Tutorial

Find Very Low Frequency Variants With QIAGEN GeneRead Panels

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Find Very Low Frequency Variants With QIAGEN GeneRead Panels

This tutorial uses the capacities of the Biomedical Genomics Workbench and the QIAGEN GeneRead Panels 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.
- Download the QIAGEN GeneRead Panels hg19 Reference Data Set.
- Create an adaptor list and use the "Prepare Raw data" Ready-to-Use Workflow to remove remaining adaptors 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 (map and perform local realignment)
  - Removal of primers and primer-dimer artefacts
  - Very Low Frequency Variant Detection: 0.5 minimum allele frequency
  - Generation of additional reports (alignment report, coverage report)
  - Annotation of variants and setting up a Genome Browser View.

Prerequisites

For this tutorial, you must be working with the Biomedical Genomics Workbench 3.0 or higher. You must also have installed the QIAGEN GeneRead Panels plugin. How to install plugins is described here: http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Installing_plugins.html.

Download and import data

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.
2. Start the *Biomedical Genomics 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**.

4. 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**.

Data management configuration

In order to do this tutorial, you need to use a Reference Data Set that contains GeneRead DNAseq Gene Panel has been used for targeted sequencing.

1. To do this, go to:

   Toolbar | **Data Management** ( )

2. This will open the wizard shown in figure 1.

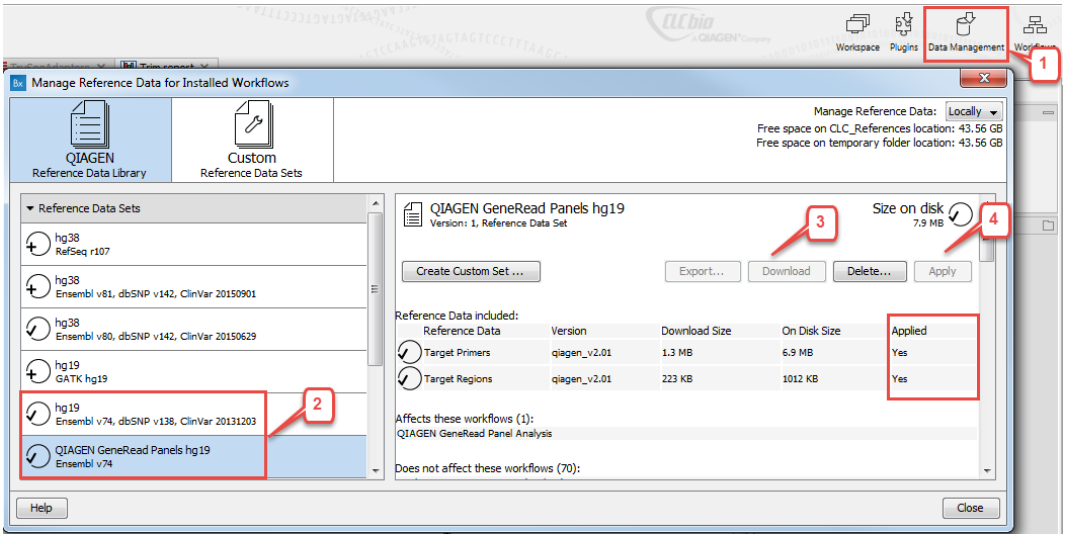

Figure 1: *Open the Data Management and download QIAGEN GeneRead Panels hg19.*

3. Select QIAGEN GeneRead Panels hg19 and click on the button labeled **Download**.
4. Once the download is complete, two extra folders are now in the CLC_References/homo_sapiens folder: "target_primers" and "target_regions" (see figure 2). Each folder contains elements specific to each commercially available QIAGEN GeneRead Panels kit.
5. Click on the button **Apply** to link the Reference Data Set to the workflow we will use later on. You can now close the Data Manager.

