

# **Long Read Support Plugin**

USER MANUAL

# User manual for *Long Read Support 24.0*

Windows, macOS and Linux

January 5, 2024

**This software is for research purposes only.**

QIAGEN Aarhus  
Silkeborgvej 2  
Prismet  
DK-8000 Aarhus C  
Denmark



# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	The concept of Long Read Support . . . . .	5
1.2	Contact information . . . . .	5
1.3	System requirements . . . . .	6
<b>2</b>	<b>Template workflows</b>	<b>7</b>
2.1	De Novo Assemble Long Reads and Polish with Short Reads . . . . .	7
<b>3</b>	<b>Import Nanopore Reads</b>	<b>10</b>
<b>4</b>	<b>De Novo Assemble Long Reads</b>	<b>11</b>
4.1	De novo assembly parameters . . . . .	11
4.2	De novo assembly output . . . . .	12
<b>5</b>	<b>Map Long Reads to Reference</b>	<b>15</b>
5.1	Map Long Reads to Reference output . . . . .	16
<b>6</b>	<b>Polish with Reads</b>	<b>19</b>
6.1	Polish with Reads output . . . . .	20
<b>7</b>	<b>Correct Long Reads</b>	<b>22</b>
7.1	Correct Long Reads output . . . . .	23
<b>8</b>	<b>RNA-Seq Analysis for Long Reads</b>	<b>25</b>
8.0.1	Reads and reference settings . . . . .	25
8.0.2	Mapping settings . . . . .	27
8.0.3	Expression settings . . . . .	27
8.1	RNA-Seq Analysis for Long Reads output . . . . .	29

---

<b>9</b>	<b>Structural Variant Caller for Long Reads</b>	<b>32</b>
9.1	Structural Variant Caller settings . . . . .	32
9.2	Structural Variant Caller output . . . . .	33
9.3	Structural Variant Caller algorithm . . . . .	37
<b>10</b>	<b>Install and uninstall plugins</b>	<b>38</b>
10.1	Installation of plugins . . . . .	38
10.2	Uninstalling plugins . . . . .	39
	<b>Bibliography</b>	<b>41</b>

# Chapter 1

## Introduction

Welcome to *Long Read Support 24.0* – a software package supporting your daily bioinformatics work.

### 1.1 The concept of Long Read Support

The Long Read Support plugin provides tools for working with long, next-generation sequencing reads such as those produced by Pacific Biosciences or Oxford Nanopore Technologies sequencing platforms.

Unless specified otherwise, all tools described in this manual can be found in the Long Read Support folder that will be placed in the Toolbox once the plugin is installed.

### 1.2 Contact information

Long Read Support is developed by:

QIAGEN Aarhus  
Silkeborgvej 2  
Prismet  
8000 Aarhus C  
Denmark

<https://digitalinsights.qiagen.com/>

Email: [ts-bioinformatics@qiagen.com](mailto:ts-bioinformatics@qiagen.com)

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

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

- Core product manuals <https://digitalinsights.qiagen.com/technical-support/manuals/>

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

## 1.3 System requirements

Long Read Support 24.0 is for use with CLC Genomics Workbench 24.0 or newer. The system requirements for CLC Genomics Workbench are provided at [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=System\\_requirements.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=System_requirements.html)

The system requirements for Long Read Support are the same as those for CLC Genomics Workbench, apart from the following:

- An AMD/Intel CPU that supports AVX2 or an Apple M series CPU is required for the below tools:
  - Polish with Reads
  - Correct Reads
  - De Novo Assemble Long Reads

### Special requirements when assembling PacBio HiFi reads with De Novo Assemble Long Reads

At least 32 GB RAM is recommended for running the De Novo Assemble Long Reads tool with PacBio HiFi reads.

### Special requirements for Correct Reads and Polish with Reads

Correct Reads and Polish with Reads use Racon [Vaser et al., 2017] for which the following is listed:

- *Memory consumption of polishing equals the size of the mandatory input files, while the memory is somewhat larger for error correction.* (<https://complex.zesoi.fer.hr/en/blog-en/58-racon-1-0-release>)

# Chapter 2

## Template workflows

Long Read Support offers one template workflow. Template workflows can be used as they are, or copies easily be opened, allowing you optimize the workflow to fit your specific application. For a general introduction to workflows, see <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>.

### 2.1 De Novo Assemble Long Reads and Polish with Short Reads

The **De Novo Assemble Long Reads and Polish with Short Reads** workflow performs de novo assembly of long reads and polishes the assembly with high-quality short reads. The workflow works best with **uncorrected** long reads why it is not recommended to run the Correct Long Reads tool before running this workflow.

#### Launching the workflow

To run the workflow, go to:

**Template Workflows** (📁) | **Long Read Workflows** (🔍) | **De Novo Assemble Long Reads and Polish with Short Reads** (🔍)

Launch the workflow and step through the wizard.

- Select the long reads to be assembled (figure 2.1).

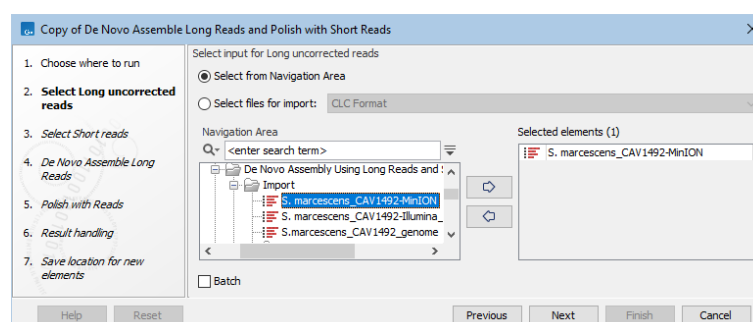


Figure 2.1: Select long, uncorrected reads to assemble

- Select the short reads to be used for polishing (figure 2.2).

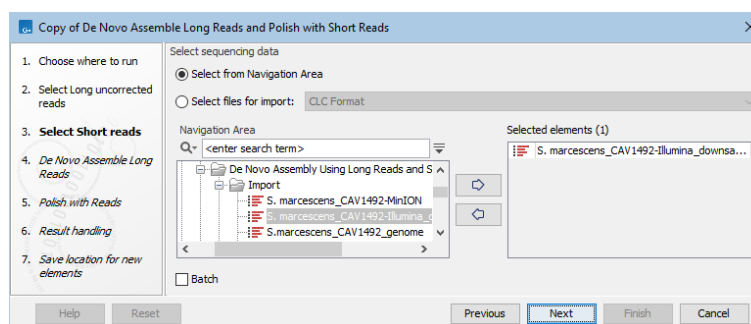


Figure 2.2: Select short reads used to polish the assembly

- If you have more than one sample (set of long and short reads) to assemble, check the Batch checkbox in the above steps. For further details on how to run workflows in batch mode when you have more than one input, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batching\\_workflows\\_with\\_more\\_than\\_one\\_input\\_changing\\_per\\_run.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batching_workflows_with_more_than_one_input_changing_per_run.html).
- Check the *PacBio HiFi* checkbox if your long reads are PacBio HiFi reads (figure 2.3).

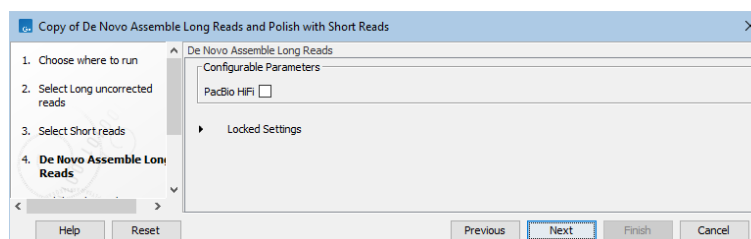


Figure 2.3: Check the checkbox if the long reads are PacBio HiFi reads.

- Check the *Include unpolished sequences* to include unpolished contigs in the output contig list (figure 2.4).

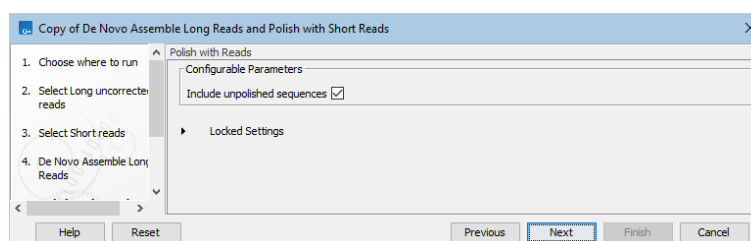


Figure 2.4: Select whether to keep unpolished contigs.

- In the final step, choose a location to save the results to.

