QIAseq Targeted Panel Analysis Plugin

USER MANUAL

User manual for

QIAseq Targeted Panel Analysis 1.2

Windows, macOS and Linux

October 30, 2018

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Chapter 1

Introduction to QIAseq Targeted Panel Analysis 1.2

QIAseq Targeted Panel Analysis 1.2 covers three different types of applications:

- Targeted DNA for variant calling
- Targeted RNA for differential expression
- Targeted RNAscan for fusion gene detection

In addition, the QIAseq Targeted Panel Analysis plugin includes the **Analyze QIAseq Panels** guide (figure 1.1) that facilitates the configuration and the running of a QIAseq panel analysis. Each secondary analysis workflow listed in the guide has been designed and calibrated for optimal variant detection on data generated by the corresponding QIAseq panel.

As the Analyze QIAseq Panels guide simplifies most steps necessary when running a QIAseq workflow, it is particularly suitable for first-time users, even when using QIAseq custom panels. However, the plugin also offers access to all application-specific workflows, as well as the independent tools that were used to build the ready-to-use workflows, or to add up to the analysis of the output data. These tools are available in a folder called QIAseq Panel Expert Tools. Note that the plugin supports panels where, when using paired reads, the UMI has been placed on Read 2. Using a different design will generate sub-optimal results.

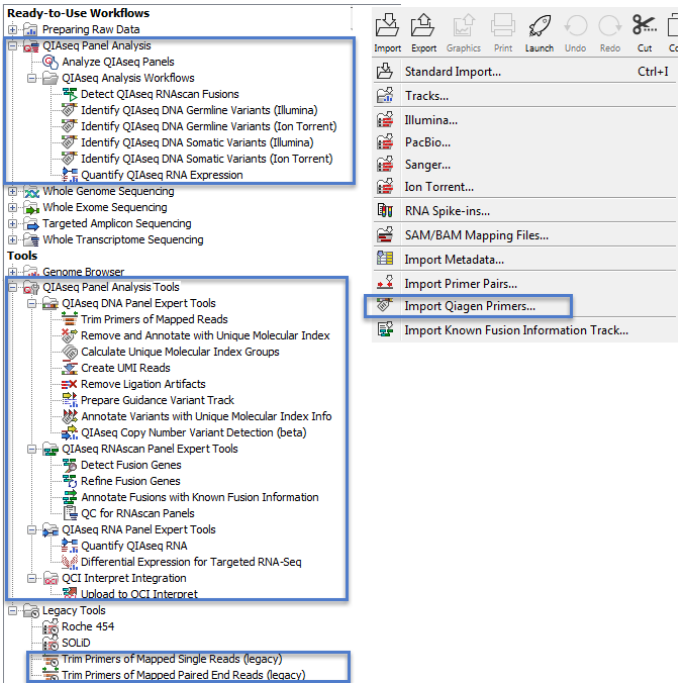


Figure 1.1: *QIAseq Targeted Panel Analysis guide, workflows and tools are available from the Toolbox after installation of the plugin.*

Chapter 2

Guide

Contents

2.1 Data management

QIAseq ready-to-use workflows need to be configured with the relevant reference data before being run for the first time. The workbench provides a Data Management interface that allows you to download easily the relevant application-specific Reference Data Set(s). If you are connected to a CLC Server via your Workbench, you need to download the references on the server even if you intend to run the ready-to-use workflows accessible from the guide locally.

To download a Reference Data Set, go to:

Toolbar | **Data Management** ()

This will open the wizard shown in figure 2.1.

On the top right corner, you can select where you want you references to be saved. If you are connected to a CLC Server via your Workbench, we recommend to choose **On Server** from the drop-down menu.

Select the Reference Data Set relevant to the application you want to use:

- QIAseq DNA Panels hg19 for Targeted DNA workflows
- QIAseq RNAscan Panels hg38 for Targeted RNAscan workflows
- QIAseq RNA Panels hg38 for Targeted RNA workflows (Note that both hg38 AND Mouse are needed to run the Mouse workflows.)

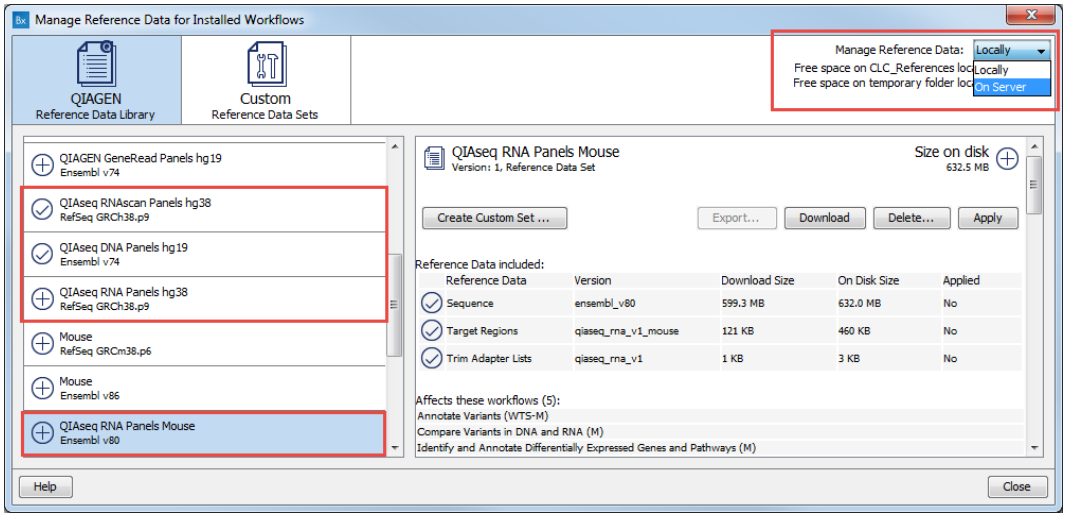


Figure 2.1: *The Data Management interface.*

Note that in the targeted regions files provided in the Reference Data Sets, target region covers 5 bp from each side of the exon.

Click on the button labeled **Download**. Once the download is complete, the Data Set icon changes from a plus to a check mark.

If you intend to run the ready-to-use workflows using the Analyze QIAseq Panels guide, you can close the Data Manager. However, if you are planning to run the workflows on their own, you need to click on **Apply** for the reference corresponding to the application you will run before closing the Data Manager.

2.2 The Analyze QIAseq Panels guide

The guide can be found in the Toolbox here:

Ready-to-Use Workflows | **QIAseq Panel Analysis** | **Analyze QIAseq Panels**

Double-click to get started with a panel analysis (figure 2.2).

2.2.1 Import your reads

Use the relevant Import button to import the reads in the Navigation Area of the workbench.

Click on the folder icon ( ) to the right of the "Select files" field, and select the files you want to import.

The Illumina High-Throughput Sequencing Import offers the following options (figure 2.3):

- **General options**
  - Discard sequence names: For high-throughput sequencing data, the naming of the individual reads is often irrelevant given the amount of reads. This option allows you to discard read names to save disk space.

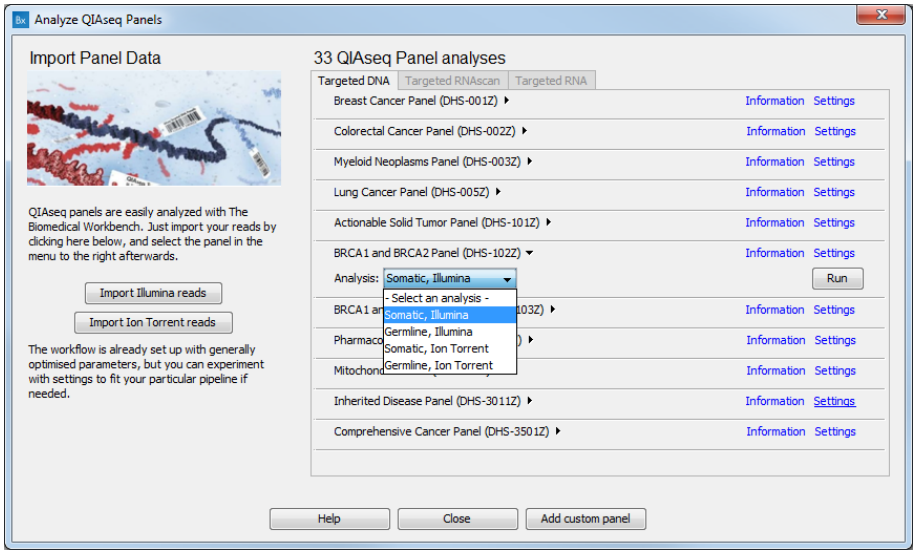


Figure 2.2: *The QIAseq Panel Analysis guide interface.*

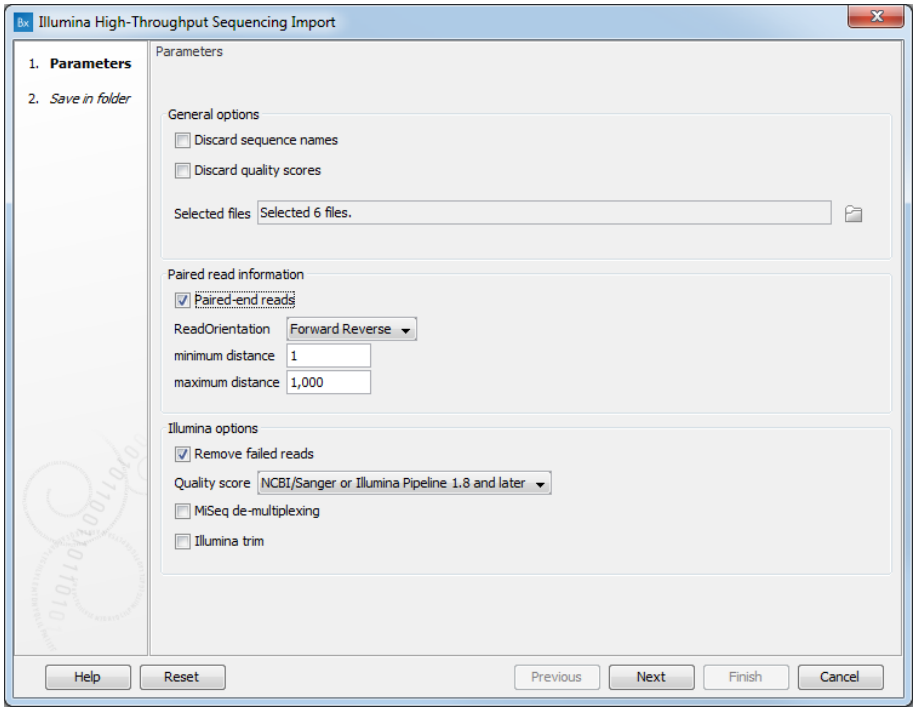


Figure 2.3: *The QIAseq Panel Analysis Guide Illumina Import interface.*

- Discard quality scores: Quality scores are visualized in the mapping view and they are used for variant and fusion detection. Therefore, we do not recommend to discard the quality scores during import.

- **Paired read information**

  - Paired-end reads: check the option if the reads are paired, then choose their orientation (Forward Reverse for end-paired or Reverse Forward for mate-pair) in the drop down menu.

- Set minimum and maximum distances. The paired read distance includes the full read sequence, which means that is from the beginning of the forward read to the beginning of the reverse read. The distances are usually defined during the library preparation of your sequencing experiment, but in doubt you can enter default values: for paired-end reads, distances are between 1 and 1000 bp while mate-pair reads typically have longer distances between 1000-5000 bp (and sometimes up to 10000).

- **Illumina options**

  - Remove failed reads: If you check **Remove failed reads**, reads that did not pass a quality filter (in qseq and fastq files) will be ignored during import. For more information on format specific quality filters see section on file format above). If you import paired data and one read in a pair is removed during import, the remaining mate will be saved in a separate sequence list with single reads.
  - Quality score: choose which Illumina pipeline version was used to sequence your reads.
  - MiSeq de-multiplexing: Using this option on MiSeq multiplexed data will divide reads into different files based on the "IndexSequence" of the read header:

    ```
    @Instrument:RunID:FlowCellID:Lane:Tile:X:Y:UMI ReadNum:FilterFlag:0:IndexSequence
    ```

    Subsequent analysis can then be executed in batch on all the files, and results can be compared at the end.
  - Illumina trim: This option applies to Illumina Pipeline 1.5 to 1.7. In this pipelines, the Phred scale (ASCII 64 to 104) disregards value 0 and 1, but the value 2 (B) has special meaning and is used as a trim clipping. This means that when selecting Illumina Pipeline 1.5 and later, the reads are trimmed when a B is encountered at either end of the reads in the input file if the **Trim reads** option is checked.

The Ion Torrent High-Throughput Sequencing Import (figure 2.4) just requires you to specify a SAM or BAM file containing unmapped reads.

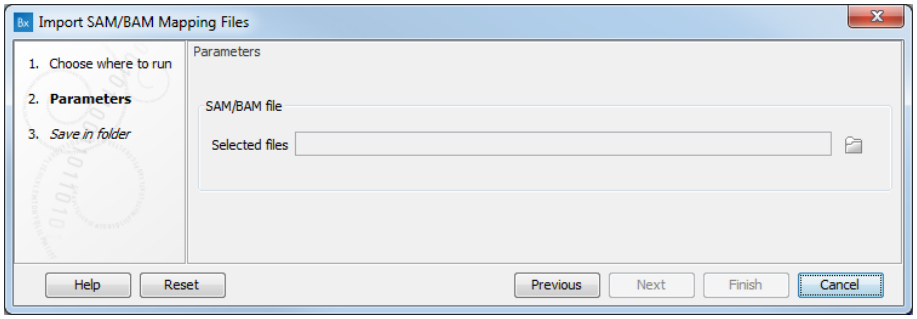


Figure 2.4: *The QIAseq Panel Analysis Guide Ion Torrent Import interface.*

2.2.2 Start a QIAseq panel analysis

The Analyze QIAseq Panel Analysis guide offers a series of pre-installed ready-to-use workflows for which all settings have already been set to optimize the results of your secondary analysis.

The Guide dialog offers three tabs, one for each category of applications: Targeted DNA, Targeted RNAscan, Targeted RNA. On each tab, all available QIAseq analysis workflows are displayed in a

list. For Targeted DNA Panels, click on the little arrow next to the analysis workflow of your choice to access configuration options and settings.

Once you are ready to start the workflow of your choice (including the relevant configuration when several are available using the drop down menu), click **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. Click **Next**.

You can then select the sequencing reads that should be analyzed (figure 2.5). Note that the workflow can be run in batch mode, i.e., you can select several read files that will be processed independently.

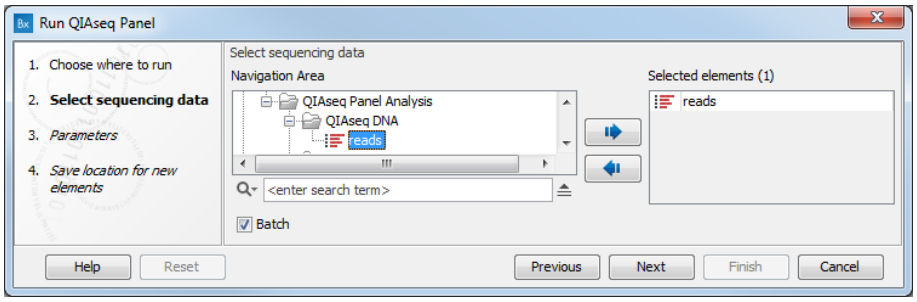


Figure 2.5: *Select the sequencing reads.*

Depending on the application, you may be offered to edit some parameters. For **Targeted DNA panels**, the parameters are as seen in figure 2.6.

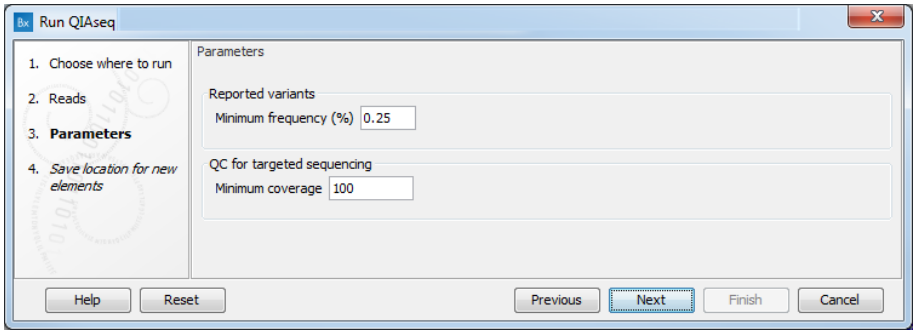


Figure 2.6: *Parameters to configure for Targeted DNA workflows.*

- Reported variants minimum frequency: the variant's frequency needs to be above that threshold for the variant to be output by the workflow in the filtered variant track.

  Note that all variants found will be output in the **Unfiltered variants** track, while only the variants passing the previous two parameters thresholds will be output in the **Variant passing filters** track. The Unfiltered variant track is produced by the Low Frequency Variant Detection tool run with a frequency cut-off value of 0.1 for somatic workflows, 1.0 for germline workflows. Therefore, it does not make sense to filter the variants with a value lower than 0.1 in case of somatic application, and 1.0 in case of germline application.

  In fact, default settings for Targeted DNA panels are set for somatic variant detection at 0.25%, and at at 20% for germline variant detection.

- QC for targeted sequencing minimum coverage: this value will be used in the Coverage Report output by the workflow: it is set at 100 for somatic workflows, at 30 for germline workflows

In the last dialog, choose where you want to save the outputs of the workflow. A description of the output are given in the following application-specific chapters.

2.3 How to create a custom panel analysis workflow

For each type of applications (DNA, RNAscan and RNA) it is possible to set up a custom workflow.

Click on the **Add custom panel** button to configure the new panel (figure 2.7).

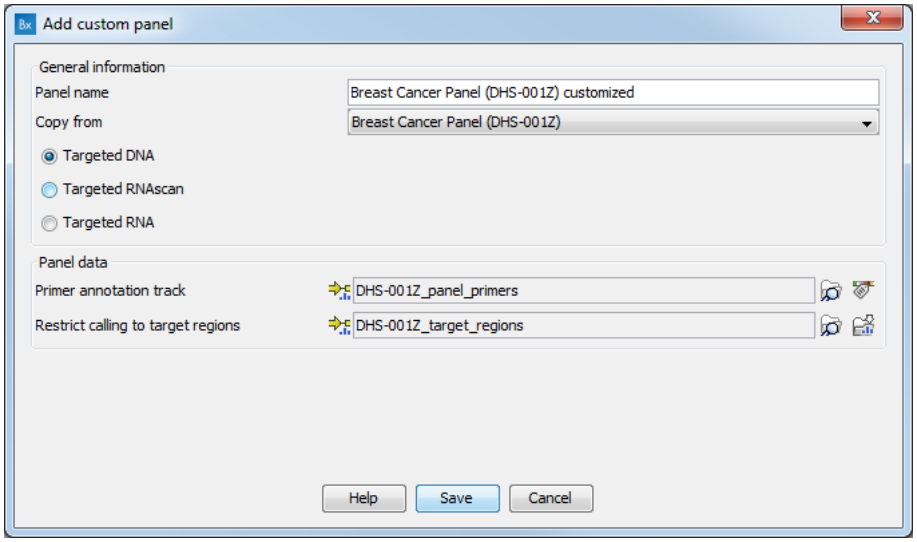


Figure 2.7: *Configure the workflow with the files corresponding to your custom panel. Here we can see an example of custom Targeted DNA workflow.*

In the **General information** section, enter the name of the custom panel. Choose the application for the panel. It is also possible to use a pre-existing panel as a template by selecting it from the drop down menu. In that case, the name of the panel chose in the drop-down menu will fill the "Panel name" field, but can also be re-edited for clarity purposes.

In the **Panel data** section, specify the following files (the type of files you have to provide depends on the application):

- Primer annotation track (Targeted DNA and RNAscan): The QIAseq DNA Panels primers are made available to you upon purchase of a kit. The Import icon allows you to import in the workbench a QIAGEN primers file previously saved on your computer (see section 2.3.1), while the Browse icon allows you to specify a file already stored in your Navigation Area.
- Restrict calling to target regions (Targeted DNA and Targeted RNA): The QIAseq DNA Panels primers target different regions of the genome that are specified in a *.bed file. This file was made available to you upon purchase of a kit. The Import icon allows you to import in the workbench a target regions BED file previously saved on your computer, while The Browse icon allows you to specify a file already stored in your Navigation Area. Note that

when importing, you also have to specify also a reference sequence (hg19 for targeted DNA panels, hg38 for Targeted RNA panels). This reference sequence is pre-set once you have downloaded and applied a Reference Data Set as described in section 2.1.

- Known Fusions track (Targeted RNAscan only). This track will be used to annotate the fusion track output by the workflow. The Import icon allows you to import in the workbench a Known Fusion Information track (see section 4.2.5). The Browse icon allows you to specify a file previously saved in the Navigation Area.
- Human or Mouse (Targeted RNA only): Specify the species of the samples that will be analyzed with your custom workflow.

