Tutorial

Resequencing Analysis using Tracks

November 28, 2018

Sample to Insight

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

Resequencing Analysis using Tracks

This tutorial takes you through some of the functionality available in the *CLC Genomics Workbench* for targeted resequencing projects, including working with track-based data. Here, we work through a basic analysis of two samples, with the intention of giving a feel for working with such data in the Workbench through a hands-on introduction to a few of the tools available for sample analysis and comparison.

This tutorial includes images and tools from the Genomics Workbench 7.5 *CLC Genomics Workbench*. For earlier versions, please use the Probabilistic Variant Detection tool in place of the Fixed Ploidy Variant Detection tool. If you are working through this tutorial with a *CLC Genomics Workbench* other than version 7.5, the precise locations and names of buttons and tools may be slightly different than described in this document, and results may vary slightly.

Overview The analyses carried out in this tutorial include:

- Mapping reads to a reference sequence
- Local realignment
- Detecting variants
- Comparison of variants
- Refinement of results

Importing the data First, we need to download and import the data.

1. Download the sample data from our website: `http://resources.qiagenbioinformatics.com/testdata/chrM-tutorial-data.zip`.
2. Start the workbench.
3. Import the data by going to:

   File | **Import** () | **Standard Import** ()
4. Choose the zip file called **chrM-tutorial-data.zip**. Leave the Import type set to **Automatic**.

The data set includes two sequencing data files (normal tissue reads and cancer tissue reads) as well as a list of tracks. After import, the files listed in the **Navigation Area** should look like figure 1.

The tracks include the human mitochondrial genome from the hg18 build, **NC_001807 (Genome)** sequence track as well as CDS, Gene and mRNA tracks for this reference. Also included are the **chrMdbSNPCommon** track, which contains the dbSNP common variants for the mitochondrial sequence.

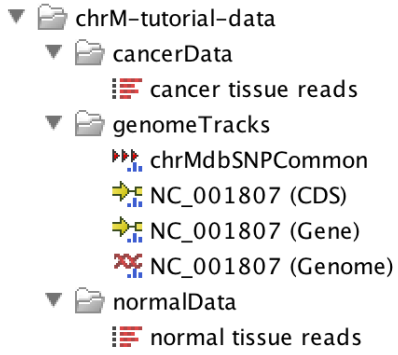


Figure 1: *Navigation area upon import of files.*

Mapping your sequences

In this section we map the reads to the reference sequence using the batch functionality that allows us to launch the mapping task for both sets of reads simultaneously.

Mapping is described in detail in the manual, starting here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Read_mapping.html

Batch functionality is also described in the manual, starting here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Batch_processing.html

1. To begin the mapping, go to:

   Toolbox | **Resequencing Analysis** | **Map Reads to Reference (**  **)**

2. In the first dialog, add a check to the box labeled **Batch**, and then select the top folder called **chrM-tutorial-data** (figure 2).

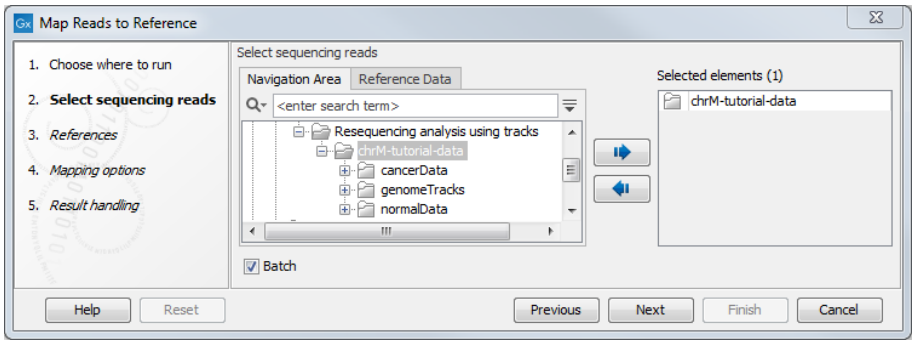


Figure 2: *Select the top level folder, under which the reads folders containing the data are contained.*

3. As we are running a batch mode job, you now see a wizard window that shows the folders that contain data that can be used in the mapping. Click on the folder called **cancerData** (figure 3).

