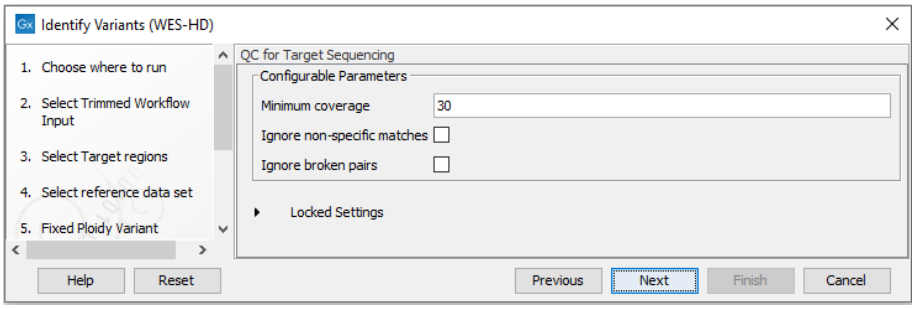


Figure 19.55: *Specify the parameters for the QC for Targeted Sequencing tool.*

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Identify Variants (WES-HD) workflow

- **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

- **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

- **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

- Two variant tracks ( ): the **Idenitified variants** track contains the variants detected by the Fixed Ploidy Variant Caller, the **Indels indirect evidence** track those detected by the Structural Variant Caller (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@_annotated_variant_table.html for a definition of the variant table content). The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

- **Genome Browser View** ( ) A collection of tracks presented together. Shows the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the indels indirect evidence variants (see figure 19.5).

19.3.3 Identify and Annotate Variants (WES-HD)

The **Identify and Annotate Variants (WES-HD)** template workflow can be used to identify and annotate variants in one sample. This workflow is a combination of the **Identify Variants** and the **Annotate Variants** template workflows.

Sequencing reads provided as input are initially mapped to the human reference sequence. Then a local realignment is carried out to improve the subsequent variant detection analysis. Variants detected are then annotated with gene names, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and information from common variants present in the common dbSNP Common, HapMap, and 1000 Genomes database. Furthermore, a targeted region report is created to inspect the overall coverage and mapping specificity.

The difference between Identify and Annotate Variants (TAS-HD) and (WES-HD) is that the **Auto-detect paired distances** has been switched off in Map Reads to Reference tool for the TAS workflows.

Run the Identify and Annotate Variants (WES-HD) workflow

1. To run the **Identify and Annotate Variants (WES-HD)** workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Exome Sequencing** ( ) | **Hereditary Disease** ( ) | **Identify and Annotate Variants (WES-HD)** ( )

2. Select the sequencing reads you want to analyze (figure 19.56).

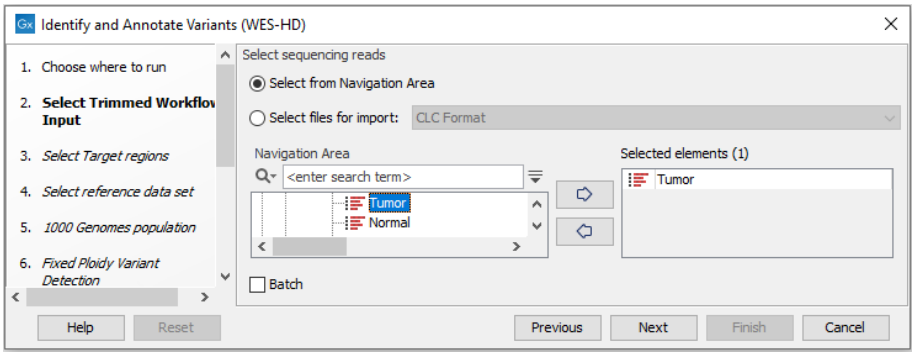


Figure 19.56: *Specify the sequencing reads.*

3. Specify the target regions. (figure 19.57).

   The targeted region file is a file that specifies which regions have been sequenced, when working with whole exome sequencing or targeted amplicon sequencing data. This file is something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents.

4. In the next dialog, you have to select which reference data set should be used in the analysis (figure 19.58).

5. Specify which 1000 Genomes population you would like to use (figure 19.59).

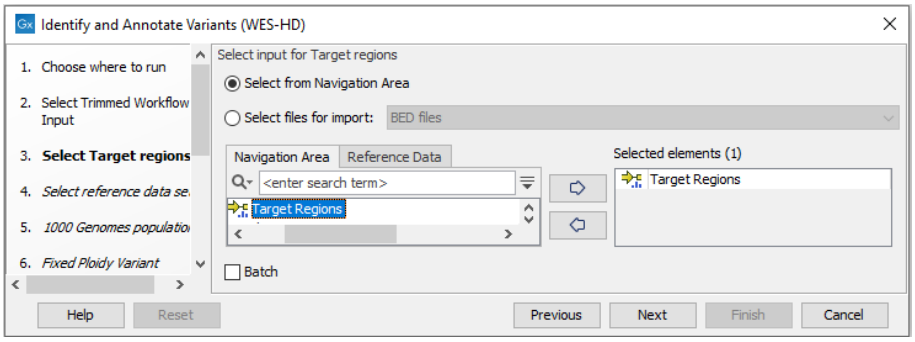


Figure 19.57: *Specify the target regions.*

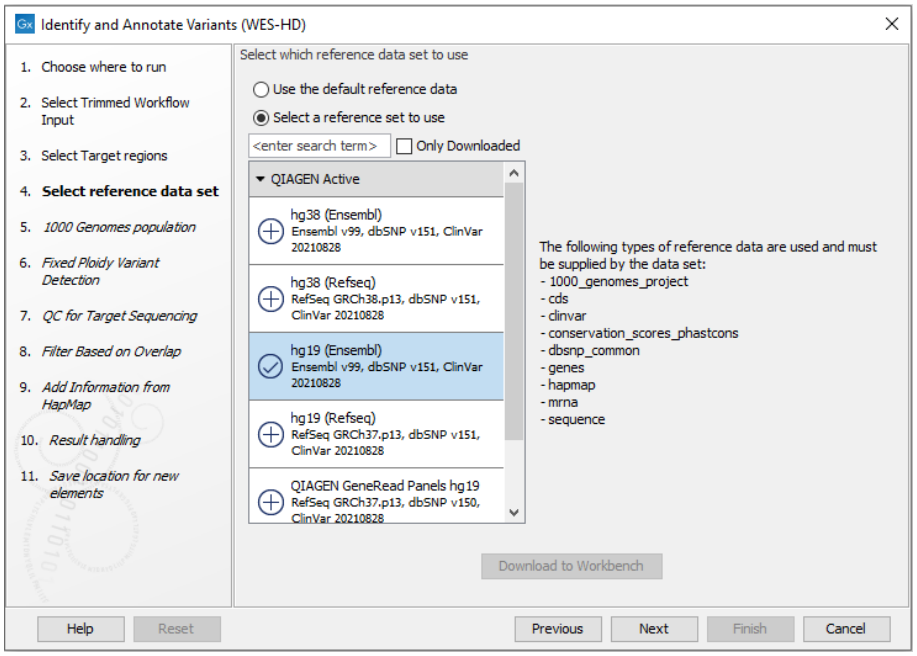


Figure 19.58: *Choose the relevant reference Data Set to identify variants.*

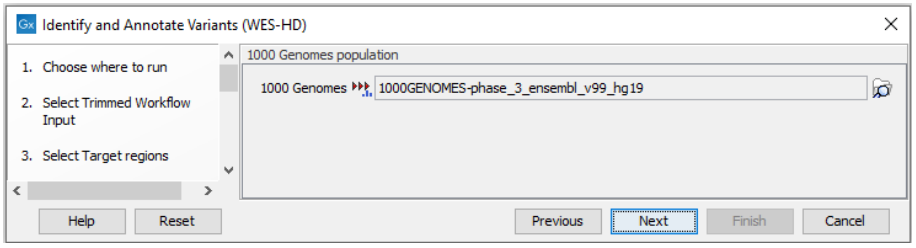


Figure 19.59: *Select the relevant 1000 Genomes population(s).*

6. Specify the Fixed Ploidy Variant Detection settings (figure 19.60).

   The parameters used by the Fixed Ploidy Variant Detection tool can be adjusted. We have optimized the parameters to the individual analyses, but you may want to tweak some of the parameters to fit your particular sequencing data. A good starting point could be to run an analysis with the default settings.

   The parameters that can be set are:

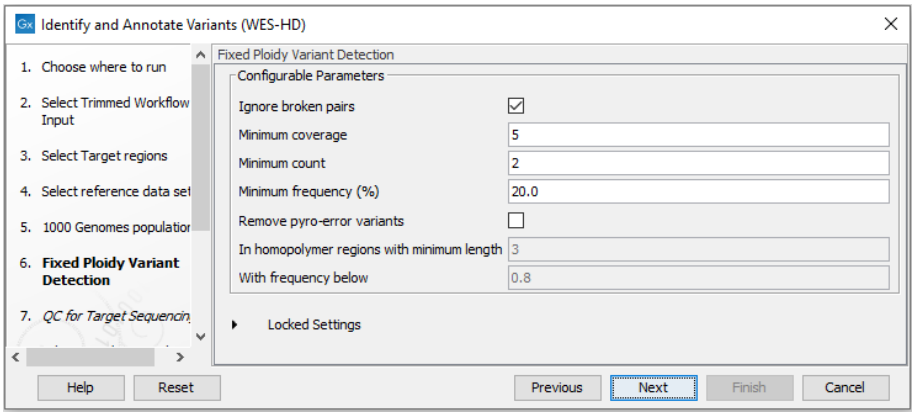


Figure 19.60: *Specify the parameters for the Fixed Ploidy Variant Detection tool.*

- **Required variant probability** is the minimum probability value of the 'variant site' required for the variant to be called. Note that it is not the minimum value of the probability of the individual variant. For the Fixed Ploidy Variant detector, if a variant site - and not the variant itself - passes the variant probability threshold, then the variant with the highest probability at that site will be reported even if the probability of that particular variant might be less than the threshold. For example if the required variant probability is set to 0.9 then the individual probability of the variant called might be less than 0.9 as long as the probability of the entire variant site is greater than 0.9.
- **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected. Please note that ignored broken pair reads will not be considered for any non-specific match filters.
- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

7. Specify the parameters for the QC for Targeted Sequencing tool (figure 19.61).

   The parameters that can be set are:

   - **Minimum coverage** provides the length of each target region that has at least this coverage.
   - **Ignore non-specific matches**: reads that are non-specifically mapped will be ignored.
   - **Ignore broken pairs**: reads that belong to broken pairs will be ignored.

8. Specify the Hapmap population that should be used to add information on variants found in the Hapmap project.

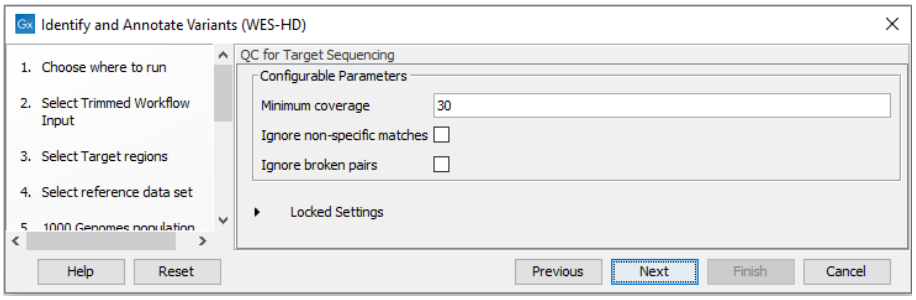


Figure 19.61: *Specify the parameters for the QC for Targeted Sequencing tool.*

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

10. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Identify and Annotate Variants (WES-HD) workflow

- **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

- **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

- **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.

- Two variant tracks ( ): the **Idenitified variants** track contains the variants detected by the Fixed Ploidy Variant Caller, the **Indels indirect evidence** track those detected by the Structural Variant Caller (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_annotated_variant_table.html http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual@@EQUALS@@_annotated_variant_table.html for a definition of the variant table content). The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.

- An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

- **Genome Browser View** ( ) A collection of tracks presented together. Shows the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the indels indirect evidence variants (see figure 19.5).

Chapter 20

Targeted amplicon sequencing (TAS)

Contents

Targeted sequencing, also known as "targeted resequencing" or "amplicon sequencing" is a focused approach to genome sequencing with only selected areas of the genome being sequenced. In cancer research and diagnostics, targeted sequencing is usually based on sequencing panels that target a number of known cancer-associated genes.

A number of template workflows are available for analysis of targeted amplicon sequencing data (figure 20.1). The concept of the pre-installed template workflows is that read data are used as input in one end of the workflow, and after running it, the workflow will output a Track List and a table with all the identified variants, which may or may not have been subjected to different kinds of filtering and/or annotation.

In this chapter we will discuss what the individual template workflows can be used for and go through step by step how to run the workflows.

Remember you will have to prepare data with the **Prepare Raw Data** workflow described in section 2 before you proceed to running any of these workflows.

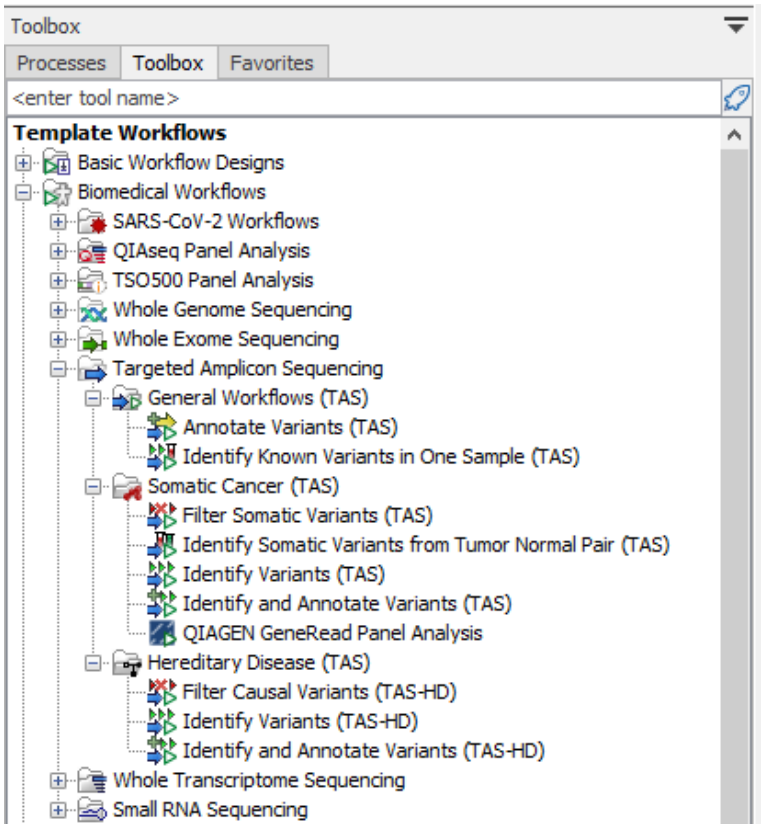


Figure 20.1: *The workflows available for analyzing targeted amplicon sequencing data.*

20.1 General Workflows (TAS)

20.1.1 Annotate Variants (TAS)

The **Annotate Variants (TAS)** template workflow can add the following annotation types to a variant track, annotation track, expression track or statistical comparison track:

- **Gene names** Adds names of genes whenever a variant is found within a known gene.
- **mRNA** Adds names of mRNA whenever a variant is found within a known transcript.
- **CDS** Adds names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Adds information about amino acid changes caused by the variants.
- **Information from ClinVar** Adds information about the relationships between human variations and their clinical significance.
- **Information from dbSNP Common** Adds information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (indels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

Run the Annotate Variants (TAS) workflow

To run the **Annotate Variants (TAS)** template workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** () | **Targeted Amplicon Sequencing** () | **General Workflows (TAS)** () | **Annotate Variants (TAS)** ()

1. Double-click on the **Annotate Variants (TAS)** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. In the first wizard step, select the input variant track, annotation track, expression track or statistical comparison track to annotate (figure 20.2).

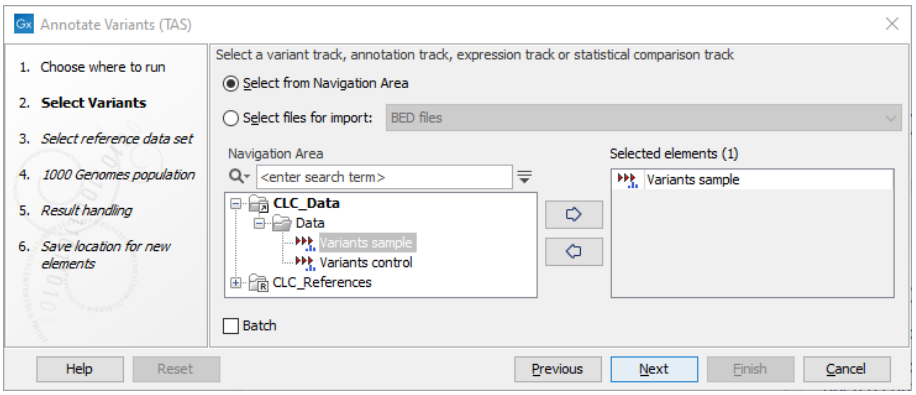


Figure 20.2: *Select the relevant track to annotate.*

3. In the next dialog, you have to select which data set should be used to annotate variants (figure 20.3).

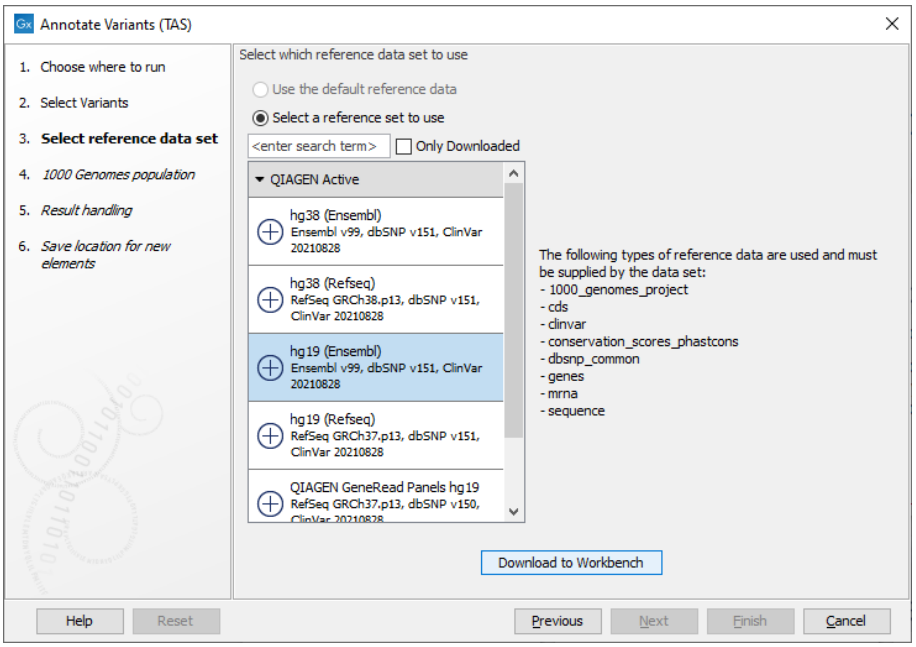


Figure 20.3: *Choose the relevant reference Data Set to annotate.*

4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 20.4).

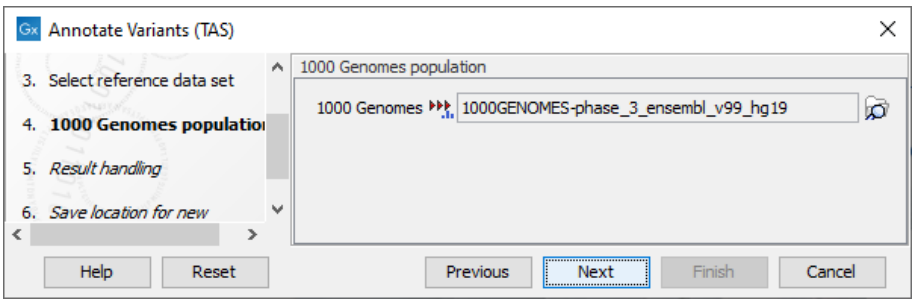


Figure 20.4: *Use the preselected 1000 Genomes population(s) or select another variant track.*

5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

6. Choose to **Save** your results and click **Finish**.

Output from the Annotate Variants (TAS) workflow

The output generated are:

1. **Filtered Annotated Variant Track** ( ) Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

2. An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

3. **Track List Annotated Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP Common, ClinVar, 1000 Genomes, and PhastCons conservation scores (see figure 20.5).

It is possible to add tracks to the Track List by dragging the track directly from the **Navigation Area** to the Track List view. On the other hand, if you delete the annotated variant track, this track will also disappear from the Track List.

Open the annotated track as a table (see figure 20.6). The table and the Track List are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Track List view.

You may be met with a warning as shown in figure 20.7. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in ClinVar or dbSNP Common, the corresponding annotation column headers are missing from the result.

Figure 20.5: *The output from the Annotate Variants template workflow is a track list containing individual tracks for all added annotations.*

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar database, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

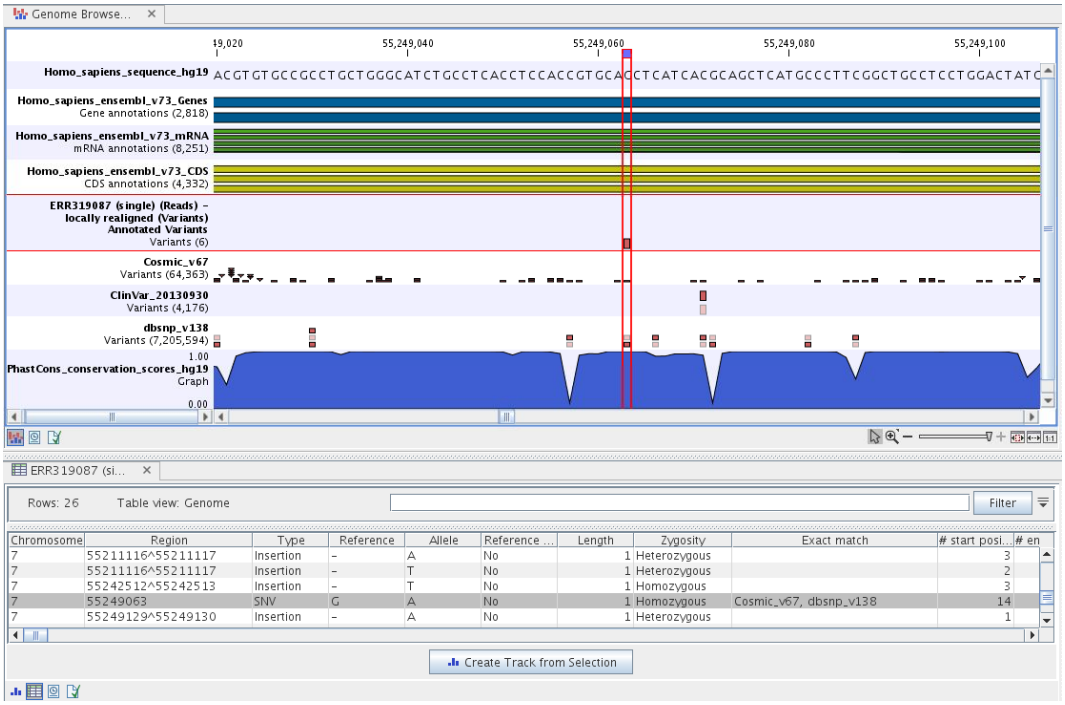


Figure 20.6: *The output from the Annotate Variants template workflow is a track list linked with the variant table view.*

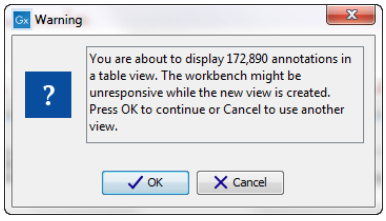


Figure 20.7: *Warning that appears when you work with tracks containing many annotations.*

20.1.2 Identify Known Variants in One Sample (TAS)

The **Identify Known Variants in One Sample (TAS)** template workflow combines data analysis and interpretation. It should be used to identify known variants as specified by the user (e.g., known breast cancer associated variants) for their presence or absence in a sample. This workflow will not identify new variants.

The workflow maps the sequencing reads to a human genome sequence and does a local realignment of the mapped reads to improve the subsequent variant detection. In the next step, only variants specified by the user are identified and annotated in the newly generated read mapping.

Before starting the workflow, you may need to import your the following files with the **Import | Tracks** tool (see `Import | Tracks`):

- **Import your known variants**. Variants can be imported in GVF or VCF format.
- **Import your targeted regions**. A file with the genomic regions targeted by the amplicon or hybridization kit can usually be provided by the vendor, either BED or GFF format.

Run the Identify Known Variants in One Sample (TAS) workflow

1. To run the **Identify Known Variants in One Sample (TAS)** template workflow, go to:

   Workflows |**Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **General Workflows (TAS)** | **Identify Known Variants from One Sample (TAS)** ( )

2. First select the reads of the sample that should be tested for presence or absence of your known variants (figure 20.8).

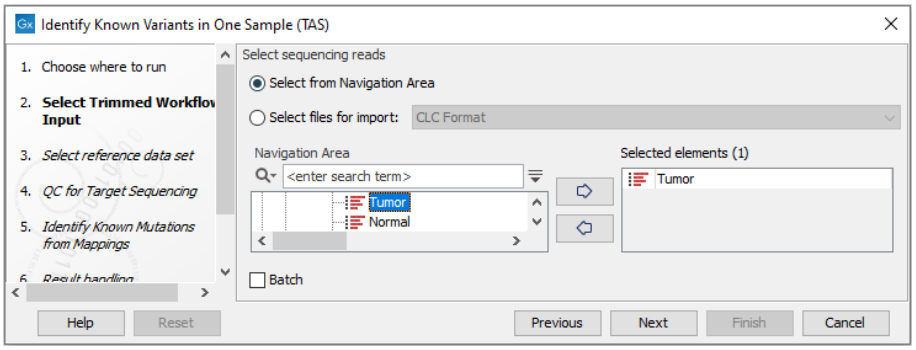


Figure 20.8: *Select the sequencing reads from the sample you would like to test for your known variants.*

   If several samples from different folders should be analyzed, the tool has to be run in batch mode. This is done by selecting "Batch" and specifying the folders that hold the data you wish to analyse.

3. In the next wizard step, select the reference data set should be used to identify the known variants (figure 20.9).

4. Specify the parameters for the QC for Targeted Sequencing tool (figure 20.10).

   When working with targeted data (WES or TAS data), quality checks for the targeted sequencing is included in the workflows. This step is not optional, and you need to specify the targeted regions file adapted to the sequencing technology you used. Choose to use the default settings or to adjust the parameters.

   The parameters that can be set are:

   - **Minimum coverage** provides the length of each target region that has at least this coverage.
   - **Ignore non-specific matches**: reads that are non-specifically mapped will be ignored.
   - **Ignore broken pairs**: reads that belong to broken pairs will be ignored.

5. In the Identify Known Mutations form Mappings, select a variant track containing the known variants you want to identify in the sample (figure 20.11).

   The parameters that can be set are:

   - **Minimum coverage** The minimum number of reads that covers the position of the variant, which is required to set "Sufficient Coverage" to YES.

