



# Tutorial

## Find Very Low Frequency Variants With QIAGEN GeneRead Panels

June 20, 2019

— Sample to Insight —

## Find Very Low Frequency Variants With QIAGEN GeneRead Panels

This tutorial uses the capabilities of the CLC Genomics Workbench and the Biomedical Genomics Analysis plugin to find very low frequency variants in Targeted Amplicons sequences generated using a QIAGEN GeneRead panel kit.

This tutorial covers in just a few steps all the following:

- Import Illumina paired reads in the workbench.
- Create an adapter list and use the Prepare Raw data Ready-to-Use Workflow to remove remaining adapters from the reads and generate QC reports.
- Find low frequency variants with the QIAGEN GeneRead Panel Analysis Ready-to-Use Workflow that performs:
  - Alignment of data to reference
  - Removal of primers and primer-dimer artifacts
  - Very Low Frequency Variant Detection
  - Generation of additional reports (alignment report, coverage report)
  - Annotation of variants and setting up a Track List.

**Prerequisites** For this tutorial, you must have installed the Biomedical Genomics Analysis plugin. How to install plugins is described here: <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html>.

### Import data and prepare your reads

This tutorial makes use of Targeted Amplicon Sequencing data provided by Dr Matthew Smith of University Hospitals Birmingham NHS Foundation Trust using HDx Reference Standards from Horizon's diagnostics division. The sample used comes from a cell line in which variants were introduced at very low frequency, and was sequenced on an Illumina platform. Go through the following steps to download and import the data into the Workbench.

1. Download the sample data from our website: [http://resources.qiagenbioinformatics.com/testdata/QIAGEN\\_GeneRead\\_panels.zip](http://resources.qiagenbioinformatics.com/testdata/QIAGEN_GeneRead_panels.zip).
2. Start the workbench.
3. Import the reads via the toolbar: **File** | **Import** (📁) | **Illumina**
  - Select the two fastq paired files.
  - Under "General Options" section, ensure the **Paired reads** and **Discard read names** checkboxes are checked.
  - In the "Paired read orientation" section, ensure the **Paired-end (forward-reverse)** option is checked.
  - Set the **Minimum distance** to 1 and the **Maximum distance** to 1000 (default values).
  - Click **Next**.

- Click on the button labeled **Save** in the wizard page that appears, choose the folder you wish to save the reads to (you can create a new folder dedicated to this tutorial for example) and click **Finish**.

As remaining adapter sequences in the reads might lead to bias in downstream data analysis, we recommend to trim them off.

For this you will have to create an adapter list using the oligonucleotide sequences disclosed in the Illumina Customer Sequence Letter found at this address: <https://support.illumina.com/downloads/illumina-adapter-sequences-document-100000002694.html>.<sup>1</sup>

- Open the workbench and go to **File | New | Trim Adapter List (🇺🇸)**
- Click on the button **Add Row (+)** found at the bottom of the View Area in the New Trim Adapter List.
- Paste the name and sequence of the TruSeq Universal Adapter. Choose to search on **All** reads, to **Remove the adapter and following sequence (3' trim)**, and to **Keep the read** found without adapters (figure 1).

