Tutorial

Resequencing: Map Reads to Reference and Variant Detection

September 15, 2016

Sample to Insight

CLC bio, a QIAGEN Company · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.clcbio.com · support-clcbio@qiagen.com

Resequencing: Map Reads to Reference and Variant Detection

This tutorial takes you through some of the tools for analyzing a typical resequencing data set from a high-throughput sequencing machine. As an example we use an *E. coli* data set consisting of just over 400,000 reads from a 454 sequencer.

Importing the data

1. Download the data set from our web site:

   `http://download.clcbio.com/testdata/raw_data/454.zip`.

2. Unzip the file somewhere on your computer (on the Desktop for example).

3. Start the *CLC Genomics Workbench* if you have not already.

4. Import the data:

   File | **Import** () | **Roche 454** ()

   This will bring up the dialog shown in figure 1

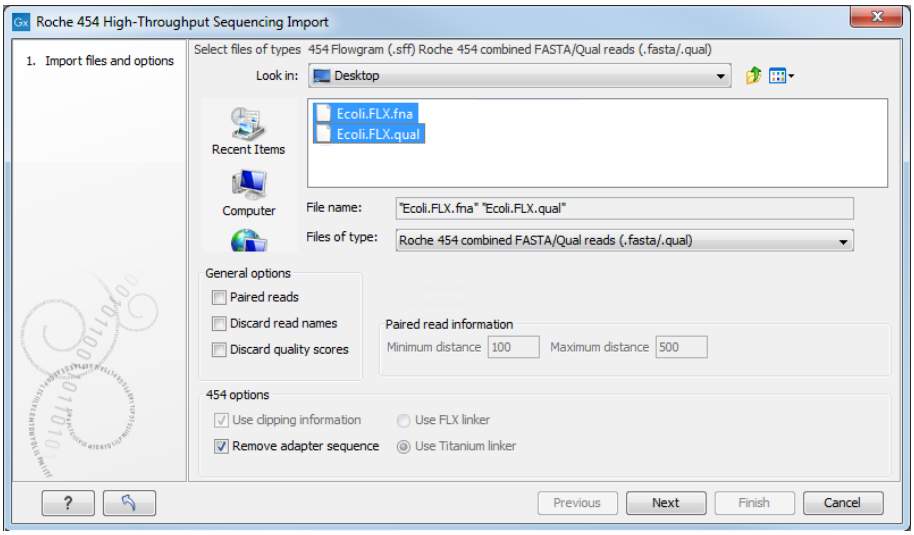


Figure 1: *Choosing what kind of data you wish to import.*

5. Select the `Ecoli.FLX.fna` and `Ecoli.FLX.qual` files that come from the downloaded zip file.

   Make sure to select the **Format** `Roche 454 combined FASTA/Qual reads (.fasta/.qual)` and the **Remove adapter sequence** checkbox is checked and that the **Paired reads** checkbox is NOT checked. The option to discard read names is not significant in this context because of the relatively small amount of reads. Click on the button labeled **Next**.

6. Choose to **Save** the data and click on the button labeled **Next**. Specify location and click on the button labeled **Finish**.

   After a short while, the reads have been imported.

   Next, import the reference genome sequence also included in the zip file. To do this:

7. Go to:

   File | **Import** () | **Standard Import**

8. Select **Locate "NC_010473.gbk"** and use **Option** `Automatic import`. Click on the button labeled **Next**.
9. Choose a location to save the data and click on the button labeled **Finished**.

We used a special import tool for the 454 data because next-generation sequencing data involves specialised information, and can involve the concerted import of more than one file. For example here, we imported two files, one with the sequence information (Ecoli.FLX.fna) and one with the quality information (Ecoli.FLX.qual). The file NC_010473.gbk is in standard GenBank format.

Mapping the reads to the E. coli reference

The data we use in this tutorial does not have any adapters, so we do not need to run a trimming step. Thus we proceed directly to mapping the reads to the reference.

During this process, you will have the option to generate a standard read mapping object or a read mapping track as output. Here, we generate a reads track object as output.

1. To begin the mapping:

   Toolbox | **NGS Core Tools (**) | **Map Reads to Reference (**)

2. This shows the dialog in figure 2.

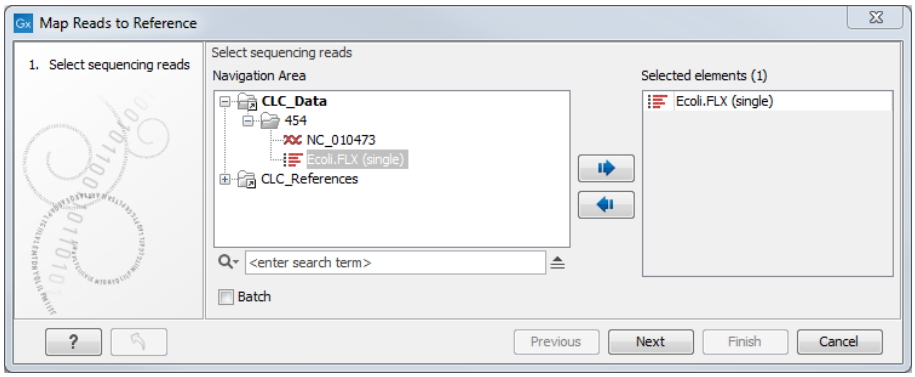


Figure 2: *Select sequence list containing the reads. The reference sequence will be selected in the next step.*