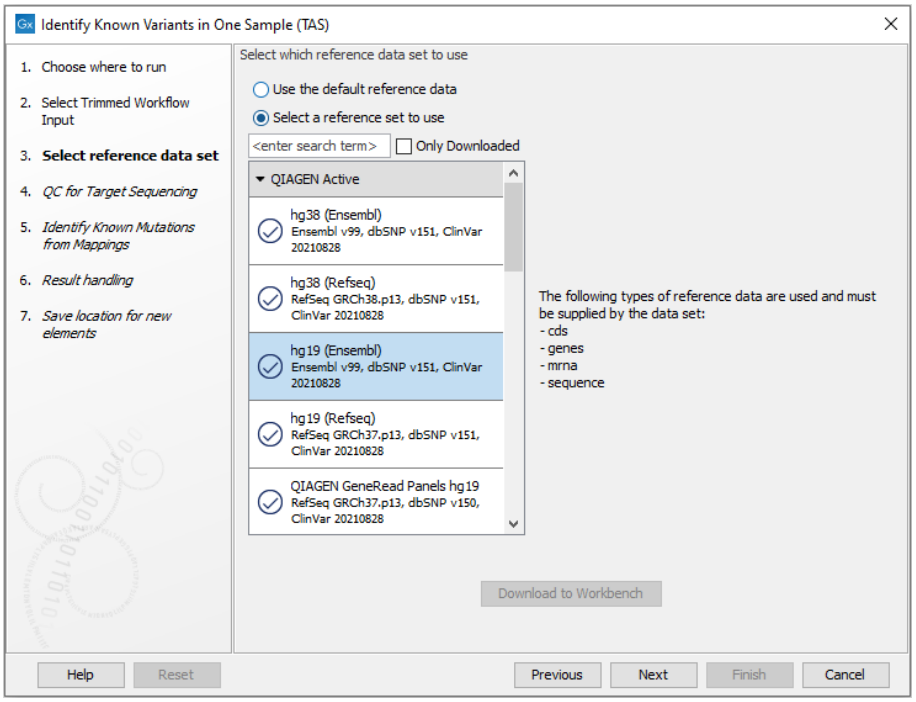


Figure 20.9: *Choose the relevant reference Data Set to identify the known variants.*

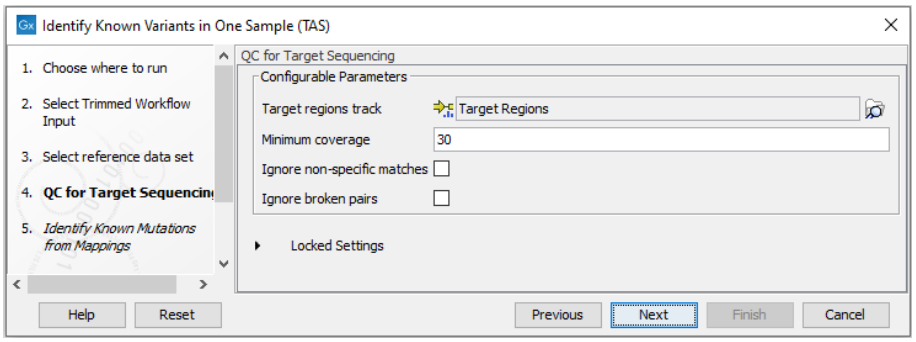


Figure 20.10: *Specify the parameters for the QC for Targeted Sequencing tool.*

- **Detection frequency** The minimum allele frequency that is required to annotate a variant as being present in the sample. The same threshold will also be used to determine if a variant is homozygous or heterozygous. In case the most frequent alternative allele at the position of the considered variant has a frequency of less than this value, the zygosity of the considered variant will be reported as being homozygous.

The parameter "Detection Frequency" will be used in the calculation twice. First, it will report in the result if a variant has been detected (observed frequency > specified frequency) or not (observed frequency <= specified frequency). Moreover, it will determine if a variant should be labeled as heterozygous (frequency of another allele identified at a position of a variant in the alignment > specified frequency) or homozygous (frequency of all other alleles identified at a position of a variant in the alignment < specified frequency).

6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the

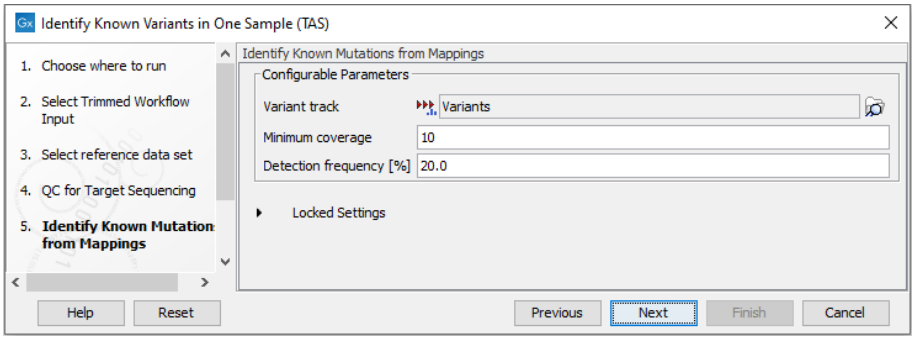


Figure 20.11: *Specify the track with the known variants that should be identified.*

settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

7. Choose to **Save** your results and click **Finish**.

Output from the Identify Known Variants in One Sample (TAS)

The **Identify Known Variants in One Sample (TAS)** tool produces five different output types:

1. **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

2. **Target Regions Coverage** ( ) A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.

3. **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways show e.g. the number, length, and coverage of the target regions and provides information about the read count per GC%.

4. **Variants Detected in Detail** ( ) Annotation track showing the known variants. Like the "Overview Variants Detected" table, this table provides information about the known variants. Four columns starting with the sample name and followed by "Read Mapping coverage", "Read Mapping detection", "Read Mapping frequency", and "Read Mapping zygosity" provides the overview of whether or not the known variants have been detected in the sequencing reads, as well as detailed information about the Most Frequent Alternative Allele (labeled MFAA).

5. **Track List Identify Known Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, target regions coverage, the mapped reads, the overview of the detected variants, and the variants detected in detail.

It is a good idea to start looking at the Target Regions Coverage Report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Please also check that at least 90%

of the reads are mapped to the human reference sequence. In case of a targeted experiment, we also recommend that you check that the majority of the reads are mapping to the targeted region.

When you have inspected the target regions coverage report you can open the Track List Identify Known Variants file (see 20.12).

The Track List includes an overview track of the known variants and a detailed result track presented in the context of the human reference sequence, genes, transcripts, coding regions, targeted regions, and mapped sequencing reads.

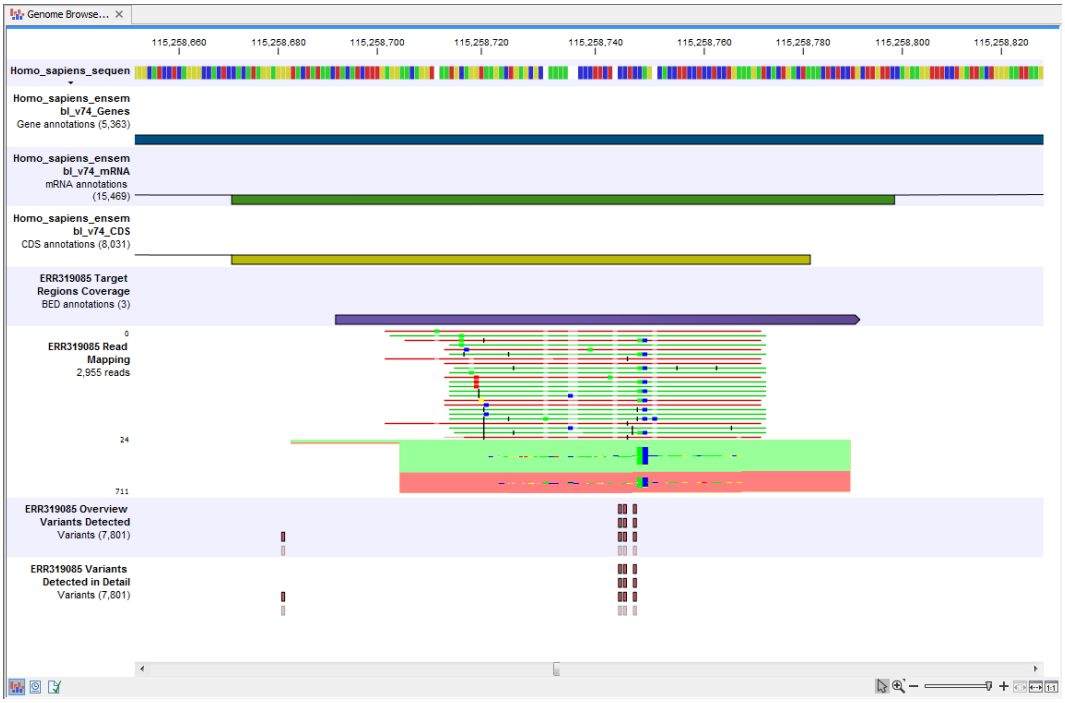


Figure 20.12: *Track List that allows inspection of the identified variants in the context of the human genome and external databases.*

Finally, a track with conservation scores has been added to be able to see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant.

Open the annotated variant as a table showing all variants and the added information/annotations (see 20.13).

Note We do not recommend that any of the produced files are deleted individually as some of them are linked to other outputs. Please always delete all of them at the same time.

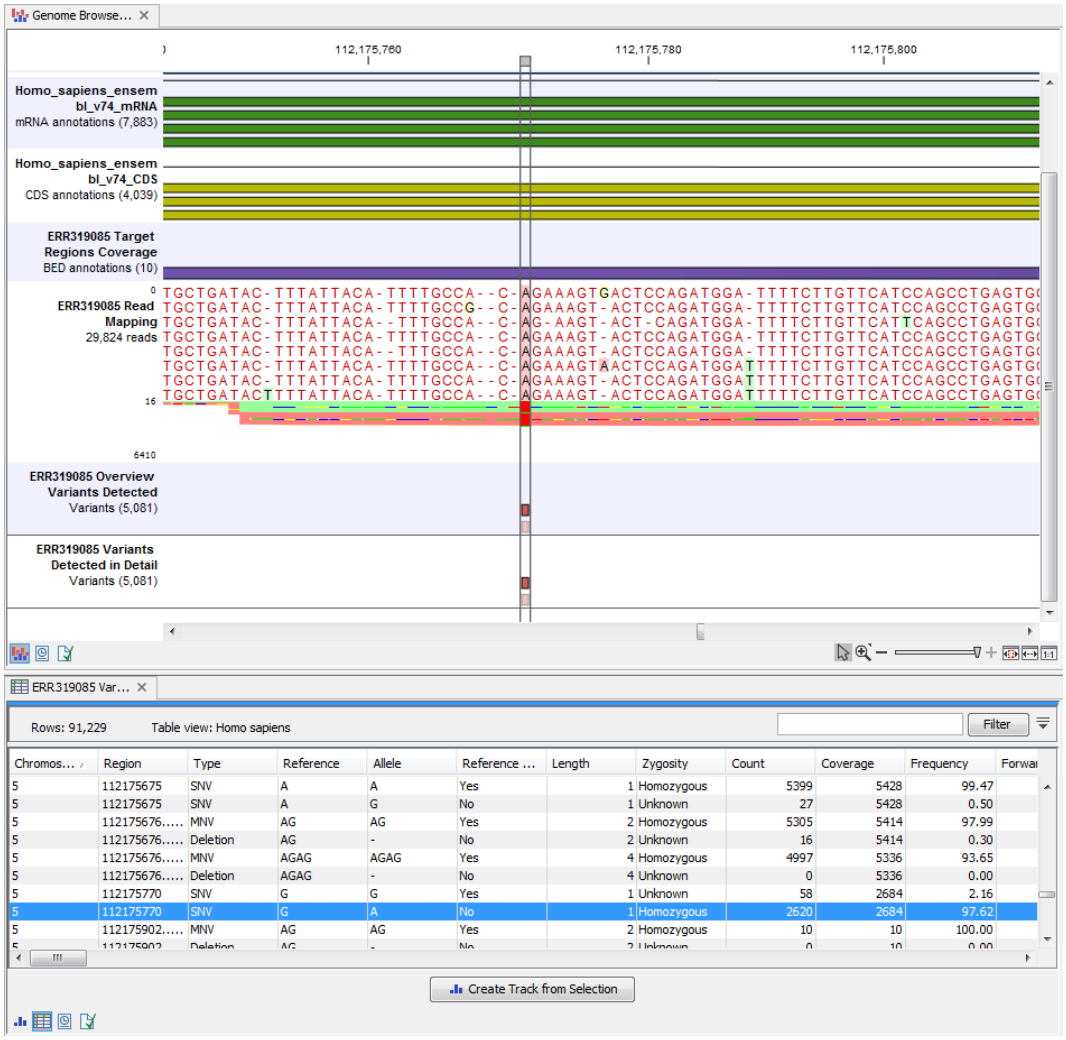


Figure 20.13: *Track List with an open overview variant track with information about if the variant has been detected or not, the identified zygosity, if the coverage was sufficient at this position and the observed allele frequency.*

20.2 Somatic Cancer (TAS)

20.2.1 Filter Somatic Variants (TAS)

If you are analyzing a list of variants that have been detected in a tumor or blood sample where no control sample is available from the same subject, you can use the **Filter Somatic Variants (TAS)** template workflow to identify potential somatic variants. The purpose of this template workflow is to use publicly available (or your own) databases, with common variants in a population, to extract potential somatic variants whenever no control/normal sample from the same subject is available.

This workflow accepts variant tracks ( ) (e.g. the output from the Identify Variants template workflow) as input. Variants that are identical to the human reference sequence are first filtered away, then variants outside the targeted region are removed, and lastly, variants found in the Common dbSNP, 1000 Genomes Project, and HapMap databases are deleted. Variants in those databases are assumed to not contain relevant somatic variants.

Please note that this tool will likely also remove inherited cancer variants that are present at a low percentage in a population.

Next, the remaining somatic variants are annotated with gene names, amino acid changes, conservation scores and information from ClinVar (known variants with medical impact) and dbSNP (all known variants).

20.2.2 Run the Filter Somatic Variants (TAS) workflow

To run the **Filter Somatic Variants (TAS)**, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **Somatic Cancer** ( ) | **Filter Somatic Variants** ( )

1. Double-click on the **Filter Somatic Variants (TAS)** tool to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Next, you will be asked to select the variant track you would like to use for filtering somatic variants (figure 20.14).

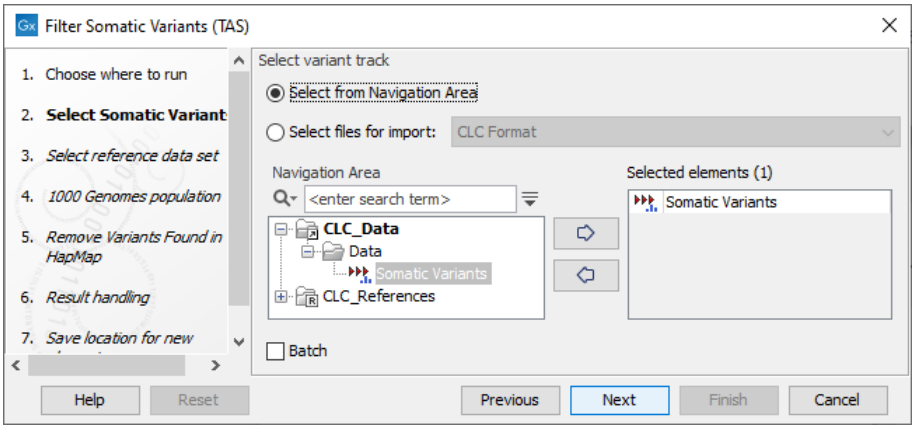


Figure 20.14: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter somatic variants (figure 20.15).
4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track to use (figure 20.16).
5. For databases that provide data from more than one population as HapMap does, the populations relevant to the data set can be specified. Click on the plus symbol ( ) and choose the population that matches the population your samples are derived from (figure 20.17).Please note that different populations are available and can be downloaded via the Reference Data Manager found in the top right corner of the CLC Workbench.
6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

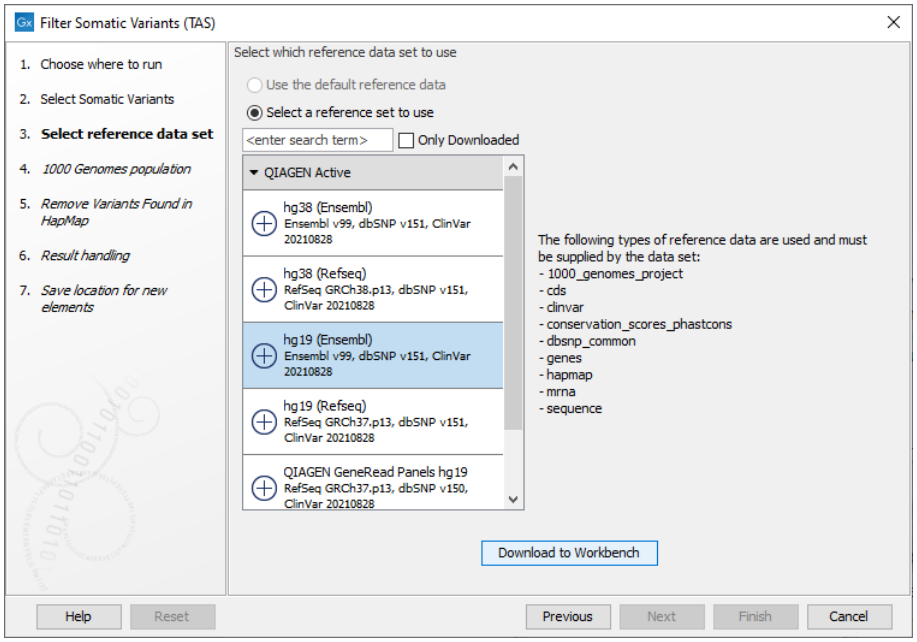


Figure 20.15: *Choose the relevant reference Data Set to annotate.*

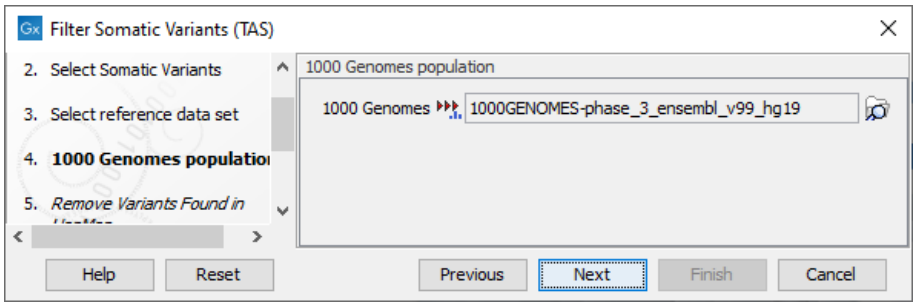


Figure 20.16: *Use the preselected 1000 Genomes population(s) or select another variant track.*

7. Choose to **Save** your results and click **Finish**.

20.2.3 Output from the Filter Somatic Variants (TAS) workflow

Two types of output are generated:

1. **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.

2. **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Track List. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.

3. **Track List Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes,

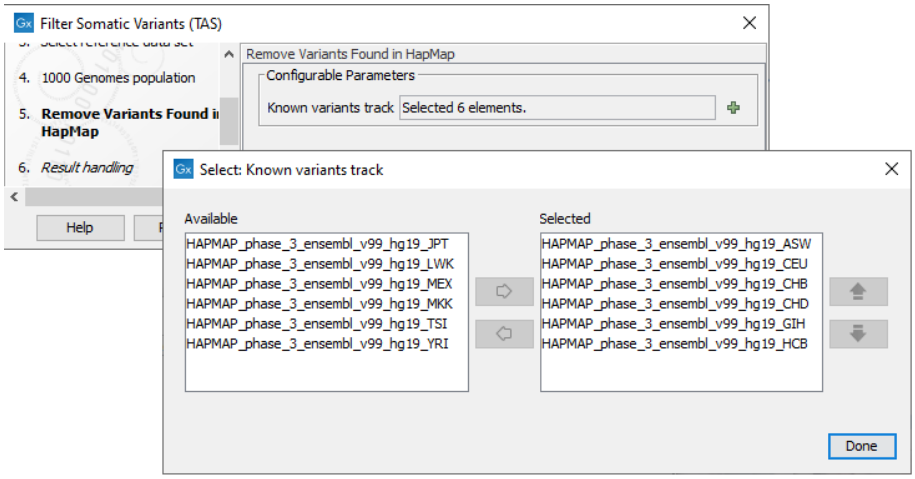


Figure 20.17: *Specify which HapMap population to use for filtering out known variants.*

transcripts, coding regions, and variants detected in ClinVar, 1000 Genomes, and the PhastCons conservation scores (see figure 20.18).

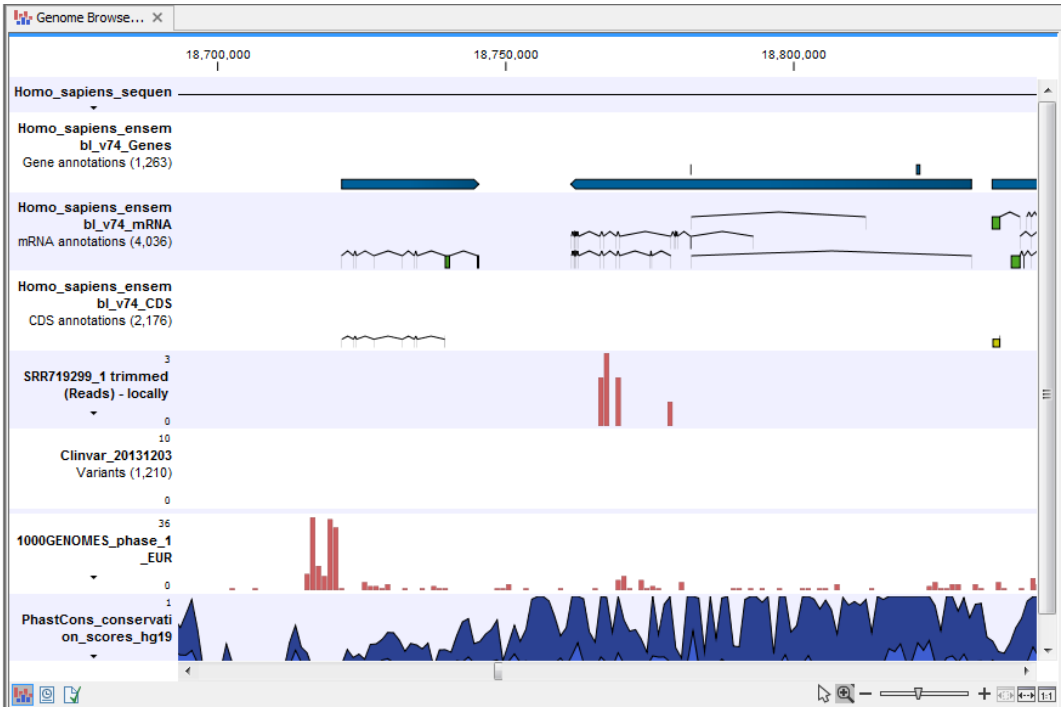


Figure 20.18: *The Track List showing the annotated somatic variants together with a range of other tracks.*

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well. Mapped sequencing reads as well as other tracks can be easily added to this Track List. Open the variant track as a table showing all variants and the added information/annotations (see figure 20.19).

Adding information from other sources may help you identify interesting candidate variants for further research. E.g. common genetic variants (present in the HapMap database) or variants

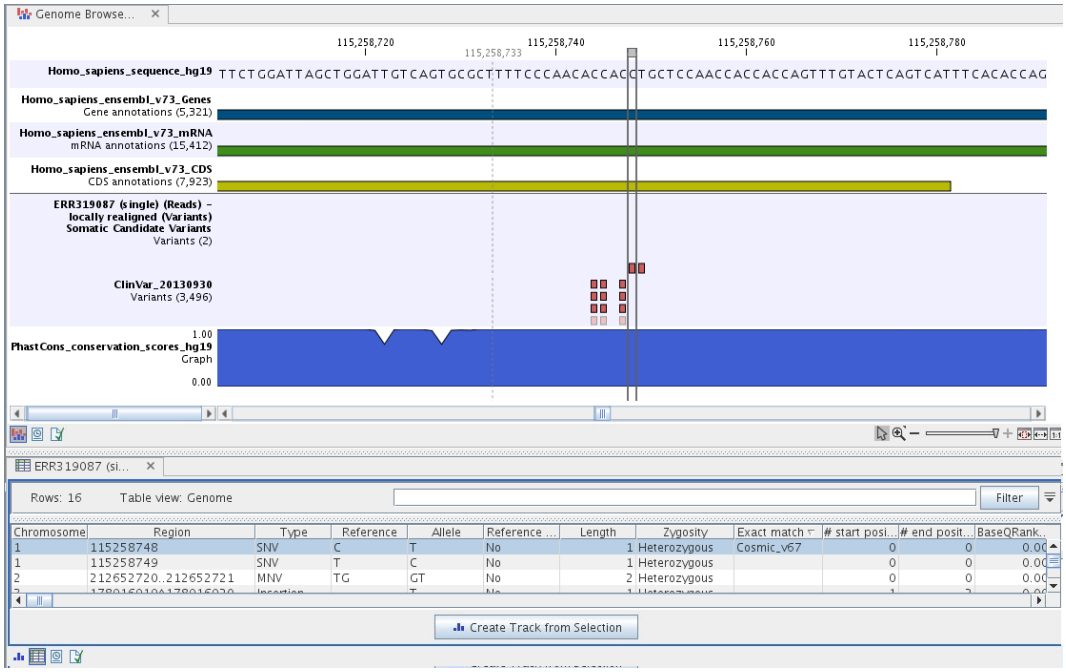


Figure 20.19: *The Track List showing the annotated somatic variants together with a range of other tracks.*

known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar databases, can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can include in a workflow the Filter on Custom Criteria tool configured with the desired set of criteria.

20.2.4 Identify Somatic Variants from Tumor Normal Pair (TAS)

The **Identify Somatic Variants from Tumor Normal Pair (TAS)** template workflow can be used to identify potential somatic variants in a tumor sample when you also have a normal/control sample from the same individual.

When running this workflow the reads are mapped and the variants identified. Germline variants that are found in the mapped reads of the normal/control sample and variants outside the target region are removed as they are likely to be false positives due to non-specific mapping of sequencing reads. Next, remaining variants are annotated with gene names, amino acid changes, conservation scores and information from relevant databases like ClinVar (variants with clinically

relevant association). Finally, information from dbSNP is added to see which of the detected variants have been observed before and which are completely new.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

Run the Identify Somatic Variants from Tumor Normal Pair (TAS) workflow

1. To run the **Identify Somatic Variants from Tumor Normal Pair (TAS)** template workflow, go to:

   Workflows | Template Workflows | Biomedical Workflows ( ) | **Targeted Amplicon Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Somatic Variants from Tumor Normal Pair (TAS)** ( )

2. First, (figure 20.20), select the tumor sample reads.

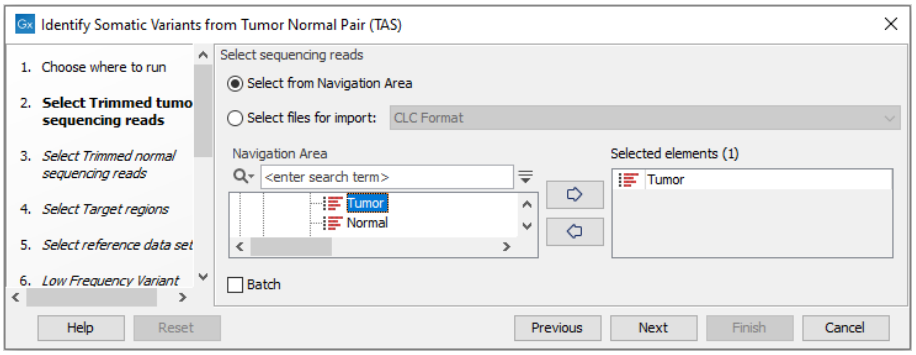


Figure 20.20: *Select the tumor sample reads.*

3. In the next wizard step, specify the normal sample reads.

4. Next, specify the target regions. Variants are only called in the target regions, both for tumor and normal reads (figure 20.21).