Once your custom panel is configured, click **Save** to add the custom workflow to the list of workflows available from the Analyze QIAseq Panels guide. Once in the list, a button **Edit** allows you to change the configuration of the custom workflow. Click **Run** to follow the instructions given in each dialog (see section 2.2.2).

2.3.1 Import QIAGEN Primers

When creating a custom analysis workflow, it is possible to specify a Primer annotation track using the import button to the right of the relevant field in the Add custom panel dialog. Importing such a file can also be done ahead of time using the Import QIAGEN Primers tool.

The **Import Primer Pairs** importer can import a QIAseq Panel primer file previously saved on your computer. During import, the primers used for targeted resequencing are saved in the Navigation Area of the workbench as a Primer track. **Note**: the Import QIAGEN Primers tool is different from the Import Primer Pairs tool because of the format of the primer file it can handle (see below for a detailed description of the format).

The **Import QIAGEN Primers** can be found in the toolbar:

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

The import wizard is shown in figure 2.8. The first step is to select the primers to import and a reference sequence.

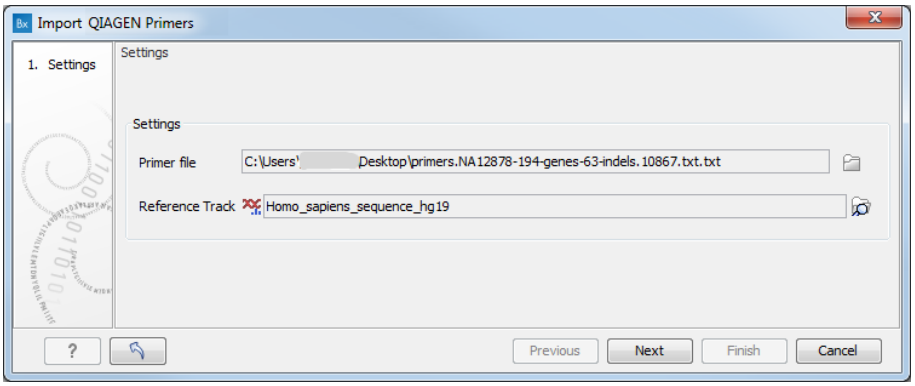


Figure 2.8: *Select files to import.*

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

- **Reference Track** Choose the hg19 or the hg38 reference sequence that can be found in the CLC_References folder after you have downloaded the hg19 (or hg38) Reference Data Set.

Click on the button labeled **Next** to go to the wizard step choose to save the imported primer location file.

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

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

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

For example, the lines
```
chr1 1887011 L AGAATATTTTCTTGCTTAACCGTCACTTAACATCGA
chr1 1900114 R GGGACAAGACCTGGAACTACATTTCTGACT
```
define the primers: On chr1, from 1886977 to 1887012 (both are 0-indexed, inclusive), on the plus strand. On chr1, from 1900114 to 1900144 (both are 0-indexed, inclusive), on the minus strand.

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

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

For example, the lines:
```
14 chr1 1886977 1887012 + AGAATATTTTCTTGCTTAACCGTCACTTAACATCGA COPA
2 chr1 1900114 1900144 - GGGACAAGACCTGGAACTACATTTCTGACT RCSD1
```
define the primers from the previous example, with their target annotation.

2.4 Create a custom Trim adapter list (optional)

All workflows included in the plugin start by removing all remaining adapters present on the reads using the "Remove read-through adapters" from the Trim Reads tool. For workflows for DNA and RNAscan applications, this means that the reads are annotated with the UMI information before the trimming removes them, as shown in figure 2.9.

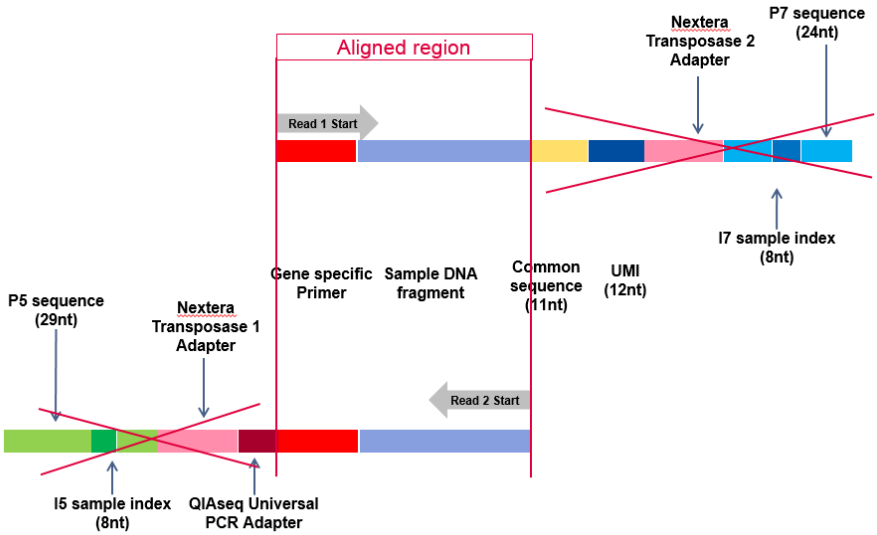


Figure 2.9: *Adapters removed automatically from the reads by the Trim Sequences tool included in the workflow. This example shows reads sequenced with the UMI technology and an Illumina sequencer.*

Despite the automatic trimming, remnants of various adapters and polyA and polyG sequences would hinder the mapping of the reads to a reference sequence if left on the reads. These pre-defined sequences are trimmed using a trim adapter list included in the reference data associated with each workflow application. Figure 2.10 shows all of the sequences potentially included in the different application-specific trim adapter list available in the CLC_References folder in the Navigation Area once the Reference Data Sets have been downloaded.

In some particular cases, such as working with data of low quality, asymmetric read lengths, mate-paired reads and single reads, as well as when working with very large inserts length, it is recommended that you specify a trim adapter list that is more detailed than the one included in the reference data used by the analysis workflow. To find out if you need to run a custom workflow with a detailed trim adapter list, run the workflow once and check the report generated by the Trim Reads tool, and in particular the last section named "5 Automatic adapter read-through trimming" (figure 2.11). If that section shows no detected adapter, you should run the workflow again with a new Trim adapter list that contains the specific adapters of the sequencing technology used to generate the reads.

- Learn how to generate a Trim adapter list here: `http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Creating_new_Trim_adapter_list.html`.
- Remember to add the adapters that were initially in the application-specific template:

Figure 2.10: *Application-specific templates for trim adapter lists.*

  - for Targeted DNA applications: a Poly G sequence and the QIAseq Universal adapter
  - for Targeted RNAscan applications: a Poly A sequence, a Poly G sequence and the QIAseq Universal adapter
  - for Targeted RNA applications:the QIAseq Universal adapter

- Learn how to create a Customized reference data here: http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Create_custom_Reference_Data_Set.html

To be able to use this list with the trimming step, **Apply** the custom Reference Data Set before starting the analysis workflow from the Navigation Area - and NOT from the Guide interface, as these are already bundled with the traditional Reference Data Set. You can read in the following chapters how to run workflows from the Navigation Area.

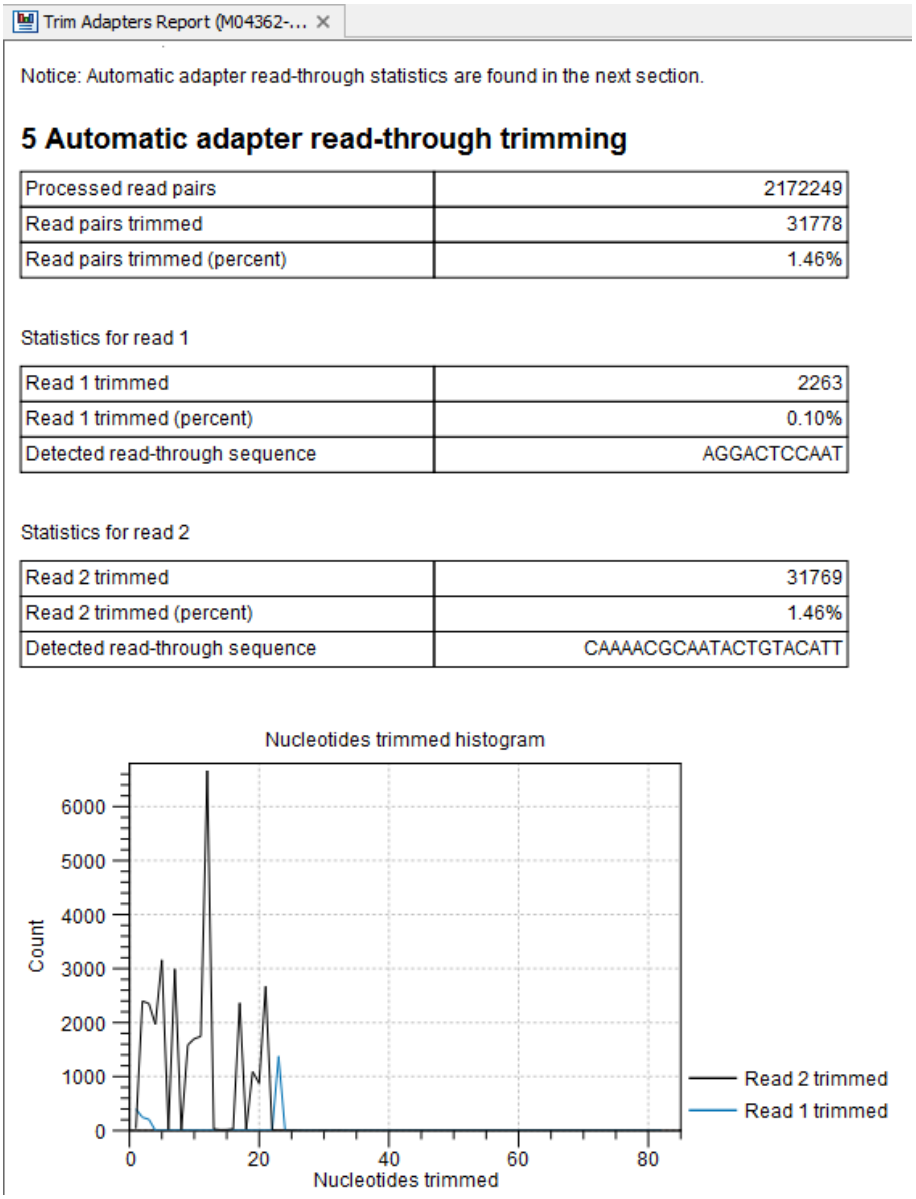


Figure 2.11: *The Trim Adapters report indicates detected read-through sequences. If this section is empty, we recommend using a custom Trim Adapter list.*

Chapter 3

Targeted DNA

Contents

QIAseq Targeted DNA Panels integrate molecular barcode technology into a highly multiplexed PCR-based target enrichment process, enabling accurate variant calling at very low frequency. The concept of molecular barcoding is that during library preparation of the samples with a QIAseq Targeted DNA Panel, a Unique Molecular Index (UMI) and a common sequence prefix are added to each read before amplification. The barcoded molecules are then amplified by PCR. Due to intrinsic noise and sequence-dependent bias, barcoded sequences may be amplified unevenly. Thus, target quantification can be better achieved by counting the number of Unique Molecular Indices (UMIs) in the reads rather than counting the number of total reads for each gene. Sequence reads having different UMIs represent different original molecules, while sequence reads having the same UMI are results of PCR duplication from one original molecule.

However, during secondary analyses of the sequenced reads, UMIs (and their attached common sequence used as identifier) will hinder the mapping of the reads to a reference sequence. The first steps in the QIAseq Targeted DNA Panel Analysis ready-to-use workflow consist in trimming remaining PCR adapters, the UMI and the common sequence while retaining the UMI barcoding information as an annotation on the read. This is followed by mapping the sequencing reads to the human reference sequence. After mapping, the Create UMI Reads tool generates a single consensus read, called a "UMI read", from reads which have the same UMI.

The workflow then removes ligation artifacts from the read mapping. Next, the Indels and Structural Variants detection step generates a guidance track used for improving the mapping

with the Local Realignment tool. Variants are then detected on the UMI reads using the Low Frequency Variant Detection tool for somatic workflows, and the Fixed Ploidy Variant Detection tool for germline applications. Finally, a series of filtering steps remove variants that are either not significant enough, likely due to artifacts, homopolymer errors, or too infrequent. The final output of the workflow includes, among other items, a list of filtered variants, including some present at very low frequency in the original dataset.

3.1 The Identify QIAseq DNA Variants ready-to-use workflows

There are four Identify QIAseq DNA Variants ready-to-use workflows, each optimized to work with either somatic or germline application, from Illumina or Ion Torrent reads.

Somatic/germline specificity: For somatic variant detection, the ready-to-use workflow uses the Low Frequency Variant Detection tool, a variant caller which does not base its statistical model on a bi-allelic assumption, thus will declare a site heterozygous if it detects more than one allele at a site even if one of the alleles is detected at very low frequency and later filtered out. For germline applications, the workflows use the Fixed Ploidy Variant Detection tool. This variant caller has higher precision than the Low Frequency Variant Detection tool, particularly at low to moderate levels of coverage (< 30x). At high levels of coverage (>100x) the Ploidy Variant Detection tool will exhibit low sensitivity for variants with allele frequencies far from what is expected for germline variants (that is 50 or 100%).

Illumina/Ion Torrent specificity: The workflow for Ion Torrent reads includes an extra step that removes non SNV type variants that are likely due to artifacts.

For each case, configuration of the parameters are optimized to obtain high sensitivity and specificity in detecting variants.

All workflows work using the same Reference Data Set: before starting any one of the workflows, open the Data Manager, select **QIAseq DNA Panels hg19**, **Download** the set if you have not done so before and click on **Apply**. You can now close the Data Management window.

The Identify QIAseq DNA Variants ready-to-use workflow can be found here:

> **Ready-to-Use Workflows** | **QIAseq Panel Analysis** | **QIAseq Analysis workflows** | **Identify QIAseq DNA (Somatic/Germline) Variants (Illumina/Ion Torrent)** ()

Double-click on the Identify QIAseq DNA Variants ready-to-use workflow relevant for your samples to run the analysis. If you wish to run a workflow in batch mode, please start the workflow using the QIAseq panel analysis.

If you are connected to a CLC Server via your Workbench, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a CLC Server when possible. Click **Next**.

Select the sequencing reads that should be analyzed (figure 3.1).

In the next dialog (figure 3.2), specify the relevant **target region BED file**. You may have imported this file in the Navigation Area yourself, or you can find it in the CLC_References folder after you downloaded the QIAseq DNA Panels hg19 Reference Data Set with the Data Manager. Note that parameters for the Low Frequency Variant Detection tool are not adjustable, but they have been set to generate an initial pool of all potential variants that will be filtered in subsequent steps.

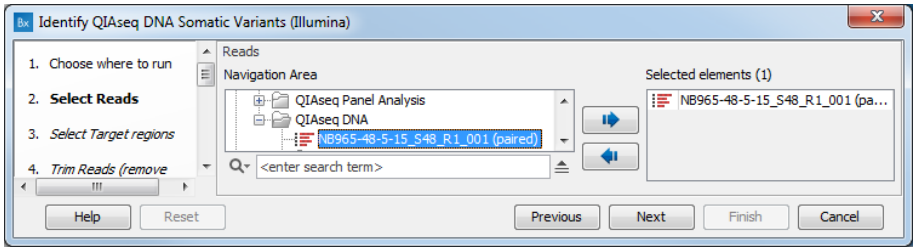


Figure 3.1: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

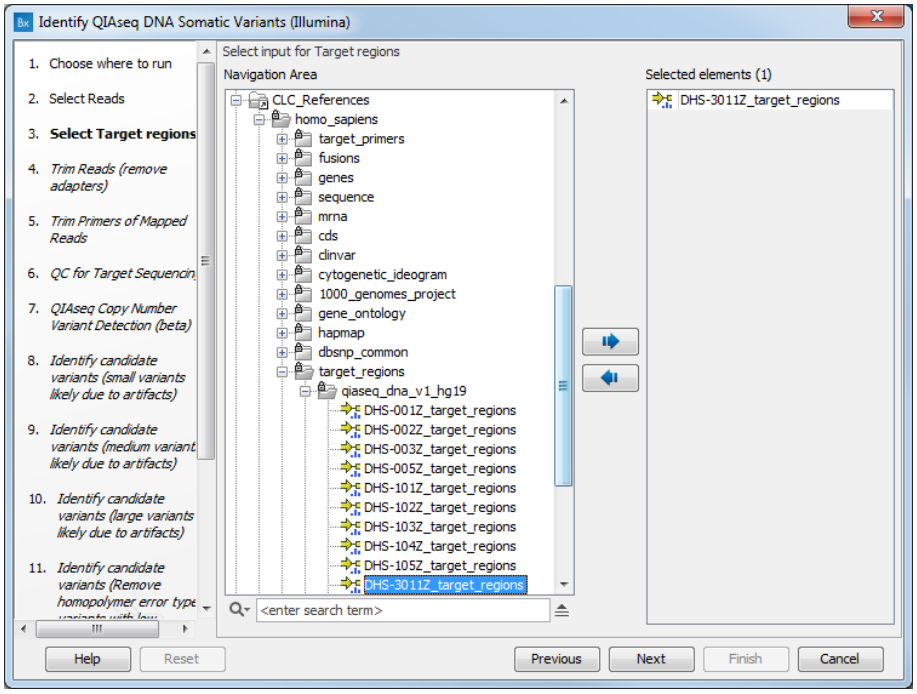


Figure 3.2: *Select the target regions file specific to the panel used.*

Select the **primer** track specific to the QIAseq DNA Panel you used to generate the sequenced reads (figure 3.3). This primer file can be imported ahead of running the workflow using the Import QIAGEN primers tool, or it can be found in the CLC_References folder after you downloaded the QIAseq DNA Panels hg19 Reference Data Set with the Data Manager.

In the dialog called "QC for Target Sequencing", you can modify the Minimum coverage needed on all positions in a target for this target to be considered covered (figure 3.4). Note that the default value for this tool depends on the application chosen (somatic or germline).

The dialog for "QIAseq Copy Number Variant Detection" allows you to specify a control mapping against which the coverage pattern in your sample will be compared, in order to call CNVs. If you do not specify a control mapping the Copy Number Variation analysis will not be carried out. If you have previously run the workflow with control data, you will find the mapping in the Reports and Data folder (Mapped UMI Reads).

Please note that if you want the copy number variation analysis to be carried out it is very important that the control mapping that you supply the tool with is meaningful as a control for the sample that you are analyzing. A meaningful control must satisfy two things: it must (1) have

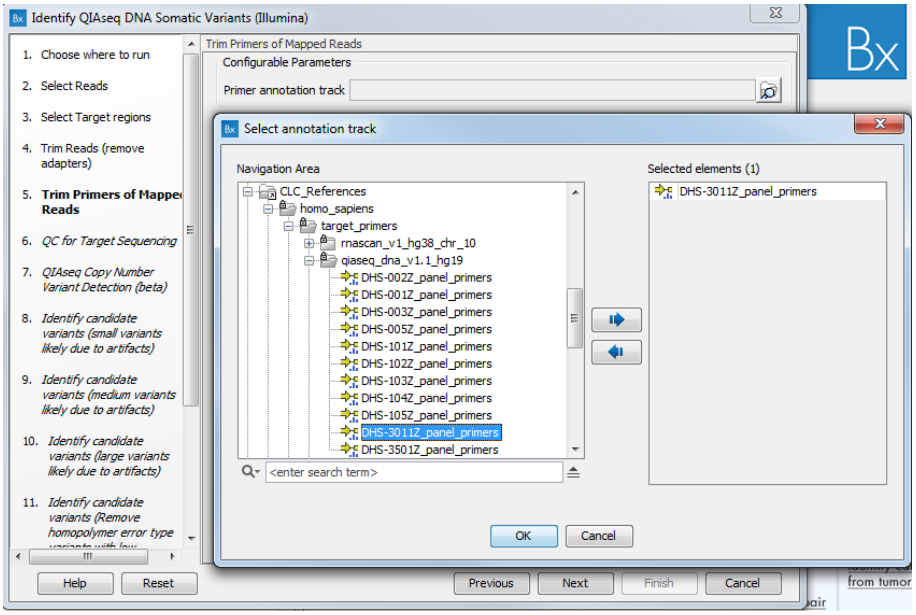


Figure 3.3: *Select the primer file specific to the panel you are analyzing.*

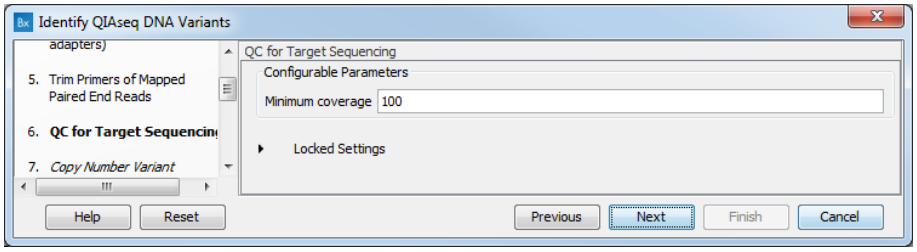


Figure 3.4: *Setting the Minimum coverage parameter of the QC for Target Sequencing.*

a copy number status that it is meaningful for you to compare you sample against and (2) it must be a read mapping that has been processed in the same manner that you sample has been processed. To ensure (1), it is important for panels with targets on the X and Y chromosomes that the control and sample are of the same gender. You can ensure (2) by processing the sample that you want to use as control with the workflow (without control mapping to the CNV detection component) and use the resulting UMI reads track as controls in subsequent runs of the workflow.

