



CLC **LightSpeed** Module

USER MANUAL

User manual for QIAGEN CLC LightSpeed Module 24.0

Windows, macOS and Linux

January 5, 2024

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark



Contents

I	Introduction	6
1	Introduction	7
1.1	Contact information	7
1.2	System requirements	8
1.3	Installing modules	8
1.3.1	Licensing modules	9
1.3.2	Uninstalling modules	10
1.4	Installing server extensions	11
1.4.1	Licensing server extensions	14
II	Methods and tools	15
2	Methods	16
2.1	Trimming	16
2.2	Readmapping	17
2.3	Deduplication	17
2.4	Local realignment	18
2.5	Structural variant detection	18
2.6	UMI grouping	19
2.7	Germline variant detection	20
2.8	Somatic variant detection	20
2.9	Tumor normal variant detection	21
2.10	Limitations	22
3	Tools	23

3.1	LightSpeed Fastq to Germline Variants	23
3.1.1	LightSpeed Fastq to Germline Variants outputs	27
3.2	LightSpeed Fastq to Somatic Variants	28
3.2.1	LightSpeed Fastq to Somatic Variants outputs	33
3.3	LightSpeed Fastq to Somatic Variants Tumor Normal	34
3.3.1	LightSpeed Fastq to Somatic Variants Tumor Normal outputs	38
3.4	Report	39
III	Template Workflows	44
4	General Template Workflows	45
4.1	Fastq to Annotated Germline Variants	45
4.1.1	Outputs from Fastq to Annotated Germline Variants	47
4.2	Fastq to Annotated Germline Variants with Coverage Analysis	48
4.2.1	Outputs from Fastq to Annotated Germline Variants with Coverage Analysis	51
4.3	Fastq to Annotated Somatic Variants	52
4.3.1	Outputs from Fastq to Annotated Somatic Variants	53
4.4	Fastq to Annotated Somatic Variants with Coverage Analysis	54
4.4.1	Outputs from Fastq to Annotated Somatic Variants with Coverage Analysis	56
4.5	Fastq to Annotated Somatic Variants (Tumor Normal)	58
4.5.1	Outputs from Fastq to Annotated Somatic Variants (Tumor Normal)	59
4.6	Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis	60
4.6.1	Outputs from Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis	62
4.7	Fastq to Germline CNV Control	63
4.7.1	Outputs from Fastq to Germline CNV Control	64
4.8	Fastq to Somatic CNV Control	65
4.8.1	Outputs from Fastq to Somatic CNV Control	67
5	Template Workflows - QIAseq Targeted DNA	69
5.1	QIAseq Fastq to Annotated Germline Variants	69
5.1.1	Outputs from QIAseq Fastq to Annotated Germline Variants	72
5.2	QIAseq Fastq to Annotated Somatic Variants	73

5.2.1	Outputs from QIAseq Fastq to Annotated Somatic Variants	76
5.3	QIAseq Fastq to Germline CNV Control	77
5.3.1	Outputs from QIAseq Fastq to Germline CNV Control	79
5.4	QIAseq Fastq to Somatic CNV Control	80
5.4.1	Outputs from QIAseq Fastq to Somatic CNV Control	82
6	Template Workflows - QIAseq Targeted DNA Pro	84
6.1	QIAseq Pro Fastq to Annotated Germline Variants	84
6.1.1	Outputs from QIAseq Pro Fastq to Annotated Germline Variants	88
6.2	QIAseq Pro Fastq to Annotated Somatic Variants	89
6.2.1	Outputs from QIAseq Pro Fastq to Annotated Somatic Variants	92
6.3	QIAseq Pro Fastq to Germline CNV Control	93
6.3.1	Outputs from QIAseq Pro Fastq to Germline CNV Control	95
6.4	QIAseq Pro Fastq to Somatic CNV Control	96
6.4.1	Outputs from QIAseq Pro Fastq to Somatic CNV Control	97

Part I

Introduction

Chapter 1

Introduction

Welcome to QIAGEN CLC LightSpeed Module 24.0 – a software package supporting your daily bioinformatics work.

The CLC LightSpeed Module provides ultra-fast secondary analysis. Raw FASTQ files are quickly processed to produce variant calls with high accuracy, without requiring specialized hardware.

1.1 Contact information

QIAGEN CLC LightSpeed Module is developed by:

QIAGEN Aarhus
Silkeborgvej 2
Prismet
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: http://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>

1.2 System requirements

A licensed *CLC Genomics Workbench* is needed to make use of the CLC LightSpeed Module. A licensed *CLC Genomics Server* is needed to make use of the CLC LightSpeed Server Extension.


System requirements for CLC software is provided on <https://digitalinsights.qiagen.com/technical-support/system-requirements/>

The system requirements for QIAGEN CLC LightSpeed Module are the same as those for other CLC Genomics Workbench, except for the following:

- All LightSpeed analyses require 32 GB RAM.
- A CPU that supports AVX2 or NEON instruction sets is required.

1.3 Installing modules

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 1.1). 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 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 QIAGEN CLC LightSpeed Module, you can install it by clicking on the **Install from File** button at the bottom of the Plugin Manager.

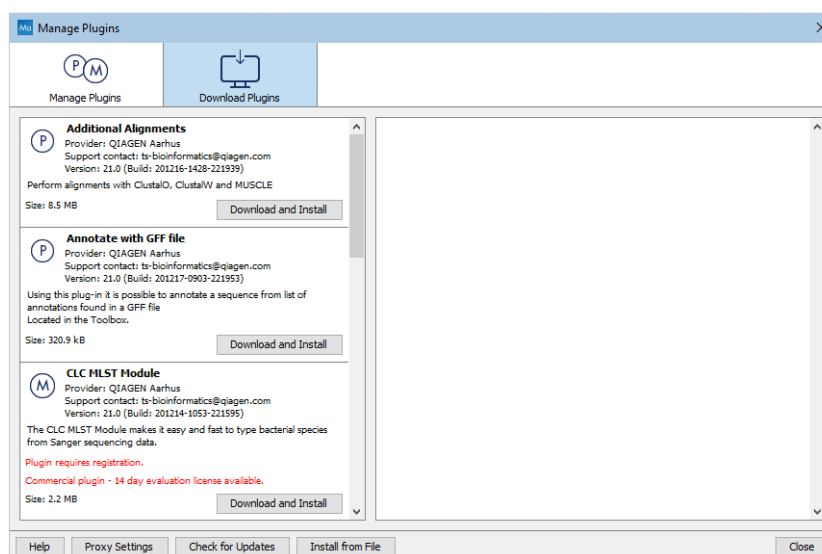


Figure 1.1: Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.

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*.

1.3.1 Licensing modules

When you have installed the QIAGEN CLC LightSpeed Module and start a tool from that module for the first time, the License Assistant will open (figure 1.2).

The License Assistant can also be launched by opening the Workbench Plugin Manager, selecting the installed module from under the Manage Plugins tab, and clicking on the button labeled *Import License*.

To install a license, 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.

The following options are available:

- **Request an evaluation license.** Request a fully functional, time-limited license.
- **Download a license.** Use the license order ID received when you purchased the software to download and install a license file.

You need a license...

In order to load the plugin "CLC Genome Finishing Module" you need a valid license.
Please choose how you would like to obtain a license for this plugin.

- ☒ **Request an evaluation license**
Choose this option if you would like to try out the plugin for 14 days.
Please note that only a single evaluation license will be allowed for each computer.
- ☐ **Download a license**
Use a license order ID to download a static license.
- ☐ **Import a license from a file**
Import a static license from an existing license file.
- ☐ **Configure License Server connection**
Configure the necessary connection for the software to connect to a CLC License Server that hosts network license(s) for this product. This option also allows you to alter or disable an existing configuration.


Figure 1.2: The License Assistant provides options for licensing modules installed on the Workbench.

- **Import a license from a file.** Import an existing license file, for example a file downloaded from the web-based licensing system.
- **Configure License Server connection.** If your organization has a *CLC Network License Manager* (or CLC License Server), select this option to configure the connection to it.

These options are described in detail in sections under http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workbench_Licenses.html.

To download licenses, including evaluation licenses, your machine must have access to the external network. To install licenses on non-networked machines, please see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Download_static_license_on_non_networked_machine.html.

1.3.2 Uninstalling modules

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 1.3). 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 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.

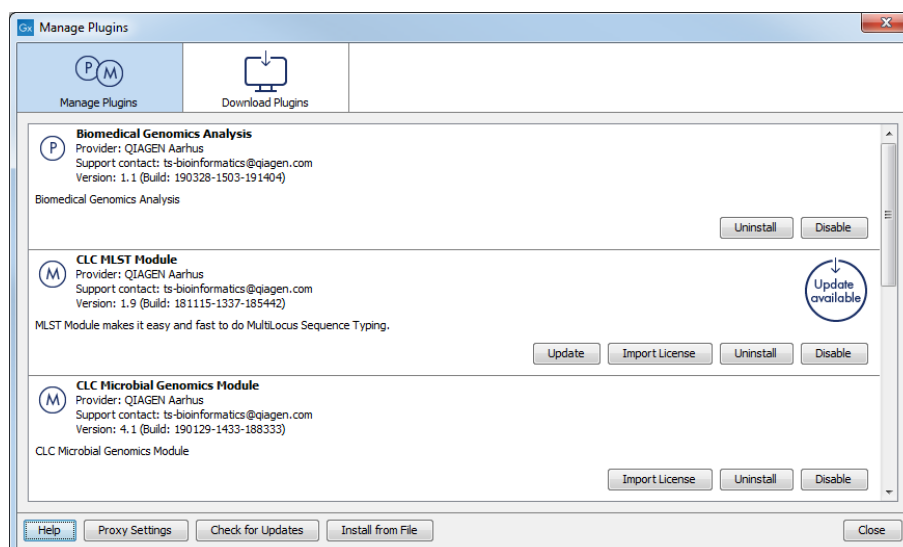


Figure 1.3: Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.

1.4 Installing server extensions

To use the tools and functionalities of QIAGEN CLC LightSpeed Module on a CLC Server:

1. You need to purchase a license to run tools delivered by the CLC LightSpeed Server Extension.
2. A CLC Server administrator must install the license on the single server, or on the master node in a job node or grid node setup, as described in section 1.4.1.
3. A CLC Server administrator must install the CLC LightSpeed Server Extension on the CLC Server, as described below.

Download and install server plugins and server extensions

Plugins, including server extensions (commercial plugins), are installed by going to the **Extensions** (🔧) tab in the web administrative interface of the single server, or the master node of a job node or grid node setup, and opening the **Download Plugins** (📄) area (figure 1.4).

If the machine has access to the external network, plugins can be both downloaded and installed via the CLC Server administrative interface. To do this, locate the plugin in the list under the **Download Plugins** (📄) area and click on the **Download and Install...** button.

To download and install multiple plugins at once on a networked machine, check the "Select for download and install" box beside each relevant plugin, and then click on the **Download and Install All...** button.

If you are working on a machine without access to the external network, server plugin (.cpa) files can be downloaded from: <https://digitalinsights.qiagen.com/products-overview/plugins/> and installed by browsing for the downloaded file and clicking on the **Install from File...** button.

The CLC Server must be restarted to complete the installation or removal of plugins and server extensions. All jobs still in the queue at the time the server is shut down will be dropped and

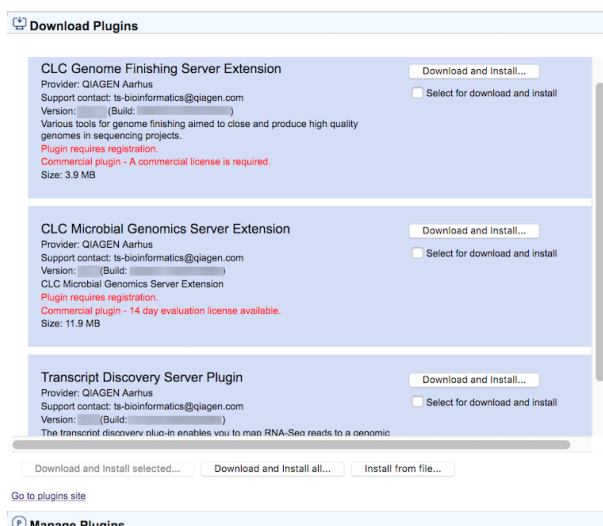


Figure 1.4: Installing plugins and server extensions is done in the Download Plugins area under the Extensions tab.

would need to be resubmitted. To minimize the impact on users, the server can be put into Maintenance Mode. In brief: running in Maintenance Mode allows current jobs to run, but no new jobs to be submitted, and users cannot log in. The CLC Server can then be restarted when desired. Each time you install or remove a plugin, you will be offered the opportunity to enter Maintenance Mode. You will also be offered the option to restart the CLC Server. If you choose not to restart when prompted, you can restart later using the option under the **Server maintenance** (🔧) tab.

For job node setups only:

- Once the *master CLC Server* is up and running normally, then restart each *job node CLC Server* so that the plugin is ready to run on each node. This is handled for you if you restart the server using the functionality under

Management (👤) | **Server maintenance** (🔧)

- In the web administrative interface on the *master CLC Server*, check that the plugin is enabled for each job node.

Installation and updating of plugins on connected job nodes requires that direct data transfer from client systems has been enabled, which is done by the CLC Server administrator, under the "External data" tab.

Grid workers will be re-deployed when a plugin is installed on the master server. Thus, no further action is needed to enable the newly installed plugin to be used on grid nodes.

Managing installed server plugins

Installed plugins can be updated or uninstalled, from under the **Manage Plugins** (Ⓟ) area (figure 1.5), under the **Extensions** (🔧) tab.

The list of tools delivered with a server plugin can be seen by clicking on the **Plugin contents** link to expand that section. Workflows delivered with a server plugin are not shown in this listing.

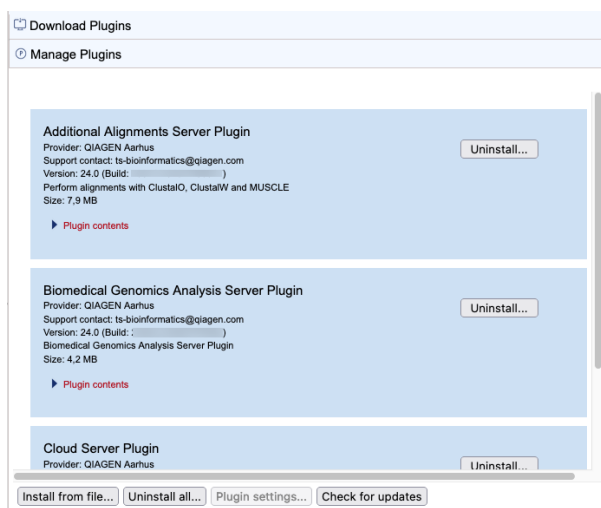


Figure 1.5: Managing installed plugins and server extensions is done in the Manage Plugins area under the Extensions tab. Clicking on Plugin contents opens a list of the tools delivered by the plugin.

Links to related documentation

- Logging into the CLC Server web administrative interface: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsserver/current/admin/index.php?manual=Logging_into_administrative_interface
- Maintenance Mode: resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server_maintenance.html
- Restarting the server: resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Starting_stopping_server.html
- Plugins on job node setups: resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Installing_Server_plugins_on_job_nodes.html
- Grid worker re-deployment: resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Overview_Model_II.html

Plugin compatibility with the server software

The version of plugins and server extensions installed must be compatible with the version of the CLC Server being run. A message is written under an installed plugin's name if it is not compatible with the version of the CLC Server software running.

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

Incompatibilities can also arise when updating to a new bug fix or minor feature release of the CLC Server. We recommend opening the **Manage Plugins** area after any server software upgrade to check for messages about the installed plugins.

Licensing server extensions is described in section [1.4.1](#).

1.4.1 Licensing server extensions

Licenses are installed on a single server or on the master node of a job node or grid node setup.

To download and install a license:

- Log into the web administrative interface of the single server or master node as an administrative user.
- Under the **Management** (🔧) tab, open the **Download License** (📄) tab.
- Enter the Order ID supplied by QIAGEN into the Order ID field and click on the "Download and Install License..." button (figure 1.6).

Please contact ts-bioinformatics@qiagen.com if you have not received an Order ID.

The CLC Server must be restarted for new license files to be loaded. Details about restarting can be found at resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Starting_stopping_server.html.

Each time you download a license file, a new file is created in the `licenses` folder under the CLC Server installation area. *If you are upgrading an existing license file, delete the old file from this area before restarting.*

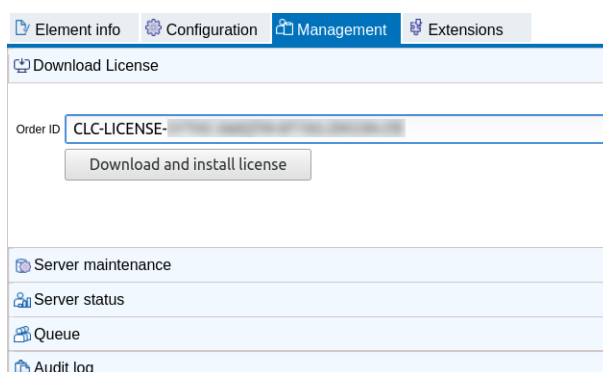


Figure 1.6: License management is done under the Management tab.

Part II

Methods and tools

Chapter 2

Methods

The CLC LightSpeed Module contains tools which facilitate end to end NGS secondary analysis using an extensive collection of algorithms. Each individual algorithm has been optimized for short runtime with minimal memory requirements while retaining accuracy of variant detection. In the following, the overall principles of core algorithms are described.

Contents

2.1 Trimming	16
2.2 Readmapping	17
2.3 Deduplication	17
2.4 Local realignment	18
2.5 Structural variant detection	18
2.6 UMI grouping	19
2.7 Germline variant detection	20
2.8 Somatic variant detection	20
2.9 Tumor normal variant detection	21
2.10 Limitations	22

2.1 Trimming

Two types of trimming are available: quality trimming and adapter trimming.

Quality trimming Raw reads are trimmed for low quality nucleotides. The method is described here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Quality_trimming.html.

For **LightSpeed Fastq to Germline Variants**, the quality limit used for trimming is 0.05 and cannot be adjusted. In **LightSpeed Fastq to Somatic Variants** and **LightSpeed Fastq to Somatic Variants Tumor Normal**, the quality limit can be adjusted.

For somatic variant calling, ensuring a high base quality is important. Therefore, the default quality limit has been set to 0.01 corresponding to a base quality of 20. This will have minimal impact on high-quality reads but can lead to markedly shorter reads and hence decreased coverage when using lower-quality reads.

Adapter trimming The algorithm can trim adapter sequences from mapped paired-end reads. For each individual read in a pair, read sequence that extends beyond the 5' end of the other read in the pair, is considered adapter sequence and is trimmed.

The consensus sequences for the removed R1 and R2 sequences, are included in the report.

It is possible to remove trimmed reads that are shorter than a defined threshold after adapter trimming.

2.2 Readmapping

Indexing When provided with a reference genome, LightSpeed first generates a Burrows-Wheeler based index of all the sequences.

Read mapping and read pairs LightSpeed maps reads to the indexed reference sequence. The quality scores are not stored.

Single reads that are part of a paired read are mapped individually. For each read, only the most likely seeds are extended using a Needleman-Wunsch based method. LightSpeed takes the relative position of individual reads in a pair into account when estimating how likely a seed is, and prefers seeds where the distance between individual reads falls within expected distance of a paired read.

Read pairs that do not map well, go through a second round of more thorough seeding.

The distance at which reads can be considered as pairs, is estimated from a subset of the reads. If there is not enough data to estimate the distance, a default insert size of 1-1000 base pairs is used. Read pairs that map within the expected distance of each other are considered pairs, read pairs that map further away from each other are considered broken pairs.

The algorithm has been optimized for the typical read length and error profile of Illumina 150 bp paired-end reads.

2.3 Deduplication

Deduplication can be used to collapse reads that likely represent the same original DNA fragment.

Reads are deduplicated through the following steps:

- Reads pairs whose outer positions are identical are considered duplicates. The outer positions are usually the 5' ends of R1 and R2 including unaligned bases.
- For each group of duplicate reads, a consensus sequence is calculated:
 - At conflicting positions, the most common base is included in the consensus read.
 - If the conflicting bases are equally represented, the consensus can be generated in two ways:
 - * When one of the bases at the conflicting position is identical to the reference symbol, the reference symbol is included in the consensus read.

- * When none of the bases at the conflicting position is identical to the reference symbol, an N is inserted in the consensus read.
- In the read mapping, the duplicate read pairs are replaced with the consensus sequence.

Because deduplication relies on the outer positions of read pairs originating from the same fragment to be identical, quality trimming can reduce the number of reads that are deduplicated.

2.4 Local realignment

Regions where the readmapping is likely to be improved through local realignment are identified and realigned. These are generally regions where reads do not align perfectly, and the imperfect read alignments are unlikely to be caused by sequencing errors. This is for example the case where long unaligned ends (potentially representing long insertions and deletions) are present in the readmapping relative to the reference.

During local realignment, the following steps are performed for each identified region:

1. A path graph of k-mers is built. The graph contains paths corresponding to all reads as well as the reference.
2. The graph undergoes refinement where paths that are unlikely to contain variants relative to the reference path are removed, and additional variants may be added.
3. Any read that intersects the region of interest is realigned against the alignment graph.

After completion of local realignment, an additional repair step of unaligned ends is performed when running the tools **LightSpeed Fastq to Somatic Variants** (section 3.2) and **LightSpeed Fastq to Somatic Variants Tumor Normal** (section 3.3). This is performed only on read pairs that are specifically mapped and serves to align unaligned ends that can be aligned if one mismatch is accepted. The mismatch is, however, not accepted on the last position of the read. Note that this process is not carried out in the first and last 10 bases of a chromosome.

2.5 Structural variant detection

The **LightSpeed Fastq to Germline Variants** and **LightSpeed Fastq to Somatic Variants** tools can infer tandem duplications and inversions from unaligned ends during the realignment step.

This works by

1. Identifying breakpoints where multiple reads share a common unaligned end at the same position.
2. Aligning the sequence of identified breakpoints (unaligned end and upstream sequence) to each other and the reference sequence up or downstream of the breakpoints to find likely matches.

Tandem duplications can be detected from pairs of breakpoints that are within 1000 base pairs of each other.

Inversions can be detected from pairs of breakpoint on the same chromosome. The tool only reports the longest possible inversion when multiple breakpoints support similar inversions.

Default inversion detection requires breakpoint support from both sides of the breakpoint, but the option **Lenient inversion detection** allows detection of inversions where each breakpoint is only supported by reads from one side of the breakpoint. Lenient inversion detection can be relevant when analyzing targeted data. Enabling lenient variant detection can lead to detection of more false positive inversions, and is also likely to increase the processing time.

Tandem duplications are reported in the variant track. Tandem duplications that are inferred from unaligned ends are annotated with **Yes** in the variant track column **Inferred from unaligned ends**. Identified inversions are reported in the inversions track.

2.6 UMI grouping

All of the LighSpeed tools can group reads based on Unique Molecular Identifiers (UMIs).

The UMI sequence is recorded and removed from the reads before trimming and mapping. After the reads have been mapped, reads with similar UMI sequence and mapping position are merged into a consensus UMI read.

The consensus is calculated following these rules:

- At conflicting positions, the most common base is included in the consensus read.
- If the conflicting bases are equally represented the consensus can be generated in two ways:
 - When one of the bases at the conflicting position is identical to the reference symbol, the reference symbol is included in the consensus read.
 - When none of the bases at the conflicting position is identical to the reference symbol, an N is inserted in the consensus read.

The following options can be used to adjust how raw reads are grouped into UMI reads:

- **Minimum group size** UMI reads must consist of at least this many raw reads. UMI reads based on fewer reads than the minimum group size are discarded.
- **Maximum UMI differences** Only reads that have this number or fewer differences between their UMI sequences can be merged into UMI reads.
- **UMI window size** Only reads that start within this many bases of each other, can be merged into UMI reads. Both R1 and R2 are considered.

Note that the maximum number of reads used for creating a UMI consensus read is 20,000. Therefore, UMI groups with more than 20,000 reads will be merged into more than one consensus UMI read.

2.7 Germline variant detection

Based on the read mapping, germline variants are identified at positions where the read alignment supports a significant difference to the reference genome.

This is achieved through a site model, where each position is first assigned a likelihood for each of the genotypes A, C, T, G, N or missing. The algorithm then iterates over the read mapping and adjusts likelihoods per position for each genotype based on observations in the data until the likelihoods no longer change. Note that broken read pairs are not considered.

Each position is then inspected, and positions where the most likely genotype(s) are different from the reference sequence are identified. As the algorithm expects the genome to be diploid and is calling germline variants, only 1 or 2 genotypes per position are considered.

Variant types LightSpeed Fastq to Germline Variants reports SNPs, MNVs and InDels and replacements provided that the variants are contained within at least one paired end read.

Variant annotations Variants identified by LightSpeed Fastq to Germline Variants are annotated with the following basic information: Chromosome, Region, Type, Reference, Allele, Reference allele, Length, Zygosity, Count, Coverage, Frequency, QUAL and Genotype. Only single base pair variants, that are not adjacent to any other variants, are assigned a QUAL score.

Read about variant annotations here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html.

2.8 Somatic variant detection

Based on the read mapping, somatic variants are identified at positions where the read alignment supports a significant difference to the reference genome. This is achieved by a significance assessment relative to the global error rate which is supplemented by a significance assessment relative to the local error rate as estimated from the data in the local vicinity of the variant. Furthermore, variants are assessed for strand imbalance significance and an additional assessment of significance of variants in low complexity contexts.

In contrast to the germline variant caller (section 2.7), the somatic variant caller makes no assumptions about the ploidy of a sample, and thus allows for sensitive detection of variant alleles at any, and low, frequencies.

Variant types LightSpeed Fastq to Somatic Variants reports SNPs, MNVs and InDels and replacements provided that the variants are contained within at least one paired end read.

Variant annotations Variants identified by LightSpeed Fastq to Somatic Variants are annotated with the following basic information: Chromosome, Region, Type, Reference, Allele, Reference allele, Length, Zygosity, Count, Coverage, Frequency, Forward read count, Reverse read count, Forward read coverage, Reverse read coverage, Forward/reverse balance and Genotype.