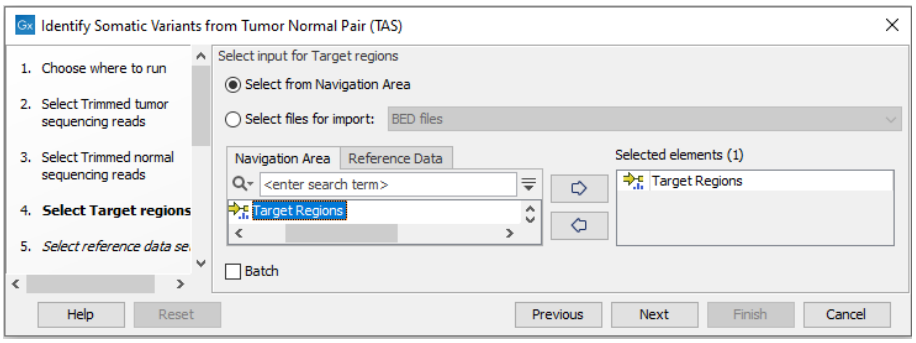


Figure 20.21: *Specify the target regions track.*

5. In the next dialog, select which reference data set should be used to identify variants (figure 20.22).

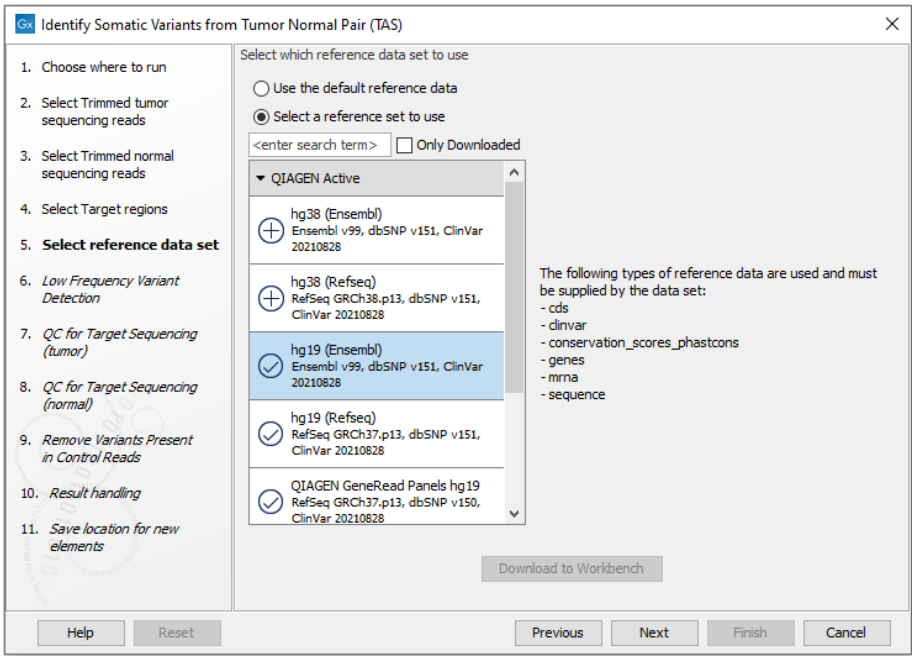


Figure 20.22: *Choose the relevant reference Data Set to identify variants.*

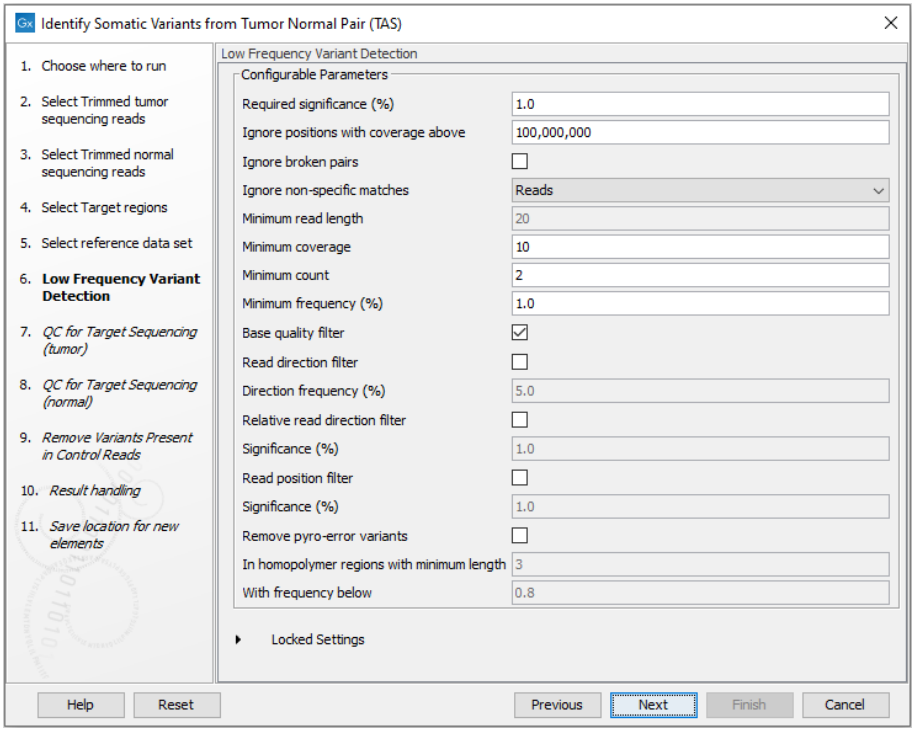


Figure 20.23: *Specify the settings for the variant detection.*

6. Set the parameters for the Low Frequency Variant Detection step (figure 20.23).

   For a description of the different parameters that can be adjusted, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html. If you click on "Locked Settings", you will be able to see all param-

eters used for variant detection in the template workflow.

7. Specify the parameters for the QC for Targeted Sequencing tool (figure 20.24). The target regions track is used for reporting the performance of the targeted re-sequencing experiment for both the tumor and normal sample.

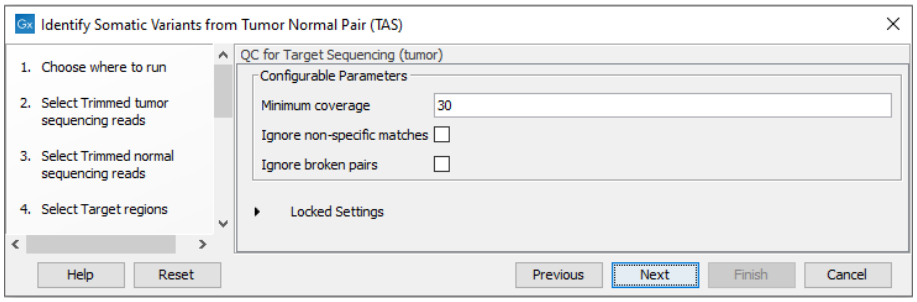


Figure 20.24: *Set the parameters for the QC for targeted regions.*

For a description of the different parameters that can be adjusted, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html. If you click on "Locked Settings", you will be able to see all parameters used for the QC for Targeted Sequencing tool in the template workflow.

8. In the Remove Variants Present in Control Reads step, you can adjust the settings for removal of germline variants (figure 20.25).

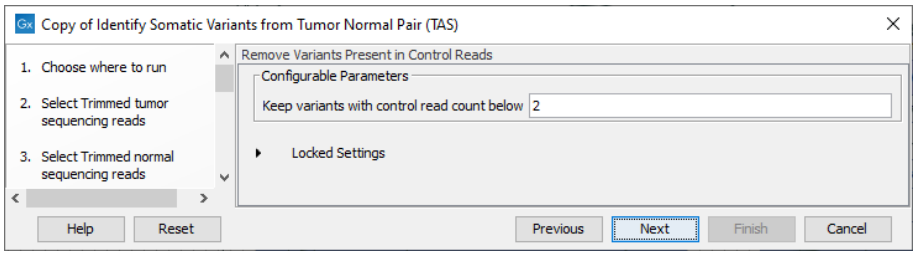


Figure 20.25: *Specify setting for removal of germline variants.*

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

10. Choose to **Save** your results and click **Finish**.

Output from the Identify Somatic Variants from Tumor Normal Pair (TAS) workflow

The following outputs are generated:

- **Read Mapping Normal** ( ) The mapped sequencing reads for the normal sample. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

- **Read Mapping Tumor** ( ) The mapped sequencing reads for the tumor sample.
- **Target Region Coverage Report Normal** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the normal sample.
- **Target Region Coverage Tumor** ( ) A track showing the targeted regions. The table view provides information about the targeted regions such as target region length, coverage, regions without coverage, and GC content.
- **Target Region Coverage Report Tumor** ( ) The report consists of a number of tables and graphs that in different ways provide information about the mapped reads from the tumor sample.
- **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
- **Variants** ( ) A variant track holding the identified variants that are found in the targeted regions. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
- **Annotated Somatic Variants** ( ) A variant track holding the identified and annotated somatic variants. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
- **Track List Tumor Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads for both normal and tumor, the annotated somatic variants, information from the ClinVar database, and finally a track showing the conservation score (see figure 20.26).

20.2.5 Identify Variants (TAS)

The **Identify Variants (TAS)** template workflow takes sequencing reads as input and returns identified variants as part of a Track List.

Sequencing reads provided as input are initially mapped to the human reference sequence. This is followed by the removal of duplicate mapped reads (to reduce biases introduced by target enrichment). The resulting read mapping is analyzed by the Structural Variant Caller to infer indels and other structural variants from unaligned end read patterns. Subsequently, the mapping is realigned, guided by the indels detected by the Structural Variant Caller. The locally realigned read mapping is analyzed by the Low Frequency Variant Detection tool. The Low Frequency Variant Detection tool produces a track of unfiltered variants; these are subjected to a number of post filters to remove variants that are likely due to artifacts or noise. The variants called by the Low Frequency Variant Detection tool that pass the post filtering criteria can be found in the Identified variants track. Variants inferred by the Structural Variant Caller, and not detected by the Low

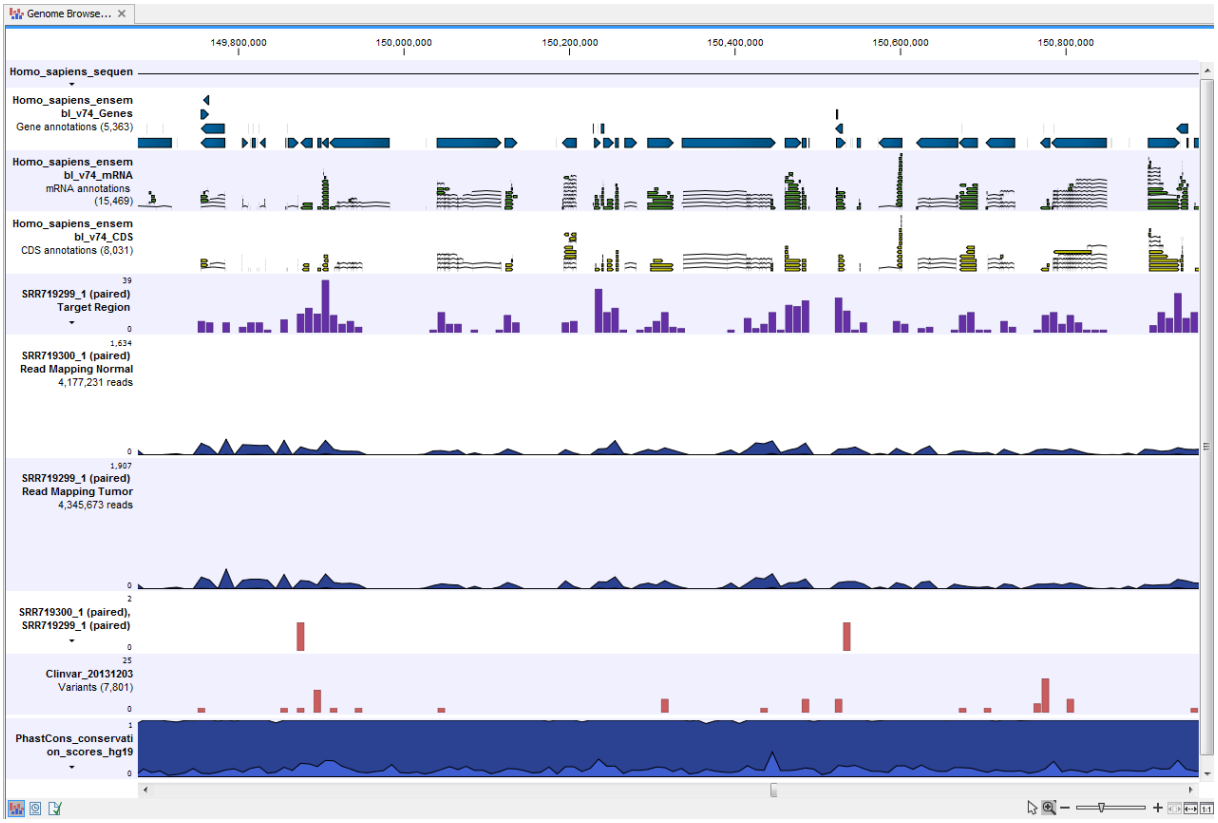


Figure 20.26: *The Track List presents all the different data tracks together and makes it easy to compare different tracks.*

Frequency Variant Detection tool, are also subjected to a number of post filters; those that pass the post filter criteria can be found in the Indels indirect evidence track.

In addition, a targeted region report is created to inspect the overall coverage and mapping specificity in the targeted regions.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

Run the Identify Variants (TAS) workflow

1. To run the **Identify Variants (TAS)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Variants (TAS)** ( )

2. Select the sequencing reads from the sample that should be analyzed (figure 20.27).

   If several samples should be analyzed, the tool has to be run in batch mode. This is done by checking "Batch" and selecting the **folder** that holds the data you wish to analyze.

3. In the Target regions dialog you specify the target regions for your application (figure 20.28). The variant calling will be restricted to these regions.

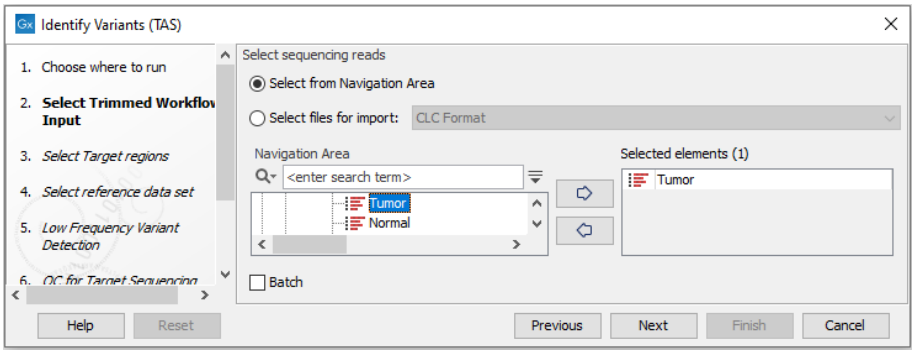


Figure 20.27: *Please select all sequencing reads from the sample to be analyzed.*

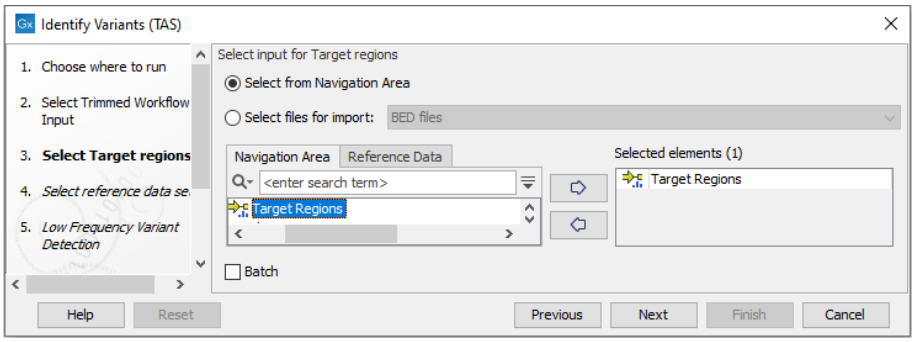


Figure 20.28: *Select the track with the targeted regions from your experiment.*

4. In the next dialog, you have to select which reference data set should be used to identify variants (figure 20.29).

5. In the next wizard step (figure 20.30) you can specify the parameters for variant detection.

6. In the next wizard step (figure 20.31) you specify the parameters for the QC reporting on the targeted regions.

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click **Finish**.

Output from the Identify Variants (TAS) workflow

The **Identify Variants (TAS)** tool produces five different types of output:

- **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html`).

- **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the

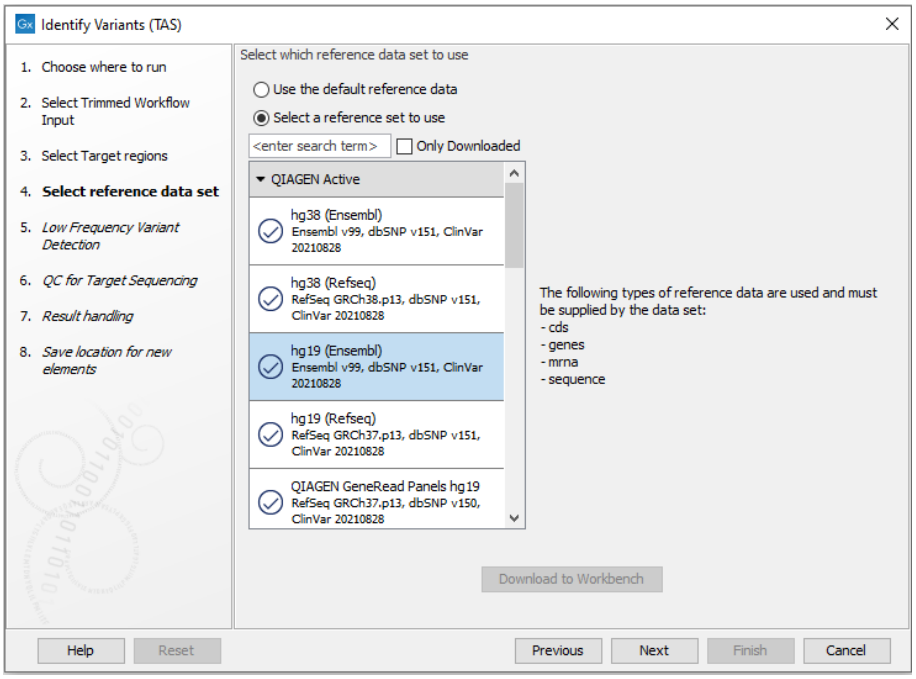


Figure 20.29: *Choose the relevant reference Data Set to identify variants in your sample.*

table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.

- **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.
- **Identified Variants and Indels Indirect Evidence** ( ) A variant track containing the variants identified by the Low Frequency Variant Detection tool, and a variant track with the indels inferred from indirect evidence by the Structural Variant Caller. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
- **Genome Browser View Identify Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 20.32).

It is important that you do not delete any of the produced files individually as some of the outputs are linked to other outputs. If you would like to delete the outputs, please always delete all of them at the same time.

First have a look at the mapping report to see if the coverage is sufficient in regions of interest (e.g. > 30). Furthermore, check that at least 90% of reads are mapped to the human reference sequence. In case of a targeted experiment, also check that the majority of reads are mapping to the targeted region.

Afterwards please open the Genome Browser View file (see 20.32).

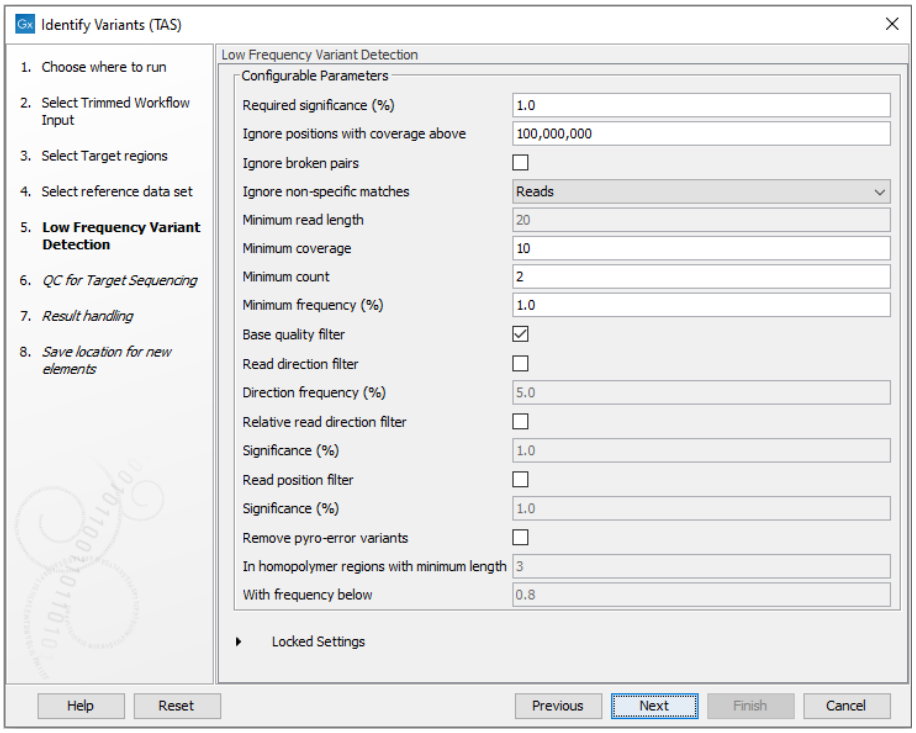


Figure 20.30: *Specify the parameters for variant detection.*

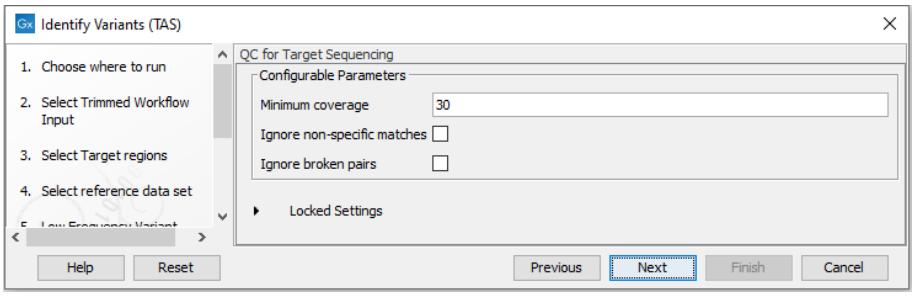


Figure 20.31: *Specify minimum coverage for the QC reporting on the targeted regions.*

The Genome Browser View includes the track of identified variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions and mapped sequencing reads.

Open the variant track as a table to see information about all identified variants (see 20.33).

20.2.6 Identify and Annotate Variants (TAS)

The **Identify and Annotate Variants (TAS)** template workflow should be used to identify and annotate variants in one sample. The workflow is a combination of the **Identify Variants** and the **Annotate Variants** workflows.

The workflow starts with mapping the sequencing reads to the human reference sequence, followed by a local realignment to improve the variant detection that is run afterwards. After the variants have been detected, they are annotated with gene names, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and

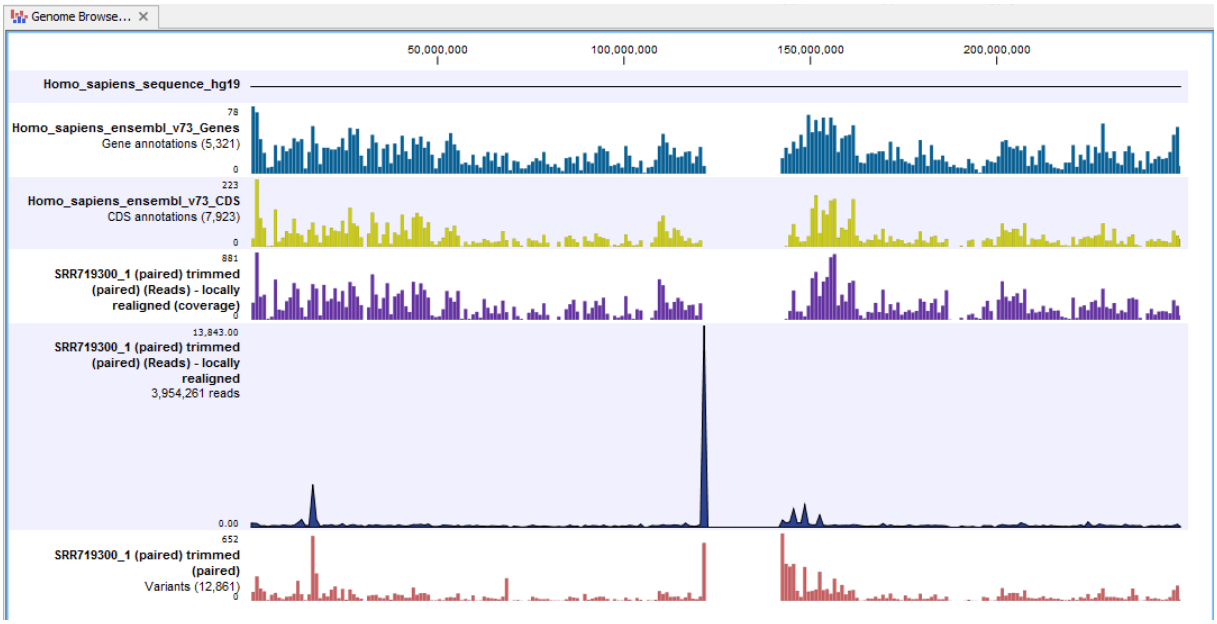


Figure 20.32: *The Genome Browser View allows you to inspect the identified variants in the context of the human genome.*

information from common variants present in the common dbSNP Common, HapMap, and 1000 Genomes database. Furthermore, a detailed targeted region coverage report is created to inspect the overall coverage and mapping specificity.

Before starting the workflow, you will need to import in the CLC Workbench a file with the genomic regions targeted by the amplicon or hybridization kit. Such a file (a BED or GFF file) is usually available from the vendor of the enrichment kit and sequencing machine. Use the **Import | Tracks** tool to import it in your Navigation Area.

Run the Identify and Annotate Variants (TAS) workflow

To run the **Identify and Annotate Variants (TAS)** workflow, go to:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify and annotate Variants (TAS)** ( )

1. Double-click on the workflow name to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. First select the sequencing reads from the sample that should be analyzed (figure 20.34).

   If several samples should be analyzed, the tool has to be run in batch mode. This is done by checking "Batch" and selecting the **folder** that holds the data you wish to analyze.
3. In the Target regions dialog (figure 20.35), you can specify the target regions track. Variants will only be called in the target regions.
4. In the next dialog, you have to select which reference data set should be used in the analysis (figure 20.36).

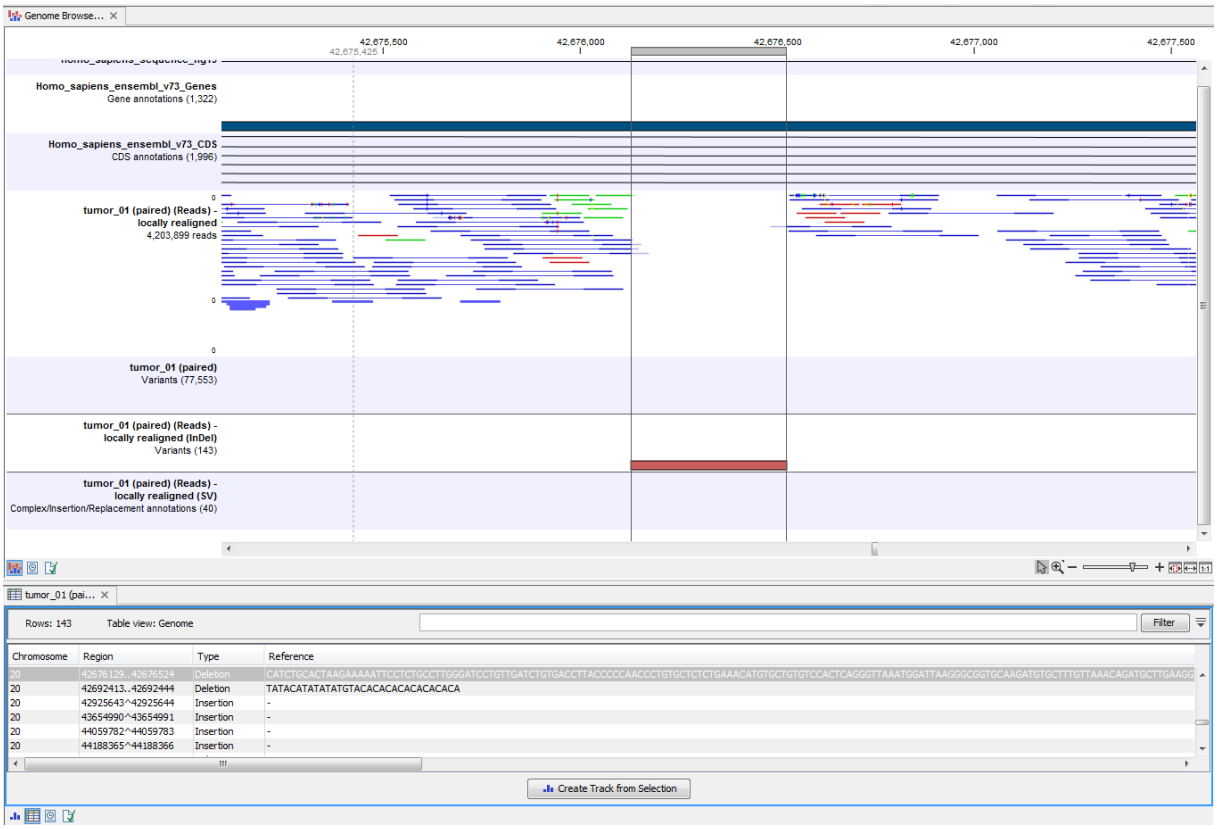


Figure 20.33: *Genome Browser View with an open track table to inspect identified variants more closely in the context of the human genome.*

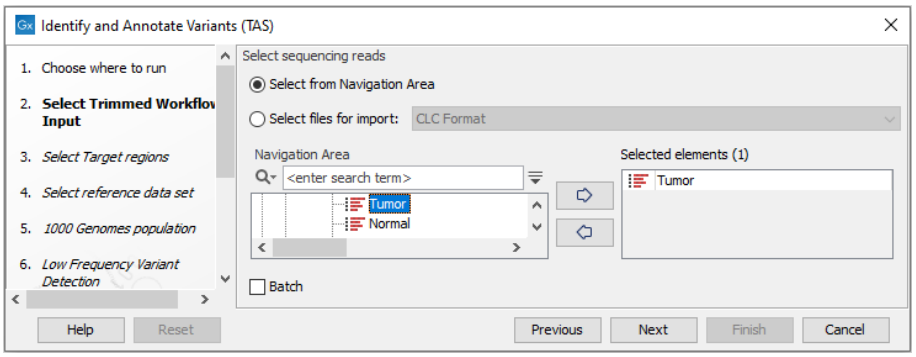


Figure 20.34: *Select all sequencing reads from the sample to be analyzed.*

5. In the next wizard step (figure 20.37) you can select the population from the 1000 Genomes project that you would like to use for annotation.