Prepare your reads

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

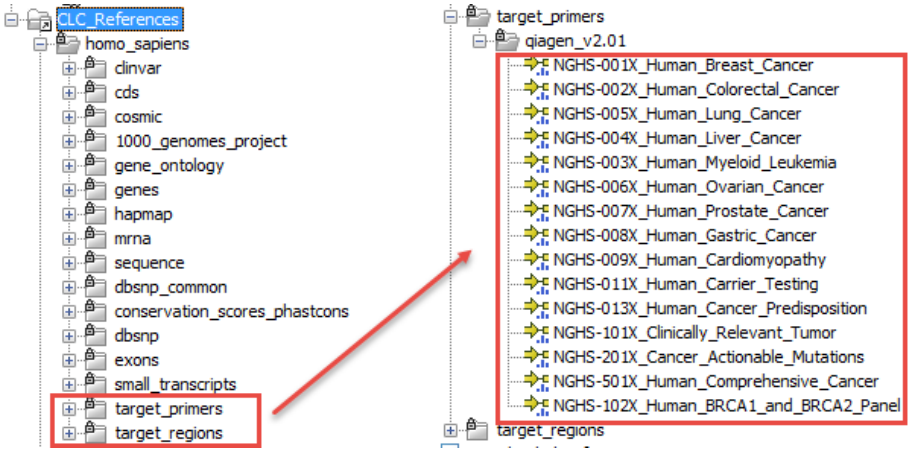

Figure 2: *The folders "target_primers" and "target_regions" are available in your CLC_References data folder.*

For this you will have to create an adaptor list using the oligonucleotide sequences disclosed in the Illumina Customer Sequence Letter found at this address: http://support.illumina.com/downloads/illumina-customer-sequence-letter.html. [1]

1. Open the workbench and go to **File** | **New** | **Trim Adapter List ( )**
2. Click on the button **Add Row** ( ) found at the bottom of the View Area in the New Trim Adapter List.
3. Paste the name and sequence of the TruSeq Universal Adapter found on page 19 of the letter. Choose to search on **All** reads, to **Remove the adapter and following sequence (3' trim)**, and to **Keep the read** found without adapters (figure 3).

   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.
4. In the Illumina letter, copy only the part of the sequence found before the index (underlined in the letter) of any TruSeq Index Adapters found on page 19 of the letter. In this way you only need to add one adapter sequence and not all the index adapters.
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**.

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

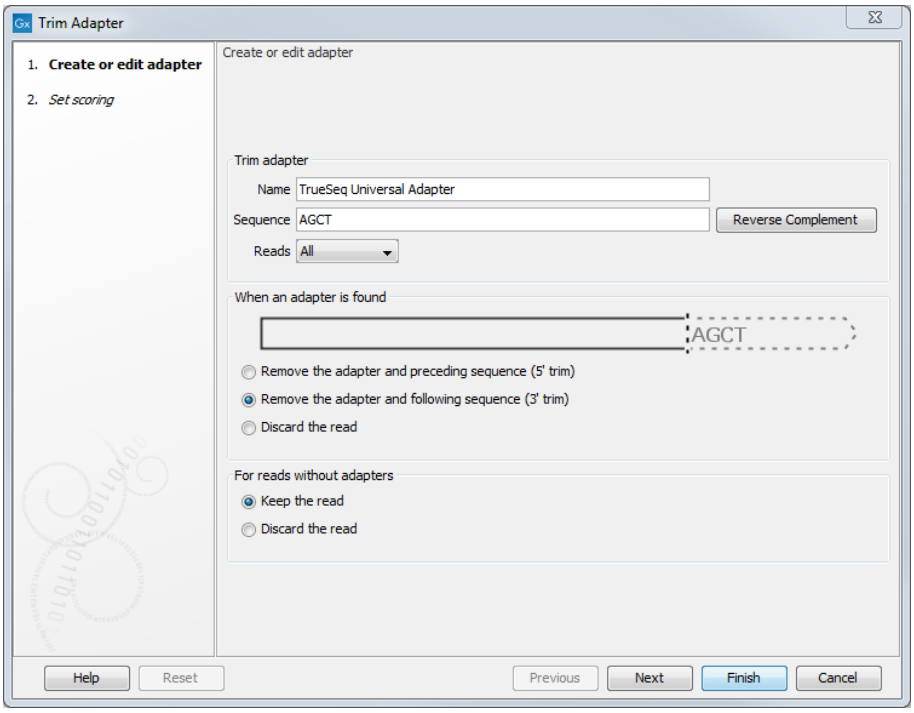

Figure 3: *Creating the adapter list. Note that the sequence written in the figure is a mock one to respect Illumina copyrights.*

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

Toolbox | **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 4.