## Workflow tools and outputs

The **De Novo Assemble Long Reads and Polish with Short Reads** workflow contains five tool (figure 2.5):



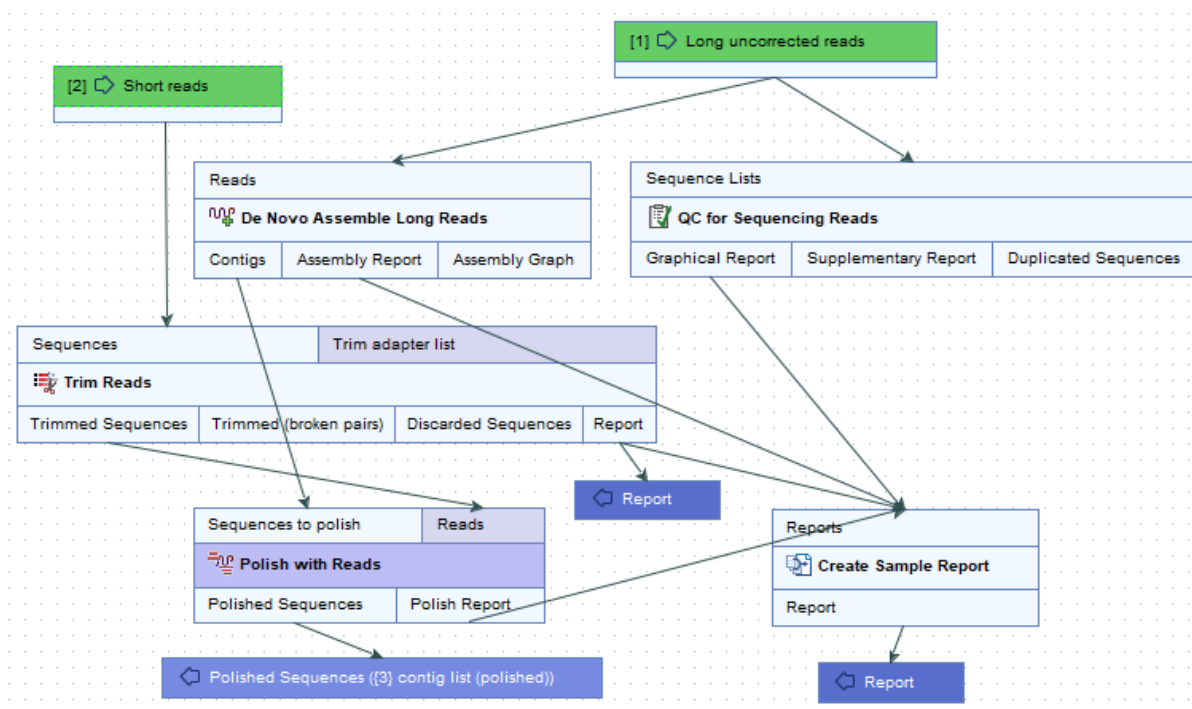


Figure 2.5: The De Novo Assemble Long Reads and Polish with Short Reads workflow.

- **QC for Sequencing Reads.** See [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC\\_Sequencing\\_Reads.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Sequencing_Reads.html).
- **Trim Reads.** See [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim\\_Reads.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html).
- **De Novo Assemble Long Reads.** See chapter 4.
- **Polish with Reads.** See chapter 6.
- **Create Sample Report.** See [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create\\_Sample\\_Report.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Create_Sample_Report.html).

The outputs provided by the workflow are:

- **Graphical QC trim report.**
- **Contig list (polished).** The output from **Polish with Reads**. Depending on the workflow settings, this may contain unpolished reads.
- **Sample report.** Contains general information about the sample and contigs produced. Includes QC for raw reads, assembly and polishing statistics.

## Chapter 3

# Import Nanopore Reads

It is possible to import Oxford Nanopore reads using the Oxford Nanopore NGS importer or by downloading them using the **Search for Reads in SRA** functionality of the Workbench.

To use the importer:

**Import | Oxford Nanopore...** (🔧)

Specify the files to import in the import wizard (figure 3.1).

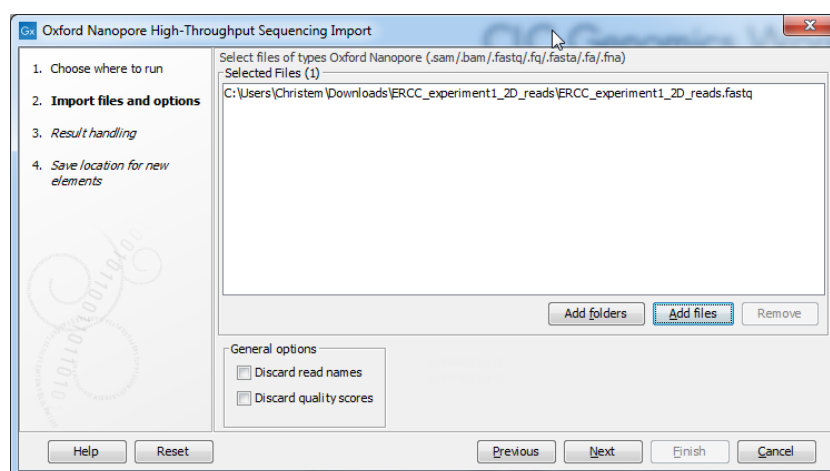


Figure 3.1: *Importing Oxford Nanopore reads*

- **Discard read names** Discard read names of the reads to import in order to save storage space.
- **Discard quality scores** Discard the quality scores of the reads to import in order to save storage space.

PacBio long reads can be imported with the CLC Genomics Workbench PacBio Long Reads importer, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PacBio\\_Long\\_Reads.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=PacBio_Long_Reads.html).

## Chapter 4

# De Novo Assemble Long Reads

The De Novo Assemble Long Reads tool facilitates the generation of de novo assemblies from long reads. It is based on open source tools: Raven [Vaser and Šiki, 2021] for analysis of Oxford Nanopore reads and Pacific Biosciences CLR reads, and hifiiasm [Cheng et al., 2021] for assembly of Pacific Biosciences HiFi reads.

The De Novo Assemble Long Reads tool works best with uncorrected long reads why it is not recommended to run the Correct Long Reads tool beforehand.

To run the De Novo Assemble Long Reads tool, go to:

**Toolbox | Long Read Support (📄) | De Novo Assemble Long Reads (🚀)**

Select one or more sequence lists containing long reads.

### 4.1 De novo assembly parameters

In the dialog seen in figure 4.1, the following parameters are available:

De Novo Assemble Long Reads

1. Choose where to run

2. Select sequencing reads

3. **De novo options**

4. Result handling

De novo options

Contig polishing

☒ Polish with reads

Output filter

Minimum contig length 1,000

☒ Keep circular contigs

PacBio HiFi options

Genome size Infer

Genome size (megabases) 5

Ploidy 2

Help Reset Previous **Next** Finish Cancel

Figure 4.1: De Novo Assemble Long Reads parameters.

- **Polish with reads.** When this parameter is set, two iterations of read polishing are run on the raw contig output, similar to running Polish with Reads (see chapter 6). Disabling this

parameter results in raw contigs with error rates similar to the error rate of the reads. This option is disabled for PacBio HiFi reads.

- **Minimum contig length.** The minimum length of contigs included in the output. Shorter contigs will be filtered.
- **Keep circular contigs.** When enabled, the minimum contig length filtering is not applied to circular contigs. This means that all circular contigs will be output regardless of length.
- **PacBio HiFi options** (enabled for PacBio HiFi reads only)
  - **Genome size.** *Infer* automatically determines the genome size as part of the analysis. *Manual* instructs the algorithm to use the genome size specified in the text field below for inferring read coverage.
  - **Genome size (megabases).** Enter the expected genome size.
  - **Ploidy.** The number of expected alleles. If it is set to >2, the quality of the assembly for polyploid genomes might be improved.

When running **De Novo Assemble Long Reads** in a workflow, the workflow element dialog will contain an additional option:

- **PacBio HiFi.** Check this option to indicate that input reads are PacBio HiFi reads. When selected, the tool will run with an algorithm optimized for HiFi reads.

When running the De Novo Assemble Long Reads tool separately, the read type is inferred from the input reads.

## 4.2 De novo assembly output

In addition to the sequence list of assembled contigs, the following outputs are available:

- **Create report.** Creates a summary report.
- **Create assembly graph.** Generates a visual representation of the assembly showing the contigs and connections between them.

### De novo assembly report

The assembly report contains information on the base and length distributions of the contigs. An example of the first sections of the report is shown in figure 4.2.

- **Nucleotide distribution.**
- **Contig measurements.** Statistics about the number and lengths of contigs.
  - **Contigs.** The number of contigs.
  - **Minimum, Maximum, Average.** Minimum, maximum and average contig length.

**1 Nucleotide distribution**

Nucleotide	Count	Frequency (%)
Adenine (A)	50,022,125	28.98
Cytosine (C)	36,377,066	21.07
Guanine (G)	36,229,844	20.99
Thymine (T)	50,008,864	28.97

**2 Contig measurements**

Contigs	441
Minimum	10,824
Maximum	24,726,942
Average	391,469
N50	15,797,497
N90	159,247
Total	172,637,899

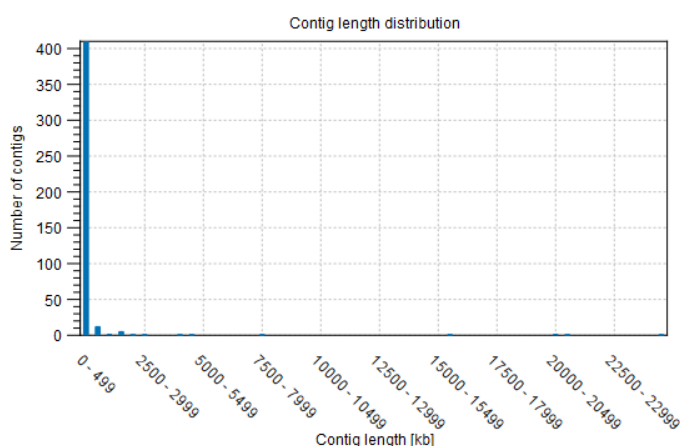


Figure 4.2: De Novo Assemble Long Reads report

- **N50.** The length of the shortest contig in sets of contigs of equal length or longer, where the summed length of contigs is at least 50% of the total contig length. As such, N50 is the shortest contig length that must be included to cover 50% of the assembly.
- **N90.** The length of the shortest contig in a set of contigs of equal length or longer, where the summed length of contigs is at least 90% of the total contig length. As such, N90 is the shortest contig length that must be included to cover 90% of the assembly. N90 will be equal to or smaller than N50.
- **Total.** The number of bases in the contigs. This can be used for comparison with the estimated genome size to evaluate how much of the genome sequence is included in the assembly.
- **Contig length distribution.** The number of contigs found at a specific length.
- **Accumulated contig length.** The y-axis shows the summed contig length, while the x-axis represents the number of contigs, arranged with the largest contigs first. This provides insight into the number of contigs required to cover, for instance, half of the genome.