Read more about these general variant annotations here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html.

In addition, the following LightSpeed specific annotations are available:

- **p-value - global error rate** p-value from binomial test given count, coverage and an error rate of 0.005. Note that if UMIs are utilized, i.e., in the UMI step a UMI preset has been selected or a custom read structure with UMIs has been specified (see section 3.2), an error rate of 0.004 is used.
- **p-value - global error rate (phred scaled)** Log transformed **p-value - global error rate**.
- **p-value - local error rate** The minimum p-value from two individual tests: 1. A binomial test given forward count, forward coverage and a local error rate for forward reads estimated from the data. 2. A binomial test given reverse count, reverse coverage and a local error rate for reverse reads estimated from the data.
- **p-value - low complexity** p-value from binomial test given count and coverage. This p-value is only calculated for variants that are located in positions where two upstream and two downstream reference symbols are identical to the variant. For sites not living up to this criteria, a p-value of 0 is reported.
- **Strand balance score** 1 - (p-value from binomial test given forward count, count, and forward count/coverage).
- **Inferred from unaligned ends** Yes/no annotation indicating if the variant is a tandem duplication inferred from unaligned ends during detection of structural variants.
- **Subtype** Annotation indicating that an insertion is a tandem duplication. This annotation is added to tandem duplications inferred from unaligned ends during detection of structural variants, but also to insertions called by the standard variant caller that perfectly match a tandem duplication called during structural variant detection.
- **Nearby similar called variant** Annotation indicating if tandem duplications inferred from unaligned ends during structural variant detection resemble, but are not identical to an insertion called by the standard somatic variant detection.

2.9 Tumor normal variant detection

Tumor normal variant detection relies on the same method as somatic variant detection (section 2.8), but has additional steps where variants that are assessed to be significantly present in the normal reads are removed.

Variant types LightSpeed Fastq to Somatic Variants Tumor Normal reports reports SNVs, MNVs, InDels and replacements, provided that the variants are contained within at least one paired end read.

Variant annotations Variants identified by LightSpeed Fastq to Somatic Variants Tumor Normal are annotated with the following basic information: Chromosome, Region, Type, Reference, Allele, Reference allele, Length, Zygosity, Count, Coverage, Frequency, Forward read count, Reverse read count, Forward read coverage, Reverse read coverage, Forward/reverse balance and Genotype.

Read more about these general variant annotations here: http://resources.qiagenbioinformatics.com/manuals/cloogenomicsworkbench/current/index.php?manual=Variant_tracks.html.

In addition, variants called by LightSpeed Fastq to Somatic Variants Tumor Normal are annotated with both the same information as variants called by LightSpeed Fastq to Somatic Variants (section 2.8) and information specific to the tumor normal analysis:

- **Count in normal** The allele count in the normal read mapping.
- **Coverage in normal** The coverage in the normal read mapping.
- **Frequency in normal** The allele frequency in the normal read mapping.
- **p-value - global error rate in normal** p-value from binomial test given normal count, normal coverage and an error rate of 0.005. Note that if UMIs are utilized, i.e., in the UMI step a UMI preset has been selected or a custom read structure with UMIs has been specified (see section 3.3), an error rate of 0.004 is used.
- **p-value - global error rate in normal (phred-scale)** Log transformed **p-value - global error rate in normal**.

2.10 Limitations

Data LightSpeed is developed for and has been optimized on Illumina paired-end short read sequencing data. Paired-end sequencing data from other platforms utilizing the same data structure and similar read lengths can be expected to perform equally well with LightSpeed unless the background error-rate is markedly different. Analysis of other types of sequencing reads may not result in similar processing times or variant calls of an equivalent quality. Reads that are longer than 800 base pairs cannot be processed.

Variant detection The germline variant detection algorithm in LightSpeed is based on a model expecting diploid genomes. Therefore, LightSpeed cannot be expected to accurately detect germline variants in genomes with other ploidies. In addition, alternate ploidies of sex chromosomes are not considered in the variant detection algorithm.

Somatic variant detection with LightSpeed is possible for variants down to a variant allele frequency of 0.1%. Variants below this frequency will not be considered. However, in order to ensure high accuracy in variant calling, we recommend only calling variants down to a variant allele frequency of approx. 1%.

Reference sequence LightSpeed considers all chromosomes to be linear. Hence, for read mapping, circular chromosomes are linearized with position 1 starting at the junction of the chromosome. No reads will be mapped across the junction of circular chromosomes.

UMI grouping Duplex UMI handling is not supported, as reads must originate from the same strand to be grouped by LightSpeed.

Chapter 3

Tools

The CLC LightSpeed Module contains tools which facilitate end to end NGS secondary analysis. In the following, tools designed for germline, somatic (tumor only) and tumor normal variant detection are described. Each tool runs a collection of underlying algorithms, that have been optimized for the specific application. For information about the underlying algorithms see section 2.

Contents

3.1 LightSpeed Fastq to Germline Variants	23
3.1.1 LightSpeed Fastq to Germline Variants outputs	27
3.2 LightSpeed Fastq to Somatic Variants	28
3.2.1 LightSpeed Fastq to Somatic Variants outputs	33
3.3 LightSpeed Fastq to Somatic Variants Tumor Normal	34
3.3.1 LightSpeed Fastq to Somatic Variants Tumor Normal outputs	38
3.4 Report	39

3.1 LightSpeed Fastq to Germline Variants

The **LightSpeed Fastq to Germline Variants** tool is designed to provide variant calls from raw sequencing data within a very short timeframe.

The tool can perform read trimming, mapping, deduplication, local realignment and germline variant calling. For a description of each step, see section 2.

LightSpeed Fastq to Germline Variants can only analyze one sample per analysis start. To analyze samples in batch, **LightSpeed Fastq to Germline Variants** must be included in a workflow. Template workflows for LightSpeed analysis are available (see chapter III), but it is also possible to create custom workflows. Read about workflows here <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

To run the LightSpeed tool go to:

Tools | LightSpeed  | LightSpeed Fastq to Germline Variants 

If you are connected to a CLC Server via your 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 first wizard step, specify fastq files and a reference sequence (figure 3.1):

- **Input data**

- **Reads (fastq)** Fastq files for analysis. At least two fastq files representing R1 and R2 reads must be provided.

- **References**

- **References** The reference sequence that reads will be mapped to.

- **Reference masking**

- **No masking** Reads are mapped to the full reference sequence.
- **Exclude annotated** Reads are mapped to the full reference sequence except regions specified in the masking track.
- **Include annotated only** Reads are only mapped to the regions specified in the masking track.
- **Masking track** The track specifying the masking regions.

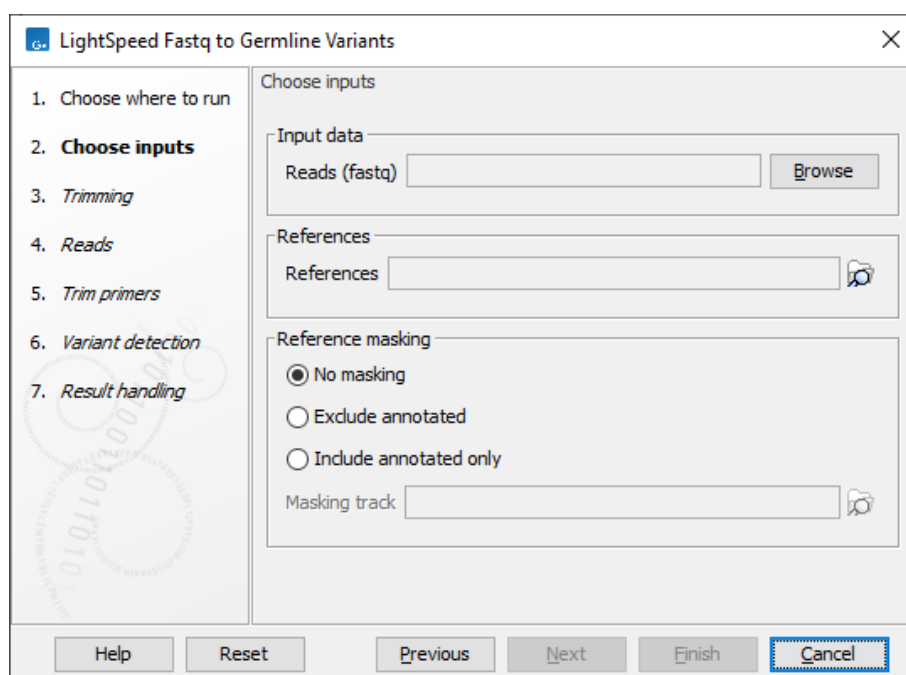


Figure 3.1: *Input fastq files and references, and, optionally, a track for reference masking.*

Next, options are available for trimming (figure 3.2):

- **Trimming**

- **Quality trim** Reads are trimmed for low quality nucleotides.

- **Minimum read length after quality trim** Trimmed reads shorter than this length are removed.
- **Adapter trim** Reads are trimmed for read-through adapter sequence.
- **Minimum read length after adapter trim** Trimmed reads shorter than this length are removed.

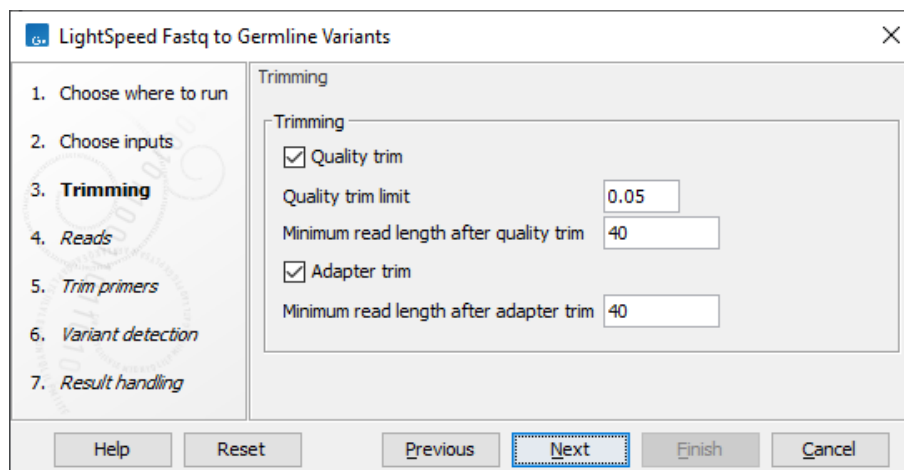


Figure 3.2: Options for trimming.

Next, options are available for UMI and duplicate reads (figure 3.3):

• UMI

- **UMI preset** Set UMI status of reads. Use the custom setting to modify all UMI settings.
- **UMI length (Read 1)** The number of nucleotides at the start of read 1 that are part of the UMI.
- **Common sequence length (Read 1)** The number of nucleotides between the UMI barcode and the fragment of the first reads from paired-end reads. These nucleotides are discarded.
- **UMI length (Read 2)** The number of nucleotides at the start of read 2 that are part of the UMI.
- **Common sequence length (Read 2)** The number of nucleotides between the UMI barcode and the fragment of the first reads from paired-end reads. These nucleotides are discarded.
- **Minimum UMI group size** Discard UMI reads created from fewer than this number of input read pairs. A UMI group is a merged read from multiple input read pairs with the same UMI barcode and mapping to the same genomic position.
- **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI barcodes nucleotides have at most this number of differences. One difference is a single nucleotide mismatch, insertion, or deletion.
- **UMI window size** Add input read pairs to the same UMI group when they map to genomic positions at most this number of bases apart.

- **Mapped read handling**

- **Discard duplicate mapped reads** Reads likely representing PCR duplicates are collapsed. This option is disabled when UMIs are used to group reads.

The screenshot shows the 'LightSpeed Fastq to Germline Variants' dialog box. On the left is a sidebar with steps: 1. Choose where to run, 2. Choose inputs, 3. Trimming, 4. **Reads**, 5. Trim primers, 6. Variant detection, 7. Result handling. The main area is titled 'Reads' and contains two sections. The 'UMI' section has a 'UMI preset' dropdown menu currently showing 'No UMI'. Below it are several input fields: 'UMI length (Read 1)' with value 0, 'Common sequence length (Read 1)' with value 0, 'UMI length (Read 2)' with value 12, 'Common sequence length (Read 2)' with value 11, 'Minimum UMI group size' with value 1, 'Maximum UMI differences' with value 1, and 'UMI window size' with value 5. The 'Mapped read handling' section below it contains a checkbox labeled 'Discard duplicate mapped reads' which is checked. At the bottom of the dialog are buttons for 'Help', 'Reset', 'Previous', 'Next' (which is highlighted with a blue border), 'Finish', and 'Cancel'.

Figure 3.3: Options for UMI and duplicate reads.

Next, options are available for primer trimming (figure 3.4):

- **Trim primers**

- **No primer trim** Disable the trim primers step.
- **Start of read 1** Primers are at the start of read 1.
- **Start of read 2** Primers are at the start of read 2.
- **Primers track** Annotation track with location and strand of primers. Unalign parts of mapped reads that overlap a primer.
- **Discard reads without primer** Discard reads that do not overlap with a primer in the primers track.
- **Additional bases to trim** Unalign this number of additional mapped bases in reads matching a primer. Bases are unaligned at the beginning of the mapped read downstream from the primer.
- **Minimum primer overlap (%)** Reads overlap a primer when the expected part of the read (start of read 1 or read 2) maps to a genomic location that overlaps at least this percentage of a primer.

Next, options are available for variant detection (figure 3.5):

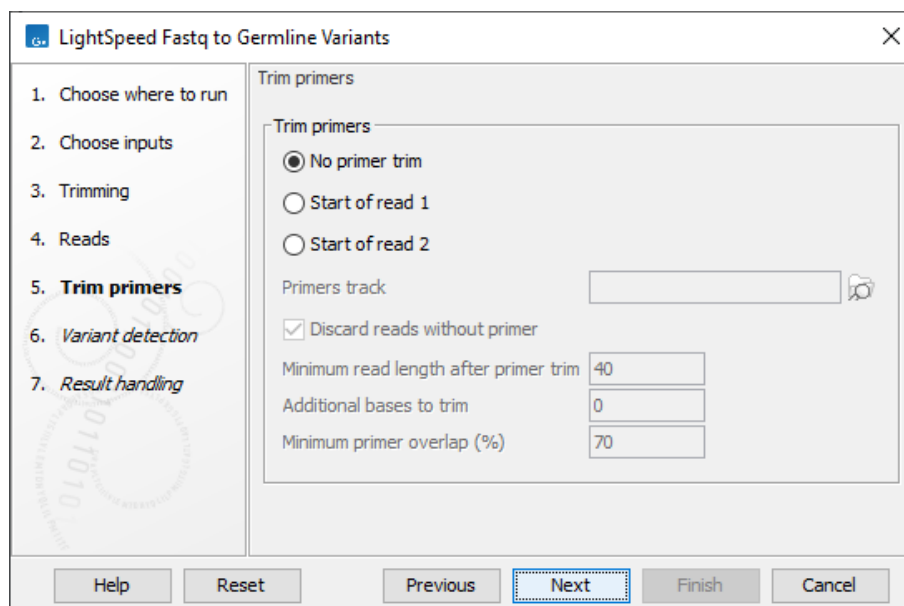


Figure 3.4: Options for primer trimming.

- **Variant detection**

- **Restrict calling to target regions** Optional. A track defining where variants are called.
- **Ignore non-specific matches** Reads that map equally well to more than one genomic position, are not used for variant calling.

- **Structural variant detection**

- **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.

In the final wizard step, choose which outputs should be generated and whether results should be saved or opened. If a reads track is selected as output, runtime will increase.

3.1.1 LightSpeed Fastq to Germline Variants outputs

LightSpeed Fastq to Germline Variants can produce the following outputs:

- **Variant track** The identified germline variants.
- **Inversions** Inversions detected from indirect evidence.
- **Report** A report providing information about each step, see section 3.4 for details.
- **Reads track** A read mapping. If a reads track is selected as output, runtime will be significantly increased.

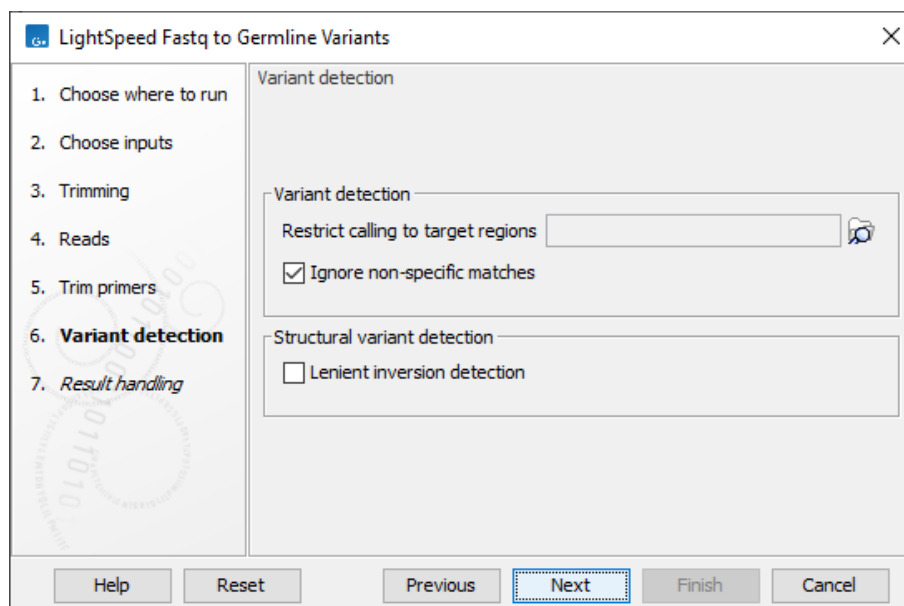


Figure 3.5: Options for variant detection.

3.2 LightSpeed Fastq to Somatic Variants

The **LightSpeed Fastq to Somatic Variants** tool is designed to provide variant calls from raw sequencing data within a very short timeframe.

The tool can perform read trimming, mapping, deduplication, local realignment and somatic variant calling. For a description of each step, see section 2.

LightSpeed Fastq to Somatic Variants can only analyze one sample per analysis start. To analyze samples in batch, **LightSpeed Fastq to Somatic Variants** must be included in a workflow. Template workflows for LigthSpeed analysis are available (see chapter III), but it is also possible to create custom workflows. Read about workflows here <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

To run the somatic LightSpeed tool go to:

Tools | LightSpeed (L) | LightSpeed Fastq to Somatic Variants (H)

If you are connected to a CLC Server via your 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 first wizard step, specify fastq files and a reference sequence (figure 3.6):

- **Input data**

- **Reads (fastq)** Fastq files for analysis. At least two fastq files representing R1 and R2 reads must be provided.

- **References**

- **References** The reference sequence that reads will be mapped to.

- **Reference masking**

- **No masking** Reads are mapped to the full reference sequence.
- **Exclude annotated** Reads are mapped to the full reference sequence except regions specified in the masking track.
- **Include annotated only** Reads are only mapped to the regions specified in the masking track.
- **Masking track** The track specifying the masking regions.

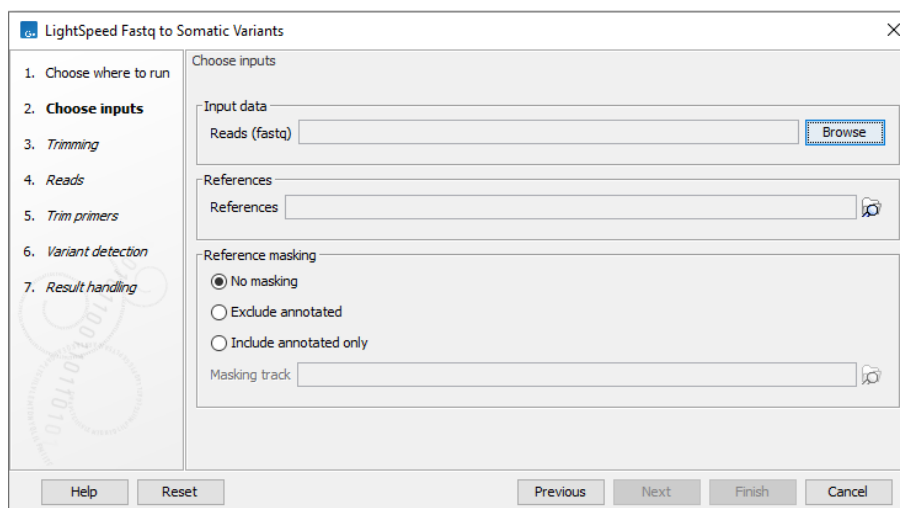


Figure 3.6: *Input fastq files and references, and, optionally, a track for reference masking.*

Next, options are available for trimming (figure 3.7):

• Trimming

- **Quality trim** Reads are trimmed for low quality nucleotides.
- **Quality trim limit** Adjust the quality trim limit for softer or harder trimming. Read more about the quality trim limit here: http://resources.qiagenbioinformatics.com/manuals/clogenomicsworkbench/current/index.php?manual=Quality_trimming.html.
- **Minimum read length after quality trim** Trimmed reads shorter than this length are removed.
- **Adapter trim** Reads are trimmed for read-through adapter sequence.
- **Minimum read length after adapter trim** Trimmed reads shorter than this length are removed.

Next, options are available for UMI and duplicate reads (figure 3.8):

• UMI

- **UMI preset** Set UMI status of reads.
 - * **No UMI** Select for reads that do not have UMIs.
 - * **Custom** Select to specify a protocol-specific read and UMI structure.

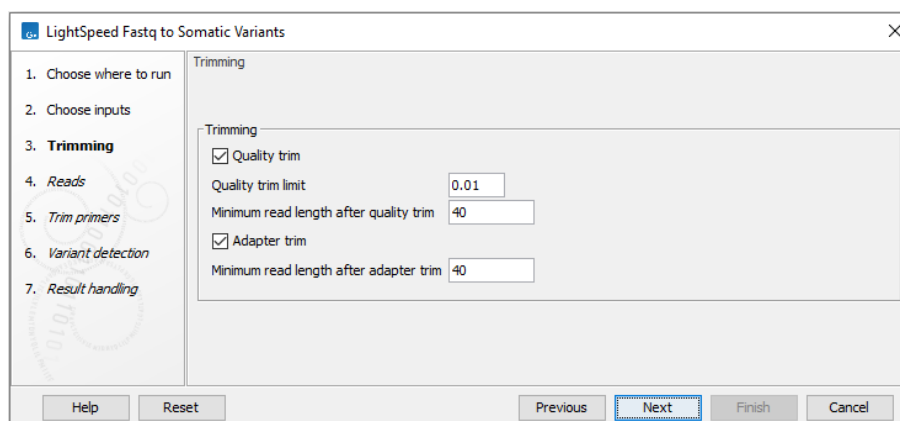


Figure 3.7: Options for trimming.

- * **QIAseq Targeted DNA** Select if you are analyzing data from a QIAseq Targeted DNA panel. The UMI and common sequence position and length are automatically adjusted to match the panel design.
- * **QIAseq Targeted DNA Pro** Select if you are analyzing data from a QIAseq Targeted DNA Pro panel. The UMI and common sequence position and length are automatically adjusted to match the panel design.
- **UMI length (Read 1)** The number of nucleotides at the start of read 1 that are part of the UMI.
- **Common sequence length (Read 1)** The number of nucleotides between the UMI and the biological sequence in read 1. These nucleotides are discarded.
- **UMI length (Read 2)** The number of nucleotides at the start of read 2 that are part of the UMI.
- **Common sequence length (Read 2)** The number of nucleotides between the UMI and the biological sequence in read 2. These nucleotides are discarded.
- **Minimum UMI group size** Only UMI groups consisting of at least this number of input read pairs will be merged to consensus UMI reads. UMI groups with fewer input read pairs than this number will be discarded. A UMI group is a group of input read pairs with the same UMI sequence that maps to the same genomic position.
- **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI sequence have at most this number of differences. One difference is a single nucleotide mismatch, insertion, or deletion.
- **UMI window size** Add input read pairs to the same UMI group when they map to genomic positions at most this number of bases apart.

- **Mapped read handling**

- **Discard duplicate mapped reads** Reads likely representing PCR duplicates are collapsed. This option is disabled when UMIs are used to group reads.

Next, options are available for primer trimming (figure 3.9):

- **Trim primers**

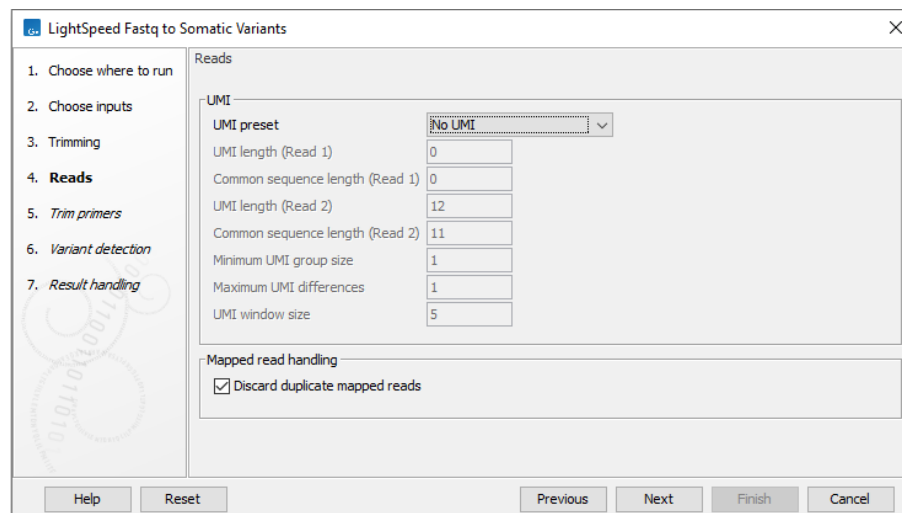


Figure 3.8: Options for UMI and duplicate reads.

- **No primer trim** Disable the trim primers step.
- **Start of read 1** Primers are at the start of read 1.
- **Start of read 2** Primers are at the start of read 2.
- **Primers track** Annotation track with location and strand of primers. Unalign parts of mapped reads that overlap a primer.
- **Discard reads without primer** Discard reads that do not overlap with a primer in the primers track.
- **Minimum read length after primer trim** Reads that are shorter than this number of nucleotides after primer trim are discarded.
- **Additional bases to trim** Unalign this number of additional mapped bases in reads matching a primer. Bases are unaligned at the beginning of the mapped read downstream from the primer.
- **Minimum primer overlap (%)** Reads overlap a primer when the expected part of the read (start of read 1 or read 2) maps to a genomic location that overlaps at least this percentage of a primer. Reads that do not meet the threshold are discarded.