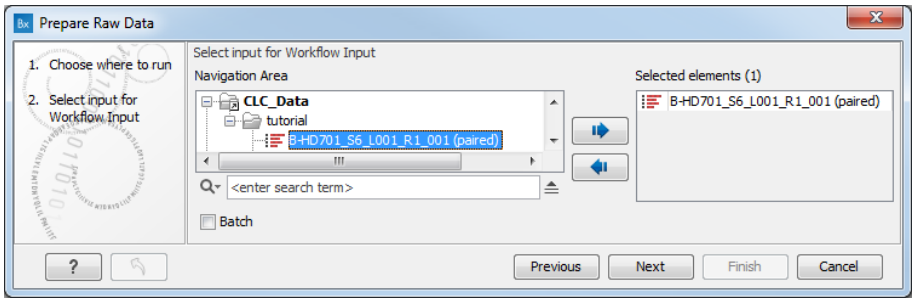

Figure 4: *Selecting the sample for extracting and counting the small RNAs.*

Click **Next** when the data is listed in the right-hand side of the dialog.

2. You are now presented with the dialog shown in figure 5. 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**.

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

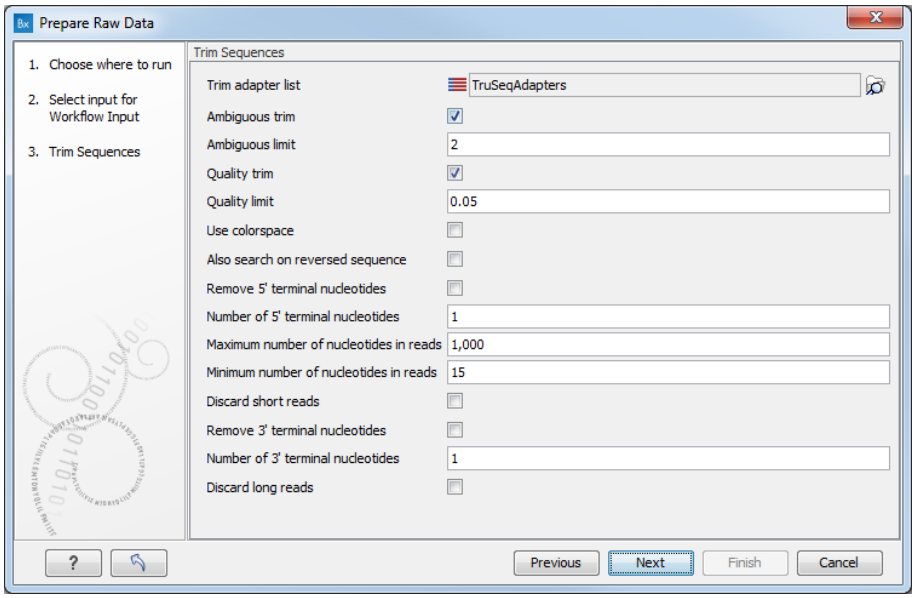

Figure 5: *Choosing to trim for adapter sequence.*

Once the analysis is complete, three reports and two trimmed sequence lists ( ) have been generated (figure 6).

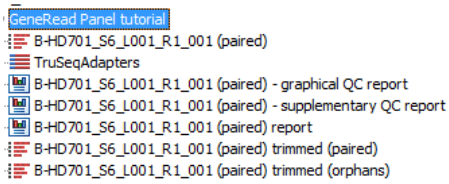

Figure 6: *Choosing to trim for adapter sequence.*

Open the Trim report (called B-HD701_S6_L001_R1_001 (paired) report) (figure 7). It is meant to be used as a quality check, mainly to see that the adapter trimming worked as expected. In this example, it shows that 99.98% of the 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	760,730	543,192	216

Figure 7: *All the reads were trimmed during the Prepare Raw Data workflow.*

Running the Ready-to-Use 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:

Toolbox | **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 8).