6. In the next dialog (figure 20.38), you have to specify the parameters for the variant detection.

   For a description of the different parameters that can be adjusted, see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html`. If you click on "Locked Settings", you will be able to see all parameters used for variant detection in the template workflow.

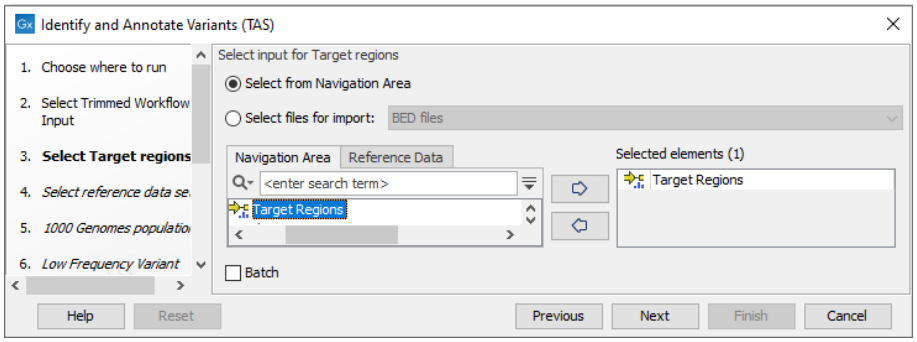


Figure 20.35: *In this wizard step you can specify the target regions track. Variants will not be called outside these regions.*

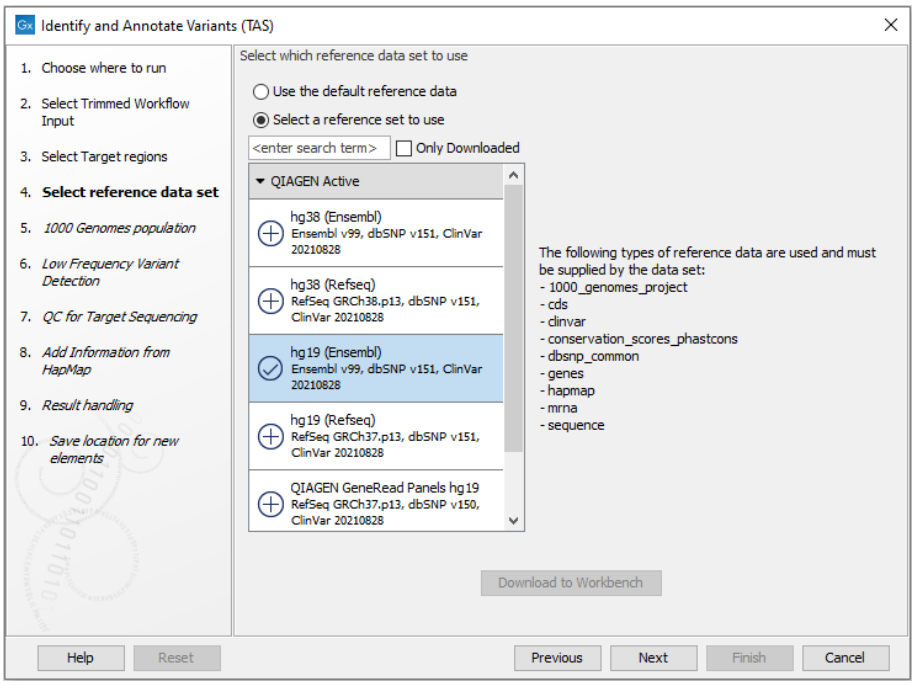


Figure 20.36: *Choose the relevant reference Data Set to identify and annotate.*

7. In the QC for Target Sequencing step (figure 20.39) you can specify the minimum read coverage that should be present in the targeted regions.

8. Finally, select a population from the HapMap database (figure 20.40). This will add information from the Hapmap database to your variants.

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

10. Choose to **Save** your results and click **Finish**.

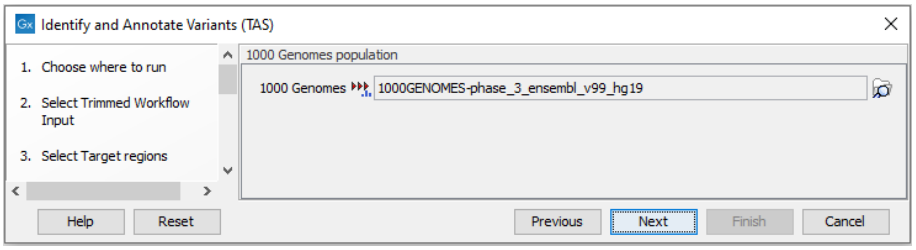


Figure 20.37: *Select the population from the 1000 Genomes project that you would like to use for annotation.*

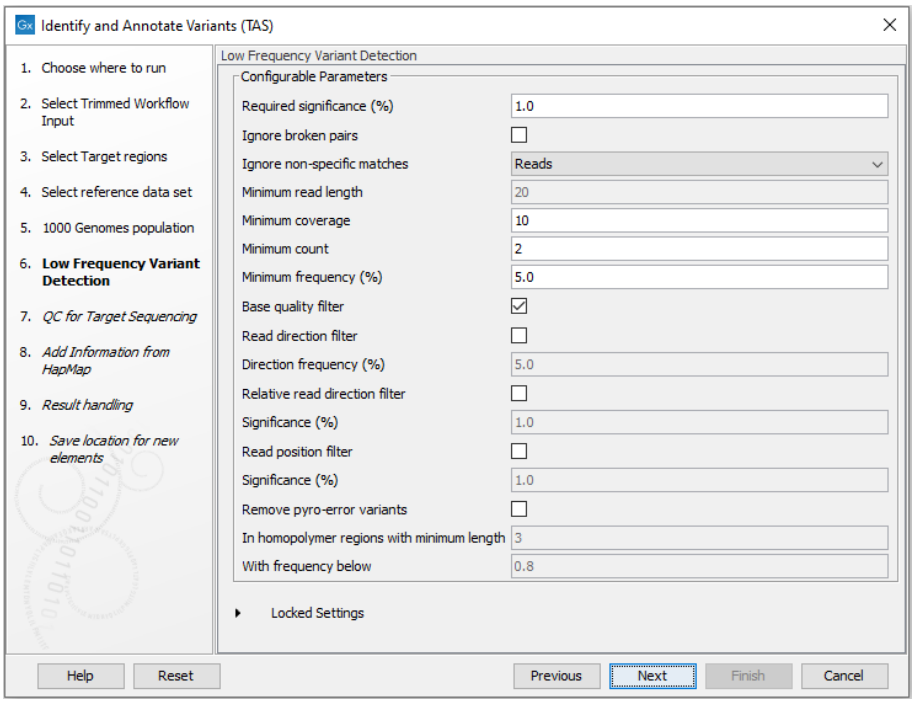


Figure 20.38: *Specify the parameters for variant calling.*

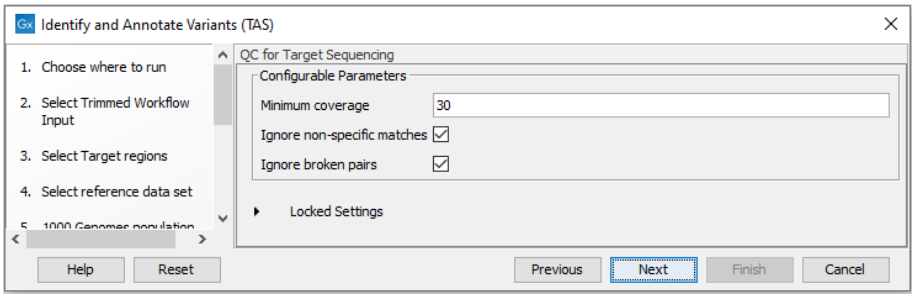


Figure 20.39: *Set the parameters for the QC for targeted regions.*

Output from the Identify and Annotate Variants (TAS) workflow

The **Identify and Annotate Variants (TAS)** tool produces several outputs.

Please do not delete any of the produced files alone as some of them are linked to other outputs. Please always delete all of them at the same time.

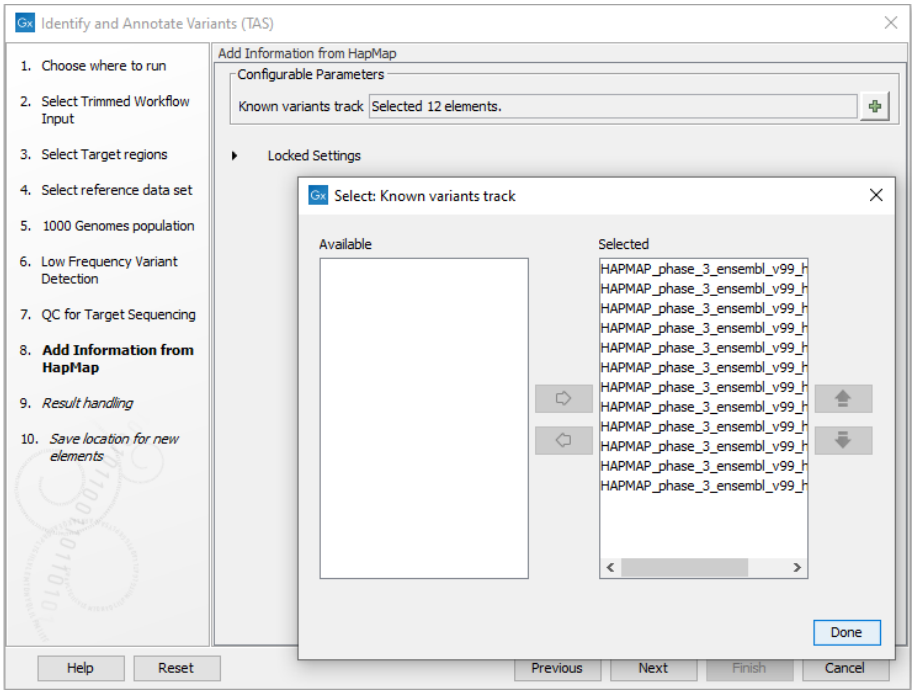


Figure 20.40: *Select a population from the HapMap database to add information from the Hapmap database to your variants.*

A good place to start is to take a look at the mapping report to see whether the coverage is sufficient in the regions of interest (e.g. > 30). Furthermore, please check that at least 90% of the reads are mapped to the human reference sequence. In case of a targeted experiment, please also check that the majority of the reads are mapping to the targeted region.

Next, open the Track List file (see figure 20.41).

The Track List includes a track of the identified annotated variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, relevant variants in the ClinVar database as well as common variants in common dbSNP Common, HapMap, and 1000 Genomes databases.

To see the level of nucleotide conservation (from a multiple alignment with many vertebrates) in the region around each variant, a track with conservation scores is added as well.

By double-clicking on the annotated variant track in the Track List, a table will be shown that includes all variants and the added information/annotations (see figure 20.42).

The added information will help you to identify candidate variants for further research. For example can common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) easily be seen.

Not identified variants in ClinVar, can for example be prioritized based on amino acid changes (do they cause any changes on the amino acid level?). A high conservation level on the position of the variant between many vertebrates or mammals can also be a hint that this region could have an important functional role and variants with a conservation score of more than 0.9 (PhastCons score) should be prioritized higher. A further filtering of the variants based on their annotations

Figure 20.41: *Track List to inspect identified variants in the context of the human genome and external databases.*

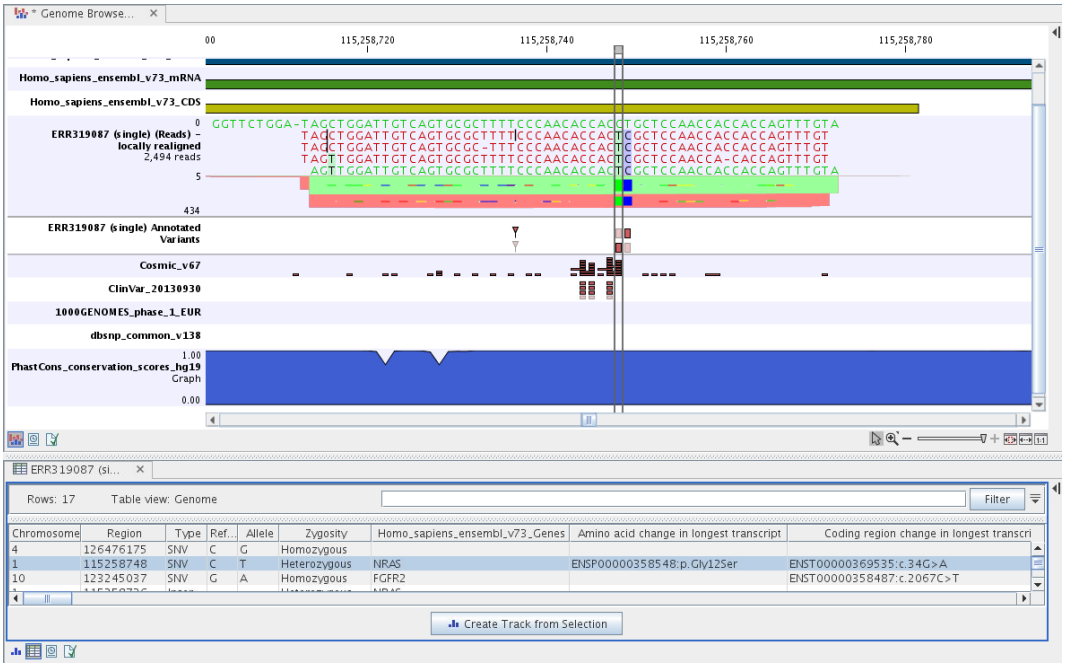


Figure 20.42: *Track List with an open track table to inspect identified somatic variants more closely in the context of the human genome and external databases.*

can be facilitated using the table filter on top of the table.

If you wish to always apply the same filter criteria, the "Create new Filter Criteria" tool should be used to specify this filter and the "Identify and Annotate" workflow should be extended by the "Identify Candidate Tool" (configured with the Filter Criterion). See the reference manual for more information on how preinstalled workflows can be edited.

Please note that in case none of the variants are present in ClinVar or dbSNP Common, the corresponding annotation column headers are missing from the result.

In case you like to change the databases as well as the used database version, please use the Reference Data Manager.

20.3 Hereditary Disease (TAS)

20.3.1 Filter Causal Variants (TAS-HD)

If you are analyzing a list of variants, you can use the **Filter Causal Variants (TAS-HD)** template workflow to remove variants that are outside the target region, as well as common variants present in publicly available databases. The workflow will annotate the remaining variants with gene names, conservation scores, and information from relevant databases.

20.3.2 Run the Filter Causal Variants (TAS-HD) workflow

To run the **Filter Causal Variants (TAS-HD)** workflow, go to:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **Hereditary Disease** ( ) | **Filter Causal Variants (TAS-HD)** ( )

1. Double-click on the workflow name to start the analysis. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Select the **variant track** you want to use for filtering causal variants (figure 20.43).

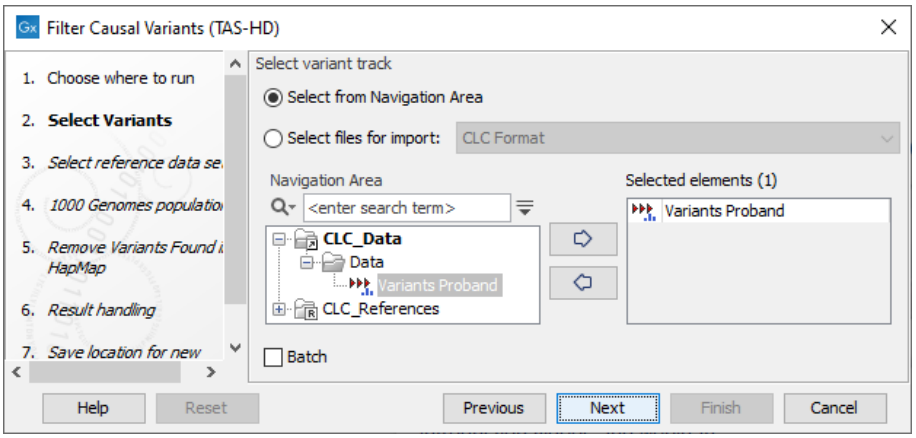


Figure 20.43: *Select the variant track from which you would like to filter somatic variants.*

3. In the next dialog, you have to select which data set should be used to filter causal variants (figure 20.44).
4. The 1000 Genomes population(s) is bundled in the downloaded reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track (figure 20.45).
5. Specify the **Hapmap populations** that should be used for **filtering out** variants found in Hapmap (figure 20.46).

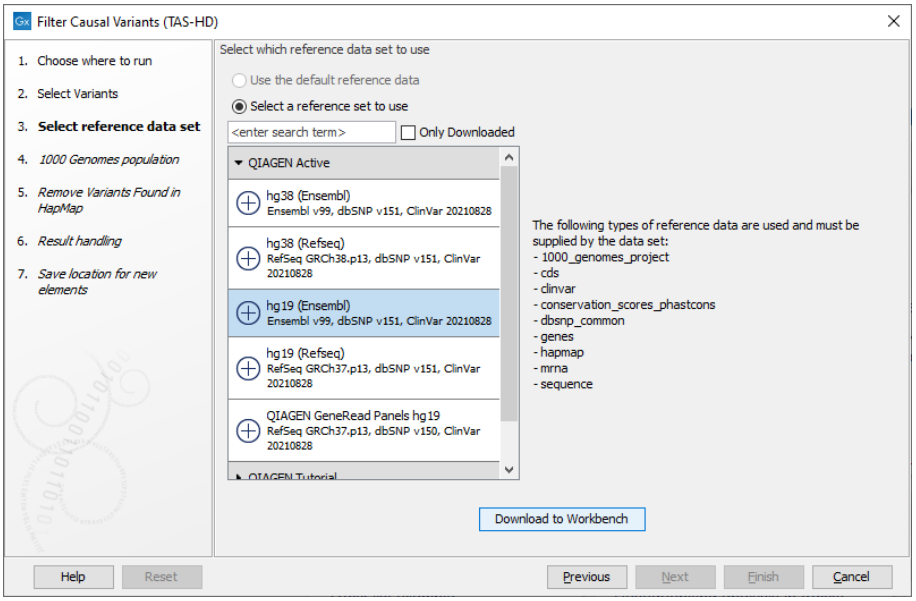


Figure 20.44: *Choose the relevant reference Data Set to annotate.*

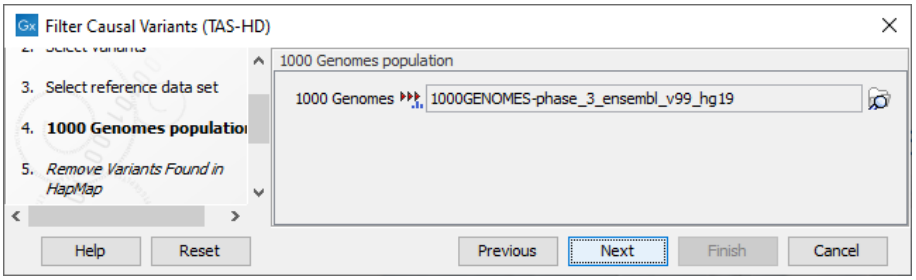


Figure 20.45: *Use the preselected 1000 Genomes population(s) or select another variant track.*

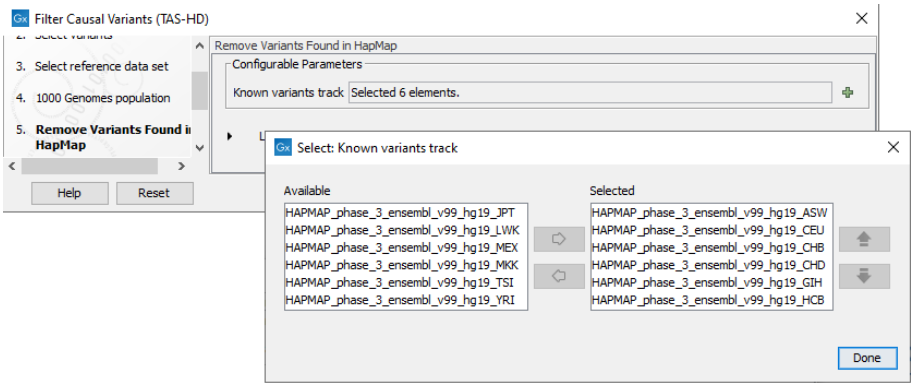


Figure 20.46: *Select the relevant Hapmap population(s).*

6. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

7. Choose to **Save** your results and click on the button labeled **Finish**.

20.3.3 Output from the Filter Causal Variants (TAS-HD) workflow

The following outputs are generated:

- **Amino Acids Changes** Track that shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
- **Somatic Candidate Variants** Track that holds the variant data. This track is also included in the Track List. If you hold down the Ctrl key (Cmd on Mac) while clicking on the table icon in the lower left side of the **View Area**, you can open the table view in split view. The table and the variant track are linked together, and when you click on a row in the table, the track view will automatically bring this position into focus.
- **Track List Filter Somatic Variants** A collection of tracks presented together. Shows the somatic candidate variants together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar, 1000 Genomes, and the PhastCons conservation scores (see figure 20.18).

20.3.4 Identify Variants (TAS-HD)

You can use the **Identify Variants (TAS-HD)** template workflow to call variants in the mapped and locally realigned reads. The workflow removes false positives and, in case of a targeted experiment, removes variants outside the targeted region. Variant calling is performed with the Fixed Ploidy Variant Detection tool and the Structural Variant Caller.

The **Identify Variants (TA-HD)** template workflow accepts sequencing reads as input.

Run the Identify Variants (TAS-HD) workflow

1. To run the **Identify Variants (TAS-HD)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** () | **Targeted Amplicon Sequencing** () | **Hereditary Disease** () | **Identify Variants (TAS-HD)** ()

2. Select the sequencing reads you want to analyze (figure 20.47).

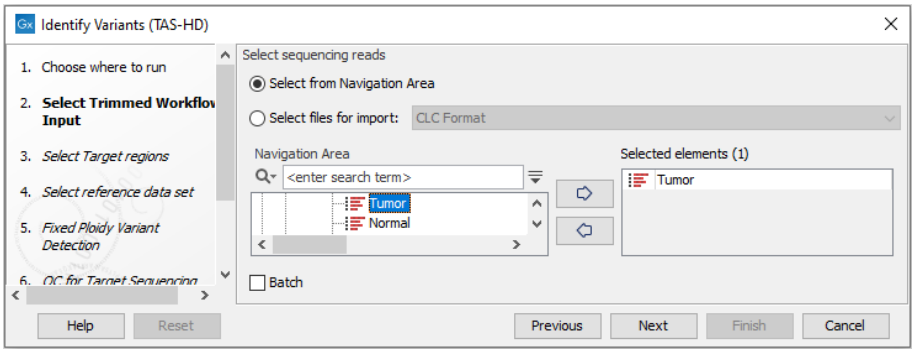


Figure 20.47: *Specify the sequencing reads for the sample.*

3. Specify a target region file for the application (figure 20.48).

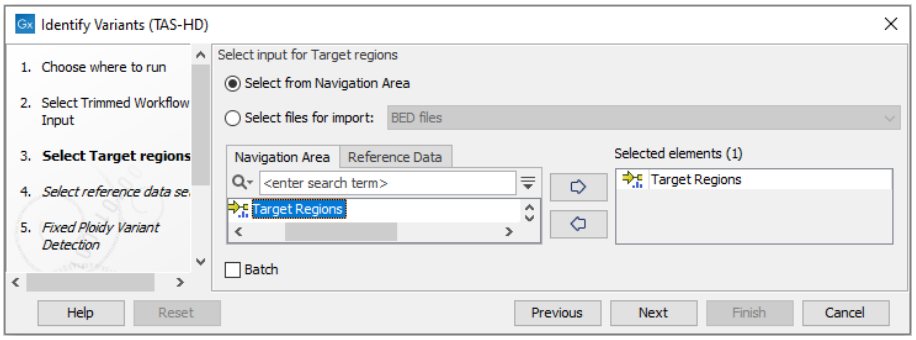


Figure 20.48: *Specify the target region file.*

4. In the next dialog, you have to select which reference data set should be used for the analysis (figure 20.49).

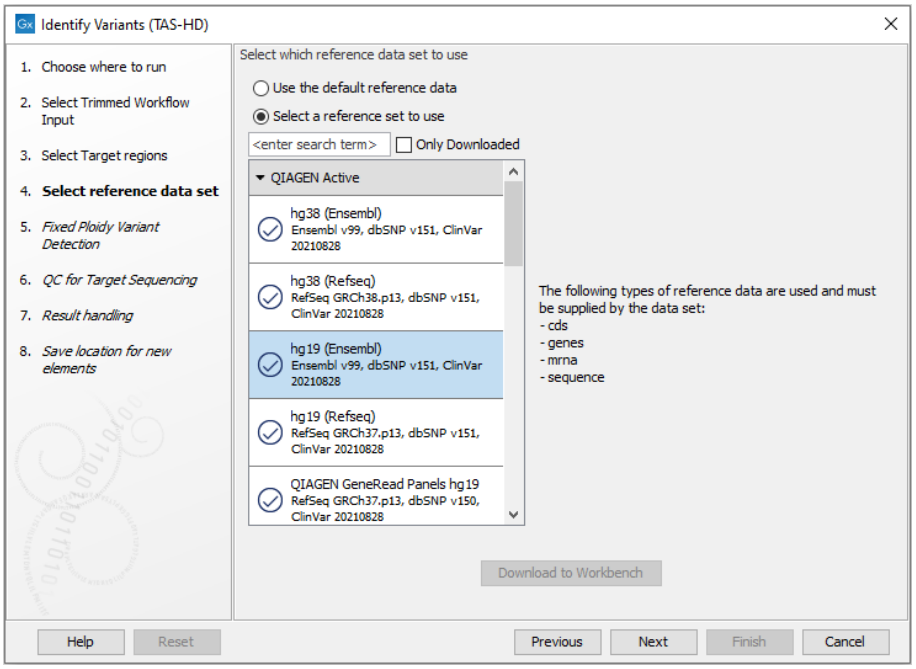


Figure 20.49: *Choose the relevant reference Data Set for the analysis.*

5. Specify the parameters for the Fixed Ploidy Variant Detection tool (figure 20.50).

   The parameters used by the Fixed Ploidy Variant Detection tool can be adjusted. We have optimized the parameters to the individual analyses, but you may want to tweak some of the parameters to fit your particular sequencing data. A good starting point could be to run an analysis with the default settings.

