CLC **LightSpeed** Module

USER MANUAL

User manual for

CLC LightSpeed Module 25.2

Windows, macOS and Linux

August 19, 2025

CLC software is intended for scientific research applications. CLC software is not intended for the diagnosis, prevention or treatment of a disease.

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

Contents

Part I

Introduction

Chapter 1

Introduction

This manual describes the functionalities that are available in the *CLC LightSpeed Module 25.2*.

1.1 Overview of CLC LightSpeed Module

The CLC LightSpeed Module offers specialized tools for fast, accurate variant calling for germline, somatic, and tumor-normal analyses, taking FASTQ files as input.

It supports whole-genome sequencing (WGS), whole-exome sequencing (WES), and QIAseq targeted panels, with an end-to-end solution supporting the following optional steps:

- Quality Trimming
- Adapter Trimming
- Read Mapping
- Deduplication
- Local Realignment
- UMI Grouping
- Primer Trimming
- Variant Calling
- QC Reporting

A set of template workflows is also available, offering enhanced capabilities for variant annotation, extended quality control (QC), and advanced data visualization features.

In addition, dedicated tools are available to facilitate the calculation of tumor mutational burden (TMB) scores and the detection of copy number variants (CNVs) from whole-genome sequencing (WGS) data.

The QIAGEN CLC LightSpeed Module is frequently updated. A detailed list of new features, improvements, bug fixes, and changes is available at https://digitalinsights.qiagen.com/clc-lightspeed-module-latest-improvements/.

1.2 Contact information

QIAGEN CLC LightSpeed Module is developed by:

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

https://digitalinsights.qiagen.com/

Email: ts-bioinformatics@qiagen.com

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

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

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

1.3 System requirements

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

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

Compatibility

CLC LightSpeed Module 25.2 and CLC LightSpeed Server Extension 25.2 can be installed on *CLC Genomics Workbench* 25.0 and *CLC Genomics Server* 25.0, respectively, and on later versions in the same major release line.

1.4 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.

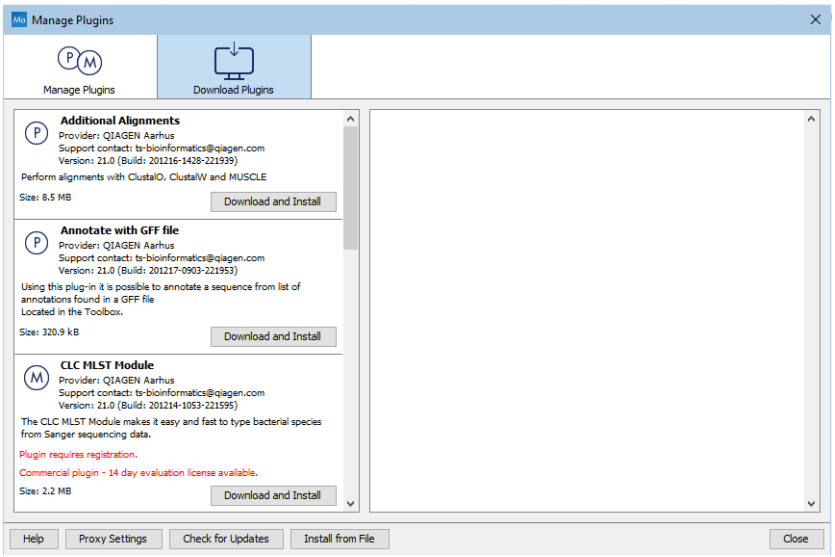

Figure 1.1: *Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.*

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.

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*.

1.4.1 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.

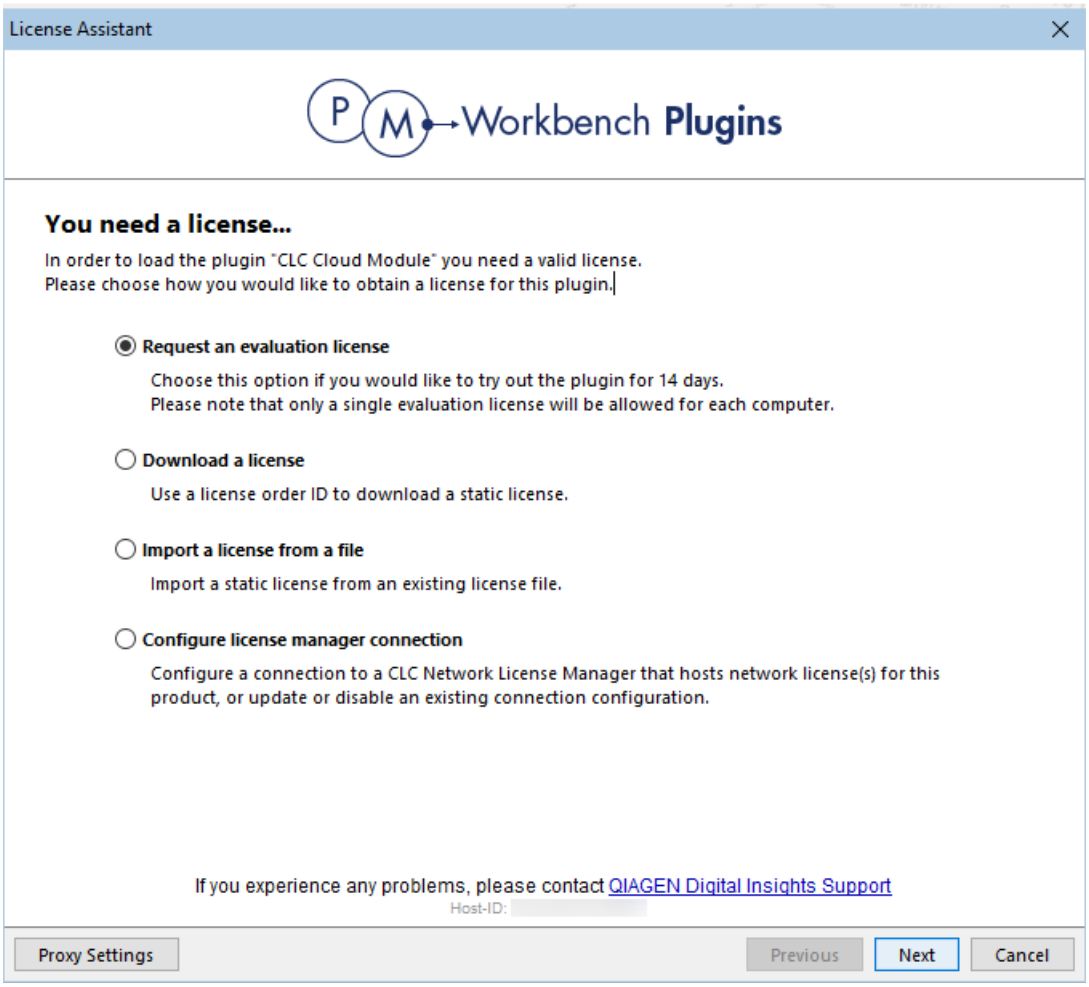

Figure 1.2: *The License Assistant provides options for licensing modules installed on the Workbench.*

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

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

1.4.2 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.

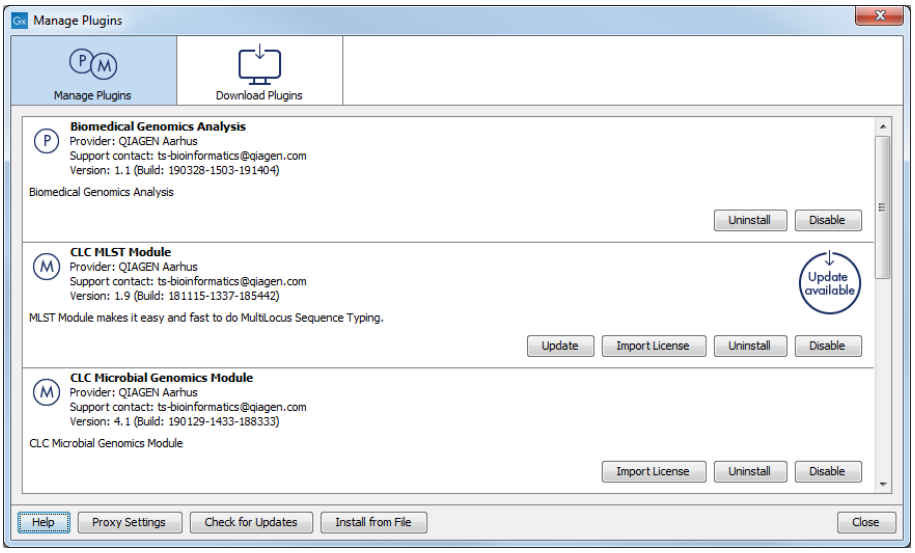

Figure 1.3: *Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.*

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.

1.5 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 1.5.1.
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 nod setup, and opening the **Download Plugins ( )** area (figure 1.4).

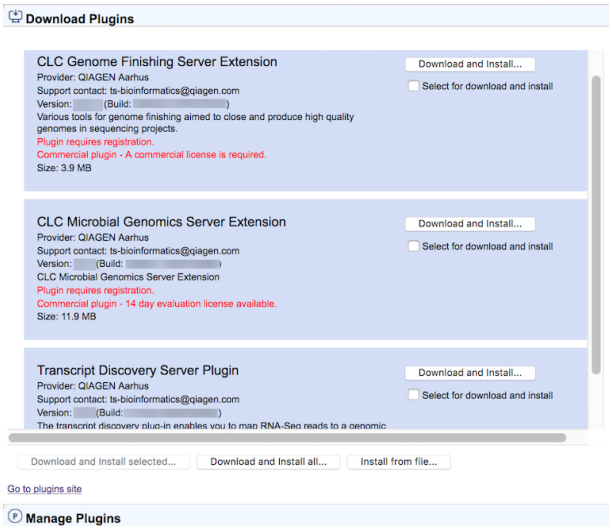

Figure 1.4: *Installing plugins and server extensions is done in the Download Plugins area under the Extensions tab.*

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 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.

Links to related documentation

- Logging into the *CLC Server* web administrative interface: https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Logging_into_administrative_interface.html
- Maintenance Mode: https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Server_maintenance.html
- Restarting the server: https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Starting_stopping_server.html

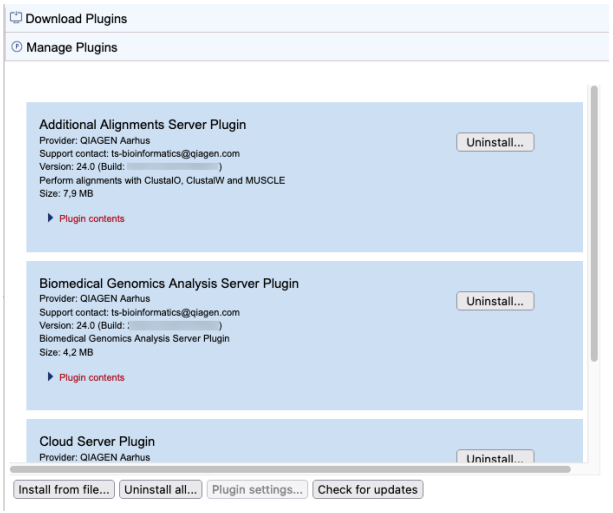

Figure 1.5: *Managing installed plugins and server extensions is done in the Manage Plugins area under the Extensions tab. Clicking on Plugin contents opens a list of the tools delivered by the plugin.*

- Plugins on job node setups: https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Installing_Server_plugins_on_job_nodes.html
- Grid worker re-deployment: https://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 1.5.1.

1.5.1 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 client 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.

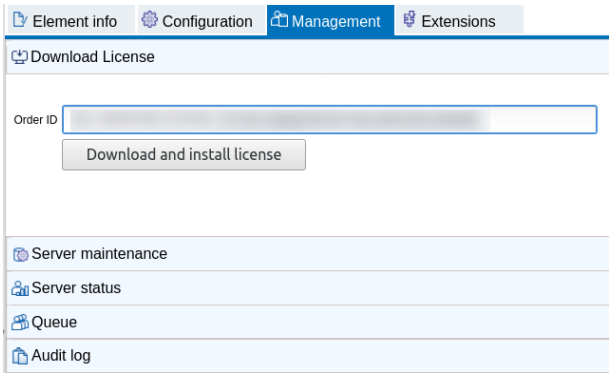

Figure 1.6: *License management is done under the Management tab.*

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. You are offered the option to restart the *CLC Server* after downloading the license file. The server can be started later instead, for example if you wish to carry out multiple administrative tasks before restarting.

nformation about restarting can be found at `https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Starting_stopping_server.html`.

If you are working on a system that does not have access to the external network, then please refer to `https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Download_static_license_on_non_networked_machine.html`.

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

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

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: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Quality_trimming.html.

For **LightSpeed Fastq to Germline Variants**, the default quality limit used for trimming is 0.05.

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 in **LightSpeed Fastq to Somatic Variants** and **LightSpeed Fastq to Somatic Variants Tumor Normal**. 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 beyound the 5' end of the other read in the pair, is considered adapter sequence and is trimmed.

If consensus sequences can be calculated from the adapter sequences removed from R1 and/or R2, these are included in the report. Provided consensus sequences start with the first base that is removed, and continue until it is no longer possible to confidently calculate a consensus.

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. After the first run, the index is cached and reused on later runs.

Read mapping and read pairs

LightSpeed maps reads to the indexed reference sequence.

Single reads that are part of a paired read are mapped individually in the following steps:

- **Seeding** All possible stretches of exact matches (seeds) to the reference are identified. Seeds that are a sub-match of a longer seed or shorter than 2/3 the length of the longest seed are skipped.
- **Extension** Seeds are extended using a Needleman-Wunsch based method. Only seeds with extensions scoring at least 4/5 the score of the highest scoring extension are kept.
- **Pairing** A search through all combinations of extensions is conducted and all proper pairs scoring the maximum score are collected.

Read pairs that did not map well or were not paired, go through a second round of more thorough seeding:

- **Secondary seeding** A search for shorter seeds, that are sub-matches of longer seeds, is conducted.
- **Secondary extension** All seeds, including the seeds shorter than 2/3 the length of the longest seed, are extended.
- **Pairing** A search through all combinations of extensions is conducted, i.e., both from the primary and secondary extension.

If there are multiple paired extensions with the highest score, one of the pairs is selected at random and the read pair is reported as non-specific.

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.

Unaligned ends of read pairs that are specifically mapped are, during read mapping, reattempted aligned by allowing for one mismatch. 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.

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.

Q-scores are assigned to the bases in the consensus read as follows:

- The Q-score assigned to a base in the consensus read is calculated as the average of the quality scores of the underlying bases.
- At conflicting positions, where there is a most common base, the Q-score assigned to a base in the consensus read is calculated as the average of the quality scores of the reads with the winning nucleotide.
- If the conflicting bases are equally represented, the consensus base will either be the reference symbol or N. In both cases no Q-score assignment is made to the base.

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 read mapping 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 read mapping relative to the reference.

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

1. A graph is built, containing nucleotide sequence paths corresponding to all reads as well as the reference. If significant unaligned end breakpoints are found in the original read mapping, the graph construction for that region may involve de-novo assembly of the reads.
2. The graph undergoes refinement where paths that are unlikely to contain variants relative to the reference path are removed, and additional variants such as structural variants inferred from indirect evidence may be added.
3. Any read that intersects the region of interest is finally realigned against the graph.

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. Tandem duplications are reported in the variant track, and those only inferred from unaligned ends are annotated with **Yes** in the variant track column **Inferred from unaligned ends**.

When tandem duplications are inferred from unaligned ends, the allele count and coverage is estimated based on the breakpoint with the highest unaligned end count, assuming the following:

- Reads are evenly distributed across the duplicated region.
- Half of the reads that cross from the insertion to the downstream duplicated region are aligned with unaligned ends at one breakpoint, and the other half at the other breakpoint.

In situations where the assumptions are not met, such as for some targeted data where the reads are not evenly distributed, the count and coverage estimates may be inaccurate. The breakpoints used for inference are detected prior to realignment, so the unaligned ends can not necessarily be found in the output read mapping where the alignment of the reads may have changed.

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. Identified inversions are reported in the inversions track.

2.6 UMI grouping

All of the LightSpeed tools can group reads based on Unique Molecular Identifiers (UMIs). Both protocols where the UMI is present on only one read in a pair or both reads in a pair (duplex UMI) are supported.

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

For duplex UMIs, UMI grouping is a two step process, where reads are first grouped to simplex reads and then to duplex reads.

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.

Q-scores are assigned to the bases in the UMI read as follows:

- For UMI groups with only one read (singleton groups), the Q-scores of the bases in the original read are used.
- For UMI groups with more than one read, and where all reads agree on the base, the average Q-score of the bases is used. However, if this value is smaller than the adaptation of the MAGERI Q-score, Q_M (described below), the Q_M value is used.
- At conflicting positions, where there is a most common base, the Q-score assigned to a base in the consensus read is calculated using an adaptation of MAGERI (a method described in [Shugay et al., 2017]). Specifically, the Q_M value is used as outlined in the following definitions:
  - count = number of reads with the winning nucleotide
  - total = total number of reads in UMI group

  - f = count/(total + 0.9)
  - Q_M = 60/3*(4*f -1)

- If the conflicting bases are equally represented, the consensus base will either be the reference symbol or N. In both cases no Q-score assignment is made to the base.

Examples of the resulting UMI read Q-scores are given in figure 2.1.