For HiFi data, the report contains an additional section:

- **K-mer coverage.**

- **K-mer plot.** Illustrates the frequency distribution of k-mers based on their occurrences. For homozygous samples, a single peak is expected around the read coverage. For heterozygous samples, two peaks should be visible, with the smaller peak corresponding to the heterozygous read coverage and the larger peak corresponding to the homozygous read coverage.
- **Homozygous coverage peak.** Estimated homozygous coverage.
- **Heterozygous coverage peak.** Estimated heterozygous coverage.

## Assembly graph

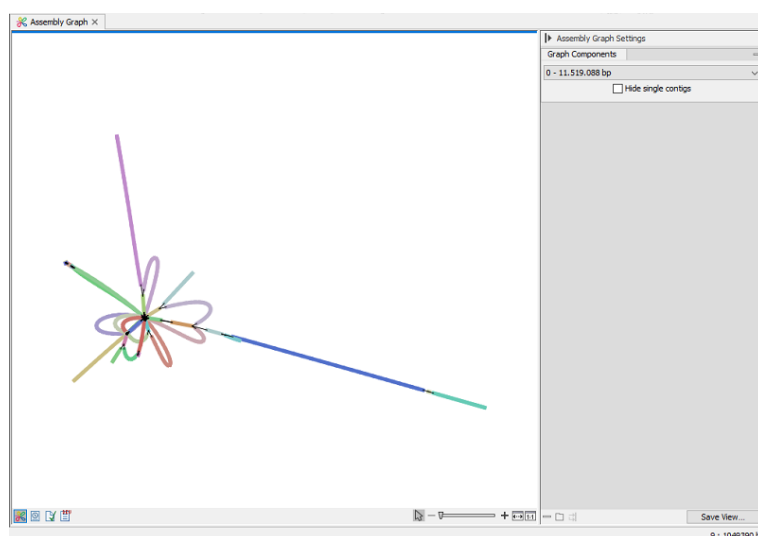


Figure 4.3: Assembly graph view

The assembly graph provides visual representations of the sequences in the contig list, and sequences that were excluded.

The viewer shows a single graph component (group of connected sequences) at a time. You switch between them using the Side Panel *Graph components* dropdown. To exclude graphs composed of a single contig, select **Hide single contigs**.

Hover your mouse over the colored line of a contig to see the name and length of the contig at the lower right corner of the view.

Dragging a contig initiates a layout animation. The animation aims to adhere to a graph layout governed by a set of force-directed rules. Consequently, when a contig is moved in a specific direction, the algorithm will strive to return to a low-energy state.

## Chapter 5

# Map Long Reads to Reference

The Map Long Reads to Reference tool enables aligning long reads to a reference with minimap2 [Li, 2018].

To run the tool, go to:

**Toolbox | Long Read Support (hnp) | Map Long Reads to Reference (hnp)**

Select one or more sequence lists containing long reads.

In the following step, select one or more reference sequences. You can select either individual sequences, a list of sequences or a sequence track as reference (figure 5.1).

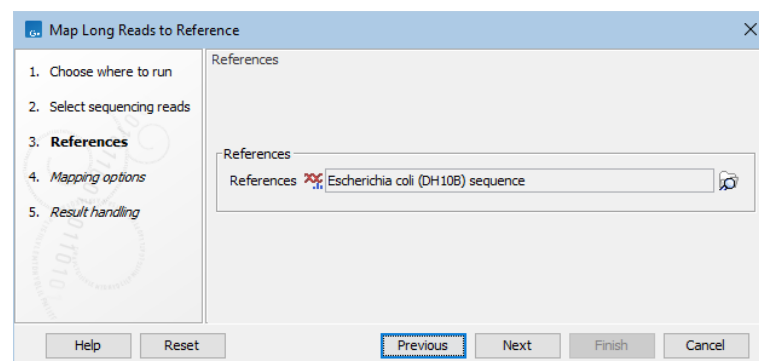


Figure 5.1: Map Long Reads to Reference reference selection.

In the Mapping options dialog, set the read mapping parameters (figure 5.2):

- **Mapping mode.** Choose between the following modes for parameter setting:
  - *Automatic.* Match cost parameters are set automatically based on the read type of the first input (e.g. Oxford Nanopore or Pacbio HiFi).
  - *Automatic spliced.* Similar to *Automatic*, except it generates spliced alignments, which can be useful for visualizing RNA-seq data.
  - *Manual.* Allows match costs to be specified manually. This will overwrite the read type specific match costs that would otherwise be used with the values entered below.
- **Match score.** Score of a match.

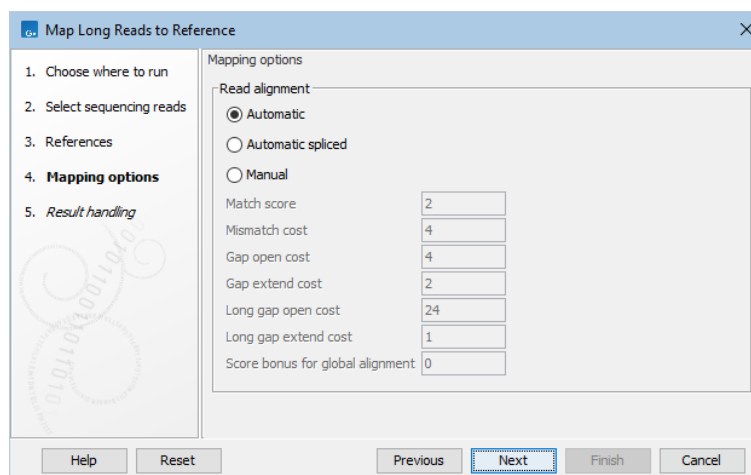


Figure 5.2: *Map Long Reads to Reference* mapping options.

- **Mismatch cost.** Cost of a mismatch.
- **Gap open cost.** Cost of starting a gap.
- **Gap extend cost.** Cost of extending a gap.
- **Long gap open cost.** Cost of starting a long gap. Long gaps are typically more expensive to open but cheaper to extend. The alignment will use the cheaper alternative.
- **Long gap extend cost.** Cost of extending a long gap.
- **Score bonus for global alignment.** A bonus that may be added to the alignment score if the alignment encompasses all nucleotides of the read.

For guidance on adjusting match cost parameters, see [http://resources.qiagenbioinformatics.com/manuals/clogenomicsworkbench/current/index.php?manual=Mapping\\_parameters.html](http://resources.qiagenbioinformatics.com/manuals/clogenomicsworkbench/current/index.php?manual=Mapping_parameters.html). Additional information on parameters are available from the minimap2 documentation.

## 5.1 Map Long Reads to Reference output

Outputs are selected from the dialog shown in figure 5.3.

The main choice in output format is at the top of the dialog - the read mapping can either be stored as a track or as a stand-alone read mapping. Both options have distinct features and advantages:

- **Reads track.** A reads track is best used in the context of a Track List, where additional information about the reference, consensus sequence or annotations can be added and viewed alongside the reads. Unless any specific functionality of the stand-alone read mapping is required, we recommend to using the tracks output for the additional flexibility it brings in further analysis.
- **Stand-alone read mapping.** This output is more elaborate than the reads track and includes the full reference sequence with annotations. A consensus sequence is created as part of



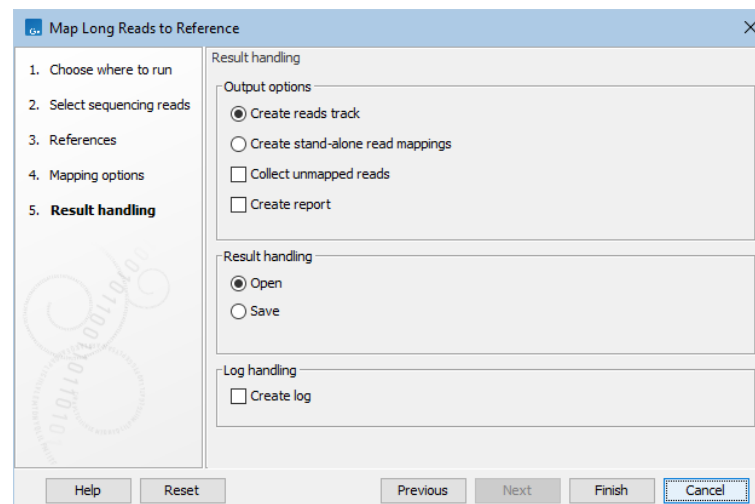


Figure 5.3: *Map Long Reads to Reference* output settings.

the output. Furthermore, the possibilities for detailed visualization and editing are richer than for the reads track. However, stand-alone read mappings do not lend themselves well to comparative analyses. Note that if multiple reference sequences are used as input, a read mapping table is created.

Read more about both output types at [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reads\\_tracks\\_stand\\_alone\\_read\\_mappings.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Reads_tracks_stand_alone_read_mappings.html).

In addition to the choice between the two main output options, the following output options are available:

- **Create report.** Creates a summary report.
- **Collect unmapped reads.** Collects all the reads that could not be mapped to the reference into a sequence list.

### Map Long Reads to Reference report

The report will summarize the results of the mapping process. An example of the first part of the report is shown in figure 5.4.

The information included in the report is:

- **Summary statistics.** A summary of the mapping statistics:
  - **References.** The count of reference sequences, the average length, and the total number of bases.
  - **Mapped reads.** The count of mapped reads, the percentage of mapped reads relative to total reads, the average read length, the total number of bases, and the percentage of bases relative to the total read count.
  - **Not mapped reads.** The count of un-mapped reads, the percentage of un-mapped reads relative to total reads, the average read length, the total number of bases, and the percentage of bases relative to the total read count.