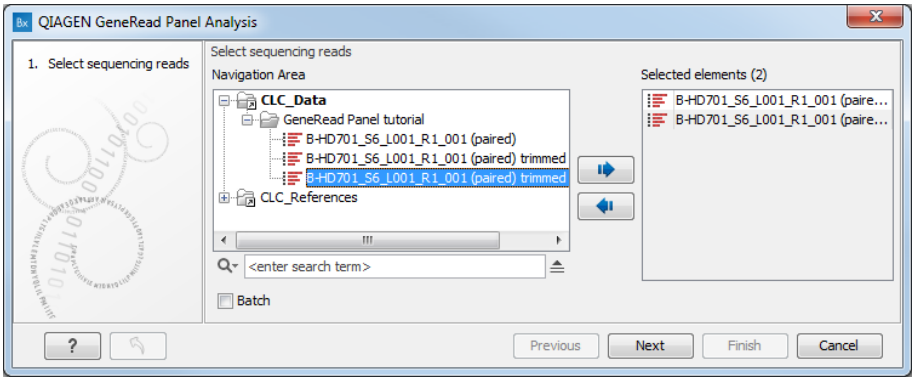

Figure 8: *Select the sequencing reads by double-clicking on the file name or by clicking once on the file name and then on the arrow pointing to the right hand side.*

2. In the next wizard (figure 9), you can specify which of the available 1000 Genomes populations to use in the analysis by clicking on the browse symbol ( ) in the right-hand side of the wizard. We choose to keep them all in this analysis, but were you to work on a specific population, you could remove the others simply by double clicking on their names in this wizard window. Click **Next**.

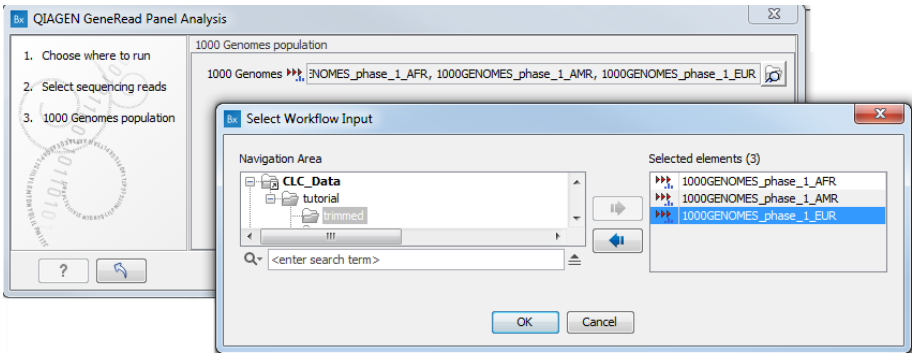

Figure 9: *Specifying the 1000 Genomes populations needed for the analysis.*

3. In the "Map Reads to reference" step (figure 10), 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**.

4. In this wizard window (figure 11), you can restrict the calling of InDels and Structural Variants to the targeted regions of your experiment by clicking on the plus symbol ( ) in the right-hand side of the wizard. In this case, the kit used to prepare the amplicons was the **Cancer Actionable Mutations**. Use the arrows to deselect all others before clicking Done. Then click **Next** to move on the the next wizard window.