   The parameters that can be set are:

   - **Required variant probability** is the minimum probability value of the 'variant site' required for the variant to be called. Note that it is not the minimum value of the probability of the individual variant. For the Fixed Ploidy Variant detector, if a variant site - and not the variant itself - passes the variant probability threshold, then the variant with the highest probability at that site will be reported even if the probability

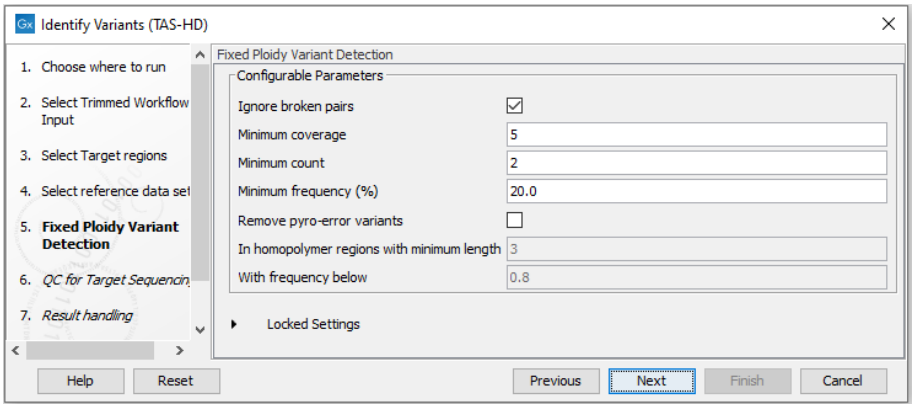


Figure 20.50: *Specify the parameters for the Fixed Ploidy Variant Detection tool.*

of that particular variant might be less than the threshold. For example if the required variant probability is set to 0.9 then the individual probability of the variant called might be less than 0.9 as long as the probability of the entire variant site is greater than 0.9.

- **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected. Please note that ignored broken pair reads will not be considered for any non-specific match filters.
- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

6. Specify the parameters for the **QC for Targeted Sequencing** tool (figure 20.51).

   When working with targeted data (WES or TAS data), quality checks for the targeted sequencing is included in the workflows. Again, you can choose to use the default settings, or you can choose to adjust the parameters.

   The parameters that can be set are:

   - **Minimum coverage** provides the length of each target region that has at least this coverage.
   - **Ignore non-specific matches**: reads that are non-specifically mapped will be ignored.
   - **Ignore broken pairs**: reads that belong to broken pairs will be ignored.

7. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

8. Choose to **Save** your results and click on the button labeled **Finish**.

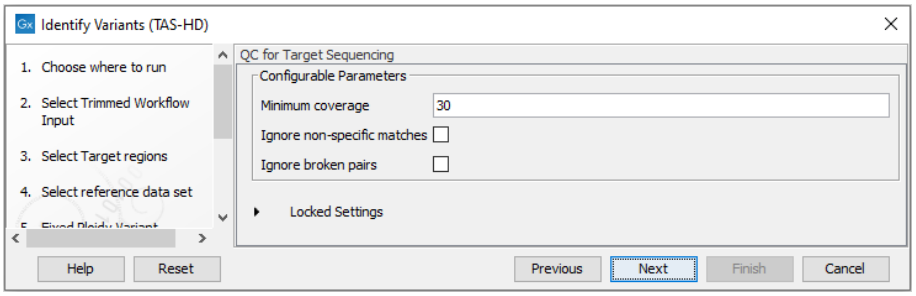


Figure 20.51: *Specify the parameters for the QC for Targeted Sequencing tool.*

Output from the Identify Variants (TAS-HD) workflow

The outputs generated are:

- **Read Mapping** ( ) The mapped sequencing reads. The reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).
- **Target Regions Coverage** ( ) The target regions coverage track shows the coverage of the targeted regions. Detailed information about coverage and read count can be found in the table format, which can be opened by pressing the table icon found in the lower left corner of the **View Area**.
- **Target Regions Coverage Report** ( ) The report consists of a number of tables and graphs that in different ways provide information about the targeted regions.
- **Identified Variants and Indels Indirect Evidence** ( ) A variant track containing the variants identified by the Fixed Ploidy Variant Detection tool, and a variant track with the indels inferred from indirect evidence by the Structural Variant Caller. The variants can be shown in track format or in table format. When holding the mouse over the detected variants in the Track List, a tooltip appears with information about the individual variants. You will have to zoom in on the variants to be able to see the detailed tooltip.
- **Genome Browser View Identify Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, the mapped reads, the identified variants, and the structural variants (see figure 20.32).

20.3.5 Identify and Annotate Variants (TAS-HD)

The **Identify and Annotate Variants (TAS-HD)** template workflow can be used to identify and annotate variants in one sample. This workflow is a combination of the **Identify Variants (TAS-HD)** and the **Annotate Variants (TAS-HD)** template workflows.

Sequencing reads provided as input are initially mapped to the human reference sequence. Then a local realignment is carried out to improve the subsequent variant detection analysis. Variants detected are annotated with gene names, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and information from common variants

present in the common dbSNP Common, HapMap, and 1000 Genomes database. Furthermore, a targeted region report is created to inspect the overall coverage and mapping specificity.

The difference between Identify and Annotate Variants (TAS-HD) and (WES-HD) is that the **Auto-detect paired distances** option has been switched off in Map Reads to Reference tool in the TAS workflows.

Run the Identify and Annotate Variants (TAS-HD) workflow

1. To run the **Identify and Annotate Variants (TAS-HD)** template workflow, go to:

   Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Targeted Amplicon Sequencing** ( ) | **Hereditary Disease** ( ) | **Identify and Annotate Variants (TAS-HD)** ( )

2. Select the sequencing reads you want to analyze (figure 20.52).

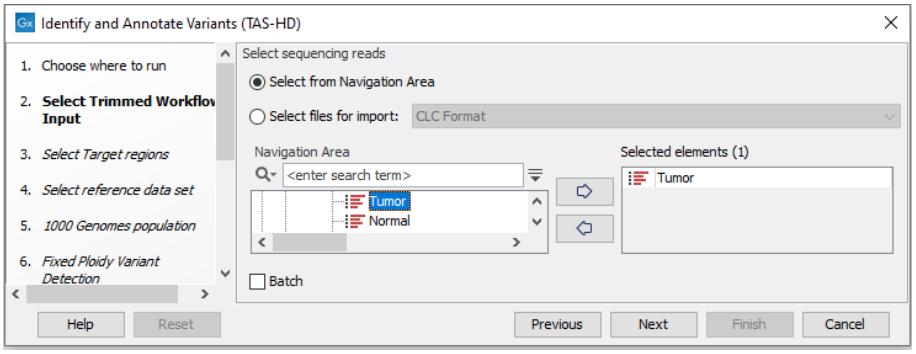


Figure 20.52: *Specify the sequencing reads for the sample.*

3. Specify a target region file (figure 20.53).

   The targeted region file is a file that specifies which regions have been sequenced, when working with whole exome sequencing or targeted amplicon sequencing data. This file is something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents.

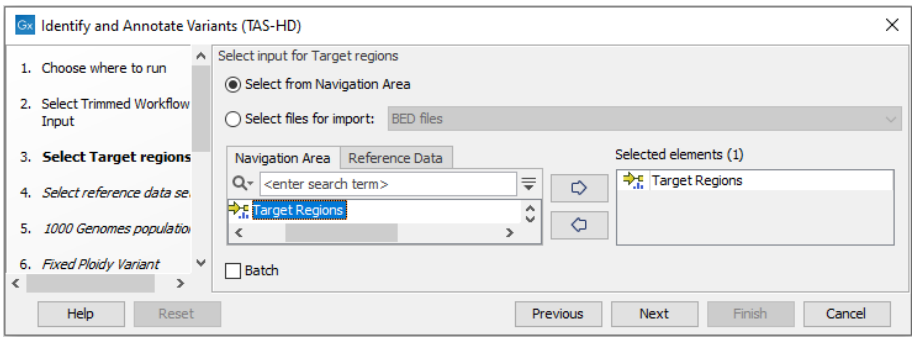


Figure 20.53: *Specify the target regions track.*

4. In the next dialog, you have to select which reference data set should be used for the analysis (figure 20.54).

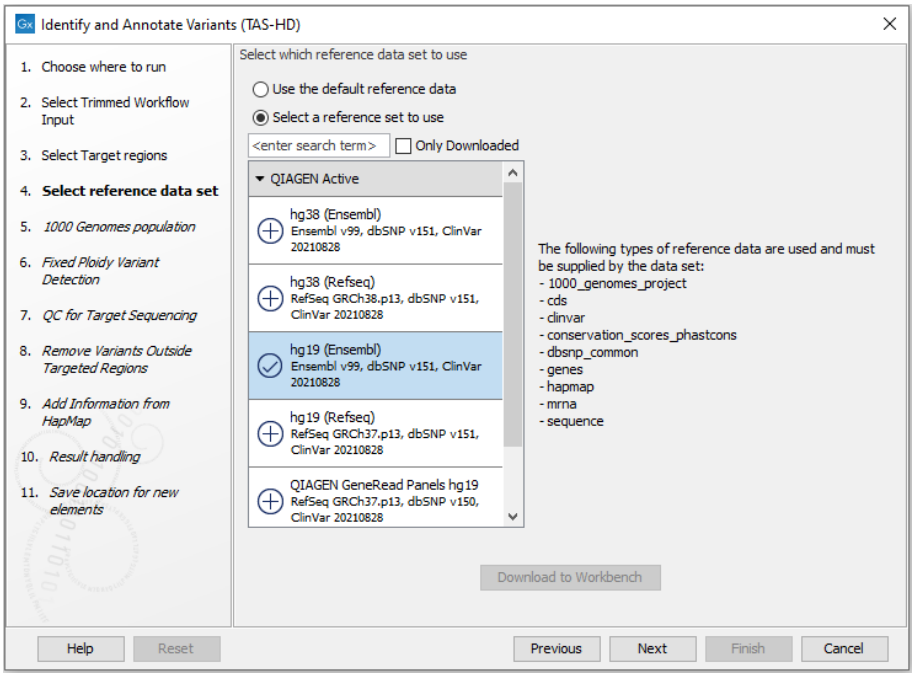


Figure 20.54: *Choose the relevant reference Data Set to identify variants.*

5. Specify which 1000 Genomes population you would like to use (figure 20.55).

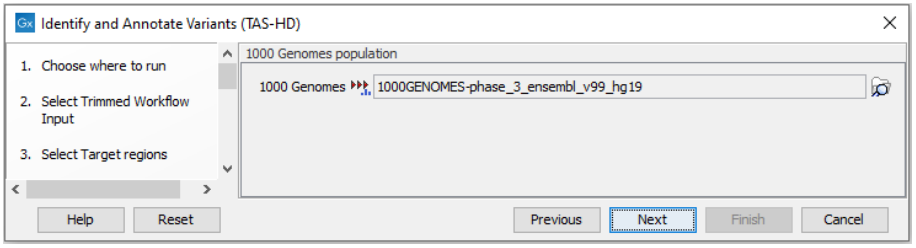


Figure 20.55: *Select the relevant 1000 Genomes population(s).*

6. Specify the Fixed Ploidy Variant Detection settings (figure 20.56).

   The parameters used by the Fixed Ploidy Variant Detection tool can be adjusted. We have optimized the parameters to the individual analyses, but you may want to tweak some of the parameters to fit your particular sequencing data. A good starting point could be to run an analysis with the default settings.

   The parameters that can be set are:

   - **Required variant probability** is the minimum probability value of the 'variant site' required for the variant to be called. Note that it is not the minimum value of the probability of the individual variant. For the Fixed Ploidy Variant detector, if a variant site - and not the variant itself - passes the variant probability threshold, then the variant with the highest probability at that site will be reported even if the probability of that particular variant might be less than the threshold. For example if the required variant probability is set to 0.9 then the individual probability of the variant called might be less than 0.9 as long as the probability of the entire variant site is greater than 0.9.

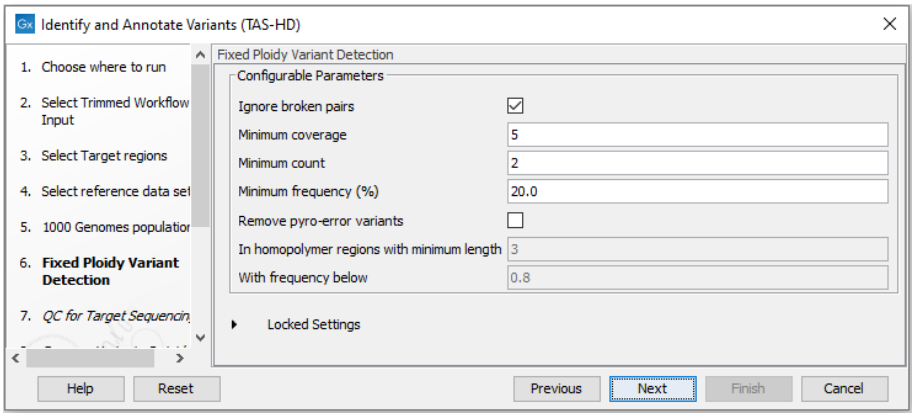


Figure 20.56: *Specify the parameters for the Fixed Ploidy Variant Detection tool.*

- **Ignore broken pairs:** When ticked, reads from broken pairs are ignored. Broken pairs may arise for a number of reasons, one being erroneous mapping of the reads. In general, variants based on broken pair reads are likely to be less reliable, so ignoring them may reduce the number of spurious variants called. However, broken pairs may also arise for biological reasons (e.g. due to structural variants) and if they are ignored some true variants may go undetected. Please note that ignored broken pair reads will not be considered for any non-specific match filters.
- **Minimum coverage:** Only variants in regions covered by at least this many reads are called.
- **Minimum count:** Only variants that are present in at least this many reads are called.
- **Minimum frequency:** Only variants that are present at least at the specified frequency (calculated as 'count'/'coverage') are called.

7. Specify the parameters for the QC for Targeted Sequencing tool (figure 20.57).

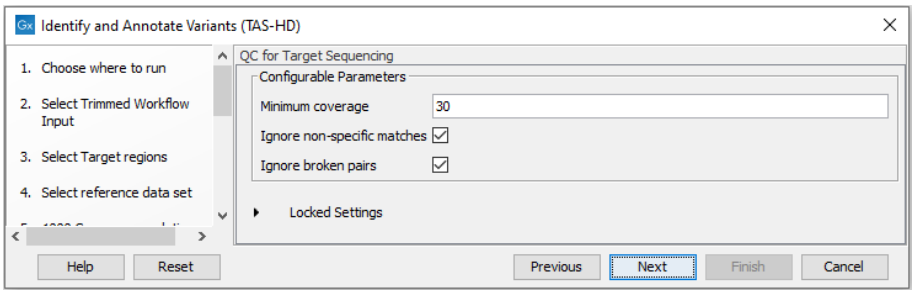


Figure 20.57: *Specify the parameters for the QC for Targeted Sequencing tool.*

The parameters that can be set are:

- **Minimum coverage** provides the length of each target region that has at least this coverage.
- **Ignore non-specific matches**: reads that are non-specifically mapped will be ignored.
- **Ignore broken pairs**: reads that belong to broken pairs will be ignored.

8. Specify the Hapmap population that should be used to add information on variants found in the Hapmap project.

9. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.
10. Choose to **Save** your results and click on the button labeled **Finish**.

Output from the Identify and Annotate Variants (TAS-HD) workflow

The following outputs are generated:

- A **Reads Track**
- A **Coverage Report Read Mapping**
- A **Per-region Statistics Track**
- A **Filtered Variant Track** Annotated variants
- A **Indels indirect evidence Variant Track** with indels inferred by the Structural Variant Caller
- An **Amino Acid Track** Shows the consequences of the variants at the amino acid level in the context of the original amino acid sequence. A variant introducing a stop mutation is illustrated with a red amino acid.
- A **Track List**

Chapter 21

Whole transcriptome sequencing (WTS)

Contents

The technologies originally developed for next-generation DNA sequencing can also be applied to deep sequencing of the transcriptome. This is done through cDNA sequencing and is called RNA sequencing or simply RNA-Seq.

One of the key advantages of RNA-Seq is that the method is independent of prior knowledge of the corresponding genomic sequences and therefore can be used to identify transcripts from unannotated genes, novel splicing isoforms, and gene-fusion transcripts [Wang et al., 2009, Martin and Wang, 2011]. Another strength is that it opens up for studies of transcriptomic complexities such as deciphering allele-specific transcription by the use of SNPs present in the transcribed regions [Heap et al., 2010].

RNA-Seq-based transcriptomic studies have the potential to increase the overall understanding of the transcriptome. However, the key to get access to the hidden information and be able to make a meaningful interpretation of the sequencing data highly relies on the downstream bioinformatic analysis.

The following template workflows are available for use with RNA-Seq data (figure 21.1):

- Differential Expression and Pathway Analysis
- Differential Expression and Pathway Analysis from Count Matrix
- For data from **Human** (), **Mouse and Rat** ():
  - Annotate Variants (WTS)
  - Compare Variants in DNA and RNA
  - Identify Candidate Variants and Genes from Tumor Normal Pair

– Identify Variants and Add Expression Values

For the workflows where variants are annotated based on information from databases available for more than one population, you will have the opportunity to select the population that is best suited for your analysis.

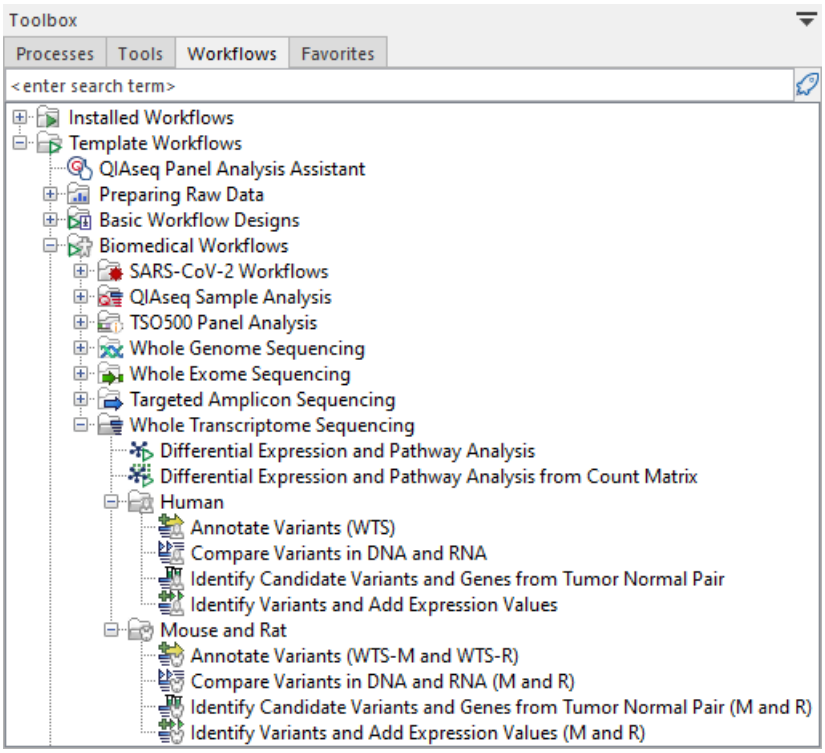


Figure 21.1: *The RNA-Seq template workflows are available under the Whole Transcriptome Sequencing folder.*

21.1 Differential Expression and Pathway Analysis

The **Differential Expression and Pathway Analysis** ( ) and the **Differential Expression and Pathway Analysis from Count Matrix** ( ) template workflows can be used for performing differential expression analysis for gene or transcript expression data from groups of samples.

The two workflows include all the necessary steps for the analysis:

- Differentially expressed features (genes or transcripts) are identified using **Differential Expression for RNA-Seq**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Differential_Expression_RNA_Seq.html.
- The differentially expressed features are optionally uploaded to QIAGEN Ingenuity Pathway Analysis (IPA) using **Upload to IPA**, see section 9.1.
- Terms from a Gene Ontology Annotation (GOA) table affected by the differentially expressed features are identified using **Gene Set Test**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Gene_Set_Test.html.

- The expression values are summarized in various plots using
  - **PCA for RNA-Seq**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PCA_RNA_Seq.html;
  - **Create Sample Level Heat Map for RNA-Seq**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Level_Heat_Map_RNA_Seq.html;
  - **Create Feature Level Heat Map for RNA-Seq**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Feature_Level_Heat_Map_RNA_Seq.html.

The two workflows differ in the accepted inputs:

- **Differential Expression and Pathway Analysis** takes as input Gene Expression (GE) or Transcript Expression (TE) tracks. These must be associated with a CLC Metadata Table, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Metadata.html for details.
- **Differential Expression and Pathway Analysis from Count Matrix** imports Gene Expression (GE) or Transcript Expression (TE) tracks, along with a CLC Metadata Table, from two separate files containing the expression data and metadata, respectively. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Import_Expression_Data.html for details.

Launching the workflows

To run these workflows, go to:

> **Template Workflows** | **Biomedical Workflows** () | **Whole Transcriptome Sequencing** ()

and select:

> **Differential Expression and Pathway Analysis** ()
>
> **Differential Expression and Pathway Analysis from Count Matrix** ()

For general information about launching workflows, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Launching_workflows_individually_in_batches.html

The input options can be configured in the following dialogs:

- For the **Differential Expression and Pathway Analysis from Count Matrix** workflow:
  - **Select Expression Tracks (TE or GE)**. Select Gene Expression (GE) or Transcript Expression (TE) tracks.
  - **Configure batching**. If running the workflow in `Batch` mode, you will be asked to define the batch units.
- For the **Differential Expression and Pathway Analysis from Count Matrix** workflow:

  - **Import Expression Data**. Select the two files containing the expression data and metadata, respectively.

Options in the following dialogs can additionally be configured for both workflows:

- **Choose where to run**. If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible.
- **Specify workflow path**. Select whether you want to upload the differential expression data to IPA. Uploading requires an IPA account.
- **Specify reference data handling**. Select the relevant Reference Data Set, see chapter 3 for details.
- **Differential Expression for RNA-Seq**. Configure the following options as needed:
  - Metadata table
  - Test differential expression due to
  - While controlling for
  - Comparisons
  - Control group

  See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Differential_Expression_RNA_Seq.html for details.
- If uploading to IPA:
  - **Filter on Custom Criteria**. Configure filtering criteria defining which features to upload to IPA. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Filter_on_Custom_Criteria.html for details.
  - **Upload to IPA**. Login to IPA and configure the options. See section 9.1 for details.
- **Result handling**. Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements**. Choose where to save the data, and press **Finish** to start the analysis.

21.1.1 Output from the Differential Expression and Pathway Analysis workflows

The following outputs are generated:

- **PCA for RNA-Seq** (): A Principal Component Analysis (PCA) plot. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PCA_RNA_Seq.html for details.
- **Sample Level Heat Map for RNA-Seq** (): A heat map of sample distances. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Level_Heat_Map_RNA_Seq.html;

- **Feature Level Heat Map for RNA-Seq** (): A heat map of expression values. Each column corresponds to one sample, and each row corresponds to a feature (a gene or a transcript). See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Feature_Level_Heat_Map_RNA_Seq.html.
- **Genome Browser View** (): A collection of output tracks, allowing multiple data types at the same genomic position to be viewed simultaneously. Note that not all output tracks are necessarily in the browser. Tracks can be added and removed from the browser.
- **Expression Tracks** (): The imported expression tracks, if running the workflow from count matrix.
- **Statistical Comparison** (): One or more comparisons containing the differential expression results. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Output_Differential_Expression_tools.html for details.
- **Expression Browser** (): A browser for inspecting feature expressions, annotations and statistics for many samples. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Expression_Browser.html.
- **Venn Diagram** (): A plot comparing the overlap of differentially expressed features in two or more statistical comparisons. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Venn_Diagram_RNA_Seq.html for details.
- **GO Enrichment Analysis** (): One or more tables containing the GO enrichment analysis results. See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Gene_Set_Test.html for details.

21.2 Annotate Variants (WTS)

Using a variant track, annotation track, expression track or statistical comparison track, the **Annotate Variants (WTS)** template workflows add the following annotations:

- **Gene names** Names of genes whenever a variant is found within a known gene.
- **mRNA** Names of mRNA whenever a variant is found within a known transcript.
- **CDS** Names of CDS whenever a variant is found within a coding sequence.
- **Amino acid changes** Information about amino acid changes caused by the variants.
- **Information from ClinVar** Information about the relationships between human variations and their clinical significance.
- **Information from dbSNP Common** Information from the "Single Nucleotide Polymorphism Database", which is a general catalog of genome variation, including SNPs, multinucleotide polymorphisms (MNPs), insertions and deletions (indels), and short tandem repeats (STRs).
- **PhastCons Conservation scores** The conservation scores, in this case generated from a multiple alignment with a number of vertebrates, describe the level of nucleotide conservation in the region around each variant.

The template workflows can be found under under the Workflows menu at:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Human** ( ) | **Annotate Variants (WTS)** ( )

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Mouse and Rat** ( ) | **Annotate Variants (WTS-M and WTS-R)** ( )

After starting the workflow:

1. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. In the first wizard step, select the variant track, annotation track, expression track or statistical comparison track to annotate (figure 21.2).

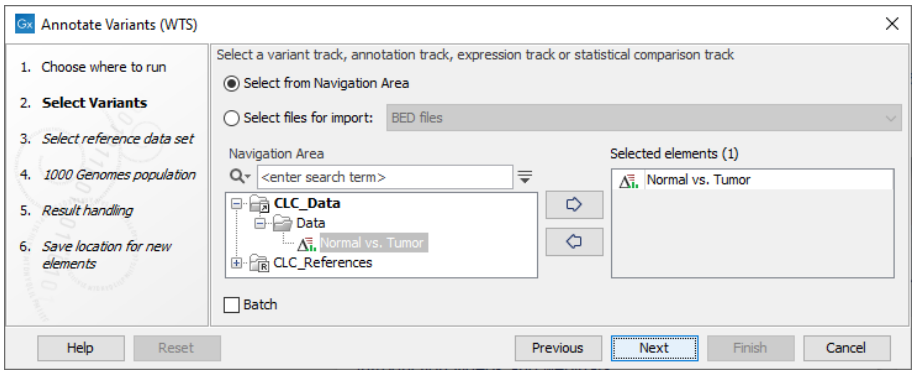


Figure 21.2: *Select the variant, annotation, expression or statistical comparison track to annotate.*

3. In the next dialog, you have to select which data set should be used to annotate variants (figure 21.3).
4. If you are using the workflow for Human, you should specify which 1000 Genomes population to use (figure 21.4).
5. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.
6. Choose to **Save** your results and click on the button labeled **Finish**.

The following outputs are generated:

1. **Annotated Variants** ( ) Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.
2. **Track List Annotated Variants** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in dbSNP Common, ClinVar, 1000 Genomes, and PhastCons conservation scores (see figure 21.5).

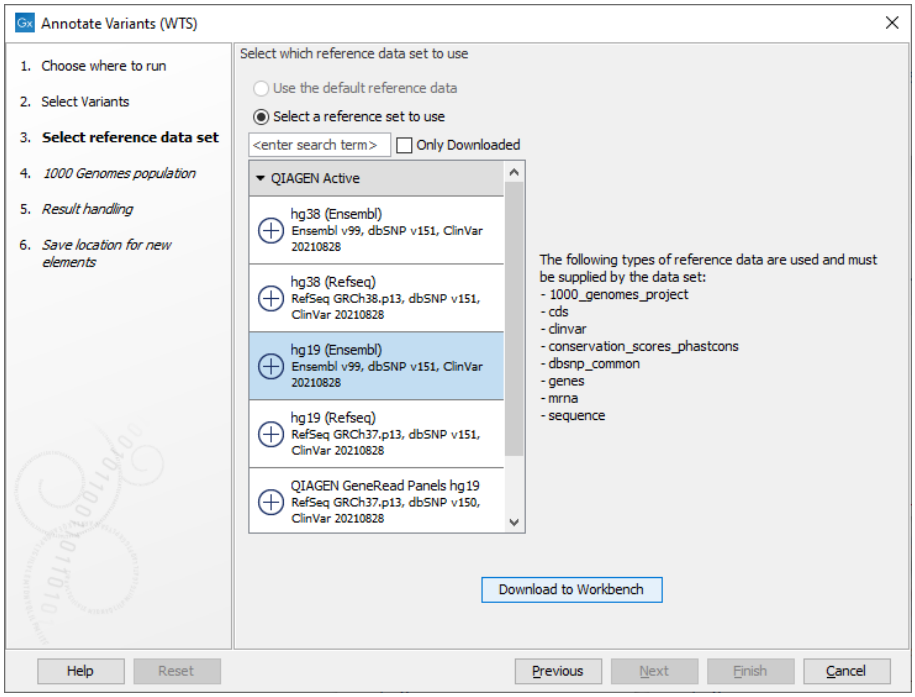


Figure 21.3: *Choose the relevant reference Data Set to annotate.*

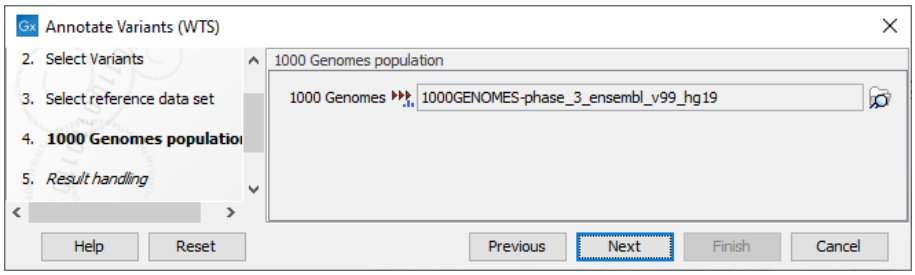


Figure 21.4: *Select the relevant 1000 Genomes population(s).*

Note! Please be aware that if you delete the annotated variant track, this track will also disappear from the track list. On the other hand, it is possible to add tracks to the Track List by dragging the track directly from the **Navigation Area** to the Track List view.