3. Select the `Ecoli.FLX (single)` ( ) sequence list and add it to the panel to the right. Click on the button labeled **Next**.
4. Click on the browse icon ( ) within the section labeled **References**.
5. Select the reference sequence `NC\_010473`. After this, you should see what is shown in figure 3.

   While we only have a single reference sequence object in this example, you are actually able to select single or multiple sequence objects or sequence lists as references.

   Click on the button labeled **Next**.
6. We will use the default mapping parameters as shown in figure 4. If the parameters shown in your Wizard window do not match those in the figure, just click on the parameter reload ( ) button to reset them and click on the button labeled **Next**.

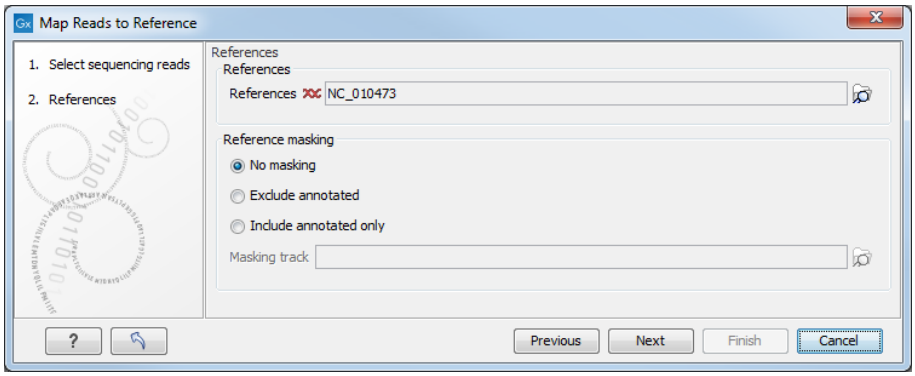


Figure 3: *Specifying the reference sequence(s) to use and the masking to apply, if desired.*

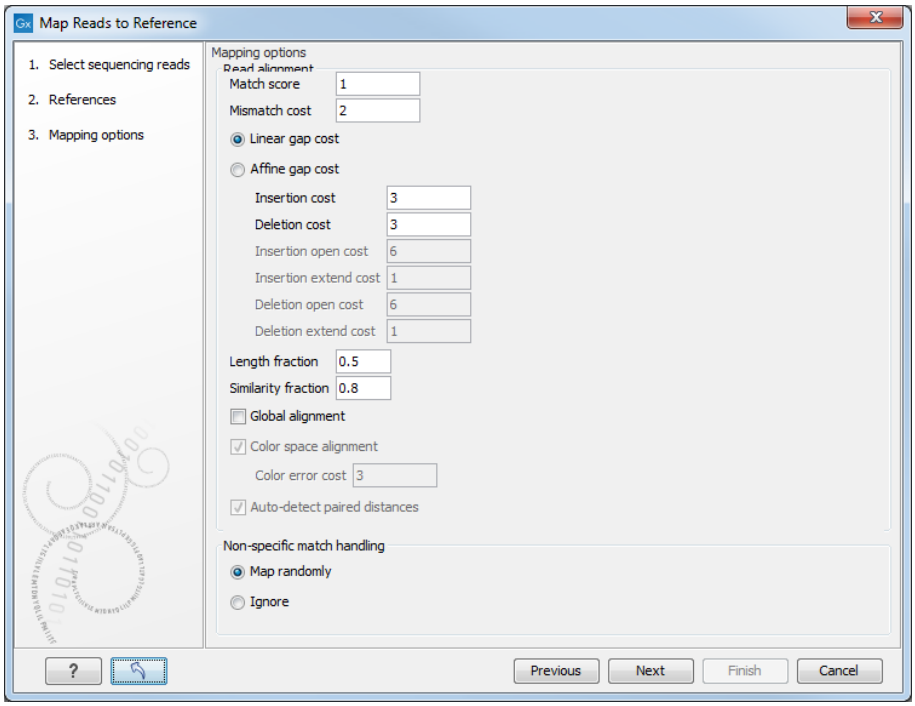


Figure 4: *Set the mapping parameters. Clicking on the parameter reload button (in the lower left corner) resets all parameters to the defaults. Click on the button with the question mark brings up the in-built help, where you can find out more about running mappings via the Workbench.*

7. Now you choose what type of mapping output you wish to create (figure 5).

   - Click in the radio button beside **Create reads track**.
   - Click in the box beside **Create report** so there is a check mark in it.
   - Click in the box beside **Collect unmapped reads** so there is a check mark in it.
   - Choose to **Save** the results.