Reads in UMI group (total)	Reads w winning nucleotide (count)	Q-score on winning nucleotides	Q sum / count	Q_M	FINAL Q-score
1	1	11	NA	NA	11
1	1	25	NA	NA	25
1	1	37	NA	NA	37
2	2	11, 37	24	35.17	35
2	2	25, 37	31	35.17	35
2	2	37, 37	37	35.17	37
3	3	11, 37, 37	28.33	41.54	42
3	3	11, 25, 37	24.3	41.54	42
3	3	37, 37, 37	37	41.54	42
3	2	37, 37	37	21.03	21
3	2	11, 37	24	21.03	21
4	4	11, 11, 11, 11	11	45.31	45
4	4	37, 37, 11, 11	24	45.31	45
4	4	37, 37, 37, 37	37	45.31	45
4	3	37, 37, 11	28.3	28.98	29
4	3	37, 37, 37	37	28.98	29
4	2	11, 25	18	12.65	13
4	2	37, 37	37	12.65	13

Figure 2.1: *Assigned Q-scores exemplified for various UMI group sizes, base quality scores and base ambiguity among contributing reads.*

When variants are called from UMI reads, additional UMI specific annotations are added, see section 2.8 or section 2.9.

For limitations in UMI grouping, see section 2.14.

Definitions

- **Duplex UMI** A protocol where both read 1 and read 2 in a pair contain a UMI. Reads originating from both strands of a DNA fragment can be grouped.
- **Singleton UMI read pairs** A UMI read pair that is based on only one input read pair.
- **Simplex UMI read pairs** For duplex protocols, the number of simplex UMI read pairs is provided. Simplex UMI read pairs are UMI read pairs where input reads all originate from the same strand. Singleton UMI read pairs are a subset of the simplex UMI read pairs.
- **Duplex UMI read pairs** UMI read pairs that are based on input reads from both strands.

2.7 Primer trimming

When primer trimming is enabled, the part of the read that overlaps a primer at the expected position (3' or 5') is unaligned. The unaligned ends are visible in the read mapping, but are not used for variant calling.

When trimming for primer sequence, it is possible to discard reads that do not match a primer. For reads to match a primer, there must af primer overlap of the percentage specified in the tool wizard. In addition, reads must not start upstream of a primer, and only one mismatch is allowed between the read and the primer sequence.

2.8 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.

At this stage, homopolymer variants with a homopolymer length of >=5 are re-called. This is done by calculating the likelihood of all possible genotypes based on the homopolymer length variants found in the reads, with the assumption that all other homopolymer variants in the reads arise by error. The likelihood is (up to a normalizing constant):

$$\prod_{j=1}^{n}(\sum_{i} P(l_j|l_i)f_j)^{c_j}$$

where $P(l_j|l_i)$ is the probability of observing a homopolymer of length l_j by error when the true length is l_i, and c_j is the number of fragments with a homopolymer of length l_j. f_j is the frequency of the homopolymer with length l_j according to the genotype G. For example, for diploid models this frequency can be 0, 0.5, or 1.0. The probabilities $P(l_j|l_i)$ are determined from the sample, by counting the number of homopolymer errors at positions that appear to be homozygous.

The final homopolymer variants are those that maximize the likelihood. However, if the maximum likelihood genotype is nearly homozygous (by which we mean all except one haplotype has the same variant), then we perform an additional test to see whether a ploidy-0.1:0.1 frequency ratio between the two variants has higher likelihood than the ploidy-1:1 frequency ratio. If it does, then we call the variant as homozygous. This ensures that low levels of noise are tolerated, and improves the accuracy of homopolymer calls.

Notes Special handling is applied to variants supported by only 1 read that have a coverage of 1 or 2. For details, see the description of the **Allele count** option under **Variant filters** in section 3.1.

For insertions only, unaligned ends that are shorter than the full insertion, but matches the

insertion sequence, contribute to the count and coverage.

A limit of maximum three alleles is enforced for each homopolymer locus and for alleles specifically marked with STR "Yes" that affect the same short tandem repeat. The alleles with the highest read counts are retained. See the description of the **STR annotations and filter** option under **Variant filters** in section 3.1 for details about STR annotation.

When enabling the option use non-specific reads for variant detection, for sites with at least 80% ambiguous reads, the sensitivity to heterozygous events is increased. The reason is, that the non-specific reads often spread variant alleles across two or more similar sites, resulting in alleles with lower than 50% allele frequency at the individual sites.

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 general variant annotations here: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html.

In addition to the basic annotations, a number of LightSpeed specific annotations are available:

- General annotations:
  - **Average quality** The average base quality score of the bases supporting a variant. The average quality score is calculated by adding the Q scores of the nucleotides supporting the variant and dividing this sum by the number of nucleotides supporting the variant. For deletions, the average quality score reported is the lowest average quality of the two bases neighboring the deleted one. For insertions, the average quality is calculated for each of the inserted bases in the reads supporting the insertion, and the minimum of the average base qualities is reported. Average quality is only calculated for non-reference alleles, for reference alleles no average quality is reported.
  - **STR** Yes/No annotation. Yes, if the variant meets minimum repeat count, minimum repeat region length and maximum repeat element length specified in the wizard when calling variants. No, if one or more of the thresholds are not met.
  - **Repeat count** The number of repeats excluding the variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele is called "AGAGAGAGAG", the repeat unit is 2 and the repeat count is 4.
  - **Repeat unit length** The length of a repeat unit. For example, for the dinucleotide repeat "AGAGAGAG", the repeat unit length is 2.
  - **Strand balance score** 1 - (p-value from binomial test given forward count, count, and forward count/coverage).
- Annotations added to variants that are called from UMI reads:

- **Count (singleton UMI)** The number of singleton UMI read pairs supporting the allele.
- **Count (big UMI)** The number of big UMI read pairs supporting the allele.
- **Proportion (singleton UMIs)** The fraction of singleton UMI read pairs relative to all UMI read pairs supporting the allele.
- **Average size (UMIs)** Average number of read pairs per UMI.
- **Average size (simplex UMIs)** Average number of read pairs per UMI for simplex UMI read pairs. The annotation is only added for duplex UMI protocols.
- **Count (duplex UMIs)** The number of duplex UMI read pairs supporting the allele. The annotation is only added for duplex UMI protocols.
- **Average size (duplex UMIs)** Average number of read pairs per UMI for duplex UMI read pairs. The annotation is only added for duplex UMI protocols.

Note that for insertions, counts from unaligned ends that are shorter than the full insertion, but matches the insertion sequence, are included in the variant annotations Count, Coverage, Frequency, Count (singleton UMI), Count (big UMI), and Proportion (singleton UMIs). Counts from unaligned ends are not included in Forward read count, Reverse read count, Forward coverage, reverse coverage and Forward/reverse balance.

2.9 Somatic variant detection

To call somatic variants, a number of steps are followed.

Firstly, positions of interest that may contain variation due to sequencing errors are identified. Two types of error are considered i) error that is due to the preceding sequenced nucleotides, and ii) error in homopolymer regions, which is typically dependent on how much the sample has been amplified. Empirical error distributions are made from both of these sequencing error types. For example, the probability of seeing a polyA homopolymer of length 18 given that the true length is 20 will be calculated.

Secondly, positions of interest that may contain variation not due to sequencing errors are identified. This identification is subject to user-controllable parameters (see the options under **Variant detection** and **Variant detection general filters**, section 3.2). Groups of adjacent positions of interest form a cluster. Often, such a cluster is just a single position, but it may be arbitrarily long.

For each of these clusters, all overlapping read fragments are reduced to their intersection with the sites of the cluster. These reduced fragments are then used in the further analysis of the cluster.

To identify which underlying haplotypes are present within a given cluster, the pairwise compatibility of the fragments is determined. Once this is known, the largest groups of such pairwise-compatible fragments are formed. Each nonconflicting group is then turned into a haplotype candidate by piecing together the information from the fragments within the group.

At this stage, homopolymer variants are filtered. The filtering is applied when a cluster contains variants that imply at least two homopolymer haplotypes, and where at least one of the homopolymers has length >=5. The filtering removes homopolymer variants that are likely to have arisen by error from the other homopolymers, as determined by the empirical homopolymer error

model. If homopolymer variants with lengths $l_1, l_2, \ldots l_n$, and counts $c_1, c_2, \ldots, c_n$ respectively are present in the sample, then the filtering proceeds as follows:

1. Calculate the frequencies of the homopolymers that would maximize the likelihood given the observed counts under the homopolymer model. The likelihood is (up to a normalizing constant):

$$\prod_{j=1}^{n}(\sum_{i=1, i \notin F}^{n} P(l_j|l_i) f_i)^{c_j}$$

   where $P(l_j|l_i)$ is the probability of observing a homopolymer of length l_j by error when the true length is l_i, and F is the set of indexes of homopolymers that have been filtered. The maximum likelihood is found by expectation-maximization.

2. Omit each homopolymer variant in turn, starting with the homopolymer with fewest counts and again calculate the frequencies of the remaining homopolymers that would maximize the likelihood.

3. If the Bayesian Information Criterion improves by removing the homopolymer variant, then it is filtered away. The Bayesian Information Criterion embodies our preference for simple explanations of the data by applying a penalty to the maximum likelihood for each unfiltered variant.

Once a list of haplotypes believed to be present in a given region is constructed, each of them needs to be assigned a count. Counts are assigned per-position to the haplotypes. In doing so, the haplotype-based per position counts are compared to the fragment-based per position counts to make sure the cumulative difference for all positions is minimized. This ensures assigning the counts that best reconcile the observed fragments with the underlying haplotypes.

Notes In contrast to the germline variant caller (section 2.8), the somatic variant caller makes no assumptions about the ploidy of a sample, and thus allows for sensitive detection of variant alleles at low frequencies.

For insertions only, unaligned ends that are shorter than the full insertion, but matches the insertion sequence, contribute to the count and coverage.

A limit of maximum three alleles is enforced for each homopolymer locus and for alleles specifically marked with STR "Yes" that affect the same short tandem repeat. The alleles with the highest read counts are retained. See the description of the **STR annotations and filter** option under **Variant filters** in section 3.2 for details about STR annotation.

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 and that their count and frequency satisfies the user-provided minimum requirements.

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: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html.

In addition, the following LightSpeed specific annotations are available:

- General annotations:
  - **Average quality** The average base quality score of the bases supporting a variant. The average quality score is calculated by adding the Q scores of the nucleotides supporting the variant and dividing this sum by the number of nucleotides supporting the variant. For deletions, the average quality score reported is the lowest average quality of the two bases neighboring the deleted one. For insertions, the average quality is calculated for each of the inserted bases in the reads supporting the insertion, and the minimum of the average base qualities is reported. Average quality is only calculated for non-reference alleles, for reference alleles no average quality is reported.
  - **p-value - global error rate** p-value from binomial test given count and coverage.
  - **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.
  - **STR** Yes/No annotation. Yes, if the variant meets minimum repeat count, minimum repeat region length and maximum repeat element length specified in the wizard when calling variants. No, if one or more of the thresholds are not met.
  - **Repeat count** The number of repeats excluding the variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele is called "AGAGAGAGAG", the repeat unit is 2 and the repeat count is 4.
  - **Repeat unit length** The length of a repeat unit. For example, for the dinucleotide repeat "AGAGAGAG", the repeat unit length is 2.
  - **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.
- Annotations added to variants that are called from UMI reads:
  - **Count (singleton UMI)** The number of singleton UMI read pairs supporting the allele.
  - **Count (big UMI)** The number of big UMI read pairs supporting the allele.
  - **Proportion (singleton UMIs)** The fraction of singleton UMI read pairs relative to all UMI read pairs supporting the allele.

  - **Average size (UMIs)** Average number of read pairs per UMI.
  - **Average size (simplex UMIs)** Average number of read pairs per UMI for simplex UMI read pairs. The annotation is only added for duplex UMI protocols.
  - **Count (duplex UMIs)** The number of duplex UMI read pairs supporting the allele. The annotation is only added for duplex UMI protocols.
  - **Average size (duplex UMIs)** Average number of read pairs per UMI for duplex UMI read pairs. The annotation is only added for duplex UMI protocols.

Note that for insertions, counts from unaligned ends that are shorter than the full insertion, but matches the insertion sequence, are included in the variant annotations Count, Coverage, Frequency, Count (singleton UMI), Count (big UMI), and Proportion (singleton UMIs). Counts from unaligned ends are not included in Forward read count, Reverse read count, Forward coverage, reverse coverage and Forward/reverse balance.

2.10 Tumor normal variant detection

Tumor normal variant detection relies on the same method as somatic variant detection (section 2.9), 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: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/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 2.9) 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 and normal coverage.

Note that for insertions, counts from unaligned ends that are shorter than the full insertion, but matches the insertion sequence, are included in count and coverage and other variant

annotations, see somatic variant detection (section 2.9). However, counts from unaligned ends are not included in the annotations Count in normal, Coverage in normal, and Frequency in normal.

2.11 Variant detection using target regions

When providing target regions to **LightSpeed Fastq to Germline Variants**, **LightSpeed Fastq to Somatic Variants**, or **LightSpeed Fastq to Somatic Variants Tumor Normal**, only variants that overlap the target regions are reported, with the following exceptions:

- Insertions occurring between the last position outside a target region and the first position inside a target region are reported.
- Deletions starting on the first position after a target region are reported. This aligns with VCF format, where the position of deletions are given as the preceding nucleotide, which is then inside the target region.
- Tandem duplications occurring outside target regions are reported if the duplicated region overlaps a target.

2.12 Batch processing

The tools **LightSpeed Fastq to Germline Variants**, **LightSpeed Fastq to Somatic Variants**, and **LightSpeed Fastq to Somatic Variants Tumor Normal** have limited batch processing functionality.

For each of the three tools, batch processing is only possible when they are run as part of a workflow.

The sections below describe additional limitations that apply to each of the tools.

LightSpeed Fastq to Germline Variants and LightSpeed Fastq to Somatic Variants When running the tools as part of a workflow, it is possible to batch over fastq files (figure 2.2), if not batching over other inputs to the LightSpeed tools. When batching over the fastq files, the names of the fastq files are used to determine which fastq files are analyzed together (see Default rules for determining pairs of files here https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Illumina.html).

When not batching over fastq files, it is possible to batch over the references, masking track, primers track and target regions tracks, when the tracks are added as input elements in a workflow (figure 2.3).

LightSpeed Fastq to Somatic Variants Tumor Normal It is not possible to batch over fastq files, whereas batching over references, masking track, primers track and target regions tracks is possible when they are added as input elements in a workflow.

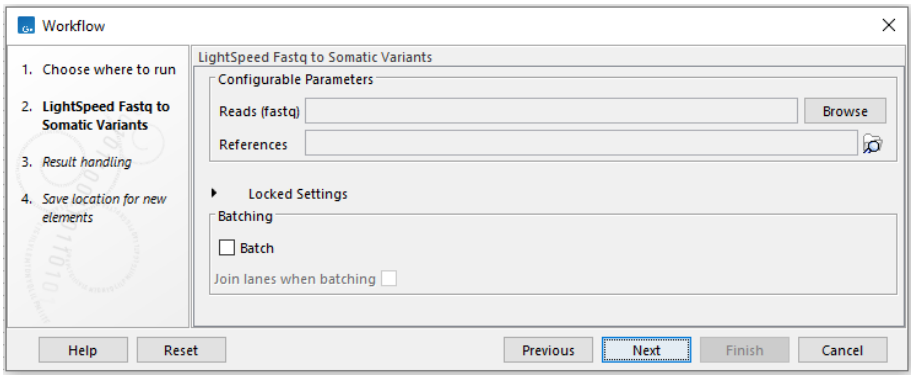

Figure 2.2: *When running LightSpeed Fastq to Germline Variants or LightSpeed Fastq to Somatic Variants in a workflow, it is possible to batch over fastq files. Check "Batch" in the wizard step where "Reads (fastq)" files are selected.*

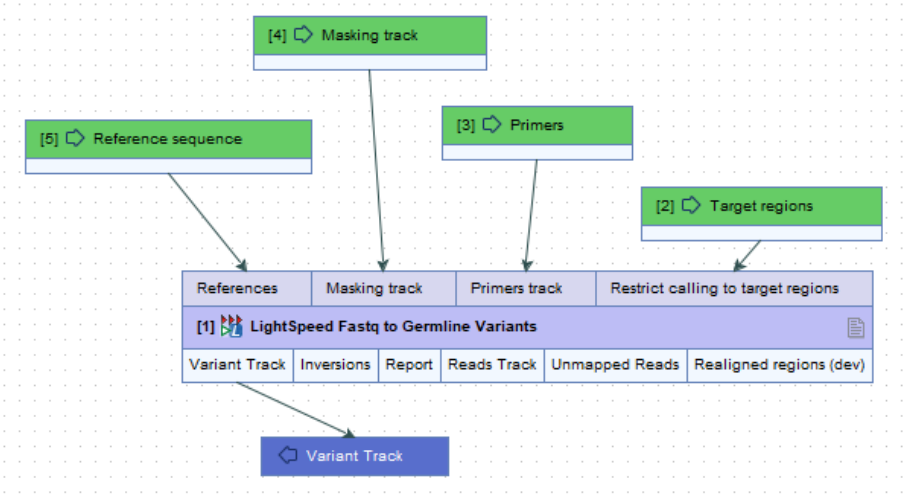

Figure 2.3: *When running LightSpeed Fastq to Germline Variants and LightSpeed Fastq to Somatic Variants in a workflow, and not batching over fastq files, it is possible to batch over the inputs references, masking track, primers track and target regions when they are added as input elements in a workflow.*

2.13 Compatibility with CLC Genomics Workbench tools

Read mappings produced by the tools **LightSpeed Fastq to Germline Variants**, **LightSpeed Fastq to Somatic Variants**, and **LightSpeed Fastq to Somatic Variants Tumor Normal** differ in some fundamental ways from read mappings produced by the **Map Reads to Reference** tool, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Map_Reads_Reference.html. The main differences are related to how base quality scores and unaligned ends are handled:

- **Base quality scores** The Lightspeed tools access and use the base quality scores during variant calling, but for efficiency reasons the base quality scores on the individual bases are not kept and stored on individual bases in the final read mapping. CLC Genomics Workbench tools that rely on quality scores being available on bases in a read mapping (such as the variant callers **Fixed Ploidy Variant Detection**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/

index.php?manual=Fixed_Ploidy_Variant_Detection.html and **Low Frequency Variant Detection**, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Low_Frequency_Variant_Detection.html) cannot be expected to perform as well on a read mapping produced by a Lightspeed tool as on a read mapping produced with the **Map Reads to Reference** tool available in CLC Genomics Workbench.