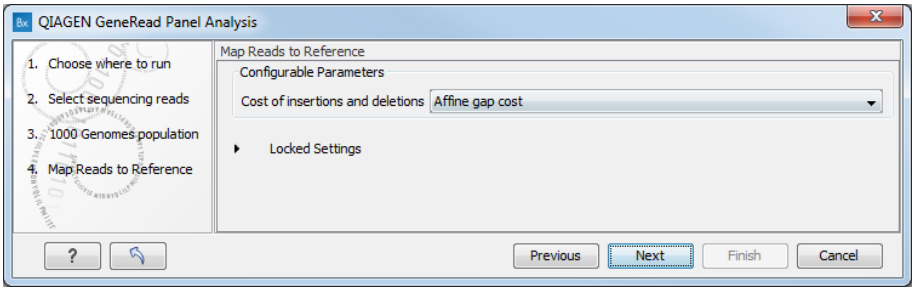

Figure 10: *Leave the "Cost of insertions and deletions" to "Affine gap cost" (default).*

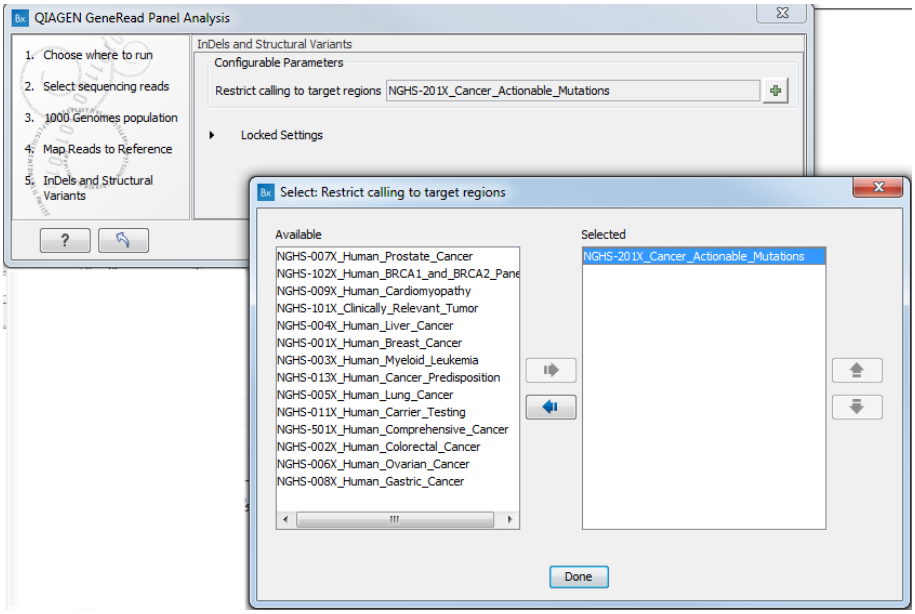

Figure 11: *In this wizard step you can specify the targeted regions matching your read mapping.*

5. In the next dialog (figure 12), 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). Click on the button labeled **Next**.

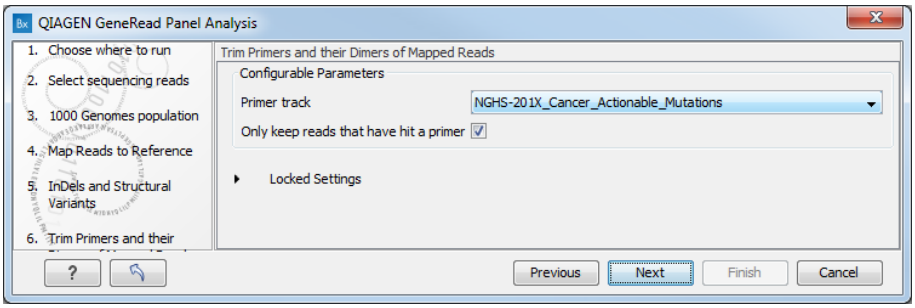

Figure 12: *Select the primer track from the drop-down list.*

6. In the next wizard step (figure 13) you can specify the parameters for variant detection. First you must specify your target region: in the same way you did in the previous step, deselect all but **NGHS201X_Cancer_Actionable_Mutations** (figure 14). 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**. Click **Next**.