   Click on the button labeled **Next**.

8. Specify a location to **Save** the results and click **Finish**.

You can follow the mapping progress both in the status bar or under the tab **Processes** at the bottom left corner.

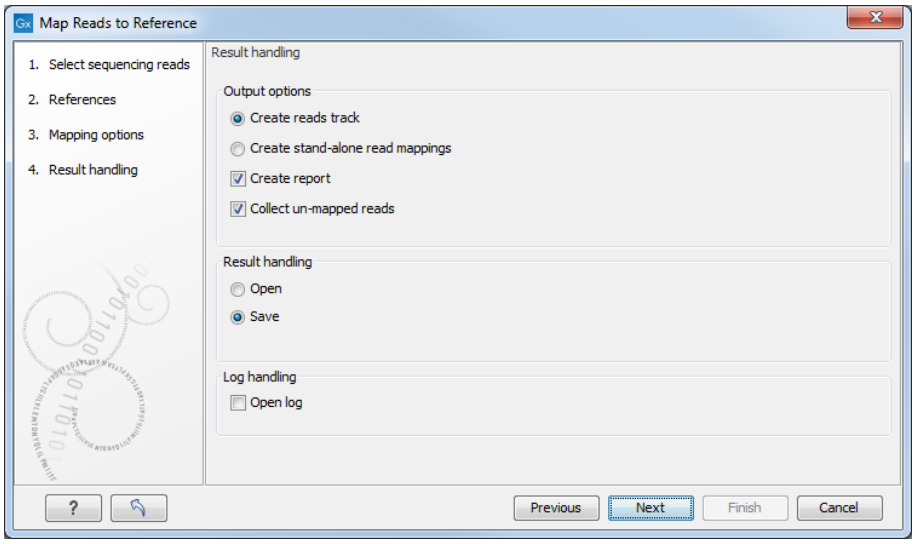


Figure 5: *Output options for read mappings.*

Interpreting the read mapping results

When the process is done, you will see the following results within your **Navigation Area**.

List of non-mapped reads ( ) These are the reads that did not match the reference sequence. You can use this list to investigate contamination in the sample or structural differences between the sequencing data and the reference sequence. Typically you will do a *de novo* assembly of these reads and then use BLAST to investigate the contigs (there is a separate tutorial showing how to do this).

Report ( ) The report shows information about the mapping. Most importantly, it shows the number of reads that matched the reference sequence.

Mapping ( ) The mapping itself shows the alignment of all the reads to the reference.

In this section, we look more closely at the mapping results.

Double click the mapping object (reads track ( )) just created. This will open it up in the viewing area of the Workbench.

By default, the mapping will open up in aggregated data view corresponding to the entire length of the reference, see figure 6.

The threshold (in bp) for when data should be aggregated can be specified with the drop-down box in the top section of the **Track Settings** pane on the right hand side of the Workbench. The threshold describes the unit (or "bucket") size in base pairs, above which the data will start being aggregated. The bucket size depends on the track length and the zoom level. Please take some time now to investigate the effects of choosing the different options available. Notice especially the effect of choosing **Show quality scores**, which allows visualisation of the quality scores at the base level.

It is possible to include the reference track in a Track list.

1. First, the reference sequence NC_010473 must be converted to tracks in beforehand by
   Toolbox | **Track Tools ( )** | **Convert to Tracks ( )**

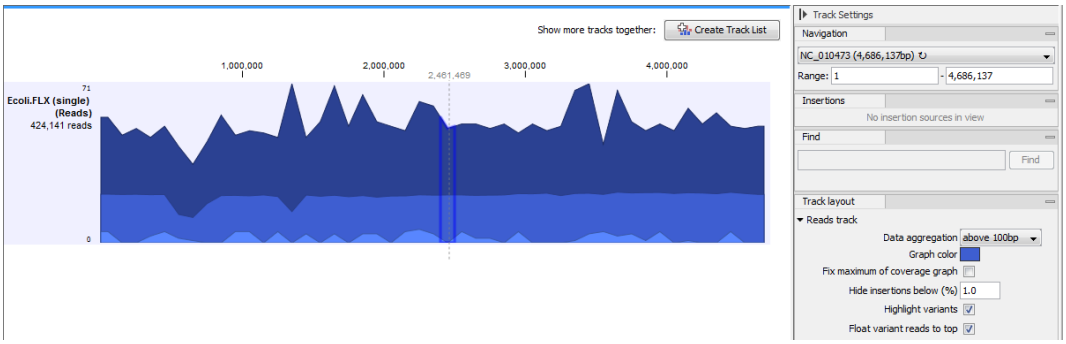


Figure 6: *Default view of a mapping object when opened in the Workbench. To add reference track, click on Create Track List button at the top right of the view area*

2. During conversion it is possible to specify which additional annotations tracks (figure 7) should be generated. In this case we choose the types specified by default in the wizard: CDS and Gene.