- **Unaligned ends** The Lightspeed tools include a structural variant detection step and a subsequent realignment step. In the structural variant detection step, structural variants are inferred from unaligned ends, and the realignment step uses the inferred structural variants during realignment. The realignment procedure differs from that used by the **Local Realignment** tool, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Local_Realignment.html (available in CLC Genomics Workbench) in that unaligned ends that equally well support different alleles are kept unaligned. CLC Genomics Workbench tools that rely on unaligned ends in a read mapping (such as the **InDels and Structural Variants** tool, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=InDels_Structural_Variants.html) will therefore behave differently on a read mapping produced with Lightspeed tools compared to a read mapping produced with the **Map Reads to Reference** tool.

2.14 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 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.

Mapping Reads with unaligned ends that extend outside chromosomes are discarded.

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 accross the junction of circular chromosomes.

UMI grouping

- The maximum number of reads used for creating a UMI consensus read is 100,000. Therefore, UMI groups with more than 100,000 reads will be merged into more than one consensus UMI read.
- LightSpeed UMI grouping requires that reads have similar mapping positions. In data from single primer extension protocols, such as many primer based QIAseq protocols, read pairs representing the same DNA fragment with the same UMI sequence can originate from

different primers. This can happen if primers in the same direction are located near each other, making it possible for a downstream primer to amplify a PCR product generated from an upstream primer. LightSpeed will not group reads originating from different primers.

- When UMIs are used to group reads, the sequence is compared base by base. If an insertion or deletion is present in the beginning of a UMI sequence, this will likely prevent the reads from being grouped because all bases after the variant will be mismatches.

Output naming support The LightSpeed tools support custom names for workflow results, however, not all CLC Genomics Workbench placeholders for workflow output elements are supported. Specifically, the following are supported:

- **{input:1}** or **{2:1}** The name of the first input to the workflow. This is the recommended output naming.
- **{name}** or **{1}** The default name for that output from that tool, i.e. the name that would be used if the tool was run outside a workflow context.
- **{metadata}** or **{3}** The batch unit identifier for workflows executed in batch mode. Depending on how the workflow was configured at launch, this value may be obtained from metadata. Workflows not executed in batch mode or without Iterate elements are not supported with this placeholder as the value will be identical to that substituted using **{input}** or **{2}**.
- **{user}** The username of the person who launched the job.
- **{host}** The name of the machine the job is run on.
- **{year}**, **{month}**, **{day}**, **{hour}**, **{minute}**, and **{second}** Timestamp information based on the time an output is created. Using these placeholders, items generated by a workflow at different times can have different file names.

For additional details, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Configuring_Workflow_Output_Export_elements.html.

Chapter 3

Tools

The CLC LightSpeed Module contains tools which facilitate general and end-to-end NGS secondary analysis. The end-to-end tools are designed for germline, somatic (tumor only) and tumor normal variant detection and each of these run 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

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 (see section 2.12). Template workflows for LightSpeed analyses are available (see chapter III), but it is also possible to create custom workflows. Read about workflows here https://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.

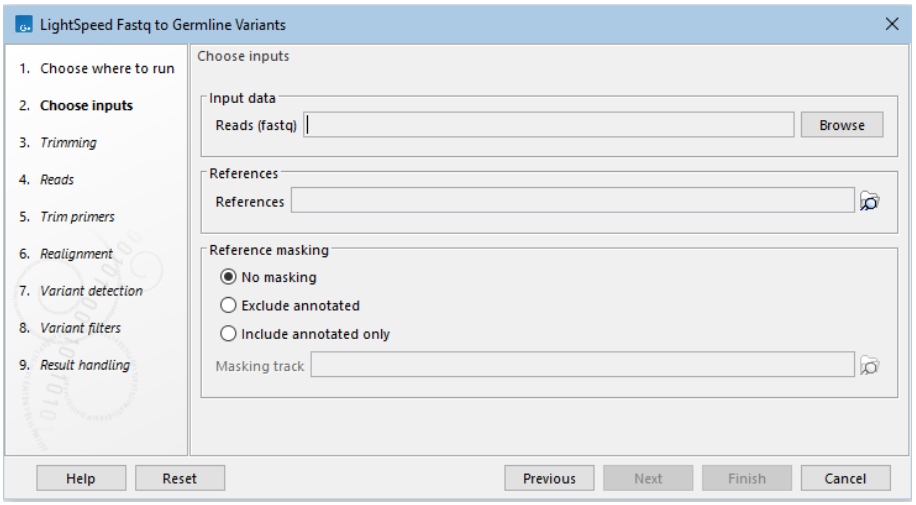

Figure 3.1: *Input fastq files and references, and, optionally, a track for reference masking.*

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

- **Quality 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: https://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.
  - **Only quality trim from 3' end** When checked, only the 3' ends of reads are trimmed for low quality nucleotides. This option is recommended for analyses where deduplication is included, because deduplication relies on the 5' positions of R1 and R2.

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

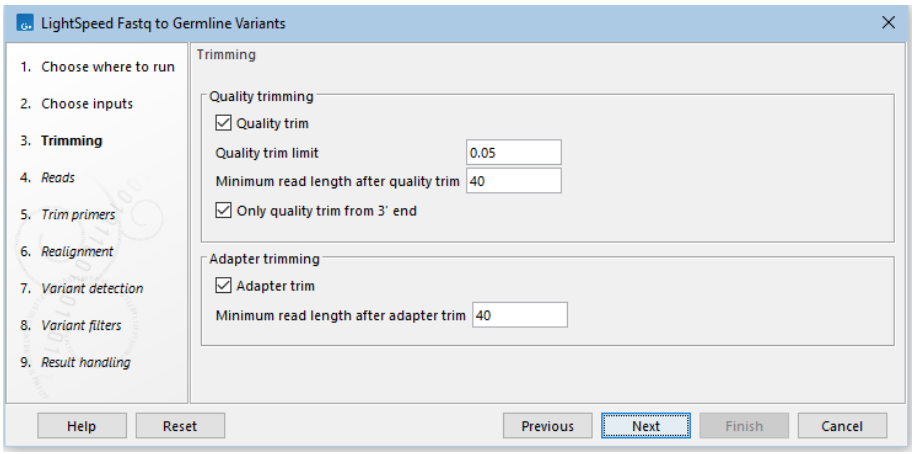

Figure 3.2: *Options for trimming.*

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

- **UMI read structure**
  - **UMI preset** Specify where the UMI is placed in the reads. If a preset for a specific panel type is selected, the UMI and common sequence position and length are automatically adjusted to match the panel design.
    - **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.
    - **QIAseq Targeted DNA Pro** Select if you are analyzing data from a QIAseq Targeted DNA Pro panel.
    - **QIAseq Multimodal DNA/RNA Library Kit** Select if you are analyzing data generated with the QIAseq Multimodal DNA/RNA Library Kit.
    - **TSO500** Select if you are analyzing data from a TSO500 protocol utilizing duplex UMIs.
    - **Twist UMI** Select if you are analyzing data from a Twist protocol utilizing duplex UMIs.
    - **Roche KAPA UMI** Select if you are analyzing data from a KAPA protocol utilizing duplex UMIs.

    * **Agilent MBC** Select if you are analyzing data from an Agilent protocol utilizing duplex UMIs.
  - **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.
  - **Calculate duplex consensus reads** Check this option to collapse simplex UMI reads sharing a UMI and mapping position, that originate from different strands, to a duplex UMI read.

- **UMI settings**
  - **Retrieve UMI from fastq header** Check this option if the UMI sequence is present in the fastq file read headers instead of being part of the sequencing reads. UMI and common sequence lengths must be specified below when this option is used.
  - **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.
  - **Minimum size of a big UMI group** Define the number of read pairs required for a UMI group to be considered a big UMI group. When variants are called, they are annotated with the number of big UMI groups supporting them.
  - **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI sequence have at most this number of mismatches. If UMIs are present on both read 1 and read 2 (duplex), this threshold operates on the sum of mismatches from the two UMIs.
  - **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.
  - **Keep duplex consensus reads only** For duplex UMI protocols, check this option if only duplex reads should be retained. Simplex reads will be discarded.

- **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 3.4):

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

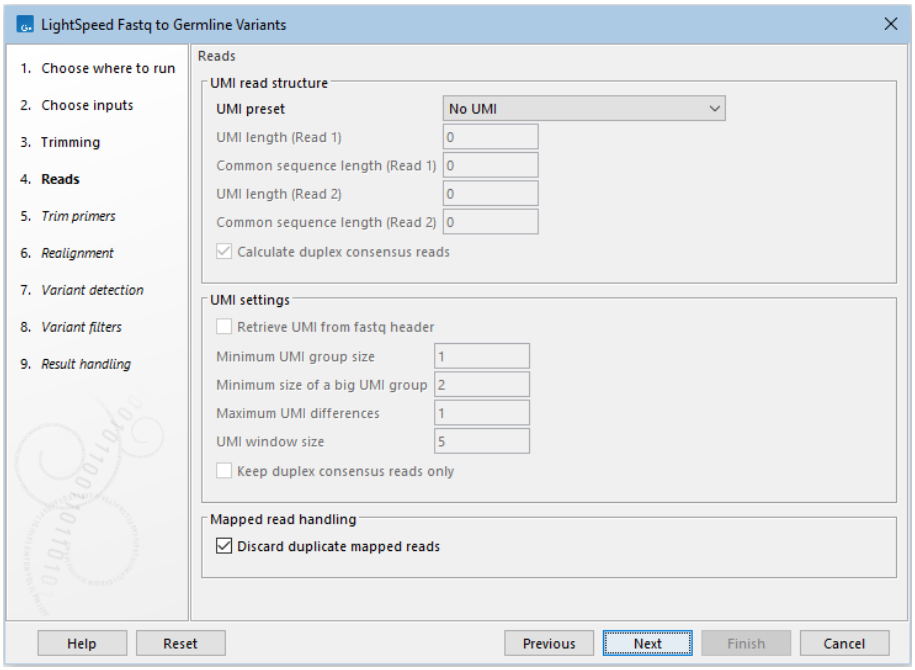

Figure 3.3: *Options for UMI and duplicate reads.*

  - **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 an option is available for realignment:

- **Realignment**
  - **Optimize for targeted data** When enabled, realignment can process complex regions with high coverage that may be present in targeted data. Enabling can cause an increase in processing time, and is not recommended for WGS data.

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

- **Variant detection**
  - **Ploidy** The expected maximum number of different alleles.

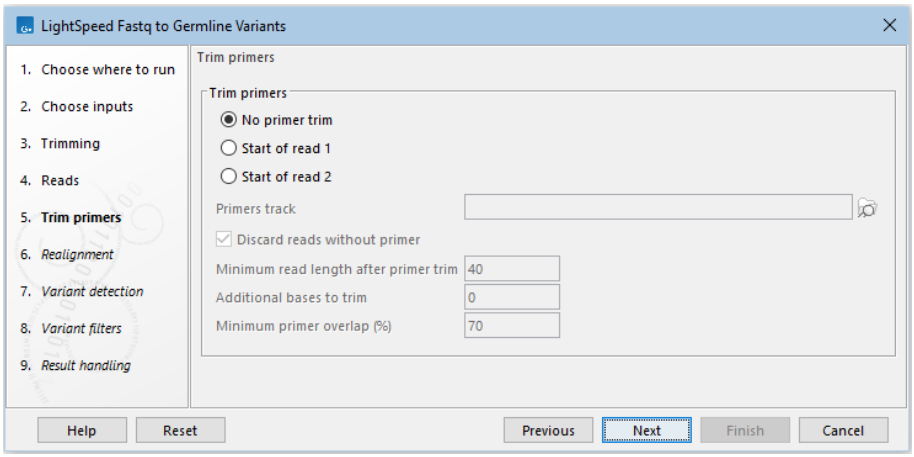

Figure 3.4: *Options for primer trimming.*

  - **Restrict calling to target regions** Optional: A track providing the regions to be inspected when calling variants. When no track is provided, all positions in the mapping are considered.
  - **Ignore non-specific matches** When enabled, reads that map to more then one genomic position equally well are not considered when calling variants.
  - **Sensitivity and precision** Set the variant calling performance priority. A variant with a frequency that deviates from the expected frequency, given the specified ploidy (e.g. 50% or 100% for a ploidy of 2), is less likely to a be real variant and more likely to be caused by artifacts. The Sensitivity and precision option controls the degree to which variants are disqualified based on their observed frequency relative to the expected frequency. **Higher sensitivity** disqualifies fewer variants, **Higher precision** disqualifies more variants and **Balanced** is the intermediate setting.

- **Structural variant detection**

  - **Lenient inversion detection** When enabled, inversions with read support in only one direction at each breakpoint can be called. Enabling this option is recommended when analysing targeted data, but can increase processing time and can result in detection of more false positive inversions.

Next, options are available for variant filtering (figure 3.6):

- **Variant filters**

  - **Minimum average quality** The minimum average quality of detected variants.
  - **Minimum QUAL** The minimum required QUAL score for detected variants. The QUAL score reflects the likelihood of the variant being a real variant.
  - **Minimum allele count** The minimum allele count for detected variants.

    When set to 1, special handling is applied to variants that are only supported by 1 read and have a coverage of 1 or 2. These variants are not assessed for significance, but are reported when the following criteria are met:

    * The average quality score is 35 or higher.

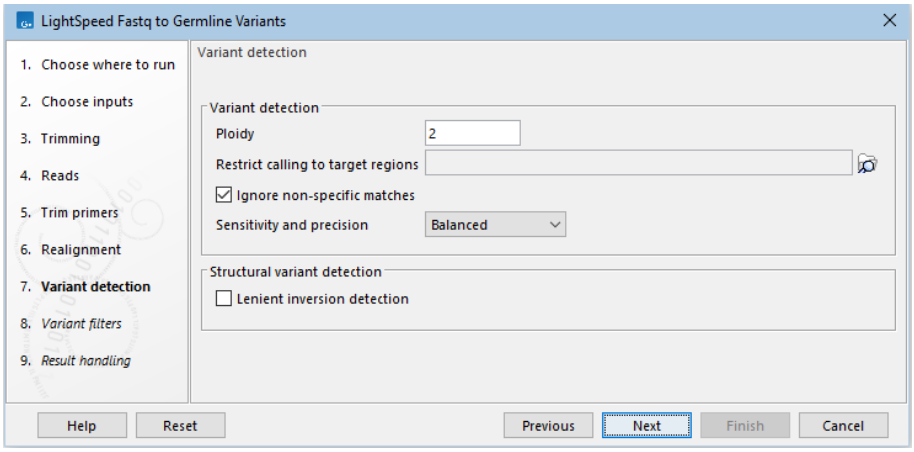

Figure 3.5: *Options for variant detection.*

  * The site has no overlapping non-specific reads.

  Further filtering of these variants may be necessary, such as excluding variants not found in databases of common variants.

- **Minimum frequency (%)** The minimum frequency for detected variants.
- **Strand balance threshold** Threshold for a strand balance score for variants. The score is calculated as 1 - (p-value from binomial test given forward count, count, and forward count/coverage). Allowed range: 0.9 - 1.

- **STR annotations and filter** Annotate and filter variants that change the number of repeats in short tandem repeat (STR) regions, excluding homopolymers.
  - **Minimum repeat count** The minimum number of repeats, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele is called "AGAGAGAGAG", the repeat count is 4.
  - **Minimum repeat region length** The minimum length of all repeats combined, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele "AGAGAGAGAG" is called, the repeat region length is 8.
  - **Maximum repeat element length** The maximum length of a repeat unit, that an allele can have to be annotated as an STR variant. For example, for the dinucleotide repeat "AGAGAGAG", the repeat unit length is 2.
  - **STR filter** Enable to remove variants annotated as STR "Yes" from the results, provided that they are below the frequency threshold defined under "Remove STR variants with frequency below".
  - **Remove STR variants with frequency below** Specify the frequency threshold below which STR variants will be removed from the results.

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.

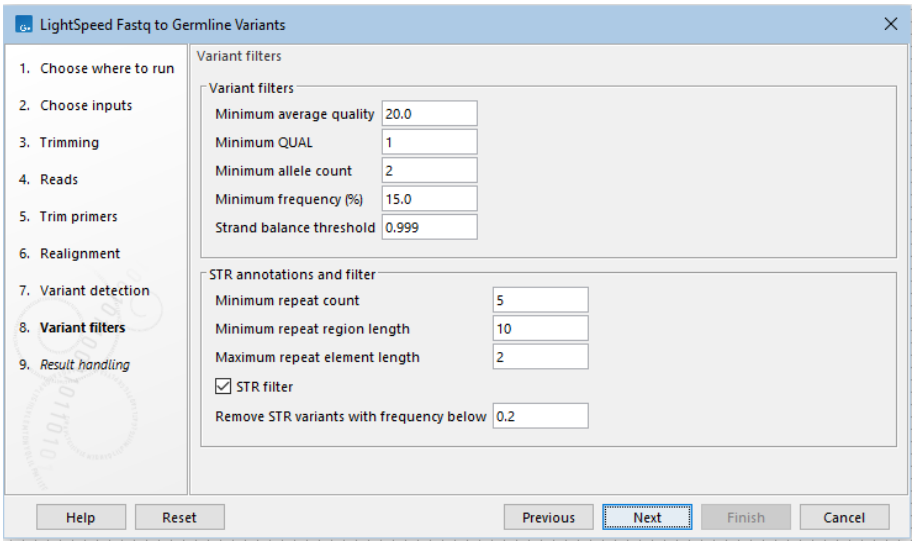

Figure 3.6: *Options for variant filtering.*

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.
- **Breakpoints** The breakpoints that form the basis for structural variant detection (see section 2.5).
- **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.
- **Unmapped reads** Sequence lists containing reads that could not be mapped, one list for intact pairs and one for single reads where the mate could not be mapped.