   You then see all the data objects within this folder that will be used in this analysis. Here, we only have one data object in each folder shown, and it is the data we wish to use. However, if you had other relevant data objects which you did not wish to use for mapping, you could use the Exclude and Include fields at the bottom of this window to ensure that

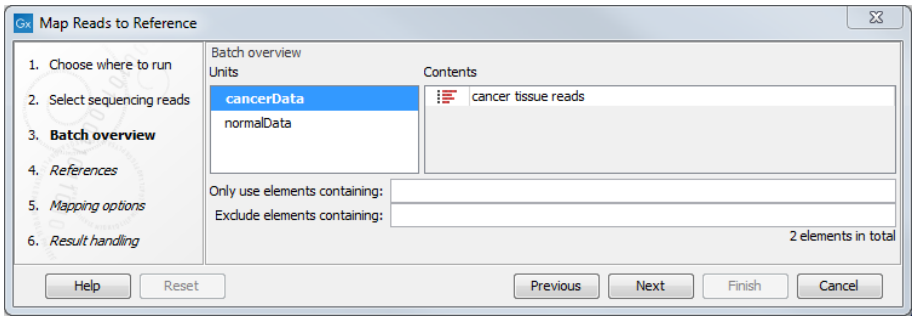


Figure 3: *Check that all the reads needed are included in the analysis.*

only data objects with names that fit the pattern you want will be included. Click on the button labeled **Next**.

4. Select the reference track NC_001807 (Genome) from the `genome Tracks` folder (figure 4).

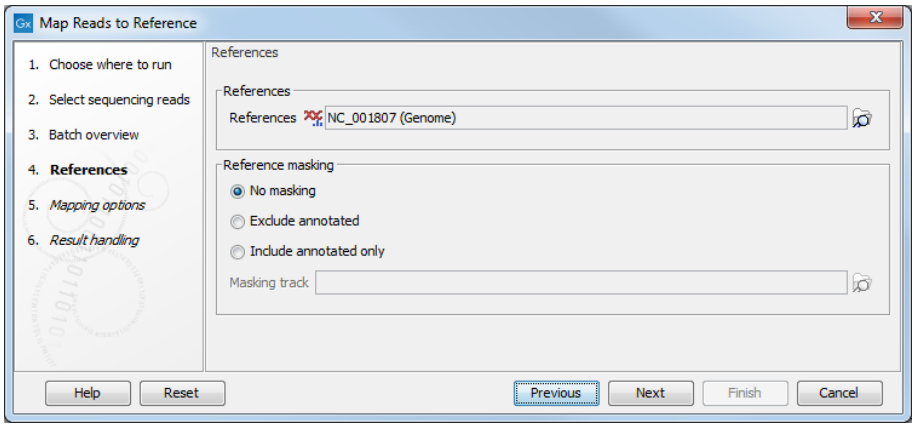


Figure 4: *Specifying the reference sequence(s) to use and the masking to apply, if desired.*

5. We will use the default mapping parameters as shown in figure 5. If the parameters shown in your wizard do not match those in the figure, just click on the **Reset** button (↶) button to reset them.

6. In this dialog, you can choose what type of mapping output you wish to create (figure 6). Click in the radio button beside **Create reads tracks**. Make sure that **Create report** has a check mark. Choose to **Save** the outputs of the mapping in a specified location (here the chrM-tutorial-data folder), and for clarity check the option to "Create subfolders per batch unit". Click **Finish**.

You have now launched a batch job, that includes two mapping jobs - one of the reads from the normal sample against the reference genome, and one of the reads from cancer sample against the reference genome. If you look at the processes tab in the bottom left hand side of the Workbench, you can see the progress of these tasks, similar to that shown in figure 7.

These mappings are being run in batch mode, so each set of results is written to the folder containing the relevant read data. When the tasks are finished, you should see something like figure 8 in the **Navigation Area** of your Workbench.

Feel free to look at the mapping reports if you are interested. This report can be very useful for