A number of options are available for variant detection (figure 3.10).

• Variant detection

- **Restrict calling to target regions** Optional. A track defining where variants are called.
- **Ignore non-specific matches** Reads that map equally well to more than one genomic position, are not used for variant calling.
- **SNV minimum allele count** Minimum allele count required for SNVs and MNVs to be called.
- **SNV minimum per-strand allele count** Minimum allele count required on each strand for SNVs and MNVs to be called.
- **Indel minimum allele count** Minimum allele count required for indels to be called.

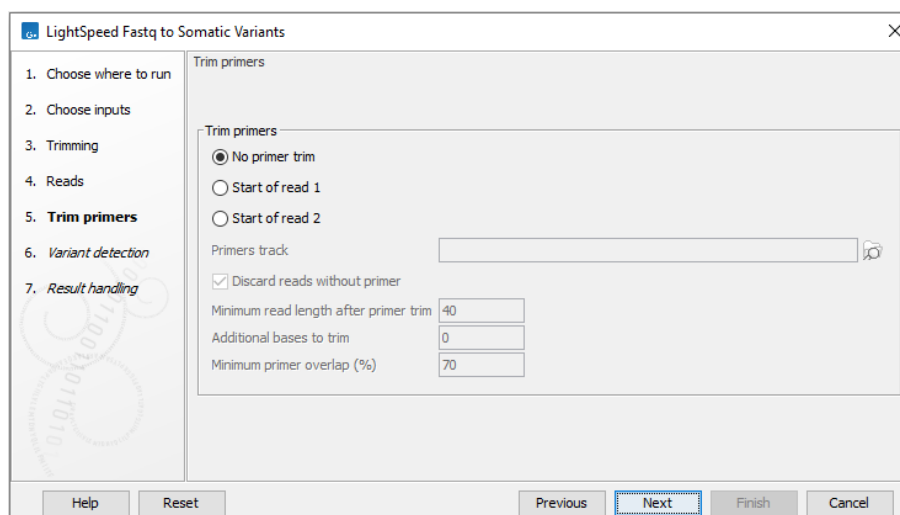


Figure 3.9: Options for primer trimming.

- **Indel minimum per-strand allele count** Minimum allele count required on each strand for indels to be called.
- **SNV significance threshold using global error rate** p-value threshold for SNVs and MNVs. The p-value is calculated from a binomial test given count, coverage and an error rate of 0.005. Allowed range: 0 - 0.1.
- **Indel significance threshold using global error rate** p-value threshold for indels. The p-value is calculated from a binomial test given count, coverage and an error rate of 0.005. Allowed range: 0 - 0.1.
- **Significance threshold using local error rate** p-value threshold for all variants. The p-value is the minimum p-value from two individual tests: 1. A binomial test given forward count, forward coverage and a local error rate for forward reads estimated from the data. 2. A binomial test given reverse count, reverse coverage and a local error rate for reverse reads estimated from the data. Allowed range: 0 - 0.1.
- **Significance threshold for low complexity regions** p-value threshold for variants that are located in positions where two upstream and two downstream reference symbols are identical to the variant. The p-value is calculated from a binomial test given count and coverage. This filter only removes alleles that have a frequency greater than 10(%). Allowed range: 0 - 1.
- **SNV strand balance threshold** Threshold for a strand balance score for SNVs and MNVs. The score is calculated as $1 - (\text{p-value from binomial test given forward count, count, and forward count/coverage})$. Allowed range: 0.9 - 1.
- **Indel strand balance threshold** Threshold for a strand balance score for indels. The score is calculated as $1 - (\text{p-value from binomial test given forward count, count, and forward count/coverage})$. Allowed range: 0.9 - 1.

For the options under **Variant detection**

- All of the options, except "SNV minimum allele count" and "Indel minimum allele count" only removes alleles with a frequency of less than 30(%).

- For all of the options that include threshold in the name, lowering the value will reduce the number of called variants.
- **Structural variant detection**
 - **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.

The screenshot shows the 'LightSpeed Fastq to Somatic Variants' wizard, specifically the 'Variant detection' step. The left sidebar lists the steps: 1. Choose where to run, 2. Choose inputs, 3. Trimming, 4. Reads, 5. Trim primers, 6. Variant detection (selected), and 7. Result handling. The main panel is titled 'Variant detection' and contains the following options:

- ☐ Restrict calling to target regions (with a file selection icon)
- ☒ Ignore non-specific matches
- SNV minimum allele count: 3
- SNV minimum per-strand allele count: 1
- Indel minimum allele count: 3
- Indel minimum per-strand allele count: 1
- SNV significance threshold using global error rate: 0.01
- Indel significance threshold using global error rate: 0.01
- Significance threshold using local error rate: 0.0025
- Significance threshold for low complexity regions: 0.01
- SNV strand balance threshold: 0.999
- Indel strand balance threshold: 0.99

Below these options is a section for 'Structural variant detection' with the option:

- ☐ Lenient inversion detection

At the bottom of the window are buttons for 'Help', 'Reset', 'Previous' (highlighted), 'Next', 'Finish', and 'Cancel'.

Figure 3.10: Options for variant detection.

In the final wizard step, choose which outputs should be generated and whether results should be saved or opened. If a reads track is selected as output, runtime will increase.

3.2.1 LightSpeed Fastq to Somatic Variants outputs

LightSpeed Fastq to Somatic Variants can produce the following outputs:

- **Variant track** The identified somatic variants.
- **Inversions** Inversions detected from indirect evidence.
- **Ignored regions** A track providing a list of regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **Report** A report providing information about each step, see section 3.4 for details.
- **Reads track** A read mapping. If a reads track is selected as output, runtime will be increased.

3.3 LightSpeed Fastq to Somatic Variants Tumor Normal

The **LightSpeed Fastq to Somatic Variants Tumor Normal** tool is designed to provide somatic variant calls from a tumor and a normal sample within a very short timeframe.

The tool can perform read trimming, mapping, deduplication, local realignment and variant calling. For a description of each step, see section 2.

LightSpeed Fastq to Somatic Variants Tumor Normal can only analyze one sample per analysis start.

To run the tumor normal LightSpeed tool go to:

Tools | LightSpeed  | LightSpeed Fastq to Somatic Variants Tumor Normal 

If you are connected to a CLC Server via your 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 first wizard step, specify tumor and normal fastq files and a reference sequence (figure 3.11):

- **Input data**

- **Tumor reads (fastq)** Tumor fastq files for analysis. At least two fastq files representing R1 and R2 reads must be provided.
- **Normal reads (fastq)** Normal fastq files for analysis. At least two fastq files representing R1 and R2 reads must be provided.

- **References**

- **References** The reference sequence that reads will be mapped to.

- **Reference masking**

- **No masking** Reads are mapped to the full reference sequence.
- **Exclude annotated** Reads are mapped to the full reference sequence except regions specified in the masking track.
- **Include annotated only** Reads are only mapped to the regions specified in the masking track.
- **Masking track** The track specifying the masking regions.

Next, options are available for trimming (figure 3.12):

- **Trimming**

- **Quality trim** Reads are trimmed for low quality nucleotides.
- **Quality trim limit** Adjust the quality trim limit for softer or harder trimming. Read more about the quality trim limit here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Quality_trimming.html.
- **Minimum read length after quality trim** Trimmed reads shorter than this length are removed.

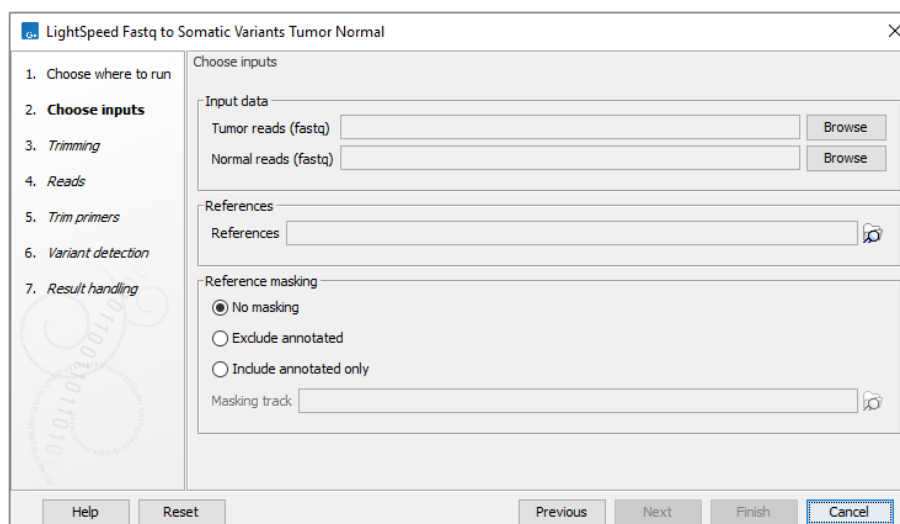


Figure 3.11: Input fastq files and references, and, optionally, a track for reference masking.

- **Adapter trim** Reads are trimmed for read-through adapter sequence.
- **Minimum read length after adapter trim** Trimmed reads shorter than this length are removed.

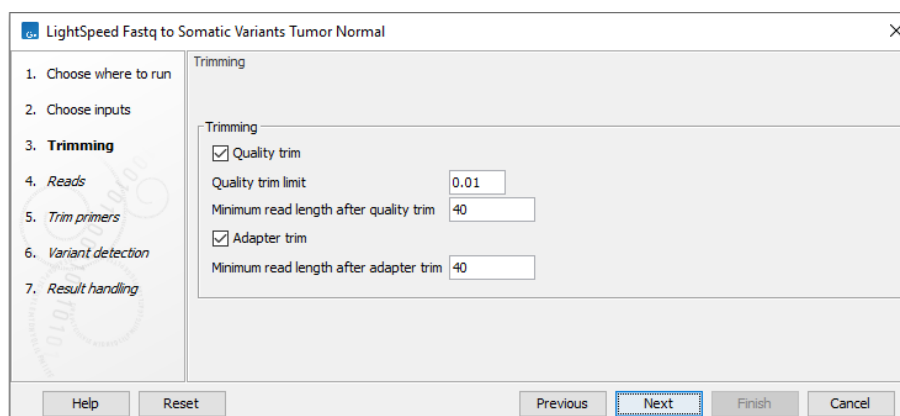


Figure 3.12: Options for trimming.

Next, options are available for UMI and duplicate reads (figure 3.13):

• UMI

- **UMI preset** Set UMI status of reads.
 - * **No UMI** Select for reads that do not have UMIs.
 - * **Custom** Select to specify a protocol-specific read and UMI structure.
 - * **QIAseq Targeted DNA** Select if you are analyzing data from a QIAseq Targeted DNA panel. The UMI and common sequence position and length are automatically adjusted to match the panel design.
 - * **QIAseq Targeted DNA Pro** Select if you are analyzing data from a QIAseq Targeted DNA Pro panel. The UMI and common sequence position and length are automatically adjusted to match the panel design.

- **UMI length (Read 1)** The number of nucleotides at the start of read 1 that are part of the UMI.
- **Common sequence length (Read 1)** The number of nucleotides between the UMI and the biological sequence in read 1. These nucleotides are discarded.
- **UMI length (Read 2)** The number of nucleotides at the start of read 2 that are part of the UMI.
- **Common sequence length (Read 2)** The number of nucleotides between the UMI and the biological sequence in read 2. These nucleotides are discarded.
- **Minimum UMI group size** Only UMI groups consisting of at least this number of input read pairs will be merged to consensus UMI reads. UMI groups with fewer input read pairs than this number will be discarded. A UMI group is a group of input read pairs with the same UMI sequence that maps to the same genomic position.
- **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI sequences have at most this number of differences. One difference is a single nucleotide mismatch, insertion, or deletion.
- **UMI window size** Add input read pairs to the same UMI group when they map to genomic positions at most this number of bases apart.

- **Mapped read handling**

- **Discard duplicate mapped reads** Reads likely representing PCR duplicates are collapsed. This option is disabled when UMIs are used to group reads.

The screenshot shows a software window titled "LightSpeed Fastq to Somatic Variants Tumor Normal". On the left is a sidebar with a list of steps: 1. Choose where to run, 2. Choose inputs, 3. Trimming, 4. **Reads**, 5. Trim primers, 6. Variant detection, and 7. Result handling. The main area is titled "Reads" and contains two sections. The first section, "UMI", has a dropdown menu for "UMI preset" set to "No UMI". Below this are six input fields: "UMI length (Read 1)" (0), "Common sequence length (Read 1)" (0), "UMI length (Read 2)" (0), "Common sequence length (Read 2)" (0), "Minimum UMI group size" (1), "Maximum UMI differences" (1), and "UMI window size" (5). The second section, "Mapped read handling", contains a checked checkbox labeled "Discard duplicate mapped reads". At the bottom are buttons for "Help", "Reset", "Previous", "Next" (which is highlighted with a blue border), "Finish", and "Cancel".

Figure 3.13: Options for UMI and duplicate reads.

Next, options are available for primer trimming (figure 3.14):

- **Trim primers**

- **No primer trim** Disable the trim primers step.
- **Start of read 1** Primers are at the start of read 1.

- **Start of read 2** Primers are at the start of read 2.
- **Primers track** Annotation track with location and strand of primers. Unalign parts of mapped reads that overlap a primer.
- **Discard reads without primer** Discard reads that do not overlap with a primer in the primers track.
- **Minimum read length after primer trim** Reads that are shorter than this number of nucleotides after primer trim are discarded.
- **Additional bases to trim** Unalign this number of additional mapped bases in reads matching a primer. Bases are unaligned at the beginning of the mapped read downstream from the primer.
- **Minimum primer overlap (%)** Reads overlap a primer when the expected part of the read (start of read 1 or read 2) maps to a genomic location that overlaps at least this percentage of a primer. Reads that do not meet the threshold are discarded.

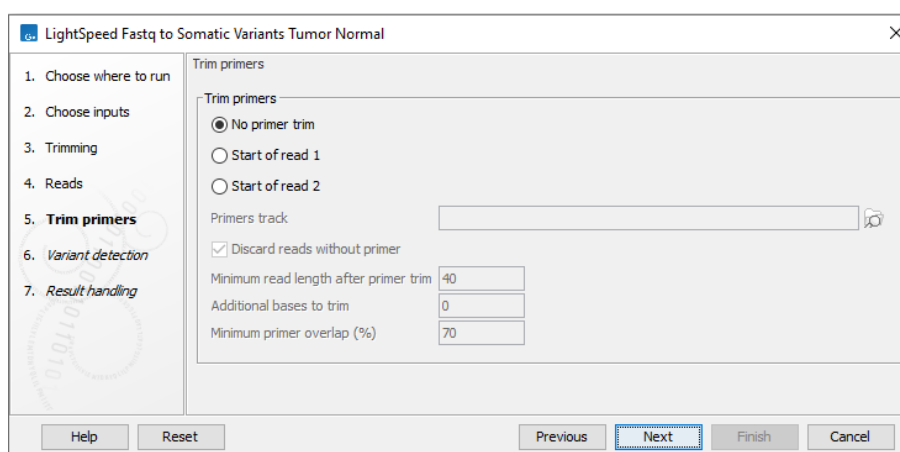


Figure 3.14: Options for primer trimming.

A number of options are available for variant detection (figure 3.15).

• Variant detection

- **Restrict calling to target regions** Optional. A track defining where variants are called.
- **Ignore non-specific matches** Reads that map equally well to more than one genomic position, are not used for variant calling.
- **SNV minimum allele count** Minimum allele count required for SNVs and MNVs to be called.
- **SNV minimum per-strand allele count** Minimum allele count required on each strand for SNVs and MNVs to be called.
- **Indel minimum allele count** Minimum allele count required for indels to be called.
- **Indel minimum per-strand allele count** Minimum allele count required on each strand for indels to be called.
- **SNV significance threshold using global error rate** p-value threshold for SNVs and MNVs. The p-value is calculated from a binomial test given count, coverage and an error rate of 0.005. Allowed range: 0 - 0.1.

- **Indel significance threshold using global error rate** p-value threshold for indels. The p-value is calculated from a binomial test given count, coverage and an error rate of 0.005. Allowed range: 0 - 0.1.
- **Significance threshold using local error rate** p-value threshold for all variants. The p-value is the minimum p-value from two individual tests: 1. A binomial test given forward count, forward coverage and a local error rate for forward reads estimated from the data. 2. A binomial test given reverse count, reverse coverage and a local error rate for reverse reads estimated from the data. Allowed range: 0 - 0.1.
- **Significance threshold for low complexity regions** p-value threshold for variants that are located in positions where two upstream and two downstream reference symbols are identical to the variant. The p-value is calculated from a binomial test given count and coverage. This filter only removes alleles that have a frequency greater than 10(%). Allowed range: 0 - 1.
- **SNV strand balance threshold** Threshold for a strand balance score for SNVs and MNVs. The score is calculated as $1 - (\text{p-value from binomial test given forward count, count, and forward count/coverage})$. Allowed range: 0.9 - 1.
- **Indel strand balance threshold** Threshold for a strand balance score for indels. The score is calculated as $1 - (\text{p-value from binomial test given forward count, count, and forward count/coverage})$. Allowed range: 0.9 - 1.

- **Normal filters**

- **Maximum count in normal** Somatic variants are not reported when the variant count in the normal is equal to or higher than this threshold.
- **Maximum frequency in normal (%)** Somatic variants are not reported when the variant frequency in the normal is equal to or higher than this threshold.
- **Significance threshold using global error rate in normal** Somatic variants are not reported when the variant p-value in the normal is lower than this threshold. Allowed range: 0 - 0.1.

For the options under **Variant detection**

- All of the options, except "SNV minimum allele count" and "Indel minimum allele count" only removes alleles with a frequency of less than 30(%)
- For all of the options that include threshold in the name, lowering the value will reduce the number of called variants.

In the final wizard step, choose which outputs should be generated and whether results should be saved or opened. If a reads track is selected as output, runtime will increase.

3.3.1 LightSpeed Fastq to Somatic Variants Tumor Normal outputs

LightSpeed Fastq to Somatic Variants Tumor Normal can produce the following outputs:

- **Somatic variant track** The identified somatic variants.

LightSpeed Fastq to Somatic Variants Tumor Normal

1. Choose where to run
2. Choose inputs
3. Trimming
4. Reads
5. Trim primers
6. **Variant detection**
7. Result handling

Variant detection

Restrict calling to target regions

☒ Ignore non-specific matches

SNV minimum allele count

SNV minimum per-strand allele count

Indel minimum allele count

Indel minimum per-strand allele count

SNV significance threshold using global error rate

Indel significance threshold using global error rate

Significance threshold using local error rate

Significance threshold for low complexity regions

SNV strand balance threshold

Indel strand balance threshold

Normal filters

Maximum count in normal

Maximum frequency in normal (%)

Significance threshold using global error rate in normal

Help Reset Previous **Next** Finish Cancel

Figure 3.15: Options for variant detection.

- **Ignored regions** A track providing a list of regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **Report** A report providing information about each step, see section 3.4 for details.
- **Tumor reads track** A read mapping of the tumor reads. If a reads track is selected as output, runtime will be increased.
- **Normal reads track** A read mapping of the normal reads. If a reads track is selected as output, runtime will be increased.

3.4 Report

The report from the LightSpeed tools provides information about each step that has been enabled in a given analysis. In the following, each section in the report is described.

Summary

- **Input read pairs** Total number of read pairs in the fastq files.
- **Read pairs removed by quality trimming** Trimmed read pairs, that after trimming are shorter than specified in the option "Minimum read length after quality trim" and have been removed.
- **Read pairs trimmed by quality trimming** Read pairs that have been trimmed and are longer than "Minimum read length after quality trim".

- **Read pairs removed by adapter trimming** Trimmed read pairs, that after trimming are shorter than specified in the option "Minimum read length after adapter trim" and have been removed.
- **Read pairs trimmed by adapter trimming** Read pairs that have been trimmed and are longer than "Minimum read length after adapter trim".
- **Average read length before trimming** Average length of reads in input.
- **Average read length after trimming** Average length of reads after quality trimming and adapter trimming.
- **Unmapped read pairs** Read pairs that did not map to the reference.
- **Non-specific mapped read pairs** Read pairs that have multiple equally good alignments to the reference.
- **Mapped broken read pairs** Mapped read pairs where the distance between the individual reads in the pair exceeded the expected distance for paired reads, or where only one of the reads in the pair was mapped.
- **Removed duplicated read pairs** Read pairs that were considered PCR duplicates of other reads and were removed during deduplication.
- **Realigned regions** The number of regions that have been locally realigned.
- **Final mapped read pairs incl. non-specific** The number of mapped read pairs excluding mapped broken reads and reads removed during deduplication.
- **Final mapped read pairs excl. non-specific** The number of mapped read pairs excluding mapped broken reads, reads removed during deduplication and non-specific mapped read pairs.

Quality trimming

- **Number of read pairs** Total number of read pairs in the fastq files.
- **Removed read pairs** Trimmed read pairs, that after trimming are shorter than specified in the option "Minimum read length after quality trim" and have been removed.
- **Trimmed read pairs** Read pairs that have been trimmed and are longer than "Minimum read length after quality trim".
- **Trimmed R1 reads** Trimmed R1 reads that are longer than "Minimum read length after quality trim".
- **Trimmed R2 reads** Trimmed R2 reads that are longer than "Minimum read length after quality trim".
- **Average read length before trim** Average read length of the raw reads in the fastq files.
- **Average read length after trim** Average read length after quality trimming. This read length may be longer than **Average read length before trim** because short reads can have been removed.

The plot **Read lengths of quality trimmed reads before / after trimming** shows the length and number of reads that were quality trimmed before and after trimming (figure 3.16).

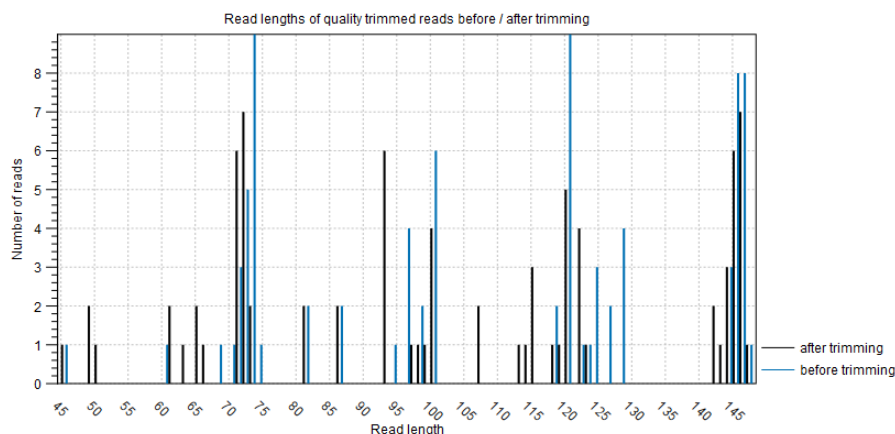


Figure 3.16: The number and length of quality trimmed reads before and after quality trimming.

Adapter trimming

- **Number of read pairs** Total number of read pairs in the fastq files.
- **Removed read pairs** Trimmed read pairs, that after trimming are shorter than specified in the option "Minimum read length after adapter trim" and have been removed.
- **Trimmed read pairs** Read pairs that have been trimmed and are longer than "Minimum read length after adapter trim".
- **Trimmed R1 reads** Trimmed R1 reads that are longer than "Minimum read length after adapter trim".
- **Trimmed R2 reads** Trimmed R2 reads that are longer than "Minimum read length after adapter trim".
- **Average read length before trim** Average length of the reads before adapter trimming. If quality trimming was enabled, read length after quality trim is given.
- **Average read length after trim** Average read length after adapter trimming. This read length may be longer than **Average read length before trim** because short reads can have been removed.
- **Detected R1 adapter** The consensus sequence of bases removed from R1 reads.
- **Detected R2 adapter** The consensus sequence of bases removed from R2 reads.

The plot **Read lengths of adapter trimmed reads before / after trimming** shows the number of reads as a function of read length before and after adapter trimming (figure 3.17).

The plot **Lengths of trimmed adapters** shows the number and lengths of trimmed adapter sequences (figure 3.18).

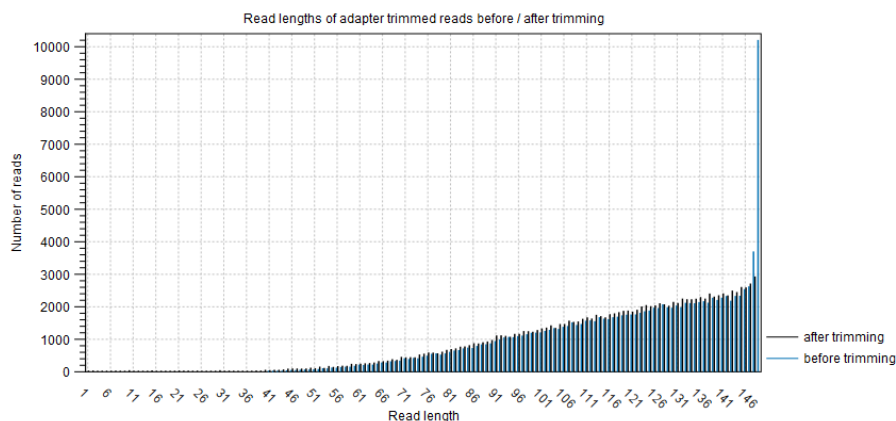


Figure 3.17: The number of reads as a function of read length before and after adapter trimming.