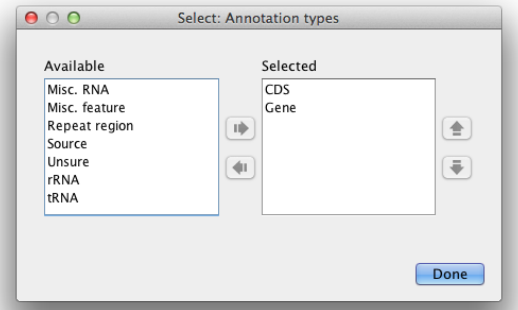


Figure 7: *During conversion of reference sequence, specification of annotation tracks as output files is needed.*

3. Save the new tracks in the Navigation Area and click on Finish.
4. You can now click on the **Create Track list** button at the top right corner of the viewing area (figure 6) or from the Toolbox.
5. Specify which tracks should be included in the list as in figure 8.

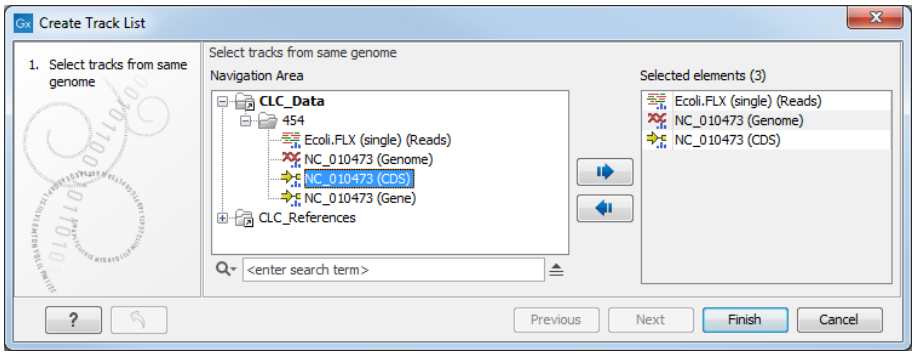


Figure 8: *Specification of tracks to be included in the Track list.*

6. Choose to **Open** the Track list and click on **Finish**.

Once the Track list view is open, the order of included tracks can be rearranged by drag-and-drop. Here, the reference track has been moved to the top, followed by the CDS annotation track, and finally the mapping at the bottom (figure 9).

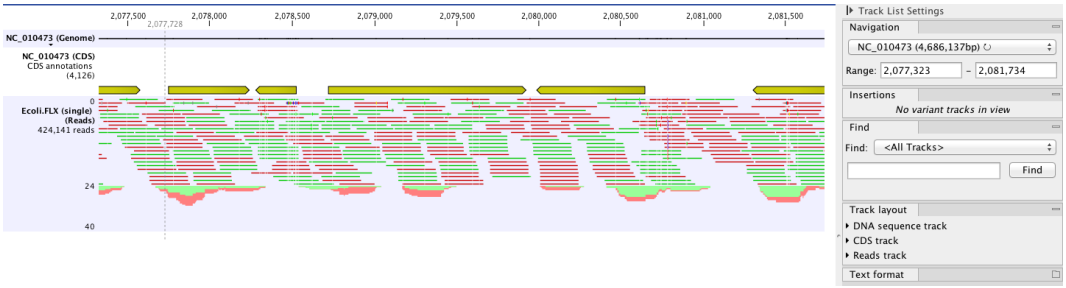


Figure 9: *Track list including reference, CDS annotation track and mapped reads. The read colors are green (forward) and red (reverse) by default.*

Take some time now to use the buttons in the bottom right hand side of the Workbench to zoom in and out, and to see the full mapping in a single view (( )), to see the mapping in a selected region (( )) or to see the mapping again in full detail (( )). When you are done, zoom out fully, by clicking on the ( ) button.

Using the **Range** of the **Navigation** pane of the **Track List Settings**, you can select regions of interest by specifying the bases (see figure 10). Try selecting the region form base 1846000 to base 1846280. The viewer now zooms to the specified region including a yellow CDS annotation. By mouse-over, annotation information e.g. the gene name (lpp) is shown.