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 (see section 2.12). Template workflows for LightSpeed analyses are available (see chapter III), but it is also possible to create custom workflows. Read about workflows here `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html`.

To run the somatic LightSpeed tool go to:

Tools | **LightSpeed** ( ) | **LightSpeed Fastq to 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, specify fastq files and a reference sequence (figure 3.7):

- **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.

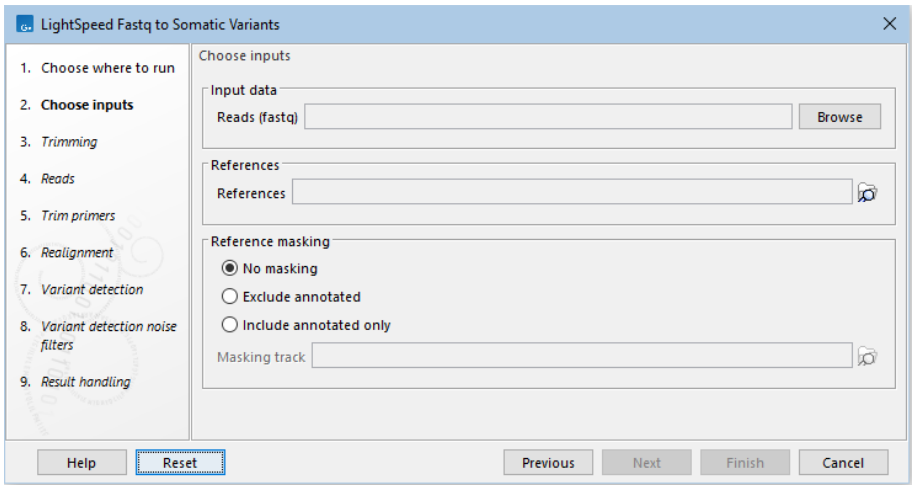

Figure 3.7: *Input fastq files and references, and, optionally, a track for reference masking.*

Next, options are available for trimming (figure 3.8):

- **Quality 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: https://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.
  - **Only quality trim from 3' end** When checked, only the 3' ends of reads are trimmed for low quality nucleotides. This option is recommended for analyses where deduplication is included, because deduplication relies on the 5' positions of R1 and R2.

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

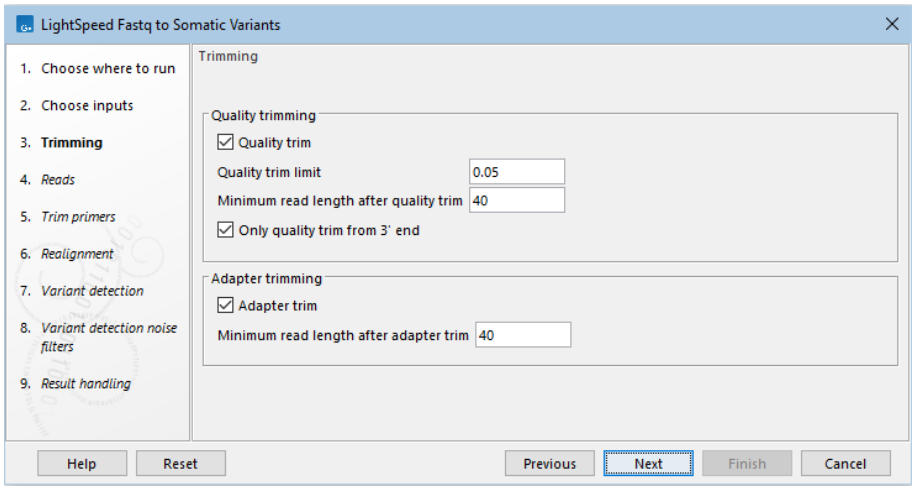

Figure 3.8: *Options for trimming.*

Next, options are available for UMI and duplicate reads (figure 3.9):

- **UMI read structure**
  - **UMI preset** Specify where the UMI is placed in the reads. If a preset for a specific panel type is selected, the UMI and common sequence position and length are automatically adjusted to match the panel design.
    - **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.
    - **QIAseq Targeted DNA Pro** Select if you are analyzing data from a QIAseq Targeted DNA Pro panel.
    - **QIAseq Multimodal DNA/RNA Library Kit** Select if you are analyzing data generated with the QIAseq Multimodal DNA/RNA Library Kit.
    - **TSO500** Select if you are analyzing data from a TSO500 protocol utilizing duplex UMIs.
    - **Twist UMI** Select if you are analyzing data from a Twist protocol utilizing duplex UMIs.

    - **Roche KAPA UMI** Select if you are analyzing data from a KAPA protocol utilizing duplex UMIs.
    - **Agilent MBC** Select if you are analyzing data from an Agilent protocol utilizing duplex UMIs.
  - **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.
  - **Calculate duplex consensus reads** Check this option to collapse simplex UMI reads sharing a UMI and mapping position, that originate from different strands, to a duplex UMI read.

- **UMI settings**
  - **Retrieve UMI from fastq header** Check this option if the UMI sequence is present in the fastq file read headers instead of being part of the sequencing reads. UMI and common sequence lengths must be specified below when this option is used.
  - **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.
  - **Minimum size of a big UMI group** Define the number of read pairs required for a UMI group to be considered a big UMI group. When variants are called, they are annotated with the number of big UMI groups supporting them.
  - **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI sequence have at most this number of mismatches. If UMIs are present on both read 1 and read 2 (duplex), this threshold operates on the sum of mismatches from the two UMIs.
  - **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.
  - **Keep duplex consensus reads only** For duplex UMI protocols, check this option if only duplex reads should be retained. Simplex reads will be discarded.

- **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 3.10):

- **Trim primers**
  - **No primer trim** Disable the trim primers step.

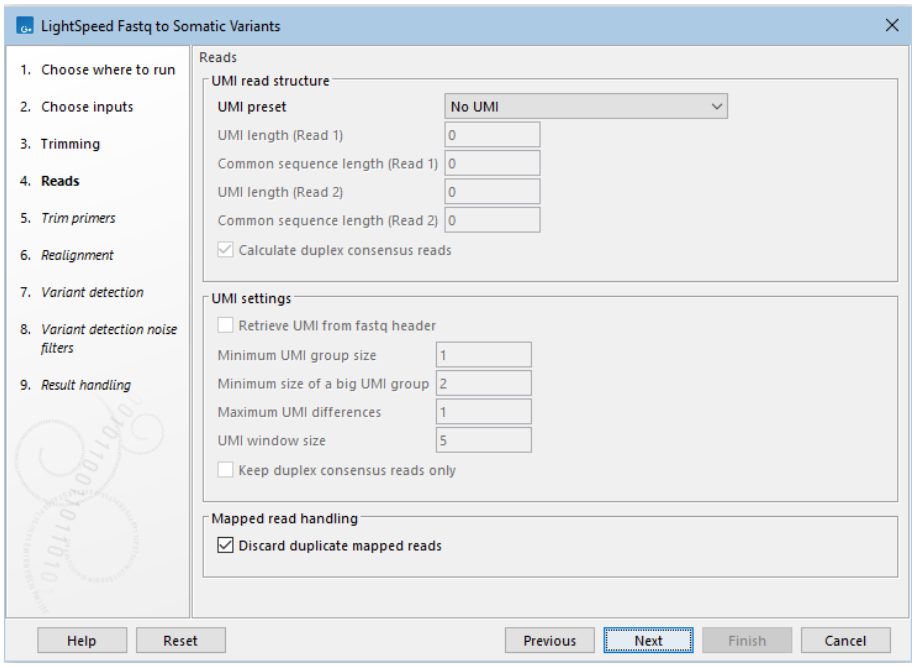

Figure 3.9: *Options for UMI and duplicate reads.*

- **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.

Next an option is available for realignment:

- **Realignment**
  - **Optimize for targeted data** When enabled, realignment can process complex regions with high coverage that may be present in targeted data. Enabling can cause an increase in processing time, and is not recommended for WGS data.

Next, options are available for variant detection (figure 3.11).

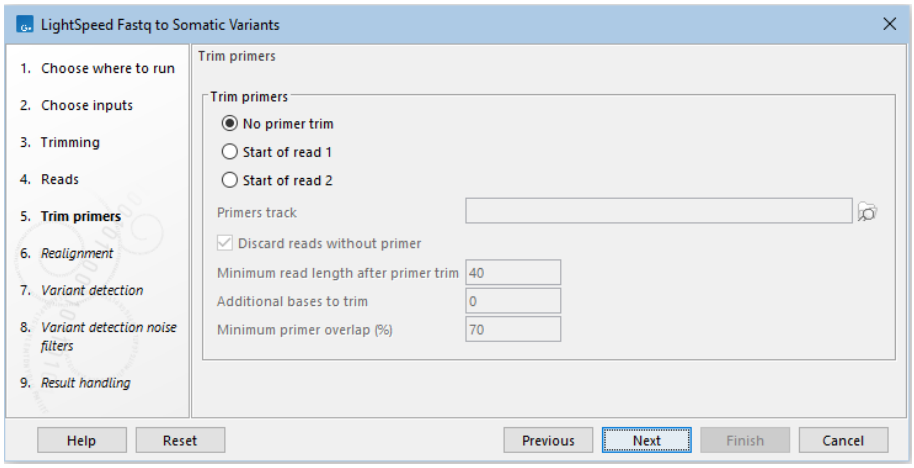

Figure 3.10: *Options for primer trimming.*

- **Variant detection**
  - **Restrict calling to target regions** Optional: A track providing the regions to be inspected when calling variants. When no track is provided, all positions in the mapping are considered.
  - **Ignore non-specific matches** When enabled, reads that map to more then one genomic position equally well are not considered when calling variants.
- **Variant detection general filters**
  - **Minimum frequency (%)** Minimum frequency for detected variants.
  - **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 - 1.0.
  - **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 - 1.0.
- **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.

Next, options are available for variant detection noise filters (figure 3.12).

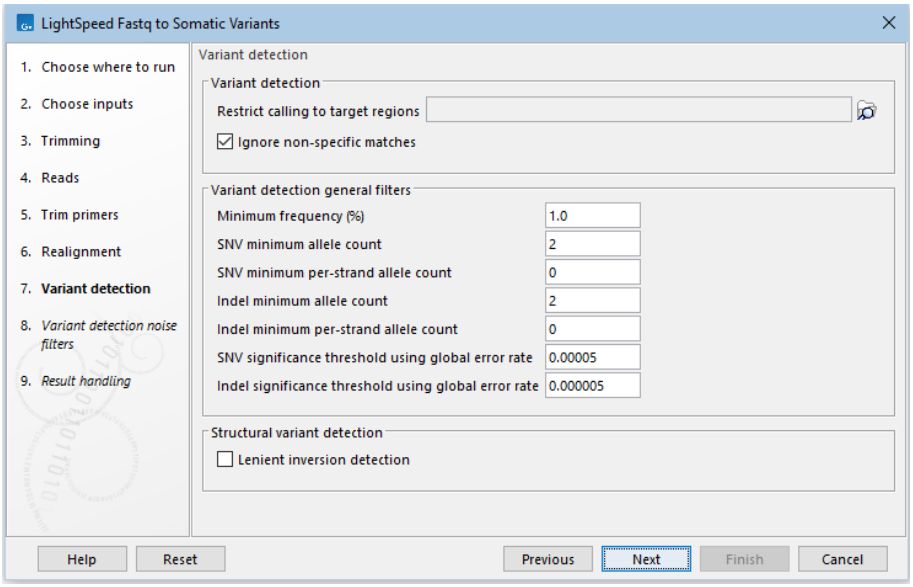

Figure 3.11: *Options for variant detection.*

- **Variant detection noise filters**
  - **Minimum average quality** The minimum average quality of detected variants.
  - **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.
  - **SNV strand balance threshold** Threshold for a strand balance score for SNVs and MNVs. The score is calculated as 1 - (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 - (p-value from binomial test given forward count, count, and forward count/coverage). Allowed range: 0.9 - 1.
- **STR annotations and filter** Annotate and filter variants that change the number of repeats in short tandem repeat (STR) regions, excluding homopolymers.
  - **Minimum repeat count** The minimum number of repeats, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele is called "AGAGAGAGAG", the repeat count is 4.
  - **Minimum repeat region length** The minimum length of all repeats combined, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele "AGAGAGAGAG" is called, the repeat region length is 8.
  - **Maximum repeat element length** The maximum length of a repeat unit, that an allele can have to be annotated as an STR variant. For example, for the dinucleotide repeat "AGAGAGAG", the repeat unit length is 2.

- **STR filter** Enable to remove variants annotated as STR "Yes" from the results, provided that they are below the frequency threshold defined under "Remove STR variants with frequency below".
- **Remove STR variants with frequency below** Specify the frequency threshold below which STR variants will be removed from the results.

For the options under **Variant detection general filters** and **Variant detection noise filters**

- 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.

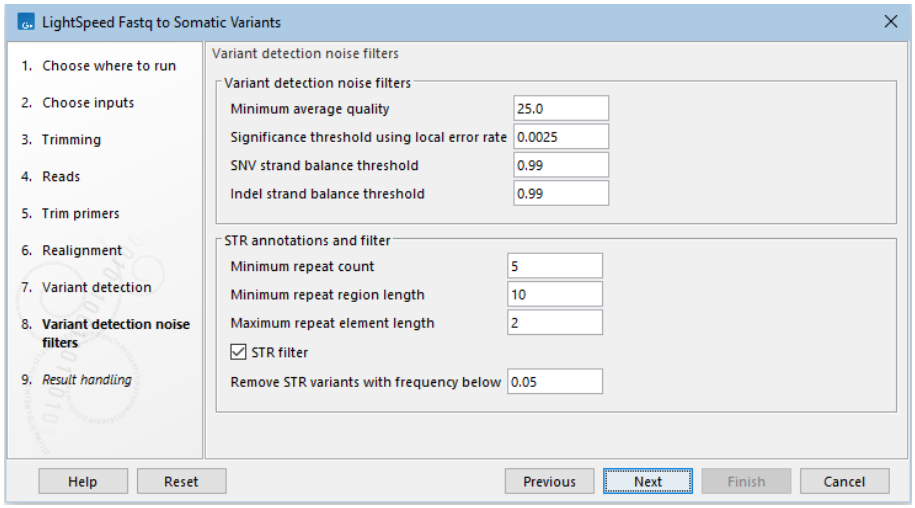

Figure 3.12: *Options for variant detection noise filters.*

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.
- **Unmapped reads** Sequence lists containing reads that could not be mapped, one list for intact pairs and one for single reads where the mate could not be mapped.

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 (see section 2.12).

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 3.13):

- **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 3.14):

- **Quality 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: https://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.

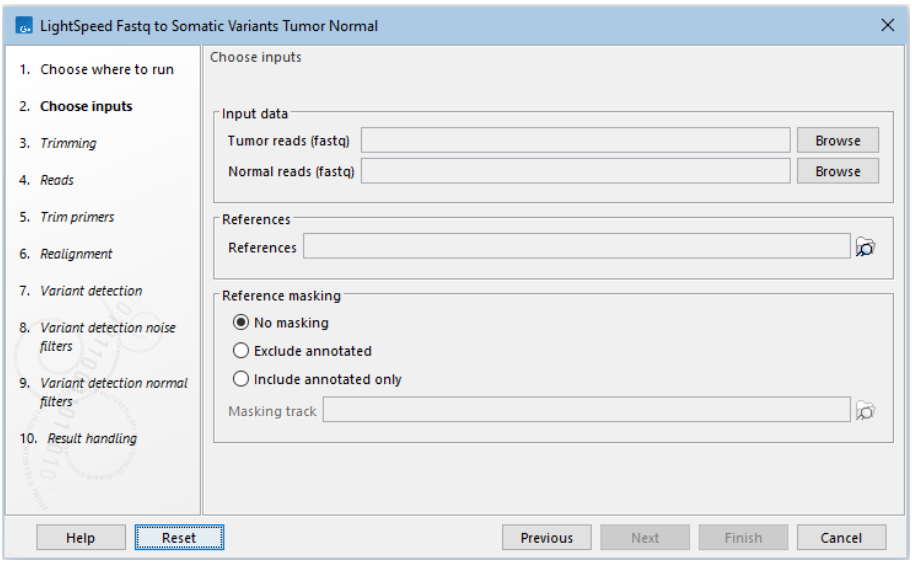

Figure 3.13: *Input fastq files and references, and, optionally, a track for reference masking.*

  - **Only quality trim from 3' end** When checked, only the 3' ends of reads are trimmed for low quality nucleotides. This option is recommended for analyses where deduplication is included, because deduplication relies on the 5' positions of R1 and R2.