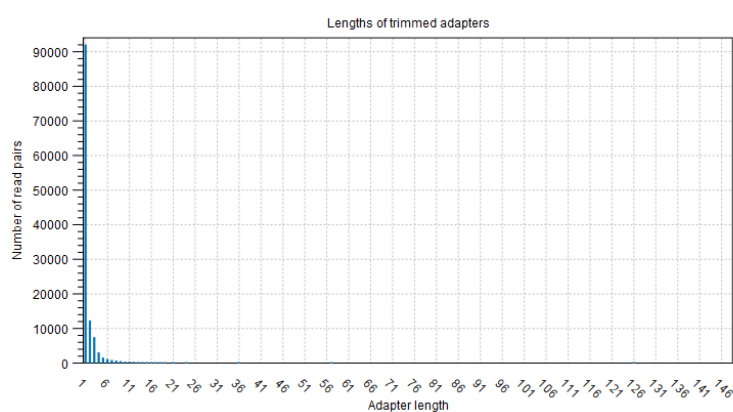


Figure 3.18: The number and length of trimmed adapter sequences.

Mapping statistics

- **References** The number of sequences in the reference genome.
- **Total read pairs** Total number of read pairs in the fastq files.
- **Read pairs attempted mapped** The number of read pairs left after trimming.
- **Mapped read pairs** The number of mapped read pairs.
- **Non-specific mapped read pairs** Read pairs that have multiple equally good alignments to the reference.
- **Mapped broken read pairs** Mapped read pairs where the distance between the individual reads in the pair exceeded the expected distance for paired reads, or where only one of the reads in the pair was mapped.
- **Mapped broken read pairs, one mapped** The number of broken read pairs where only one of the reads in the read pair was mapped.
- **Mapped broken read pairs, both mapped** The number of broken read pairs where both reads in the read pair were mapped.

- **Unmapped read pairs** Read pairs that did not map to the reference.

Part III

Template Workflows

Chapter 4

General Template Workflows

The QIAGEN CLC LightSpeed Module comes with a series of template workflows. In this chapter, template workflows that can be used to analyse WES and or WGS data, where no specialized trimming is needed, are described.

All workflows output variant tracks. Workflows that include coverage analysis also output a read mapping and a coverage report, but have longer processing time than workflows without coverage analysis.

Contents

4.1 Fastq to Annotated Germline Variants	45
4.1.1 Outputs from Fastq to Annotated Germline Variants	47
4.2 Fastq to Annotated Germline Variants with Coverage Analysis	48
4.2.1 Outputs from Fastq to Annotated Germline Variants with Coverage Analysis	51
4.3 Fastq to Annotated Somatic Variants	52
4.3.1 Outputs from Fastq to Annotated Somatic Variants	53
4.4 Fastq to Annotated Somatic Variants with Coverage Analysis	54
4.4.1 Outputs from Fastq to Annotated Somatic Variants with Coverage Analysis	56
4.5 Fastq to Annotated Somatic Variants (Tumor Normal)	58
4.5.1 Outputs from Fastq to Annotated Somatic Variants (Tumor Normal)	59
4.6 Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis	60
4.6.1 Outputs from Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis	62
4.7 Fastq to Germline CNV Control	63
4.7.1 Outputs from Fastq to Germline CNV Control	64
4.8 Fastq to Somatic CNV Control	65
4.8.1 Outputs from Fastq to Somatic CNV Control	67

4.1 Fastq to Annotated Germline Variants

The **Fastq to Annotated Germline Variants** template workflow identifies germline variants and annotates these with exon number and amino acid changes.

The workflow can be used to identify and annotate variants in both targeted sequencing and whole genome sequencing pipelines.

The workflow can be found at:

Template Workflows | LightSpeed Workflows | Fastq to Annotated Germline Variants

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 4.1). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

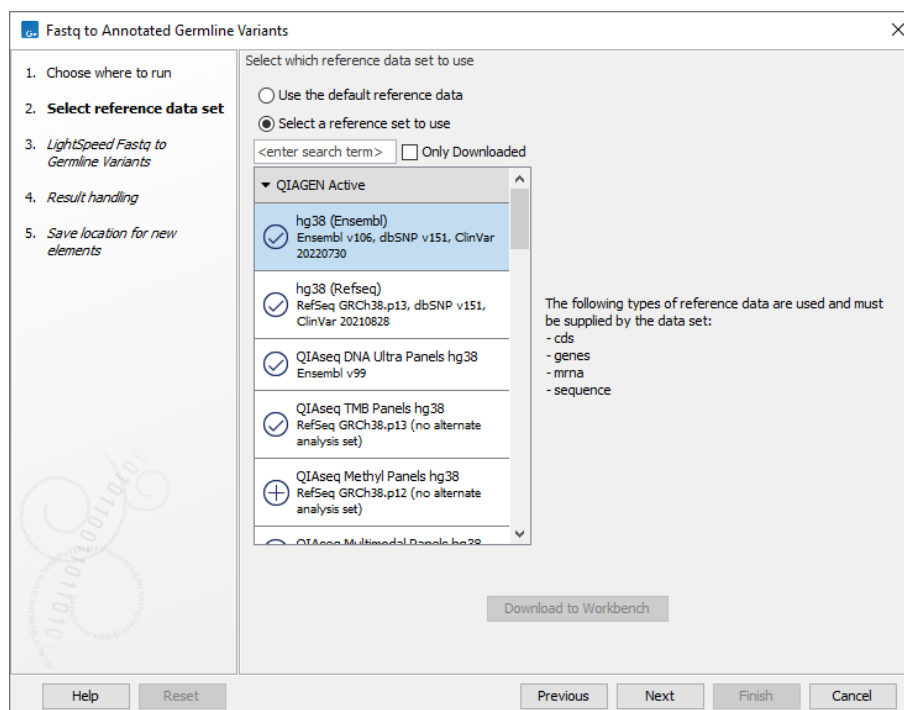


Figure 4.1: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 4.2) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.

- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Restrict calling to target regions** Optional. If a targeted protocol is used, provide target regions here.
- **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

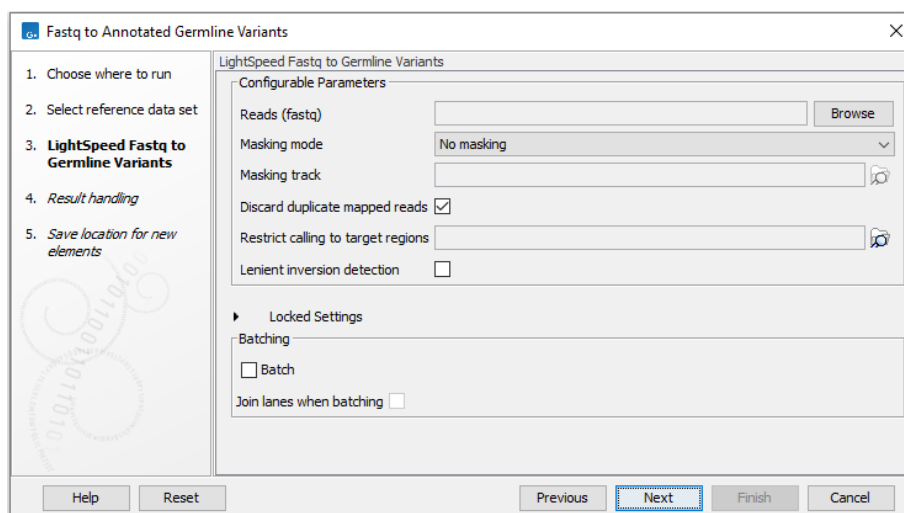




Figure 4.2: Select fastq files.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.1.1 Outputs from Fastq to Annotated Germline Variants

The **Fastq to Annotated Germline Variants** template workflow produces the following outputs:

- **Germline Variants** The variant track (🔍) with the annotated variants.
- **Inversions** An annotation track (📊) providing the called inversions.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.

- **Amino Acid Track** A track () providing a graphical representation of identified amino acid changes.
- **Genome Browser View** A track list () containing the Germline Variants, the Amino Acid Track as well as the Reference sequence and the Genes, mRNA and CDS tracks.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

4.2 Fastq to Annotated Germline Variants with Coverage Analysis

The **Fastq to Annotated Germline Variants with Coverage Analysis** template workflow:

- Identifies germline variants and annotates these with exon number and amino acid changes.
- Produces a read mapping.
- Reports coverage at target and gene level.
- Optionally identifies copy number variants (CNVs).

The workflow can only be used with targeted data.

The runtime of this workflow is significantly longer than the runtime of **Fast to Annotated Germline Variants** (section 4.1), because a read mapping track is saved.

Fastq to Annotated Germline Variants with Coverage Analysis can be found at:

Template Workflows | LightSpeed Workflows () | **Fastq to Annotated Germline Variants with Coverage Analysis** ()

If you are connected to a CLC Server via your 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 first wizard step, select the target regions (figure 4.3).

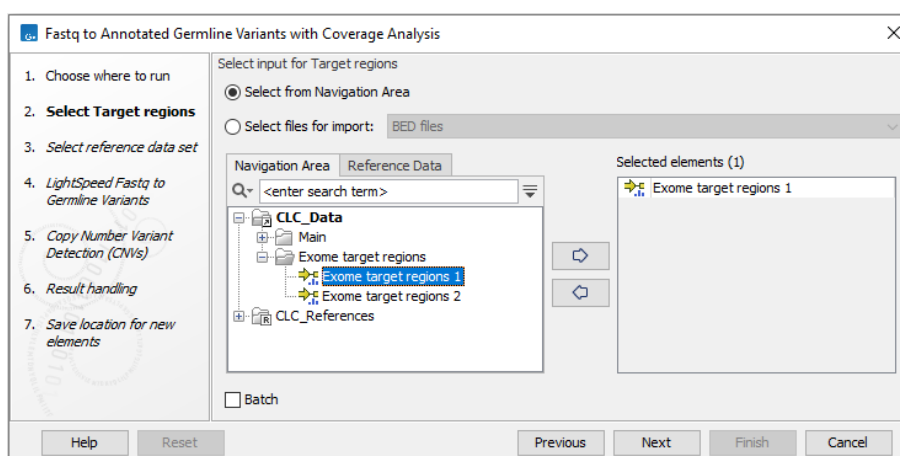


Figure 4.3: Select the target regions.

Next, select a Reference Data Set (figure 4.4). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

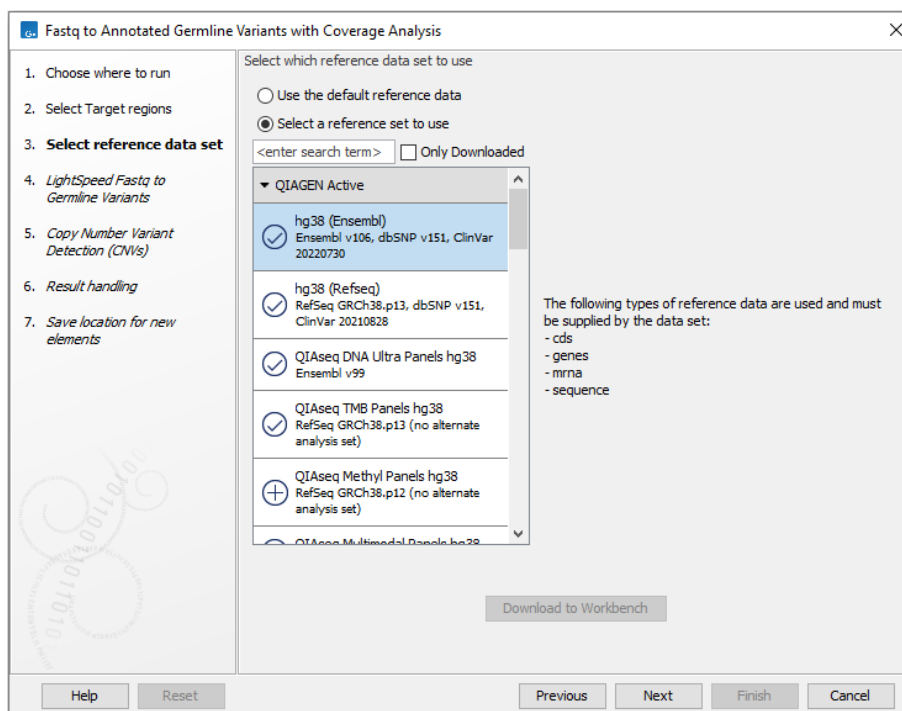


Figure 4.4: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 4.5) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.

- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

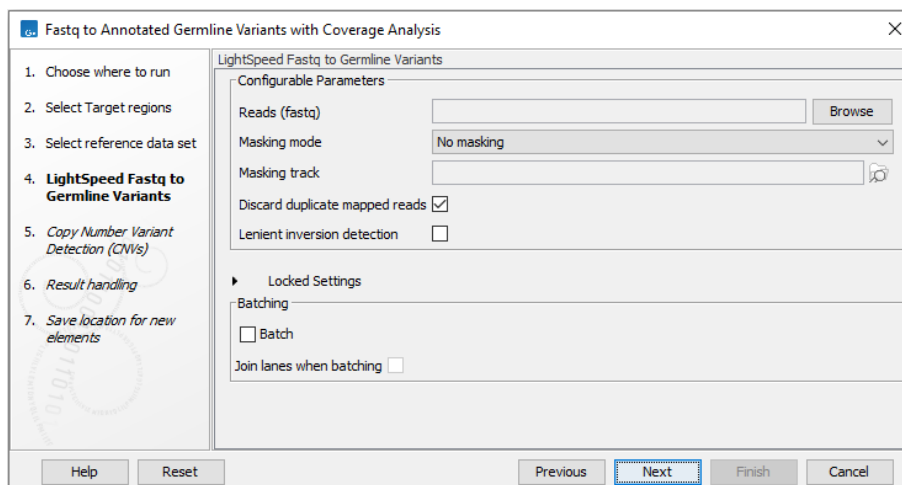


Figure 4.5: Select fastq files.

In the wizard step Copy Number Variant Detection (CNVs), it is possible to specify control coverage tables or read mappings for copy number variant detection (figure 4.6). If controls are not provided, copy number variant detection will not be performed. Read about copy number variant detection here http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

Note that for CNV detection it is important that the same processing is applied to control samples and the sample that is tested for CNVs. We recommend using the LightSpeed template workflow **Fastq to Germline CNV Control** to create appropriate control coverage tables, see section 4.7.

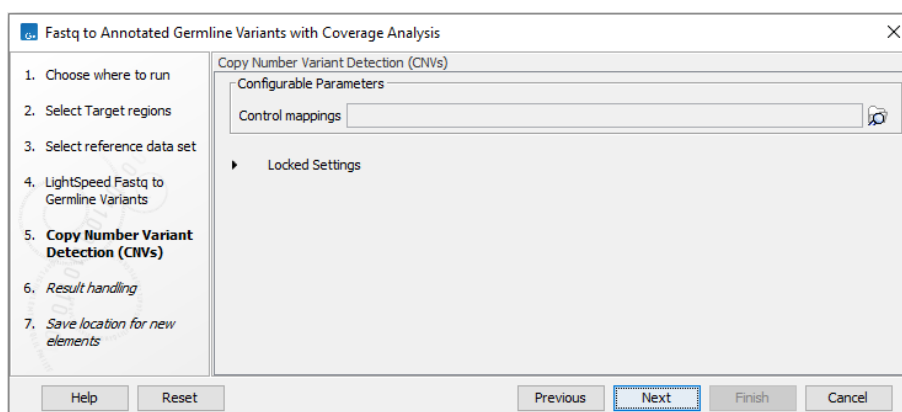
















Figure 4.6: Select control coverage tables or read mappings for copy number variant detection.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.2.1 Outputs from Fastq to Annotated Germline Variants with Coverage Analysis

The **Fastq to Annotated Germline Variants with Coverage Analysis** template workflow produces the following outputs:

- **Germline Variants** The variant track () with the annotated variants.
- **Inversions** An annotation track () providing the called inversions.
- **LightSpeed Report** A report () summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.
- **Genome Browser View** A track list () containing the Germline Variants, the Amino Acid Track, the Target regions, the Target Region Statistics Track, the Gene-level CNV Track, the Read Mapping as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Read Mapping** A read mapping track ().
- **Amino Acid Track** A track () providing a graphical representation of identified amino acid changes.
- **CNV Results Report** A report () providing an overview of identified CNVs.
- **Target-level CNV Track** An annotation track () providing CNV results per target.
- **Gene-level CNV Track** An annotation track () providing CNV results per gene.
- **Region-level CNV Track** An annotation track () providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Coverage Report** A report () summarizing coverage.
- **Target Region Statistics Track** A track () providing coverage information per target region.
- **Gene Coverage Track** A track () providing coverage information per gene.
- **Sample Report** A report () containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

The **CNV Results Report**, and the **Target, Gene and Region-level CNV Tracks** are produced by **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).

The **Coverage Report**, **Target Region Statistics Track** and the **Gene Coverage Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

4.3 Fastq to Annotated Somatic Variants

The **Fastq to Annotated Somatic Variants** template workflow identifies somatic variants and annotates these with exon number and amino acid changes.

The workflow can be used to identify and annotate variants in both targeted sequencing and whole genome sequencing pipelines.

The workflow can be found at:

Template Workflows | LightSpeed Workflows  | **Fastq to Annotated Somatic Variants** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 4.7). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

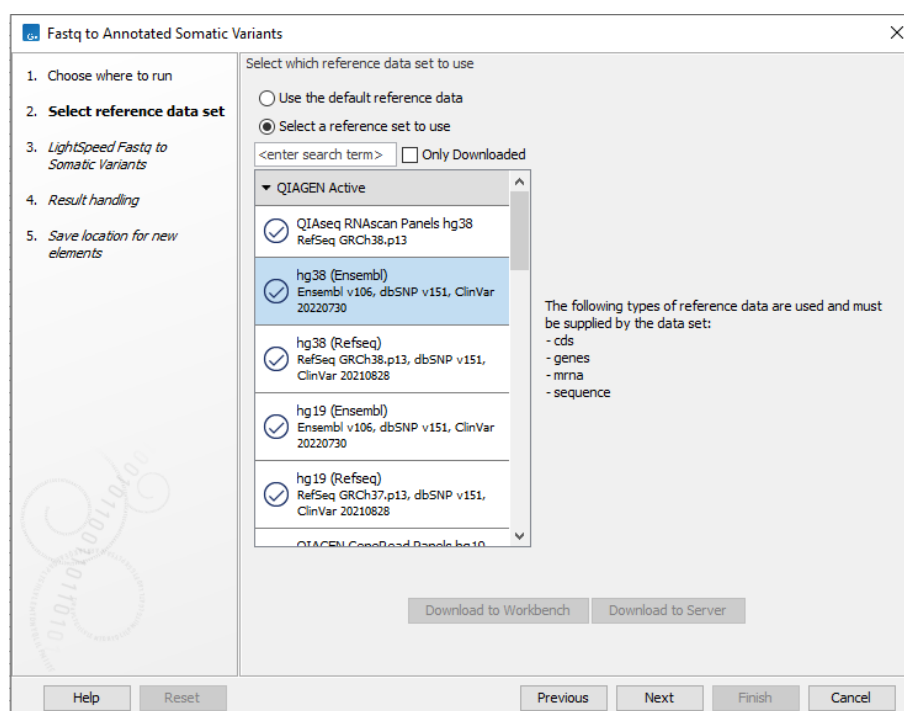


Figure 4.7: Select a reference data set.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 4.8) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.

- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Restrict calling to target regions** Optional. If a targeted protocol is used, provide target regions here.
- **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

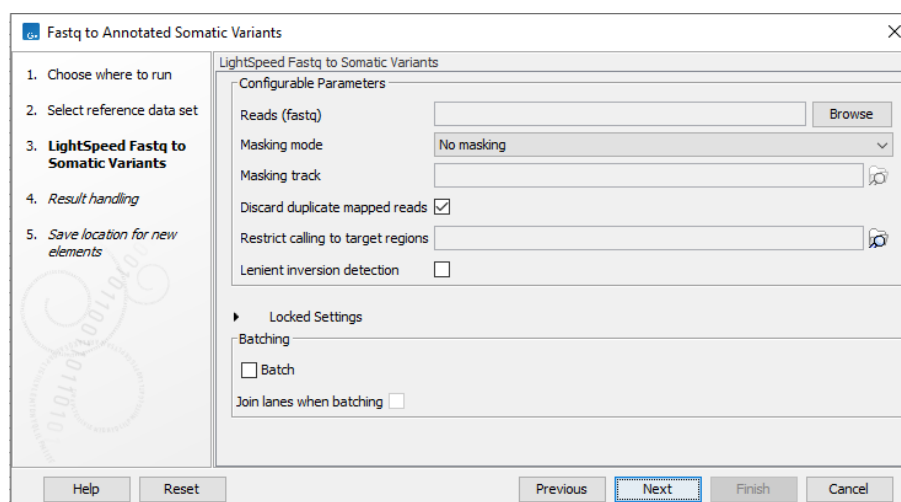






Figure 4.8: Select fastq files.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.3.1 Outputs from Fastq to Annotated Somatic Variants

The **Fastq to Annotated Somatic Variants** template workflow produces the following outputs:

- **Somatic Variants** The variant track (🔍) with the annotated variants.
- **Inversions** An annotation track (📊) providing the called inversions.

- **Ignored Regions** An annotation track () providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **LightSpeed Report** A report () summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.
- **Amino Acid Track** A track () providing a graphical representation of identified amino acid changes.
- **Genome Browser View** A track list () containing the Somatic Variants, the Ignored Regions, the Amino Acid Track as well as the Reference sequence and the Genes, mRNA and CDS tracks.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

4.4 Fastq to Annotated Somatic Variants with Coverage Analysis

The **Fastq to Annotated Somatic Variants with Coverage Analysis** template workflow:

- Identifies somatic variants and annotates these with exon number and amino acid changes.
- Produces a read mapping.
- Reports coverage at target and gene level.
- Optionally identifies copy number variants (CNVs).

The workflow can only be used with targeted data.

The runtime of this workflow is significantly longer than the runtime of **Fast to Annotated Somatic Variants** (section 4.3), because a read mapping track is saved.

Fastq to Annotated Somatic Variants with Coverage Analysis can be found at:

Template Workflows | LightSpeed Workflows () | **Fastq to Annotated Somatic Variants with Coverage Analysis** ()

If you are connected to a CLC Server via your 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 first wizard step, select the target regions (figure 4.9).

Next, select a Reference Data Set (figure 4.10). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 4.11) you have the following options:

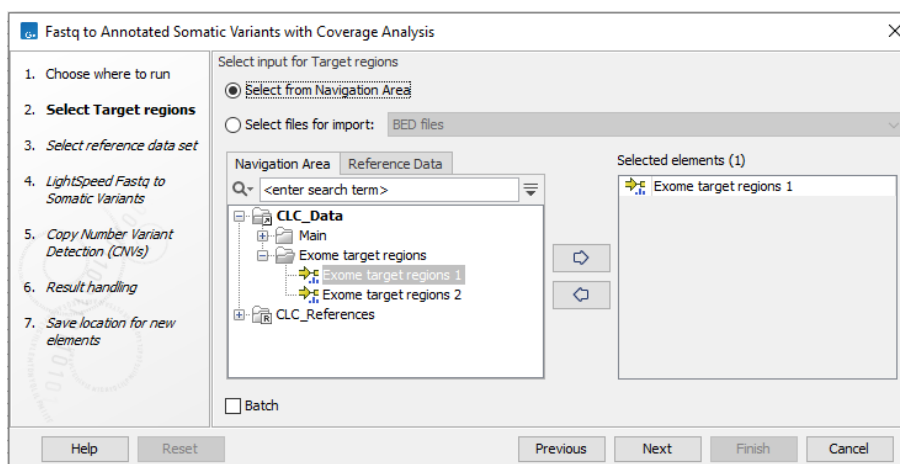


Figure 4.9: Select the target regions.

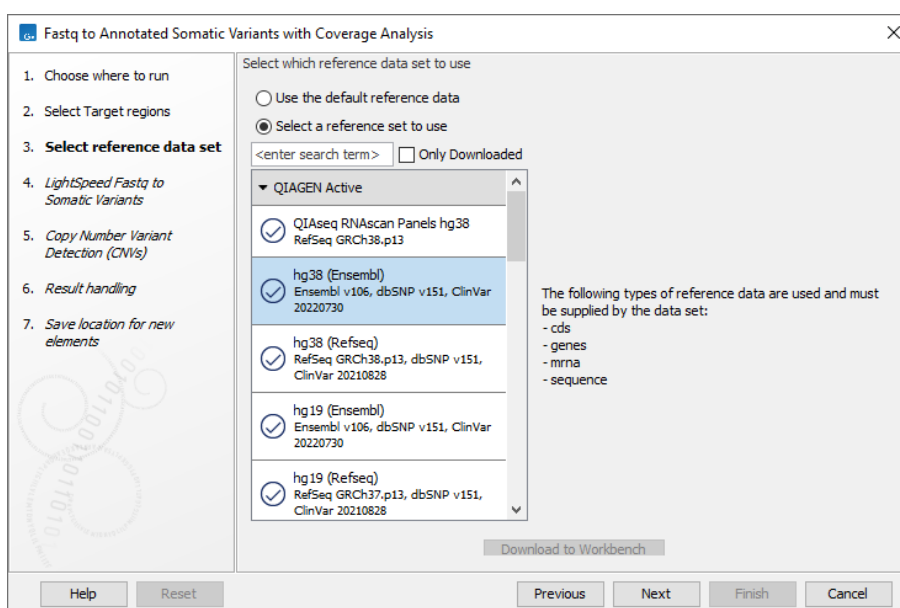


Figure 4.10: Select a reference data set.

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Lenient inversion detection** Enable lenient inversion detection to allow detection of inversions which only has read support in one direction on each of the breakpoints. This is

recommended for targeted data. Enabling this option can increase processing time and can result in detection of more false positive inversions.

- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

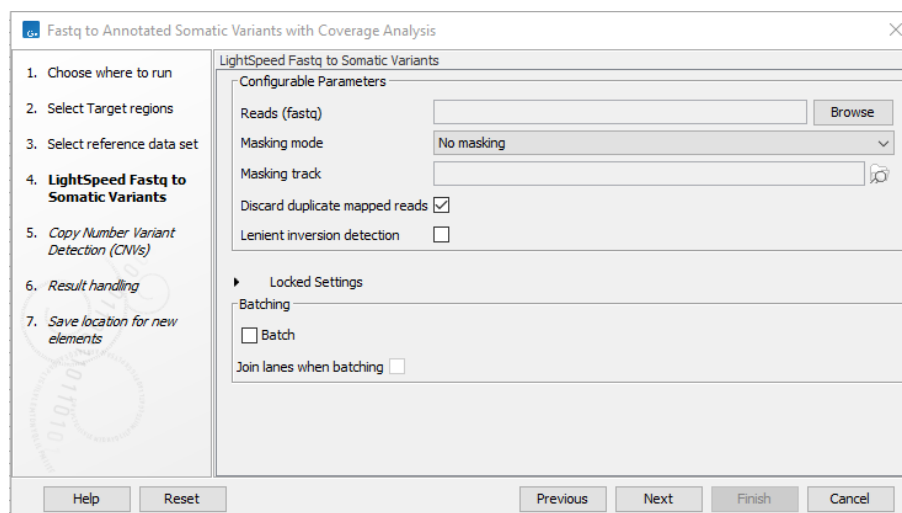


Figure 4.11: Select fastq files.