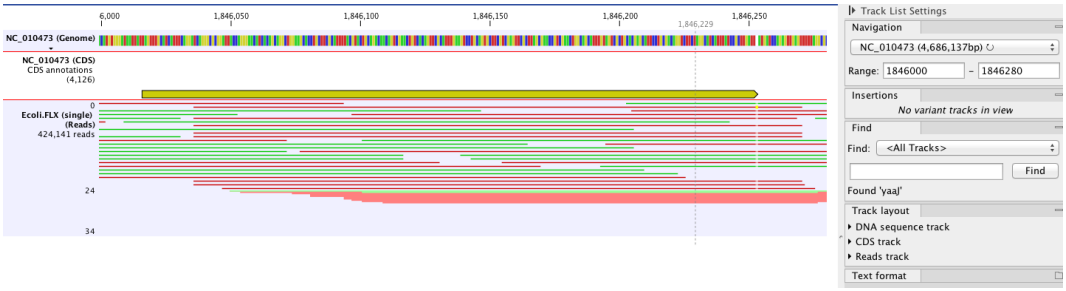


Figure 10: *Jump to a particular location in an object using the Range functionality of the view settings pane.*

The **Find** section of the **Track List Settings** pane allows you to look for subregions within the specified range, based on annotation information or name. Try to find the lpp gene by entering **lpp** in the search box and clicking on the button labeled **Find**. Now the selection base region is narrowed the actual CDS region (see figure 11.

Variant detection

There are three tools you can use to find areas where there is evidence to suggest that there are differences in your sample data compared to the reference sequence. These are called **Basic Variant Detection**, **Fixed Ploidy Variant Detection** and **Low Frequency Variant Detection**. For information on how these tools work, please refer to the manual:

http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Variant_Detectors_overview.html.

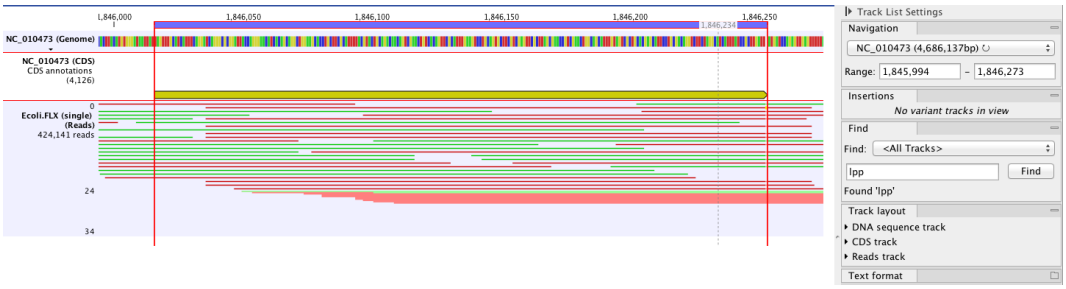


Figure 11: *Selection of the sub-region representing the CDS region of the lpp gene using the Find functionality of the view settings pane.*

The variant detection tools report on positions where there may be a Single Nucleotide Variations (SNV), Insertion, Deletion, or Replacement.

Here we will run the Fixed Ploidy Variant Detection tool, and then view the results as a table linked with the mapping.

1. To run the Fixed Ploidy Variant Detection tool, go to:

   Toolbox | **Resequencing Analysis (**  **)** | **Variant Detectors (**  **)** | **Fixed Ploidy Variant Detection (****)**

2. In the Wizard window that opens, select the mapping result object () **Ecoli.FLX (single) (Reads)** you created earlier. Click on the button labeled **Next**.

3. The default parameter values in this window have a **Ploidy** of 2, and the **Required variant probability** of 90%. Set the Ploidy to **1** as E.coli has a single chromosome (figure 12). For more information about the settings, feel free to click the **Help** (?) button. Click on the button labeled **Next**.

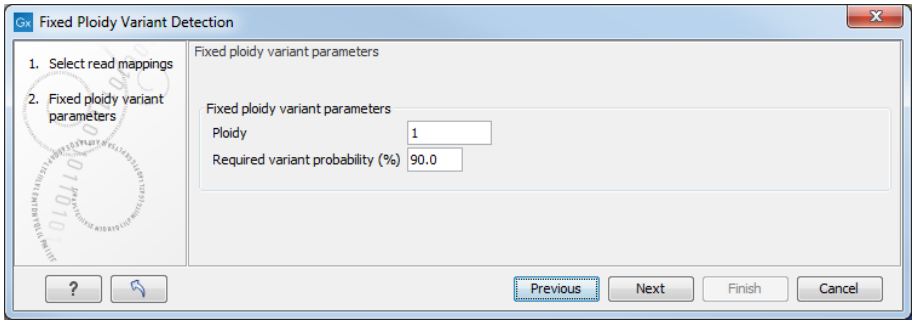


Figure 12: *Setting variation detection parameters.*

4. In the General filters window, set most parameters to their default values. If you are not sure if the settings you have are the defaults, just click the bottom left button with the () icon to reset to the defaults (see figure 13). Change however the Minimum coverage to 8 and the Minimum frequency 15% before clicking **Next**.

   Note that to detect true variants, you do need to ensure that the settings you choose are relevant for your dataset and for your study. If the **Minimum coverage** is set to 50 but you have a mapping with an average coverage of 15, a lot of potential SNPs will not be reported. The **Minimum variant frequency** needs to be adjusted when working with mixed samples or non-haploid organisms. For diploids, it should be set below 50 % in order to report heterozygote SNPs.