- **Adapter trimming**

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

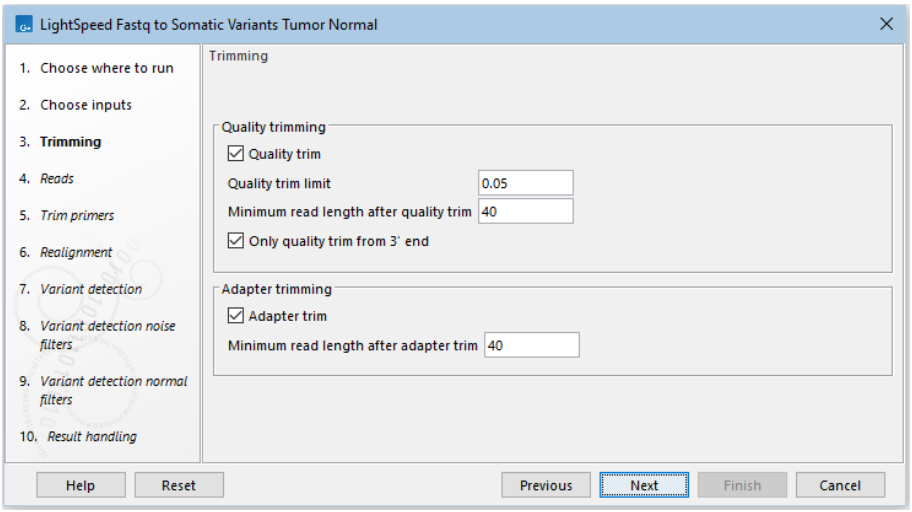

Figure 3.14: *Options for trimming.*

Next, options are available for UMI and duplicate reads (figure 3.15):

- **UMI read structure**

  - **UMI preset** Specify where the UMI is placed in the reads. If a preset for a specific panel type is selected, the UMI and common sequence position and length are automatically adjusted to match the panel design.
    * **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.
    * **QIAseq Targeted DNA Pro** Select if you are analyzing data from a QIAseq Targeted DNA Pro panel.
    * **QIAseq Multimodal DNA/RNA Library Kit** Select if you are analyzing data generated with the QIAseq Multimodal DNA/RNA Library Kit.
    * **TSO500** Select if you are analyzing data from a TSO500 protocol utilizing duplex UMIs.
    * **Twist UMI** Select if you are analyzing data from a Twist protocol utilizing duplex UMIs.
    * **Roche KAPA UMI** Select if you are analyzing data from a KAPA protocol utilizing duplex UMIs.
    * **Agilent MBC** Select if you are analyzing data from an Agilent protocol utilizing duplex UMIs.
  - **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.
  - **Calculate duplex consensus reads** Check this option to collapse simplex UMI reads sharing a UMI and mapping position, that originate from different strands, to a duplex UMI read.

- **UMI settings**
  - **Retrieve UMI from fastq header** Check this option if the UMI sequence is present in the fastq file read headers instead of being part of the sequencing reads. UMI and common sequence lengths must be specified below when this option is used.
  - **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.
  - **Minimum size of a big UMI group** Define the number of read pairs required for a UMI group to be considered a big UMI group. When variants are called, they are annotated with the number of big UMI groups supporting them.
  - **Maximum UMI differences** Add input read pairs to the same UMI group when their UMI sequence have at most this number of mismatches. If UMIs are present on both read 1 and read 2 (duplex), this threshold operates on the sum of mismatches from the two UMIs.

  - **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.
  - **Keep duplex consensus reads only** For duplex UMI protocols, check this option if only duplex reads should be retained. Simplex reads will be discarded.

- **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.

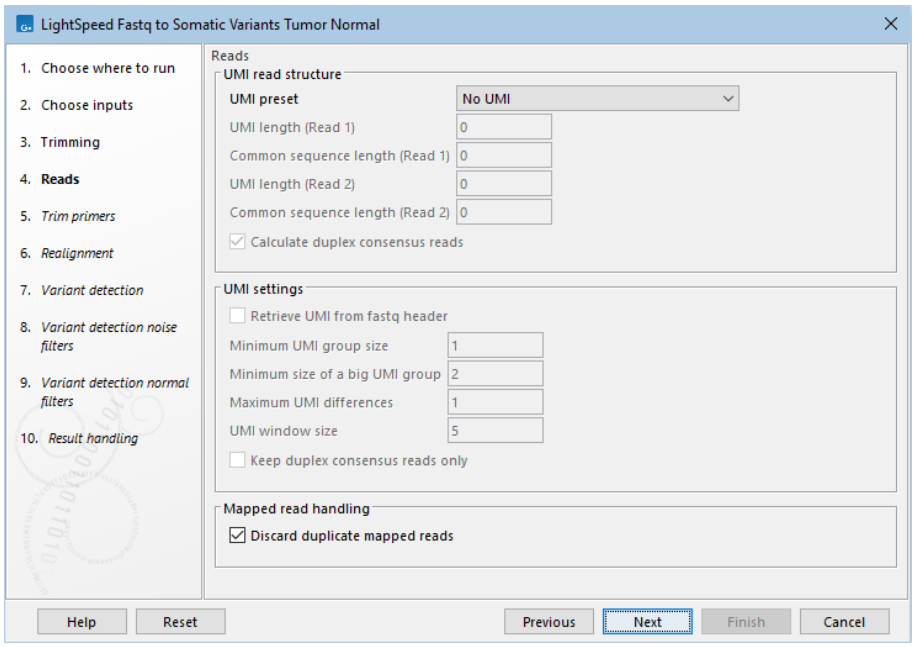

Figure 3.15: *Options for UMI and duplicate reads.*

Next, options are available for primer trimming (figure 3.16):

- **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.

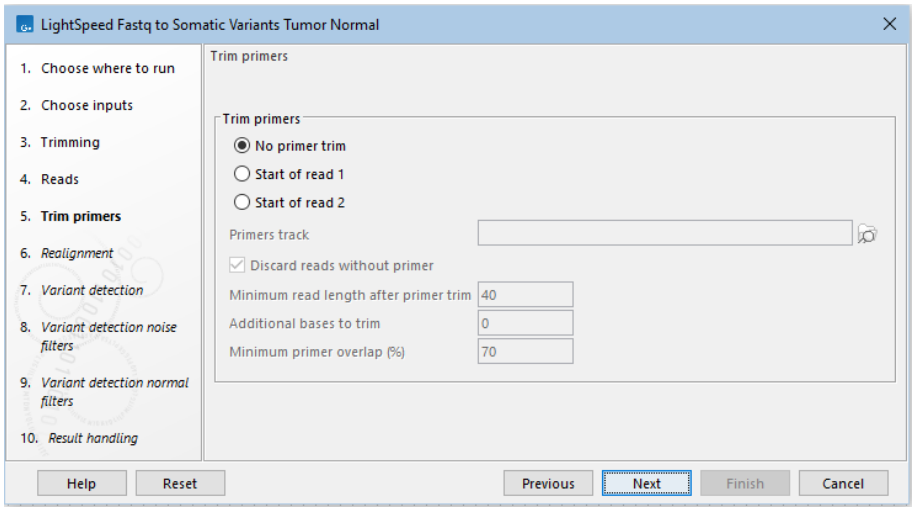

Figure 3.16: *Options for primer trimming.*

Next an option is available for realignment:

- **Realignment**
  - **Optimize for targeted data** When enabled, realignment can process complex regions with high coverage that may be present in targeted data. Enabling can cause an increase in processing time, and is not recommended for WGS data.

Next, options are available for variant detection (figure 3.17).

- **Variant detection**
  - **Restrict calling to target regions** Optional: A track providing the regions to be inspected when calling variants. When no track is provided, all positions in the mapping are considered.
  - **Ignore non-specific matches** When enabled, reads that map to more then one genomic position equally well are not considered when calling variants.
- **Variant detection general filters**, see figure 3.17
  - **Minimum frequency (%)** Minimum frequency for detected variants.
  - **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 - 1.0.
- **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 - 1.0.

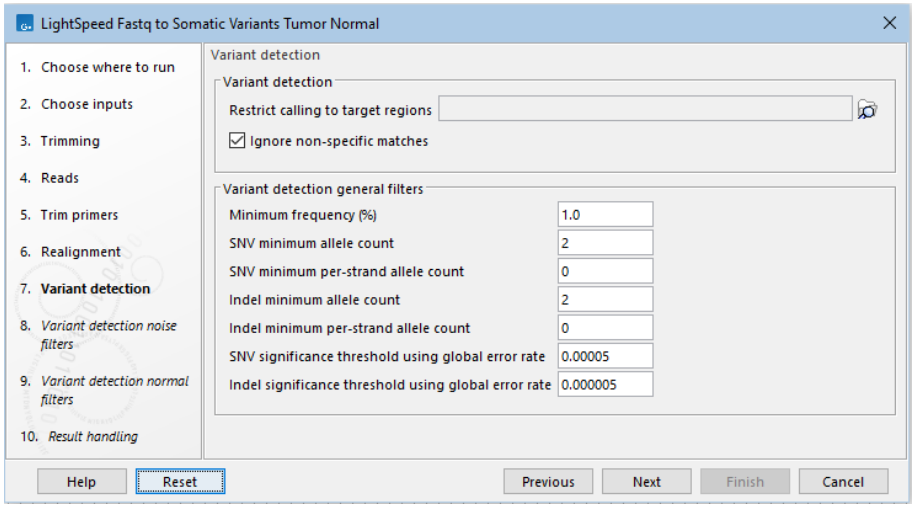

Figure 3.17: *Options for variant detection.*

Next, options are available for variant detection noise filters (figure 3.18).

- **Variant detection noise filters**
  - **Minimum average quality** The minimum average quality of detected variants.
  - **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.
  - **SNV strand balance threshold** Threshold for a strand balance score for SNVs and MNVs. The score is calculated as 1 - (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 - (p-value from binomial test given forward count, count, and forward count/coverage). Allowed range: 0.9 - 1.
- **STR annotations and filter** Annotate and filter variants that change the number of repeats in short tandem repeat (STR) regions, excluding homopolymers.
  - **Minimum repeat count** The minimum number of repeats, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference

allele "AGAGAGAG" is called, and a low frequent stutter insertion allele is called "AGAGAGAGAG", the repeat count is 4.

- **Minimum repeat region length** The minimum length of all repeats combined, excluding the variant, that an allele must have to be annotated as an STR variant. For example if a reference allele "AGAGAGAG" is called, and a low frequent stutter insertion allele "AGAGAGAGAG" is called, the repeat region length is 8.
- **Maximum repeat element length** The maximum length of a repeat unit, that an allele can have to be annotated as an STR variant. For example, for the dinucleotide repeat "AGAGAGAG", the repeat unit length is 2.
- **STR filter** Enable to remove variants annotated as STR "Yes" from the results, provided that they are below the frequency threshold defined under "Remove STR variants with frequency below".
- **Remove STR variants with frequency below** Specify the frequency threshold below which STR variants will be removed from the results.

For the options under **Variant detection general filters** and **Variant detection noise filters**

- 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.

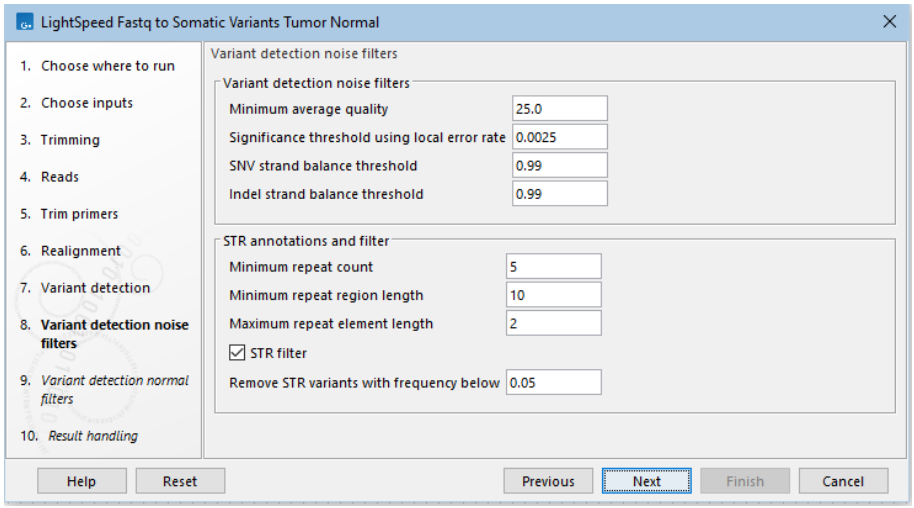

Figure 3.18: *Options for variant detection noise filters.*

In the following wizard step, specify the maximum variant count, frequency and significance in the normal read mapping (figure 3.19).

- **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 - 1.0.

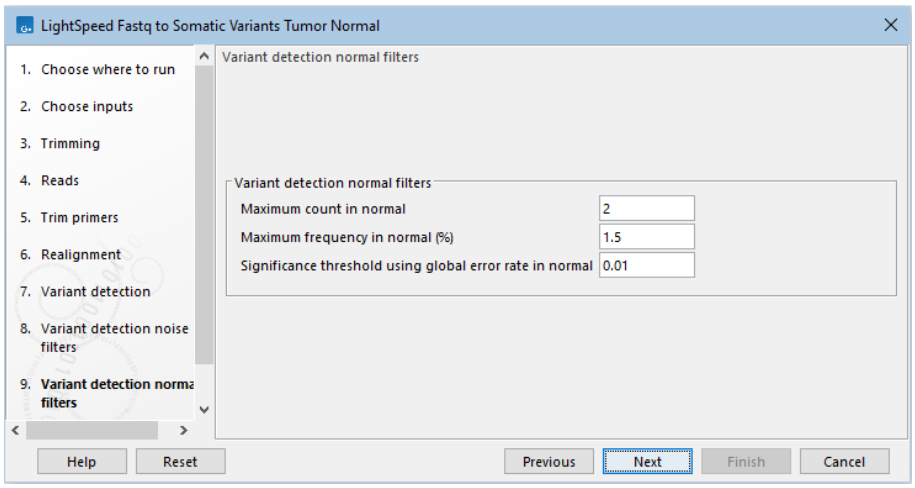

Figure 3.19: *Options for variant detection normal filters.*

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.
- **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 from LightSpeed Fastq to Variants tools

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

The following terms are used in many sections of the report:

- **Specific read pairs** Read pairs where one best match for mapping was identified.
- **Non-specific read pairs** Reads that map equally well to more than one genomic position.
- **Proper read pairs** Read pairs where the distance between read 1 and read 2 are within the expected range for a pair.
- **Broken read pairs** Read pairs where the distance between read 1 and read 2 exceed the expected distance for a read pair and read pairs where only one of the reads were mapped.

Summary

- **Input read pairs** Total number of read pairs in the fastq files.
- **Read pairs discarded 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 discarded.
- **Read pairs trimmed by quality trimming** Read pairs that have been trimmed and are longer than "Minimum read length after quality trim".
- **Read pairs discarded 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 discarded.
- **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.
- **Read pairs remaining after trimming** Read pairs remaining after quality and adapter trimmming. These are the read pairs that are mapped.
- **Unmapped read pairs** Read pairs that did not map to the reference.
- **Mapped read pairs** The total number of mapped read pairs including specific, non-specific and broken pairs.
- **Proper read pairs** Read pairs that are mapped as pairs. The percentage is calculated relative to "Mapped read pairs".
- **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. The percentage is calculated relative to "Mapped read pairs".
- **Specific proper read pairs** Read pairs that are mapped as pairs and are specific. The percentage is calculated relative to "Mapped read pairs".
- **Non-specific proper read pairs** Read pairs that are mapped as pairs, but are non-specific. The percentage is calculated relative to "Mapped read pairs".

- **Average insert size** Average insert size calculated from specific proper read pairs.
- **Median insert size** Median insert size calculated from specific proper read pairs.
- **Read pairs after deduplication** Read pairs after deduplication.
- **UMI read pairs** Read pairs after UMI grouping.
- **Singleton UMI read pairs** UMI read pairs generated from only one read pair. The percentage is calculated relative to "UMI read pairs".
- **Simplex UMI read pairs** UMI read pairs where input reads all originate from the same strand. Singleton UMI read pairs are a subset of the simplex UMI read pairs. The percentage is calculated relative to "UMI read pairs"
- **Duplex UMI read pairs** UMI read pairs that are based on input reads from both strands. The percentage is calculated relative to "UMI read pairs"
- **Average number of reads per UMI** The average number of read pairs per UMI read pair.
- **Median number of reads per UMI** The median number of read pairs per UMI read pair.
- **Average number of reads per duplex UMI** The average number of read pairs per duplex UMI read pair.
- **Median number of reads per duplex UMI** The median number of read pairs per duplex UMI read pair.
- **Read pairs after primer trimming** Read pairs remaining after primer trimming.

Input read QC This section contains information about the input reads before quality and adapter trimming. Full descriptions of the per-sequence plots are available at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Per_sequence_analysis.html and full descriptions of the per-sequence plots are available at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Per_base_analysis.html.

Per-sequence analysis

- **Lengths distribution** Plot showing the distribution of R1 and R2 read lengths.
- **GC-content** Plot showing the distribution of GC-content in R1 and R2 reads.
- **Ambiguous base-content** Plot showing the distribution of ambiguous base-content in R1 and R2 reads.
- **Quality distribution** Plot showing the distribution of average quality per read for R1 and R2 reads.
- **Reads passing average quality thresholds** Table providing the percentage of R1 and R2 reads with average quality above 25, 30 and 35.