A series of dialog will now filter the initial pool of variants found in the data to remove variants that are either:

- not significant enough: filtering is performed based on the QUAL value, value that weighs count against coverage and error rate.
- likely due to artifacts: a called variant must be of sufficient quality, and have an un-biased read direction or read position presence (using the values "Average quality", "Read position test probability" and "Read direction test probability").
- homopolymer errors (indels occuring in a homopolymer stretch) using a too low frequency cut-off value.

- too infrequent.

Some filters only filters alternative alleles - and not reference alleles - as this potentially lead to wrong interpretation of variants by the VCF exporter where such variants could be misinterpreted as hemizygote when the reference allele is missing.

Note that each filter has specific default values depending on the technology (Illumina / Ion torrent) and application (somatic or germline) chosen. These filtering steps have been configured to provide the best sensitivity and precision in the variants output by each workflow. However, benchmarking was performed on samples of relatively high coverage. Therefore, additional filtering might be needed, or filtering values adjusted when working with low coverage samples. This can only be done by running the workflows listed in the Toolbox, and not by using the panel guide. When configuring filters, do not load any annotations, nor try to change the name of the filters in the first column, as it would disable the filter completely.

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

3.1.1 Output from the Identify QIAseq DNA Variants workflow

The Identify QIAseq DNA Variants workflow produces a Genome Browser View () as well as the following files, available in a subfolder as seen in figure 3.5):

- a Trim Reads Report () where you can check that adapters which adapters were detected by the automatic detection option
- a UMI Groups Report () containing a breakdown of UMI groups with different number of reads, along with percentage of groups and reads
- a Create UMI Report () that indicates how many reads were ignored and the reason why they were not included in a UMI read.
- a read mapping of the UMI Reads ()
- a coverage report and a coverage track from the QC for Target Sequencing tool (see `http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=QC_Target_Sequencing.html`)
- two variant tracks: the Unfiltered Variants is output before the filtering steps, the Variants passing filters is the one used in the Genome Browser View (see `http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=_variant_track_output.html` for a definition of the variant table content).
- one amino acid track
- when a read mapping was submitted in the Copy Number Variant Detection dialog, the workflow also ouptputs three CNV tracks (Target-, Region- and Gene-level) and a CNV Results report.

The Unfiltered variants track is included in the output so you can also review why a variant that was expected in the output would have been filtered out of the Variants passing filters track. The

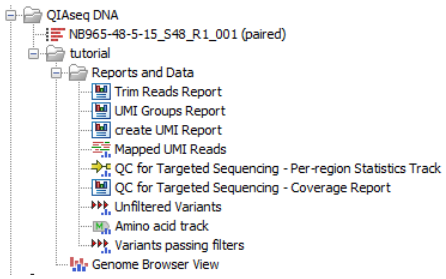


Figure 3.5: *Output from the Identify QIAseq DNA Variants workflow without CNV detection.*

difference between the Unfiltered variants track and the Variants passing filters track depends on the following options available in the filtering steps:

- Filter based on quality criteria: Average Quality (quality of the sequenced bases that carry the variant), QUAL (significance of the variant), Read Position Test Probability (relative location of the variant in the reads that cover the variant position) and Read Direction Test Probability (relative presence of the variant in the reads from different directions that cover the variant position). Additional criteria can be added when working with a copy of Identify QIAseq DNA Variants workflow (see http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Open_copy_ready_to_use_workflow.html).
- Remove homopolymer error type variants, i.e., errors of the indel type that occur in homopolymer regions. These regions are known to be harder to sequence than non-homopolymeric regions. Note that the definition of homopolymere regions differs between the pipelines due to differences in sequencing technology.
- Remove false positive based on frequency: the variant's frequency needs to be above that threshold for the variant to be output by the workflow in the filtered variant track. Note that because the unfiltered variant track is produced by the Low Frequency Variant Caller run with a frequency cut-off value of 0.1, it does not make sense to filter the variants with a value lower than 0.1. In fact, we recommend filtering with a value that is at least twice the frequency cut-off value of 0.1.

The workflow also produces a QC report for the target enrichment that offers statistics on the numbers of targets for which all positions are covered by the "Minimum coverage" threshold set in the QC for targeted sequencing dialog.

The read mapping of the merged UMI groups will let you verify the found variants, and examine why expected variants were not found. The UMI Groups Report gives information about the number of UMI groups found, and how many reads are in each. It includes the following information:

- How many reads were aligned to the reference (Reads in input).
- How many reads were mapped in multiple places and thus discarded.
- Groups merged: How many groups were created by merging singleton groups with other groups.

- Number of groups that were discarded for being too small (by default 0 but the option "Minimum group size" of the Calculate Unique Molecular Index Groups can be set up to discard small groups), and how many reads were thus discarded.
- How many groups were created, and of these how many were singletons groups (groups made with sequences sharing identical UMI).
- How many reads are in the largest group.
- How many different UMIs are in the most divergent group (different sequences with different UMIs can be in the same group, if they start on the same position and if they have UMIs that only differ with one character).
- Statistics about the number of reads in the groups.
- Statistics about groups size and reads not included in these groups (also available as graphs below the table).

Note about interpreting the QC for DNA Panels Report

The relationship between the number of reads per UMI and the original coverage is not straightforward: if the original molecule has been amplified many times, the resulting, seemingly deep, coverage will not add much information. The plots in figure 3.6 illustrate the relationship between sensitivity and UMI coverage, and precision and UMI coverage, with sensitivity and precision defined as:

Sensitivity = $TP/(TP + FN)$

Precision = $TP/(TP + FP)$

and TP: True Positives, FN: False Negatives, FP: False Positives.

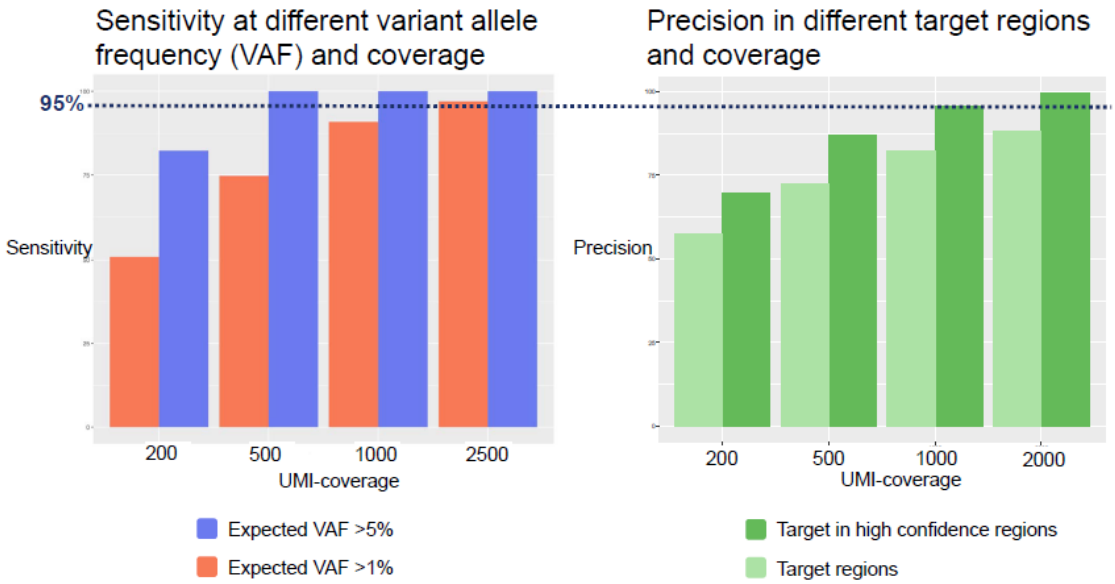


Figure 3.6: *Sensitivity at different variant allele frequency and precision in different target regions at various UMI coverage values.*

Note about exporting output files in SAM/BAM format Exporting UMI reads as BAM files will show UMI described as Unique_Molecular_Index=[number1]_count=[number2], where number1 is a UMI ID (just a unique UMI group number), and number2 is the number of reads that are in that UMI group.

3.2 Targeted DNA tools detailed description

3.2.1 Remove and Annotate with Unique Molecular Index

During library preparation of the samples with a QIAseq Targeted DNA Panel, a UMI and a common sequence prefix are added to each read before amplification. While the UMI is essential in identifying reads that originate from the same fragment, retaining it as such on the sequenced reads would hinder the subsequent read mapping efficiency and accuracy. Therefore, the **Remove and Annotate with Unique Molecular Index** tool removes the UMI and the common sequence prefix from the reads, while annotating each read with the UMI to retain the fragment identity as annotation.

The tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq DNA Panel Expert Tools** () | **Remove and Annotate with Unique Molecular Index** ()

In the first dialog (figure 3.7), select the reads saved as a sequence list.

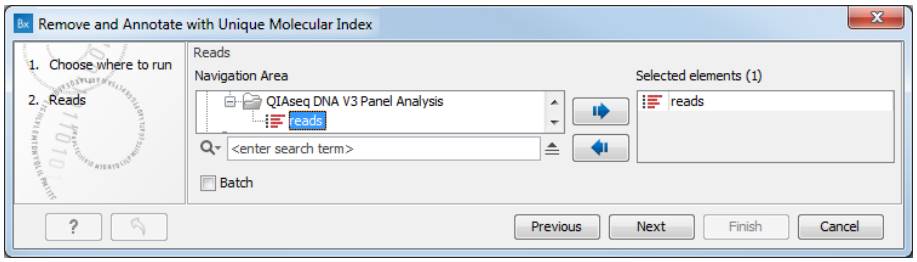


Figure 3.7: *Select the reads generated using the QIAseq DNA Panel.*

In the "Settings" dialog (figure 3.8), the following options are available.

- **Unique Molecular Index read** defines on which read (for paired reads) the UMI will be found. For QIAseq panels, the UMI is situated on Read 2. Changing this parameter to Read 1 may compromise the results of the tools used subsequently, and in particular the results from the Calculate Unique Molecular Index Groups tool that assumes that in paired reads, the UMI will be found on Read2. Keep the value at Read 2 when using single reads.
- **Number of bases to remove** should be the length of the UMI and of the common sequence. It is set by default to 23.
- **Start position of Unique Molecular Index** is set by default to 0.
- **Length of Unique Molecular Index** is set by default to 12.
- **Common sequence verification** can be performed but not on reads generated with an Illumina Miseq sequencer. When this option is enabled, only reads for which the common sequence is recognized will be kept, and the following parameters need to be set:

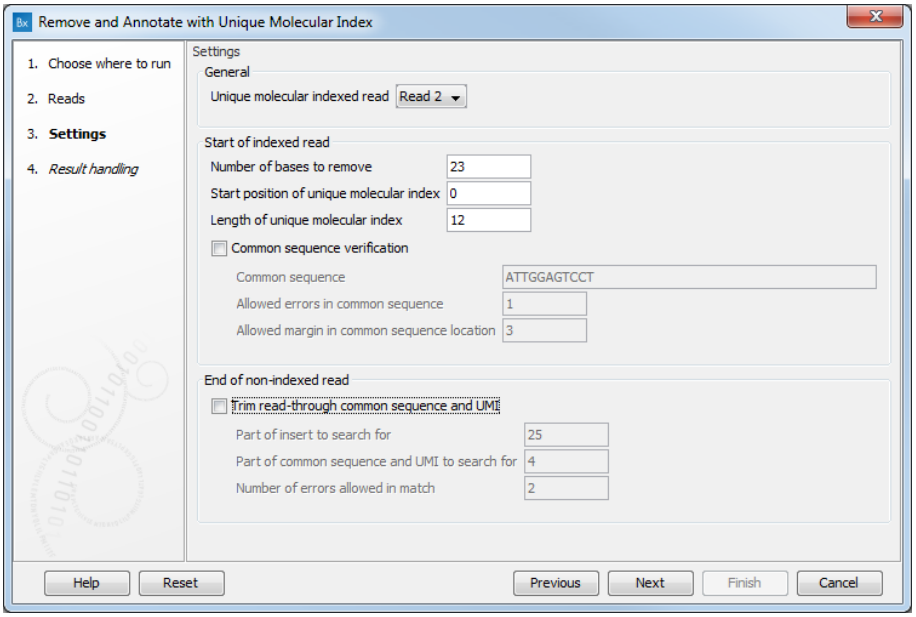


Figure 3.8: *Settings.*

  - Common sequence: Type in the sequence used right next to the UMI to ensure that the UMI will be recognized as such.
  - Allowed errors in common sequence: Number of insertion/deletion/mismatches we allow in the common sequence of reads and still recognize it as such.
  - Allowed margin in common sequence location: Number of nucleotides that separates the actual position of the common sequence from its intended location.

- **Trim read through common sequence and UMI**: If this option is enabled, then for each read pair, first a sequence is extracted from the indexed read consisting of a part of the insert sequence and a part of the adjacent common sequence and UMI. Then, the reverse complement of this sequence is used to search the non-indexed read of a read pair, and if a match is found, the non-indexed read will be trimmed at the boundary between the insert and the common sequence.

  - Part of insert to search for: Number of nucleotides from the sample sequence insert used to identify read-through. Increase this value to get more specific matches, decrease it if the indexed reads are very short, or to improve speed.
  - Part of common sequence and UMI to search for: Number of nucleotides from the common sequence and UMI used to identify read-through. Increase this value to get more specific matches and avoid truncation at repetitive instances, decrease it to trim off shorter partial occurrences of common sequence and UMI.
  - Number of errors allowed in match: Number of insertion, deletion, or mismatch errors allowed when looking for read-through sequences.

Click **Next** to choose whether to **Open** or **Save** your results. By default (i.e., with the Common sequence verification option unchecked), the tool outputs the same number of reads as was present in the input, where reads have had their UMI and common sequence trimmed off. It is also possible to output a report that will inform about the total number of reads processed, the total number of reads found to have UMIs and the fraction of reads that have UMIs.

3.2.2 Calculate Unique Molecular Index Groups

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

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq DNA Panel Expert Tools** ( ) | **Calculate Unique Molecular Index Groups** ( )

In the first dialog (figure 3.9), select a read mapping of reads that were previously annotated with UMI annotations.

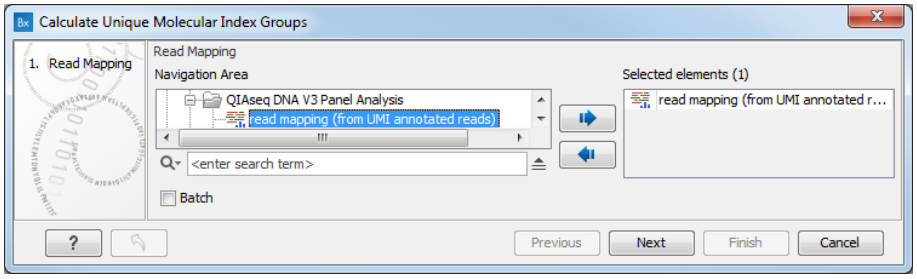


Figure 3.9: *Select a read mapping made from reads whose UMI was removed and annotated on the sequences.*

The grouping of reads into UMI groups works as follows:

1. The tool groups reads that
   - start at the same position based on the end of the read to which the UMI is ligated (i.e., Read2 for paired data),
   - are from the same strand, and
   - have identical UMIs.

   Groups that contain only one read are called singletons.
2. The tool then fuzzy merges singleton groups into non-singleton groups, if the UMI of the singleton group can be made into UMI of non-singleton group by introducing a SNP, and if the non-singleton group is the biggest of such group (i.e., if two different introduced SNPs yields two different non-singleton groups, the biggest one is chosen).
3. Additional merging of singletons and small groups into bigger ones can happen depending on the parameters set for the tool.

It is possible to change the following parameters:

- **Fuzzy match Unique Molecular Indices**: checked by default. Unchecking this option means that the second step of the grouping will not take place.
- **Fuzzy match Unique Molecular Indices (enable indels)**: the tool looks for a group with a UMI one SNP and/or one insertion or deletion away from the current groups UMI.
- **Exclude ambiguously mapped reads**: is checked by default.

- **Maximum relative size difference between merged groups**: will merge small groups into bigger ones if the size difference between the two groups is smaller than a certain value (set at 0.1 per default).
- **Always merge singleton groups**: if this option is not checked, a singleton barcode group is merged into a bigger group only if the singleton UMI is a "neighbor" of the big groups UMI, and the location of the reads are similar.

Click **Next** to choose whether to **Open** or **Save** the resulting read mapping of reads which now have had a "UMI group ID" annotation. It is also possible to choose to generate a report containing a breakdown of UMI groups with different number of reads (along with percentage of groups and reads).

It is also possible to choose to generate a UMI Groups report containing the following information:

- Reads in input: reads that were aligned to the reference
- Reads mapped multiple places (discarded): reads that aligned to the reference in multiple places, and thus discarded
- Groups merged
- Groups not merged due to >1 candidate of equals size
- Groups not merged due to parameter thresholds
- Number of groups that were too small (discarded)
- Number of reads in groups that were too small (discarded)
- Output groups
- Singleton groups
- Reads in largest group
- Number of UMI in most divergent group
- Average, Median and Standard deviation of reads per group
- 5, 20 or 30% biggest group size
- % of reads NOT in 5, 20 or 30% biggest size groups
- Groups with size x (% of groups) (% of reads)

The following graphs are also available:

- Group Sizes graphs, the first including all groups, and the second including only groups with less than 50 reads in them;

- Relative size of biggest group member (figure 3.10). This plot shows in percentage the number of reads that have the most commonly seen UMI sequence in a UMI group. First, UMI groups are created by grouping the reads that start at the same position, have the same strand, and have the exact same UMI. Since sequencing errors can happen in the UMI sequence, we then fuzzy merge singleton groups into non-singleton groups if the UMI of the singleton group can be made into UMI of non-singleton group by introducing one SNP, and if the non-singleton group is the biggest of such group. Furthermore, we merge small groups into bigger ones if the size difference between the two groups (number of reads in the small group divided with number of reads in the large group) is smaller than 0.1. Finally, we always try to merge singletons into a larger group. Hence, not all reads in a UMI group have the exact same sequence.

  The x-axis of the plot shows the percentage of reads have the most common UMI sequence out of all reads in the group, calculated as reads the have most common UMI sequence in the UMI group divided by the number all reads in the group multiplied by 100 to get the value as a percentage. The values in the y-axis is counting UMI groups that have the percentage on the x-axis. Thus, the "Relative size of biggest group member (%)" plot show the percentage of reads in a UMI group that have the most common UMI sequence for the UMI group.

In the example below, we see that for most UMI groups (1725 UMI groups), 97% of the reads have the most common UMI sequence for the group. Furthermore, we see that in the most diverse UMI group only 22% of the reads in the group had the most common UMI sequence for that group.

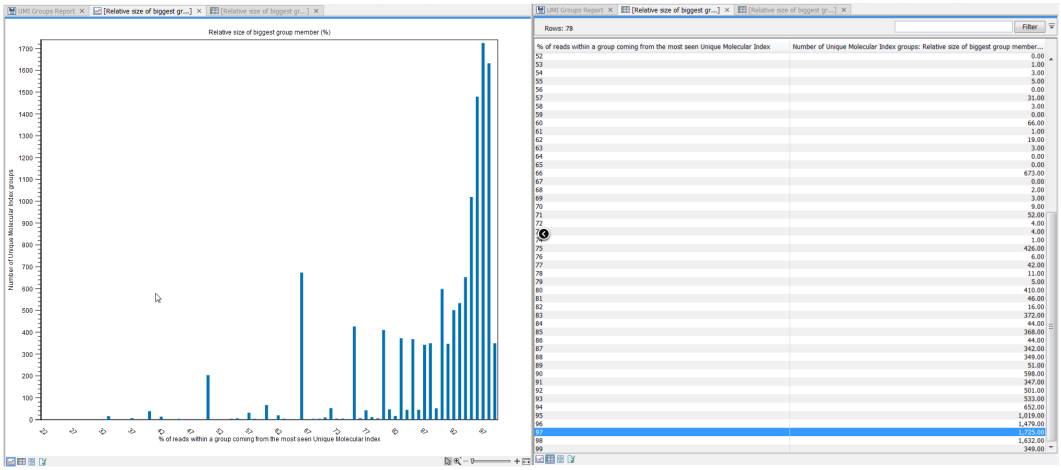


Figure 3.10: *Relative size of biggest group member graph as seen in the UMI Groups Report: click on the plot to open it in a larger view, and view the values from the plot in the table view.*

3.2.3 Create UMI Reads

The tool **Create UMI Reads** generates a single consensus read, called a UMI read, from reads which have the same UMI, and places the UMI read in a read mapping at the location of the original reads. Therefore, the output of the tool is a read mapping of generated UMI reads.