In the wizard step Copy Number Variant Detection (CNVs), it is possible to specify control coverage tables or read mappings for copy number variant detection (figure 4.12). If controls are not provided, copy number variant detection will not be performed. Read about copy number variant detection here http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

Note that for CNV detection it is important that the same processing is applied to control samples and the sample that is tested for CNVs. We recommend using the LightSpeed template workflow **Fastq to Somatic CNV Control** to create appropriate control coverage tables, see section 4.7.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.4.1 Outputs from Fastq to Annotated Somatic Variants with Coverage Analysis

The **Fastq to Annotated Somatic Variants with Coverage Analysis** template workflow produces the following outputs:

- **Somatic Variants** The variant track (🔍) with the annotated variants.
- **Inversions** An annotation track (📊) providing the called inversions.

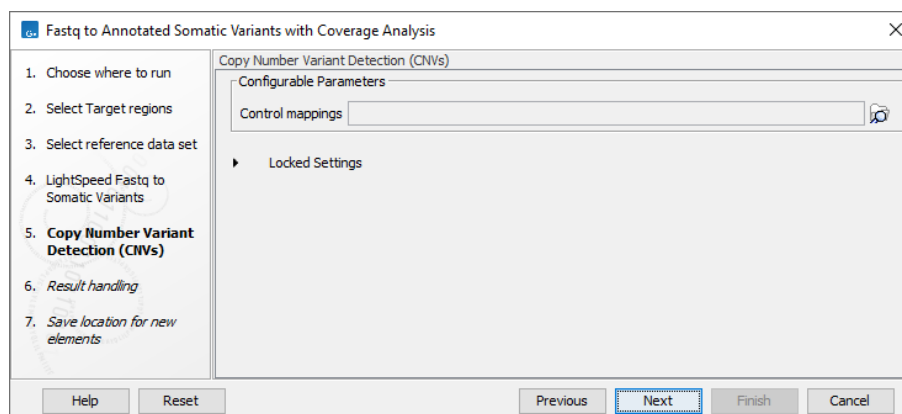


Figure 4.12: Select control coverage tables or read mappings for copy number variant detection.

- **Ignored Regions** An annotation track (📊) providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.
- **Genome Browser View** A track list (📋) containing the Somatic Variants, the Ignored Regions, the Amino Acid Track, the Target regions, the Target Region Statistics Track, the Gene-level CNV Track, the Read Mapping as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Read Mapping** A read mapping track (📊).
- **Amino Acid Track** A track (📊) providing a graphical representation of identified amino acid changes.
- **CNV Results Report** A report (📄) providing an overview of identified CNVs.
- **Target-level CNV Track** An annotation track (📊) providing CNV results per target.
- **Gene-level CNV Track** An annotation track (📊) providing CNV results per gene.
- **Region-level CNV Track** An annotation track (📊) providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Coverage Report** A report (📄) summarizing coverage.
- **Target Region Statistics Track** A track (📊) providing coverage information per target region.
- **Gene Coverage Track** A track (📊) providing coverage information per gene.
- **Sample Report** A report (📄) containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

The **CNV Results Report**, and the **Target, Gene and Region-level CNV Tracks** are produced by **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).

The **Coverage Report**, **Target Region Statistics Track** and the **Gene Coverage Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

4.5 Fastq to Annotated Somatic Variants (Tumor Normal)

The **Fastq to Annotated Somatic Variants (Tumor Normal)** template workflow identifies somatic variants from tumor normal reads and annotates these with exon number and amino acid changes.

The workflow can be used to identify and annotate variants in both targeted sequencing and whole genome sequencing pipelines.

The workflow can be found at:

Template Workflows | LightSpeed Workflows  | **Fastq to Annotated Somatic Variants (Tumor Normal)** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 4.13). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

In the LightSpeed Fastq to Annotated Somatic Variants (Tumor Normal) wizard step (figure 4.14) you have the following options:

- **Tumor reads (fastq)** Press **Browse** to select fastq files for the tumor reads.
- **Normal reads (fastq)** Press **Browse** to select fastq files for the normal reads.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Restrict calling to target regions** Optional. If a targeted protocol is used, provide target regions here.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

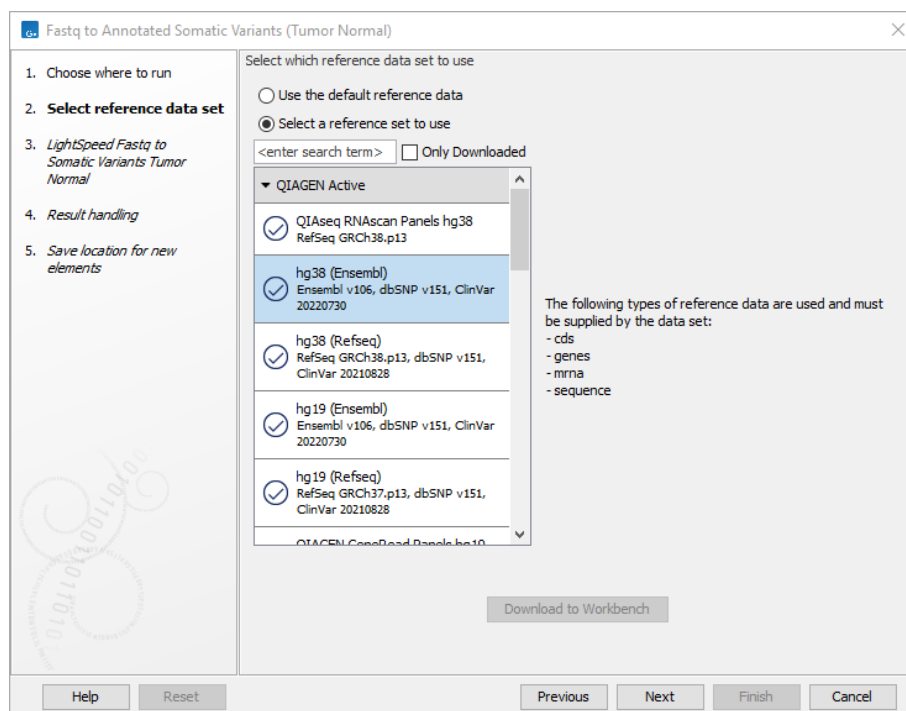


Figure 4.13: Select a reference data set.

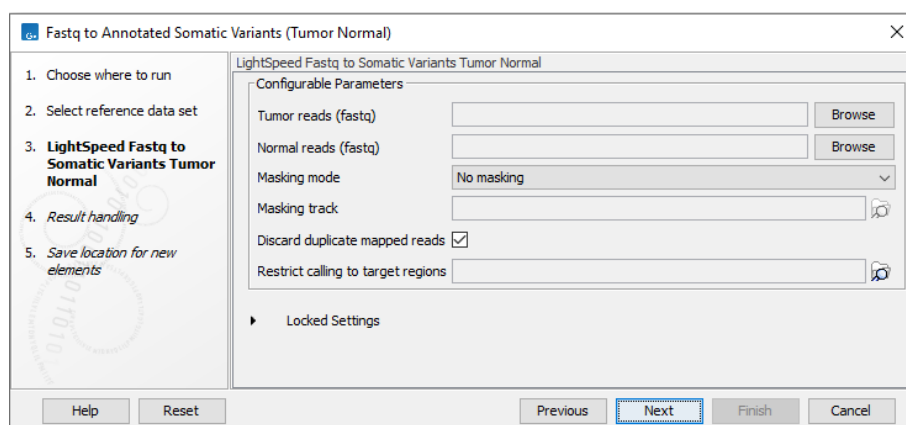




Figure 4.14: Select fastq files.

4.5.1 Outputs from Fastq to Annotated Somatic Variants (Tumor Normal)

The **Fastq to Annotated Somatic Variants (Tumor Normal)** template workflow produces the following outputs:

- **Somatic Variants** The variant track (🔍) with the annotated variants.
- **Ignored Regions** An annotation track (📌) providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.

- **Amino Acid Track** A track () providing a graphical representation of identified amino acid changes.
- **Genome Browser View** A track list () containing the Somatic Variants, the Ignored Regions, the Amino Acid Track as well as the Reference sequence and the Genes, mRNA and CDS tracks.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

4.6 Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis

The **Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis** template workflow:

- Identifies somatic variants from tumor normal reads and annotates these with exon number and amino acid changes.
- Produces a tumor read mapping and a normal read mapping.
- Reports coverage at target and gene level for tumor and normal.

The workflow can only be used with targeted data.

The runtime of this workflow is significantly longer than the runtime of **Fast to Annotated Somatic Variants (Tumor Normal)** (section 4.3), because a read mapping track is saved.

Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis can be found at:

Template Workflows | LightSpeed Workflows () | **Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis** ()

If you are connected to a CLC Server via your 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 first wizard step, select the target regions (figure 4.15).

Next, select a Reference Data Set (figure 4.16). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

In the LightSpeed Fastq to Somatic Variants (Tumor Normal) wizard step (figure 4.17) you have the following options:

- **Tumor reads (fastq)** Press **Browse** to select fastq files for the tumor reads.
- **Normal reads (fastq)** Press **Browse** to select fastq files for the normal reads.

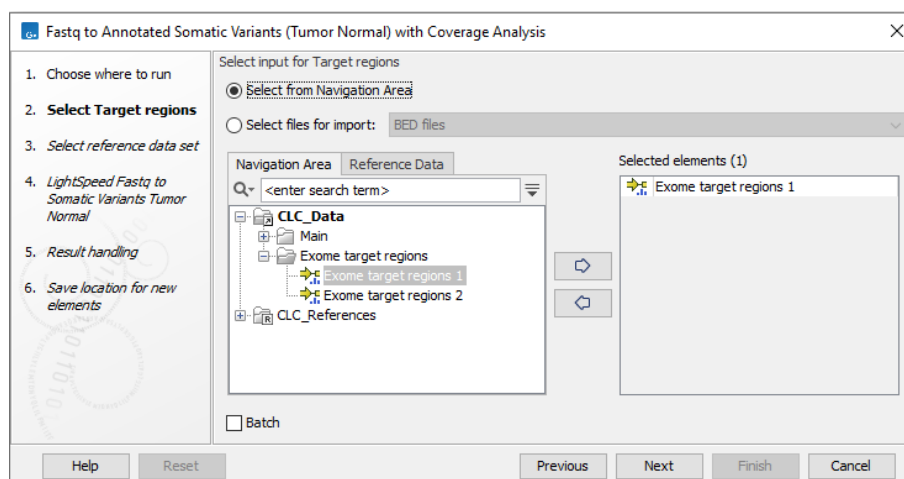


Figure 4.15: Select the target regions.

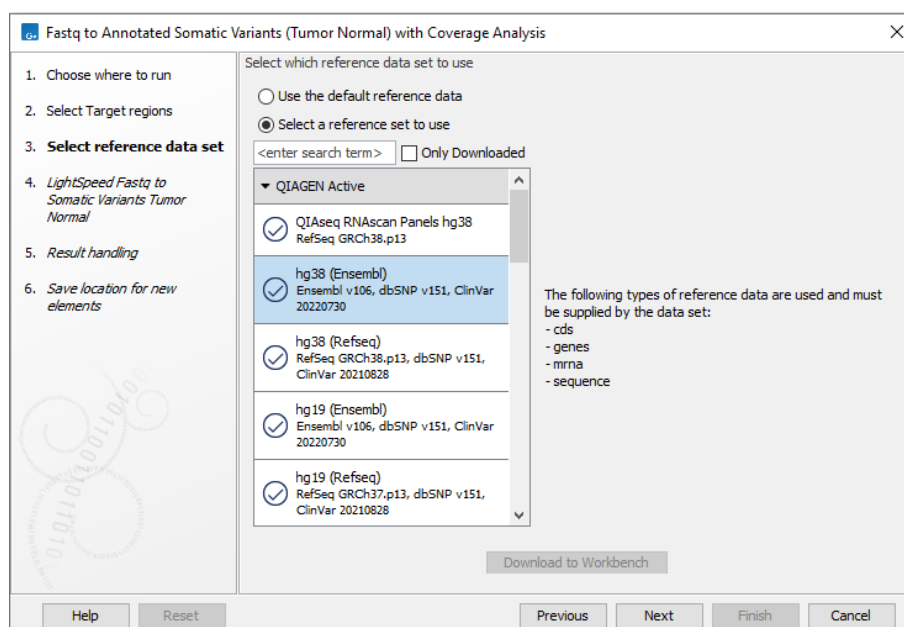


Figure 4.16: Select a reference data set.

- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

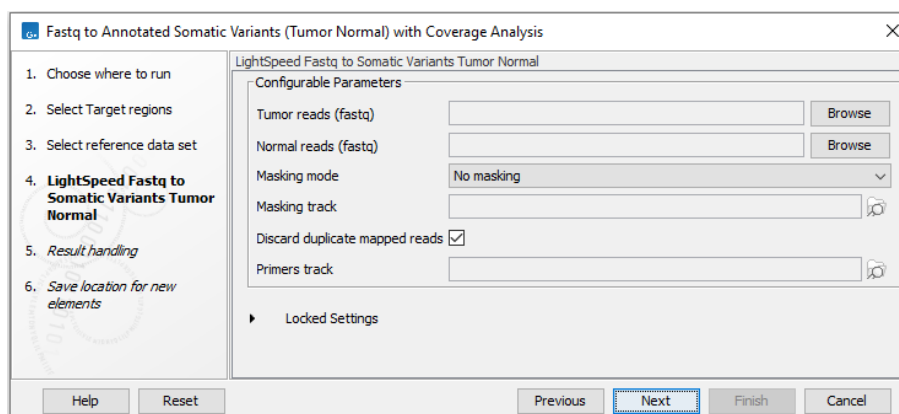


Figure 4.17: Select fastq files.

4.6.1 Outputs from Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis

The **Fastq to Annotated Somatic Variants (Tumor Normal) with Coverage Analysis** template workflow produces the following outputs:

- **Somatic Variants** The variant track (🔍) with the annotated variants.
- **Ignored Regions** An annotation track (🔍) providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants Tumor Normal tool.
- **Genome Browser View** A track list (📊) containing the Somatic Variants, the Ignored Regions, the Amino Acid Track, the Target regions, the tumor Target Region Statistics Track, the tumor and normal Read Mappings as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Read Mapping** A read mapping track (📊).
- **Amino Acid Track** A track (📊) providing a graphical representation of identified amino acid changes.
- **Coverage Report (Normal)** A report (📄) summarizing coverage.
- **Target Region Statistics Track (Normal)** A track (🔍) providing coverage information per target region.
- **Gene Coverage Track (Normal)** A track (🔍) providing coverage information per gene.
- **Coverage Report (Tumor)** A report (📄) summarizing coverage.
- **Target Region Statistics Track (Tumor)** A track (🔍) providing coverage information per target region.
- **Gene Coverage Track (Tumor)** A track (🔍) providing coverage information per gene.

The **Amino Acid Track** is produced by **Amino Acid Changes** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Amino_Acid_Changes.html).

The **Coverage Reports**, **Target Region Statistics Tracks** and the **Gene Coverage Tracks** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

4.7 Fastq to Germline CNV Control

The **Fastq to Germline CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

The workflow can only be used with targeted data.

Use the workflow to generate coverage tables for the **Fastq to Annotated Germline Variants with Coverage Analysis** template workflow (section 4.2).

Fastq to Germline CNV Control can be found at:

Template Workflows | LightSpeed Workflows  **| Fastq to Germline CNV Control**


If you are connected to a CLC Server via your 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 first wizard step, select the target regions (figure 4.18).

The target regions must be identical to the target regions that will later be used for copy number variant detection together with the control coverage tables.

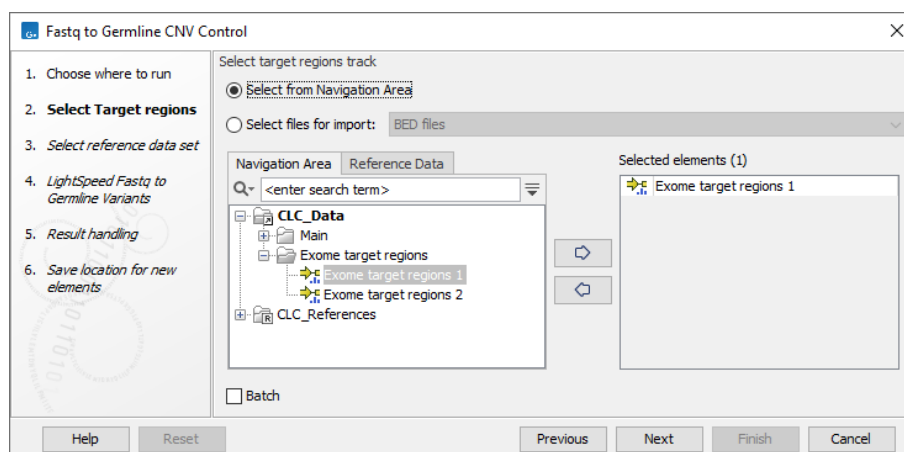


Figure 4.18: Select the target regions.

Next, select a Reference Data Set (figure 4.19). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

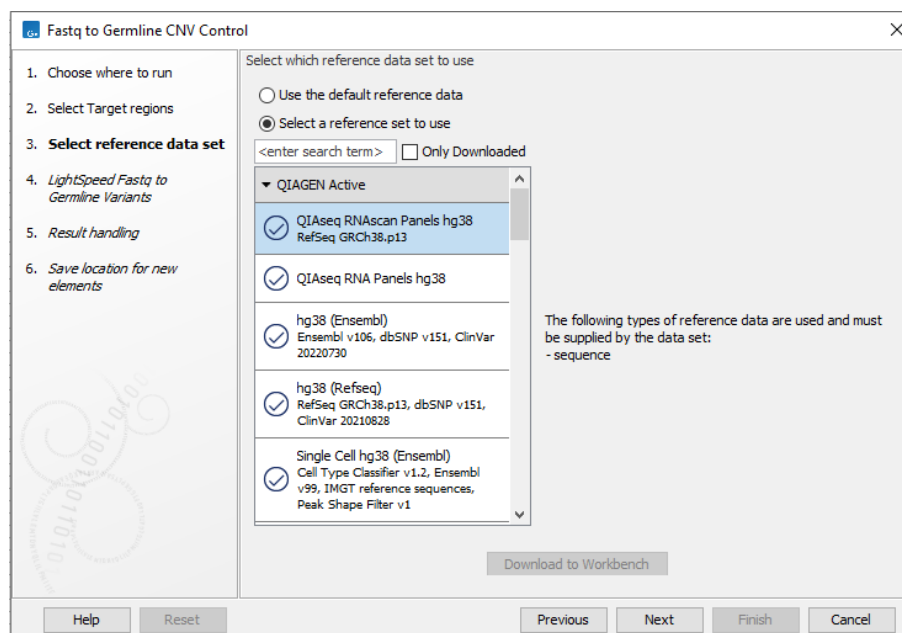


Figure 4.19: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 4.20) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.7.1 Outputs from Fastq to Germline CNV Control

The **Fastq to Germline CNV Control** template workflow produces the following outputs:

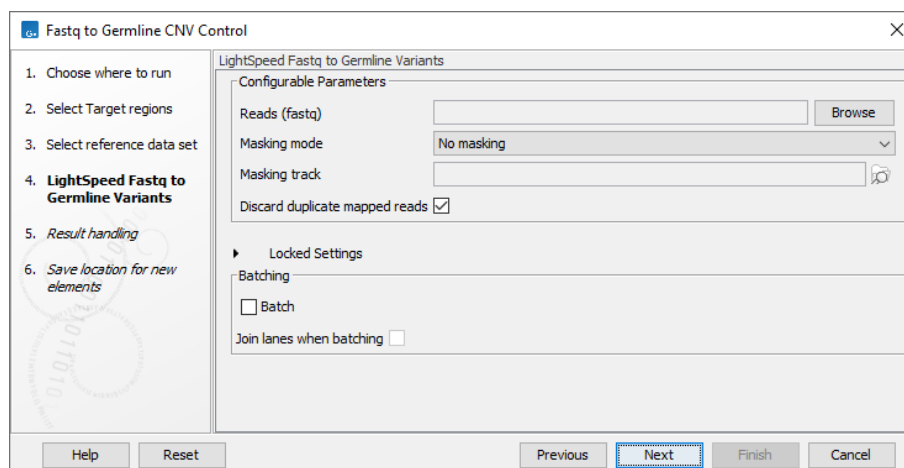






Figure 4.20: Select fastq files.

- **Coverage Table** A table  providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection, either in the **Fastq to Annotated Germline Variants with Coverage Analysis** (section 4.2) template workflow, or directly in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).
- **LightSpeed Report** A report  summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.
- **Coverage Report** A report  summarizing coverage.
- **Target Region Statistics Track** A track  providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

4.8 Fastq to Somatic CNV Control

The **Fastq to Somatic CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

The workflow can only be used with targeted data.

Use the workflow to generate coverage tables for the **Fastq to Annotated Somatic Variants with Coverage Analysis** template workflow (section 4.4).

Fastq to Somatic CNV Control can be found at:

Template Workflows | LightSpeed Workflows  | Fastq to Somatic CNV Control 

If you are connected to a CLC Server via your 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 first wizard step, select the target regions (figure 4.21).

The target regions must be identical to the target regions that will later be used for copy number variant detection together with the control coverage tables.

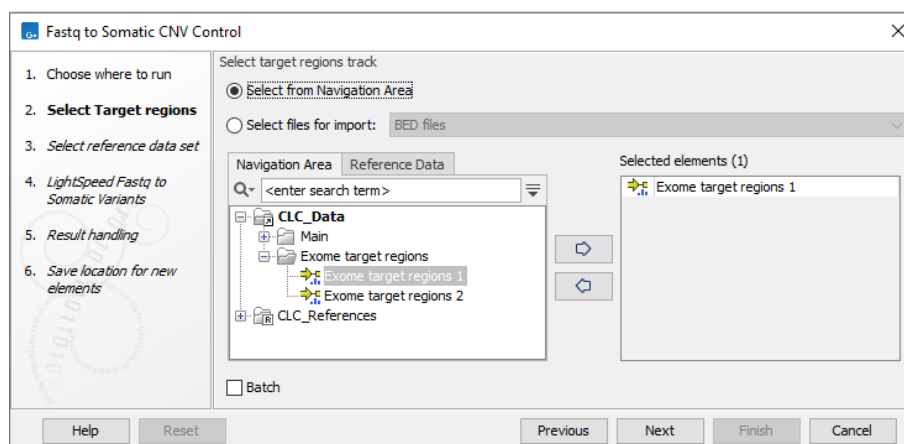


Figure 4.21: Select the target regions.

Next, select a Reference Data Set (figure 4.22). 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.

If none of the available reference data sets are appropriate, custom reference data sets can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

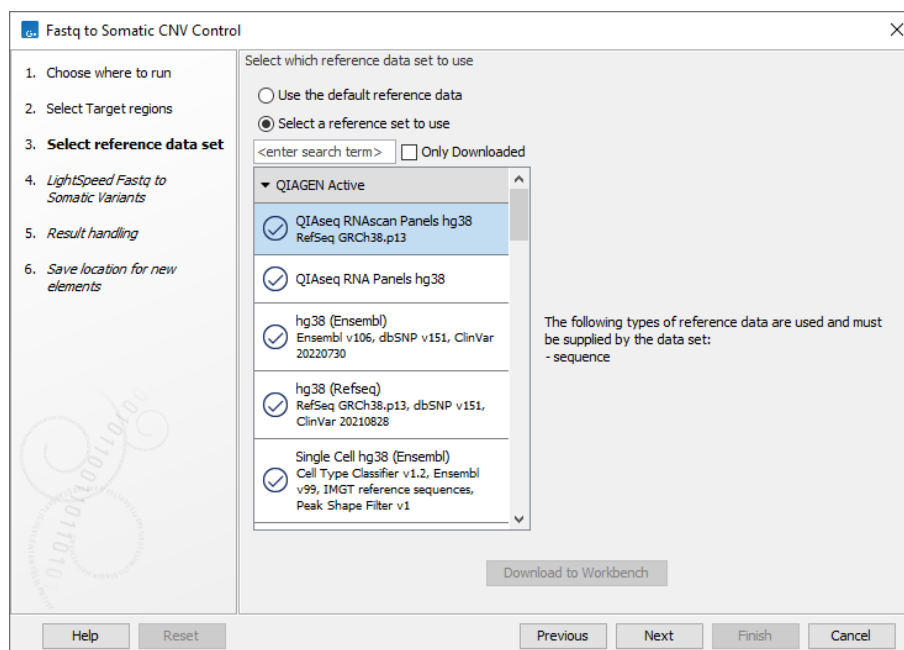


Figure 4.22: Select a reference data set.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 4.23) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Discard duplicate mapped reads** Duplicate mapped reads are per default replaced with a consensus read. Untick if duplicate mapped reads should be retained. See section 2.3 for additional details.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

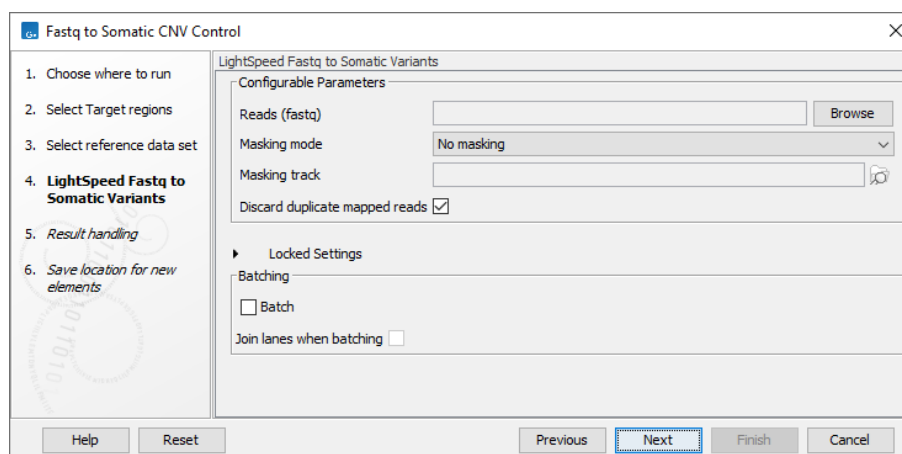




Figure 4.23: Select fastq files.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

4.8.1 Outputs from Fastq to Somatic CNV Control

The **Fastq to Somatic CNV Control** template workflow produces the following outputs:

- **Coverage Table** A table (📊) providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection, either in the template workflow **Fastq to Annotated Somatic Variants with Coverage Analysis** (section 4.4), or directly in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.