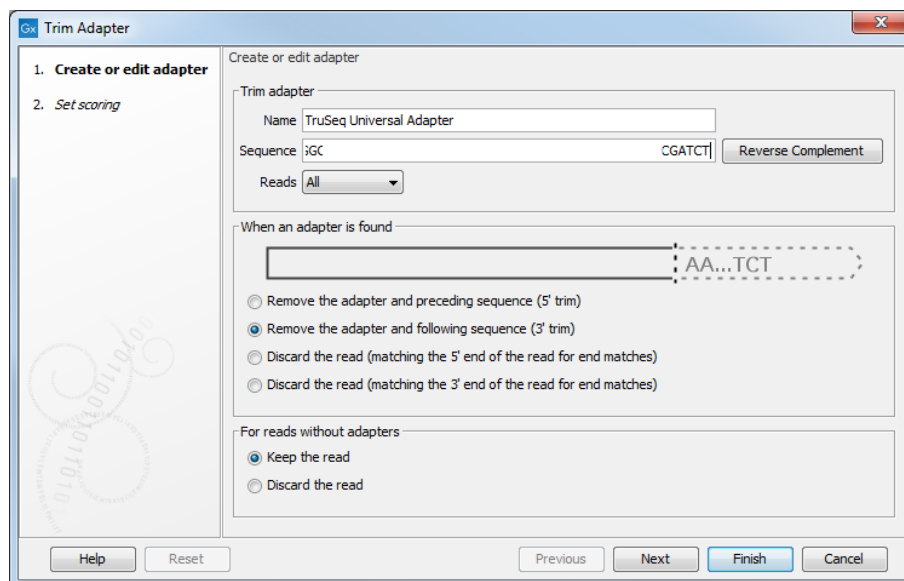


Figure 1: Creating the adapter list. Note that the sequence written in the figure is a mock one to respect Illumina copyrights.

Leave the scores set by default, i.e., **Mismatch cost** is set at 2 and **Gap cost** at 3. **Allow both internal and end matches** is checked and Minimal score are set at 10 and 4 respectively.

Click on the button labeled **Finish** to add the adapter to the trim list.

- In the Illumina letter, copy only the part of the sequence found before the index (underlined in the letter) of any TruSeq Index Adapters. In this way you only need to add one adapter sequence and not all the index adapters.

<sup>1</sup>Oligonucleotide sequences © 2007-2013 Illumina, Inc. All rights reserved. No sponsorship or affiliation. Link provided for convenience. QIAGEN not responsible for content at link.

5. Back to the New Trim Adapter List, click on the button **Add Row** (+) again.
6. Name the new adapter TruSeq Index Adapter, and paste the sequence you just copied and click on **Reverse complement**. Choose to search on **All** reads, to **Remove the adapter and following sequence (3' trim)**, and to **Keep the read** found without adapters.
7. Leave the other scoring parameters as default. Click on the button labeled **Finish** to add the adapter to the list.
8. Save the generated trim adapter list as **TrueSeqAdapters** in the Navigation Area. You can do this by going to **File** in the menu bar and the choose **Save as**.

Now we are ready to trim the adapters from our reads. Go to

**Ready-to-Use Workflows | Preparing Raw Data** (📁) | **Prepare Raw Data** (🔧)

1. This opens a dialog where you select the HD701 sample as shown in figure 2.

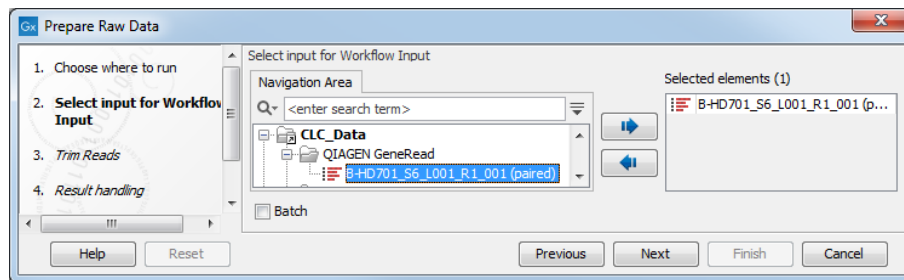


Figure 2: Selecting the sample for extracting and counting the small RNAs.

2. In the next dialog (figure 3). Click on the browse icon next to the "Trim Adapter list" field, and in the dialog that opens, select the **TrueSeqAdapters** that you have just created and click OK. Leave all parameters as they are set by default and click **Next**.