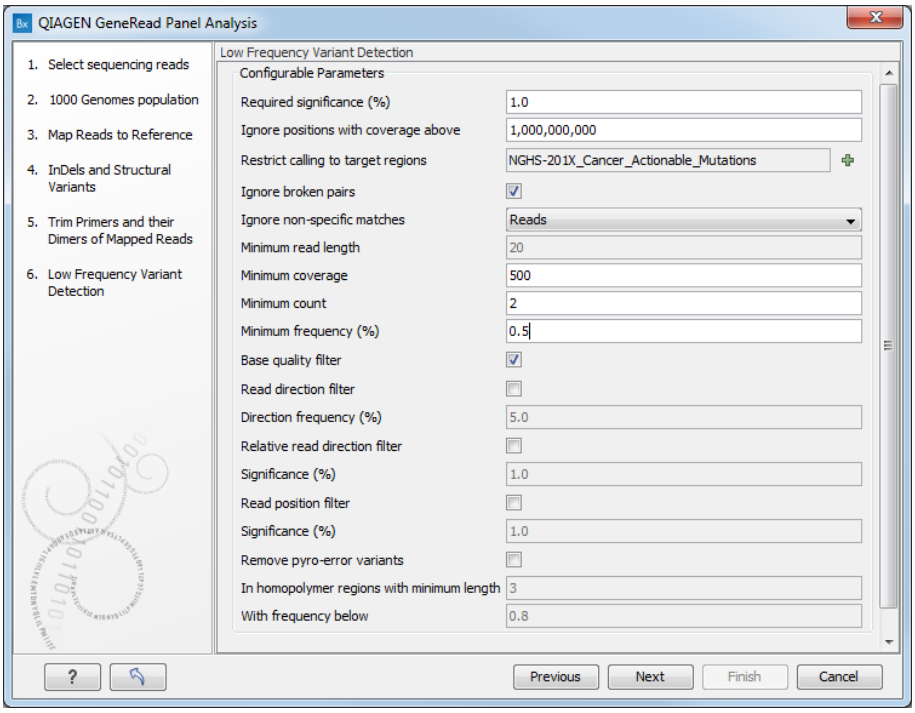

Figure 13: *In this wizard step the parameters for variant detection can be adjusted.*

7. In the "QC for Target Sequencing" step, you must specify your target region. Again, choose to keep only **NGHS201X_Cancer_Actionable_Mutations** (figure 14). Set the Minimum coverage to 500 and leave the other parameters as default before clicking **Next**.

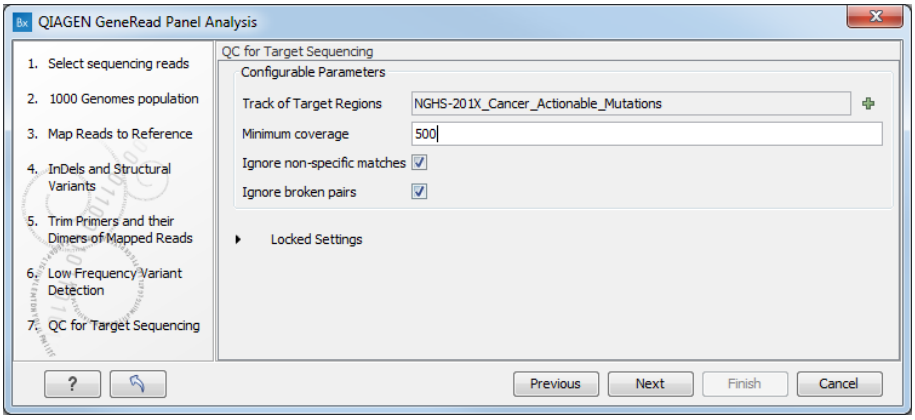

Figure 14: *Specify your target regions and leave the other parameters as default.*

8. The next two wizard steps deals with annotating the variants using the HapMap and the 1000 Genomes Project databases. We will leave all possible options as they are already selected by default for both databases (figure 15). Click on the button labeled **Next** to go to the last wizard step.