## 1 Mapping summary report

### 1.1 Summary statistics

Input type: Single reads

	Count	Percentage of reads (%)	Average length	Number of bases	Percentage of bases (%)
References	1	-	4,686,137.00	4,686,137	-
Mapped reads	4	66.67	1,367,462.00	5,469,848	99.64
Not mapped reads	2	33.33	9,952.00	19,904	0.36
Total reads	6	100.00	914,958.67	5,489,752	100.00

### 1.2 Distribution of read length

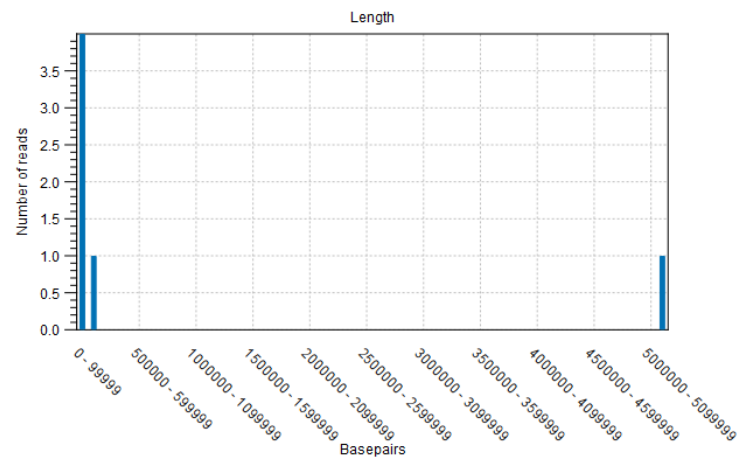


Figure 5.4: Map Long Reads to Reference summary report

- **Total reads.** The total count of reads, the average read length, and the total number of bases.
- **Distribution of read length.** A visual representation of the occurrence of each specific read length, offering insights into the overall pattern and variability in the dataset.
- **Distribution of mapped reads lengths.** Equivalent to the above, but including mapped reads only.
- **Distribution of un-mapped reads lengths.** Equivalent to the above, but including un-mapped reads only.

## Chapter 6

# Polish with Reads

The Polish with Reads tool facilitates the process of refining a set of sequences with high-quality reads. This enables the creation of hybrid assemblies by first creating an assembly from long, error prone reads, and subsequently using high-quality Illumina reads to polish the contigs.

Before polishing, high-quality reads should be stripped of adapters and lower quality bases. This can be done using **Trim Reads** ([http://resources.qiagenbioinformatics.com/manuals/cleogenomicsworkbench/current/index.php?manual=Trim\\_Reads.html](http://resources.qiagenbioinformatics.com/manuals/cleogenomicsworkbench/current/index.php?manual=Trim_Reads.html)).

The tool uses Racon [Vaser et al., 2017] with additional improvements inspired by Minipolish [Wick and Holt, 2019]. Racon uses a divide-and-conquer approach for rapid consensus calling. The partial order alignment (POA) of the reads against the target sequences occurs in non-overlapping windows on the target sequences. This approach has the consequence that Racon may not always use a globally optimal alignment of reads for consensus calling.

Polishing is conducted in two steps. Following each step, a set of corrections, inspired by minipolish [Wick and Holt, 2019], are carried out to improve contig quality:

- Contig ends that were truncated by Racon are reintroduced.
- Circular contigs are rotated by half the sequence length in each of the iterations.
- For circular contigs, the mapping is corrected for reads that span the junction.

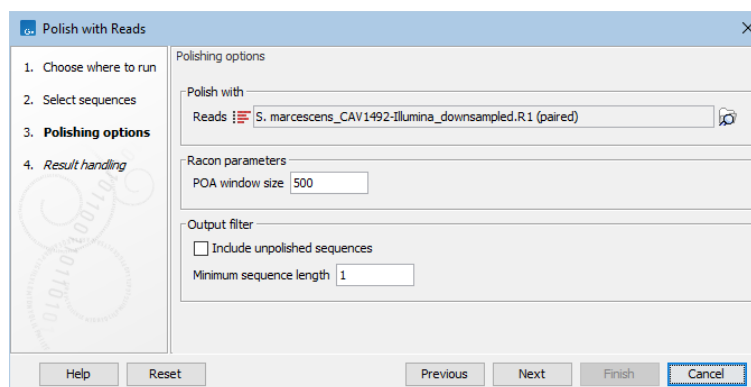
To run the Polish with Reads tool, go to:

**Toolbox | Long Read Support (🔧) | Polish with Reads (🔧)**

Select a sequence list containing contigs or long reads.

In the next dialog, set the polishing parameters (figure 6.1):

- **Reads.** Select a sequence list of trimmed, high-quality reads to be used for polishing.
- **POA window size.** The window size for which Racon computes partial order alignments (POA). A larger window size enhances the ability to capture more global structure during the polishing process, but it also increases the memory requirement.
- **Include unpolished sequences.** Check the checkbox to keep sequences for which polishing was not possible.

Figure 6.1: *Polish with Reads* parameters

- **Minimum sequence length.** The minimum length of sequences to be included in the output.

## 6.1 Polish with Reads output

In addition to the sequence list of polished sequences, a summary report is available via the output option **Create report**.

### Polish with Reads report

An example of the first sections of the Polish with Reads summary report is shown in figure 6.2.

The report contains the following information:

- **Polishing summary**
  - **Input sequences.** Number of target input sequences.
  - **Input reads.** Number of high-quality input reads.
  - **Alignments (includes multiple alignments).** The number of alignments between target sequences and high-quality reads.
  - **Polished sequences.** Number of sequences that were polished.
  - **Polished windows (%).** The percentage of windows that were polished.
- **Sequences.** For each target input sequence, the following information is given:
  - **Name.** Name of the sequence.
  - **Nucleotides.** Number of nucleotides in the sequence.
  - **Mapped reads.** Number of high-quality reads that mapped to the sequence.
  - **Polished windows (%).** The percentage of windows that were polished.
- **Sequence statistics**
  - **Sequences.** Number of sequences.
  - **Minimum, Maximum, Average.** Minimum, maximum and average sequence length.

**1 Polishing summary**

Input sequences	6
Input reads	300,000
Alignments (includes multiple alignments)	284,345
Polished sequences	6
Polished windows (%)	99.51

**2 Sequences**

Name	Nucleotides	Mapped reads	Polished windows (%)
Utg2138	5,204,845	151,491	99.82
Utg2140	268,426	6,321	99.63
Utg2144	94,163	17,726	81.48
Utg2132	69,147	2,250	100.00
Utg2136	3,229	84,858	100.00
Utg2142	199,410	21,699	99.75

**3 Sequence statistics**

Nucleotide	Count	Frequency (%)
Adenine (A)	1,205,076	20.64
Cytosine (C)	1,708,305	29.26
Guanine (G)	1,712,751	29.33
Thymine (T)	1,213,088	20.77

Sequences	6
Minimum	3,229
Maximum	5,204,845
Average	973,203
N50	5,204,845
N90	268,426
Total	5,839,220

Figure 6.2: *Polish with Reads report*

- **N50.** The length of the shortest sequence in sets of sequences of equal length or longer, where the summed length of sequences is at least 50% of the total sequence length.
- **N90.** The length of the shortest sequence in a set of sequences of equal length or longer, where the summed length of sequence is at least 90% of the total sequence length. N90 will be equal to or smaller than N50.
- **Total.** The number of bases in the sequences.
- **Sequence length distribution.** A graph depicting the number of sequences found at a specific length.

## Chapter 7

# Correct Long Reads

The Correct Long Reads tool enables the correction of a set of error-prone long reads by finding overlaps between the reads, and performing a consensus error correction using Racon [Vaser et al., 2017].

Note that if the aim is to create an assembly from a set of error-prone long reads it is not recommended to run **Correct Long Reads** prior to **De Novo Assemble Long Reads**. For other applications, however, it may be beneficial to correct the reads before further analysis.

To run the Correct Long Reads tool, go to:

**Toolbox | Long Read Support (hnp) | Correct Long Reads (nsp)**

Select one or more sequence lists containing reads.

In the next dialog, set the following parameters (figure 7.1):

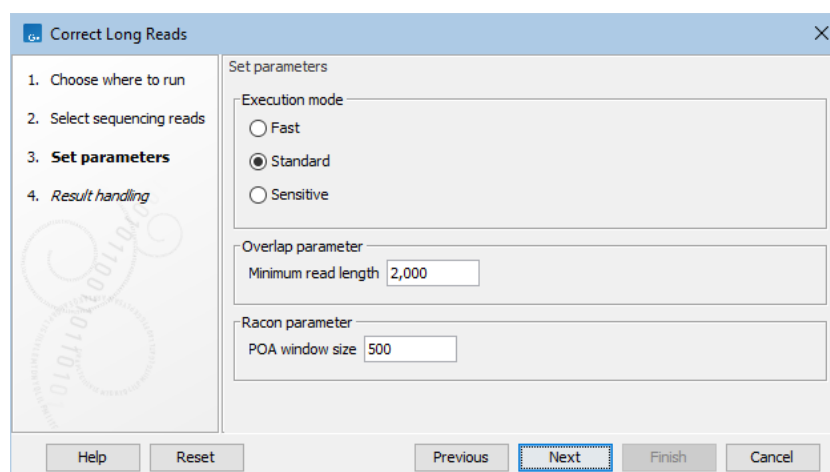
The image shows a software dialog box titled "Correct Long Reads". On the left, there is a vertical list of four steps: "1. Choose where to run", "2. Select sequencing reads", "3. Set parameters", and "4. Result handling". Step 3 is currently selected and highlighted. The main area of the dialog is titled "Set parameters" and contains three sections: "Execution mode" with three radio buttons (Fast, Standard, Sensitive), where "Standard" is selected; "Overlap parameter" with a text box for "Minimum read length" set to "2,000"; and "Racon parameter" with a text box for "POA window size" set to "500". At the bottom of the dialog, there are five buttons: "Help", "Reset", "Previous", "Next" (which is highlighted with a blue border), "Finish", and "Cancel".