Per-base analysis

- **Coverage** Plot showing the coverage for individual base positions of R1 and R2.
- **Nucleotide contributions** Plots for R1 and R2 showing the nucleotide contributions per position.
- **GC-content** Plot showing the GC content in R1 and R2 reads per position.
- **Ambiguous base-content** Plot showing the ambiguous base-content in R1 and R2 reads per position.
- **Quality distribution** Plots for R1 and R2 individually and combined showing the quality per position.

Quality trimming

- **Input read pairs** Total number of read pairs in the fastq files.
- **Read pairs discarded 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 discarded.
- **Read pairs trimmed by quality trimming** Read pairs that have been trimmed and are longer than "Minimum read length after quality trim".
- **R1 reads trimmed by quality trimming** Number of R1 reads trimmed by quality trimming.
- **R2 reads trimmed by quality trimming** Number of R2 reads trimmed by quality trimming.
- **Average read length before quality trimming** Average read length of the raw reads in the fastq files.
- **Average read length after quality trimming** Average read length after quality trimming. This read length may be longer than **Average read length before quality trimming** 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 3.20).

Adapter trimming

- **Input read pairs** Total number of read pairs in the fastq files.
- **Read pairs discarded 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 discarded.
- **Read pairs trimmed by adapter trimming** Read pairs that have been trimmed and are longer than "Minimum read length after adapter trim".

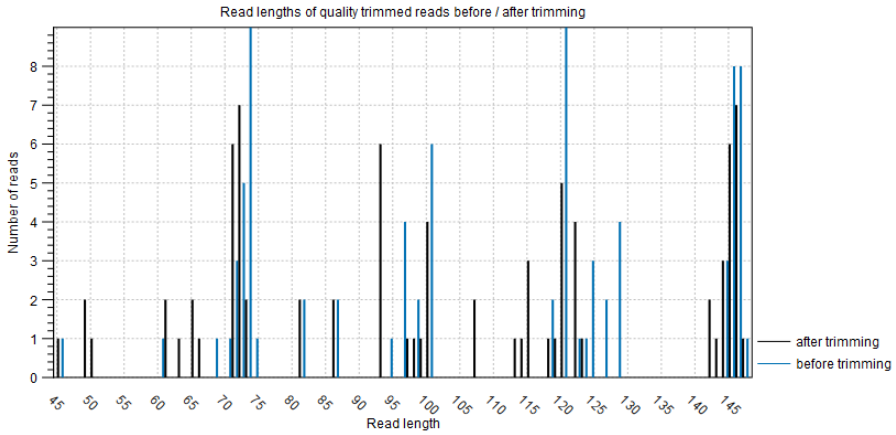

Figure 3.20: *The number and length of quality trimmed reads before and after quality trimming.*

- **R1 reads trimmed by adapter trimming** Trimmed R1 reads that are longer than "Minimum read length after adapter trim".
- **R2 reads trimmed by adapter trimming** Trimmed R2 reads that are longer than "Minimum read length after adapter trim".
- **Average read length before adapter trimming** Average length of the reads before adapter trimming. If quality trimming was enabled, read length after quality trim is given.
- **Average read length after adapter trimming** Average read length after adapter trimming. This read length may be longer than **Average read length before adapter trimming** because short reads can have been discarded.
- **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 funtion of read length before and after adapter trimming (figure 3.21).

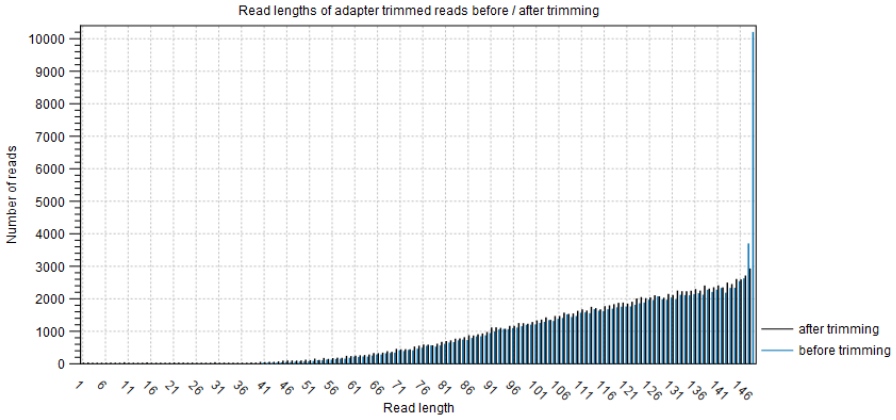

Figure 3.21: *The number of reads as a function of read length before and after adapter trimming.*

The plot **Lengths of trimmed adapters** shows the number and lengths of trimmed adapter sequences (figure 3.22).

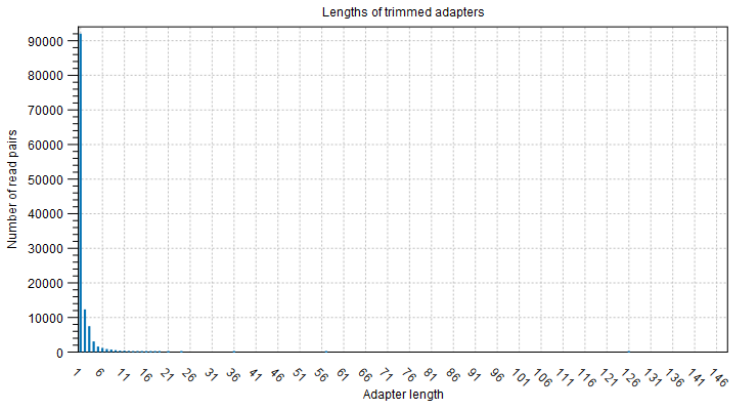

Figure 3.22: *The number and length of trimmed adapter sequences.*

Mapping statistics

- **References** The number of sequences in the reference genome.
- **Input read pairs** Total number of read pairs in the fastq files.
- **Read pairs remaining after trimming** The number of read pairs left after trimming.
- **Unmapped read pairs** The number of read pairs that could not be mapped to the reference.
- **Mapped read pairs** The number of mapped read pairs including specific, non-specific and broken read pairs.
- **Mapped proper read pairs** Read pairs that are mapped as pairs. The percentage is calculated relative to "Mapped read pairs".
- **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. The percentage is calculated relative to "Mapped read pairs".
- **Mapped specific proper read pairs** Read pairs that are mapped as pairs and are specific. The percentage is calculated relative to "Mapped read pairs".
- **Mapped non-specific proper read pairs** Read pairs that are mapped as pairs, but are non-specific. The percentage is calculated relative to "Mapped read pairs".

Insert size distribution Plot showing the distribution of insert sizes in specific proper read pairs. The insert is defined as the distance between the 5' ends of R1 and R2. If reads are quality trimmed from the 5' end or are trimmed for UMI and common sequence, the removed bases are not included when calculating the insert size.

Deduplication

- **Mapped read pairs** The number of mapped read pairs including specific, non-specific and broken read pairs before deduplication.

- **Duplicate read pairs** Read pairs considered PCR duplicates.
- **Read pairs after deduplication** The number of mapped read pairs including specific, non-specific and broken read pairs.
- **Proper read pairs after deduplication** Read pairs that are mapped as pairs. The percentage is calculated relative to "Read pairs after deduplication".
- **Broken read pairs after deduplication** 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. The percentage is calculated relative to "Read pairs after deduplication".
- **Specific proper read pairs after deduplication** Read pairs that are mapped as pairs and are specific. The percentage is calculated relative to "Read pairs after deduplication".
- **Non-specific proper read pairs after deduplication** Read pairs that are mapped as pairs, but are non-specific. The percentage is calculated relative to "Read pairs after deduplication".

UMI

- **Mapped read pairs** The number of mapped read pairs including specific, non-specific and broken read pairs before UMI grouping.
- **UMI read pairs** Read pairs after UMI grouping.
- **Singleton UMI read pairs** UMI read pairs generated from only one read pair. The percentage is calculated relative to "UMI read pairs".
- **Simplex UMI read pairs** UMI read pairs where input reads all originate from the same strand. Singleton UMI read pairs are a subset of the simplex UMI read pairs. The percentage is calculated relative to "UMI read pairs".
- **Duplex UMI read pairs** UMI read pairs that are based on input reads from both strands. The percentage is calculated relative to "UMI read pairs".
- **Proper UMI read pairs** UMI read pairs that are mapped as pairs. The percentage is calculated relative to "UMI read pairs".
- **Broken UMI read pairs** UMI 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. The percentage is calculated relative to "UMI read pairs".
- **Specific proper UMI read pairs** UMI read pairs that are mapped as pairs and are specific. The percentage is calculated relative to "UMI read pairs".
- **Non-specific proper UMI read pairs** UMI read pairs that are mapped as pairs, but are non-specific. The percentage is calculated relative to "UMI read pairs".
- **Average number of reads per UMI** The average number of read pairs per UMI read pair.
- **Median number of reads per UMI** The median number of read pairs per UMI read pair.

- **Average number of reads per duplex UMI** The average number of read pairs per duplex UMI read pair.
- **Median number of reads per duplex UMI** The median number of read pairs per duplex UMI read pair.

When calculating average and median number of read pairs per UMI, broken pairs are not included.

The plot **Reads by group size** shows the number of proper input read pairs distributed by the size of the UMI groups that they have been grouped to. When running duplex UMI analyses, input read pairs that are grouped to duplex UMI read pairs are not represented in this plot.

The plot **Groups by group size** shows the number of UMI groups distributed by the UMI group sizes. When running duplex UMI analyses, duplex groups are not represented in this plot.

The plot **Reads by group size (duplex)** shows the number of proper input read pairs distributed by the size of the duplex UMI groups that they have been grouped to. If reads are not assigned to a duplex UMI group, they are not represented in this plot.

The plot **Groups by group size (duplex)** shows the number of duplex UMI groups distributed by the duplex UMI group sizes.

Realignment

- **Realigned regions** The number of regions that were subjected to local realignment, e.g. regions with long unaligned ends.
- **Combined length of realigned regions** The combined length of the locally realigned regions.
- **Reassembled regions** The number of regions that were subjected to reassembly, e.g. regions with significant unaligned end breakpoints.
- **Combined length of reassembled regions** The combined length of the reassembled regions.

Primer trimming

- **Mapped read pairs** The number of mapped read pairs including specific, non-specific and broken read pairs before UMI grouping.
- **Discard read pairs without primer** An option used when trimming for primer sequence, can be Yes or No.
- **Read pairs without primers** Read pairs that could not be assigned to a primer. If "Discard read pairs without primer" is set to yes, these reads will be discarded.
- **Primer not found** Read pairs not overlapping a primer.
- **Inside alignment** Read pairs overlapping a primer, but the primer is inside the alignment, not at the end.
- **Not enough overlap** Read pairs overlapping a primer, but the overlap is less than required as defined in the wizard step "Minimum primer overlap (%)".

- **Read too short** Read pairs where the remaining sequence of read 1 or read 2 is shorter than the threshold defined in the wizard step "Minimum read length after primer trim".
- **Too many read mismatches** Read pairs with at least 2 mismatches between the overlapping parts of the read and the primer.
- **No primers on chromosome** Read pairs mapped to chromosomes where there are no primers.
- **Read pairs after primer trimming** Read pairs remaining after primer trimming.
- **Proper read pairs after primer trimming** Read pairs that are mapped as pairs. The percentage is calculated relative to "Read pairs after primer trimming".
- **Broken read pairs after primer trimming** 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. The percentage is calculated relative to "Read pairs after primer trimming".
- **Specific proper read pairs after primer trimming** Read pairs that are mapped as pairs and are specific. The percentage is calculated relative to "Read pairs after primer trimming".
- **Non-specific proper read pairs after primer trimming** Read pairs that are mapped as pairs, but are non-specific. The percentage is calculated relative to "Read pairs after primer trimming".

Variant detection

- **Ignored bases due to complex regions** The number of bases where it was not possible to detect variants due to high complexity of the region.
- **Ignored intervals due to complex regions** The number of intervals where it was not possible to detect variants due to high complexity of the region.

3.5 Calculate TMB Score (LightSpeed)

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

It is recommended that target regions with a coverage lower than 100X are discarded before running the tool. To do so, a workflow including the tools Create Mapping Graph and Identify Graph Threshold Area can be used to generate a target region file only containing target regions with at least 100X coverage (see figure 3.23).

The **Calculate TMB Score (LightSpeed)** tool selects variants for analysis based on the following criteria:

- Variants must be SNVs.
- Variants must be within target regions.
- Variants must be within exonic regions.

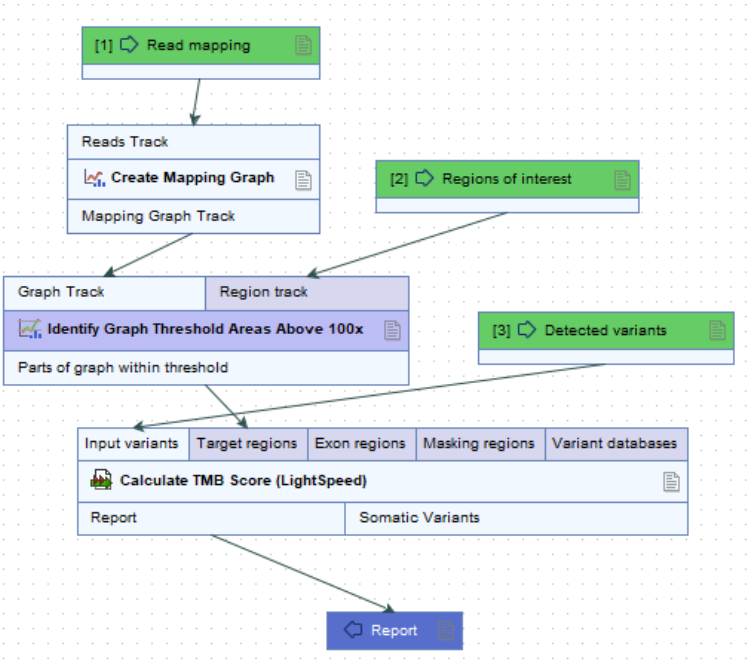

Figure 3.23: *Workflow to discard low coverage target regions.*

- Variants must be outside masking regions.

Resulting SNVs are filtered using quality, germline and non-synonymous filters before calculating the TMB score as the number of somatic variants multiplied by 1 million bases and divided by the length of the target regions minus the length of masking regions.

To run **Calculate TMB Score (LightSpeed)**, go to:

Tools | **LightSpeed** ( ) | **Calculate TMB Score (LightSpeed)** ( )

The tool takes a variant track as input which is provided in the first dialog (figure 3.24):

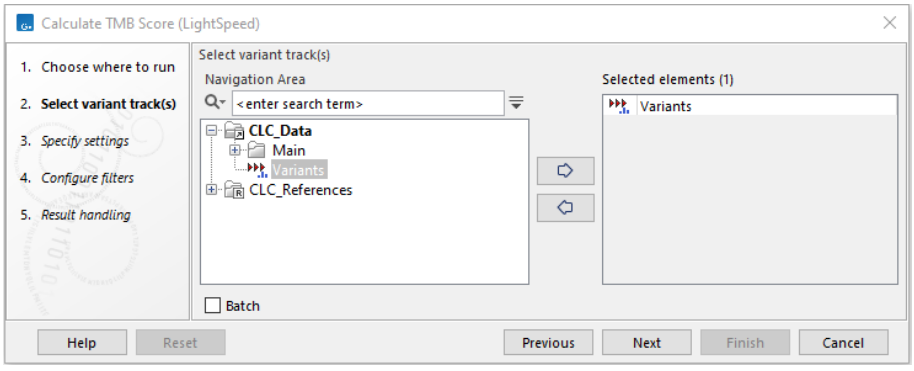

Figure 3.24: *Select a variant track.*

In the next dialog, tracks relevant to the analysis are specified (figure 3.25):

- **Target regions** A track containing the target regions.
- **Exon regions** An mRNA track containing exons.

- **Masking regions** Regions that should not be considered.

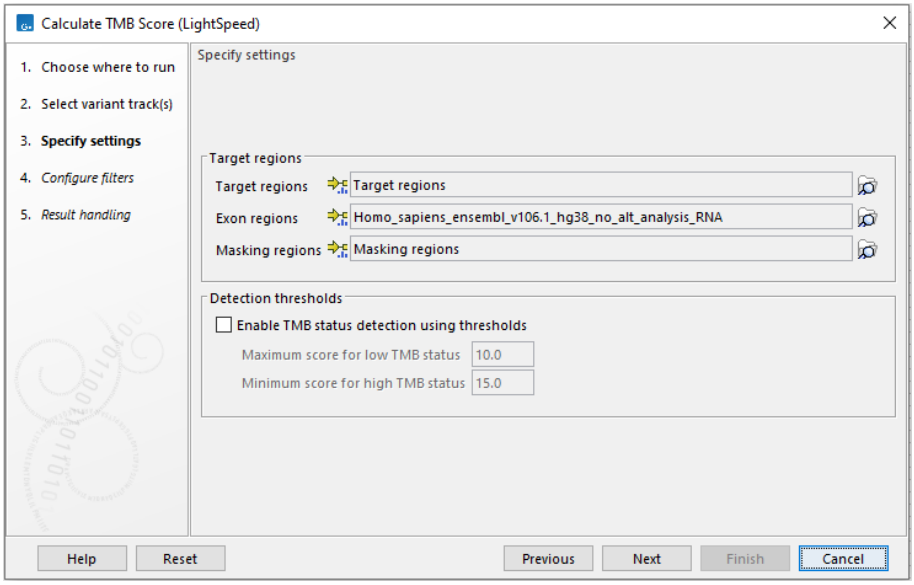

Figure 3.25: *Specify tracks and parameters for calculating a TMB score.*

It is also possible to calculate a TMB status based on thresholds for a low and a high TMB status. This will appear as an additional item in the TMB report. The default values of 10 and 15 have been chosen based on internal benchmark analyses. Samples with a score between the low and high threshold, will be assigned TMB status intermediate. Given the lack of standardization of TMB scoring methods and the heterogeneity of TMB across tumor types, we recommend manually specifying values that are suitable for the analyzed sample and tumor type.