The tool can be found in the Toolbox here:

Tools | QIAseq Panel Analysis Tools | QIAseq DNA Panel Expert Tools () **| Create UMI Reads** ()

In the first dialog (figure 3.11), select a read mapping of the original reads *with UMI annotations* that was previously handled with the Calculate Unique Molecular Index Groups tool.

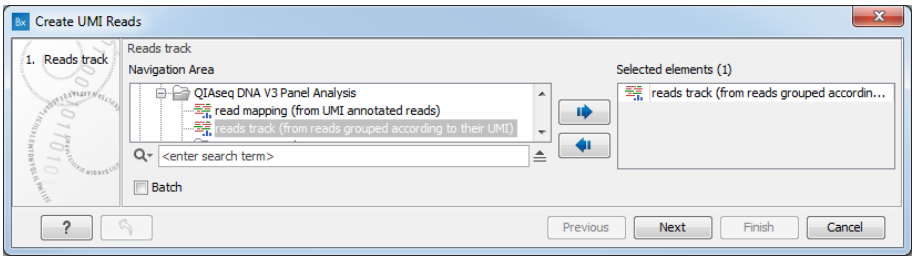


Figure 3.11: *Select a read mapping of the original reads with UMI annotations.*

The second dialog of the wizard (figure 3.12) offers the following options.

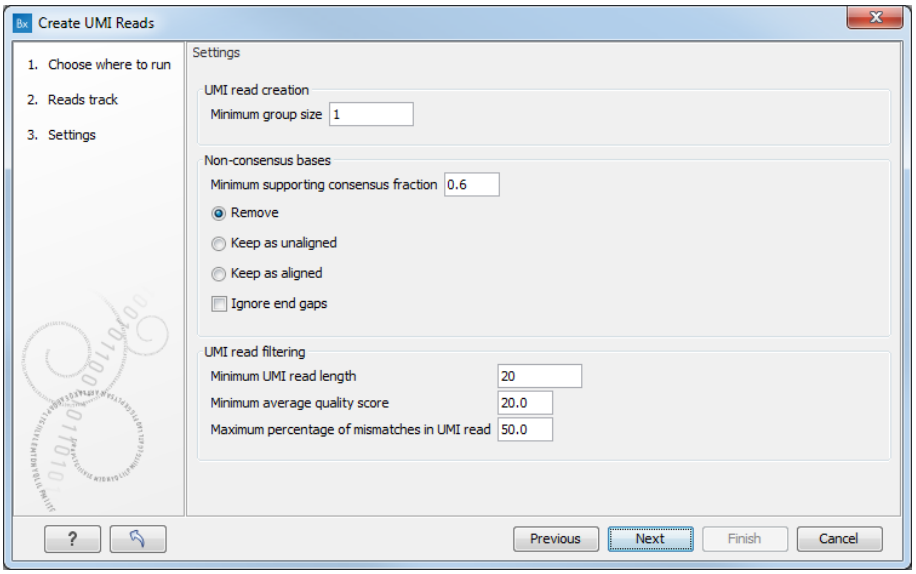


Figure 3.12: *Settings for the Create UMI Reads tool.*

- **UMI read creation**
  - **Minimum group size**: The tool will only create a UMI read if the number of reads in the UMI is at least "Minimum group size".
- **Non-consensus bases**
  - **Minimum supporting consensus fraction** set at 0.6 by default. At each position in the UMI read, the consensus nucleotide is chosen to be the nucleotide with the highest probability of being correct (see the Consensus nucleotide calculation section below). If this probability is higher than "Minimum supporting consensus fraction", a Q score for the consensus nucleotide is calculated. The positions in UMI reads that do not have a consensus nucleotide will be an unaligned end (if it is near the ends of the read), and a base with Q score 0 (if it is in the middle of the read).

  - There is a choice between 3 methods of handling non-consensus bases: **Remove** removes the bases, **Keep as unaligned** (set by default) keeps the bases as unaligned ends, and **Keep as aligned** keeps the bases as aligned bases (but with a Q-score of 0).
  - The last option enables you to **Ignore end gaps** for the calculation of quality scores: gaps are introduced at the end of raw reads to have them of equal size when building an UMI. This option is disabled by default, meaning that the quality scores at the end of the UMI will be rather low due to the presence of the gaps. Enabling this option will result in quality scores of the consensus bases that are at the end of an UMI read close to the quality scores of the raw reads.

- **UMI read filtering**
  - **Minimum UMI read length**: UMI shorter than this value will be discarded.
  - **Minimum average quality score**: UMI reads will be discarded, if their average Q-score is lower than "Minimum average quality score".
  - **Maximum percentage of mismatches in UMI read**: UMI reads will be discarded, if more than 50% of the bases are mismatches.

Click **Next** to **Open** or **Save** the resulting read mapping of UMI reads, i.e., a read mapping of the merged UMI groups. It is also possible to generate a report that will indicate how many reads were ignored and the reason why they were not included in a UMI read. This data will let you verify the found variants, and examine why expected variants were not found.

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

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

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

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

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

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

And by assuming each read base observation is independent,

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

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

The Q-score is now simply the probability of making a wrong call, i.e.

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

which means that the Q-score is

$$-10\, log_{10}(1 - P(C|O_1O_2 \ldots O_k)).$$

Q-scores are capped at 60.

3.2.4 Remove Ligation Artifacts

During the adapter ligation step of the library preparation, it can happen that two different DNA sequences also get ligated together. These ligation artifacts are more prone to occur between short DNA fragments, such as the ones generated from FFPE samples. The tool **Remove Ligation Artifacts** removes reads which are likely the result of ligation artifacts. In addition, in cases of short fragments, a remnant of the common sequence can be found at the end of R1 reads. The tool will also remove these common sequence artifacts.

The tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq DNA Panel Expert Tools** ( ) | **Remove Ligation Artifacts** ( )

In the first dialog (figure 3.13), select a read mapping (it can also be an rna-seq read mapping).

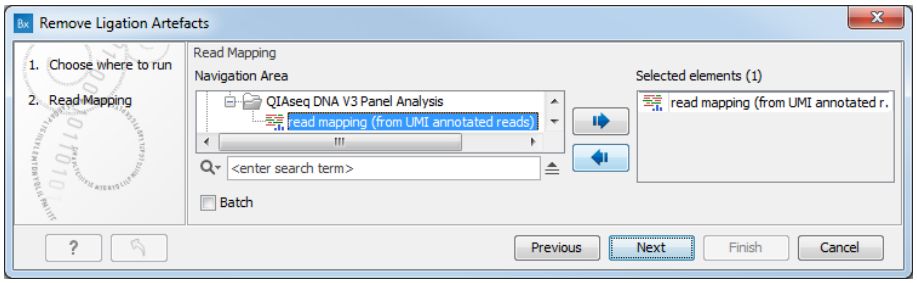


Figure 3.13: *Select a read mapping.*

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

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

- If the option "Remove entire Unique Molecular Index" is checked, all reads in a UMI group are removed if at least one ligation artifact read is found in the group.

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

- The tool looks at the 23 first and last bases (the window) in the read (defined by the "Full length common sequence search limit" parameter) and searches for the common sequence and the reverse complemented common sequence in the window. It is possible to allow a single mismatch between the common sequence and the read window and still call it a match (check "Allow mismatch" in the Common sequence artifacts section of the dialog).
- If no match is found, the tool searches for sub-strings down to a minimum of 4 bases ("Minimal partial common sequence length" parameter) on the read: When searching the first bases of a read, the tool checks if suffixes of the common sequence match the start of the read. When searching for the last bases of a read, the tool checks if prefixes of the common sequence match the end of the read. It is here again possible to allow a single SNP in the common sequence sub-string and the read and still call them a match.
- The tool then counts mismatches in the window (the window is from the match (including the common sequence) out to the end of the read). If the percentage of mismatches in the window is less than 50% (defined by the "Minimum mismatch percentage" parameter), the read is accepted.
- If there are more than 50% mismatches in the window, the read is trimmed from the bases in the read in the window. Both unaligned and aligned bases will be removed. If there are no aligned bases left after trimming, the read is removed.

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

The setting options for the Remove Ligation Artifacts tools are as follow (figure 3.14).

- Ligation artifacts
  - **Remove ligation artifacts**: uncheck this option to keep ligation artifacts in your data.
  - **Minimum mismatches**: define the thresholds of mismatching characterizing a potential ligation artifact.
  - **Ligation artifact recognition length**: defines the size of the window being searched for mismatches.
  - **Allow mismatch**: checking this option will allow a single mismatch between the sequence window and the main sequence while still calling it a match.
  - **Remove entire Unique Molecular Index**: remove all reads in a UMI group if at least one ligation artifact read is found in the group.
- Remove common sequence artifacts from
  - **Single reads**
  - **Paired reads**

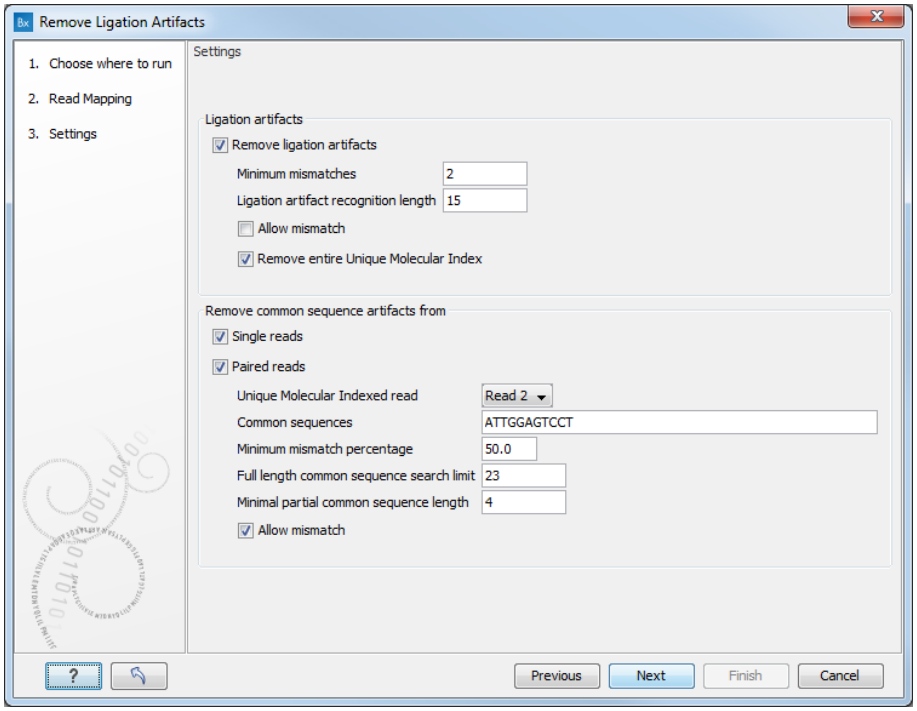


Figure 3.14: *Set the parameters for the remove Ligation Artifacts tool.*

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

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

3.2.5 Prepare Guidance Variant Track

The Prepare Guidance Variant Track tool generates a guidance variant track from the two outputs of the InDels and Structural Variants tool: a Structural variants feature track and an InDels variant track This guidance track can be used for better realignment.

The tool can be found in the Toolbox here:

Tools | QIAseq Panel Analysis Tools | QIAseq DNA Panel Expert Tools ( ) **| Prepare Guidance Variant Track** ( )

In the first dialog (figure 3.15), select an Indels variant track (it usually has the name of the read mapping it originates from, with (Indel) at the end of the full name).

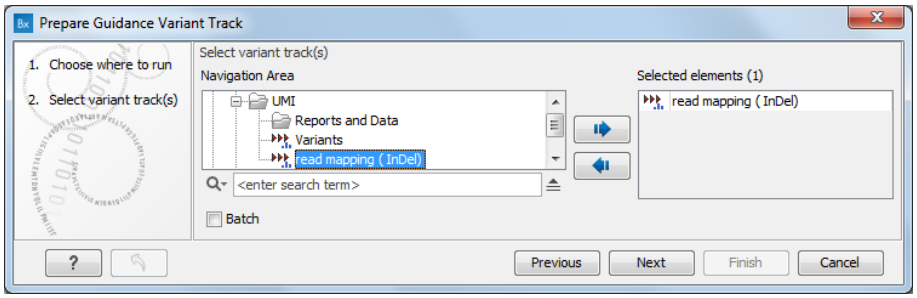


Figure 3.15: *Select a variant track.*

In the second dialog, select the Structural variant track (SV) that was output with the Indels variant (InDel) track from the same read mapping by the InDels and Structural Variants tool (as seen on figure 3.16). You can optionally specify a reference sequence. If set, it will be used to left-align indels to the extent possible. This will make a subsequent local alignment more in line with the recommendation/convention to left-align and that might in turn affect the downstream variant detection to also be left-aligned.

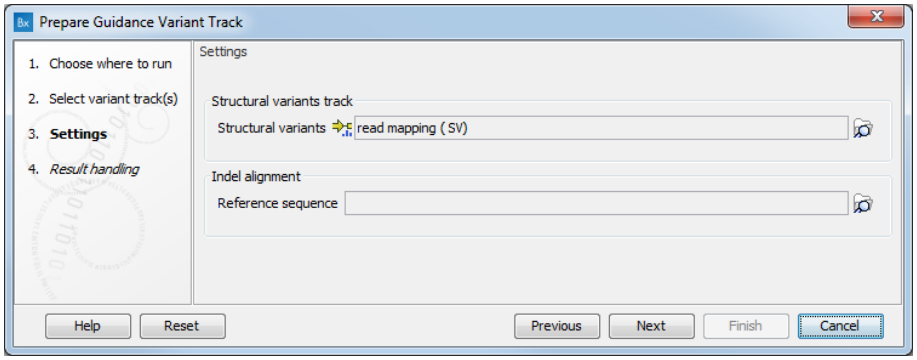


Figure 3.16: *Select a Structural variant track.*

The output is a guidance track (Indel, guidance track), that combines valuable information from the indels and structural variants tracks (such as all the replacements that the tool detects), and that can be used as input for the Local Realignment tool. You can read more about the Indel and Structural variant tracks here: http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=_Structural_Variants_InDels_output.html

3.2.6 Trim Primers of Mapped Reads

This tool replaces the previous tools called Trim Primers of Mapped Single Reads and Trim Primers of Mapped Paired End Reads. Both have been moved to the Legacy folder of the Toolbox (see section 7).

The tool **Trim Primers of Mapped Reads** removes the primer parts of mapped reads, as they

reflect the primer that was added and not the actual sample. Note that the tool also trim primers from RNA-seq mapped reads, except for primers that span intron boundary.

The tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq DNA Panel Expert Tools** () | **Trim Primers of Mapped Reads** ()

In the first dialog (figure 3.17), select a read mapping.

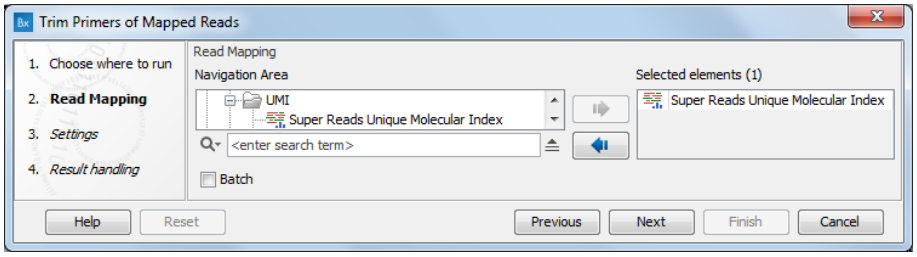


Figure 3.17: *Select a read mapping.*

In the second dialog (figure 3.18), select the primer annotation track that was provided with the QIAseq DNA Panel.

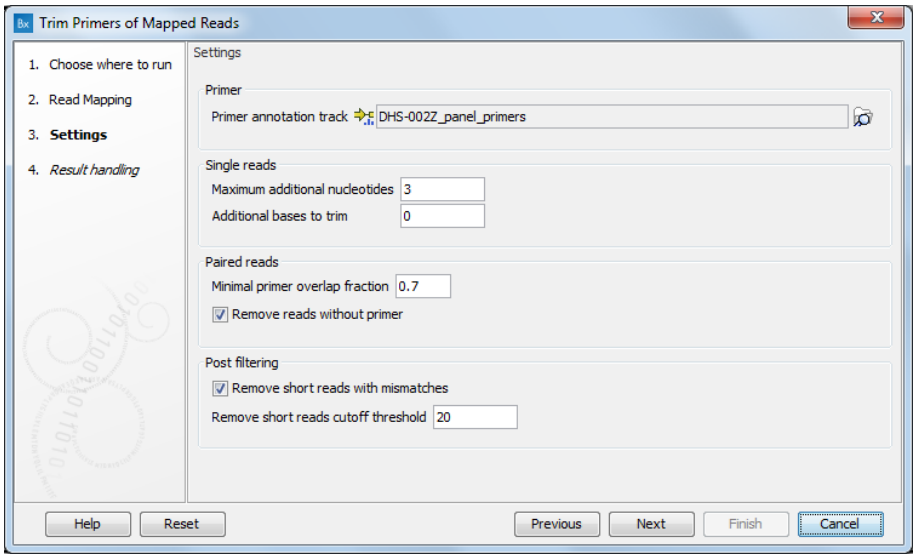


Figure 3.18: *Select the primer annotation track specific to the panel, and add the parameters needed to deal with the type of reads you are working with.*

In addition, set the following parameters:

- Single reads. This tool works on reads that potentially ends in a primer (rather than starts in a primers). The tool aims to unalign the primer parts of reads that came from that primer. It approximates this by only unaligning reads that end inside of a primer or up to 3 extra bases after the primer, as set by the parameter **Maximal additional nucleotides**. You can also decide on how many **Additional bases to trim** right after the primer. This trimming is not done on reads for which dimer artifacts were identified.
- Paired reads. If an aligned read starts within the span of a primer, and if it overlaps the

primer with at least 70% (set by default for the **Minimal primer overlap fraction** option), then it is said to "hit" the primer. For reads "hitting" a primer, the part of the read that overlaps the primer will be unaligned. For reads not "hitting" a primer, the read will either be removed from or retained in the read mapping, depending on the option **Remove reads without primer**. If the option is checked, the tool will remove reads that do not "hit" a primer.

- Post filtering. After unaligning the primer part of the reads, the tool removes reads where less than 20 aligned bases remain, as set by the parameter **Remove short reads cutoff threshold**. This feature can be disabled by unchecking the parameter **Remove short reads with mismatches**.

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

3.2.7 Annotate Variants with Unique Molecular Index Info

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

The tool can be found in the Toolbox here:

Tools | **QIAseq Panel Analysis Tools** | **QIAseq DNA Panel Expert Tools** () | **Annotate Variants with Unique Molecular Index Info** ()

In the first dialog (figure 3.19), select a variant track.

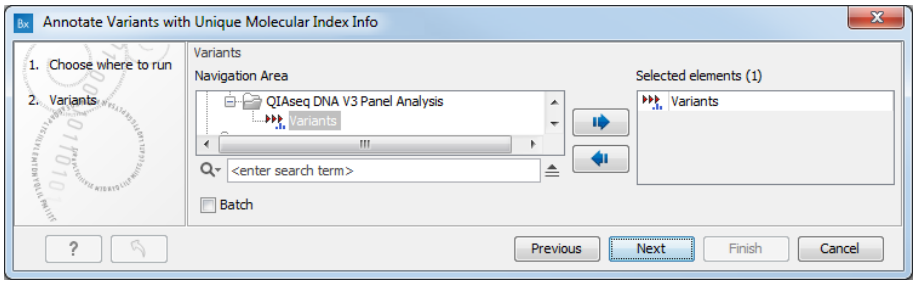


Figure 3.19: *Select a variant track.*

In the second dialog, select a read mapping. The tool works on any read mapping on which UMI groups have been calculated, i.e. a read mapping consisting of raw reads or a read mapping consisting of UMI consensus reads generated by the Create UMI Reads tool (as seen on figure 3.20). If the read mapping consists of UMI reads, check the "Mapping consists of UMI reads" option.