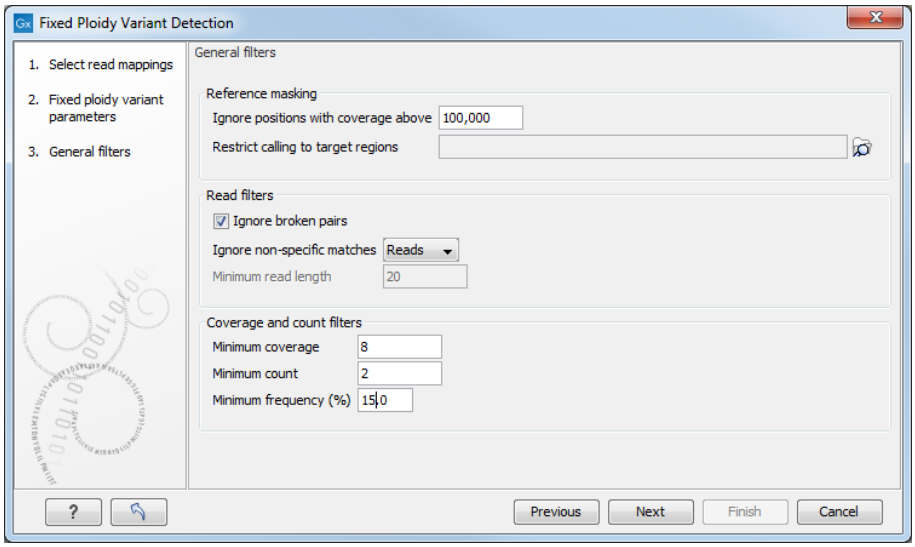


Figure 13: *Setting general filtering parameters.*

5. In the Noise filters window, reset all parameters to their default values with the Reset button ( ) (see figure 14). Click on the button **Next**.

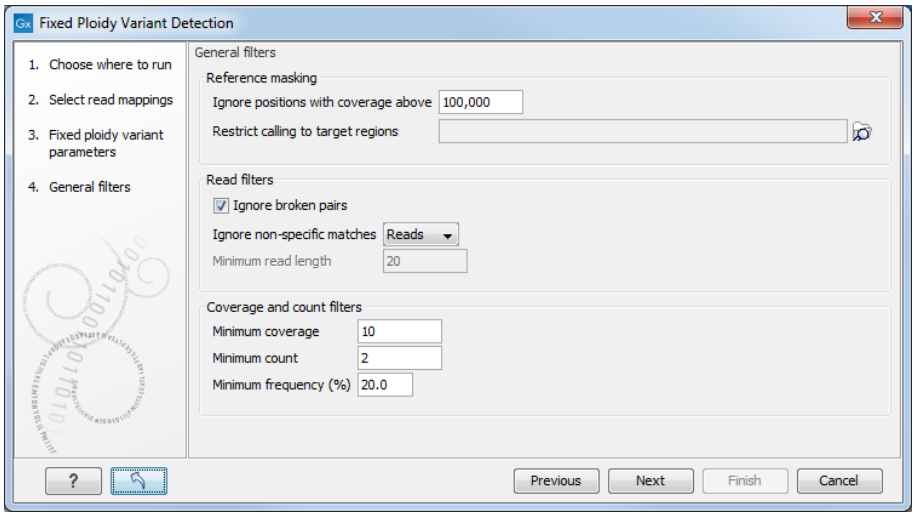
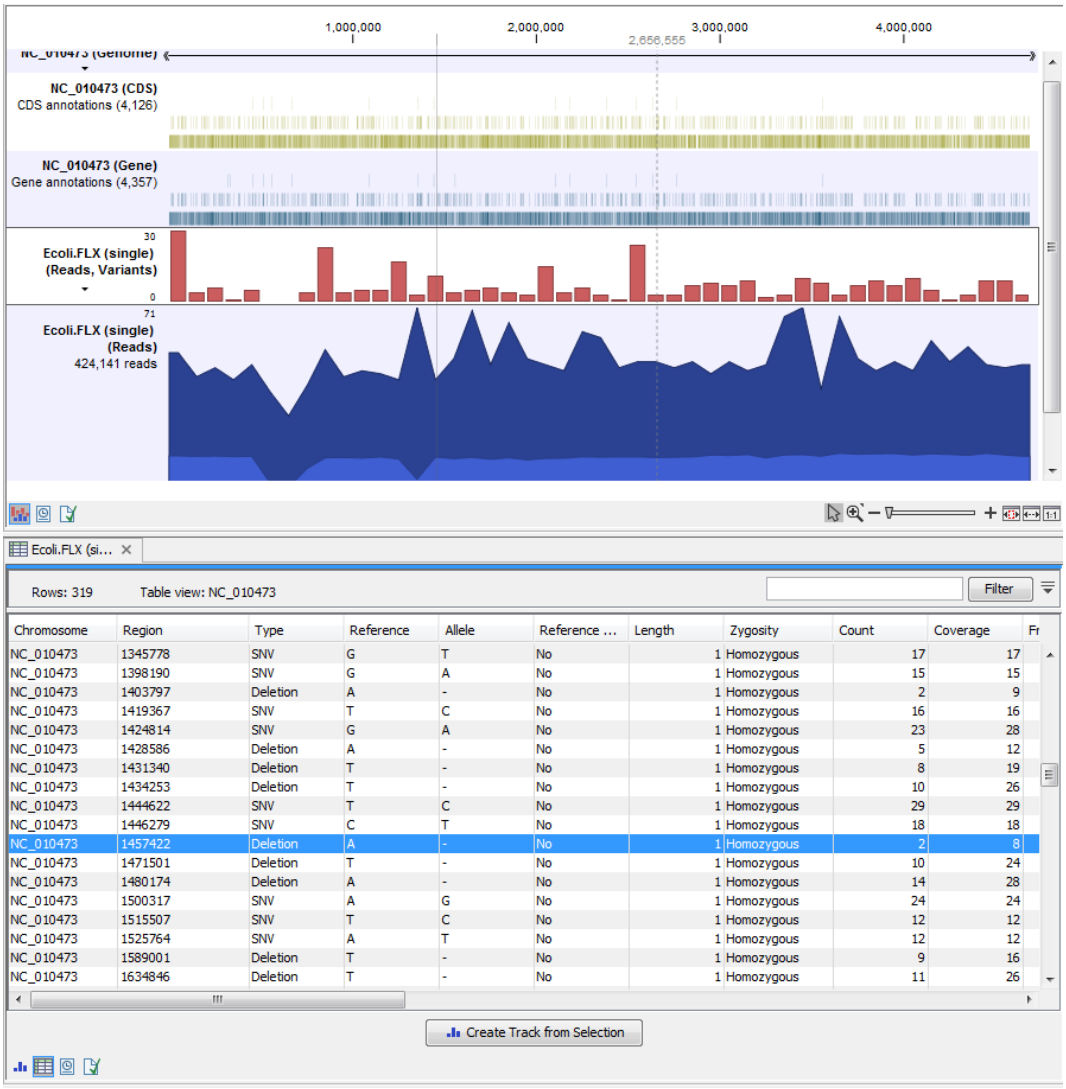