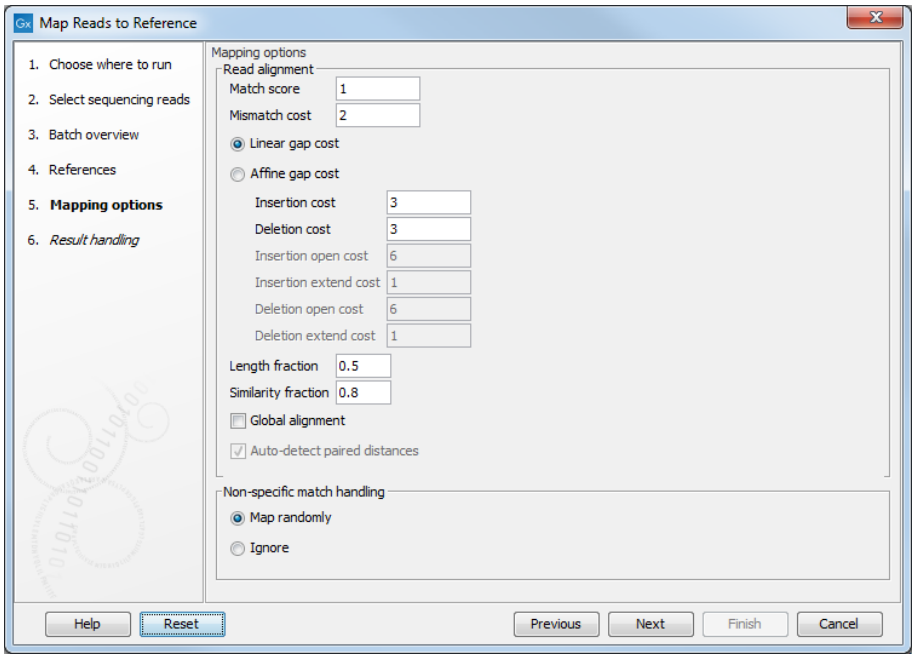


Figure 5: *Set the mapping parameters. Clicking on the parameter reload button resets all parameters to the defaults. Click on the button with the question mark brings up the in-built help, where you can find out more about running mappings via the Workbench.*

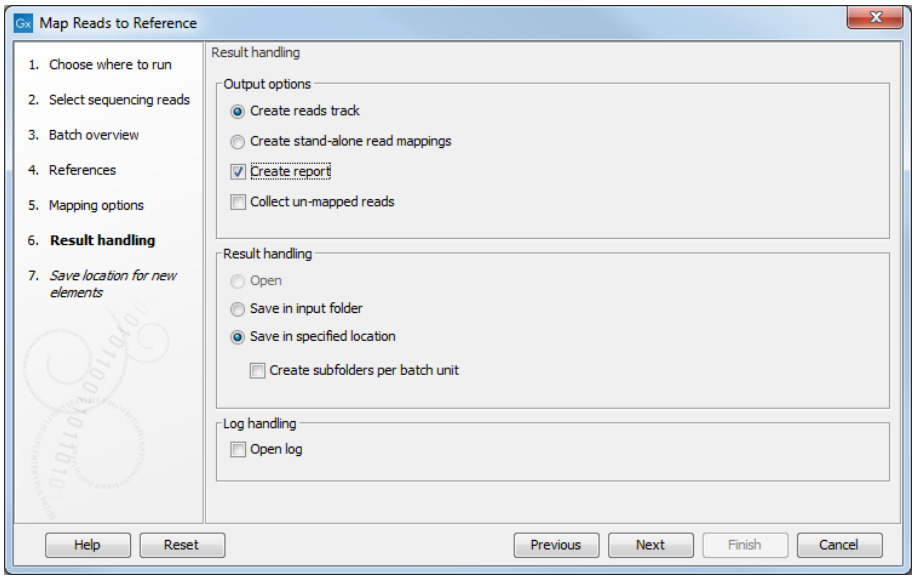


Figure 6: *Results handling options when working in batch.*

considering your mapping results before investing time in carrying out downstream analyses. You can also generate a detailed mapping report using the **QC for Read Mapping** tool.

Local realignment

To improve on the alignments of the reads in the generated read mapping we will perform a **Local realignment**. Local realignment will typically have an effect on any read mapping, whether the reads were mapped using a local or global alignment algorithm (i.e. with the Global alignment option of the mapping tool unchecked (the default) or checked, respectively). It will improve

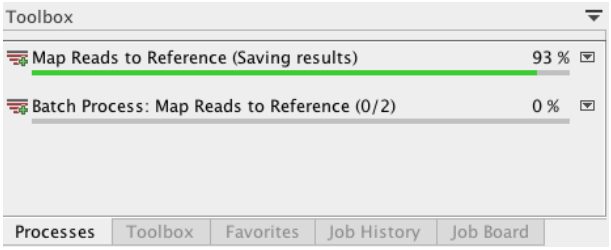


Figure 7: *The progress of the jobs launched can be viewed in the Progress tab in the Workbench. Here, the batch process, and the two mapping jobs it launches, are listed.*