Open the annotated variant track as a table; it includes all variants and the added information/annotations (see figure 21.6). The table and the Track List are linked; if you click on an entry in the table, this particular position in the genome will automatically be brought into focus in the Track List view.

You may be met with a warning as shown in figure 21.7. This is simply a warning telling you that it may take some time to create the table if you are working with tracks containing large amounts of annotations. Please note that in case none of the variants are present in ClinVar or dbSNP Common, the corresponding annotation column headers are missing from the result.

Adding information from other sources may help you identify interesting candidate variants for further research. For example, known common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be identified. Further, variants not found in the ClinVar database

Figure 21.5: *The output from the Annotate Variants (WTS) template workflow is a track list containing individual tracks for all added annotations.*

can be prioritized based on amino acid changes in case the variant causes changes on the amino acid level.

A high conservation level between different vertebrates or mammals, in the region containing the variant, can also be used to give a hint about whether a given variant is found in a region with an important functional role. If you would like to use the conservation scores to identify interesting variants, we recommend that variants with a conservation score of more than 0.9 (PhastCons score) is prioritized over variants with lower conservation scores.

It is possible to filter variants based on their annotations. This type of filtering can be facilitated using the table filter found at the top part of the table. If you are performing multiple experiments where you would like to use the exact same filter criteria, you can create a filter that can be saved and reused. To do this, use the Filter on Custom Criteria tool.

Figure 21.6: *The variant track in the Track List is open as a table in split view.*

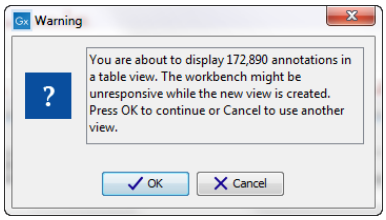


Figure 21.7: *Warning that appears when you work with tracks containing many annotations.*

21.3 Compare Variants in DNA and RNA

Integrated analysis of genomic and transcriptomic sequencing data is a powerful tool that can help increase our current understanding of genomic variants. The **Compare Variants in DNA and RNA** workflows identify variants in DNA and RNA and studies the relationship between the identified genomic and transcriptomic variants.

The workflows can be found under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Human** ( ) | **Compare Variants in DNA and RNA** ( )
>
> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Mouse and Rat** ( ) | **Compare Variants in DNA and RNA (M and R)** ( )

After starting the workflow:

1. If you are connected to a server, you will first be asked where you would like to run the analysis.

2. Select the **RNA reads** that you would like to analyze (figure 21.8).

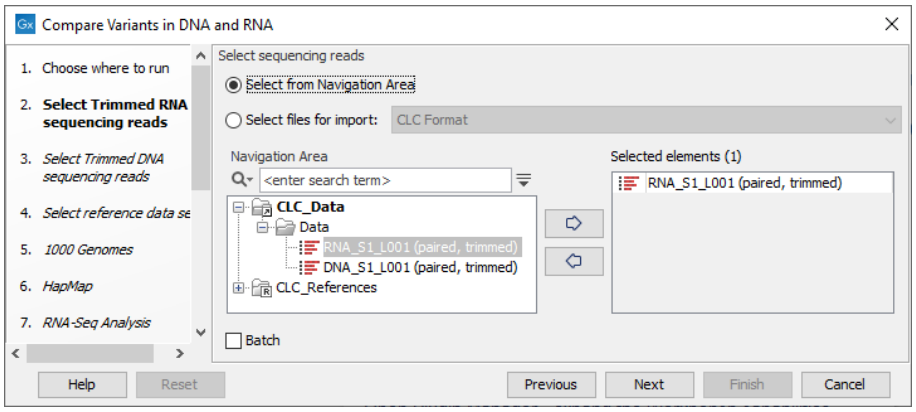


Figure 21.8: *Select the RNA reads to analyze.*

3. Select now the **DNA reads** to analyze (see figure 21.9).

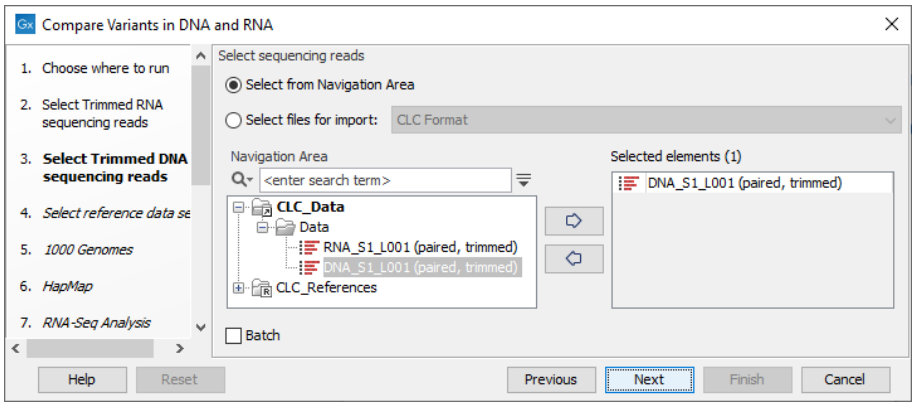


Figure 21.9: *Select the DNA reads to analyze.*

4. Select the Reference Data Set that is relevant to your study (figure 21.10).

5. The 1000 Genomes population(s) is bundled in the downloaded human reference dataset and therefore preselected in this step. If you want to use another variant track you can browse in the navigation area for the preferred track (see figure 21.11).

6. In this step you should specify the **Hapmap population(s)** that should be used for **filtering out** variants found in Hapmap (see figure 21.12).

7. Configure the parameters for the RNA-Seq Analysis (figure 21.13).

   If you wish to use spike-in controls, add the relevant file in the "Spike-in controls" field.

   You can also specify that the reads should be mapped only in their forward or reverse orientation (it is by default set to both). Choosing to restrict mapping to one direction is typically appropriate when a strand specific protocol for read generation has been used,

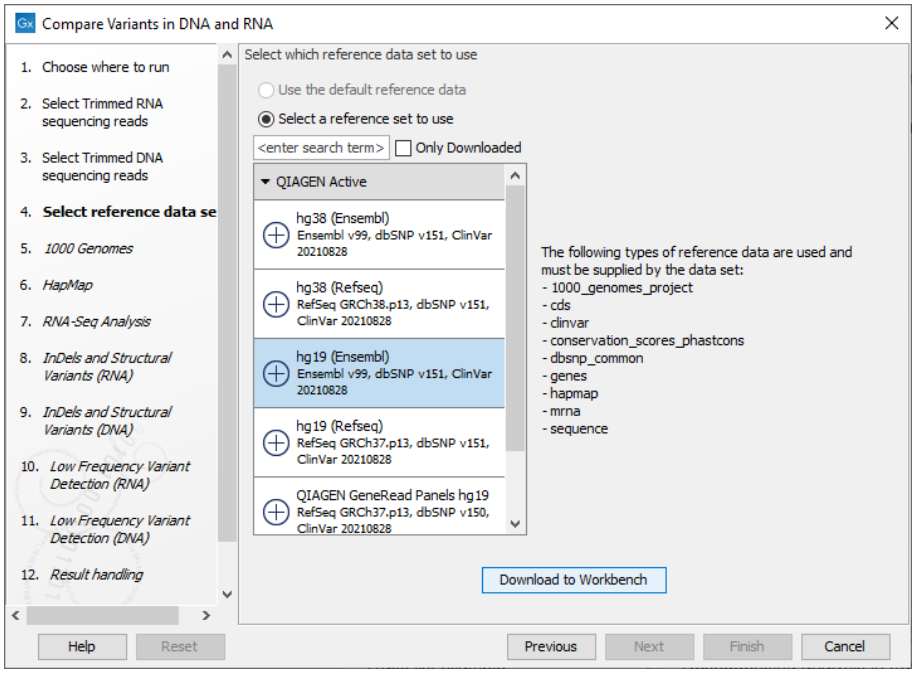


Figure 21.10: *Select the relevant data set for the samples being studied.*

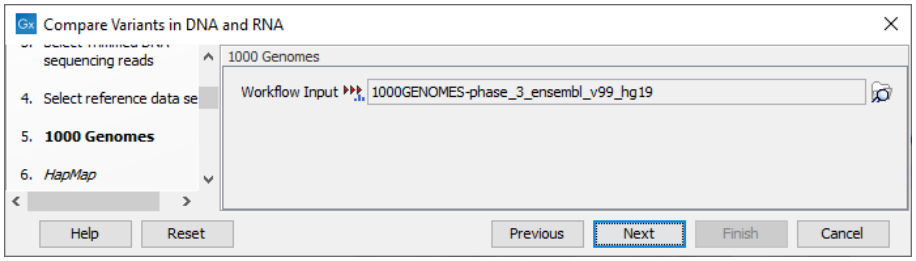


Figure 21.11: *Use the preselected 1000 Genomes population(s) or select another variant track (human workflow only).*

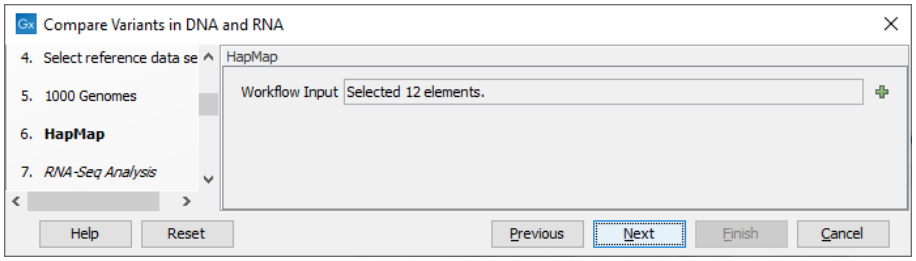


Figure 21.12: *Select the relevant Hapmap population(s).*

as it allows assignment of the reads to the right gene in cases where overlapping genes are located on different strands. Also, applying the 'strand specific' 'reverse' option in an RNA-Seq run could allow the user to assess the degree of antisense transcription. Note that mate pairs are not supported when choosing the forward only or reverse only option.

8. Specify a **target region** for the analysis of the **RNA** sample with the Indels and Structural Variants tool (figure 21.14). Repeat for the **DNA** sample at the next step.

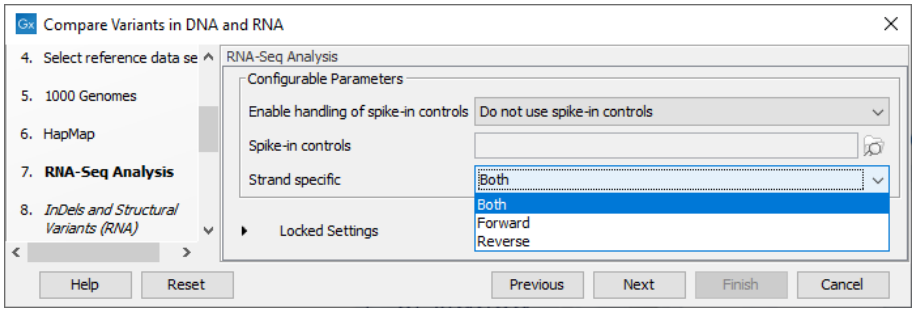


Figure 21.13: *Configure the RNA-Seq Analysis. Here we specified left the parameters to their default value.*

The targeted region file is a file that specifies which regions have been sequenced. This file is something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents. Remember that you have a hg38-specific BED file when using hg38 as reference, and hg19-specific BED file when using hg19 as reference.

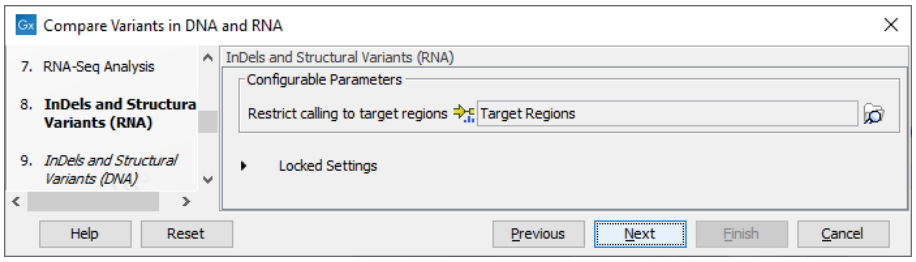


Figure 21.14: *Specify the target region for the Indels and Structural Variants tool.*

9. Set the parameters for the Low Frequency Variant Detection step for your **RNA** sample (see figure 21.15), and for the **DNA** sample at the next step. For a description of the different parameters that can be adjusted in the variant detection step, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html.

10. Click **Next** to go to the result handling step. **Preview All Parameters** allows you to view all parameters, but not edit them. Choose to save the results and click **Finish** to select a location to save the results to and start the analysis.

The following outputs are generated:

1. A **DNA Read Mapping** and a **RNA Read Mapping** ( ) The mapped DNA or RNA sequencing reads. The sequencing reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

2. A **DNA Mapping Report** and a **RNA Mapping Report** ( ) This report contains information about the reads, reference, transcripts, and statistics (seehttps://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html).

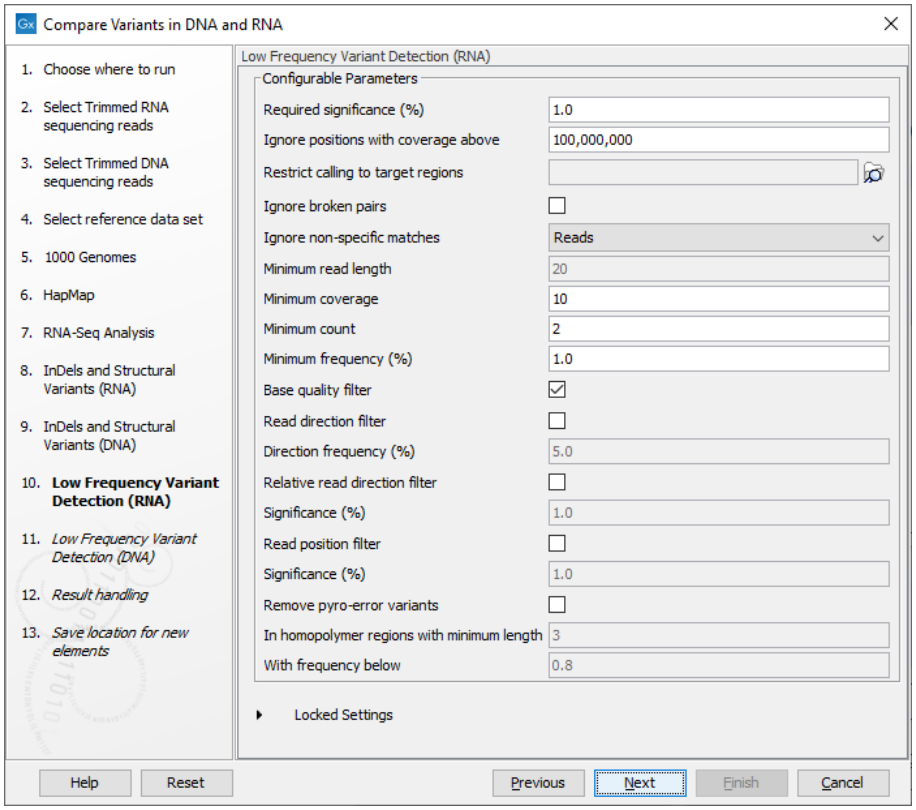


Figure 21.15: *Specify the parametes for transcriptomic variant detection.*

3. An **RNA Gene Expression** ( ) A track showing gene expression annotations. Hold the mouse over or right-click on the track: if you have zoomed in to nucleotide level, a tooltip will appear with information about gene name and expression values.

4. An **RNA Transcript Expression** ( ) A track showing transcript expression annotations.

5. A **Filtered Variant Track with All Variants Found in DNA or RNA** ( ) This track shows all variants that have been detected in either RNA, DNA or both.

6. A **Filtered Variant Track with Variants Found in Both DNA and RNA** ( ) This track shows only the variants that are present in both DNA and RNA. With the table icon ( ) found in the lower left part of the **View Area** it is possible to switch to table view. The table view provides details about the variants such as type, zygosity, and information from a range of different databases.

7. A **Track List Variants Found in DNA and RNA** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar and dbSNP Common (see figure 21.16).

The three most important tracks generated are the **Variants found in both DNA and RNA track**, **All variants found in DNA or RNA track**, and the **Track List**. The Track List view makes it easy to get an overview in the context of a reference sequence, and compare variant and expression tracks with information from different databases. The two other tracks (**Variants found in both**

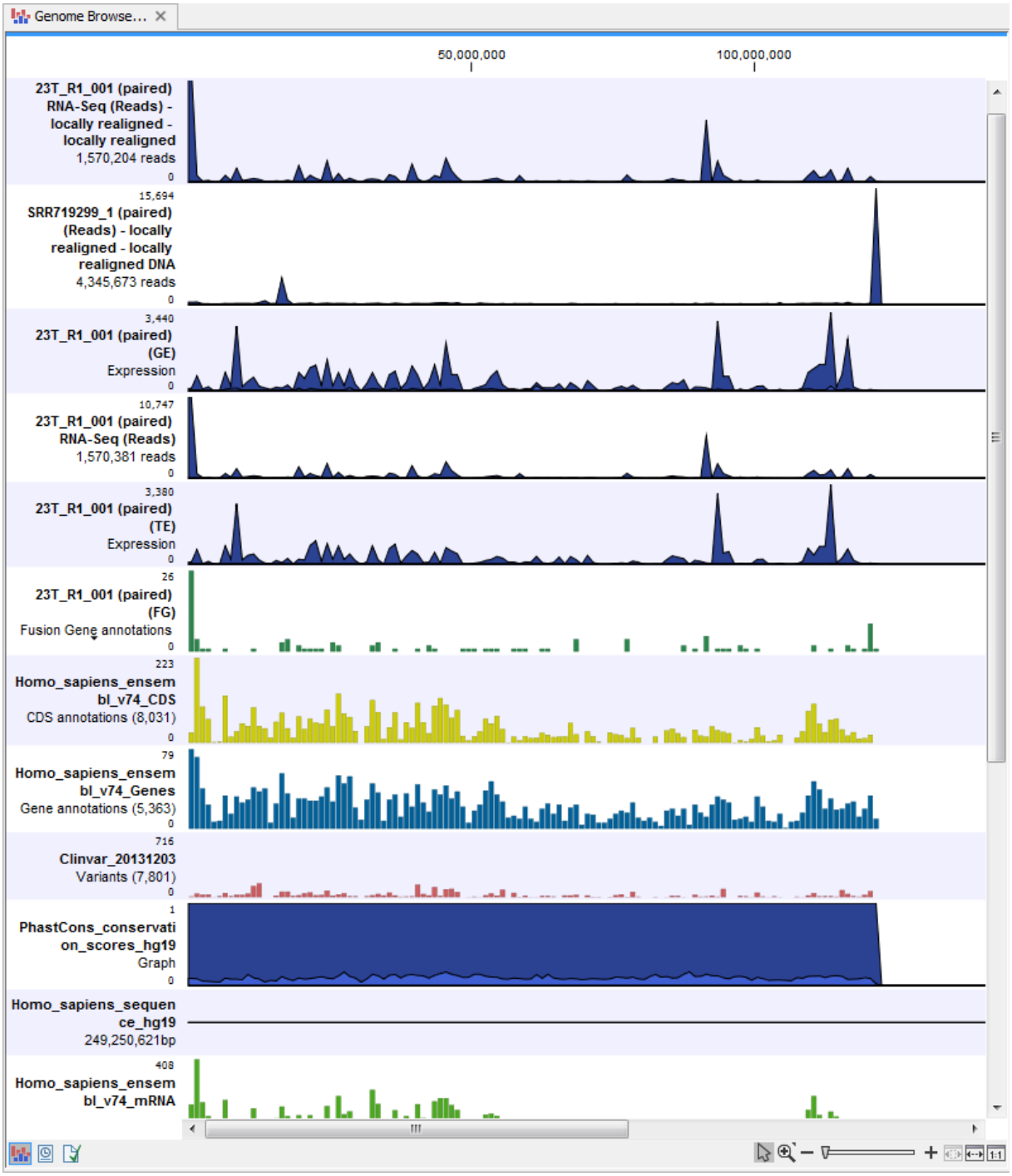


Figure 21.16: *The Track List view makes it easy to compare a range of different data.*

DNA and RNA track and **All variants found in DNA or RNA track**) provides detailed information about the detected variants when opened in table view.

21.4 Identify Candidate Variants and Genes from Tumor Normal Pair

The **Identify Candidate Variants and Genes from Tumor Normal Pair** workflows identify somatic variants and differentially expressed genes in a tumor normal pair. One tumor normal pair can be compared at the time. If you would like to compare more than one pair you must repeat the analysis with the next tumor normal pair.

The workflows can be found under the Workflows menu at:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Human** ( ) | **Identify Candidate Variants and Genes from Tumor Normal Pair** ( )

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Mouse and Rat** ( ) | **Identify Candidate Variants and Genes from Tumor Normal Pair (M and R)** ( )

After starting the workflow:

1. If you are connected to a server, you will first be asked where you would like to run the analysis.
2. Specify the RNA-Seq reads from the tumor sample (the panel in the left side of the wizard shows the kind of input that should be provided as in figure 21.17). Click **Next**.

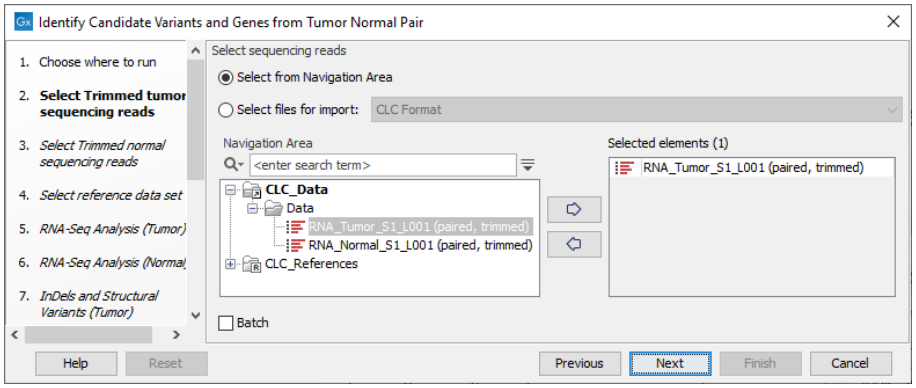


Figure 21.17: *Select the RNA-Seq reads from the tumor sample.*

3. In the next step you will be asked to select the RNA-Seq reads from the normal sample (see figure 21.18). Click **Next**.
4. Select the Reference Data Set that is relevant to your study (figure 21.19).
5. Configure the parameters for the RNA-Seq Analysis (figure 21.20), first for the tumor sample, and then for the normal sample in the following step.

   If you wish to use spike-in controls, add the relevant file in the "Spike-in controls" field.

   You can also specify that the reads should be mapped only in their forward or reverse orientation (it is by default set to both). Choosing to restrict mapping to one direction is typically appropriate when a strand specific protocol for read generation has been used, as it allows assignment of the reads to the right gene in cases where overlapping genes are located on different strands. Also, applying the 'strand specific' 'reverse' option in an

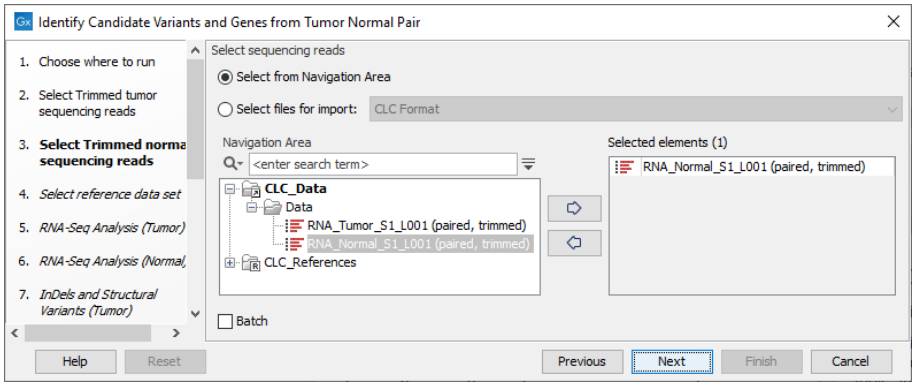


Figure 21.18: *Select the RNA-Seq reads from the normal sample.*

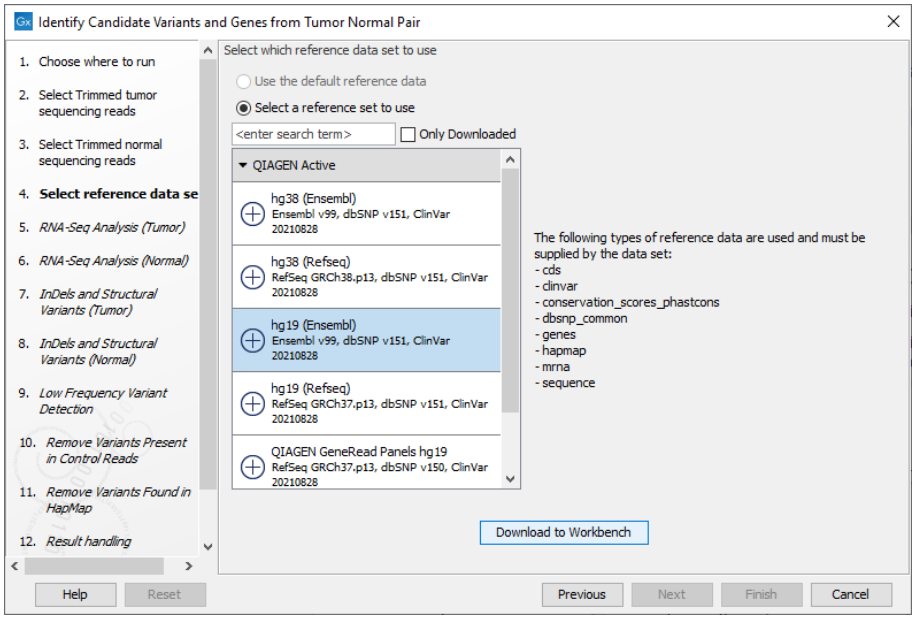


Figure 21.19: *Select the relevant data set for the samples being studied.*

RNA-Seq run could allow the user to assess the degree of antisense transcription. Note that mate pairs are not supported when choosing the forward only or reverse only option.

6. Specify in the next two dialog a **target region** for the analysis of the sample with the Indels and Structural Variants tool, first for the **tumor** sample, followed by the **normal** sample (figure 21.12).

   The targeted region file is a file that specifies which regions have been sequenced. This file is something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents. Remember that you have a hg38-specific BED file when using hg38 as reference, and hg19-specific BED file when using hg19 as reference.

7. Set the parameters for the Low Frequency Variant Detection step (see figure 21.22). For a description of the different parameters that can be adjusted in the variant detection step, see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html`.