- **Coverage Report** A report () summarizing coverage.
- **Target Region Statistics Track** A track () providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

Chapter 5

Template Workflows - QIAseq Targeted DNA

The QIAGEN CLC LightSpeed Module comes with a series of template workflows facilitating both somatic and germline variant detection. The workflows are pre-configured to process reads from different protocols. In this chapter, template workflows that are specifically designed to analyse data generated with **QIAseq Targeted DNA** panels are described.

Note that the read structure differs between **QIAseq Targeted DNA** and **QIAseq Targeted DNA Pro** panels (see chapter 6). It is therefore important to choose the appropriate template workflow.

Contents

5.1 QIAseq Fastq to Annotated Germline Variants	69
5.1.1 Outputs from QIAseq Fastq to Annotated Germline Variants	72
5.2 QIAseq Fastq to Annotated Somatic Variants	73
5.2.1 Outputs from QIAseq Fastq to Annotated Somatic Variants	76
5.3 QIAseq Fastq to Germline CNV Control	77
5.3.1 Outputs from QIAseq Fastq to Germline CNV Control	79
5.4 QIAseq Fastq to Somatic CNV Control	80
5.4.1 Outputs from QIAseq Fastq to Somatic CNV Control	82

5.1 QIAseq Fastq to Annotated Germline Variants

The **QIAseq Fastq to Annotated Germline Variants** template workflow identifies germline variants from **QIAseq Targeted DNA** data and annotates these with exon number and amino acid changes. The workflow also produces a read mapping and a coverage report, and if provided with a baseline, copy number variation is also calculated.

The workflow can be found at:

Template Workflows | **LightSpeed Workflows**  | **QIAseq workflows**  | **QIAseq Targeted DNA**  | **QIAseq Fastq to Annotated Germline Variants** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 5.1).

This workflow has been set up to process data generated with QIAseq Targeted DNA panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA panels are available in the **QIAseq DNA Panels hg19** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Panels hg19** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

Reference data sets for QIAseq Targeted DNA Pro and QIAseq Targeted DNA Ultra panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

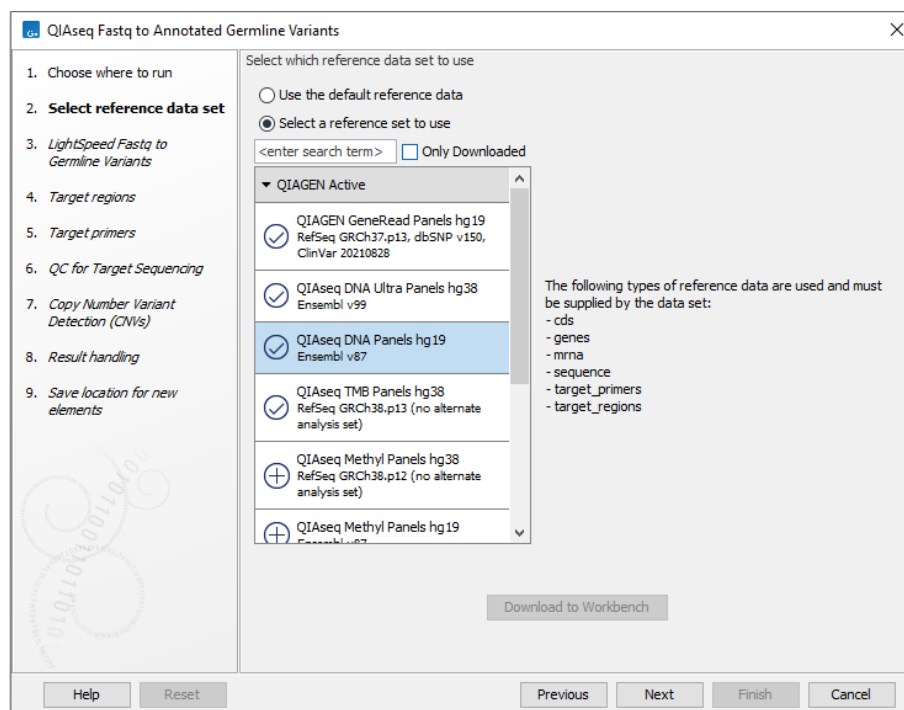


Figure 5.1: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 5.2) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.

- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

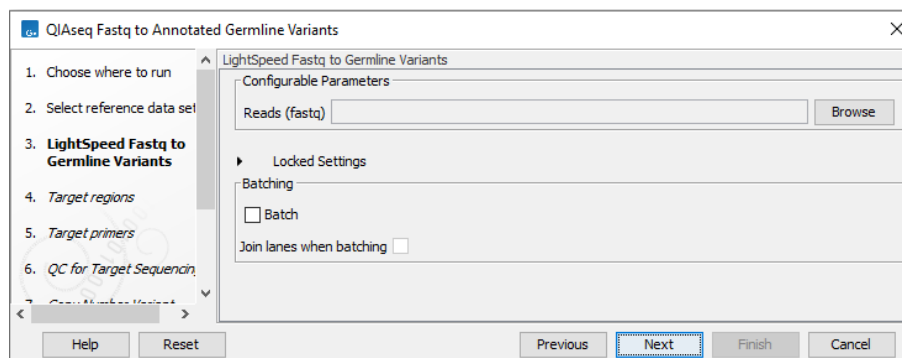


Figure 5.2: Select fastq files.

In the next dialog (figure 5.3), specify the relevant target regions from the drop down list.

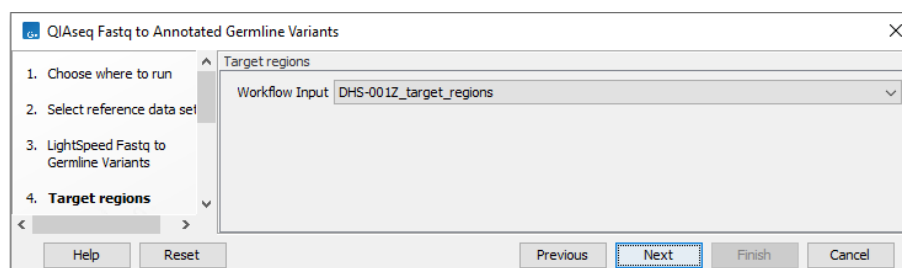


Figure 5.3: Select target regions.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 5.4).

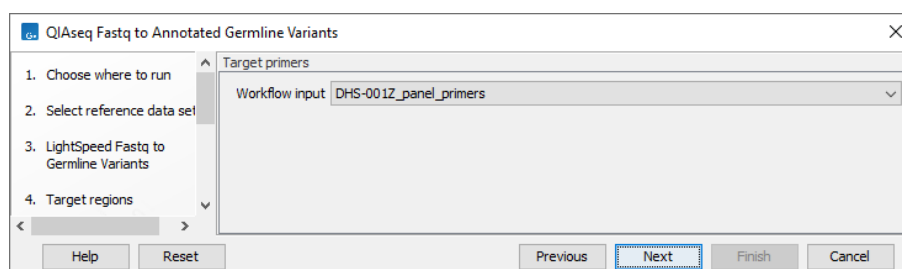


Figure 5.4: Select target primers.

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 5.5). Note that the default value for this tool depends on the application chosen (somatic or germline).

The dialog for Copy Number Variant Detection allows you to specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs (figure 5.6). 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.

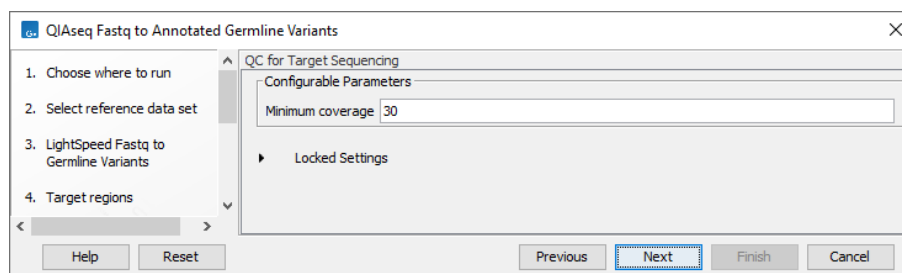


Figure 5.5: Set the Minimum coverage parameter of the QC for Target Sequencing.

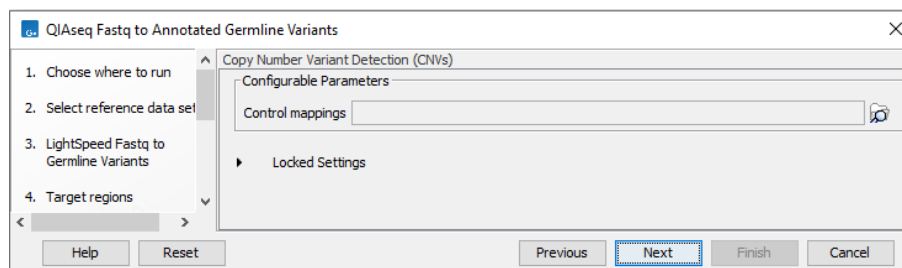


Figure 5.6: Select control coverage tables or read mappings for copy number variant detection.

Please note that if you want the copy number variation 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 workflow described in section 5.3.










A meaningful control must satisfy two conditions: (1) It must have a copy number status that it is meaningful for you to compare your sample 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. One way to achieve this is to process the control using the workflow (without providing a control mapping for the CNV detection component) and then to use the resulting UMI reads track as the control in subsequent workflow runs.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

5.1.1 Outputs from QIAseq Fastq to Annotated Germline Variants

The **QIAseq Fastq to Annotated Germline Variants** template workflow produces the following outputs:

- **Germline Variants** The variant track (📊) with the annotated variants.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.
- **Inversions** An annotation track (📊) providing the called inversions.
- **Mapped UMI Reads** A read mapping track (📊) with the mapped UMI reads.
- **Amino Acid Track** A track (📊) providing a graphical representation of identified amino acid changes.

- **Genome Browser View** A track list () containing the Variants, the Inversions, the Amino Acid Track, the Mapped UMI Reads, the Target Region Statistics Track, the Gene-level CNV Track, the Target regions as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Sample Report** A report () containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Target Region Statistics Track** A track () providing coverage information per target region.
- **Coverage Report** A report () summarizing coverage.
- **Gene Coverage Track** An annotation track () providing coverage information at the gene level.
- **Target-level CNV Track** An annotation track () providing CNV results per target.
- **Region-level CNV Track** An annotation track () providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Gene-level CNV Track** An annotation track () providing CNV results per gene.
- **CNV Results Report** A report () providing an overview of identified CNVs.

5.2 QIAseq Fastq to Annotated Somatic Variants

The **QIAseq Fastq to Annotated Somatic Variants** template workflow identifies somatic variants from **QIAseq Targeted DNA** data and annotates these with exon number and amino acid changes. The workflow also produces a read mapping and a coverage report, and if provided with a baseline, copy number variation is also calculated.

The workflow can be found at:

Template Workflows | LightSpeed Workflows () | **QIAseq workflows** () | **QIAseq Targeted DNA** () | **QIAseq Fastq to Annotated Somatic Variants** ()

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 5.7).

This workflow has been set up to process data generated with QIAseq Targeted DNA panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA panels are available in the **QIAseq DNA Panels hg19** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Panels hg19** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

Reference data sets for QIAseq Targeted DNA Pro and QIAseq Targeted DNA Ultra panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

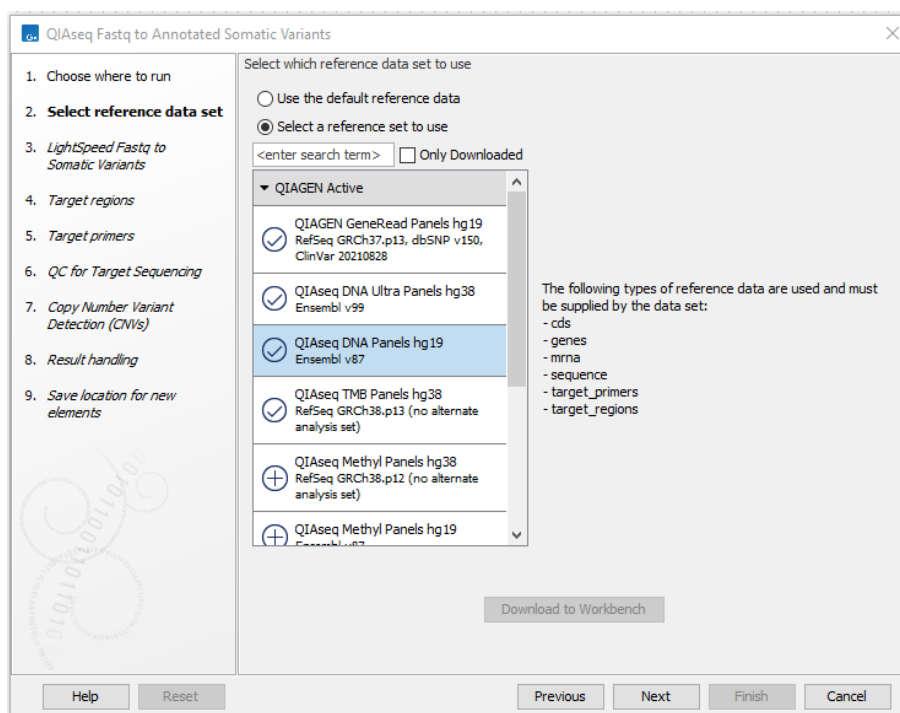


Figure 5.7: Select a reference data set.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 5.8) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

In the next dialog (figure 5.9), specify the relevant target regions from the drop down list.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 5.10).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 5.11). Note that the default value for this tool depends on the application chosen (somatic or germline).

The dialog for Copy Number Variant Detection allows you to specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs (figure 5.12). If

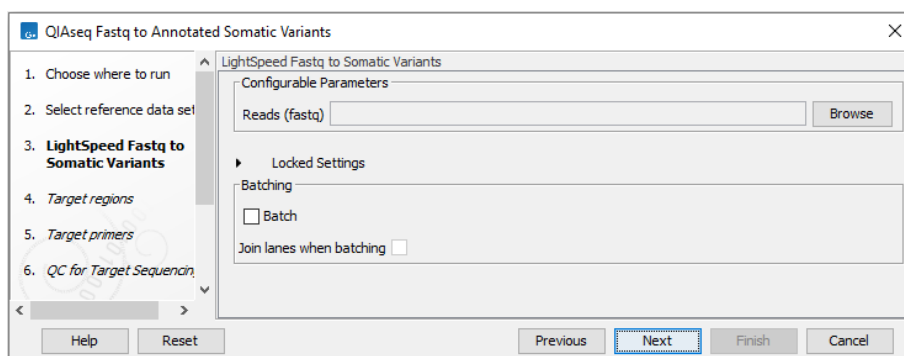


Figure 5.8: Select fastq files.

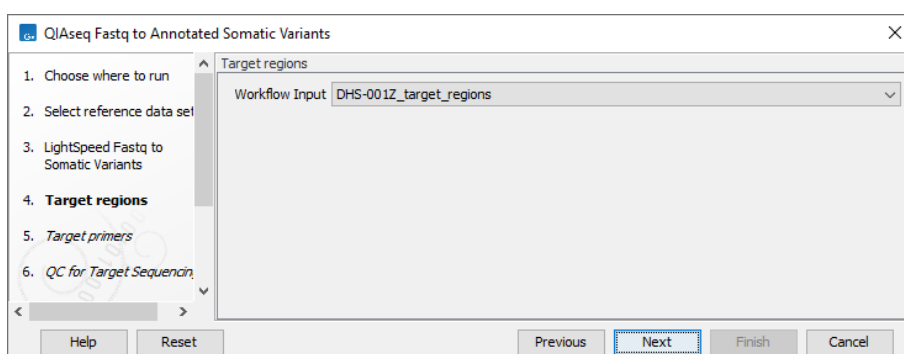


Figure 5.9: Select target regions.

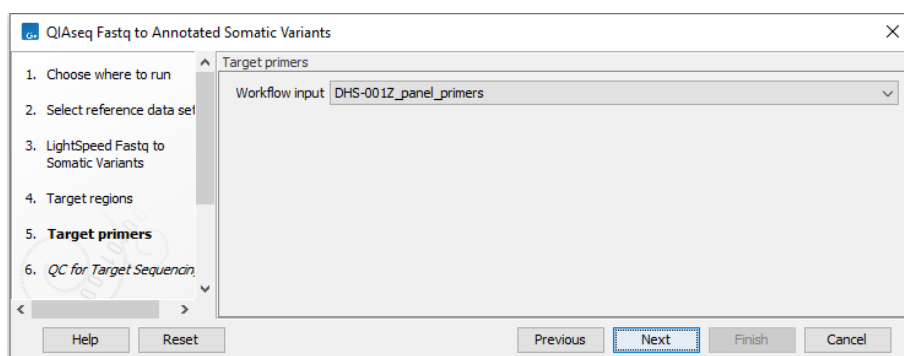


Figure 5.10: Select target primers.

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.

Please note that if you want the copy number variation 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 workflow described in section 5.4.

A meaningful control must satisfy two conditions: (1) It must have a copy number status that it is meaningful for you to compare your sample 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. One way to achieve this is to process the control using the workflow (without providing a control

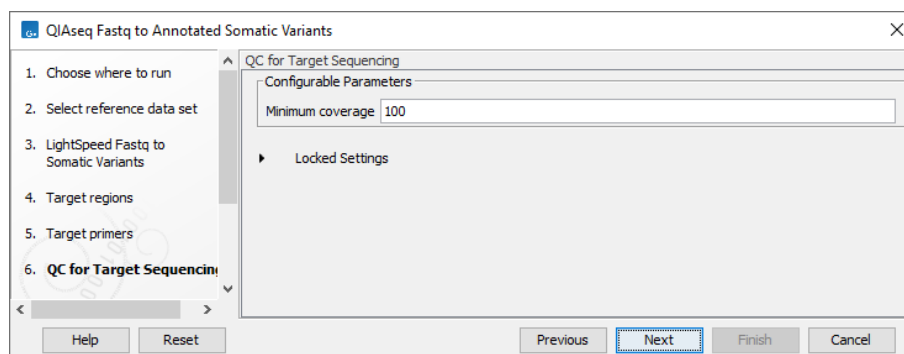


Figure 5.11: Set the Minimum coverage parameter of the QC for Target Sequencing.

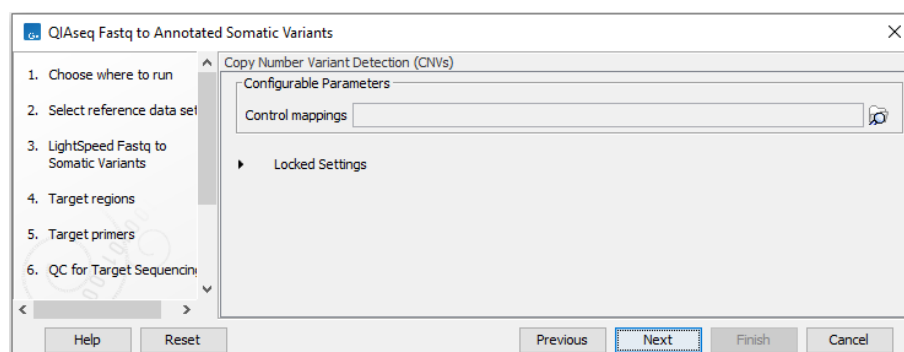


Figure 5.12: Select control coverage tables or read mappings for copy number variant detection.










mapping for the CNV detection component) and then to use the resulting UMI reads track as the control in subsequent workflow runs.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

5.2.1 Outputs from QIAseq Fastq to Annotated Somatic Variants

The **QIAseq Fastq to Annotated Somatic Variants** template workflow produces the following outputs:

- **Somatic Variants** The variant track (🔍) with the annotated variants.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.
- **Inversions** An annotation track (📊) providing the called inversions.
- **Ignored Regions** An annotation track (📊) providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **Mapped UMI Reads** A read mapping track (📊) with the mapped UMI reads.
- **Amino Acid Track** A track (📊) providing a graphical representation of identified amino acid changes.

- **Genome Browser View** A track list () containing the Variants, the Inversions, the Ignored regions, the Amino Acid Track, the Mapped UMI Reads, the Target Region Statistics Track, the Gene-level CNV Track, the Target regions as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Sample Report** A report () containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Target Region Statistics Track** A track () providing coverage information per target region.
- **Coverage Report** A report () summarizing coverage.
- **Gene Coverage Track** An annotation track () providing coverage information at the gene level.
- **Target-level CNV Track** An annotation track () providing CNV results per target.
- **Region-level CNV Track** An annotation track () providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Gene-level CNV Track** An annotation track () providing CNV results per gene.
- **CNV Results Report** A report () providing an overview of identified CNVs.

5.3 QIAseq Fastq to Germline CNV Control

The **QIAseq Fastq to Germline CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

Use the workflow to generate coverage tables for the **QIAseq Fastq to Annotated Germline Variants** (section 5.1) template workflow.

QIAseq Fastq to Germline CNV Control can be found at:

Template Workflows | LightSpeed Workflows () | **QIAseq workflows** () | **QIAseq Targeted DNA** () | **QIAseq Fastq to Germline CNV Control** ()

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 5.13).

This workflow has been set up to process data generated with QIAseq Targeted DNA panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA panels are available in the **QIAseq DNA Panels hg19** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Panels hg19** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

Reference data sets for QIAseq Targeted DNA Pro and QIAseq Targeted DNA Ultra panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

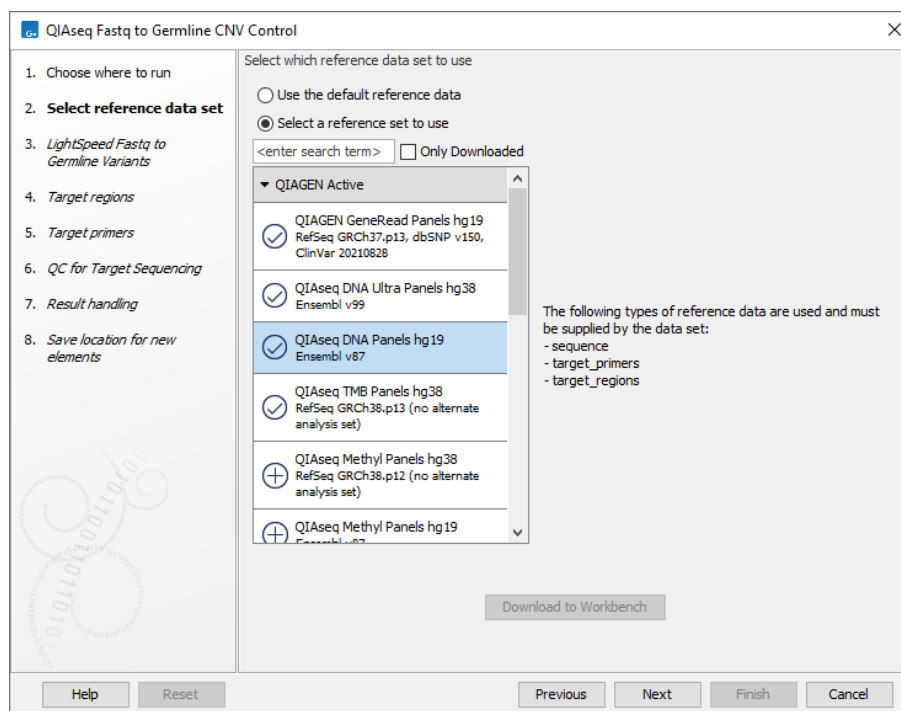


Figure 5.13: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 5.14) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

In the next dialog (figure 5.15), specify the relevant target regions from the drop down list.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 5.16).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 5.17). Note that the default value for this tool depends on the application chosen (somatic or germline).

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

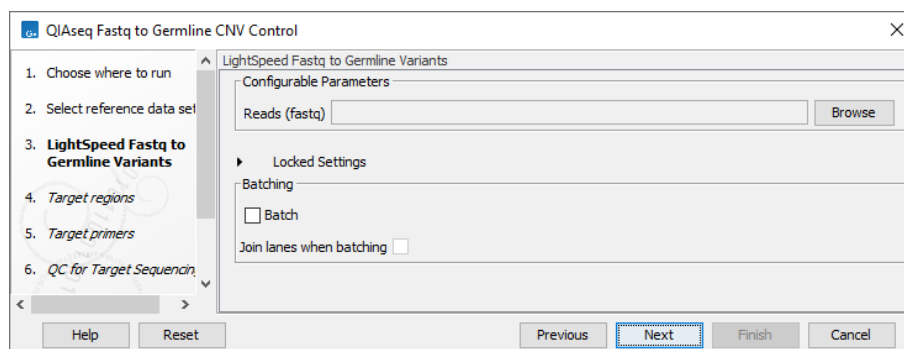


Figure 5.14: Select fastq files.