Figure 14: *Setting noise filtering parameters.*

6. In the Result handling window, we choose the type of output to generate. As we are working with track-based objects in this tutorial, please click in the boxes next to **Create track** and **Create report**, and ensure the box next to **Create annotated table** is unchecked. Finally choose to **Save** your results, specify the location you want them to be saved and click on the button labeled **Finish**.

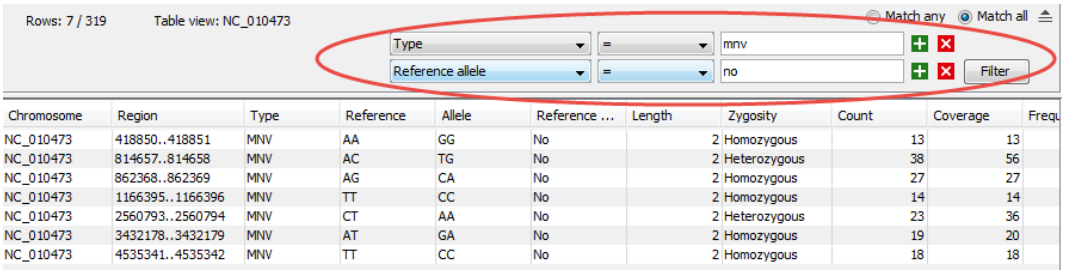
You can check the progress of the detection analysis in the Processes tab of the toolbox. When it is completed, add the generated Variant track **Ecoli.FLX (single) (Reads, Variants)** to your Track List by drag-and-drop or by right clicking in the view area, and choosing the option **Include More Tracks**. To view variant data in table view, double click on the Variant Track. The Variant Table now linked to the Mapping Track (see figure 15). This means that if you click in the row of the Variant Table, the cursor will jump to the point in the mapping where that variant has been called. You may need to use the Zoom buttons to view the variant in detail.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Fr
NC_010473	1345778	SNV	G	T	No	1	Homozygous	17	17	
NC_010473	1398190	SNV	G	A	No	1	Homozygous	15	15	
NC_010473	1403797	Deletion	A	-	No	1	Homozygous	2	9	
NC_010473	1419367	SNV	T	C	No	1	Homozygous	16	16	
NC_010473	1424814	SNV	G	A	No	1	Homozygous	23	28	
NC_010473	1428586	Deletion	A	-	No	1	Homozygous	5	12	
NC_010473	1431340	Deletion	T	-	No	1	Homozygous	8	19	
NC_010473	1434253	Deletion	T	-	No	1	Homozygous	10	26	
NC_010473	1444622	SNV	T	C	No	1	Homozygous	29	29	
NC_010473	1446279	SNV	C	T	No	1	Homozygous	18	18	
NC_010473	1457422	Deletion	A	-	No	1	Homozygous	2	8	
NC_010473	1471501	Deletion	T	-	No	1	Homozygous	10	24	
NC_010473	1480174	Deletion	A	-	No	1	Homozygous	14	28	
NC_010473	1500317	SNV	A	G	No	1	Homozygous	24	24	
NC_010473	1515507	SNV	T	C	No	1	Homozygous	12	12	
NC_010473	1525764	SNV	A	T	No	1	Homozygous	12	12	
NC_010473	1589001	Deletion	T	-	No	1	Homozygous	9	16	
NC_010473	1634846	Deletion	T	-	No	1	Homozygous	11	26	