The parameters below are used to calculate the annotations:

- **Minimum size of a Big Unique Molecular Index**: Minimum number of reads in a UMI group for it to be considered Big.
- **Minimum consensus % of a Consistent Unique Molecular Index**: Minimum percentage of

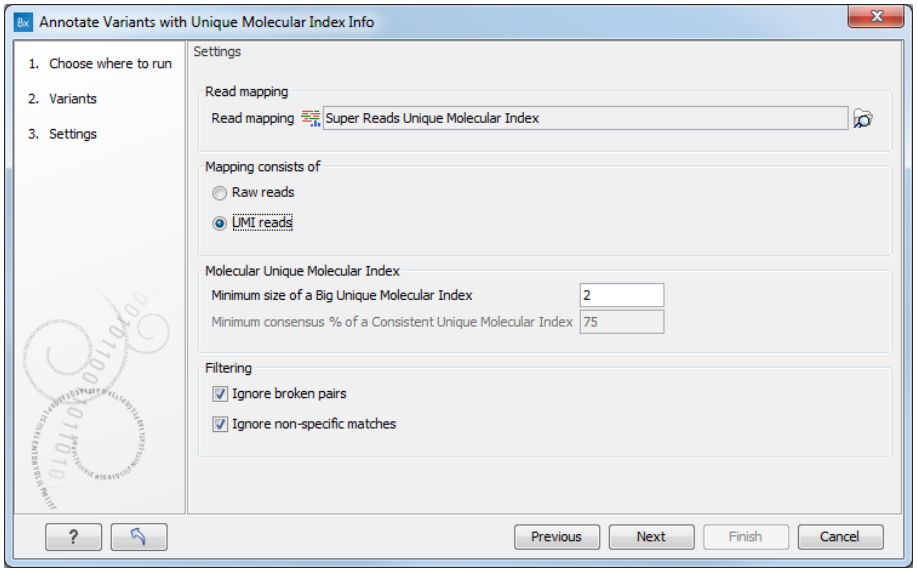


Figure 3.20: *Select a read mapping.*

reads in a UMI group that should support a variant for the UMI to be considered Consistent for that variant. This option is valid only if the read mapping chosen is made of raw reads.

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

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

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

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

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

The annotations differ when the tool is used with UMI reads as it is not possible to calculate annotations involving how many reads in a read group have the variant from the UMI consensus reads. This is for example "Found matching groups with good part matching", "Found matching groups by reads and group sizes", etc.. These columns then assume that all reads in a read group have the variant, that means that all matching UMI groups are Consistent. It is also not possible to calculate how many reads match a variant and how many do not. Many columns show a frequency of reads and also a frequency of groups, e.g. "Found matches" and "Found matches UMI-groups". When running on UMI reads, these two numbers will be the same.

3.2.8 QIAseq CNV Detection

This tool is similar to the Copy Number Variant Detection tool already present in the workbench Toolbox, with the added advantage of making that detection step optional. It means that when using the QIAseq CNV Detection tool in a workflow, users can choose to provide a read mapping or not. To read more about CNV detection, please see http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Copy_Number_Variant_Detection.html

Chapter 4

Targeted RNAscan

Contents

QIAseq Targeted RNAscan Panels use molecular barcode technology to quantify a large number of fusion genes and identify new fusion gene partners.

The concept of molecular barcoding is that during library preparation of the samples with a QIAseq Targeted RNAscan Panel, a Unique Molecular Index (UMI) is added to each read before amplification. The barcoded molecules are then amplified by PCR. Due to intrinsic noise and sequence-dependent bias, barcoded sequences may be amplified unevenly. Thus, target quantification can be better achieved by counting the number of Unique Molecular Indices (UMIs) in the reads rather than counting the number of total reads for each gene. Sequence reads having different UMIs represent different original molecules, while sequence reads having the same UMI are results of PCR duplication from one original molecule.

However, during secondary analyses of the sequenced reads, UMIs (and their attached common sequence used as identifier) can hinder the mapping of the reads to a reference sequence. The first steps in the Detect QIAseq RNAscan Fusions ready-to-use workflow consists in trimming the UMI while retaining the UMI barcoding information as an annotation on the read. Remaining PCR adapters are also trimmed before mapping the sequencing reads to the human transcriptome to perform a RNA-seq analysis of the reads.

In the latest stage of the workflow, the Detect QIAseq RNAscan Fusions workflow works in two major steps: first it detects all potential fusion genes, and then it evaluates (refines) the identified fusion events to increase the sensitivity and specificity of the calls.

Detection. The workflow first trims all remaining adapters from the reads. The trimmed reads are then mapped to the reference transcriptome sequence, and reads are grouped according to

their Unique Molecular Index. The Detect Fusion Genes tool will identify fusion events based primarily on the number of fusion crossing reads, and subsequently on the number of fusion spanning reads. However, when determining whether a read actually crosses the fusion, the tool takes into account the length of the unaligned end, as well as exon boundaries (as at the RNA level, fusions usually happen at exon boundaries). Finally, other evidence, such as whether the unaligned end maps many places in the genome, are considered. Note that the parameters of the Define Fusion Genes tool when included in the workflow are configured differently than the default values of the tool used on its own. In particular, filters have been relaxed to not overlook any fusion.

The Detect Fusion Genes tool uses a binomial model to evaluate the fusions. The null hypothesis is that there is no fusion, i.e., the reads originate from the wild type transcript. Hence, a small p-value suggests a fusion transcript. Reads are assigned to either come from fusion or wild type transcripts based on how well they map to either. This assignment is based on mapping, and it will have an error rate (e) that we estimate from test data. In addition, we require a minimum number of reads to support any fusion breakpoint before considering it as a fusion. This guards against false positives due to low coverage. In addition, we require a minimum number of reads to support any fusion breakpoint before considering it as a fusion. This guards against false positives due to low coverage.

The Z-score and p-value are then calculated using a standard one-tailed binomial test and an "Assumed error rate". This Assumed error rate is a mapping error rate, i.e., the probability of an unaligned end mapping to another gene by random. The p-value represents the probability of spanning/crossing reads (indicating a fusion), under the null hypothesis where a fraction (i.e., the "Assumed error rate") of reads map there by chance.

The Detect Fusion Genes tool outputs a maximum of 200 identified fusions (the ones with the lowest p-value, i.e., the highest Z-score) in a fusion track. In addition to that track, the tool will also generate a set of "fusion references", i.e., a version of the input sequence track, gene track, mRNA track, CDS track and primer track, as well as the fusion breakpoints track, which are mapped on an artifical genome that includes both the wildtype and the fusion chromosomes. Note that the number of fusions output by the tool when used in the workflow has been limited to 200 to avoid having to compute too many subsequent fusion reference chromosomes.

Refinement. The workflow then re-maps the original trimmed reads against the fusion references (i.e., the transcriptome including putative fusion transcripts), and regroups all reads in new UMI groups. We expect that some previously unmapped or poorly mapped reads will now map directly to the fusion transcripts, resulting in a more accurate detection of fusion supporting reads.

The Refine Fusion Genes tool takes the fusions previously identified by the Detect Fusion Genes tool, and re-counts the number of fusion crossing reads as well as wildtype supporting reads. It then calculates the "refined" Z-score and p-value using the same binomial model as the Detect Fusion Genes tool.

4.1 The Detect QIAseq RNAscan Fusions ready-to-use workflow

Before starting the workflow, make sure the correct Reference Data Set is applied: open the Data Manager, select **QIAseq RNAscan Panels hg38**, **Download** the set if you have not done so before and click on **Apply**. You can now close the Data Management window.

The Detect QIAseq RNAscan Fusions ready-to-use workflow can be found in the Toolbox here:

Toolbox | Ready-to-Use Workflows | QIAseq Panel Analysis | Detect QIAseq RNAscan Fusions

Double-click on the Detect QIAseq RNAscan Fusions ready-to-use workflow to run the analysis.

If you are connected to a CLC Server via your Workbench, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a CLC Server when possible.

Select the sequencing reads that should be analyzed (figure 4.1).

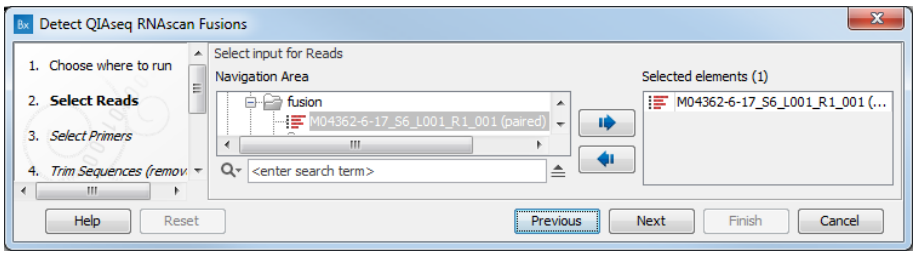


Figure 4.1: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

In the Select primers dialog, chooses the file that corresponds to the panel that was used to generate the reads (figure 4.2).

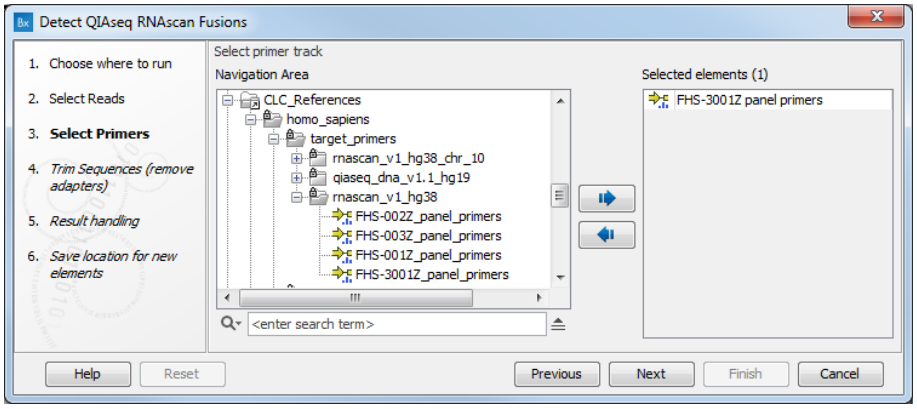


Figure 4.2: *Select the primer track for the relevant panel. You can find it in the CLC_References folder after you have downloaded the QIAseq RNAscan Panels hg38 Reference Data Set.*

In the Detect Fusion Genes dialog, it is possible to change the Promiscuity threshold, i.e., the minimum number of different fusion partners needed to deem a gene promisucous. You can also check the *Detect exon skipping* option to consider same-gene fusions, i.e., exon-skipping fusions. Note however that same-gene fusions where the 5' breakpoint is downstream of the 3' breakpoint will not be considered. Also excluded are exon-skipping fusions that were already annotated on other transcripts from the mrNA track selected.

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

4.1.1 Output from the Detect QIAseq RNAscan Fusions workflow

The Detect QIAseq RNAscan Fusions workflow's main outputs are two Genome Browser Views.

Genome Browser View (WT) displays the following tracks:

- Reference mRNA tracks
- Reference gene tracks
- Reference sequence tracks
- The mapping of the UMI reads
- The detected fusion genes track (based on the wild type genome sequence as reference))

Genome Browser View (Fusions) displays the following tracks:

- The detected fusion genes track (based on a fusion genome sequence as reference)
- The mapping of the UMI reads
- Fusion sequence tracks
- Fusion mRNA tracks (with a representation showing no intron between fusion exons as inserting an artificial intronic sequence could be misleading).
- Fusion gene tracks
- Fusion CDS tracks: each fusion CDS includes a "Frame aligns" annotation, which shows whether the CDS stays in the reading frame across the fusion breakpoint, allowing users to visualize fusions that are more likely to be translated to protein.
- The Fusion panel primers track

In addition, the workflow outputs a series of folders:

"Tracks", "Read Mappings" and "Fusions References" contains the tracks included in the Genome Browser both for Wild Type and Fusions.

"QC and Reports" contain the following reports:

- A QC for Sequencing Reads Graphical Report (see http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=QC_Sequencing_Report_Content.html)
- A Remove and Annotate with UMI Report
- A Trim Adaptors Report (see http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Trim_output.html)
- A QC for RNAscan Panels Report, containing the following information:
  - Total number of mapped reads, counting paired reads as two.
  - Total number of mapped reads(-pair)s, counting paired reads as one.
  - Target genes: genes targeted by the primers.

- Number of target genes.
- Primers in target genes: how many primers are within the target genes regions.
- Mean read coverage per target gene primer (exact start position match): coverage from reads that start exactly at the primer start site.
- Fraction of mapped read(-pair)s exactly matching primers [%].
- Primers outside gene regions: gDNA control primers outside target gene regions, used for detection of DNA contamination of samples.
- Mean read coverage per non-gene primer: a number higher than 0 indicates DNA contamination of the sample. A mean gDNA coverage around 50 reads or higher may increase false positive signal level.
- DNA contamination [%]: calculated using the Mean read coverage per target gene primer and the Mean read coverage per non-gene primer metrics. In general, if the Mean read coverage per non-gene primer metric is around 50 or higher, the chances for a false positive may be higher.
- Primers in reference genes [COPA, MRPS14, CIAO1, UBE3C]: QIAseq RNAscan panels typically include four reference gene primers.
- Mean read coverage per reference gene control primer: the reference genes should have a mean coverage of at least 300 reads, otherwise the effective input is too low, and false negatives are expected.
- Target gene versus reference gene coverage ratio [%]

The easiest way to review the results is to open the Genome Browser Views.

The read mapping track displays the mapping of UMI reads to the transcript sequences. It contains only reads that provide conclusive evidence for or against fusion (figure 4.3).

Figure 4.3: *An example of fusion in a Genome Browser View.*

Double-click on the fusion track name (to the right of the Genome Browser View) while pressing the Ctrl button. The fusion track will open as a table in split view, below the Genome Browser View. Clicking on a fusion event in the table will zoom in to its location in the UMI read mapping, allowing you to review the UMI reads supporting the detected fusion. Note that each fusion event is represented by two lines in the table sharing the same Fusion number. Each line corresponds to a fusion breakpoint.

The fusion table contains the following information:

- Chromosome. Chromosome where "Gene" and "Transcript" are located.
- Region. Breakpoint position of the fusion event relative to the reference sequence hg38.
- Name. Short name of the fusion event, 5' gene–3' gene.
- Fusion number. This number allows to identify easily the two rows that represents a single fusion event.
- 5' or 3' Gene. The fusion gene that corresponds to the "Chromosome" and "Region" fields.
- Breakpoint type. 3' or 5'.
- Fusion Crossing Reads. Number of reads crossing the fusion breakpoint.
- 5' or 3' Read Coverage. Number of reads (unaligned ends and pairs) that cover the 5' or 3'-transcript breakpoint, including normal transcripts and fusion transcripts.
- Z-score. Converted from the P-value using the inverse distribution function for a standard Gaussian distribution.
- P-value. A measure of certainty of the call calculated using a binomial test, it is calculated as the probability that an observation indicating a fusion event occurs by chance when there is no fusion. The closer the value is to 0, the more certain the call. Although one should avoid strictly interpret the p-value as the true false positive rate, our test data show that the p-value seems to be appropriately calibrated using standard parameter settings.
- Filter. Contains the names of the filters applicable to the fusion, or the value "PASS" if it passed all filters.
- Exon skipping. Whether the fusion is a same-gene fusions where the 5' breakpoint is upstream of the 3' breakpoint.
- Compatible Transcripts. All possible and known transcripts (as identified by their transcript IDs) that the fusion reads are compatible with, such as the 5' and 3' gene transcripts that include the exons which are expressed in the fusion.
- Translocation Name. Description of the fusion in the HGVS format (http://cancer.sanger.ac.uk/cosmic/help/fusion/summary) using the preferred transcript.
- Original chromosome. (only in tracks based on a fusion reference)
- Original breakpoint region. (only in tracks based on a fusion reference)
- Known Fusion. Indicates the Fusion ID Number of the matching fusion in the known fusion database. If the fusion is not found in the database, then -1 is reported. By default, the tool uses a QIAGEN known fusion database, but you can replace this database with another known fusions database relevant to your assay. Note that this database is not used to detect the fusions, but only for annotating the identified fusions output from the Detect Fusion Genes before refinement.

- Found in-frame CDS. This column is present when a CDS track was specified as input. It contains "Yes" if at least one fusion CDS that stays in frame across the fusion breakpoints has been found. Note that the in-frame calculation only takes into account the frame of the last included exon in the 5' gene and the first included exon in the 3' gene, and ignores more complex factors that might affect frame, such as frameshift mutations or stop codons due to variants around the fusion breakpoints.
- Promiscuity. Number of different potential fusion partners found for this gene.

Note about exporting output files in SAM/BAM format Exporting UMI reads as BAM files will show UMI described as Unique_Molecular_Index=[number1]_count=[number2], where number1 is a UMI ID (just a unique UMI group number), and number2 is the number of reads that are in that UMI group.

Note about interpreting the QC for RNAscan Panels Report The relationship between input, amplification and therefore number of reads per UMI with the original coverage is complex. For example, a "good" original coverage can be deceiving, if the user has amplified many times the original molecule: in that case, the information added to the experiment is close to null.

Note about known false positive fusions found when running QIAseq Targeted Fusion Panels

The following fusions can be disregarded as common read-through mRNAs or false fusions due to gene homology.

- HALC1-COLQ, common read through
- BCR-BCRP3 (BCRP3-BCR), known false positive
- TMP3-TMP4 (TMP4-TMP3), homologous genes

Note that these are found when running QIAseq Targeted Fusion Panels; additional false positive fusions may be found when running custom made panels.

4.2 Targeted RNAscan tools detailed description

4.2.1 Detect Fusion Genes

The Detect Fusion Genes tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq RNAscan Panel Expert Tools** | **Detect Fusion Genes**

The **Detect Fusion Genes** tool takes an RNA-seq read mapping as input (figure 4.4).

In the next dialog (figure 4.5), specify the reference sequence, gene and mRNA track that are saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNAscan Panels hg38** Reference Data Set. It is possible - but optional - to add a CDS or primer track to run the analysis.

The additional parameters to set are:

- **Maximum number of fusions**: The maximum number of putative fusions that will be evaluated.

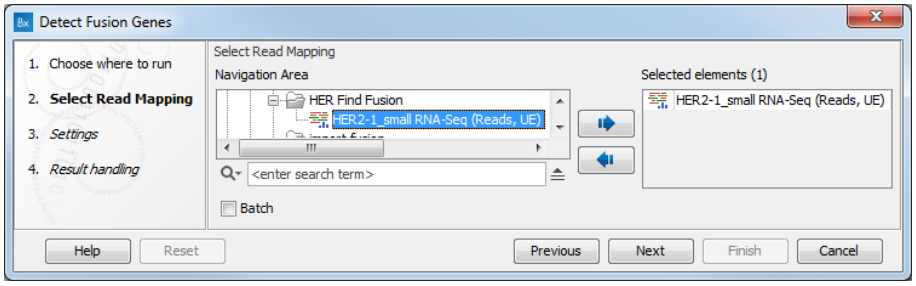


Figure 4.4: *Select an RNA-seq read mapping.*

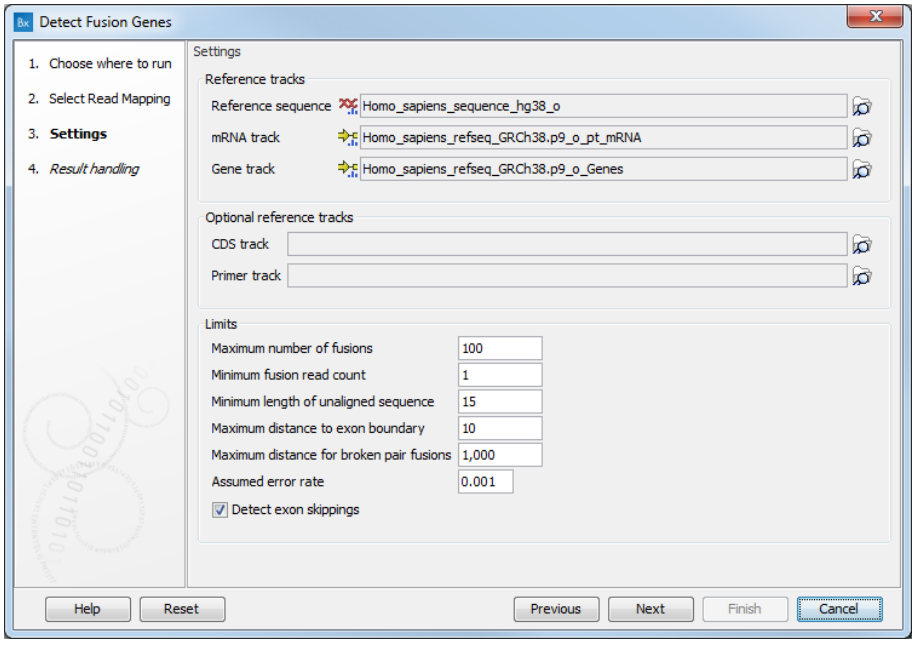


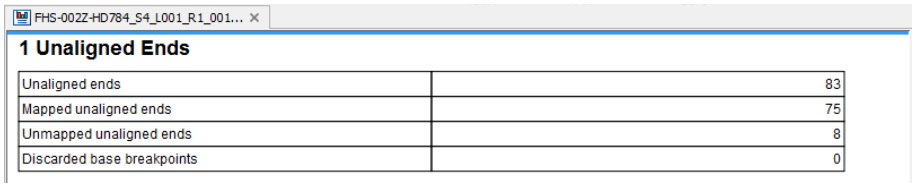
Figure 4.5: *Specify references and parameters for the detection.*

- **Minimum fusion read count**: This value is used to calculate Z-score and p-value, by subtracting that number from the total read count before doing the statistics.
- **Minimum length of unaligned sequence**: Only unaligned ends longer than this will be used for detecting fusions.
- **Maximum distance to exon boundary**: Reads must break within this distance of an exon boundary to consider the read and exon compatible.
- **Maximum distance for broken pairs fusions**: The algorithm uses broken pairs to find additional support for fusion events. If a pair of reads originally mapped as a broken pair, but would not be considered broken if mapped across the fusion breakpoints (because the two reads in the pair then get close enough to each other), then that pair of reads supports the fusion event as "fusion spanning reads". The "Maximum distance for broken pairs fusions" parameter specifies how close to each other two broken pairs must map across the fusion breakpoints in order for them to be considered fusion spanning reads. This is usually set to the maximum paired end distance used for the Illumina import of reads.
- **Assumed error rate**: Value used to calculate Z-score and p-value.