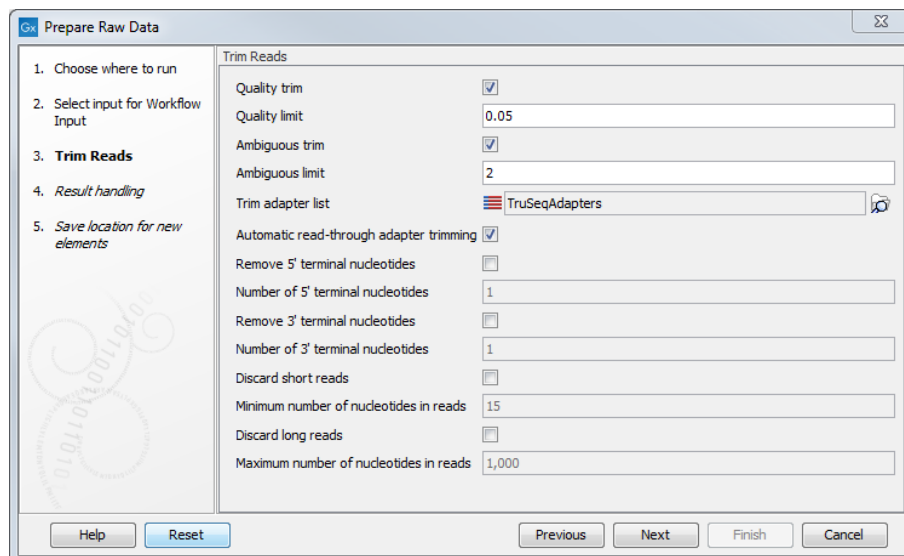


Figure 3: Choosing to trim for adapter sequence.

3. In the Result handling window, choose to **Save** the outputs files, and click on the button labeled **Finish**.

Once the analysis is complete, three reports and two trimmed sequence lists (📄) have been generated (figure 4).

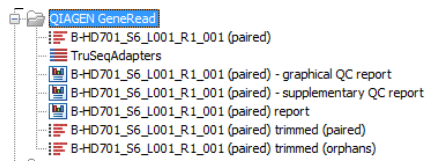


Figure 4: Choosing to trim for adapter sequence.

Open the Trim report (called B-HD701\_S6\_L001\_R1\_001 (paired) report) (figure 5). It is meant to be used as a quality check, mainly to see that the adapter trimming worked as expected. In this example, nearly all reads were trimmed. Our reads are ready to be used for variant detection.

#### 4 Detailed trim results

Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	1,304,138	1,284,224	19,914	0
Ambiguity trim	1,304,138	1,304,138	0	0
Adapter trimming	1,304,138	1,275,052	29,085	1

Figure 5: All the reads were trimmed during the Prepare Raw Data workflow.

## Running the QIAGEN GeneRead Panel Analysis workflow

The **QIAGEN GeneRead Panel Analysis** workflow is an automated way to run a series of tools using the output of one as input for the next. The first step in the ready-to-use workflow is mapping of the sequencing reads to the human reference sequence. This is followed by a local realignment step, which is included to improve the variant detection that follows directly after a primer trimming step. After variant detection, the variants are annotated with gene names, exon numbers, amino acid changes, conservation scores, information from relevant variants present in the ClinVar database, and information from common variants present in the common dbSNP, HapMap, and 1000 Genomes database. Furthermore, a detailed target regions mapping report is created that allows inspection of the coverage and mapping specificity in the target regions.

To run this workflow, go to:

**Ready-to-Use Workflows | Targeted Amplicon Sequencing** (📁) | **Somatic Cancer** (📁) | **QIAGEN GeneRead Panel Analysis** (📄)

1. Select the sequencing reads that should be analyzed, i.e., the trimmed ones you generated previously (figure 6).
2. In the next dialog, select the QIAGEN GeneRead Panels hg19 set as seen in figure 7
3. In the next two dialogs, all possible QIAGEN Generead Panels and all possible Hapmap populations are pre-selected. You can choose to deselect the panels/populations you are not interested in but we will keep them all selected in the case of this tutorial, so click **Next** twice.
4. In the "Map Reads to reference" step (figure 8), leave the read mapper configured as it is by default, i.e., the "Cost of insertions and deletions" is set to "Affine gap cost". Click **Next**.