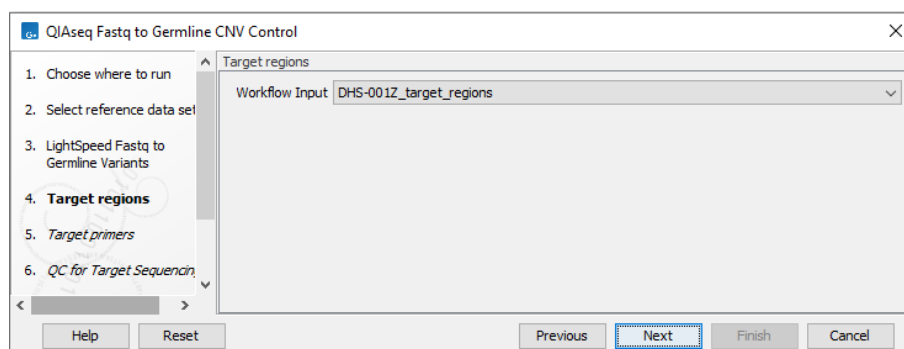


Figure 5.15: Select target regions.

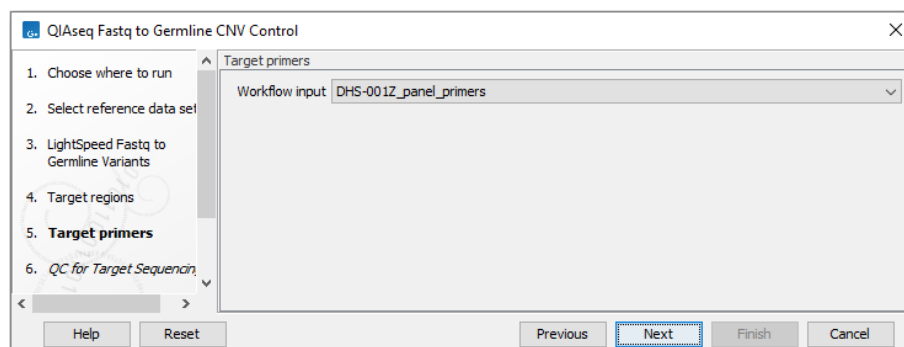


Figure 5.16: Select target primers.

5.3.1 Outputs from QIaseq Fastq to Germline CNV Control

The **QIaseq Fastq to Germline CNV Control** template workflow produces the following outputs:

- **Coverage Table** A table (📊) providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection in the **QIaseq Fastq to Annotated Germline Variants** (section 5.1) template workflow or directly in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.

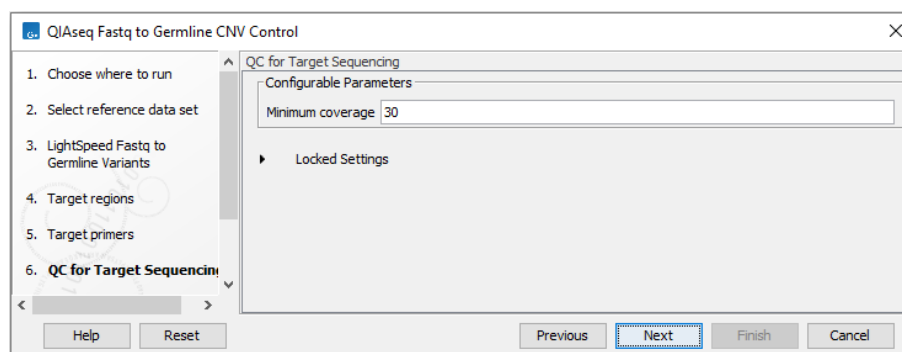




Figure 5.17: Set the *Minimum coverage* parameter of the *QC for Target Sequencing*.

- **Coverage Report** A report () summarizing coverage.
- **Target Region Statistics Track** A track () providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

5.4 QIAseq Fastq to Somatic CNV Control

The **QIAseq Fastq to Somatic CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

Use the workflow to generate coverage tables for the **QIAseq Fastq to Annotated Somatic Variants** (section 5.2) template workflow.

QIAseq Fastq to Somatic CNV Control can be found at:

Template Workflows | **LightSpeed Workflows** () | **QIAseq workflows** () | **QIAseq Targeted DNA** () | **QIAseq Fastq to Somatic CNV Control** ()

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 5.18).

This workflow has been set up to process data generated with QIAseq Targeted DNA panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA panels are available in the **QIAseq DNA Panels hg19** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Panels hg19** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

Reference data sets for QIAseq Targeted DNA Pro and QIAseq Targeted DNA Ultra panels should *not* be used with this workflow. The differences in read structure will for example prevent primers

from being correctly trimmed.

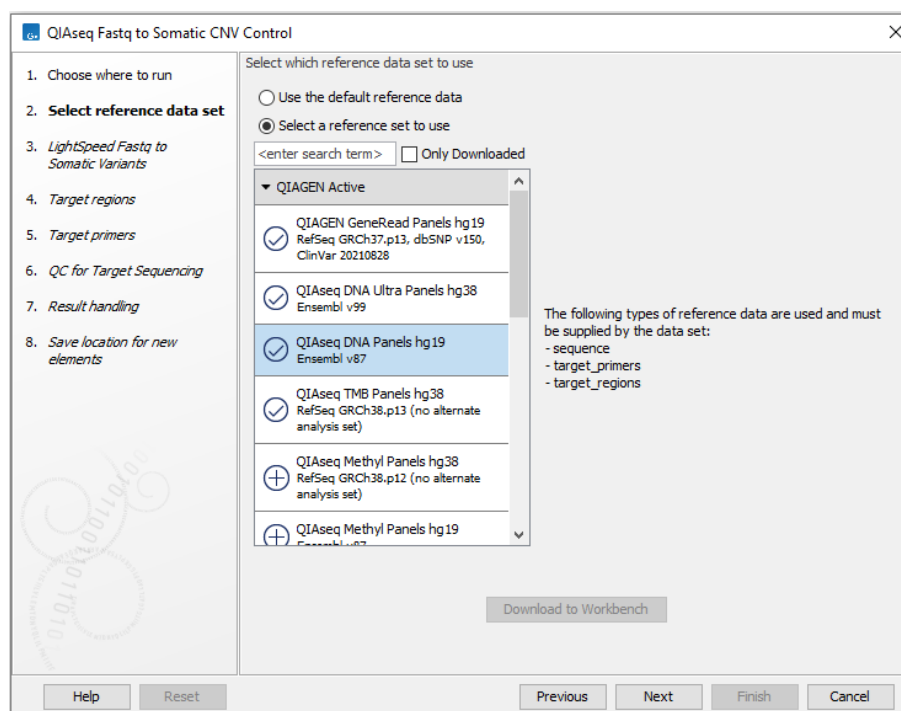


Figure 5.18: Select a reference data set.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 5.19) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

In the next dialog (figure 5.20), specify the relevant target regions from the drop down list.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 5.21).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 5.22). Note that the default value for this tool depends on the application chosen (somatic or germline).

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

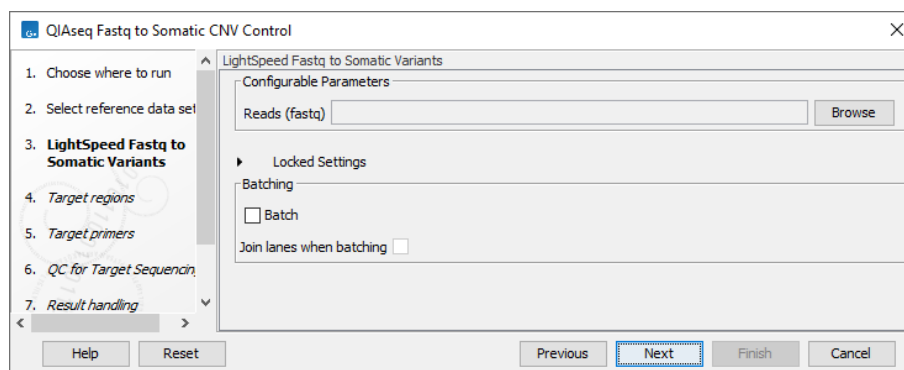


Figure 5.19: Select fastq files.

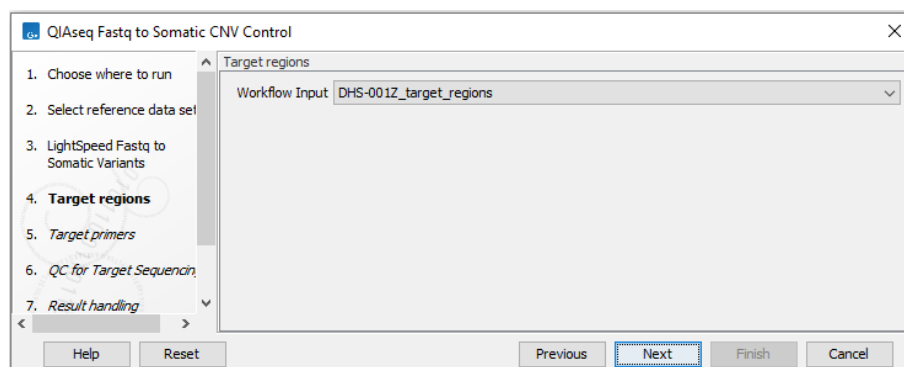


Figure 5.20: Select target regions.

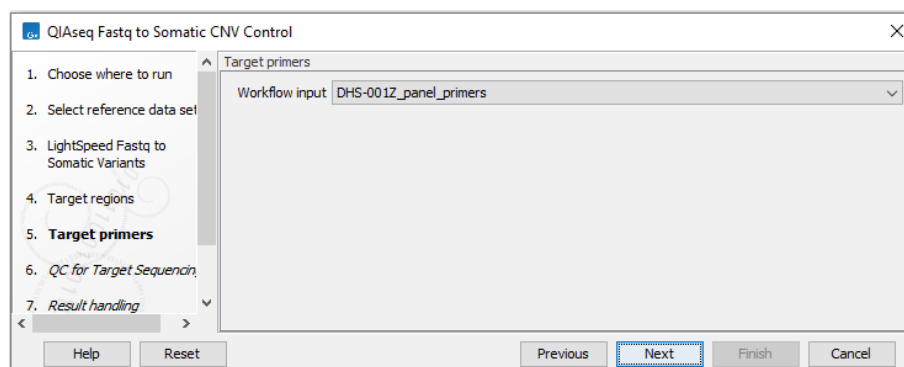


Figure 5.21: Select target primers.

5.4.1 Outputs from QIAseq Fastq to Somatic CNV Control

The **QIAseq Fastq to Somatic CNV Control** template workflow produces the following outputs:

- **Coverage Table** A table (📊) providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection in the **QIAseq Fastq to Annotated Somatic Variants** (section 5.2) template workflow or directly in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by

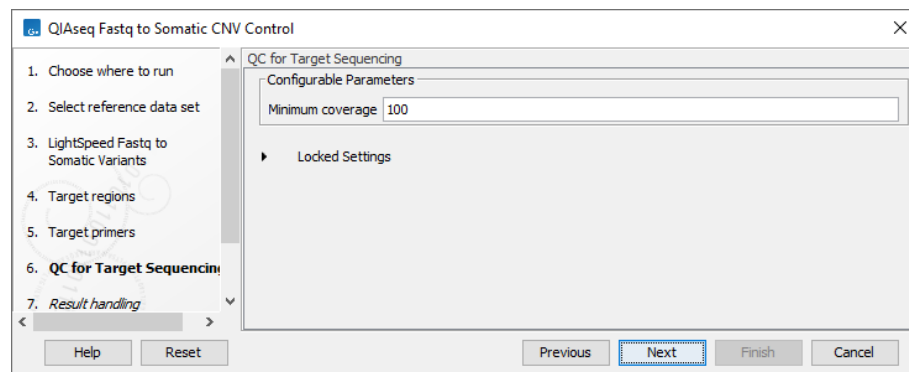




Figure 5.22: Set the *Minimum coverage* parameter of the *QC for Target Sequencing*.

the LightSpeed Fastq to Somatic Variants tool.

- **Coverage Report** A report () summarizing coverage.
- **Target Region Statistics Track** A track () providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted_Sequencing.html).

Chapter 6

Template Workflows - QIAseq Targeted DNA Pro

The QIAGEN CLC LightSpeed Module comes with a series of template workflows facilitating both somatic and germline variant detection. The workflows are pre-configured to process reads from different protocols. In this chapter, template workflows that are specifically designed to analyse data generated with **QIAseq Targeted DNA Pro** panels are described.

Note that the read structure differs between **QIAseq Targeted DNA Pro** and **QIAseq Targeted DNA** panels (see chapter 5). It is therefore important to choose an appropriate template workflow.

Contents

6.1 QIAseq Pro Fastq to Annotated Germline Variants	84
6.1.1 Outputs from QIAseq Pro Fastq to Annotated Germline Variants	88
6.2 QIAseq Pro Fastq to Annotated Somatic Variants	89
6.2.1 Outputs from QIAseq Pro Fastq to Annotated Somatic Variants	92
6.3 QIAseq Pro Fastq to Germline CNV Control	93
6.3.1 Outputs from QIAseq Pro Fastq to Germline CNV Control	95
6.4 QIAseq Pro Fastq to Somatic CNV Control	96
6.4.1 Outputs from QIAseq Pro Fastq to Somatic CNV Control	97

6.1 QIAseq Pro Fastq to Annotated Germline Variants

The **QIAseq Pro Fastq to Annotated Germline Variants** template workflow identifies germline variants from **QIAseq Targeted DNA Pro** data and annotates these with exon number and amino acid changes. The workflow also produces a read mapping and a coverage report, and if provided with a baseline, copy number variation is also calculated.

The workflow can be found at:

Template Workflows | **LightSpeed Workflows**  | **QIAseq workflows**  | **QIAseq Targeted DNA Pro**  | **QIAseq Pro Fastq to Annotated Germline Variants** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 6.1).

This workflow has been set up to process data generated with QIAseq Targeted DNA Pro panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA Pro panels are available in the **QIAseq DNA Pro Panels hg38** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Pro Panels hg38** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

The reference data set for QIAseq Targeted DNA panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

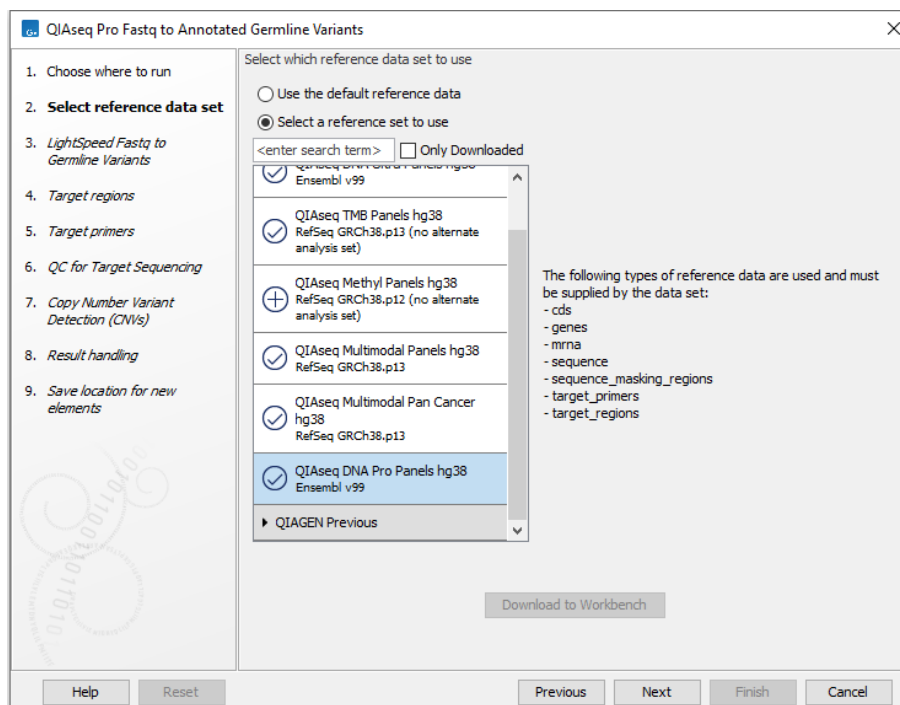


Figure 6.1: Select a reference data set.

In the LightSpeed Fastq to Germline Variants wizard step (figure 6.2) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.

- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

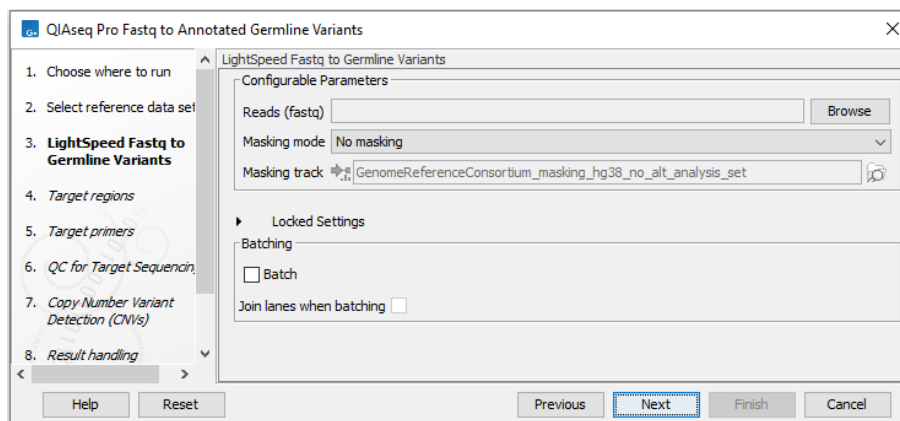


Figure 6.2: Select fastq files.

In the next dialog (figure 6.3), specify the relevant target regions from the drop down list.

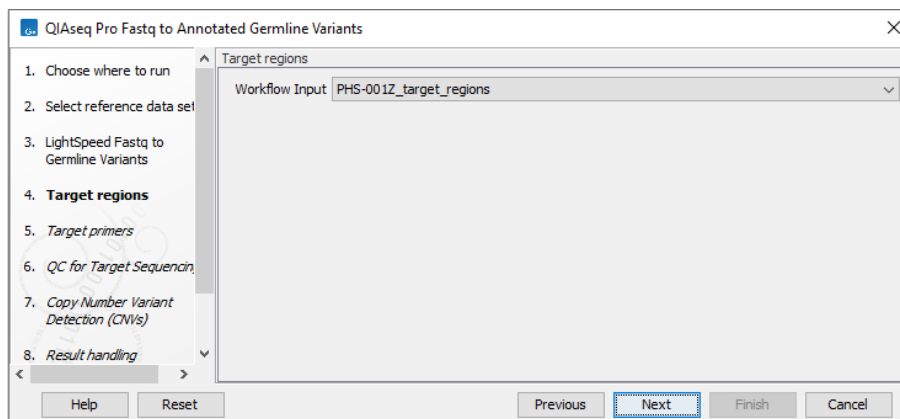


Figure 6.3: Select target regions.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 6.4).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 6.5). Note that the default value for this tool depends on the application chosen (somatic or germline).

The dialog for Copy Number Variant Detection allows you to specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs (figure 6.6). 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.

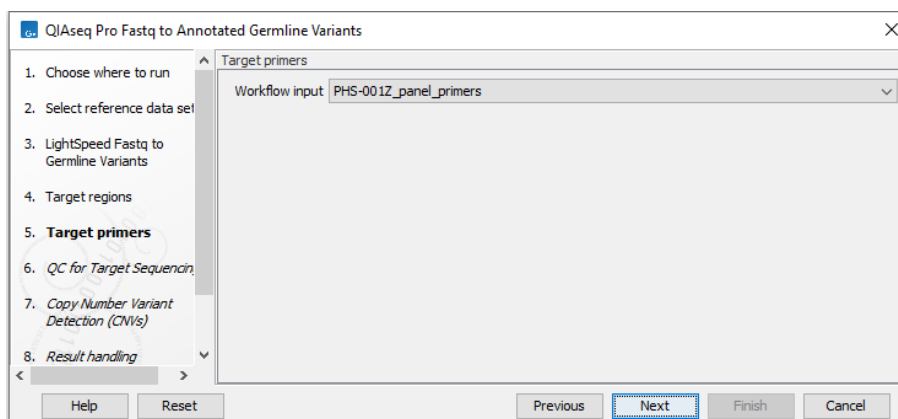


Figure 6.4: Select target primers.

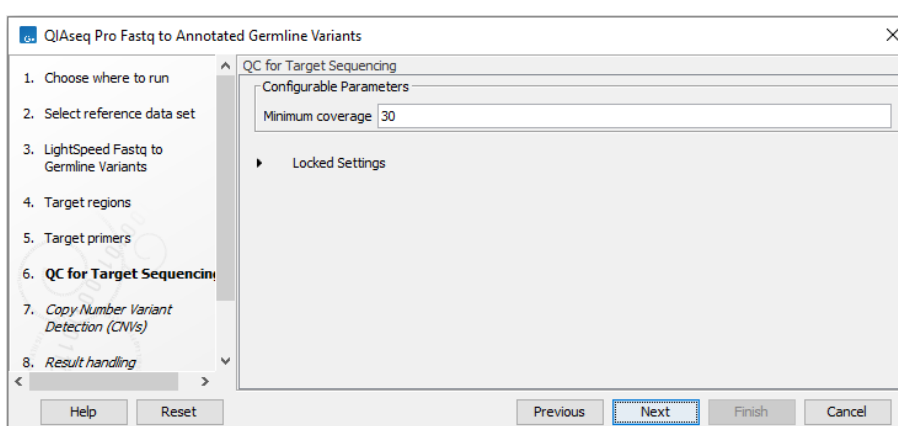


Figure 6.5: Set the Minimum coverage parameter of the QC for Target Sequencing.

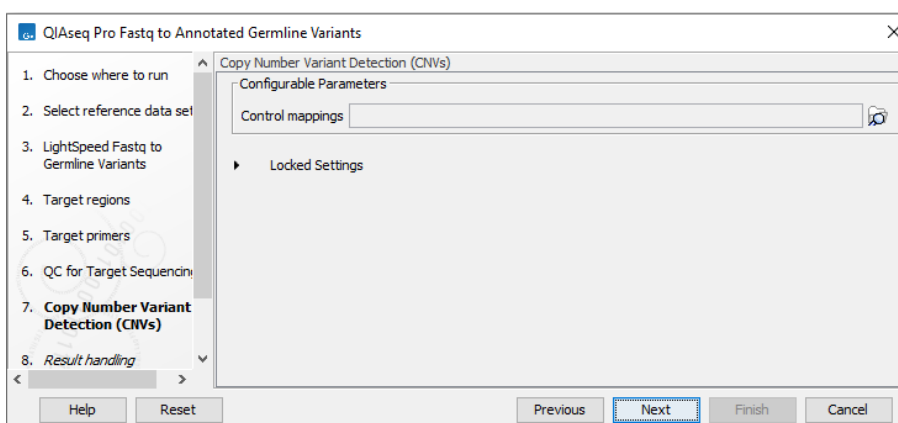


Figure 6.6: Select control coverage tables or read mappings for copy number variant detection.

Please note that if you want the copy number variation 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 workflow described in section 6.3.















A meaningful control must satisfy two conditions: (1) It must have a copy number status that it is meaningful for you to compare your sample 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. One way to achieve this is to process the control using the workflow (without providing a control mapping for the CNV detection component) and then to use the resulting UMI reads track as the control in subsequent workflow runs.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

6.1.1 Outputs from QIAseq Pro Fastq to Annotated Germline Variants

The **QIAseq Pro Fastq to Annotated Germline Variants** template workflow produces the following outputs:

- **Germline Variants** The variant track () with the annotated variants.
- **LightSpeed Report** A report () summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.
- **Inversions** An annotation track () providing the called inversions.
- **Mapped UMI Reads** A read mapping track () with the mapped UMI reads.
- **Amino Acid Track** A track () providing a graphical representation of identified amino acid changes.
- **Genome Browser View** A track list () containing the Variants, the Inversions, the Amino Acid Track, the Mapped UMI Reads, the Target Region Statistics Track, the Gene-level CNV Track, the Target regions as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Sample Report** A report () containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Target Region Statistics Track** A track () providing coverage information per target region.
- **Coverage Report** A report () summarizing coverage.
- **Gene Coverage Track** An annotation track () providing coverage information at the gene level.
- **Target-level CNV Track** An annotation track () providing CNV results per target.
- **Region-level CNV Track** An annotation track () providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Gene-level CNV Track** An annotation track () providing CNV results per gene.
- **CNV Results Report** A report () providing an overview of identified CNVs.

6.2 QIAseq Pro Fastq to Annotated Somatic Variants

The **QIAseq Pro Fastq to Annotated Somatic Variants** template workflow identifies somatic variants from **QIAseq Targeted DNA Pro** data and annotates these with exon number and amino acid changes. The workflow also produces a read mapping and a coverage report, and if provided with a baseline, copy number variation is also calculated.

The workflow can be found at:

Template Workflows | **LightSpeed Workflows**  | **QIAseq workflows**  | **QIAseq Targeted DNA Pro**  | **QIAseq Pro Fastq to Annotated Somatic Variants** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 6.7).

This workflow has been set up to process data generated with QIAseq Targeted DNA Pro panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA Pro panels are available in the **QIAseq DNA Pro Panels hg38** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Pro Panels hg38** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

The reference data set for QIAseq Targeted DNA panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

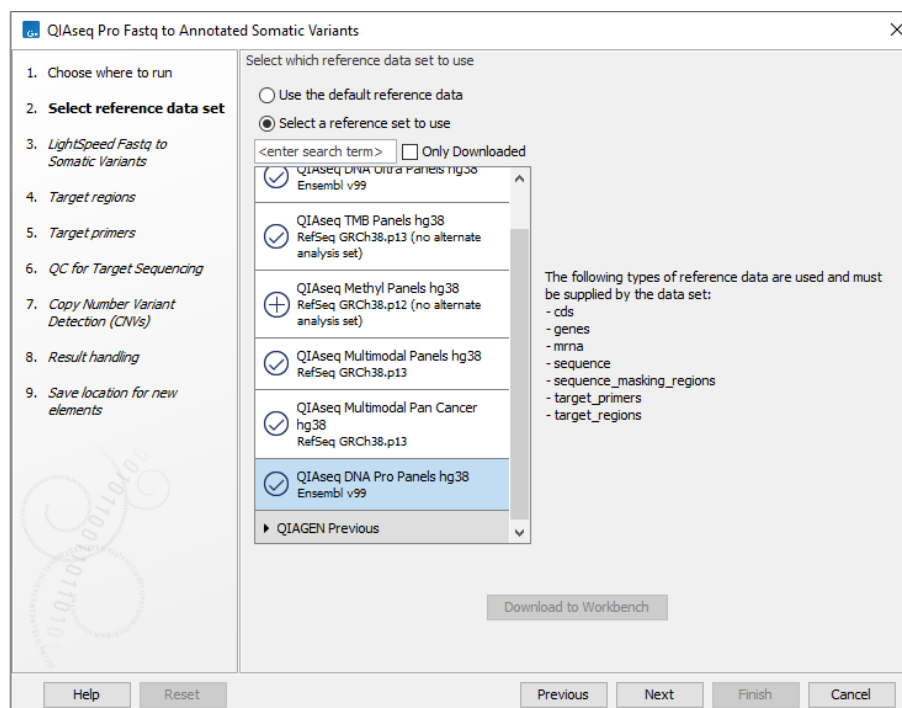


Figure 6.7: Select a reference data set.