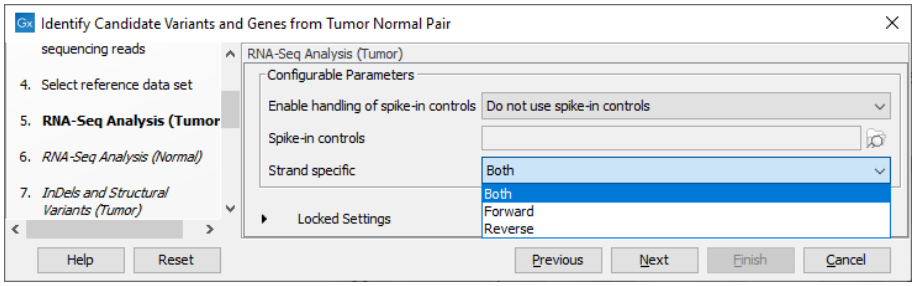


Figure 21.20: *Configure the RNA-Seq Analysis. Here we specified a file for spike-in control but left the strand specific parameter to its default value.*

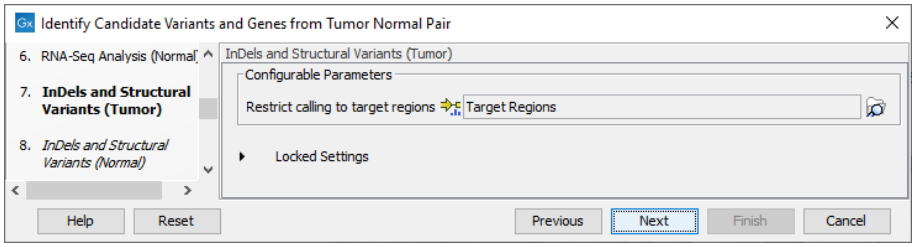


Figure 21.21: *Specify the target region for the Indels and Structural Variants tool.*

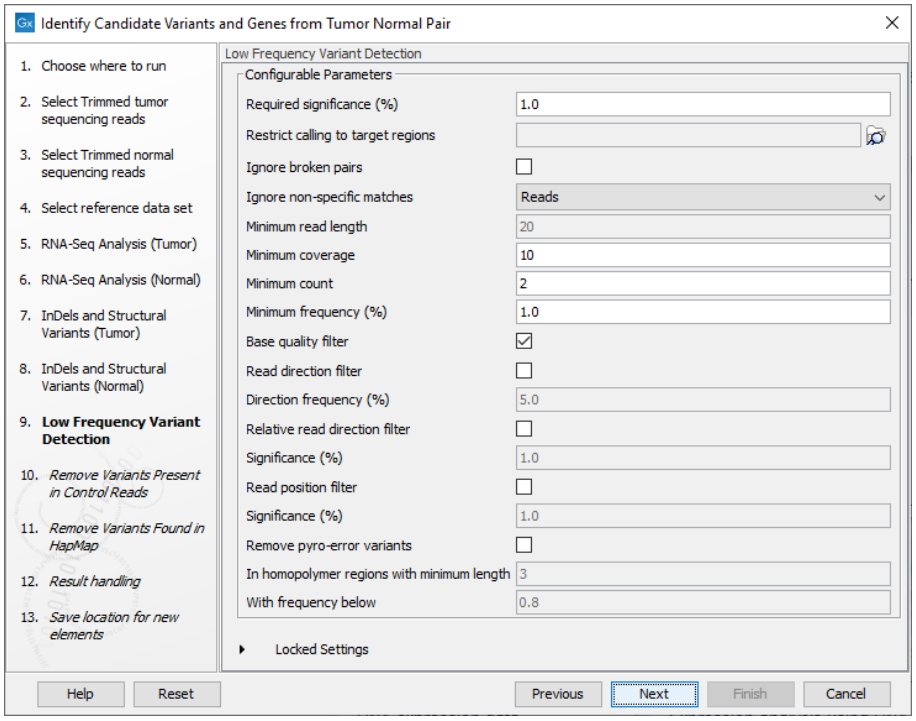


Figure 21.22: *Specify the parameters for variant calling.*

8. The next dialog called Remove Variants Present in Control Reads (figure 21.23) concerns removal of germline variants. You are asked to supply the number of reads in the control data set that should support the variant allele in order to include it as a match. All the variants where at least this number of control reads show the particular allele will be filtered away in the result track.

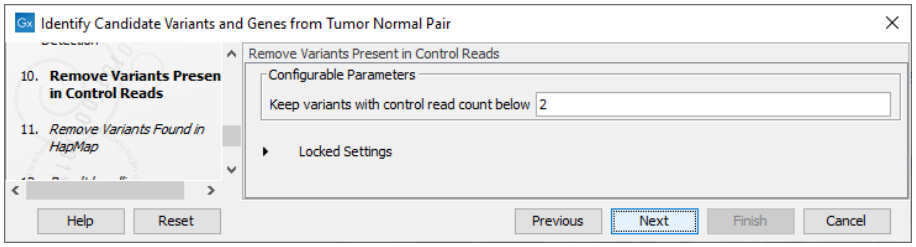


Figure 21.23: *Specify the number of reads to use as cutoff for removal of germline variants.*

9. Finally, for the Remove Variants Found in HapMap (figure 21.24), you can also specify which specific Hapmap population(s) characterize(s) best the samples.

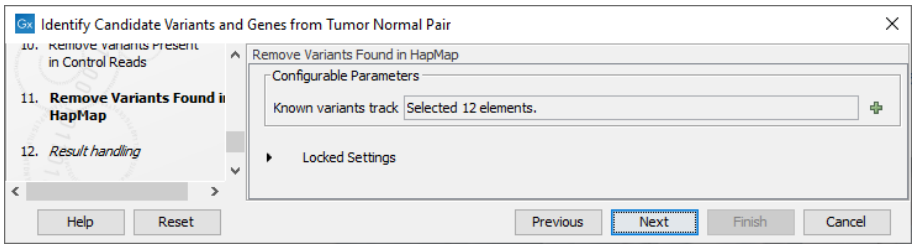


Figure 21.24: *Remove Hapmap variants.*

10. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

11. Choose to **Save** your results and click on the button labeled **Finish**.

The following outputs are generated:

1. **Gene Expression Normal** and **Gene Expression Tumor** ( ) A track showing gene expression annotations. Hold the mouse over or right-click on the track: a tooltip will appear with information about e.g. gene name and gene expression values.

2. **Transcript Expression Normal** and **Transcript Expression Tumor** ( ) A track showing transcript expression annotations.

3. **RNA-Seq Mapping Report Normal** and **RNA-Seq Mapping Report Tumor** ( ) This report contains information about the reads, reference, transcripts, and statistics. This is explained in more detail in the CLC Workbench reference manual in section **RNA-Seq report** (`https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html`).

4. **Read Mapping Normal** and **Read Mapping Tumor** ( ) The mapped RNA-Seq reads. The RNA-Seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html`).

5. **Differentially Expressed Genes** file ( ) A track showing the differentially expressed genes. The table view provides information about fold change, difference in expression, the maximum expression (observed in either the case or the control), the expression in the case, and the expression in the control.

6. **Variant Calling Report Tumor** ( ) Report showing error rates for quality categories, quality of examined sites, and estimated frequencies of actual to called bases for different quality score ranges.

7. **Annotated Somatic Variants with Expression Values** ( ) A variant track showing the somatic variants. When mousing over a variant, a tooltip will appear with information about the variant.

8. **Amino Acid Track**

9. **Track List RNA-Seq Tumor_Normal Comparison** ( ) A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar and dbSNP Common (see figure 21.25).

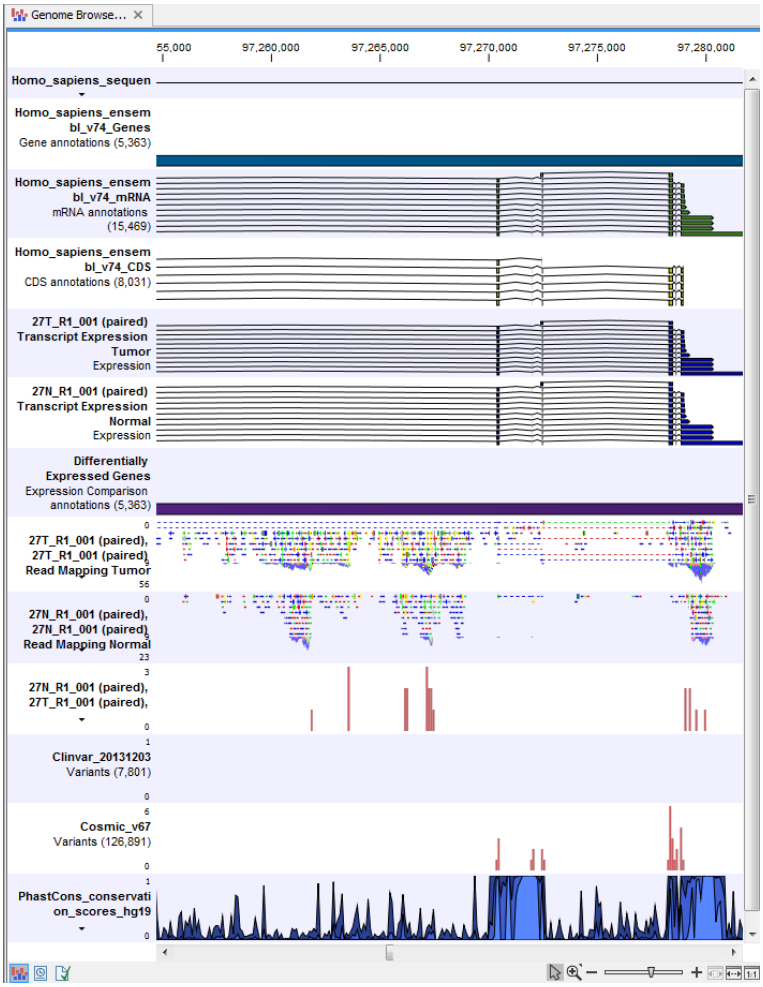


Figure 21.25: *The Track List is a collection of tracks that makes it easy to compare them to each other. Each track kan be opened individually by double-clicking on the track name in the left side of the Track List view.*

21.5 Identify Variants and Add Expression Values

The **Identify Variants and Add Expression Values** workflows can be used to identify novel and known mutations in RNA-Seq data, automatically map, quantify, and annotate the transcriptomes, and compare the mutational patterns in the samples with the expression values of the corresponding transcripts and genes.

The workflows can be found under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Human** ( ) | **Identify Variants and Add Expression Values** ( )
>
> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **Whole Transcriptome Sequencing** ( ) | **Mouse and Rat** ( ) | **Identify Variants and Add Expression Values (M and R)** ( )

After starting the workflow:

1. If you are connected to a server, you will first be asked where you would like to run the analysis.

2. Specify the **RNA-Seq reads** to analyze. The reads can be selected by double-clicking on the reads file name or clicking once on the file and then clicking on the arrow pointing to the right side in the middle of the wizard (figure 21.26).

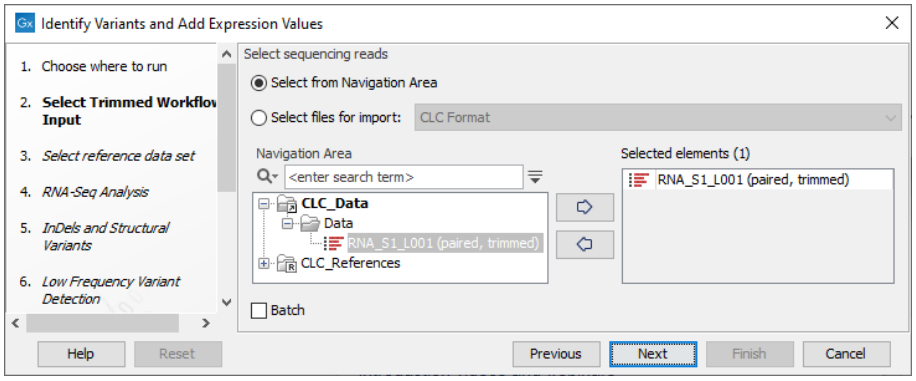


Figure 21.26: *Select the sequencing reads to analyze.*

3. Select the Reference Data Set that is relevant to your study (figure 21.27).

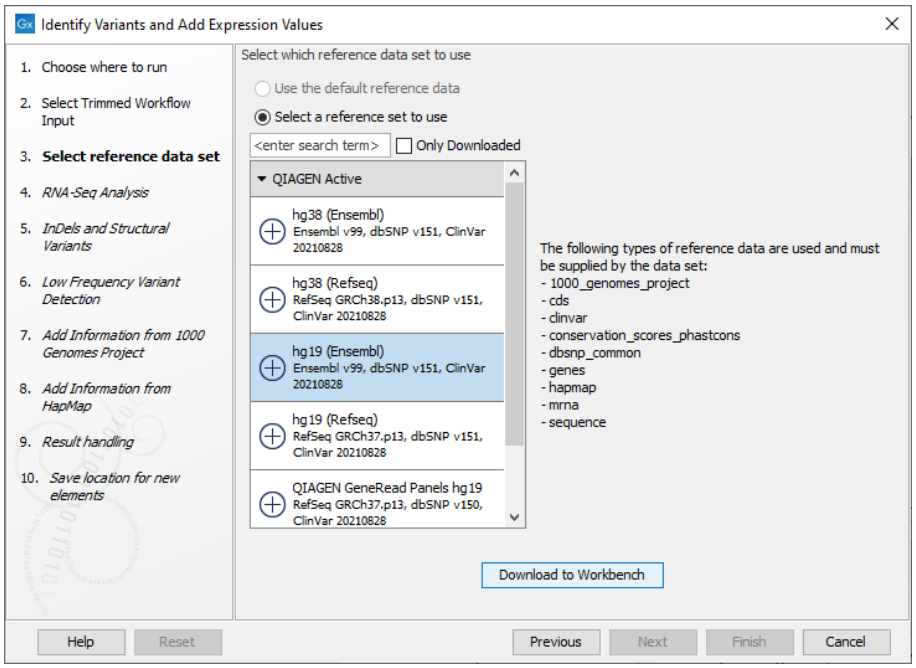


Figure 21.27: *Select the relevant data set for the samples being studied.*

4. Configure the parameters for the RNA-Seq Analysis (figure 21.28).

   If you wish to use spike-in controls, add the relevant file in the "Spike-in controls" field.

   You can also specify that the reads should be mapped only in their forward or reverse orientation (it is by default set to both). Choosing to restrict mapping to one direction is typically appropriate when a strand specific protocol for read generation has been used, as it allows assignment of the reads to the right gene in cases where overlapping genes

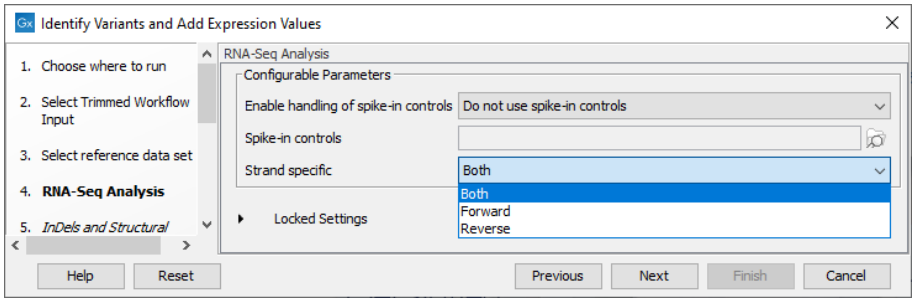


Figure 21.28: *Configure the RNA-Seq Analysis.*

are located on different strands. Also, applying the 'strand specific' 'reverse' option in an RNA-Seq run could allow the user to assess the degree of antisense transcription. Note that mate pairs are not supported when choosing the forward only or reverse only option.

5. Specify a **target region** for the Indels and Structural Variants tool (figure 21.29).

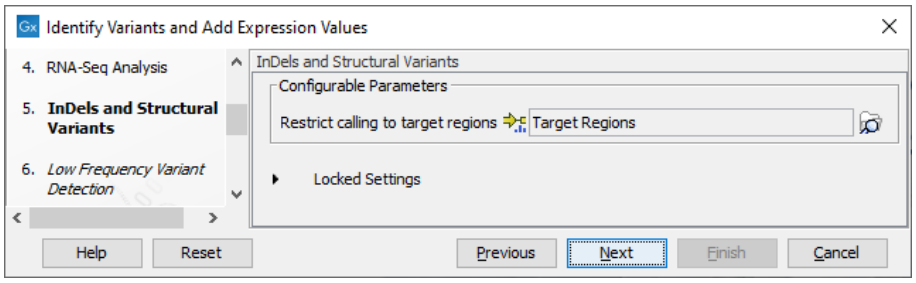


Figure 21.29: *Specify the target region for the Indels and Structural Variants tool.*

The targeted region file is a file that specifies which regions have been sequenced. This file is something that you must provide yourself, as this file depends on the technology used for sequencing. You can obtain the targeted regions file from the vendor of your targeted sequencing reagents. Remember that you have a hg38-specific BED file when using hg38 as reference, and hg19-specific BED file when using hg19 as reference.

6. Set the parameters for the Low Frequency Variant Detection step (see figure 21.30). For a description of the different parameters that can be adjusted in the variant detection step, see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html`.

7. If you are working with the workflow for Human, specify here the relevant **1000 Genomes** population (and **HapMap** populations at the next step) (see figure 21.31). Choose the population that matches best the population your samples are derived from.

   Under "Locked settings" you can see that "Automatically join adjacent MNVs and SNVs" has been selected. The reason for this is that many databases do not report a succession of SNVs as one MNV as is the case for CLC Workbench, and as a consequence it is not possible to directly compare variants called with CLC Workbench with these databases. In order to support filtering against these databases anyway, the option to **Automatically join adjacent MNVs and SNVs** is enabled. This means that an MNV in the experimental data will get an exact match, if a set of SNVs and MNVs in the database can be combined to provide the same allele. This assumes that SNVs and MNVs in the track of known variants

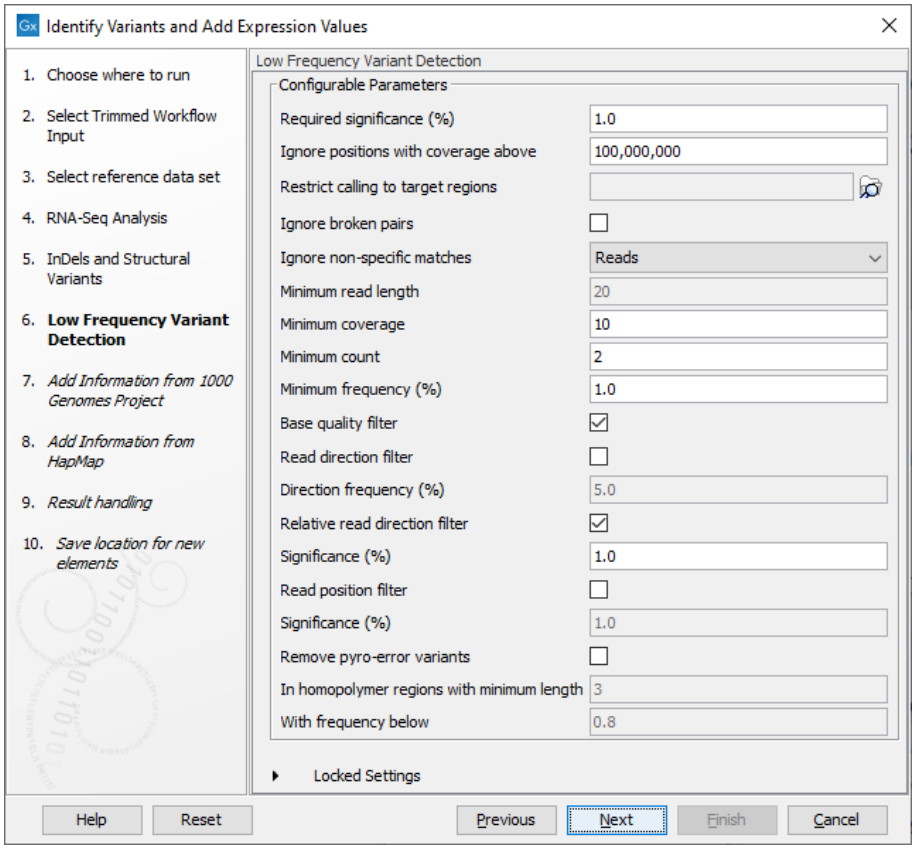


Figure 21.30: *Specify the parametes for transcriptomic variant detection.*

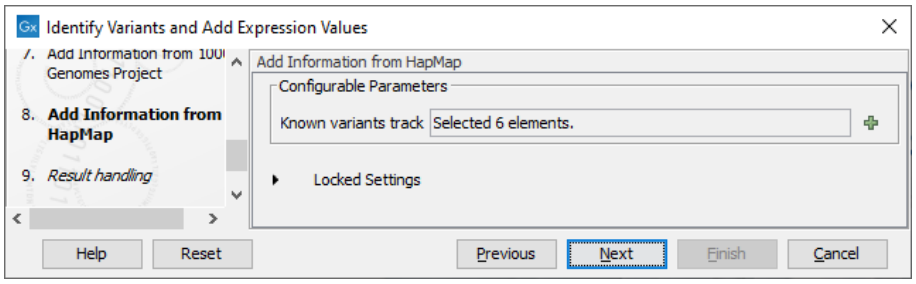


Figure 21.31: *Select the relevant population from the drop-down list for Hapmap databases.*

represent the same allele, although there is no evidence for this in the track of known variants.

8. In the last wizard step you can check the selected settings by clicking on the button labeled **Preview All Parameters**. In the **Preview All Parameters** wizard you can only check the settings, and if you wish to make changes you have to use the **Previous** button from the wizard to edit parameters in the relevant windows.

9. Choose to **Save** your results and click on the button labeled **Finish**.

The following outputs are generated:

1. **Gene expression** ( ) A track showing gene expression annotations. Hold the mouse over

or right-click on the track: a tooltip will appear with information about e.g. gene name and expression values.

2. **Transcript expression** () A track showing transcript expression annotations.

3. **RNA-Seq Mapping Report** () This report contains information about the reads, reference, transcripts, and statistics. This is explained in more details here: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html.

4. **Read Mapping** () The mapped RNA-Seq reads. The RNA-Seq reads are shown in different colors depending on their orientation, whether they are single reads or paired reads, and whether they map unambiguously (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Coloring_mapped_reads.html).

5. **Annotated Variants with Expression Values** () Annotation track showing the variants. Hold the mouse over one of the variants or right-clicking on the variant. A tooltip will appear with detailed information about the variant.

6. **RNA-Seq Track List** () A collection of tracks presented together. Shows the annotated variant track together with the human reference sequence, genes, transcripts, coding regions, and variants detected in ClinVar and dbSNP (see figure 21.16).

7. **Log** () A log of the workflow execution.

Part V

Legacy tools and workflows

Chapter 22

Legacy workflows

Contents

22.1 QIAGEN GeneRead Panel Analysis (legacy)

The **QIAGEN GeneRead Panel Analysis (legacy)** is a template workflow that can identify and annotate variants in Targeted Amplicon Sequencing data generated with GeneRead DNAseq Gene Panels. The GeneRead DNAseq Gene Panels can either be standard panels focused on a specific set of genes or can be customized to include genes tailored to specific research interests.

The first step in the template workflow is mapping of the sequencing reads to the human reference sequence. This is followed by a local realignment step, which is included to improve the variant detection that follows directly after a primer trimming step. After variant detection, the variants are annotated with gene names, exon numbers, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and information from common variants present in the common dbSNP, HapMap, and 1000 Genomes database. Furthermore, a detailed target regions mapping report is created that allows inspection of the coverage and mapping specificity in the target regions.

The **QIAGEN GeneRead Panel Analysis (legacy)** template workflow assumes that the sequences used as input do not contain adapters as the removal of adapters is often done directly on the sequencing machine. If adapters have not been trimmed off, please do so before proceeding with your analysis by using the Trim Reads tool (from the "Prepare Raw Data" folder) with the "Automatic read-through adapter trimming" option enabled.

The **QIAGEN GeneRead Panel Analysis (legacy)** template workflow is available under the Workflows menu at:

> **Workflows** | **Template Workflows** | **Legacy Template Workflows** | **QIAGEN GeneRead Panel Analysis (legacy)** ()

Double-click on the **QIAGEN GeneRead Panel Analysis (legacy)** workflow to run the analysis.

If you are connected to a *CLC Server* via the *CLC Workbench*, you will be asked where you would

like to run the analysis. We recommend that you run the analysis on a *CLC Server* when possible. Click **Next**.

Select the sequencing reads that should be analyzed (figure 22.1).

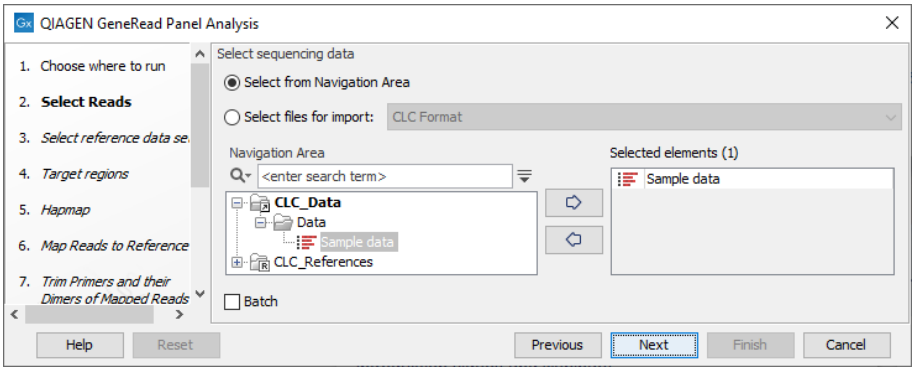


Figure 22.1: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

If you would like to analyze more than one sample you can choose to run the analysis in batch mode. This is done by ticking "Batch" in the lower left side of the wizard and selecting the folder(s) that holds the data you wish to analyze. If you have your sequencing data in separate folders, you should choose to run the analysis in batch mode. You can learn more about batch analysis in the CLC Workbench user manual (see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batch_processing.html`).

In the next window, specify the relevant QIAGEN reference data set to be used with the workflow (figure 22.2)

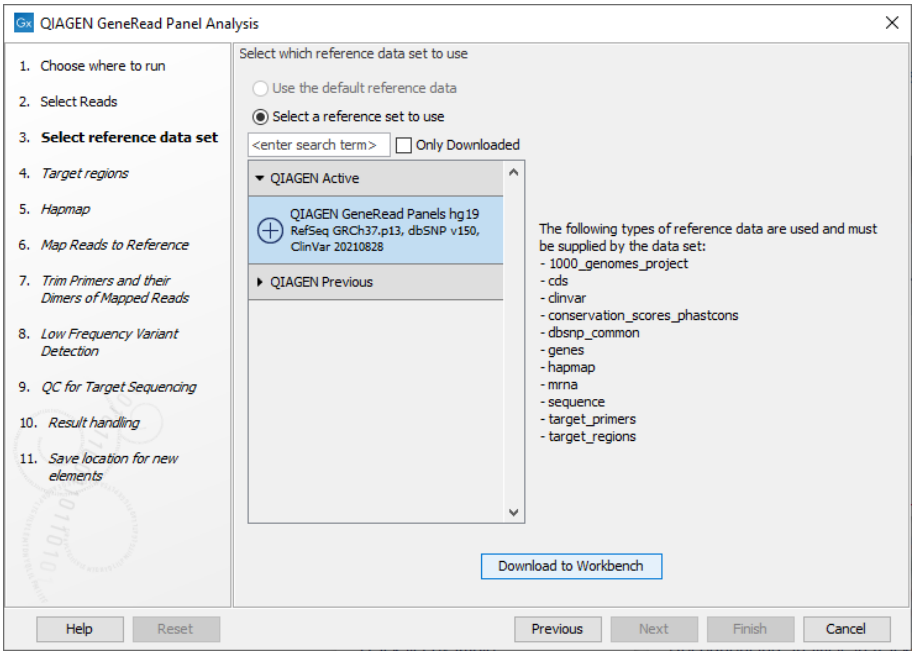


Figure 22.2: *The default reference data set that can be selected is called QIAGEN GeneRead Panels hg19.*

In the next wizard window (figure 22.3), you must specify the target regions fitting your sample from the drop down menu.

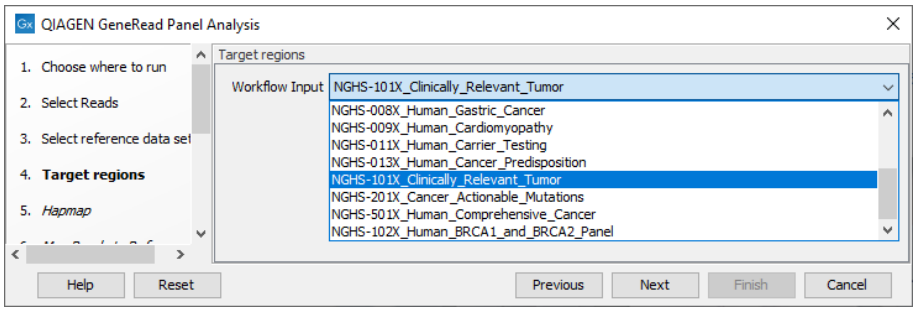


Figure 22.3: *In this wizard step you can specify the targeted regions matching your read mapping.*

In the next dialog, **Hapmap**, you can specify the populations that fit your dataset. Indeed, detected variants are annotated with a range of different data in this template workflow, but for databases that provide data from more than one population as HapMap does, the populations relevant to the data set can be specified by the user (figure 22.4).