Figure 7.1: *Correct Long Reads* parameters

- **Execution mode.**
  - **Fast.** An overlap mapper computes overlaps using a coarse all-to-all sequence alignment, prior to running Racon. This option will use the least amount of memory.

- **Standard.** Similar to the *Fast* , but with a higher sensitivity.
- **Sensitive.** Minimap2 is used to compute a more precise all-to-all sequence alignment prior to running Racon. This option should only be used for low-coverage datasets.
- **Minimum read length.** Reads shorter than this value will not be included in the correction process.
- **POA window size.** The window size for which Racon computes partial order alignments (POA). A larger window size enhances the ability to capture more global structure during the polishing process, but it also increases the memory requirement.

## 7.1 Correct Long Reads output

In addition to the sequence list of corrected reads, a summary report is available via the output option **Create report**.

### Correct Long Reads report

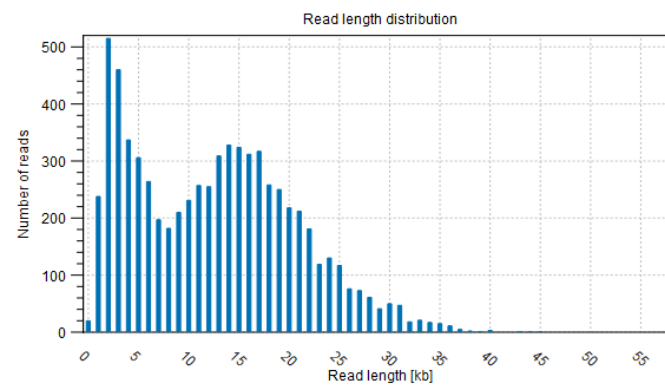
An example of the top part of the Correct Long Reads summary report is shown in figure 7.2.

The report is divided into two sections, *Input statistics* and *Output statistics*, both containing the following information:

- **Reads.** Number of reads.
- **Minimum, Maximum, Average.** Minimum, maximum and average sequence length.
- **N50.** The length of the shortest sequence in sets of sequences of equal length or longer, where the summed length of sequences is at least 50% of the total sequence length.
- **N90.** The length of the shortest sequence in a set of sequences of equal length or longer, where the summed length of sequence is at least 90% of the total sequence length. N90 will be equal to or smaller than N50.
- **Total.** The number of bases in the sequences.
- **Sequence length distribution.** A graph depicting the number of sequences found at a specific length.

**1 Input statistics**

Nucleotide	Count	Frequency (%)
Adenine (A)	18,412,271	20.66
Cytosine (C)	25,880,634	29.04
Guanine (G)	26,068,337	29.25
Thymine (T)	18,749,472	21.04
Reads		7,039
Minimum		232
Maximum		57,649
Average		12,660
N50		17,593
N90		7,743
Total		89,110,714

**2 Output statistics**

Nucleotide	Count	Frequency (%)
Adenine (A)	7,734,576	21.86
Cytosine (C)	9,991,371	28.24
Guanine (G)	9,925,504	28.05
Thymine (T)	7,734,838	21.86
Reads		3,343
Minimum		1,973
Maximum		40,760
Average		10,585
N50		14,537
N90		5,383

Figure 7.2: *Correct Long Reads report.*



## Chapter 8

# RNA-Seq Analysis for Long Reads

The RNA-Seq Analysis for Long Reads tool supports analysis of RNA-Seq data by mapping sequencing reads to an annotated reference genome with minimap2 [Li, 2018] and distributing and counting the reads across genes and transcripts. Subsequently, the results can be used for expression analysis.

RNA-Seq analysis with long reads is done in several steps: First, all annotated transcripts or genes are extracted. If there are several annotated splice variants, they are all extracted. Next, the reads are mapped against all the transcripts, and to the whole genome using minimap2. For more information about the read mapper, see chapter 5.

From this mapping, the reads are categorized and assigned to the transcripts using the EM estimation algorithm, and expression values for each gene are obtained by summing the transcript counts belonging to the gene.

Detailed information on RNA-Seq analysis including the EM algorithm is found at [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=\\_EM\\_estimation\\_algorithm.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=_EM_estimation_algorithm.html).

The results can be used as input for expression analysis and other downstream RNA-Seq analysis tools in *CLC Genomics Workbench*, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA\\_Seq\\_Tools.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_Tools.html).

### 8.0.1 Reads and reference settings

To run the RNA-Seq Analysis for Long Reads tool, go to:

**Toolbox | Long Read Support  | RNA-Seq Analysis for Long Reads **

Select one or more sequence lists containing long reads.

In the *Reference settings* dialog (figure 8.1), at the top there are three options concerning how the reference sequences are annotated.

- **Genome annotated with genes and transcripts.** This option should be used when both gene and mRNA annotations are available. When this option is enabled, the EM will distribute the reads over the transcripts. Gene counts are then obtained by summing over the (EM-distributed) transcript counts. The mRNA annotations are used to define how the

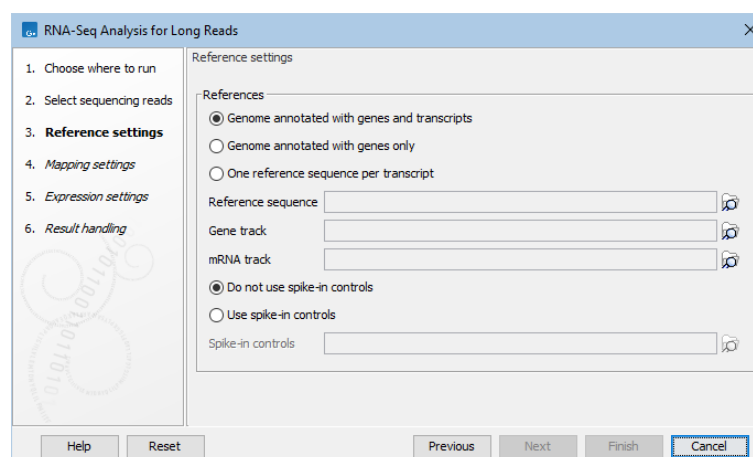


Figure 8.1: RNA-Seq Analysis for Long Reads reference settings.

transcripts are spliced. This option should be used for Eukaryotes since it is the only option where splicing is taken into account. Note that genes and transcripts are linked by name only (not by position, ID etc).

When this option is selected, both a *Gene* and an *mRNA* track should be provided in the boxes below. Annotated reference genomes can be obtained in various ways:

- Directly downloaded as tracks using the Reference Data Manager.
- Imported as tracks from fasta and gff/gtf files.
- Imported from Genbank or EMBL files and converted to tracks.
- Downloaded from Genbank.

When using this option, Expression values, RPKM and TPM are calculated based on the lengths of the transcripts provided by the mRNA track. If a gene's transcript annotation is absent from the mRNA track, all values will be set to zero unless the option "Calculate expression for genes without transcript" is checked in a later dialog.

- **Genome annotated with genes only.** This option should be used for Prokaryotes where transcripts are not spliced. When this option is selected, a *Gene* track should be provided in the box below. The data can be obtained in the same ways as described above.

When using this option, Expression values, RPKM and TPM are calculated based on the lengths of the genes provided by the Genes track.

- **One reference sequence per transcript.** This option is suitable for situations where the reference is a list of sequences. Each sequence in the list will be treated as a "transcript" and expression values are calculated for each sequence. This option is most often used if the reference is a product of a *de novo* assembly of RNA-Seq data. It is also a suitable option for references where genes are particularly close to each other or clustered in operon structures. When this option is selected, only the reference sequence should be provided, either as a sequence track or a sequence list. Expression values, RPKM and TPM are calculated based on the lengths of sequences from the sequence track or sequence list.

At the bottom of the dialog you can choose between these two options:

- **Do not use spike-in controls.**
- **Use spike-in controls.** When selected, you provide a spike-in control file in the field below.

To learn how to import spike-in control files, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Import\\_RNA\\_spike\\_in\\_controls.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Import_RNA_spike_in_controls.html).

## 8.0.2 Mapping settings

In the next dialog, set the mapping parameters (figure 8.2).

Figure 8.2: Define mapping settings for RNA-Seq Analysis for Long Reads.

The generic mapping parameters are identical to those applied for **Map Long Reads to Reference**, see chapter 5.

In addition, the following RNA-Seq specific setting is available:

- **Maximum number of hits for a read.** Reads mapping equally well to more places than this number will be ignored. If a read matches to multiple distinct places, but less than or equal to the specified maximum number, it will be assigned to one of these places by the EM algorithm.

## 8.0.3 Expression settings

Expression settings are defined in the dialog shown in figure 8.3.

These parameters determine the way expression values are counted.

- **Strand setting**
  - **Both.** Reads are mapped both in the same and reversed orientation as the transcript from which they originate. This is the default.
  - **Forward.** Reads are mapped in the same orientation as the transcript from which they originate.
  - **Reverse.** Reads are mapped in the reverse orientation as the transcript from which they originate.