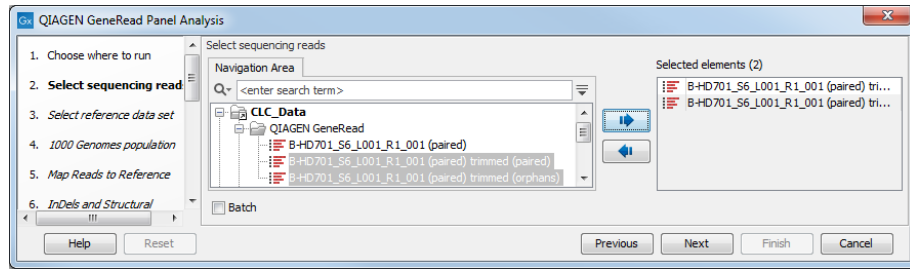


Figure 6: Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.

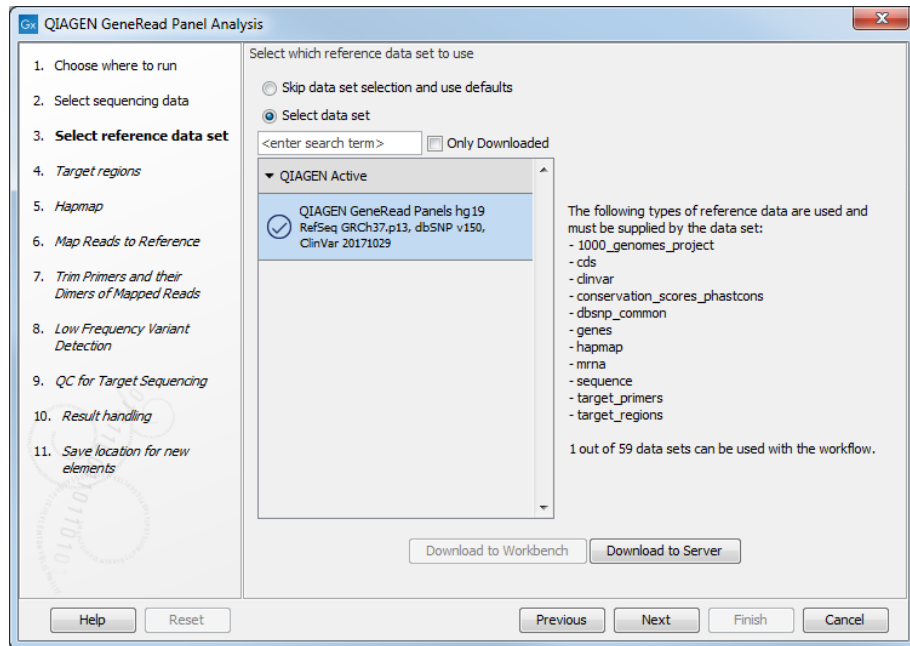


Figure 7: Specifying the 1000 Genomes populations needed for the analysis.

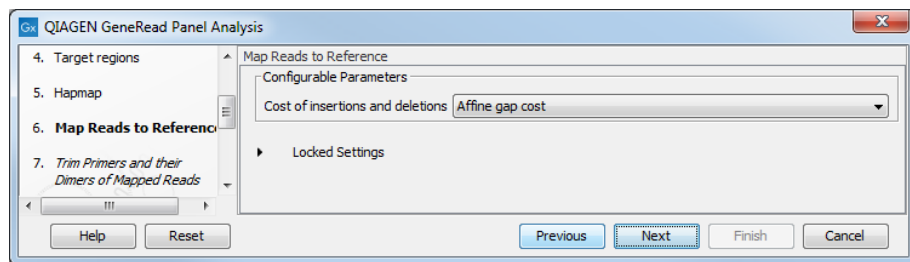


Figure 8: Leave the "Cost of insertions and deletions" to "Affine gap cost" (default).