In the next dialog (figure 3.26), variant filters can be configured:

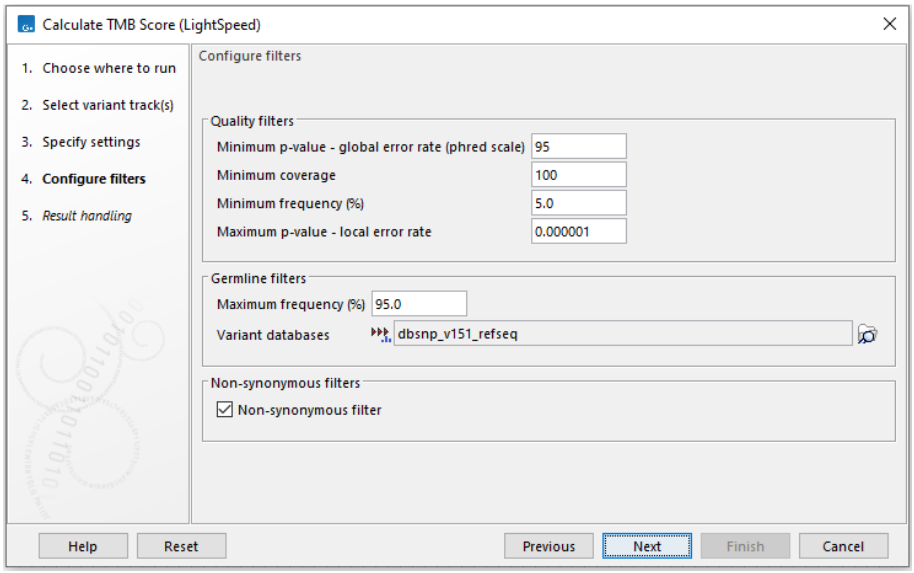

Figure 3.26: *Configure variant filters.*

- **Quality filters**

  - **Minimum p-value - global error rate (phred scale)** Only variants that have achieved this level of significance when considering the global error rate are considered.
  - **Minimum coverage** Only variants in regions covered by at least this many reads are considered.
  - **Minimum frequency (%)** Only variants with a frequency above this value are considered.
  - **Maximum p-value - local error rate** Only variants that have achieved this level of significance when considering the local error rate are considered.

- **Germline filters**
  - **Maximum frequency (%)** Only variants with a frequency equal to or lower than the specified value will be considered. Variants with a frequency above this value are considered germline.
  - **Variant databases** Specify a variant database such as dbSNP. Although dbSNP is thought to contain many erroneous calls, these may still be useful for removing variants that are not somatic, e.g. if they arise from common sequencing artifacts. It is mandatory to provide a variant database of known germline variants.

- **Non-synonymous filters** Only amino acid changing variants are kept and considered for the TMB score calculation.

The default quality filter thresholds have been determined based on internal benchmarking using samples from the Tumor Mutational Burden Harmonization Project led by Friends of Cancer Research [Vega et al., 2021].

The tool outputs a track of filtered somatic variants, i.e., the variants that remained after the filtering and that were included in the TMB score calculation. However, the main output is a report that includes filtering statistics and the calculated TMB score. It will also include a TMB status if the option was enabled (as shown in figure 3.27).

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

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

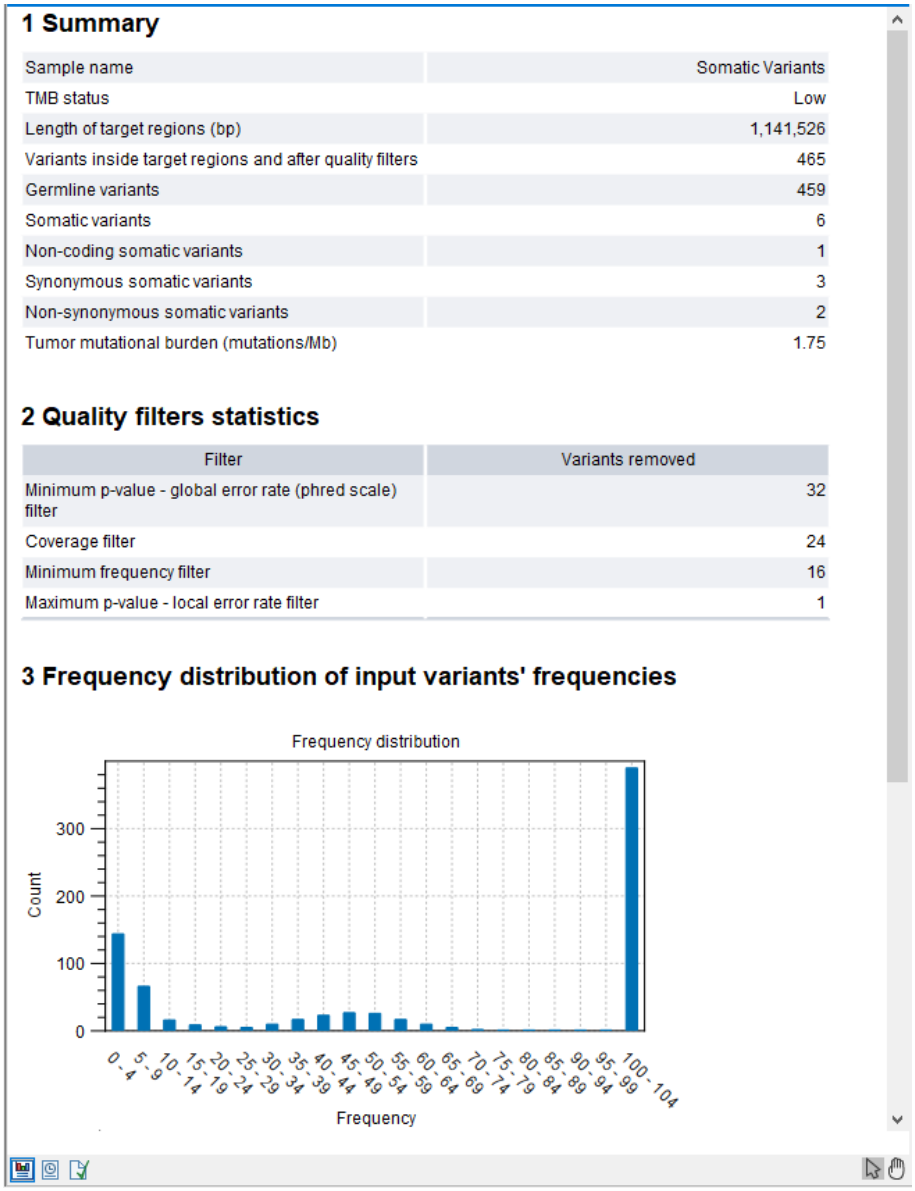

Figure 3.27: *Part of a TMB report where the option to detect TMB status was enabled with default threshold values.*

3.6 Copy Number Variant Detection (WGS)

The **Copy Number Variant Detection (WGS)** tool is designed to identify copy number variants (CNVs) from whole genome sequencing (WGS) data including low-pass WGS data.

The tool takes a read mapping as input and is designed to not rely on control samples. This is achieved by estimating the expected coverage for diploid regions via the following steps:

- The mean and standard deviation for the coverage is calculated for each chromosome. Only coverage values between the 10th and 90th percentile are included.
- Chromosomes are clustered based on their mean and standard deviation for the coverage.
- The cluster with the greatest density and highest number of chromosomes is selected using

a weighted density value.

- Finally, the mean coverage for the selected cluster of chromosomes is calculated and used as the expected coverage for diploid regions.

If the tool is unable to find a robust cluster, the median coverage for all chromosomes will be used as the expected coverage for diploid regions.

The presence of XX or XY chromosomes is automatically determined by the tool based on the observed coverage. Note that only XX or XY can be assigned by the tool.

The tool defines non-overlapping windows that mapped reads are divided into, and calculates a coverage in each window that is adjusted for mapping quality and GC content. The resulting coverage in each window is normalized using the expected coverage for diploid regions. If any masking tracks are provided, the coverage will be ignored in the regions defined by these.

CNVs are detected using the normalized window coverage values with a hidden Markov model (HMM). The HMM consists of 11 different states for diploid chromosomes that each represent a copy number (0-10). Only 10 states (0-9) are used for haploid chromosomes, i.e. when there is both an X and a Y chromosome. For each window, we calculate a probability for each state based on the normalized coverage value and the sample purity. The HMM considers each window as an event, and then tries to find the most likely sequence of copy number states that explains these events. The boundaries of detected CNVs are refined by using a window of half the size of the original window.

A coefficient of variation is calculated based on the coverage windows across all chromosomes. This is used to set the copy number state transition probabilities in the HMM. This serves the purpose of making it less likely to detect a CNV when using noisy data. The coefficient of variation is also used for automatically determining the window size.

The HMM calculates a CNV score by initially obtaining the probability of the sequence of events that occurred (i.e. the states it traversed). Next, the probability that there is no CNV is found using the same calculations, but where only copy neutral states are traversed. The CNV score is finally calculated as the log-ratio between these two probabilities.

To run **Copy Number Variant Detection (WGS)**:

Tools | **LightSpeed** () | **Copy Number Variant Detection (WGS)** ()

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 read mapping.

Next, options are available for window size selection and sequence masking (figure 3.28):

- **Window size**
  - **Automatic window size** or **Specify window size**. Choose whether the tool should estimate the optimal window size based on the data or use a pre-specified window size.
  - **Window size (kb)** Manually specify the window size. This affects the size of CNVs that can be detected. The lower the coverage, the larger the window size should be. CNVs typically need to span at least 2-3 consecutive windows to be called.

- **Masking regions**
  - **Centromeres** Specify regions defining centromeres. Centromere regions are repetitive regions that can result in the detection of false positive CNVs.
  - **Pseudoautosomal regions (PAR)** Specify PAR. PAR are homologous between the X and Y chromosome which similarly can result in the detection of false positive CNVs.
  - **Umap mappability** Specify a graph track with position-wise mappability scores. Specific regions of the reference sequence are in a problematic way too likely/unlikely to have reads mapped resulting in false positive CNVs. This likelihood can be reflected by position-wise mappability scores. When a mappability track is provided the windows with an average mappability score below the provided minimum mappability score will not be used for estimating the coverage for diploid regions. In addition, identified CNVs with an average mappability score below the minimum mappability score value will be removed. We recommend using mappability scores from the tool Umap [Karimzadeh et al., 2018].
  - **Minimum mappability score** Specify a minimum average mappability score for windows and identified CNVs.

Tracks containing centromere regions, PAR and Umap mappability scores are available for hg19 and hg38_no_alt_analysis_set in the Reference Data Manager. The tracks contain regions that are known to exhibit systematic abnormal coverage for different reasons, and we recommend masking with all three tracks when calling CNVs. Note that windows with more than 25% ambiguous nucleotides, i.e. N, will automatically be masked. In addition, detected CNVs are required to have less than 50% overlap with centromere and PAR regions and less than 40% overlap with N-masked regions.

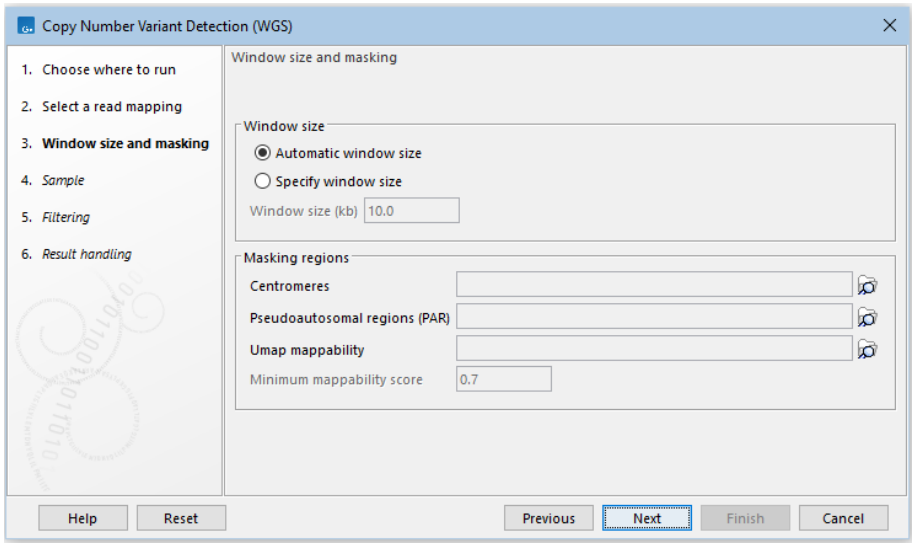

Figure 3.28: *Specify how the window size should be determined and provide tracks for sequence masking.*

In the sample step, the following options are available (figure 3.29):

- **Sample type**

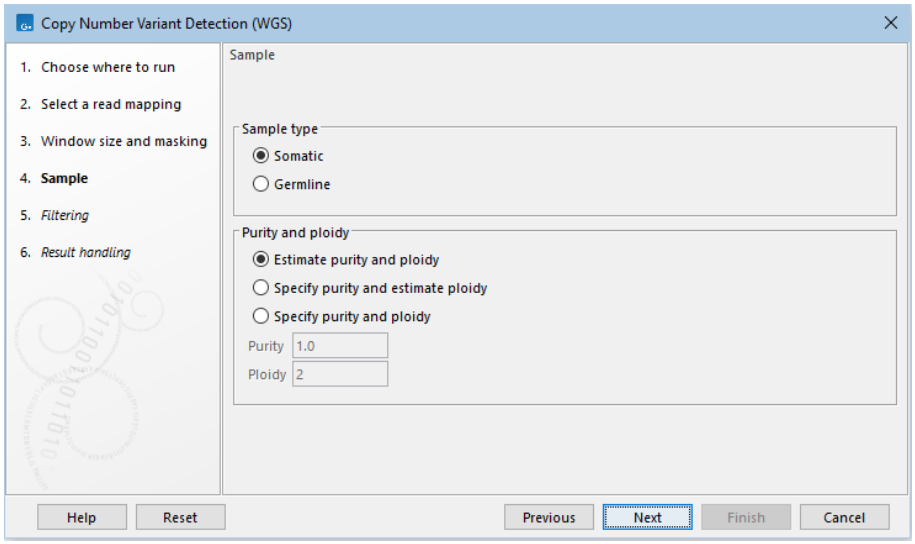

Figure 3.29: *Provide sample information for CNV detection.*

  - Specify whether the sample should be considered germline or somatic. Germline samples are assumed to have ploidy 2 and purity 1.0. For somatic samples the purity and ploidy can be estimated or manually specified.

- **Purity and ploidy**

  - Choose to manually specify the purity and ploidy or have the tool automatically estimate the values. Purity is taken into account when predicting the copy number states of individual windows. A lower purity will result in the detection of more CNVs. Note that the automatic estimation of purity and ploidy is performed by fitting the observed CNVs to a range of purity and ploidy values and selecting the best fit. It therefore only serves as an estimation. The automatic estimation will accept a maximum ploidy value of 6.

Next, options that can be used to filter CNVs are available (figure 3.30):

- **Filtering and probability**

  - **Minimum score** Remove CNVs with a CNV score below this cutoff (see how the CNV scores are calculated by the HMM above).
  - **Maximum output probability** The maximum copy number state output probabilities calculated by the HMM can be limited to this threshold. The threshold also sets a minimum output probability that is equal to 1.0 - maximum output probability. Decreasing the maximum output probability will typically result in fewer and shorter CNVs, and increasing the maximum output probability will typically result in more and longer CNVs.
  - **Remove CNVs in regions with many non-specific reads** Enabling this option will remove CNVs where more than half of the underlying windows contain a high fraction of non-specifically mapped reads. The threshold for a high fraction of non-specifically mapped reads is 25% for CNVs with a score below 25 and 50% for CNVs with a score above 25.

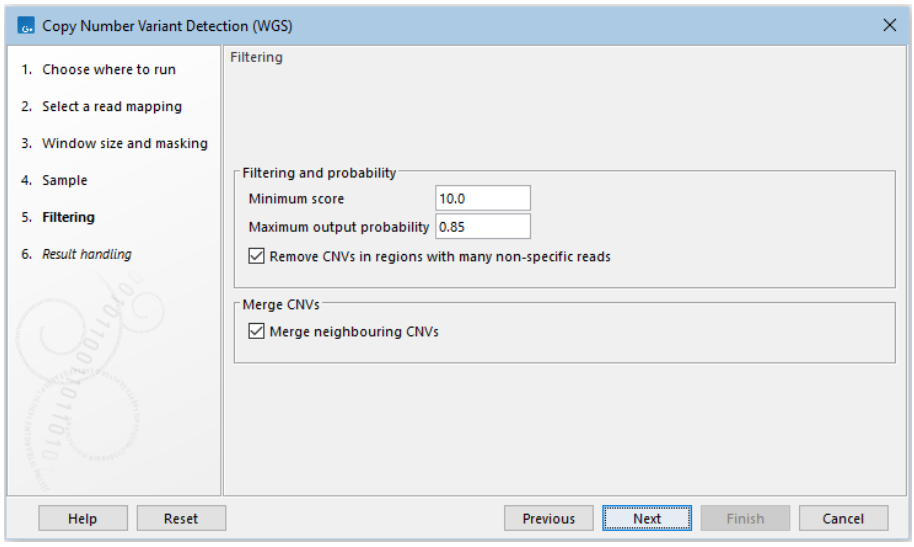

Figure 3.30: *Specify filtering for CNV detection.*

- **Merge CNVs**
  - Merge CNVs that are in close relation and have been identified to have the same copy number.