- **Promiscuity threshold**: Minimum number of different fusion partners to consider a gene promiscuous. Fusions with a promiscuous gene will get the annotation "Promiscuous" in the Filter column.
- **Detect exon skippings**: Check this option to consider same-gene fusions, i.e., exon-skipping fusions. Note however that same-gene fusions where the 5' breakpoint is downstream of the 3' breakpoint will not be considered.

In the Result handling dialog, it is possible to choose to output a report with unaligned ends information (figure 4.6):

- Unaligned ends: number of found unaligned ends.
- Mapped unaligned ends: number of unaligned ends which could be mapped
- Unmapped unaligned ends: number of unaligned which could not be mapped.
- Discarded base breakpoints: when two transcripts of the same gene overlap so that two breakpoints are found next to each other, one of them will be discarded.


FHS-002Z-HD784_S4_L001_R1_001...

1 Unaligned Ends

Unaligned ends	83
Mapped unaligned ends	75
Unmapped unaligned ends	8
Discarded base breakpoints	0

Figure 4.6: *Unaligned ends report.*

The tool otherwise generates a read mapping and a fusion track (see the details of the fusion track in section 4.1.1.

4.2.2 Refine Fusion Genes

The Refine Fusion Genes tool takes as input the fusions identified by the Detect Fusion Genes tool, and re-counts the number of fusion crossing reads as well as the wildtype supporting reads using the RNA-Seq mapping against the wild type and fusion references. The fusion reference is an artificial reference sequence that "assumes" the detected fusions by generating new chromosomes corresponding to each fusion in addition to the original chromosomes (figure 4.7).

RNASeq analysis is used to re-map the reads to the artificial reference, with the expectation that reads that were used to detect the fusion will now map to the fusion transcript with a spliced read. In addition, some reads that did not originally map at all will now map to the artifical reference sequence, increasing evidence for the fusion event. The tool then calculates the "refined" Z-score and p-value using the same binomial model as the Detect Fusion Genes tool.

The Refine Fusion Genes tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq RNAscan Panel Expert Tools** | **Refine Fusion Genes**

The **Refine Fusion Genes** tool takes a fusion track as input (figure 4.8).

The next dialog of the tool allows you to configure the following parameters (figure 4.9):

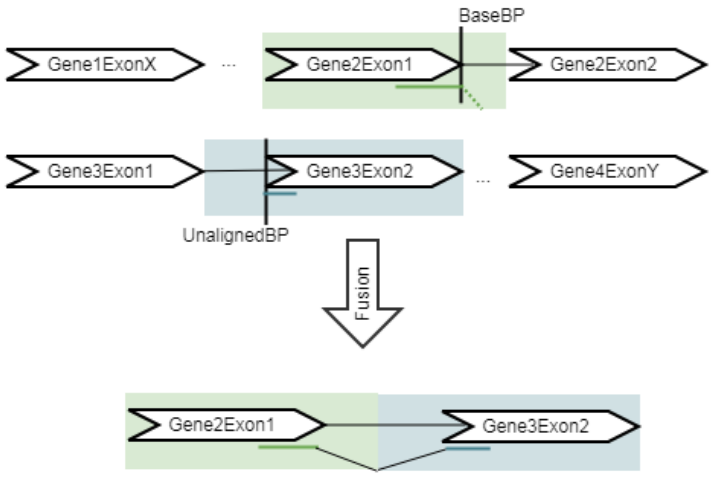


Figure 4.7: *An artificial chromosome is created consisting of the vicinity of both ends of the fusion. It will be output without any intron sequence in the fusion track.*

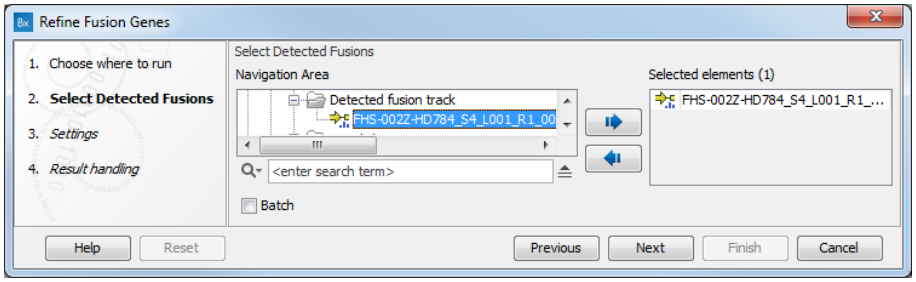


Figure 4.8: *Select a fusion track.*

- Reference tracks (wildtype + fusion): Specify a read mapping, a reference sequence with fusions, a mRNA track with fusions, or a detected fusions track: these files are output by the Define Fusion Genes.
- Optional reference tracks (wildtype + fusion): You can also add a Gene track, CDS track or Primer track with fusions
- Thresholds:
  - Breakpoint distance: Minimum distance from a read mapping exon position to a fusion breakpoint position required to count a read as supporting.
  - Assumed error rate: Probability of an unaligned end mapping to another gene by random.
  - Minimum number of supporting reads that should support a fusion. Fusions with fewer supporting reads will get a corresponding filter annotation.
  - Maximum p-value: Fusions with p-value above this threshold will get a corresponding filter annotation.

The output of the Refine Fusion Genes tool is the same fusion track as the one that was input, but now including a FILTER column, where fusions "PASS" only if they fulfill the p-value and minimum number of fusion crossing reads requirements supplied by the user as parameters.

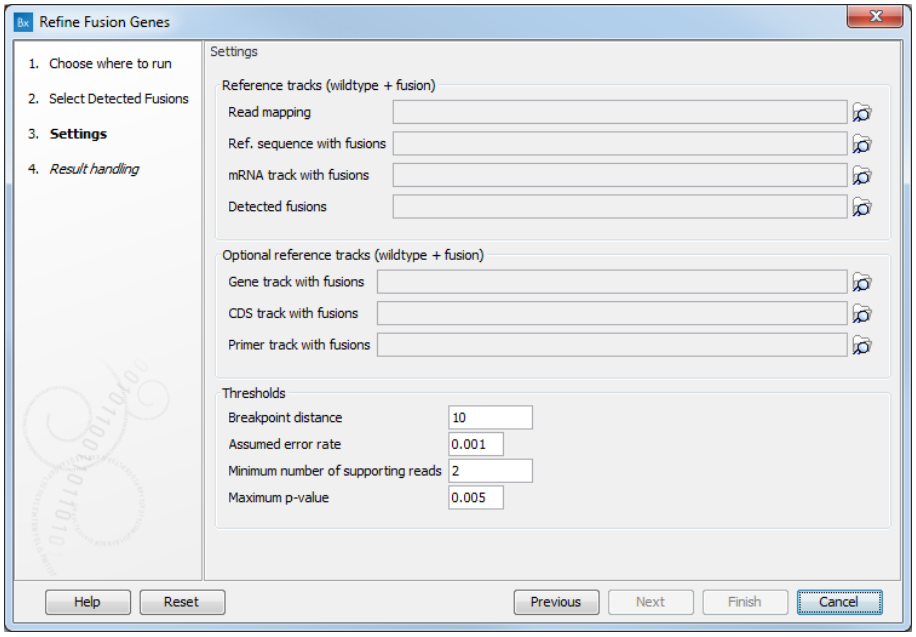


Figure 4.9: *Set the parameters for the Refine Fusion Genes tool.*

4.2.3 Annotate Fusions with Known Fusion Information

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

The Annotate Fusions with Known Fusion Information tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Analysis Tools** | **QIAseq RNAscan Panel Expert Tools** | **Annotate Fusions with Known Fusion Information**

Specify a fusion track as input (figure 4.10).

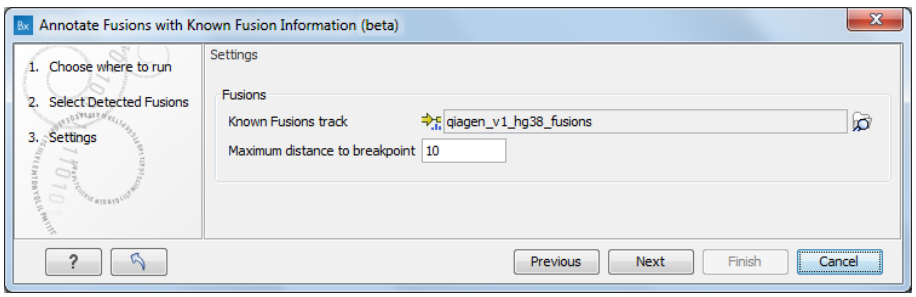


Figure 4.10: *Select a fusion track, and set up the maximum distance to breakpoint.*

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

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

The tool outputs a fusion track. It is similar to the one that was input, but includes now additional information from the known fusion track. For example, when using the known fusion track included in the **QIAseq RNAscan Panels hg38** Reference Data Set, this output track will have two additional columns for Catalog Panels IDs and Cancer Types.

4.2.4 QC for RNAscan Panels

The QC for RNAscan Panels tool can be found in the Toolbox here:

Tools | QIAseq Panel Analysis Tools | QIAseq RNAscan Panel Expert Tools | QC for RNAscan Panels

Specify a RNA-seq read mapping as input (figure 4.11).

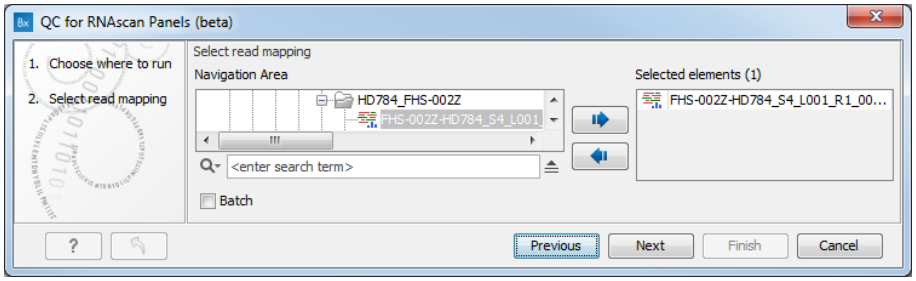


Figure 4.11: *Select a UMI read mapping.*

In the next dialog (figure 4.12), specify the mRNA track and the primer track that are saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNAscan Panels hg38** Reference Data Set. You can also set a maximal distance between a read and a primer start for them to be considered matching. It is set by default to 0, which means that a read will not be considered as starting in the primer unless it maps exactly to the start of the primer.

The tool outputs a primer track with annotated read coverage and a report that recapitulates QC data (figure 4.13). The primer track gives information about each primer, as well as their read coverage, whether they overlap with target or housekeeping genes.

See more details on the QC report in section 4.1.1.

4.2.5 Import Known Fusion Information Track

The Import Known Fusion Information Track tool can be found in the Import menu.

The tool takes a TAB separated text file that describes the known fusion information as input (figure 4.14). You can then choose the appropriate reference gene track, saved in the CLC_References folder of the Navigation Area when downloading the **QIAseq RNAscan Panels hg38** Reference Data Set.

The TAB separated text file format is as follow: The first row must contain the header which defines column names and number of columns in the table. Each subsequent row describes a known fusion information record consisting of a 5'prime and 3'prime gene name, an optional breakpoint position and a list of custom annotations.

The header must include include two specific columns: 5'prime and 3'prime gene names, and two optional columns: 5'prime and 3'prime breakpoint position. Column names are:

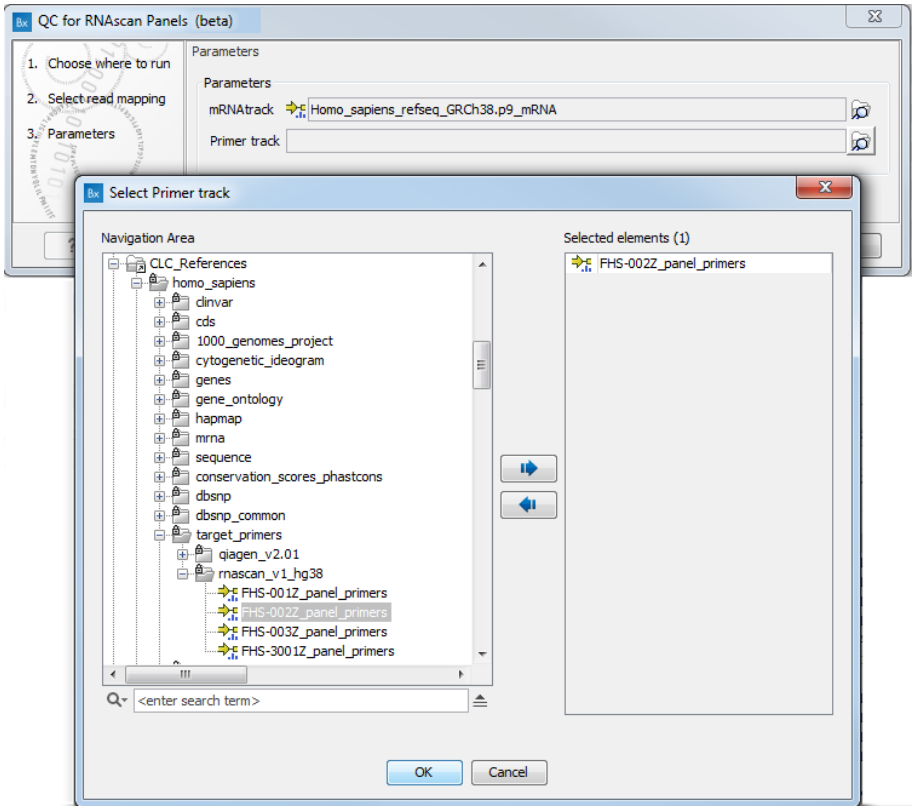


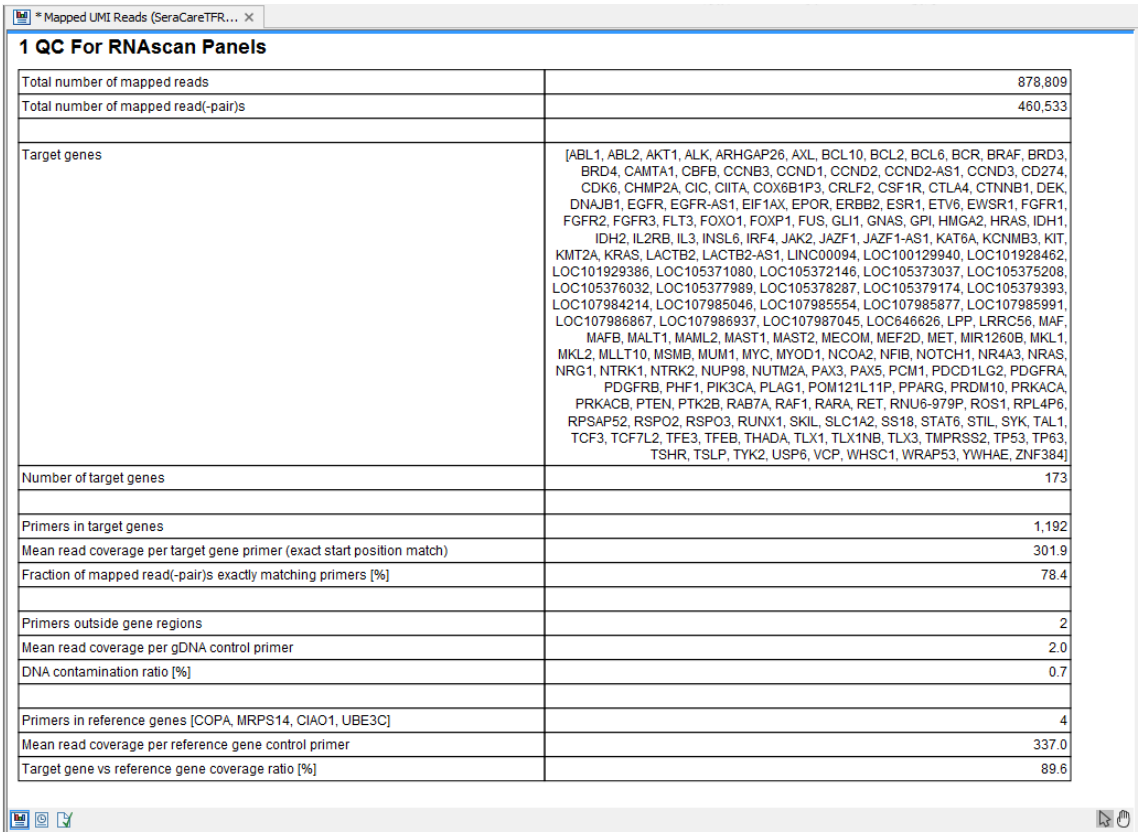
Figure 4.12: *Specify mRNA and primer tracks.*

- "5prime_gene" : 5'prime gene name.
- "5prime_bp" : 5'prime breakpoint position (optional).
- "3prime_gene" : 3'prime gene name.
- "3prime_bp" : 3'prime breakpoint position (optional).

Any additional columns are added as annotations for the known fusion information record. The order of the columns are not important since the header column names are used to identify the contents of the fusion information.

The 5'prime and 3'prime genes are looked up in the provided Gene track and must therefore be present.

Choose to save in the Navigation Area the newly created track element annotated with (KFI) and recapitulating all the information from the input file. This track can now be used with the Annotate Fusions with Known Fusion Information.



* Mapped UMI Reads (SeraCareTFR... ×

1 QC For RNAscan Panels

Total number of mapped reads	878,809
Total number of mapped read(-pair)s	460,533
Target genes	[ABL1, ABL2, AKT1, ALK, ARHGAP26, AXL, BCL10, BCL2, BCL6, BCR, BRAF, BRD3, BRD4, CAMTA1, CBFB, CCNB3, CCND1, CCND2, CCND2-AS1, CCND3, CD274, CDK6, CHMP2A, CIC, CIITA, COX6B1P3, CRLF2, CSF1R, CTLA4, CTNNB1, DEK, DNAJB1, EGFR, EGFR-AS1, EIF1AX, EPOR, ERBB2, ESR1, ETV6, EWSR1, FGFR1, FGFR2, FGFR3, FLT3, FOXO1, FOXP1, FUS, GLI1, GNAS, GPI, HMGA2, HRAS, IDH1, IDH2, IL2RB, IL3, INSL6, IRF4, JAK2, JAZF1, JAZF1-AS1, KAT6A, KCNMB3, KIT, KMT2A, KRAS, LACTB2, LACTB2-AS1, LINC00094, LOC100129940, LOC101928462, LOC101929386, LOC105371080, LOC105372146, LOC105373037, LOC105375208, LOC105376032, LOC105377989, LOC105378287, LOC105379174, LOC105379393, LOC107984214, LOC107985046, LOC107985554, LOC107985877, LOC107985991, LOC107986867, LOC107986937, LOC107987045, LOC646626, LPP, LRRC56, MAF, MAFB, MALT1, MAML2, MAST1, MAST2, MECOM, MEF2D, MET, MIR1260B, MKL1, MKL2, MLLT10, MSMB, MUM1, MYC, MYOD1, NCOA2, NFIB, NOTCH1, NR4A3, NRAS, NRG1, NTRK1, NTRK2, NUP98, NUTM2A, PAX3, PAX5, PCM1, PDCD1LG2, PDGFRA, PDGFRB, PHF1, PIK3CA, PLAG1, POM121L11P, PPARG, PRDM10, PRKACA, PRKACB, PTEN, PTK2B, RAB7A, RAF1, RARA, RET, RNU6-979P, ROS1, RPL4P6, RPSAP52, RSPO2, RSPO3, RUNX1, SKIL, SLC1A2, SS18, STAT6, STIL, SYK, TAL1, TCF3, TCF7L2, TFE3, TFEB, THADA, TLX1, TLX1NB, TLX3, TMPRSS2, TP53, TP63, TSHR, TSLP, TYK2, USP6, VCP, WHSC1, WRAP53, YWHAE, ZNF384]
Number of target genes	173
Primers in target genes	1,192
Mean read coverage per target gene primer (exact start position match)	301.9
Fraction of mapped read(-pair)s exactly matching primers [%]	78.4
Primers outside gene regions	2
Mean read coverage per gDNA control primer	2.0
DNA contamination ratio [%]	0.7
Primers in reference genes [COPA, MRPS14, CIAO1, UBE3C]	4
Mean read coverage per reference gene control primer	337.0
Target gene vs reference gene coverage ratio [%]	89.6

Figure 4.13: *QC for RNAscan Panels report.*

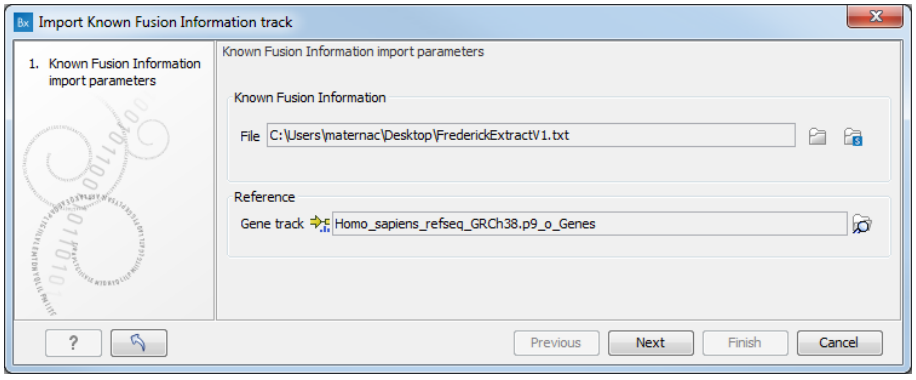


Figure 4.14: *Select a text file containing the previously known fusion information, and a reference gene track.*

Chapter 5

Targeted RNA

Contents

5.1 The Quantify QIAseq RNA Expression ready-to-use workflow

Before starting the workflow, make sure the correct Reference Data Set is applied: open the Data Manager, select **QIAseq RNA Panels hg38**, **Download** the set if you have not done so before and click on **Apply**. You can now close the Data Management window.