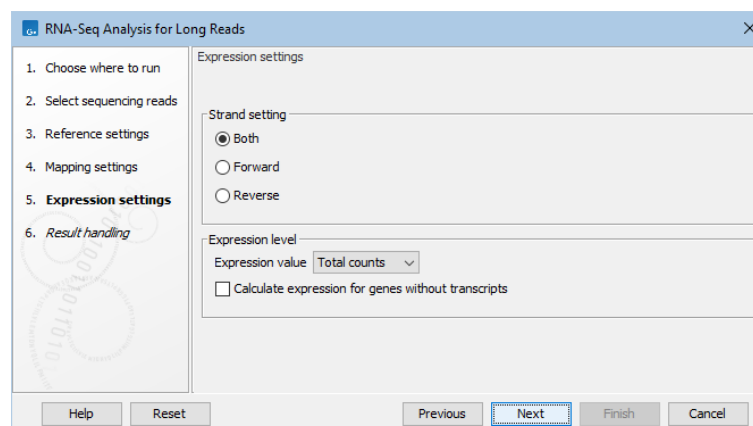


Figure 8.3: Set strand setting and define how expression values should be calculated.

If a strand specific protocol for read generation was used, choose the corresponding setting. This allows assignment of the reads to the right gene in cases where overlapping genes are located on different strands.

- **Expression value.** This parameter describes how expression per gene or transcript can be defined in different ways on both levels:
  - **Total counts.** When the reference is annotated with genes only, this value is the total number of reads mapped to the gene. For un-annotated references, this value is the total number of reads mapped to the reference sequence. For references annotated with transcripts and genes, the value reported for each gene is the number of reads that map to the exons of that gene. The value reported per transcript is the total number of reads mapped to the transcript.
  - **Unique counts.** This is similar to the above, except only reads that are uniquely mapped are counted.
  - **TPM.** (Transcripts per million). This is computed as  $\frac{RPKM \cdot 10^6}{\sum RPKM}$ , where the sum is over the RPKM values of all genes/transcripts.
  - **RPKM.** This is a normalized form of the "Total counts" option (see more under *Definition of RPKM* below).

All values are present in the output. The choice of expression value only affects how Expression Tracks are visualized in the track view but the results will not be affected by this choice as the most appropriate expression value is automatically selected for the analysis being performed. For detection of differential expression this is the "Total counts" value, and for the other tools this is a normalized and transformed version of the "Total counts" as described below.

- **Calculate expression for genes without transcripts.** For genes without annotated transcripts, the RPKM cannot be calculated since the total length of all exons is needed. By selecting this option, the length of the gene will be used in place of an "exon length". If the option is not checked, there will be no RPKM value reported for those genes.

### Definition of RPKM

RPKM, Reads Per Kilobase of exon model per Million mapped reads, is defined in this way [Mortazavi et al., 2008]:

$$RPKM = \frac{\text{total exon reads}}{\text{mapped reads(millions)} \times \text{exon length (KB)}}$$

For prokaryotic genes and other non-exon based regions, the calculation is performed in this way:

$$RPKM = \frac{\text{total gene reads}}{\text{mapped reads(millions)} \times \text{gene length (KB)}}$$

**Total exon reads.** This value can be found in the column with header **Total exon reads** in the expression track. This is the number of reads that have been mapped to exons (either within an exon or at the exon junction). When the reference genome is annotated with gene and transcript annotations, the mRNA track defines the exons, and the total exon reads are the reads mapped to all transcripts for that gene. When only genes are used, each gene in the gene track is considered an exon. When an un-annotated sequence list is used, each sequence is considered an exon.

**Exon length.** This is the number in the column with the header **Exon length** in the expression track, divided by 1000. This is calculated as the sum of the lengths of all exons (see definition of exon above). Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region.

**Mapped reads.** The sum of all mapped reads as listed in the RNA-Seq analysis report. For more information on how expression is calculated in this case, see above.

## 8.1 RNA-Seq Analysis for Long Reads output

Output options are specified in the dialog shown in figure 8.4.

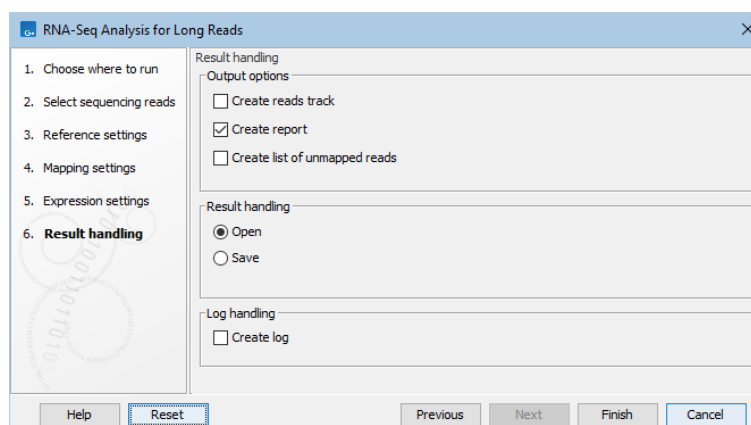


Figure 8.4: The output options for RNA-Seq using long reads.

The main results of the RNA-Seq analysis are Expression Tracks. A track with a name ending with **(GE)** summarizes expression at the gene level. If the "Genome annotated with genes and

transcripts" option was selected, a second track, with a name ending in **(TE)**, is produced, which summarizes expression at the transcript level.

In addition, the following outputs are available:

- **Create reads track.** This track contains the mapping of the reads to the references.
- **Create report.** A summary report.
- **Create list of unmapped reads.** This list is made of reads that did not map to the reference, or that were non-specific matches with more placements than specified.

## Expression tracks

Expression tracks can be shown in a **Table** (📊) and a **Track** (📈) view.

The expression track table view has the following options (figure 8.5).

- The "Filter to selection" only displays pre-selected rows in the table.
- The "Create track from Selection" will create a new Track with the selected rows.
- The "Select Genes/Transcripts in Other Views" button finds and selects the currently selected genes or transcripts in all other open expression track table views.
- The "Copy Gene/Transcript Names to Clipboard" button copies the currently selected gene or transcript names to the clipboard.

Rows: 33,538		Table view: Genome		Filter to Selection...	Filter
Name	TPM	Exons	Biotype	Transcripts annota...	Uniquely identified ...
gene07104-v2.0.a2-hybrid	7,737.70		11 Gene	1	1
gene08023-v2.0.a2-hybrid	7,175.25		14 Gene	3	3
gene26994-v2.0.a2-hybrid	7,114.26		5 Gene	1	1
gene19551-v2.0.a2-hybrid	6,864.27		7 Gene	3	3
gene22974-v1.0-hybrid	6,122.62		12 Gene	6	6
gene22465-v1.0-hybrid	6,032.70		3 Gene	1	1
gene01495-v1.0-hybrid	5,845.13		5 Gene	2	2
gene21573-v1.0-hybrid	5,819.01		8 Gene	3	3
gene18240-v2.0.a2-hybrid	5,668.02		5 Gene	2	2
gene30512-v1.0-hybrid	5,413.05		11 Gene	1	1
gene13212-v1.0-hybrid	5,192.39		1 Gene	1	1
gene01608-v1.0-hybrid	4,871.66		7 Gene	2	2
gene20927-v1.0-hybrid	4,863.98		6 Gene	2	2
gene03809-v1.0-hybrid	4,535.41		6 Gene	3	3
gene09659-v2.0.a2-hybrid	4,431.43		1 Gene	1	1
gene26083-v2.0.a2-hybrid	4,253.76		11 Gene	3	1
gene25105-v2.0.a2-hybrid	4,126.40		11 Gene	2	2
gene28639-v1.0-hybrid	4,080.65		5 Gene	2	2
gene26908-v1.0-hybrid	3,939.64		2 Gene	1	1
gene23293-v1.0-hybrid	3,799.04		6 Gene	2	2
gene06563-v1.0-hybrid	3,686.90		6 Gene	2	2
gene11464-v1.0-hybrid	3,648.93		7 Gene	1	1
gene20983-v1.0-hybrid	3,528.43		5 Gene	1	1

📈 Create Track from Selection
Select Genes in Other Views
Copy Gene Names to Clipboard

Figure 8.5: A gene-level expression track shown in Table view.

For detailed information on expression tracks, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Expression\\_tracks.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Expression_tracks.html).

## RNA-Seq reads track

A track containing the mapped reads can be generated by the tool if the option to do so is enabled. The graphical view of the mapped reads can be shown together with the expression tracks in a **Track list**. For additional information on this output, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Expression\\_tracks.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Expression_tracks.html).

## RNA-Seq report

An example of first part of the RNA-Seq report is shown in figure 8.6.

For a detailed description of the different sections in the report, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA\\_Seq\\_report.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=RNA_Seq_report.html).

### 1 Selected input sequences

#### 1.1 Sequence reads

Name	Number of reads	Longest read	Paired
PacBio SRR3497430	442,601	11,501	no

For paired data, there are two reads in a pair.

#### 1.2 Reference sequences

References	Length	Genes	Transcripts
8	211,673,467	33,538	50,732

### 2 References

#### 2.1 Transcripts per gene

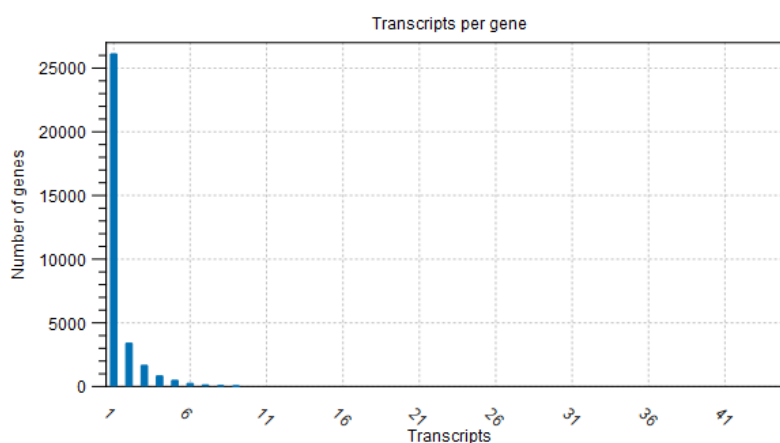


Figure 8.6: Report of an RNA-Seq Analysis for Long Reads.