5. In the next dialog (figure 9), the target primers for primer trimming can be specified using a drop down menu. Select again the **NGHS-201X\_Cancer\_Actionable\_Mutations** and leave the parameter "Only keep reads that have hit a primer" enabled (it is already checked by default).
6. In the next wizard step (figure 10) you can specify the parameters for variant detection. As we wish to detect very low frequency variant, set the **Minimum frequency** to **0.5%**. And to avoid having too many false positive, we will raise the **Minimum coverage** to **500**.

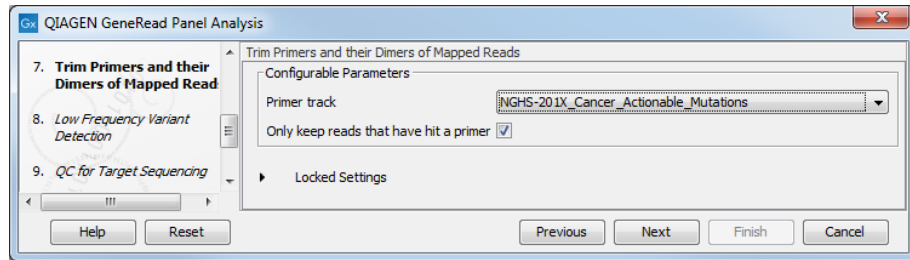


Figure 9: Select the primer track from the drop-down list.

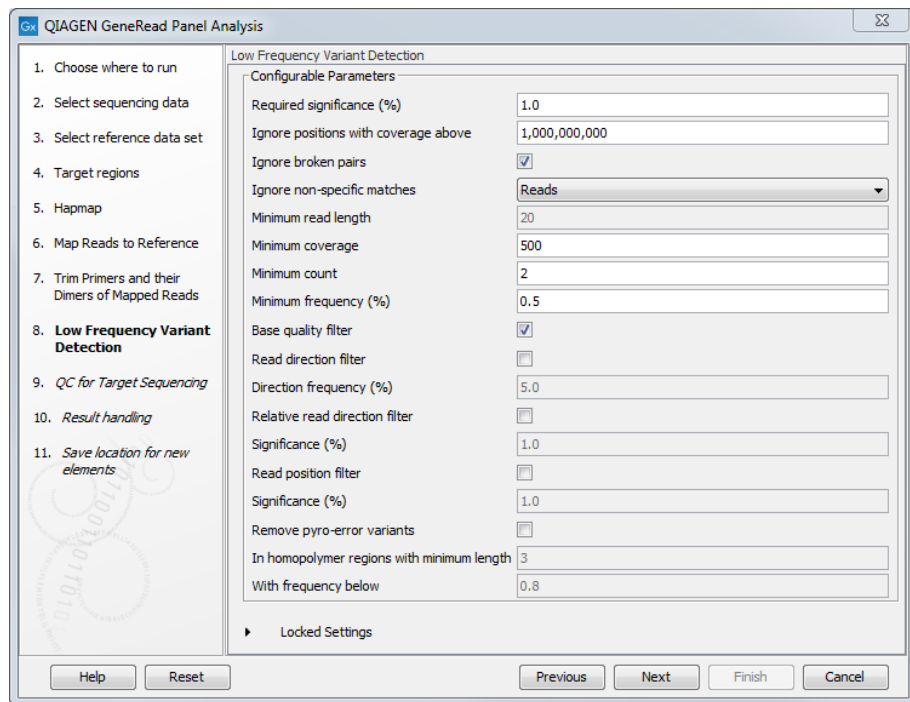


Figure 10: In this wizard step the parameters for variant detection can be adjusted.

7. In the "QC for Targeted Sequencing" step, set the Minimum coverage to 500 and leave the other parameters as default before clicking **Next**.