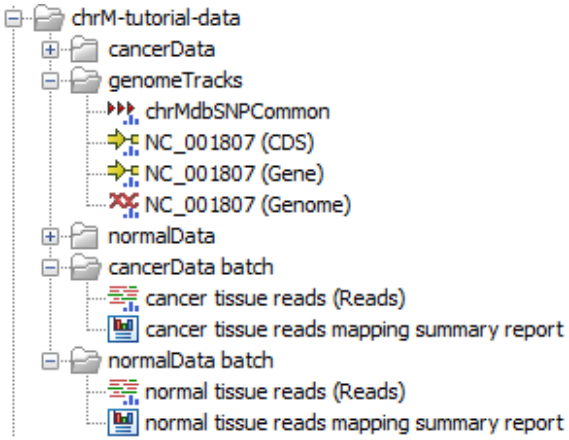


Figure 8: *The mapping track and mapping report for each mapping are saved in data input specific subfolder, within the chrM-tutorial-data folder.*

mapping in areas around insertions and deletions in the sample reads relative to the reference. Please note that the variants reported by the structural variation tool can be fed into the local realignment tool to re-adjust the alignment of the reads to span the indels, making some of the indels detected by the structural variation ready to be picked up by the basic variant detection. We recommend that you refer to the manual for further details on how this tool works before running them on your own datasets: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Local_realignment.html

1. To run the Local realignment tool, go to:

   Toolbox | **Resequencing Analysis** | **Local realignment ( )**

2. The following steps are similar to those performed during initiation of mapping: add a check to the box **Batch** and add the **chrM-tutorial-data** folder.

3. If you decide to inspect the folders, you can see that the tool automatically chose the read mappings that were generated at the previous steps and that were saved in the batch subfolder.

4. In the Realignment settings wizard (figure 9). If you are not sure if the settings you have are the defaults, just click on the **Reset** button ( ) to reset all fields to the default values.

5. Click in the radio button beside **Create reads tracks**. Check the box beside **Output track of realigned regions**. Choose to **Save** the outputs of the mapping in the input folders and click **Finish** (figure 10).

Progress of these tasks can again be viewed in the Progress tab.