## Chapter 9

# Structural Variant Caller for Long Reads

The Structural Variant Caller for Long Reads tool calls structural variants of length  $\geq 35$  from long reads. The input read mapping must have been produced by the Map Long Reads to Reference tool using Long Read Support 24 or later.

The tool is designed for Whole Genome Sequencing data. Targeted sequencing is not explicitly supported. If this is necessary, a workaround may be to generate a pseudo reference genome consisting purely of the targeted regions using the Extract Annotated Regions tool (see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Extract\\_Annotated\\_Regions.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Extract_Annotated_Regions.html)).

The tool does not have specific handling for circular chromosomes. Circular chromosomes are treated as if they were linear. This means that variants near the origin may go undetected because the part of the read wrapping around the origin is ignored.

### 9.1 Structural Variant Caller settings

To run the Structural Variant Caller for Long Reads tool:

**Toolbox | Long Read Support (📖) | Structural Variant Caller for Long Reads (🔧)**

The tool accepts a single read mapping as input.

There are two calling modes, as shown in figure 9.1.

- **Germline.** Recommended for diploid samples. Variants with a frequency inconsistent with a diploid organism are filtered away.
- **Somatic.** Recommended for calling on samples where variants may appear at low frequencies. When this is enabled:
  - The diploid filtering of the **Germline** mode is disabled.
  - Variants are never filtered based on having too few supporting reads.
  - Strict filtering of deletions and duplications that are longer than 50,000bp is disabled.
  - Variants with frequencies between 5% and 30% are subjected to stricter filtering. This includes:



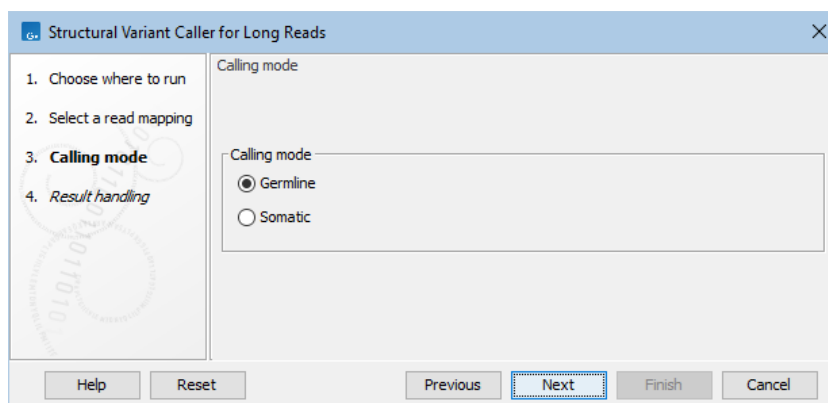


Figure 9.1: The calling mode options for Structural Variant Caller for Long Reads.

- \* All variants other than long insertions are only reported if evidence for them is observed on reads mapping in both orientations.
- \* Variants are filtered away if they are supported by alignments with many mismatches and indels compared to the reference.
- \* Inversions and duplications are only reported if they are at least 500bp long.
- \* Variants whose coverage changes by more than 10% when moving from upstream of the variant to the start position, or from the start position to the center of the variant, or from the center of the variant to the end position, or from the end position to downstream of the variant are filtered away.

## 9.2 Structural Variant Caller output

The tool has the following output options, as shown in figure 9.2:

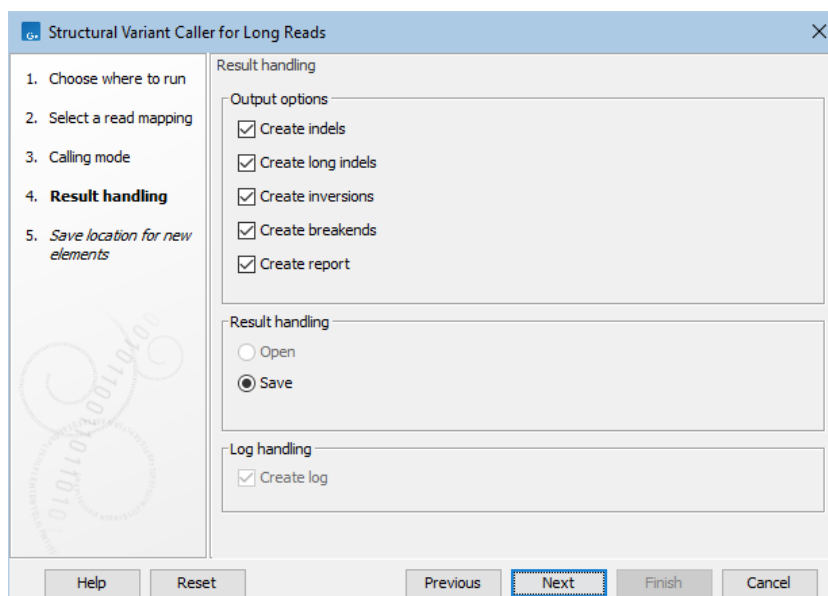







Figure 9.2: The output options for Structural Variant Caller for Long Reads.

- **Indels (Indels)** . A variant track with indels (deletions and insertions - including tandem duplications) that have lengths up to 100,000 bp.
- **Long indels (Indels long)** . An annotation track with long indels (those with lengths larger than 100,000 bp). The reason for putting indels larger than 100,000 bp in a separate annotation track, is that these variants have either long allele or reference entries in the variant track, which make them challenging to work with in the track viewer.
- **Inversions (Inv)** . An annotation track with inversions.
- **Breakends (Breakends)** . An annotation track with a row for each breakend in a translocation.
- **Report** . A report giving an overview over analyzed references and found structural variants.

All outputs (other than the Report) can be exported together to a single VCF file, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Export\\_in\\_VCF\\_format.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Export_in_VCF_format.html). The VCF-exportable outputs contain the following annotations:

- **Chromosome**. The chromosome on which the variant is located.
- **Region**. The location of the variant.
- **Zygosity**. The zygosity of the variant called, as determined by the variant detection tool. This will be either 'Homozygous', where there is only one variant called at that position or 'Heterozygous' where more than one variant was called at that position.
- **Count**. The estimated count for the structural variant. In the case of insertions longer than 2500 bp, this is increased by an ad hoc procedure to account for reads that are not observed because they align within the insertion. If the coverage cannot be determined then this is reported as '0'. It is therefore advised to perform any desired filtering of variants on the 'Raw count' annotation rather than this annotation.
- **Coverage**. The estimated coverage for the structural variant. This is determined in different ways for different structural variant types. For insertions, this is the coverage at the location of the insertion; for duplications it is the average of the coverage at the two ends of the duplication; for inversions it is the average at positions upstream and downstream of the inversion; for deletions and breakends it is the average of the coverage at the start, end, and center of the variant. When taking averages, locations with no reads are ignored. In rare cases the coverage cannot be determined, because none of the locations used for calculating it have any reads. In these cases, the coverage will be reported as '0'.
- **Frequency**. The count divided by the coverage, reported as a percentage.
- **Average quality**. The average mapping quality of the alignments supporting the variant. The mapping quality is reported on the Phred scale, and describes the probability that the alignment is incorrect. The estimates are made by minimap2. A value of 30 means a 1 in 1000 chance of incorrect alignment. Very low average qualities are not seen, because the algorithm filters alignments with low mapping quality. Note that other variant calling tools, such as the Low Frequency Variant Detection tool, use this annotation to report the average quality scores of reads on the Phred scale rather than the average quality of alignments on the Phred scale.

- **Stdev position.** An estimate of the standard deviation of the position of the variant.

Additional annotations present on more than one output are:

- **Length.** The length of the variant. For deletions, it is the length of the deleted sequence, and for insertions and duplications it is the length of the inserted sequence. For inversions, it is the length of the inverted region.
- **Allele.** The inserted sequence. This is only reported for insertions and duplications.
- **Forward read count.** For inversions, the number of countable reads supporting the 5' side of the inversion on the reference. For other types of variant, the number of reads supporting the allele and mapping in the forward direction. The 'countable' reads are those that are used by the variant detection tool when calling the variant.
- **Reverse read count.** For inversions, the number of countable reads supporting the 3' side of the inversion on the reference. For other types of variant, the number of reads supporting the allele and mapping in the reverse direction. The 'countable' reads are those that are used by the variant detection tool when calling the variant.
- **Raw count.** The number of countable reads supporting the allele. The 'countable' reads are those that are used by the variant detection tool when calling the variant.
- **Detailed type.** This is only present for insertions and is set to "Insertion" if the variant describes a novel insertion, and "Duplication" if the insertion instead describes a duplication.
- **Repeat unit size.** Duplications are sometimes detected as insertions within a read. In these cases, their reported "Length" is estimated from these insertions. Because the reference sequence may be duplicated more than once, this can be larger than the duplicated region on the reference. This is an estimate of the size of one copy on the reference. It is equal to the "Length" if it cannot be estimated. In rare circumstances, the "Repeat unit size" may be estimated to exceed the "Length". This is usually caused when a repeat is inserted between two existing repeats of the same kind.
- **Stdev length.** An estimate of the standard deviation of the length of the variant.

### Indels variant track

The indels track uses many of the standard variant annotations, see [http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant\\_tracks.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Variant_tracks.html).