Figure 15: *Track list with its associated variant table.*

All tables in the Workbench can be filtered. In this section, we use the filter tools to help concentrate of multiple nucleotide variant different from the reference. To do this, set up filtering on your table, as shown in figure 16.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequ
NC_010473	418850..418851	MNV	AA	GG	No	2	Homozygous	13	13	
NC_010473	814657..814658	MNV	AC	TG	No	2	Heterozygous	38	56	
NC_010473	862368..862369	MNV	AG	CA	No	2	Homozygous	27	27	
NC_010473	1166395..1166396	MNV	TT	CC	No	2	Homozygous	14	14	
NC_010473	2560793..2560794	MNV	CT	AA	No	2	Heterozygous	23	36	
NC_010473	3432178..3432179	MNV	AT	GA	No	2	Homozygous	19	20	
NC_010473	4535341..4535342	MNV	TT	CC	No	2	Homozygous	18	18	

Figure 16: *Filtering for single nucleotide variants where an amino acid change would result.*

In the mapping, zoom in to the nucleotide level with this button (1:1), and in the variant table, double click on some of the remaining 7 rows to hop to the locations in the mapping track where the multiple nucleotide variants are. Using the **Reads Track List Settings** layout in the **Track List Settings** pane on the right hand side of the Track list view, you can choose to **Show quality**

scores as color coding of your nucleotides as seen in figure 17. Click on the color legend to choose the color shades you want to use.

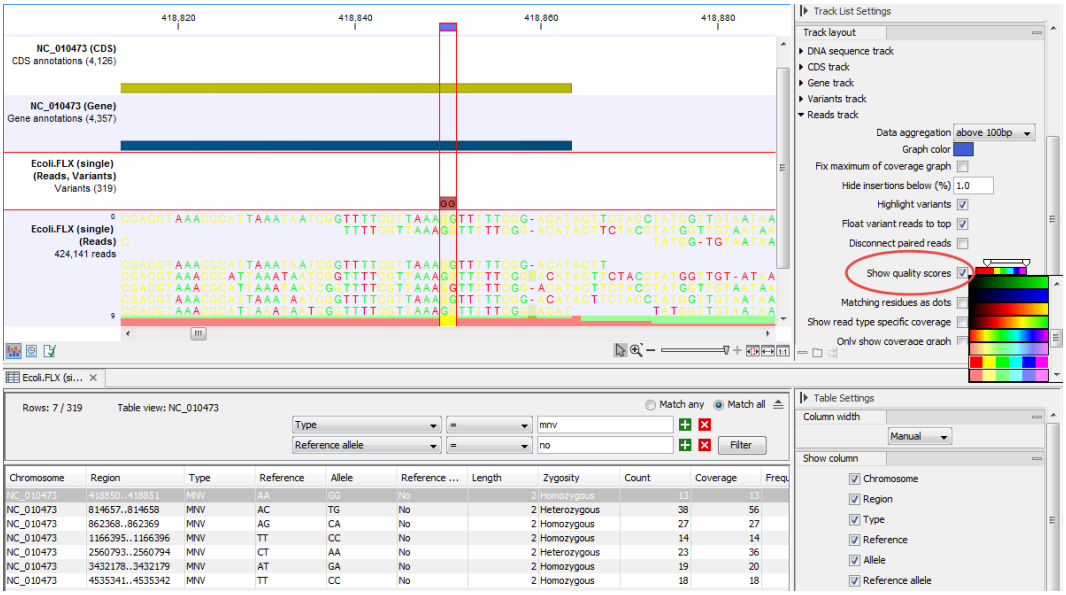


Figure 17: *Viewing the full detail of the selected multiple nucleotide variant.*

You can also try to change the Variant type in the table filter from MNV (multiple nucleotide variant) to **SNV** (single nucleotide variant) and take a look at the mapping.

Note: You can export tables from the Workbench in Excel formats or text formats. Just click on the view of the table, so it is the selected view, and then use the **Export** ( ) button in the toolbar. For small tables, you can also just **Copy** ( ) the contents of the table and paste into a spreadsheet for further processing.