The **Copy Number Variant Detection (WGS)** tool has the following limitations:

- Only chromosomes larger than 10Mb are considered.
- Broken reads are ignored because they can accumulate in specific places, e.g. at the boundaries of deletions.
- If the majority of chromosomes are affected by chromosome-wide CNVs, the expected coverage for diploid regions might be suboptimally estimated as affected chromosomes may define the cluster used to estimate the expected coverage.
- The tool is not able to discern female X loss from male Y loss.
- The tool is only able to call up to a copy number of 10. A CNV with copy number 10 should therefore be interpreted as 10 or more.

3.6.1 Copy Number Variant Detection (WGS) outputs

The **Copy Number Variant Detection (WGS)** tool produces the following outputs:

- **CNVs from WGS** A track containing all detected CNVs with the CNV region and the following information:
  - **Length** The length of the CNV in base pairs.
  - **Median read counts per window** The median read count per window that is part of the CNV.
  - **Consequence** The consequence of the detected CNV, i.e. "Gain" or "Loss".

  - **Copy number** The predicted copy number of the CNV.
  - **Score** A score reflecting the amount of evidence of the CNV being a true event.
  - **Average mappability score** The average Umap mappability score of the CNV. This annotation is only present if a Umap mappability track was provided when running the tool.
  - **Distance to centromere** The distance to the nearest centromere.

- **N-masked and low mappability regions** A track containing N-masked regions and, if a Umap mappability track was provided, regions with low mappability. It outlines the masked regions and the following information:
  - **Type** Specification of the type of masking.
  - **Average mappability score** The average Umap mappability score of the masked region.

- **CNV WGS report** A detailed report about the detected CNVs. It contains the following sections:
  - **Summary** A summary table containing general information about the sample, the window size and the identified CNVs.
  - **References** A table listing each of the analyzed chromosomes together with the number of reads mapped to the chromosome and the average coverage.
  - **CNVs per chromosome** A table listing each of the analyzed chromosomes together with the number of losses and gains and the number of positions affected by gains and losses.
  - **Visualization of CNVs** An overview plot depicting identified CNVs on all the chromosomes in the genome.
  - **Chromosome window read count distributions** Plot of the normalized mean read count per window versus the standard deviation for the read count per window. Each chromosome is represented by one dot. The selected cluster is black, and chromosomes that are not part of the cluster are grey. A red dot is positioned at the center of the selected cluster. The mean coverage at the center of the selected cluster is defined as the expected coverage for a diploid region.
  - **Size distribution** A table summarizing the sizes of identified CNVs.
  - **Coverage per window and identified CNVs** A section containing a plot for all chromosomes combined and one for each chromosome. The plots show normalized read counts per window versus the genomic position. Identified CNVs and masked regions are also illustrated. The plots contain one dot for each window, but dots may be collapsed if they cannot be visually discerned. Per default, the Y-axis is scaled to fit 95% of all dots; however, identified CNVs will always be included and this can result in more than 95% of all dots being visible in the plot.

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 WGS or targeted data, including WES, where no specialized trimming is needed, are described.

Contents

4.1 Fastq to Germline Variants (WGS)

The **Fastq to Germline Variants (WGS)** template workflow identifies and annotates germline variants and generates various QC metrics. It is intended for analysis of whole genome sequencing (WGS) data.

The workflow can be found at:

> **Template Workflows** | **LightSpeed Workflows** ( ) | **Fastq to Germline Variants (WGS)** ( )

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.8 for additional details.
  - **Minimum allele count** Specify the minimum number of reads supporting an identified variant.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

4.1.1 Outputs from Fastq to Germline Variants (WGS)

The **Fastq to Germline Variants (WGS)** template workflow produces the following outputs:

- **Read Mapping** A read mapping track ().

- **Germline Variants** A 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.
- **Read Mapping Report** A report () summarizing QC metrics, including coverage, for the read mapping.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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, the Read Mapping as well as the Reference sequence and the Genes, mRNA and CDS tracks.

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

The **Read Mapping Report** is produced by **QC for Read Mapping** (https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html).

4.2 Fastq to Germline Variants (WES)

The **Fastq to Germline Variants (WES)** template workflow identifies and annotates germline variants and generates various QC metrics. It is intended for analysis of data generated with target enrichment, including whole exome sequencing (WES) data, and therefore requires target regions to be provided.

Fastq to Germline Variants (WES) can be found at:

> **Template Workflows** | **LightSpeed Workflows** () | **Fastq to Germline Variants (WES)** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Select Target regions** Specify target regions for analysis.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.8 for additional details.
  - **Lenient inversion detection** When enabled, inversions with read support in only one direction at each breakpoint can be called. Enabling this option is recommended when analysing targeted data, but can increase processing time and can result in detection of more false positive inversions.
  - **Minimum allele count** Specify the minimum number of reads supporting an identified variant.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 4.7. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.

- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

4.2.1 Outputs from Fastq to Germline Variants (WES)

The **Fastq to Germline Variants (WES)** template workflow produces the following outputs:

- **Germline Variants** A 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.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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.

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

4.3 Fastq to Somatic Variants (WGS)

The **Fastq to Somatic Variants (WGS)** template workflow identifies and annotates somatic variants and generates various QC metrics. It is intended for analysis of whole genome sequencing (WGS) data.

The workflow can be found at:

> **Template Workflows** | **LightSpeed Workflows** () | **Fastq to Somatic Variants (WGS)** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **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.
  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.9 for additional details.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.

- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

4.3.1 Outputs from Fastq to Somatic Variants (WGS)

The **Fastq to Somatic Variants (WGS)** template workflow produces the following outputs:

- **Read Mapping** A read mapping track ().
- **Somatic Variants** A 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.
- **Read Mapping Report** A report () summarizing QC metrics, including coverage, for the read mapping.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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, the Read Mapping as well as the Reference sequence and the Genes, mRNA and CDS tracks.

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

The **Read Mapping Report** is produced by **QC for Read Mapping** (https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html).

4.4 Fastq to Somatic Variants (WES)

The **Fastq to Somatic Variants (WES)** template workflow identifies and annotates somatic variants and generates various QC metrics. It is intended for analysis of data generated with target enrichment, including whole exome sequencing (WES) data, and therefore requires target regions to be provided.

Fastq to Somatic Variants (WES) can be found at:

> **Template Workflows** | **LightSpeed Workflows** () | **Fastq to Somatic Variants (WES)** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Select Target regions** Specify target regions for analysis.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **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.
  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.9 for additional details.
  - **Lenient inversion detection** When enabled, inversions with read support in only one direction at each breakpoint can be called. Enabling this option is recommended when analysing targeted data, but 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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 4.8. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is

meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

4.4.1 Outputs from Fastq to Somatic Variants (WES)

The **Fastq to Somatic Variants (WES)** template workflow produces the following outputs:

- **Somatic Variants** A 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.
- **Sample Report** A report ( ) that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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.

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

4.5 Fastq to Somatic Variants (Tumor Normal) (WGS)

The **Fastq to Somatic Variants (Tumor Normal) (WGS)** template workflow identifies and annotates somatic variants from tumor normal reads and generates various QC metrics. It is intended for analysis of whole genome sequencing (WGS) data.

The workflow can be found at:

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

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Somatic Variants Tumor Normal** Specify options for the LightSpeed Fastq to Somatic Variants Tumor Normal tool:
  - **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.
  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.10 for additional details.

- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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

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

- **Tumor Read Mapping** A read mapping track for the tumor sample ().
- **Normal Read Mapping** A read mapping track for the normal sample ().
- **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.
- **Tumor Read Mapping Report** A report () summarizing QC metrics, including coverage, for the tumor read mapping.
- **Normal Read Mapping Report** A report () summarizing QC metrics, including coverage, for the normal read mapping.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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, the tumor and normal Read Mappings as well as the Reference sequence and the Genes, mRNA and CDS tracks.

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

The **Read Mapping Report** is produced by **QC for Read Mapping** (https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html).

4.6 Fastq to Somatic Variants (Tumor Normal) (WES)

The **Fastq to Somatic Variants (Tumor Normal) (WES)** template workflow identifies and annotates somatic variants and generates various QC metrics. It is intended for analysis of data generated with target enrichment, including whole exome sequencing (WES) data, and therefore requires target regions to be provided.

Fastq to Somatic Variants (Tumor Normal) (WES) can be found at:

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

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Select Target regions** Specify target regions for analysis.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Somatic Variants Tumor Normal** Specify options for the LightSpeed Fastq to Somatic Variants Tumor Normaltool:
  - **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.
  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.10 for additional details.

- **QC for Targeted Sequencing (Normal)** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool for the normal read mapping. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **QC for Targeted Sequencing (Tumor)** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool for the tumor read mapping. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

Note that to be able to add information from both the normal and the tumor reports from QC for Targeted Sequencing to the sample report, the reports are retyped using the **Modify Report Type** tool. To select additional summary items for the two reports, the following steps are needed: After pressing **Add...**, select the **QC for Targeted Sequencing** report type. Then check "Apply to custom report type" and write either "QC for Targeted Sequencing (Normal)" or "QC for Targeted Sequencing (Tumor)" as relevant before selecting summary items.

4.6.1 Outputs from Fastq to Somatic Variants (Tumor Normal) (WES)

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

- **Somatic Variants** A 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.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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.

- **Tumor Read Mapping** A read mapping track for the tumor sample ().
- **Normal Read Mapping** A read mapping track for the normal sample ().
- **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** (https://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** (https://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 Germline Variants (WES)** template workflow (section 4.2).

Fastq to Germline CNV Control can be found at:

> **Template Workflows** | **LightSpeed Workflows** () | **Fastq to Germline CNV Control** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Select Target regions** Specify target regions for analysis.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the

available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

4.7.1 Outputs from Fastq to Germline CNV Control

The **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, either in the **Fastq to Germline Variants (WES)** (section 4.2) template workflow, or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.

- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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** (https://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 Somatic Variants (WES)** template workflow (section 4.4).

Fastq to Somatic CNV Control can be found at:

> **Template Workflows** | **LightSpeed Workflows** () | **Fastq to Somatic CNV Control** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Select Target regions** Specify target regions for analysis.
- **Specify reference data handling** Select a Reference Data Set. If you have not downloaded the Reference Data Set yet, the dialog will suggest the relevant data set and offer the opportunity to download it using the Download to Workbench button. If none of the available reference data sets are appropriate, custom reference data sets can be created, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.

- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.

- **Result handling** Choose if a workflow result metadata and/or log should be saved.

- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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 Somatic Variants (WES)** (section 4.4), or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.

- **Sample Report** A report summarizing the most important metrics from the LightSpeed Report and the Coverage Report. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.

- **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** (https://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 Germline Variants

The **QIAseq Fastq to 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 Germline Variants**

Options in the following dialogs can be configured:

- **Choose where to run** 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.

- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Panels hg19** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_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.

- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:

  - **Reads (fastq)** Press **Browse** to select fastq files for analysis.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.8 for additional details.
  - **Minimum allele count** Specify the minimum number of reads supporting an identified variant.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

- **Target regions** Choose the relevant target regions from the drop down list.

- **Target primers** Choose the relevant target primers from the drop down list.

- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.

- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 5.3. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

5.1.1 Outputs from QIAseq Fastq to Germline Variants

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

- **Germline Variants** A 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 () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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 informtion 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.

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

5.2 QIAseq Fastq to Somatic Variants

The **QIAseq Fastq to 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 Somatic Variants**

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Panels hg19** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_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.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **Reads (fastq)** Press **Browse** to select fastq files for analysis.
  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.9 for additional details.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).

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

- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 5.4. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

5.2.1 Outputs from QIAseq Fastq to Somatic Variants

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

- **Somatic Variants** A 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 () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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 informtion 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.

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

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 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** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Panels hg19** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_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.
- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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 Germline Variants (section 5.1) template workflow or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.
- **Sample Report** A report ( ) that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (https://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 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** ( )

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Panels hg19** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_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.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:

  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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 Somatic Variants** (section 5.2) template workflow or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (https://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 Germline Variants

The **QIAseq Pro Fastq to 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 Germline Variants**

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Pro Panels hg38** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html. Reference data sets for QIAseq Targeted DNA 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.
- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.8 for additional details.
  - **Minimum allele count** Specify the minimum number of reads supporting an identified variant.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 6.3. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must

result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.

- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

6.1.1 Outputs from QIAseq Pro Fastq to Germline Variants

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

- **Germline Variants** A 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 ( ) that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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 informtion 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.

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

6.2 QIAseq Pro Fastq to Somatic Variants

The **QIAseq Pro Fastq to 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 Somatic Variants** ()

Options in the following dialogs can be configured:

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Pro Panels hg38** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html. Reference data sets for QIAseq Targeted DNA 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.
- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **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.

  - **Minimum frequency (%)** Specify the minimum variant allelle frequency for detected variants.
  - **Minimum average quality** Specify the minimum average quality of detected variants. See section 2.9 for additional details.
  - **Batch** Select if fastq files from different samples are used as input, and each sample should be analyzed individually (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.

- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Copy Number Variant Detection (Targeted)** Specify **Controls** against which the coverage pattern in your sample will be compared in order to call CNVs. If you do not specify a control mapping the CNV analysis will not be carried out. Please note that if you want the CNV analysis to be done, it is important that the control mapping supplied is a meaningful control for the sample being analyzed. Mapping of control samples for the CNV analysis can be done using the workflows described in section 6.4. A meaningful control must satisfy two conditions: (1) It must have a copy number status that is meaningful to compare against. For panels with targets on the X and Y chromosomes, the control and sample should be matched for gender. (2) The control read mapping must result from the same type of processing that will be applied to the sample. For more information about CNV detection see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

6.2.1 Outputs from QIAseq Pro Fastq to Somatic Variants

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

- **Somatic Variants** A 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 () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.
- **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 informtion 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.

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

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 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**

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Pro Panels hg38** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html. Reference data sets for QIAseq Targeted DNA 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.
- **LightSpeed Fastq to Germline Variants** Specify options for the LightSpeed Fastq to Germline Variants tool:
  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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 Germline Variants** (section 6.1) template workflow or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.

The **Coverage Table**, **Coverage Report**, and the **Target Region Statistics Track** are produced by **QC for Targeted Sequencing** (https://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 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** ()

- **Choose where to run** 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.
- **Specify reference data handling** Select the relevant Reference Data Set. **QIAseq DNA Pro Panels hg38** will be pre-selected and is recommended for running this workflow. 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reference_Data_Sets_defining_Custom_Sets.html.

Reference data sets for QIAseq Targeted DNA 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.

- **LightSpeed Fastq to Somatic Variants** Specify options for the LightSpeed Fastq to Somatic Variants tool:
  - **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 (for information about batching see section 2.12).
  - **Join lanes when batching** Select to join fastq files from the same sample that were sequenced on different lanes.
- **Target regions** Choose the relevant target regions from the drop down list.
- **Target primers** Choose the relevant target primers from the drop down list.
- **QC for Targeted Sequencing** Set the Minimum coverage parameter of the QC for Targeted Sequencing tool. Using default settings, samples where 90 percent of target region positions do not meet this threshold will be flagged in the sample report generated by the workflow.
- **Create Sample Report** Select relevant summary items and specify thresholds for quality control. Summary items, thresholds and an indication of whether specified thresholds were met, will be shown in the quality control section of the sample report. The default summary items are appropriate for many data sets, but may need to be adjusted. For additional information, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html.
- **Result handling** Choose if a workflow result metadata and/or log should be saved.
- **Save location for new elements** Choose where to save the data, and press **Finish** to start the analysis.

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:

- **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 Somatic Variants** (section 6.2) template workflow or directly in the tool **Copy Number Variant Detection (Targeted)** (https://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.
- **Sample Report** A report () that contains compiled QC metrics from other reports and provides an overview of a given sample. The report contains a quality control section reflecting the summary items specified in the Create Sample Report wizard step.

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

Bibliography

[Karimzadeh et al., 2018] Karimzadeh, M., Ernst, C., Kundaje, A., and Hoffman, M. (2018). Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, 46(20):e120.

[Shugay et al., 2017] Shugay, M., Zaretsky, A. R., Shagin, D. A., Shagina, I. A., Volchenkov, I. A., Shelenkov, A. A., Lebedin, M. Y., Bagaev, D. V., Lukyanov, S., and Chudakov, D. M. (2017). Mageri: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS computational biology*, 13(5):e1005480.

[Vega et al., 2021] Vega, D., Yee, L., McShane, L., Williams, P., Chen, L., Vilimas, T., Fabrizio, D., Funari, V., Newberg, J., Bruce, L., Chen, S.-J., Baden, J., Barrett, J. C., Beer, P., Butler, M., Cheng, J.-H., Conroy, J., Cyanam, D., Eyring, K., Garcia, E., Green, G., Gregersen, V., Hellmann, M., Keefer, L., Lasiter, L., Lazar, A., Li, M.-C., MacConaill, L., Meier, K., Mellert, H., Pabla, S., Pallavajjalla, A., Pestano, G., Salgado, R., Samara, R., Sokol, E., Stafford, P., Budczies, J., Stenzinger, A., Tom, W., Valkenburg, K., Wang, X., Weigman, V., Xie, M., Xie, Q., Zehir, A., Zhao, C., Zhao, Y., Stewart, M., and on behalf of the TMB Consortium, J. A. (2021). Aligning tumor mutational burden (tmb) quantification across diagnostic platforms: phase ii of the friends of cancer research tmb harmonization project. *Annals of Oncology*, 132(12):1626–1636.