### Long indels variant track

This track contains insertions and deletions larger than 100,000 bp. It is often enriched for false positive calls. This is because deletions and duplications are called when a read maps to two disjoint locations. If these two locations are at either end of a chromosome, then a near-whole chromosome deletion (or duplication) will be called. In many cases, it is more likely that the read maps to two places because an insertion is present that shares homology with one of the locations.

It is sometimes possible to detect false positives. For example, if the sample is germline, and other structural variants are called within a long homozygous deletion, then it is likely that the long homozygous deletion is a false positive.

### Inversions track

This track is often enriched for false positive calls. This is because an inversion may be called when a read maps to two disjoint locations on the same chromosome and in different orientations. If these two locations are at either end of a chromosome, then a near-whole chromosome inversion will be called. In many cases, it is more likely that the read maps to two places because an insertion is present that shares homology with one of the locations.

When coverage is high, it is often possible to detect false positives by requiring that there is support for both sides of the inversion. Reads supporting the 5' side of the inversion on the reference are counted as "forward" reads, and reads supporting the 3' side of the inversion on the reference are counted as "reverse" reads. Each variant reports these in "Forward read count" and "Reverse read count" annotations respectively.

Another class of false positives are inversions that start or end at the same location as an insertion. This is sometimes a signature of an inverted repeat.

### Breakend track

The breakend track can be used to look for translocations and other complex rearrangements that involve more than one chromosome. The definition of a breakend that we use here closely follows that from the VCF specification. Please refer to Section 5.4 "Specifying complex rearrangements with breakends" of <https://samtools.github.io/hts-specs/VCFv4.4.pdf>. Specifically we support the cases shown in figures 1, 4, 5, and 7 of that section.

Annotations that are only present on the breakends output are:

- **Filter.** This is always PASS. It is necessary for VCF export.
- **Breakpoint type.** 5' if the breakend is on an earlier chromosome than its mate, 3' if it is on a later chromosome.
- **Type.** Either "donor" or "acceptor". Reads mapping to a "donor" breakend are mapped to the left of the breakend, disappear at the breakend location, and reappear on another chromosome. Reads mapping to an "acceptor" breakend appear at the breakend location and continue to the right of the location.
- **Fusion crossing reads.** The number of distinct reads that support the breakend. Note that this is not necessarily the number of reads that support the fusion, because the same breakend may take part in more than one fusion.
- **Fusion number.** Two breakends that share a fusion number describe one fusion. The read aligns up to one breakend on one chromosome, and then continues aligning at the other breakend on a different chromosome. If a breakend is involved in more than one fusion, it will appear more than once in the track, with a different fusion number for each fusion.

A simple reciprocal translocation involves 4 breakends: an acceptor and donor on each of the two chromosomes involved in the translocation. The 4 breakends will have different combinations of Name and Type, and two different Fusion numbers.

The easiest way to find translocations is to:

- Open the table view of the Breakend track and sort by **Fusion number** by clicking the column header.
- Look for two nearby fusion pairs in the sorted table with the same **Chromosome** and similar **Region**.
- Verify that each fusion pair consists of one donor and one acceptor type, and that each chromosome contains one donor and one acceptor type.

Note that the **Region** for a breakend is sometimes on the plus strand (e.g. 123456^123457) and sometimes on the minus strand (e.g. complement(123456^1234567)). There is no significance to the reported strand - it is used by the VCF exporter.

### 9.3 Structural Variant Caller algorithm

The tool is based on Sniffles2 v2.2. Results are therefore expected to be similar but not identical to those of Sniffles2 v2.2. For details on the algorithm, please refer to the Sniffles2 preprint <https://www.biorxiv.org/content/10.1101/2022.04.04.487055v2.full>.

The principal differences from Sniffles2 v2.2 are outlined below:

- Insertions contained within a primary alignment of a read are remapped to determine if they are duplications. This reduces the number of duplications that are erroneously reported as both an insertion and a duplication.
- Breakends that are close to an insertion of known length are not called when the longest read supporting the breakend is shorter than the shortest estimate of the insertion length. This is because it is likely that the inserted sequence is homologous to a region on another chromosome, and that the reads supporting the breakend are the reads that did not extend through the insertion.
- The full supplementary alignments for a read are always used, rather than the summary of the alignments provided by minimap2 in the SA SAM flag. Supplementary alignments describe the alignment of parts of a read that are not aligned in the primary alignment. This change is expected to give more precise structural variant locations and lengths in some cases.
- The consensus sequence of insertions is determined from all the reads supporting the insertion.
- Breakends are paired, and are only reported if the two locations supported by a read passing through the breakend are called. This leads to more interpretable breakends, but at the cost that rearrangements where only one breakend meets quality control cutoffs are not visible.


# Chapter 10

## Install and uninstall plugins

Long Read Support is installed as a plugin.

### 10.1 Installation of plugins

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

Plugins and modules are installed and uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins** (  ) **button** in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins... (  )**

The Plugin Manager has two tabs at the top:

- **Manage Plugins** An overview of your installed plugins and modules is provided under this tab.
- **Download Plugins** Plugins and modules available to download and install are listed in this tab.

To install a plugin, click on the **Download Plugins** tab (figure 10.1). Select a plugin. Information about it will be shown in the right hand panel. Click on the **Download and Install** button to install the plugin.

#### Accepting the license agreement

The End User License Agreement (EULA) must be read and accepted as part of the installation process. Please read the EULA text carefully, and if you agree to it, check the box next to the text **I accept these terms**. If further information is requested from you, please fill this in before clicking on the **Finish** button.

#### Installing a cpa file

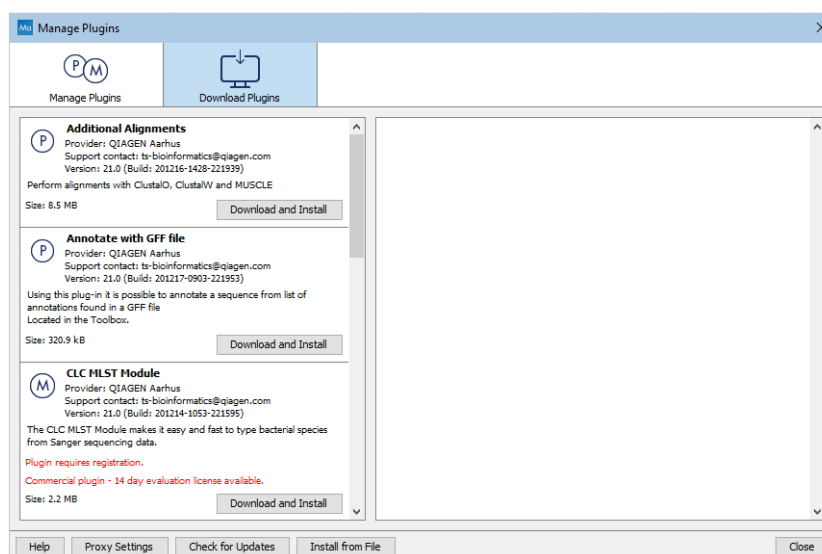


Figure 10.1: Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.

If you have a .cpa installer file for Long Read Support, 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.

## 10.2 Uninstalling plugins

Plugins and modules are uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins (P)** button in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins... (P)**

This will open the Plugin Manager (figure 10.2). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

To uninstall a plugin or module, click on its entry in the list, and click on the **Uninstall** button.

Plugins and modules are not uninstalled until the Workbench is restarted. When you close the 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

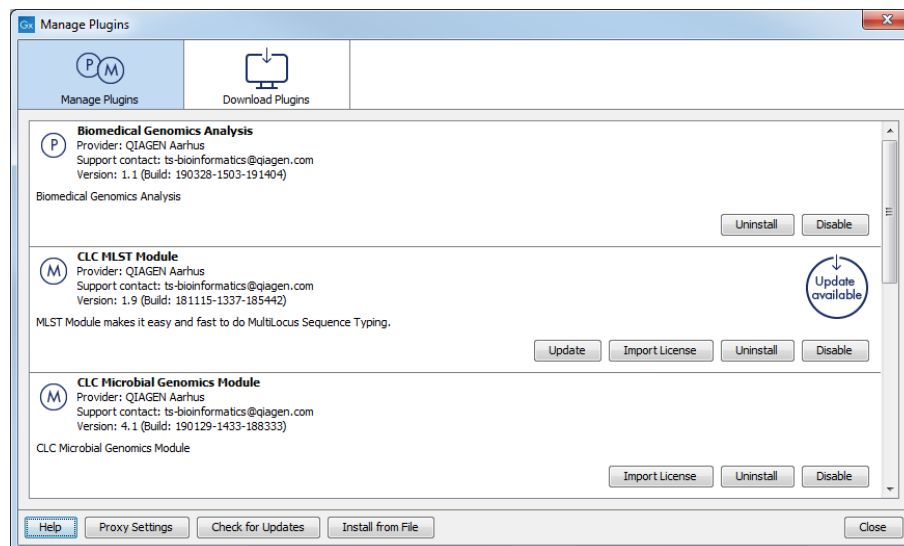


Figure 10.2: Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.

list under the Manage Plugins tab and click on the **Disable** button.



# Bibliography

- [Cheng et al., 2021] Cheng, H., Concepcion, G., Feng, X., Zhang, H., and H., L. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, 18(2):170–175.
- [Li, 2018] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.
- [Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628.
- [Vaser and Šiki, 2021] Vaser, R. and Šiki, M. (2021). Time-and memory-efficient genome assembly with raven. *Nature Computational Science*, 1(5):332–336.
- [Vaser et al., 2017] Vaser, R., Sovi, I., Nagarajan, N., and Šiki, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746.
- [Wick and Holt, 2019] Wick, R. R. and Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, 8:2138.