In the LightSpeed Fastq to Somatic Variants wizard step (figure 6.8) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

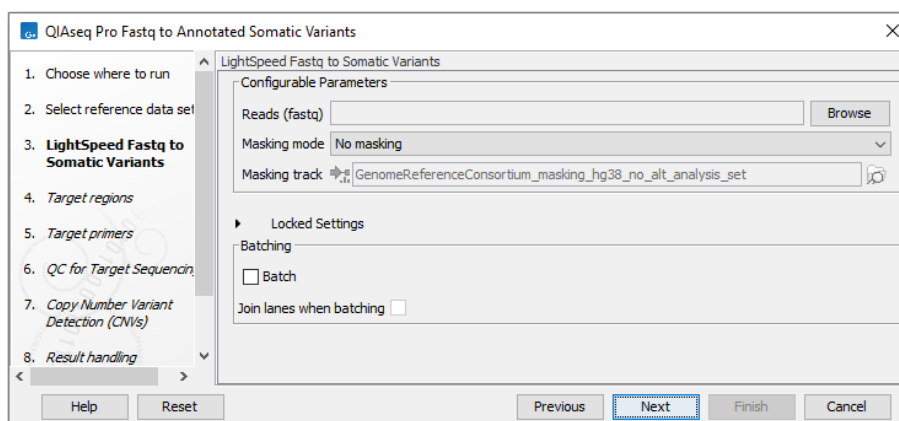


Figure 6.8: Select fastq files.

In the next dialog (figure 6.9), specify the relevant target regions from the drop down list.

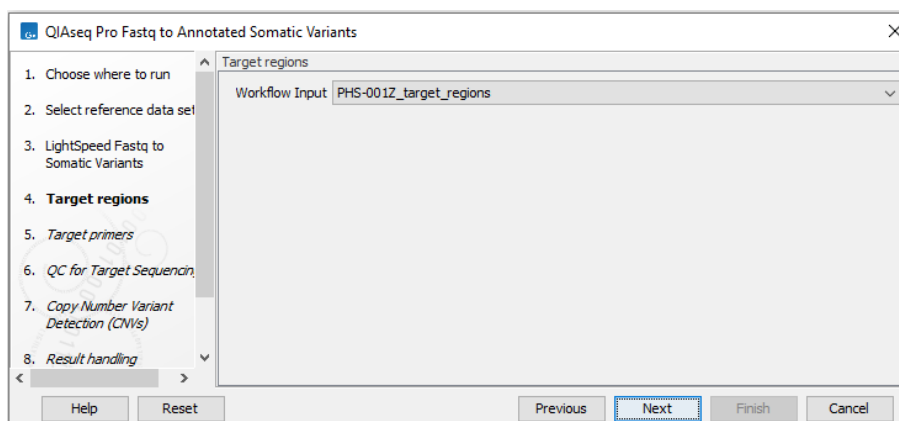


Figure 6.9: Select target regions.

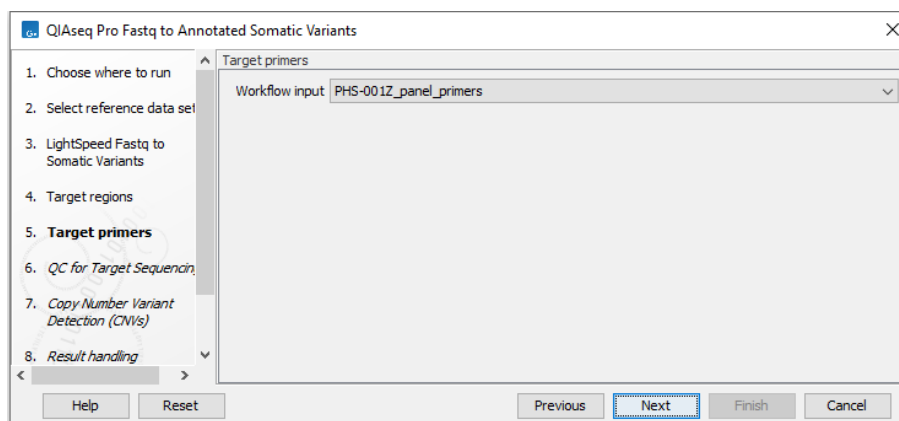


Figure 6.10: Select target primers.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 6.10).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 6.11). Note that the default value for this tool depends on the application chosen (somatic or germline).

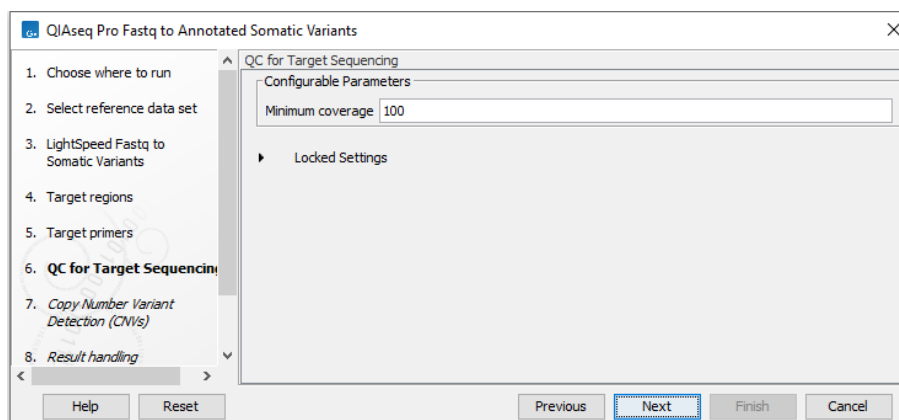


Figure 6.11: Set the Minimum coverage parameter of the QC for Target Sequencing.

The dialog for Copy Number Variant Detection allows you to specify a control mapping against which the coverage pattern in your sample will be compared in order to call CNVs (figure 6.12). 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.

Please note that if you want the copy number variation 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 workflow described in section 6.4.

A meaningful control must satisfy two conditions: (1) It must have a copy number status that it is meaningful for you to compare your sample 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. One way to achieve this is to process the control using the workflow (without providing a control

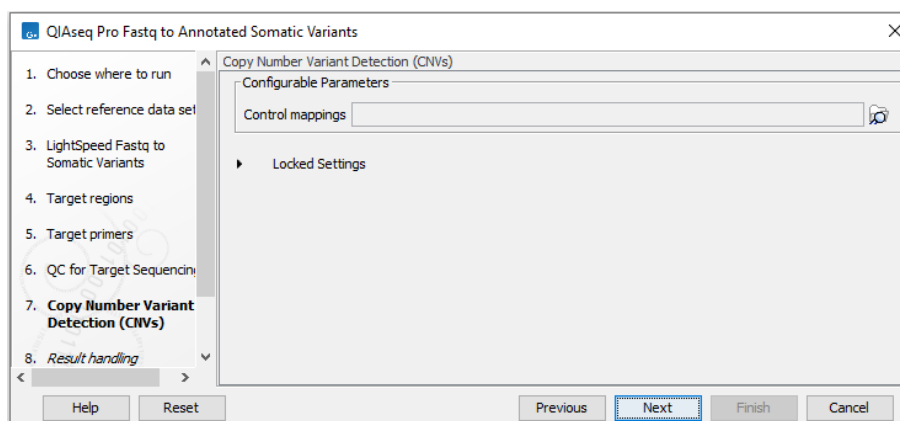


Figure 6.12: Select control coverage tables or read mappings for copy number variant detection.







mapping for the CNV detection component) and then to use the resulting UMI reads track as the control in subsequent workflow runs.

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

6.2.1 Outputs from QIAseq Pro Fastq to Annotated Somatic Variants

The **QIAseq Pro Fastq to Annotated Somatic Variants** template workflow produces the following outputs:

- **Somatic Variants** The variant track (📊) with the annotated variants.
- **LightSpeed Report** A report (📄) summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.
- **Inversions** An annotation track (📊) providing the called inversions.
- **Ignored Regions** An annotation track (📊) providing regions where it was not possible to detect variants due to high complexity among the initial variants being tested.
- **Mapped UMI Reads** A read mapping track (📊) with the mapped UMI reads.
- **Amino Acid Track** A track (📊) providing a graphical representation of identified amino acid changes.
- **Genome Browser View** A track list (📊) containing the Variants, the Inversions, the Ignored regions, the Amino Acid Track, the Mapped UMI Reads, the Target Region Statistics Track, the Gene-level CNV Track, the Target regions as well as the Reference sequence and the Genes, mRNA and CDS tracks.
- **Sample Report** A report (📄) containing essential information from all reports produced by the workflow. For further details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Target Region Statistics Track** A track (📊) providing coverage information per target region.

- **Coverage Report** A report () summarizing coverage.
- **Gene Coverage Track** An annotation track () providing coverage information at the gene level.
- **Target-level CNV Track** An annotation track () providing CNV results per target.
- **Region-level CNV Track** An annotation track () providing CNV results per region, where regions are formed from adjacent targets with similar CNV states.
- **Gene-level CNV Track** An annotation track () providing CNV results per gene.
- **CNV Results Report** A report () providing an overview of identified CNVs.

6.3 QIAseq Pro Fastq to Germline CNV Control

The **QIAseq Pro Fastq to Germline CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

Use the workflow to generate coverage tables for the **QIAseq Pro Fastq to Annotated Germline Variants** (section 6.1) template workflow.

QIAseq Pro Fastq to Germline CNV Control can be found at:

Template Workflows | LightSpeed Workflows () | **QIAseq workflows** () | **QIAseq Targeted DNA Pro** () | **QIAseq Pro Fastq to Germline CNV Control** ()

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 6.13).

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 workflow has been set up to process data generated with QIAseq Targeted DNA Pro panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA Pro panels are available in the **QIAseq DNA Pro Panels hg38** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Pro Panels hg38** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

The reference data set for QIAseq Targeted DNA panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

In the LightSpeed Fastq to Germline Variants wizard step (figure 6.14) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.
- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.

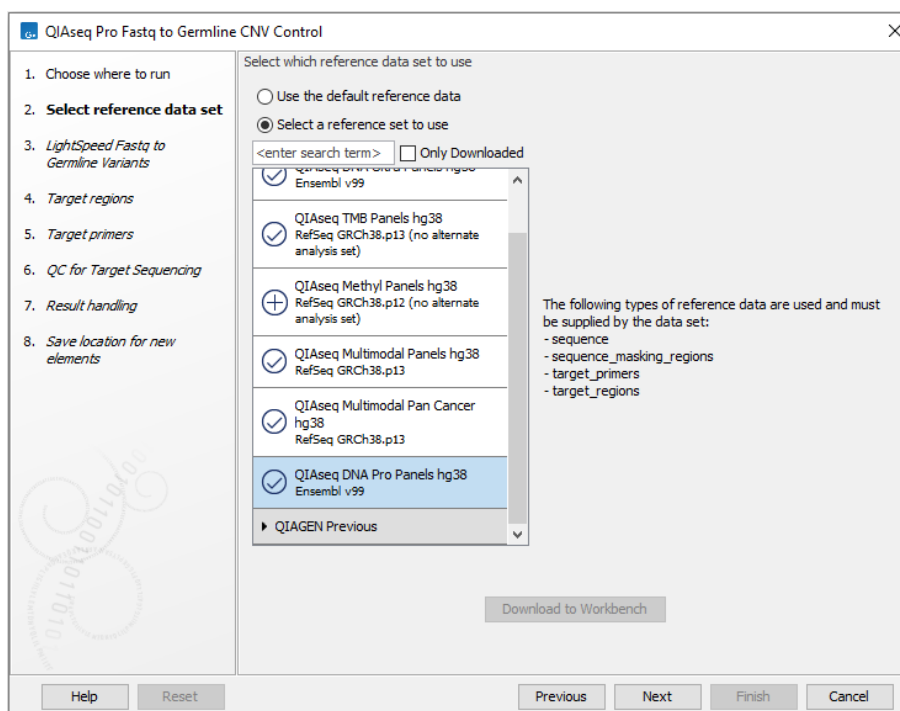


Figure 6.13: Select a reference data set.

- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

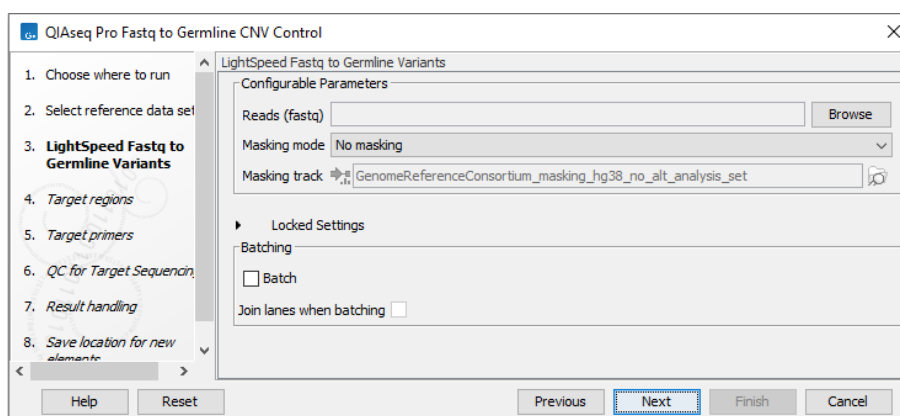


Figure 6.14: Select fastq files.

In the next dialog (figure 6.15), specify the relevant target regions from the drop down list.

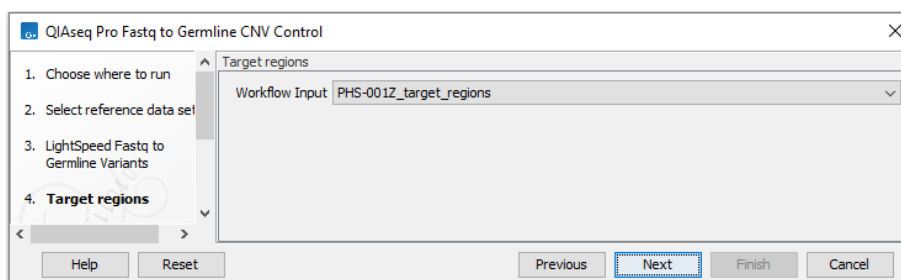


Figure 6.15: Select target regions.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 6.16).

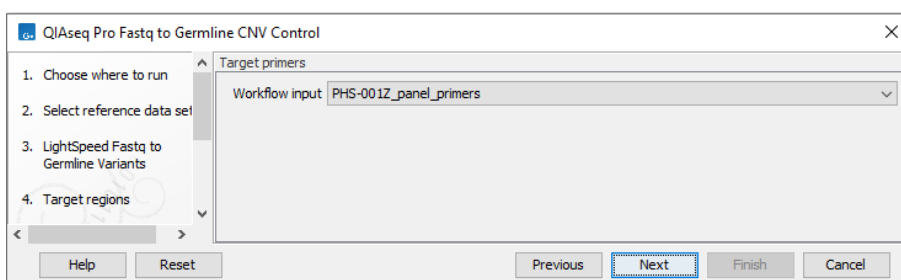


Figure 6.16: Select target primers.

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 6.17). Note that the default value for this tool depends on the application chosen (somatic or germline).

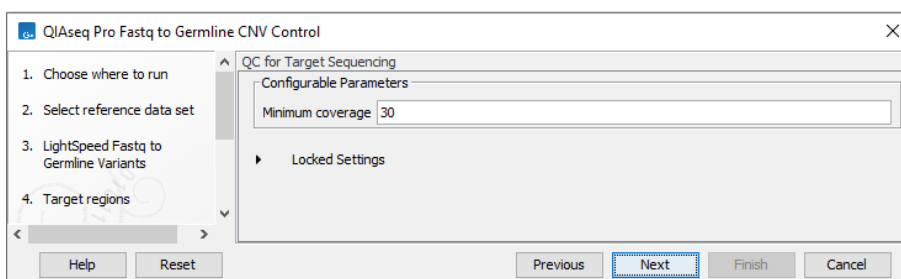


Figure 6.17: Set the Minimum coverage parameter of the QC for Target Sequencing.




In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

6.3.1 Outputs from QIAseq Pro Fastq to Germline CNV Control

The **QIAseq Pro Fastq to Germline CNV Control** template workflow produces the following outputs:

- **Coverage Table** A table (📊) providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection in the **QIAseq Pro Fastq to Annotated Germline Variants** (section 6.1) template workflow or directly

in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).

- **LightSpeed Report** A report  summarizing details of each analysis step performed by the LightSpeed Fastq to Germline Variants tool.
- **Coverage Report** A report  summarizing coverage.
- **Target Region Statistics Track** A track  providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).

6.4 QIAseq Pro Fastq to Somatic CNV Control

The **QIAseq Pro Fastq to Somatic CNV Control** template workflow produces coverage tables that can be used as controls for copy number variant detection.

Use the workflow to generate coverage tables for the **QIAseq Pro Fastq to Annotated Somatic Variants** (section 6.2) template workflow.

QIAseq Fastq to Somatic CNV Control can be found at:

Template Workflows | **LightSpeed Workflows**  | **QIAseq workflows**  | **QIAseq Targeted DNA Pro**  | **QIAseq Pro Fastq to Somatic CNV Control** 

If you are connected to a CLC Server via your 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 first wizard step, select a Reference Data Set (figure 6.18).

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 workflow has been set up to process data generated with QIAseq Targeted DNA Pro panels, and it is important to choose the right reference data to get the reads correctly processed.

The off-the-shelf QIAseq Targeted DNA Pro panels are available in the **QIAseq DNA Pro Panels hg38** reference data set. If you have not downloaded the Reference Data Set yet, the dialog will offer the opportunity to download it using the Download to Workbench button.

If the **QIAseq DNA Pro Panels hg38** reference data set does not contain the needed primers and target regions, a custom reference data set can be created, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Custom_Sets.html.

The reference data set for QIAseq Targeted DNA panels should *not* be used with this workflow. The differences in read structure will for example prevent primers from being correctly trimmed.

In the LightSpeed Fastq to Germline Variants wizard step (figure 6.19) you have the following options:

- **Reads (fastq)** Press **Browse** to select fastq files for analysis.

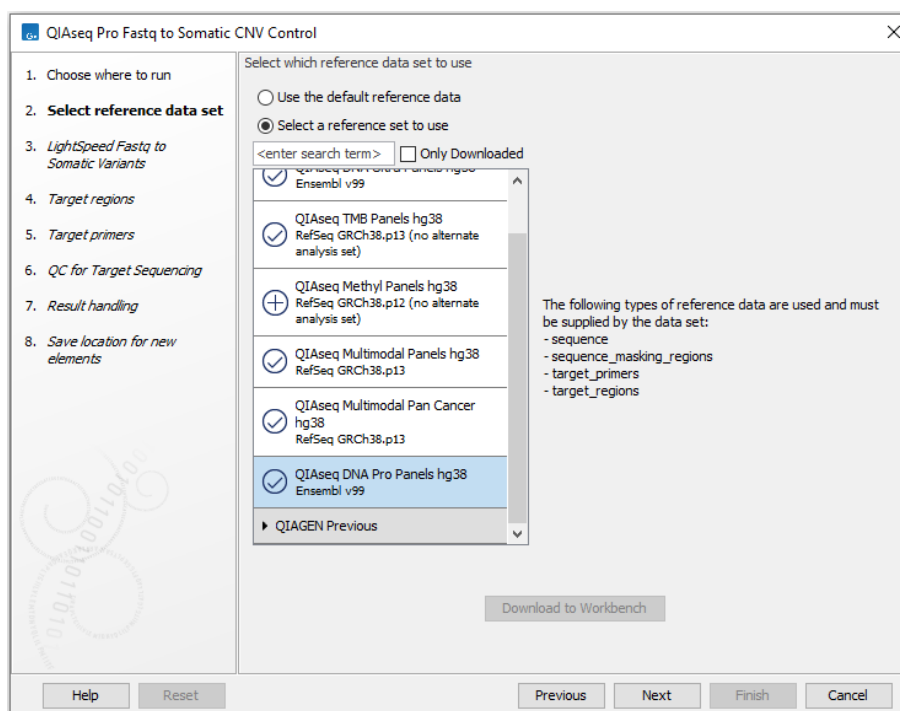


Figure 6.18: Select a reference data set.

- **Masking mode** To enable reference masking when mapping reads, set this option and select a masking track.
- **Masking track** Provide a masking track for the chosen reference genome if reference masking has been enabled.
- **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually. The names of the fastq files must follow standard Illumina naming scheme to allow the tool to identify individual fastq files as belonging to the same sample.
- **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

In the next dialog (figure 6.20), specify the relevant target regions from the drop down list.

Repeat the selection of the appropriate track for Target primers in the subsequent dialog (figure 6.21).

In the dialog called QC for Target Sequencing, you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 6.22). Note that the default value for this tool depends on the application chosen (somatic or germline).

In the final wizard step, choose to **Save** the results of the workflow and specify a location in the Navigation Area before clicking **Finish**.

6.4.1 Outputs from QIAseq Pro Fastq to Somatic CNV Control

The **QIAseq Pro Fastq to Somatic CNV Control** template workflow produces the following outputs:

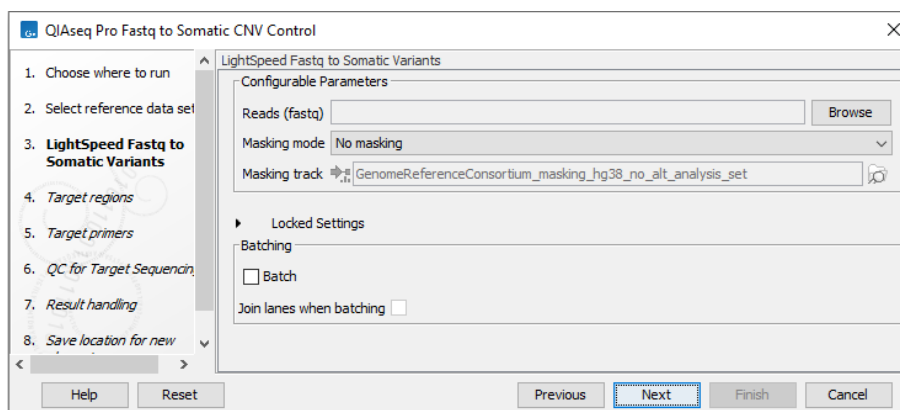


Figure 6.19: Select fastq files.

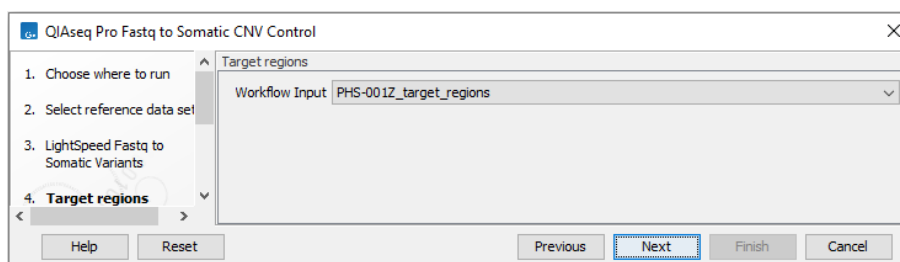


Figure 6.20: Select target regions.

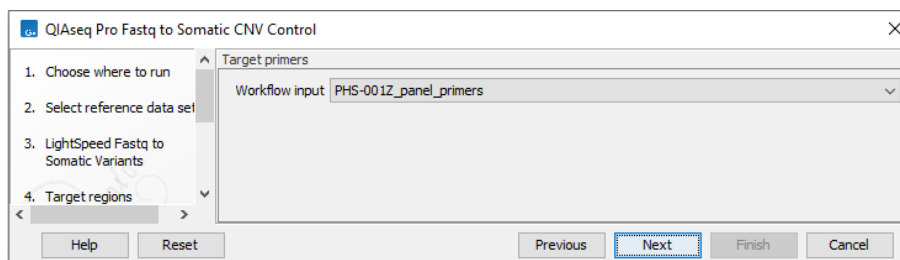


Figure 6.21: Select target primers.

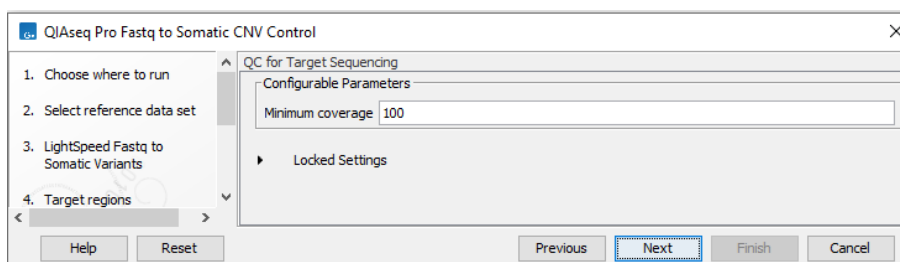





Figure 6.22: Set the Minimum coverage parameter of the QC for Target Sequencing.

- **Coverage Table** A table (📊) providing coverage information per position in the target regions. The coverage table can be used as control for copy number variant detection in the **QIAseq Pro Fastq to Annotated Somatic Variants** (section 6.2) template workflow or directly in the tool **Copy Number Variant Detection (CNVs)** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html).

- **LightSpeed Report** A report () summarizing details of each analysis step performed by the LightSpeed Fastq to Somatic Variants tool.
- **Coverage Report** A report () summarizing coverage.
- **Target Region Statistics Track** A track () providing coverage information per target region.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Targeted-Sequencing.html).