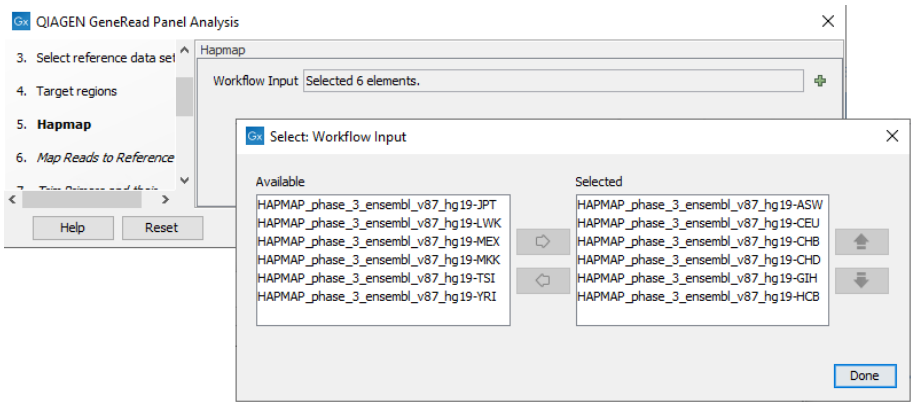


Figure 22.4: *Select the relevant population from the list or use all populations that have already been selected.*

From the list that can be accessed by clicking on the plus symbol ( ) you can choose the population that matches the population your samples are derived from. Please note that the populations available from the drop-down list can be specified with the Reference Data Manager found in the top right corner of the CLC Workbench.

In the **Map Reads to Reference** wizard step (figure 22.5), you can configure the read mapper by setting the "Cost of insertions and deletions" to either "Affine gap cost" (default) or "Linear gap cost".

- **Linear gap cost** The cost of a gap is computed directly from the length of the gap and the insertion or deletion cost. This model often favors small, fragmented gaps over long contiguous gaps.
- **Affine gap cost** An extra cost associated with opening a gap is introduced such that long contiguous gaps are favored over short gaps.

Specify the target primers for primer trimming in the **Trim Primers and their Dimers of**

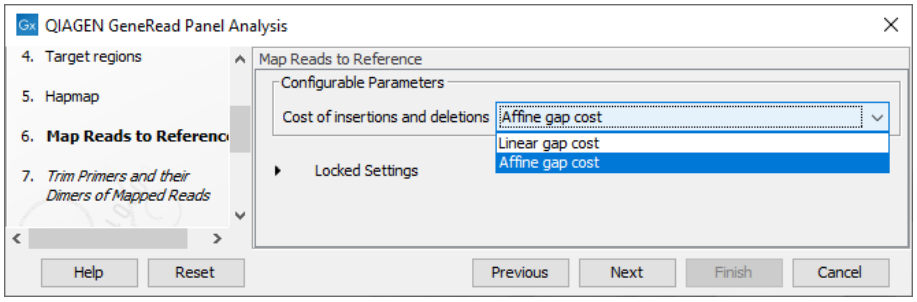


Figure 22.5: *In this wizard step you can set the "Cost of insertions and deletions" to either "Affine gap cost" (default) or "Linear gap cost".*

Mapped Reads window (figure 22.6). If you would like to add more GeneRead DNAseq Gene Panel target primers, this can be done using the Reference Data Manager as described in https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html. It is also possible to either enable or disable the parameter "Only keep reads that have hit a primer". Note that it is enabled by default.

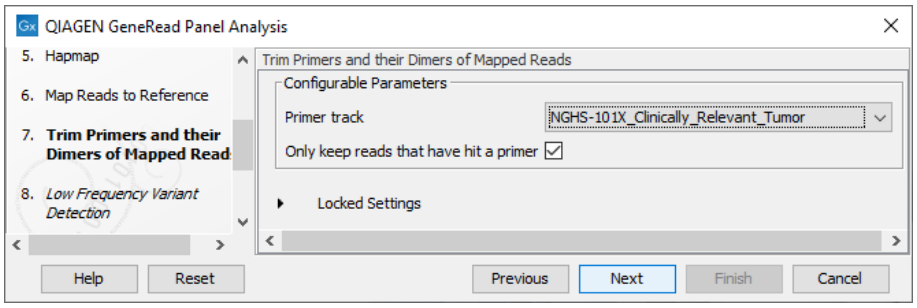


Figure 22.6: *Select the primer track from the drop-down list.*

In the **Low Frequency Variant Detection** wizard step (figure 22.7), you can specify the parameters for variant detection.

Please see the CLC Workbench user manual for a description of the different parameters that can be adjusted in the variant detection step. A description of the "Low Frequency Variant Detection" tool can be found here: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html.

In the **QC for Targeted Sequencing** wizard step (figure 22.8), you can specify:

- **Minimum coverage** i.e., the minimum coverage needed on all positions in a target, in order for that target to be considered covered.
- **Ignore non-specific matches and/or broken pairs** When these are applied reads that are non-specifically mapped or belong to broken pairs will be ignored.

Finally, in the last wizard step, pressing the button Preview All Parameters allows you to preview all parameters, but to make any changes, you must use the button **Previous** and **Next** to reach the relevant wizard window. If no change is necessary, choose to save the results and click **Finish**.

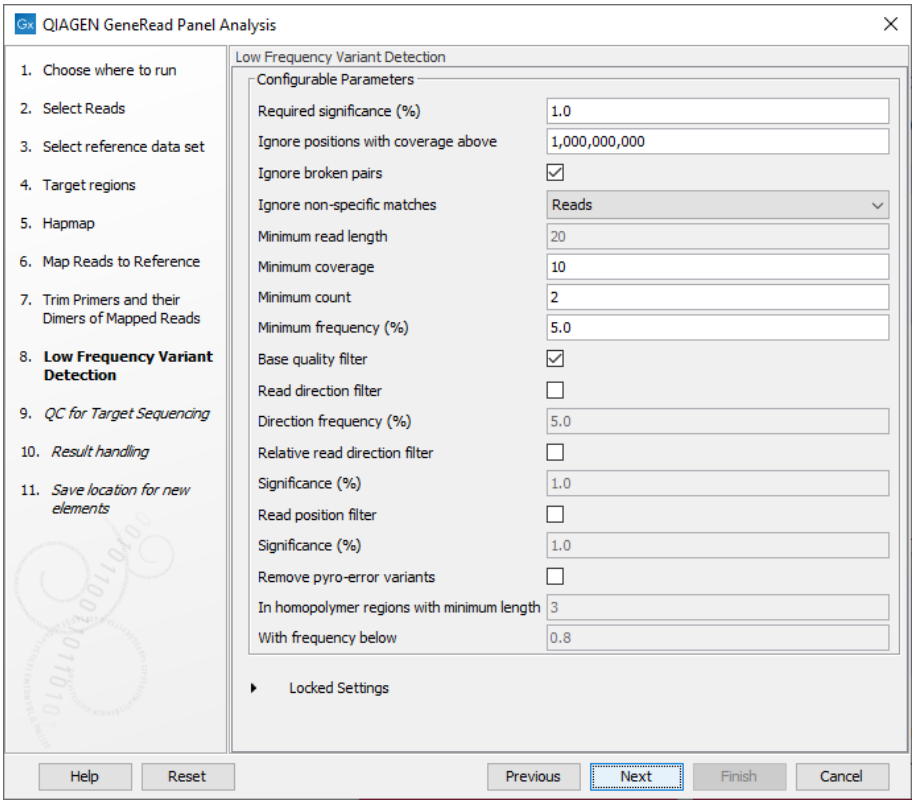


Figure 22.7: *In this wizard step the parameters for variant detection can be adjusted.*

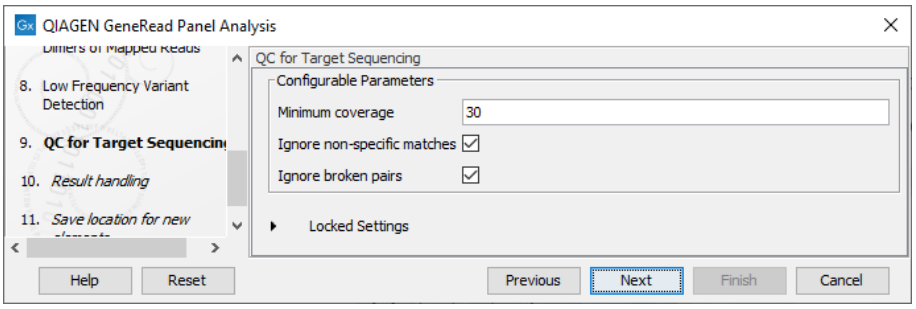


Figure 22.8: *Adjust the parameters if desired.*

22.1.1 Output from QIAGEN GeneRead Panel Analysis (legacy)

The QIAGEN GeneRead Panel Analysis (legacy) workflow produces seven different outputs:

- Trimmed Reads report
- Trimmed reads mapping
- Target region coverage track
- Coverage report
- Amino Acid Changes track
- Annotated variant track

- Indels indirect evidence track ( )
- Genome Browser View ( )

Note! We advise you to not delete any of the produced files individually as some of them are linked to each other. If you would like to delete an experiment, please always delete all of generated files from one experiment at the same time.

When looking at the results of the analysis, a good place to start is the target region coverage report ( ) to see whether the coverage is sufficient in the regions of interest (e.g. >30). Please also check that at least 90% of the reads are mapped to the human reference sequence and that the majority of the reads map to the targeted region.

Open the Genome Browser View file ( ) to get an overview of the identified variants (see 22.9).

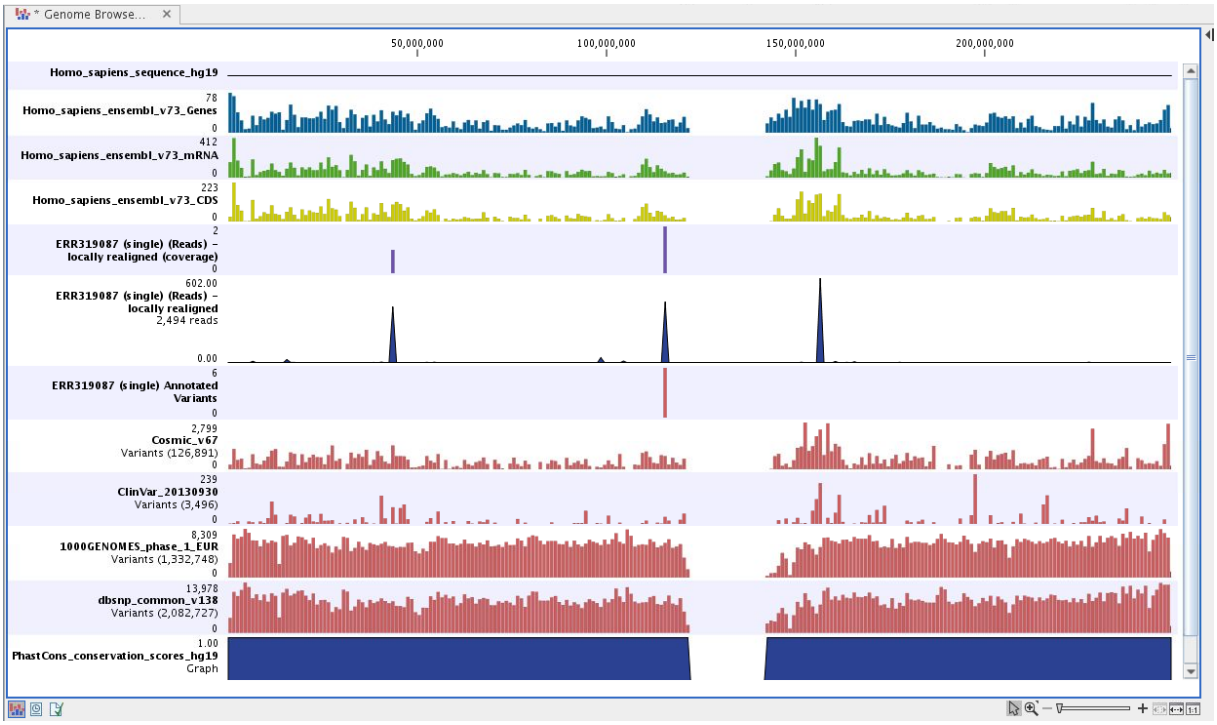


Figure 22.9: *Genome Browser View to inspect identified variants in the context of the human genome and external databases.*

The Genome Browser View includes the annotated variants in context to the human reference sequence, genes, transcripts, coding regions, targeted regions, mapped sequencing reads, relevant variants in the ClinVar database as well as common variants in common dbSNP, HapMap and 1000 Genomes databases. Finally, a track with conservation scores shows the level of nucleotide conservation around each variant.

The conservation scores are based on a multiple alignment with a range of different vertebrates. The conservation in the region around each variant is particularly relevant when you consider the potential importance of the individual variants. A high conservation score could indicate that the variant is located in a region of the genome that is of great importance.

The annotated variant track can also be shown in table view. To open the table, double-click on the name of the variant track in the left side of the Genome Browser View (when opened in the View

Area). The annotated variant table includes all variants and the added information/annotations (see 22.10).

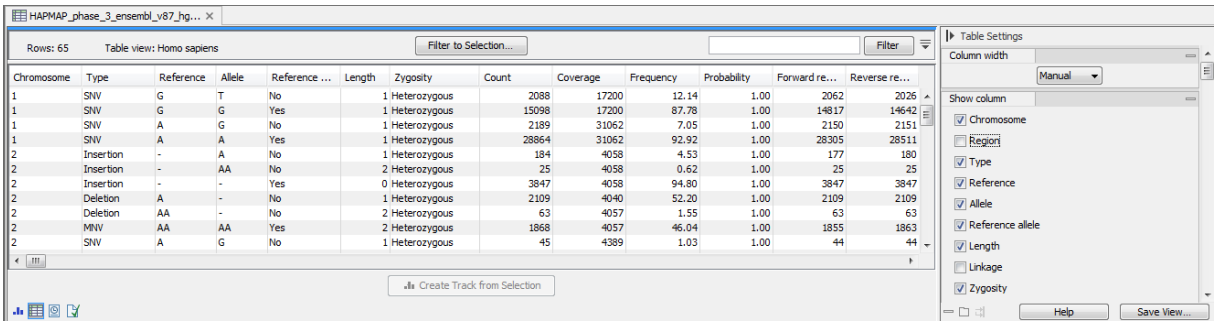


Figure 22.10: *The annotated variant track opened in table view from the Genome Browser View. The table makes it easier to inspect identified variants in detail.*

In figure 22.11 the annotated variant table and the Genome Browser View are shown in split view. The annotated variant table and the Genome Browser View are connected and when selecting a variant from the table by clicking on a row in the table, the Genome Browser View will automatically put the selected variant into focus. In figure 22.11 the "Zoom to base level" function (1:1), marked with a red arrow in the lower right corner of the View Area, has been used to zoom in on the variant.

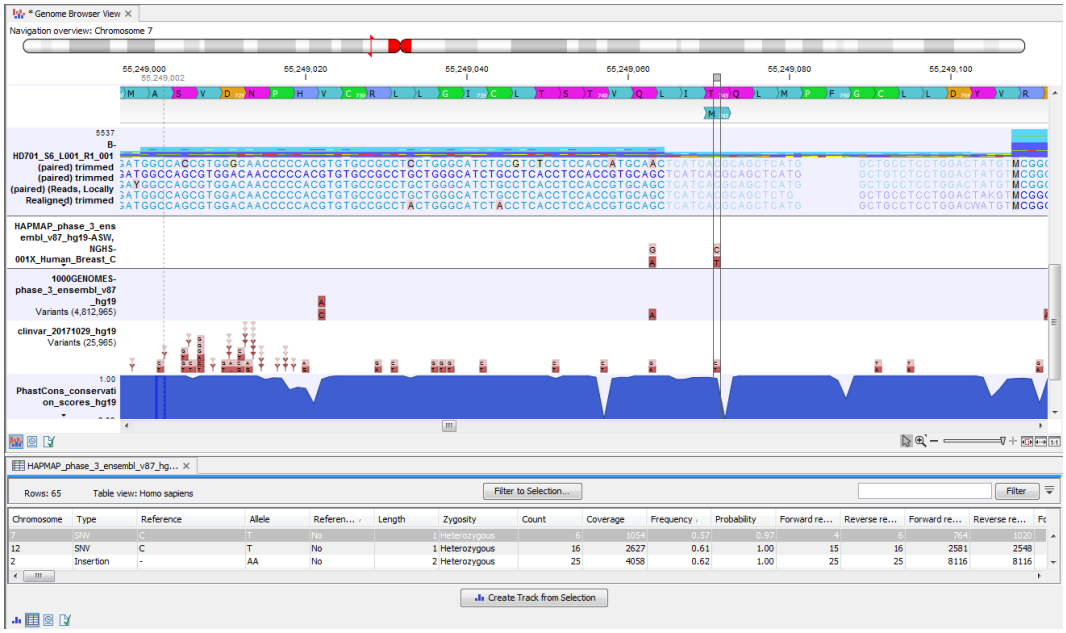


Figure 22.11: *The annotated variant table and the Genome Browser View shown in split view.*

The added information can support identification of candidate variants for further research. For example common genetic variants (present in the HapMap database) or variants known to play a role in drug response or other relevant phenotypes (present in the ClinVar database) can easily be singled out using the table.

Also, identified variants that are unknown in the ClinVar database can be for example prioritized based on amino acid changes. A high conservation level on the position of the variant between

many vertebrates or mammals can also be a hint that this region could have an important functional role, with variants with a conservation score of more than 0.9 (PhastCons score) that should be prioritized higher. Filtering of the variants based on their annotations can be facilitated using the table filter in the top right side of the table.

Please note that in case none of the variants are present in ClinVar or dbSNP, the corresponding annotation column headers are missing from the result.

Part VI

Appendices

Chapter 23

Install and uninstall plugins

Contents

Biomedical Genomics Analysis is installed as a plugin.

23.1 Installation of plugins

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

Plugins and modules are installed and uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins ( ) button** in the top Toolbar, or go to the menu option:

Utilities | **Manage Plugins... ( )**

The Plugin Manager has two tabs at the top:

- **Manage Plugins** An overview of your installed plugins and modules is provided under this tab.
- **Download Plugins** Plugins and modules available to download and install are listed in this tab.

To install a plugin, click on the **Download Plugins** tab (figure 23.1). Select a plugin. Information about it will be shown in the right hand panel. Click on the **Download and Install** button to install the plugin.

Accepting the license agreement

The End User License Agreement (EULA) must be read and accepted as part of the installation process. Please read the EULA text carefully, and if you agree to it, check the box next to the

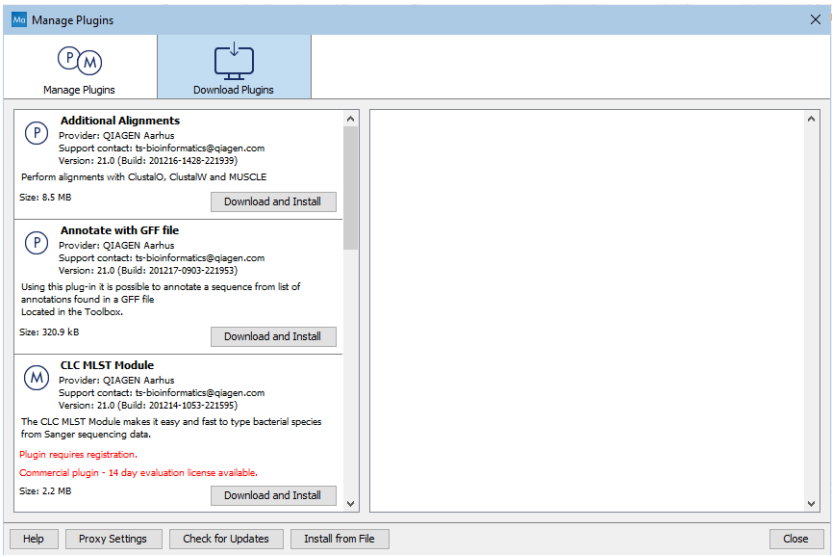


Figure 23.1: *Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.*

text **I accept these terms**. If further information is requested from you, please fill this in before clicking on the **Finish** button.

Installing a cpa file

If you have a .cpa installer file for Biomedical Genomics Analysis, you can install it by clicking on the **Install from File** button at the bottom of the Plugin Manager.

If you are working on a system not connected to the internet, plugin and module .cpa files can be downloaded from `https://digitalinsights.qiagen.com/products-overview/plugins/`using a networked machine, and then transferred to the non-networked machine for installation.

Restart to complete the installation

Newly installed plugins and modules will be available for use after restarting the software. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the *CLC Workbench*.

23.2 Uninstalling plugins

Plugins and modules are uninstalled using the Workbench Plugin Manager.To open the Plugin Manager, click on the **Plugins ( ) button** in the top Toolbar, or go to the menu option:

Utilities | Manage Plugins... ( )

This will open the Plugin Manager (figure 23.2). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

To uninstall a plugin or module, click on its entry in the list, and click on the **Uninstall** button.

Plugins and modules are not uninstalled until the Workbench is restarted. When you close the

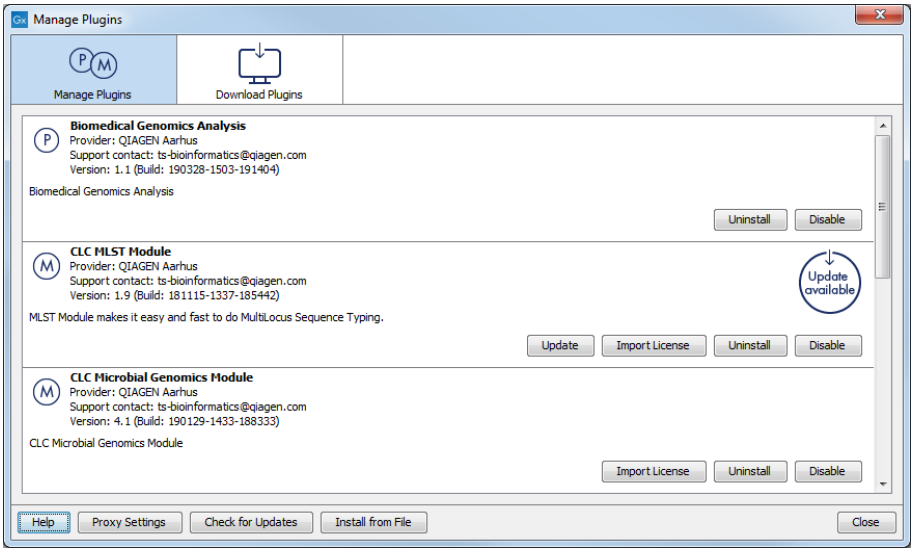


Figure 23.2: *Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.*

Plugin Manager, a dialog appears offering the opportunity to restart the *CLC Workbench*.

Disabling a plugin without uninstalling it

If you do not want a plugin to be loaded the next time you start the Workbench, select it in the list under the Manage Plugins tab and click on the **Disable** button.

Bibliography

[Abkevich et al., 2012] Abkevich, V., Timms, K. M., Hennessy, B. T., Potter, J., Carey, M. S., Meyer, L. A., Smith-McCune, K., Broaddus, R., Lu, K. H., Chen, J., Tran, T. V., Williams, D., Iliev, D., Jammulapati, S., FitzGerald, L. M., Krivak, T., DeLoia, J. A., Gutin, A., Mills, G. B., and Lanchbury, J. S. (2012). Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British Journal of Cancer*, 107(10):1776–1782.

[Beroukhim et al., 2006] Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L. A., Fox, E. A., Hochberg, E. P., Mellinghoff, I. K., Hofer, M. D., et al. (2006). Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide snp arrays. *PLoS Comput Biol*, 2(5):e41.

[Birkbak et al., 2012] Birkbak, N. J., Wang, Z. C., Kim, J.-Y., Eklund, A. C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greene-Colozzi, A., Iglehart, J. D., Tung, N., Ryan, P. D., Garber, J. E., Silver, D. P., Szallasi, Z., and Richardson, A. L. (2012). Telomeric allelic imbalance indicates defective dna repair and sensitivity to dna-damaging agents. *Cancer Discovery*, 2(4):366–375.

[Chao, 1987] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791.

[Chao et al., 2014] Chao, A., Gotelli, N. J., Hsieh, T., Sander, E. L., Ma, K., Colwell, R. K., and Ellison, A. M. (2014). Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological monographs*, 84(1):45–67.

[Chao et al., 2013] Chao, A., Wang, Y., and Jost, L. (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, 4(11):1091–1100.

[Choi et al., 2009] Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*, 106(45):19096–19101.

[de Luca et al., 2020] de Luca, X. M., Newell, F., Kazakoff, S. H., Hartel, G., McCart Reed, A. E., Holmes, O., Xu, Q., Wood, S., Leonard, C., Pearson, J. V., Lakhani, S. R., Waddell, N., Nones, K., and Simpson, P. T. (2020). Using whole-genome sequencing data to derive the homologous recombination deficiency scores. *npj Breast Cancer*, 6(1):33.

[Glusman et al., 2001] Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J. C., Smit, A. F., Wang, K., Koop, B. F., and Hood, L. (2001). Comparative genomics of the human and mouse t cell receptor loci. *Immunity*, 15(3):337–349.

[Heap et al., 2010] Heap, G. A., Yang, J. H. M., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., Franke, L., Dubois, P. C., Mein, C. A., Dobson, R. J., Albert, T. J., Rodesch, M. J., Clayton, D. G., Todd, J. A., van Heel, D. A., and Plagnol, V. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*, 19(1):122–134.

[Hiatt et al., 2013] Hiatt, J., Pritchard, C., Salipante, S., O'Roak, B., and Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Research*, (23):843–854.

[Kuchenbecker et al., 2015] Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A. U., Babel, N., Reinert, K., and Robinson, P. N. (2015). Imseq–a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics*, 31(18):2963–2971.

[Lefranc et al., 2009] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. (2009). Imgt®, the international immunogenetics information system®. *Nucleic acids research*, 37(suppl_1):D1006–D1012.

[Martin and Wang, 2011] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, 12(10):671–682.

[Ng et al., 2009] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

[Peng et al., 2015] Peng, Q., Satya, R. V., Lewis, M., Randad, P., and Wang, Y. (2015). Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics*, (16):589.

[Shugay et al., 2015] Shugay, M., Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., Putintseva, E. V., Pogorelyy, M. V., Nazarov, V. I., Zvyagin, I. V., Kirgizova, V. I., et al. (2015). Vdjtools: unifying post-analysis of t cell receptor repertoires. *PLoS computational biology*, 11(11):e1004503.

[Shugay et al., 2017] Shugay, M., Zaretsky, A. R., Shagin, D. A., Shagina, I. A., Volchenkov, I. A., Shelenkov, A. A., Lebedin, M. Y., Bagaev, D. V., Lukyanov, S., and Chudakov, D. M. (2017). Mageri: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS computational biology*, 13(5):e1005480.

[Vega et al., 2021] Vega, D., Yee, L., McShane, L., Williams, P., Chen, L., Vilimas, T., Fabrizio, D., Funari, V., Newberg, J., Bruce, L., Chen, S.-J., Baden, J., Barrett, J. C., Beer, P., Butler, M., Cheng, J.-H., Conroy, J., Cyanam, D., Eyring, K., Garcia, E., Green, G., Gregersen, V., Hellmann, M., Keefer, L., Lasiter, L., Lazar, A., Li, M.-C., MacConaill, L., Meier, K., Mellert, H., Pabla, S., Pallavajjalla, A., Pestano, G., Salgado, R., Samara, R., Sokol, E., Stafford, P., Budczies, J., Stenzinger, A., Tom, W., Valkenburg, K., Wang, X., Weigman, V., Xie, M., Xie, Q., Zehir, A., Zhao, C., Zhao, Y., Stewart, M., and on behalf of the TMB Consortium, J. A. (2021). Aligning tumor mutational burden (tmb) quantification across diagnostic platforms: phase ii of the friends of cancer research tmb harmonization project. *Annals of Oncology*, 132(12):1626–1636.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.