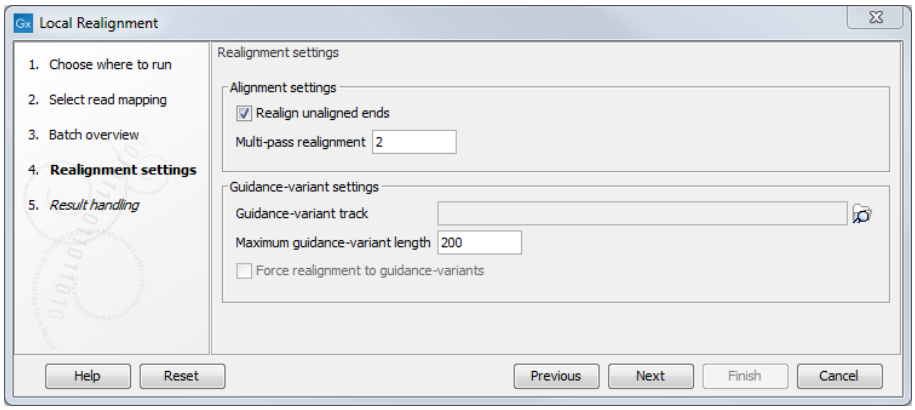


Figure 9: *Set the realignment options.*

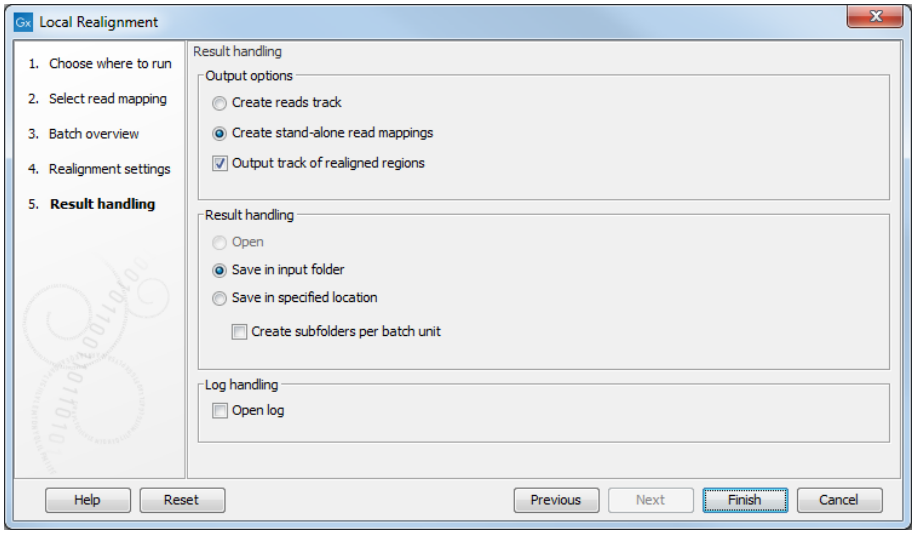


Figure 10: *Result handling options.*

Variation detection

There are different tools in the *CLC Genomics Workbench* for variant detection. Here, we will run the Fixed Ploidy Variant Detection tool. We recommend that you refer to the manual for further details on how this tools work before running them on your own datasets: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Fixed_Ploidy_Variant_Detection.html

The variant detection tool will report Single- and Multiple Nucleotide Variations (SNVs and MNVs), Insertions, Deletions, and Replacements.

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

   Resequencing Analysis | **Variant Detectors** | **Fixed Ploidy Variant Detection**

2. This time we will not run in batch mode. The reason is that after we performed the local realignment, a new read mapping was generated with the locally realigned reads and added to the folder. When running in batch mode, only the first read mapping found in each folder is taken into consideration. So we will run the variant calling separately for the cancer sample first and after that repeat the procedure for the normal sample.

3. Find the **cancerData batch** folder and select the locally realigned reads (**cancer tissue reads (Reads) - locally realigned**).

4. 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 the mitochondrial chrM has a single chromosome (figure 11). For more information about the settings, feel free to click the **Help** button. Otherwise, click **Next**.

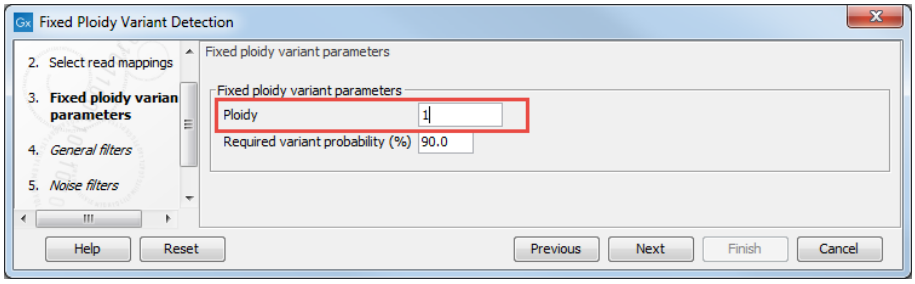


Figure 11: *Setting variation detection parameters.*

5. Keep the default general filter parameters as in figure 12 for this tutorial. If you are not sure if the settings you have are the defaults, just click the **Reset** ( ) button. In general, remember that for good variation detection analyses, you do need to ensure that the settings you choose are relevant for your data set 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. Similarly, the "Minimum variant frequency" would need to be changed when working with mixed samples or non-haploid organisms (i.e., it should be set below 50 % for diploids in order to report heterozygote SNPs).