9. Choose to save the results, create a new folder called results within your tutorial folder,

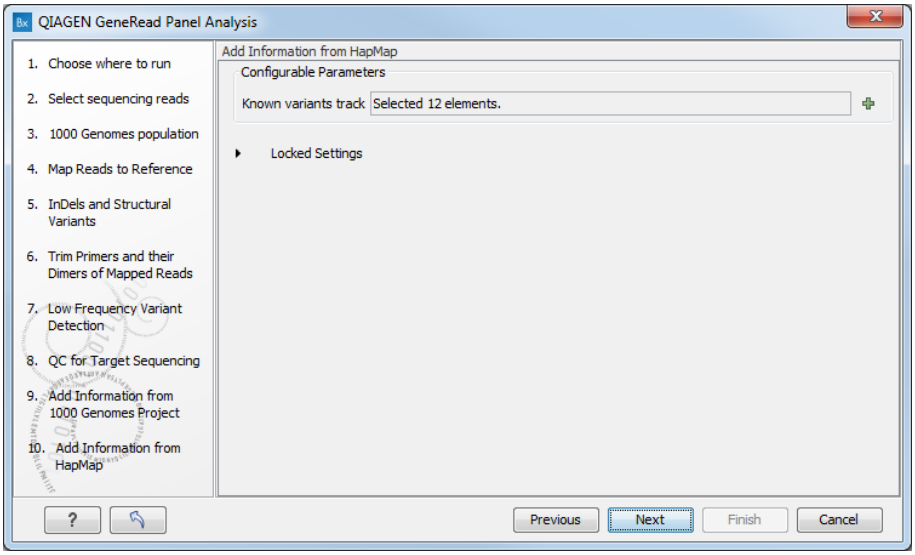

Figure 15: *Leave all populations that have already been selected for 1000 Genomes Project and Hapmap databases.*

and click on the button labeled **Finish**.

Note that some steps in this workflow can be preset using the Data Manager: if you intend to work with always the same panel and samples coming from specific populations, you can create a custom reference data set containing only the relevant panel and populations needed for analysis. Once you apply this custom Reference Data Set, some steps of the workflow will be be skipped. Creating a custom Reference Data Set is described in the manual here: http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.php?manual=Create_custom_Reference_Data_Set.html

Output from the QIAGEN generead panel workflow

Open the Genome Browser View file ( ), and focus for now on the variant table that opens automatically in split view to get an overview of the identified variants (see 16).

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 38 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 Genome Browser 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.

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency	Probability
5	180055897	Deletion	G	-	No	1	Heterozygous	53	10197	0.52	1.00
7	55241605	Deletion	C	-	No	1	Heterozygous	25	4712	0.53	1.00
7	55249071	SNV	C	T	No	1	Heterozygous	6	1022	0.59	0.96
2	209113048^209113049	Insertion	-	AA	No	2	Heterozygous	24	4002	0.60	1.00
12	25398271	SNV	C	T	No	1	Heterozygous	16	2484	0.64	1.00
12	25378565	SNV	A	T	No	1	Heterozygous	4	566	0.71	1.00
12	25378565	SNV	A	G	No	1	Heterozygous	5	566	0.88	0.80
2	209113058	SNV	A	G	No	1	Heterozygous	40	4276	0.94	1.00
15	39313413	Deletion	T	-	No	1	Heterozygous	11	1153	0.95	1.00
22	41558756^41558757	Insertion	-	A	No	1	Heterozygous	16	1441	1.11	1.00
2	209113049..209113050	Deletion	AA	-	No	2	Heterozygous	59	3998	1.48	1.00
7	55242465..55242479	Deletion	GGAATTAAGAGAAGC	-	No	15	Heterozygous	42	2117	1.98	1.00
22	41558794^41558795	Insertion	-	C	No	1	Heterozygous	31	1439	2.15	1.00
7	55259515	SNV	T	G	No	1	Heterozygous	152	5212	2.92	1.00
2	209113048^209113049	Insertion	-	A	No	1	Heterozygous	186	4002	4.65	1.00
1	160786670	SNV	A	G	No	1	Heterozygous	70	1132	6.18	1.00
12	25398284	SNV	C	T	No	1	Heterozygous	154	2487	6.19	1.00
4	46329655	SNV	A	T	No	1	Heterozygous	1841	27429	6.71	1.00
2	209113192	SNV	G	A	No	1	Heterozygous	196	2509	7.81	1.00

Figure 16: *Filter and order the variant table to find the very low frequency variants.*