The Quantify QIAseq RNA Expression ready-to-use workflow can be found here:

> **Toolbox** | **Ready-to-Use Workflows** | **QIAseq Panel Analysis** | **Quantify QIAseq RNA Expression**

Double-click on the Quantify QIAseq RNA Expression ready-to-use workflow to run the analysis.

If you are connected to a CLC Server via your Workbench, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a CLC Server when possible.

The workflow can and should be run in batch mode, allowing the analysis of several samples at once. Once you have checked the **Batch** option, you can select the **folder** holding the samples that should be analyzed (figure 5.1).

When working in batch mode, it is important to select the folder containing the samples, and not the subfolders containing the sequence lists, nor the reads themselves. In the two latter cases, each sequence list would be considered as an independent sample, when in fact, individual samples are usually made of several sequence lists.

The batch overview dialog that comes next in the wizard allows you to check that the batch unit is the sample, as opposed to independent sequence lists. You can take advantage of this dialog to exclude some samples from your analysis (figure 5.2).

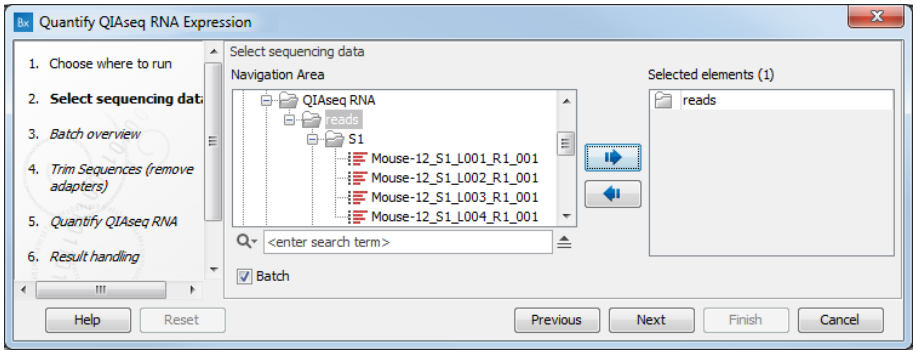


Figure 5.1: *Select the samples to analyze by working in batch mode and choosing the top folder holding all samples.*

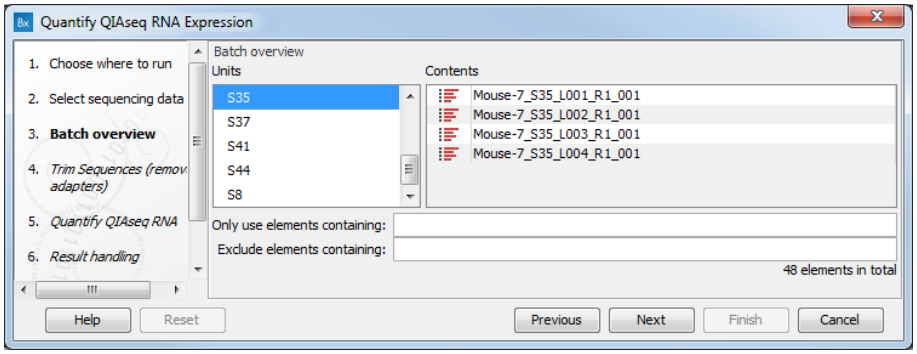


Figure 5.2: *Check in this dialog that the batch unit is the sample and not the reads.*

Next, the **Trim adapter list** available in the CLC_References folder once you downloaded the QIAseq RNA Panels hg38 or Mouse Reference Data Set shoudl be preselected provided that you applied the Reference Data Set before starting the workflow (figure 5.3).

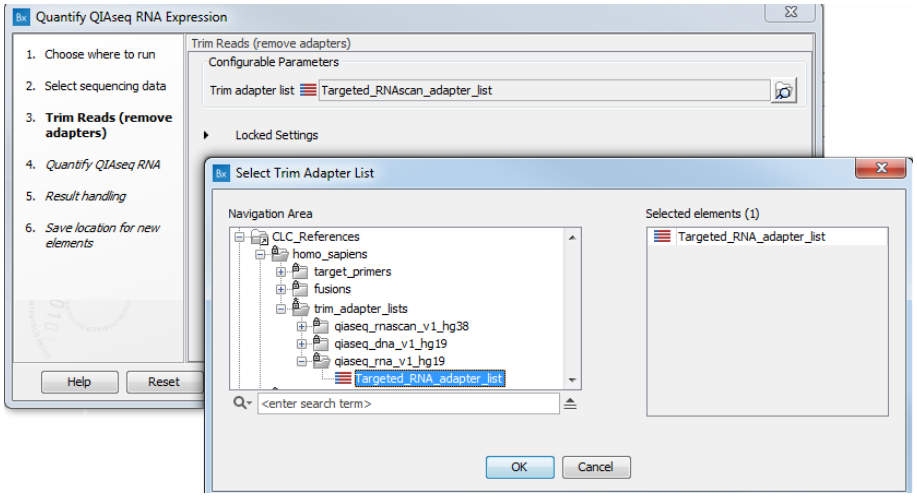


Figure 5.3: *The relevant Targeted RNA adapter list should be preselected.*

The Trim Sequences tool included in the workflow will trim the reads based on the Universal PCR adapter list that was specified as well all reads that are shorter than 60bp as these are likely to be oligo dimer products.

In the Quantify QIAseq RNA Expression dialog, select the reference relevant for the panel used. If you have configured Data Management properly, the Reference sequence should already be specified in the first field. In the field below, specify the Target regions file that correspond to the panel used. You can find this reference in the folder CLC_References as indicated on figure 5.4.

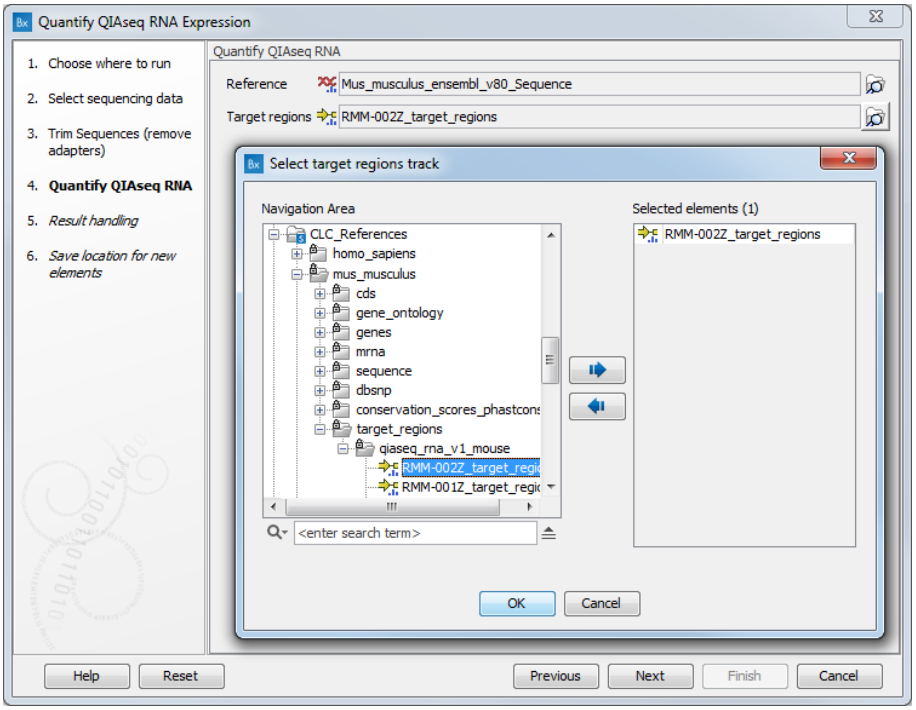


Figure 5.4: *Select the Target regions track that correspond to the panel used.*

Finally, in the last wizard step, choose to **Save** the results of the workflow before clicking **Finish**.

5.1.1 Output from the Quantify QIAseq RNA Expression workflow

The **Quantify QIAseq RNA Expression** ready-to-use workflow produces expression tracks ( ) for each sample analyzed.

Expression tracks are displayed as tables listing genes included in the panel and several measures of their expression levels.

- **Expression value** The number of distinct UMIs seen for this gene.
- **TPM** (Transcripts per Million) The number of transcripts per million that come from this gene. This is computed as the relative abundance per million $(X_i / \sum_j X_j)$.
- **RPKM** (Reads per Kilobase Million) There is no good definition of RPKM for targeted amplicon data. We therefore define RPKM to be equal to TPM, which preserves the expected property that RPKM is proportional to TPM.
- **Total gene reads** The number of distinct UMIs seen for this gene.
- **Read counts** The total number of reads mapping to this gene. Several reads may have the same UMI.

- **UMI counts** The number of distinct UMIs seen for this gene.
- **Mean reads per UMI** The mean number of reads for each UMI seen for this target.

Expression tracks be further analyzed using statistical tools from the RNA-Seq Analysis folder of the workbench, such as PCA for RNA-Seq and Create Heat Map for RNA-Seq.

Why does the workflow not produce a read mapping? The very short target regions in a QIAseq Targeted RNA Panel are not suited to downstream analyses that require a read mapping, such as variant calling. If a read mapping is desired, for example to investigate suspected off-target effects, we recommend using the Map Reads to Reference tool.

5.2 Targeted RNA tools detailed description

5.2.1 The Quantify QIAseq RNA tool

The Quantify QIAseq RNA tool is available as a stand alone tool in the Toolbox.

> **Tools | QIAseq Panel Analysis Tools | QIAseq RNA Panel Expert Tools | Quantify QIAseq RNA**

It is there solely for the purpose of building additional workflows and we do not recommend to use the tool on its own as the trimming step included in the Quantify QIAseq RNA Expression ready-to-use workflow increases the accuracy of the results generated.

The tool works in a three-step process: 1) Mapping, 2) Filtering of mapped reads, 3) Merging of reads with similar UMIs.

Mapping RNA-Seq reads are mapped to the target regions. All other regions of the genome are ignored.

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

- The sequence of the target regions is determined from the BED file.
- The exons of multi-exon targets are concatenated into one single target sequence.
- 12 ambiguous "N" nucleotides are prepended to the 5’ region to allow reads with UMIs to map without mismatch penalties.

Filtering Reads mapping to the targets are checked to ensure that the UMI has the expected length. Reads with UMIs that are as little as one base pair too long or too short are filtered away.

Merging To account for the fact that PCR and sequencing errors also happen in the UMI region, some UMIs may be merged.

To be considered for a merge, a UMI has to have a count below 5% of the maximum count.

To be merged, the count difference between the UMIs must be at least 6 fold, and they must differ from from another UMI by at most one base pair. The algorithm is described in more detail in Peng et al., 2015.

5.2.2 Differential Expression for Targeted RNA-Seq

The **Differential Expression for Targeted RNA-Seq** tool can be found inthe Toolbox here:

> **Tools | QIAseq Panel Analysis Tools | QIAseq RNA Panel Expert Tools | Differential Expression for Targeted RNA-Seq**

Select at least 2 expression tracks (figure 5.5), but do not use the "batch mode" in this tool.

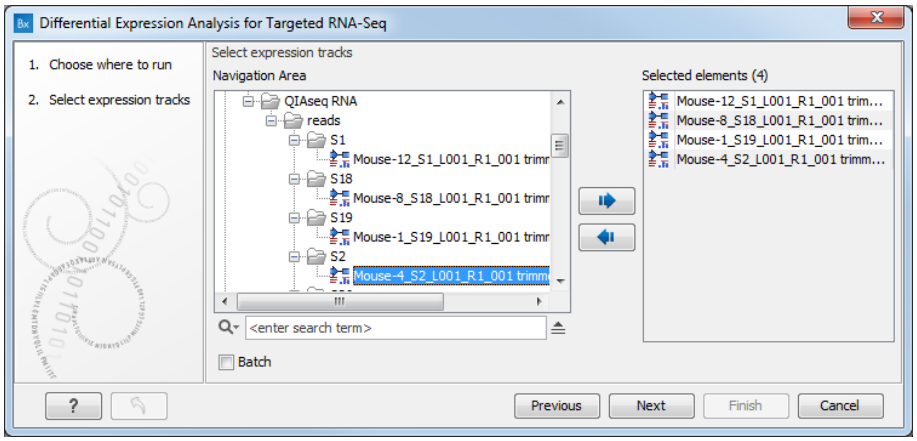


Figure 5.5: *Select at least two expression tracks.*

In the next wizard step you can set up the experimental design associated with the data (figure 5.6).

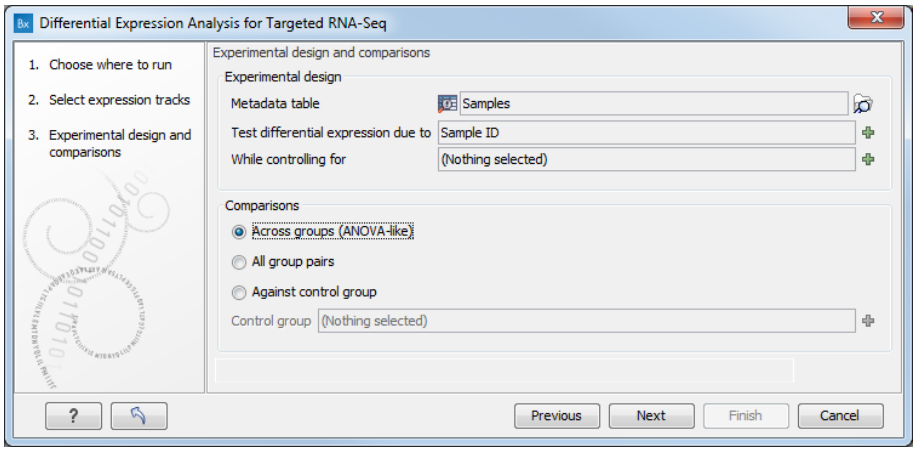


Figure 5.6: *Setting up the experimental design and comparisons.*

In the **Experimental design** panel, a Metadata table must be selected that describes the factors and groups for all the samples

- **Metadata table** The metadata table describing the factors for the selected Expression tracks.
- **Test differential expression due to** Specify the one factor differential expression is tested for.
- **While controlling for** Specify confounding factors, i.e., factors that are not of primary interest, but may affect gene expression.

The **Comparisons** panel determines the number and type of statistical comparison tracks output by the workflow (see http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Output_Differential_Expression_RNA_Seq_tool.html for more details).

- **Across groups (ANOVA-like)** This mode tests for the effect of a factor across all groups.
- **All group pairs** tests for differences between all pairs of groups in a factor.
- **Against control group** This mode tests for differences between all the groups in a factor and the named reference group. In this example the reference group is skin.

In the next step you can configure the normalization method (figure 5.7).

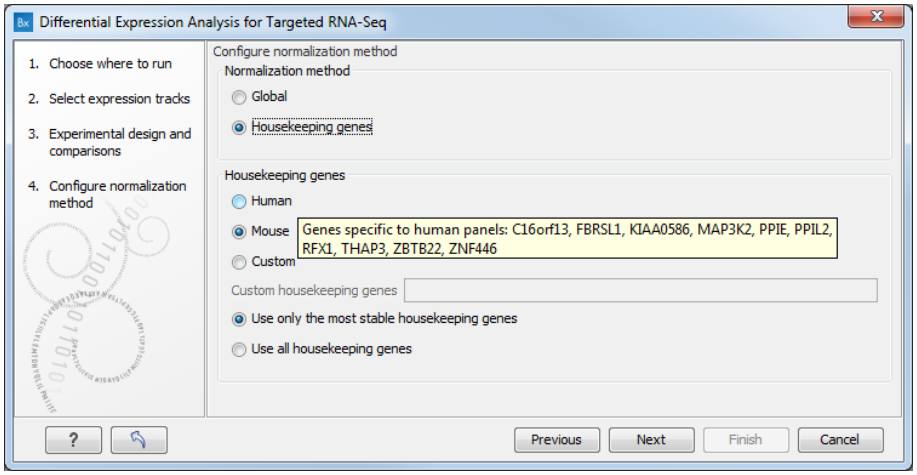


Figure 5.7: *Setting up the experimental design and comparisons.*

The option "Global" uses the Global TMM Normalization (Trimmed Mean of M values) to calculate effective libraries sizes, which are then used as part of the per-sample normalization. TMM normalization adjusts library sizes based on the assumption that most genes are not differentially expressed. It is the normalization used by the Differential Expression for RNA-Seq tool.

When working with Targeted RNA Panels, it is recommended to use Housekeeping genes for normalization. In this case, specify which set of housekeeping genes to use. Hover on the dialog to find the list of genes included in the Human and Mouse sets. If you are working with a custom panel, you can also provide the corresponding set of housekeeping genes in the "Custom housekeeping genes" field: type the name of the genes separated by a space.

Finally choose between these two options:

- **Use only the most stable housekeeping genes** will use a subset (at least three) of the most stable genes for normalization, these being defined using the GeNorm algorithm [Vandesompele et al., 2002].
- **Use all housekeeping genes** keep all housekeeping genes for normalization.

In the last wizard window, choose to preview the settings and Save the results. Click **Finish** to start the analysis. The tool will output statistical comparisons () that can be further analyzed with the Create Venn Diagram for RNA-Seq tool.

Chapter 6

QCI Interpret Integration

Contents

6.1 Configuration of the Upload to QCI Interpret tool

The workbench can be configured to allow seamless upload of analysis results to QCI Interpret. Configuration of the Upload to QCI Interpret tool depends on having an active QCI Interpret subscription (not a trial license), and a QCI Interpret account with API license privileges.

If an API user account has not been set up for you, please contact support-license@qiagen.com. In your email, ask to have an API enabled QCI Interpret account set up for the Biomedical Workbench integration. The license team can then send you your API Key ID and API Key Secret.

To retrieve yourself your user's API Key ID and an API Key Secret number, log in to QCI Interpret using the following links:

- QCI Interpret US: `https://apps.ingenuity.com/qcibridge`
- QCI Interpret UK: `https://apps.qiagenbioinformatics.eu/qcibridge`
- QCI Interpret China: `https://variants.qiagenbioinformatics.cn/qcibridge`

You can then access the **API Explorer** (as seen in figure 6.1 and figure 6.2):

Once you have the necessary information, you can start using the Upload to QCI Interpret tool.