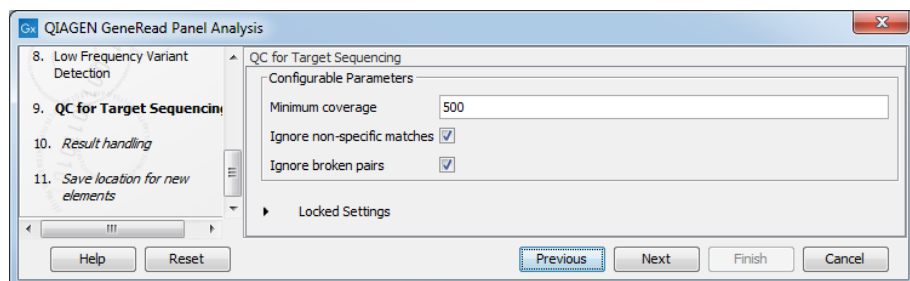



Figure 11: Specify your target regions and leave the other parameters as default.

8. Choose to save the results, create a new folder called results within your tutorial folder, and click on the button labeled **Finish**.

**Output from the QIAGEN generead panel workflow**

Open the Track List () , and focus for now on the variant table that opens automatically in split view to get an overview of the identified variants (see 12).

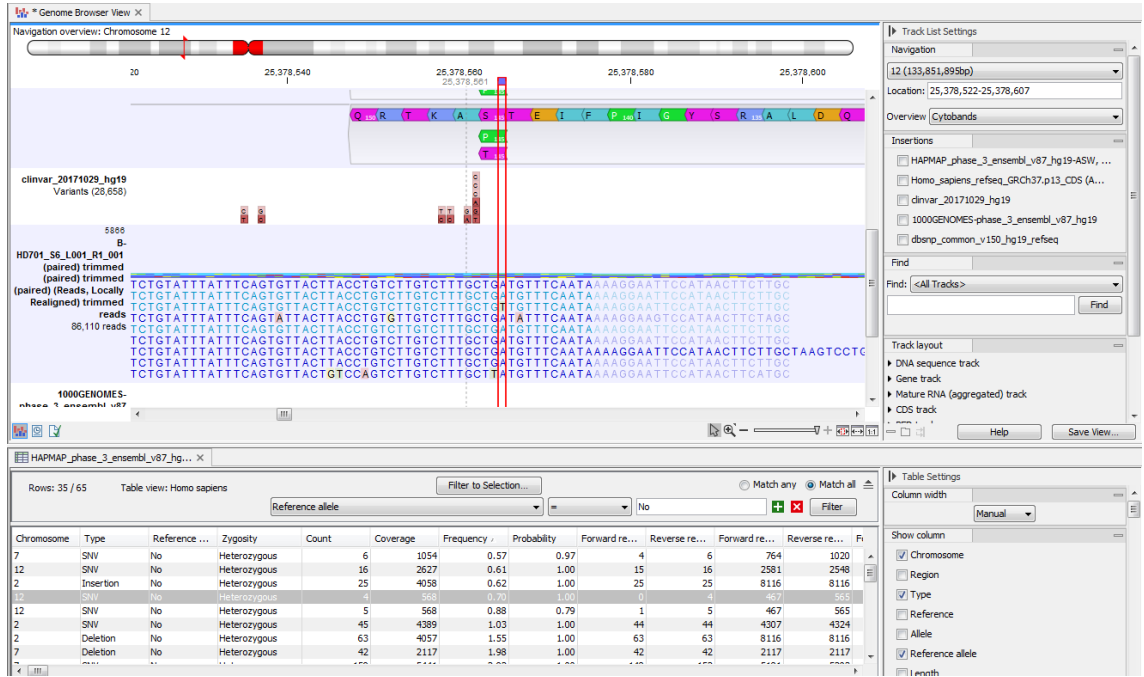


Figure 12: Filter and order the variant table to find the very low frequency variants.

In the table, filter for variants for which the allele is not identical to the reference: open the advanced filter by clicking on the arrow in the top right corner of the table. Fill in the filtering fields with "Reference allele", "=", and "No". Only 35 variants are now left in the table. To find the very low frequency variants, click on the header "Frequency". You can now easily check coverage for each of these variants by clicking on a table row. The Track List will zoom in to the chosen variant at the nucleotide level, allowing you to see the variant in the context of the mapped reads, an amino acid track, and various databases.