6.2 The Upload to QCI Interpret tool

The Upload to QCI Interpret tool can be found here:

> **Tools** | **QIAseq Panel Analysis Tools** | **Variant Interpretation Tools** | **Upload to QCI Interpret**

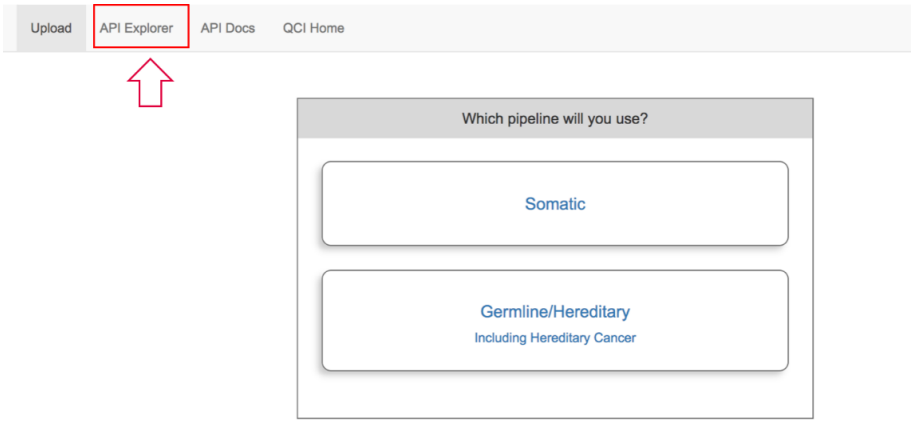


Figure 6.1: *Accessing the QCI Interpret API Explorer.*

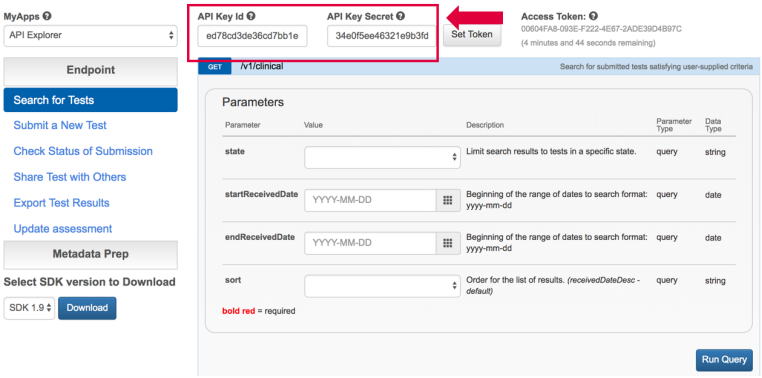


Figure 6.2: *Finding the API Key Id and API Key Secret in the API Explorer.*

Double-click on the Upload to QCI Interpret tool to send a VCF file and possibly metadata information to QCI Interpret.

Select the variant track you want to send (as VCF) to QCI Interpret (figure 6.3), and then select the reference sequence (hg19 by default, but hg38 is also supported). You can then choose which application you would like for the QCI Interpret results (Somatic or Hereditary).

In case you are working with RNAscan panels analysis, select a Fusions track to upload to QCI Interpret.

QCI Interpret configuration requires the user to fill the QCI Interpret dialog (figure 6.4) with the following information:

- Server location. QCI Interpret is hosted in multiple locations. Your QCI Interpret account is created for a specific QCI Interpret server that should be specified here.
- API key ID and API key secret. If you do not know your API credentials, contact support-license@qiagen.com.
- Username. Insert the username associated with the QCI Interpret account.

For additional information on QCI Interpret user account and the QCI Interpret uploader, please

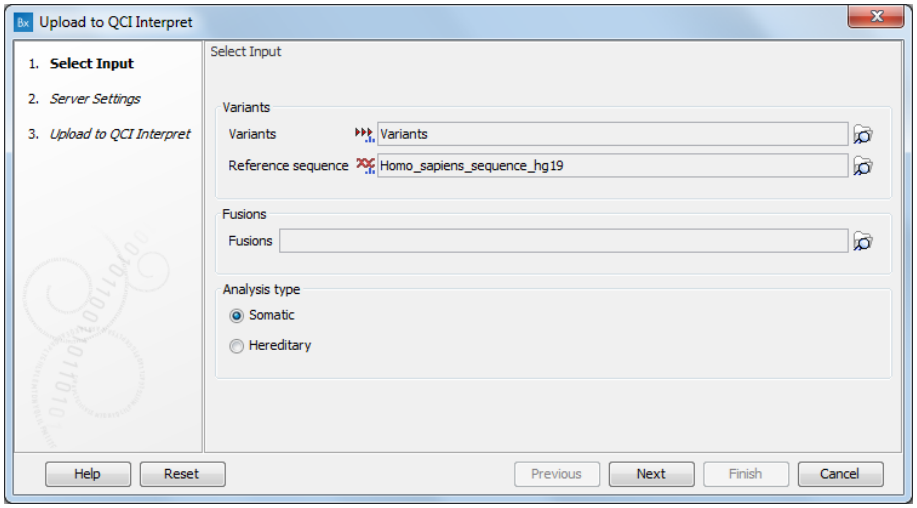


Figure 6.3: *The Upload to QCI Interpret dialog.*

refer to the QCI Interpret documentation.

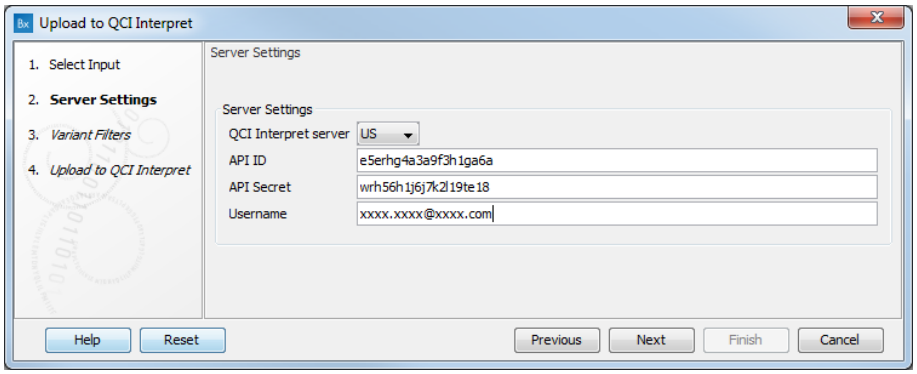


Figure 6.4: *Configure the connection to the QCI Interpret server dialog.*

In the case of uploading a variant file to QCI Interpret, a dialog called "Variants Filters" will appear next (figure 6.5).

Larger VCF files may need to be filtered to remove the excess variants found during second analysis. Use the next dialog to reduce the number of variants uploaded to QCI Interpret, in particular for somatic pipeline where the upload of variants is currently limited to 400.

Different types of filter are available:

- Genes and chromosomes. Select from the list that opens when clicking to the right of the relevant field all genes or chromosomes of interest. By clicking OK, the pop up window closes and reveals how many variants passed the filtered in the Preview section of the dialog.
- Coding changes. Choose between *No restriction*, *Only gene region changes*, *Only amino acid changes*.
- Frequency and quality. Modify the different filters to reduce the list of variants passing filters (see the Preview section of the dialog).

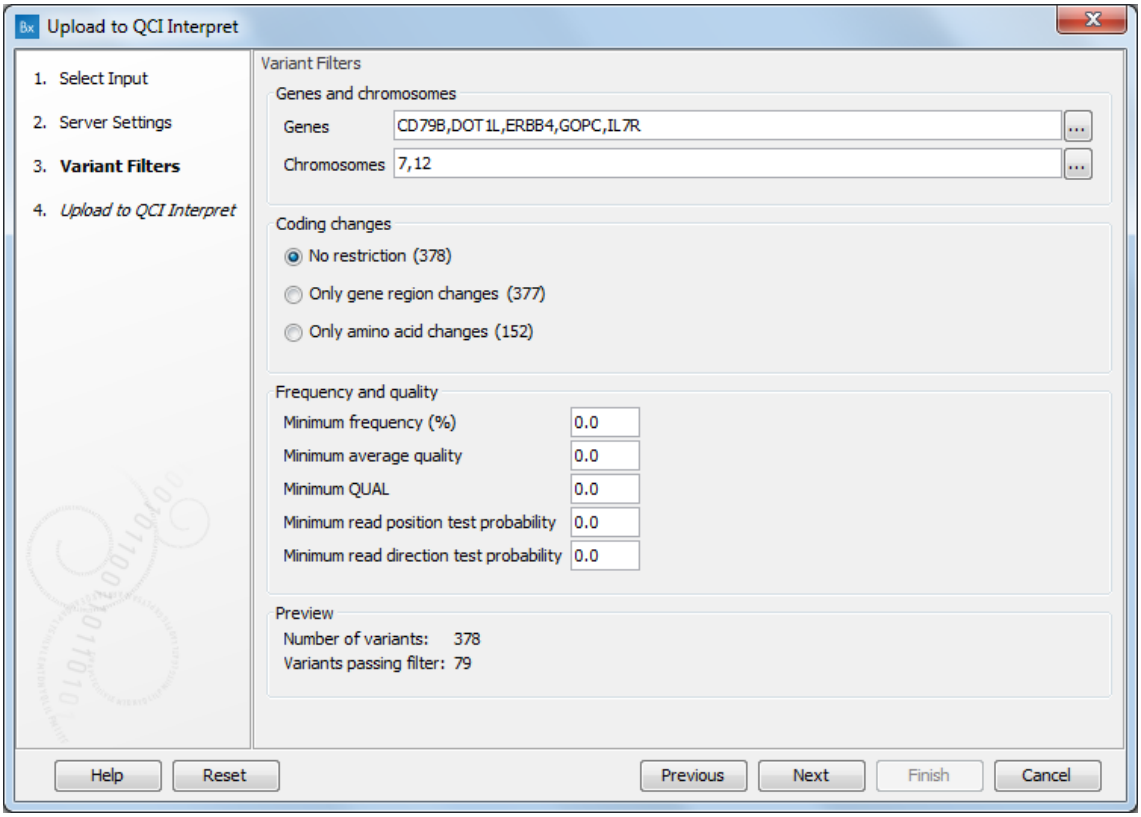


Figure 6.5: *The Variants Filters dialog, with the pop up window that allows users to select variants based on their chromosomal location opened as example.*

In the last wizard window, click on **Finish** to initiate the transfer of variants or fusions to QCI Interpret. The tool will open a browser window with the QCI Interpret interview page. This page must be filled in and the button Submit must be pressed to complete the upload of the files and metadata to QCI Interpret. In cases where a browser window fails to open automatically, the tool will also output a report containing the link to the QCI Interpret interview page.

Chapter 7

Legacy tools

Contents

7.1 Trim Primers of Mapped Single Reads

The tool **Trim Primers of Mapped Single Reads** removes the primer parts of reads as they reflect the primer that was added and not the actual sample.

The tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Expert Tools** | **QIAseq DNA Panel Expert Tools** () | **Trim Primers of Mapped Single Reads** ()

In the first dialog (figure 7.1), select a read mapping.

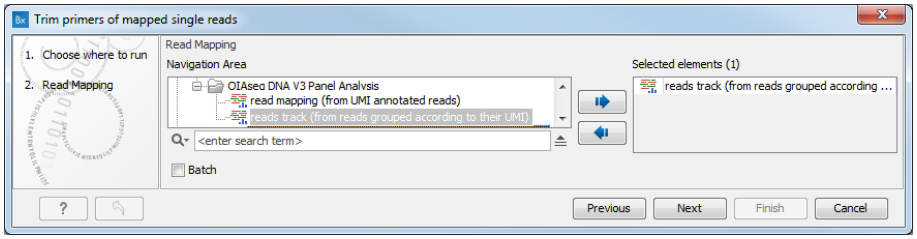


Figure 7.1: *Select a read mapping.*

In the second dialog (figure 7.2), select the primer annotation track that was provided with the QIAseq DNA V3 Panel.

Unlike other similar tools, this tool works on reads that potentially ends in a primer (rather than starts in a primers). The tool aims to unalign the primer parts of reads that came from that primer. It approximates this by only unaligning reads that end inside of a primer or up to 3 (set by the parameter "**Extra width at end**") extra bases after the primer. After unaligning the primer part of the reads, the tool removes reads where less than 20 (set by the parameter "**Removal Cutoff Threshold**") aligned bases remain. This feature can be disabled by unchecking the parameter

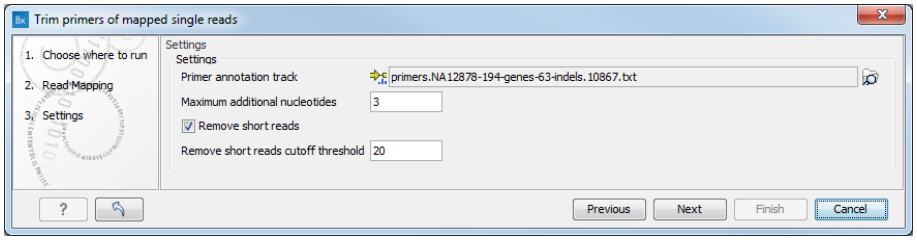


Figure 7.2: *Select the primer annotation track specific to the panel.*

"**Remove Reads that are Small After unaligning**".

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

7.2 Trim Primers of Mapped Paired End Reads

The tool **Trim Primers of Mapped Paired End Reads** removes the primed parts of reads as they reflect the primer that was added and not the actual sample.

The tool can be found in the Toolbox here:

> **Tools** | **QIAseq Panel Expert Tools** | **QIAseq DNA Panel Expert Tools** () | **Trim Primers of Mapped Paired End Reads** ()

In the first dialog (figure 7.3), select a read mapping.

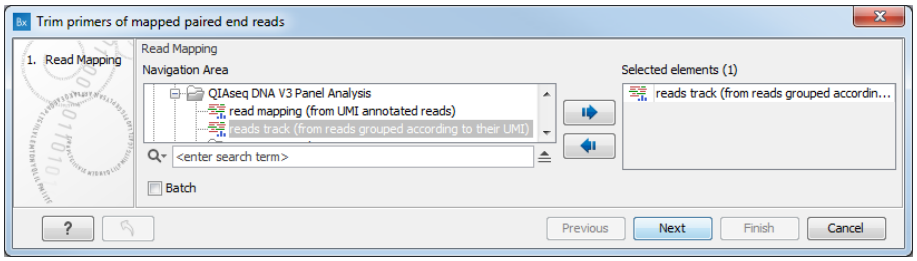


Figure 7.3: *Select a read mapping.*

In the second dialog (figure 7.4), select the primer annotation track that was provided with the QIAseq DNA V3 Panel.

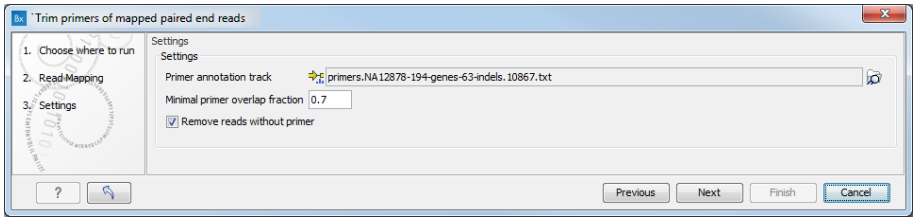


Figure 7.4: *Select the primer annotation track specific to the panel.*

If an aligned read starts within the span of a primer, and if it overlaps the primer with at least 70% (set by default for the "**Minimal primer overlap fraction**" option), then it is said to "hit"

the primer. For reads "hitting" a primer, the part of the read that overlaps the primer will be unaligned. For reads not "hitting" a primer, the read will either be removed from or retained in the read mapping, depending on the option "**Remove reads without primer**". If the option is checked, the tool will remove reads that do not "hit" a primer.

Chapter 8

Install and uninstall plugins

QIAseq Targeted Panel Analysis is installed as a plugin.

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

Plugins are installed and uninstalled using the plugin manager.

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

The plugin manager has two tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins on QIAGEN Aarhus server.

8.1 Install

To install a plugin, click the **Download Plugins** tab. This will display an overview of the plugins that are available for download and installation (see figure 8.1).

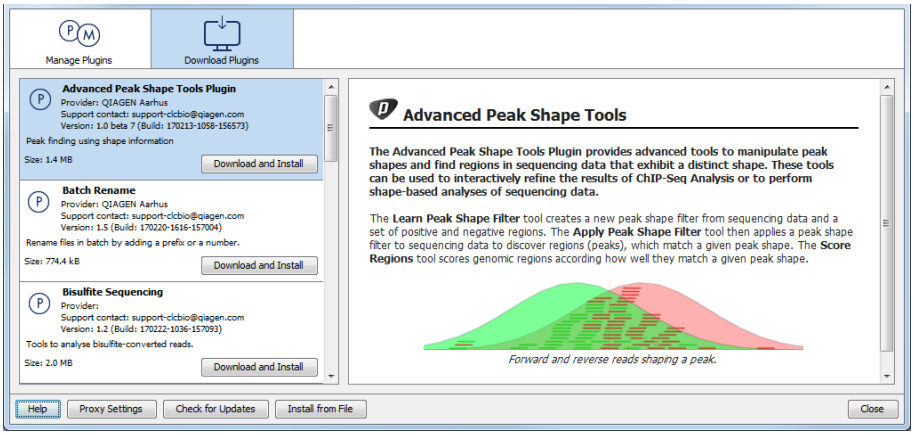


Figure 8.1: *The plugins that are available for download.*

Select QIAseq Targeted Panel Analysis to display additional information about the plugin on the

right side of the dialog. Click **Download and Install** to add the plugin functionalities to your workbench.

Accepting the license agreement

Part of the installation involves checking and accepting the end user license agreement (EULA) as seen in figure 8.2.

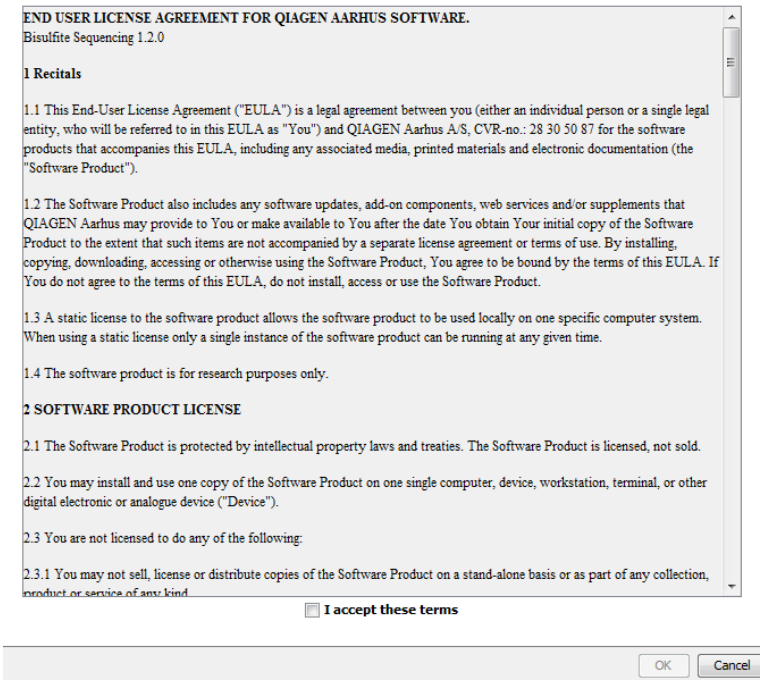


Figure 8.2: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept. If requested, fill in your personal information before clicking **Finish**.

If QIAseq Targeted Panel Analysis is not shown on the server but you have the installer file on your computer (for example if you have downloaded it from our website), you can install the plugin by clicking the **Install from File** button at the bottom of the dialog and specifying the plugin *.cpa file saved on your computer.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be ready for use until you have restarted.

8.2 Uninstall

Plugins are uninstalled using the plugin manager:

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

This will open the dialog shown in figure 8.3.

The installed plugins are shown in the **Manage plugins** tab of the plugin manager. To uninstall, select QIAseq Targeted Panel Analysis and click **Uninstall**.

If you do not wish to completely uninstall the plugin, but you do not want it to be used next time you start the Workbench, click the **Disable** button.

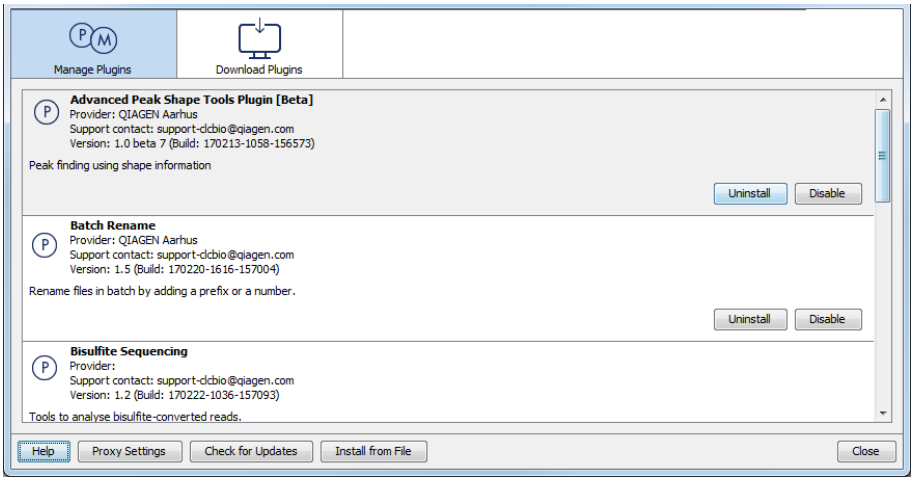


Figure 8.3: *The plugin manager with plugins installed.*

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be uninstalled until the workbench is restarted.

Bibliography

[Hiatt et al., 2013] Hiatt, J., Pritchard, C., Salipante, S., O'Roak, B., and Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Research*, (23):843–854.

[Peng et al., 2015] Peng, Q., Satya, R. V., Lewis, M., Randad, P., and Wang, Y. (2015). Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics*, (16):589.

[Vandesompele et al., 2002] Vandesompele, J., Preter, K. D., Pattyn, F., Poppe, B., Roy, N. V., Paepe, A. D., and Speleman, F. (2002). Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. *Genome Biol.*