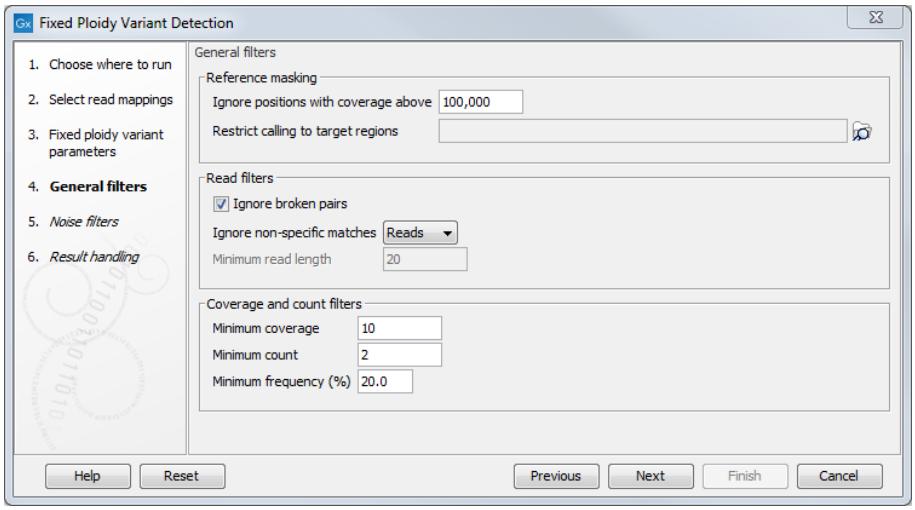


Figure 12: *Setting general filtering parameters.*

6. Choose to accept the default parameter values in the "Noise filters" dialog (figure 13).

7. Here, we choose the type of output to generate. As we are working with track-based objects in this tutorial, click in the boxes next to **Create track** and **Create report**, and ensure the box next to **Create annotated table** is unchecked.

8. Choose to **Save** your results, specify the appropriate subfolder (here **cancerData batch**) and click **Finish**.

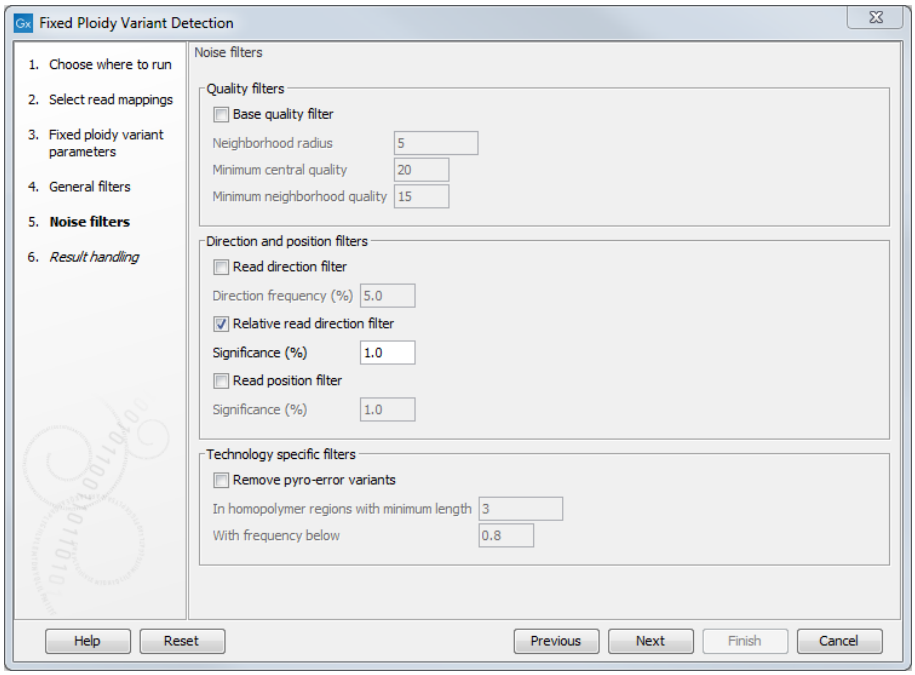


Figure 13: *Setting general filtering parameters.*

You have now launched the Fixed Ploidy Variant Detection tasks for the cancer sample.

Repeat the Fixed Ploidy Variant Detection with the normal sample by following the previous steps using the **normal tissue reads (Reads) - locally realigned** element as input, and saving the output to the **normalData batch** folder.

Viewing the data - working with track lists

Create a track list The output of each of the tasks above included a track object (a read mapping ( ), a realigned region ( ), or a variant track ( )). Each individual track can be viewed by opening it in the viewing area. However, the power of track visualization comes when working with track lists. Track lists allow you to view the reference sequence together with the various annotations, mapped reads and variant calls, and so on, in a single view. From this view, you can easily open up linked tables, allowing you to navigate easily between positions of potential interest, and visually compare information in different tracks for the same position.

1. To create a track list, go to:

   File | **New** | **Track List ( )**

2. Add the variant and mapping tracks you have created, as well as the annotation tracks and dbSNP track into the Selected elements pane, as shown in figure 14.

3. Click **Finish**.

The Track List opens in the View area. You will need to explicitly save it if you wish to access it again later by clicking the view tab and dragging it into the **Navigation Area** to the location where you want to save it. Track list names are rather generic by default. We recommend changing the name to something as meaningful as possible. You can change the name of the track list, or any other data object in your **Navigation Area** by:

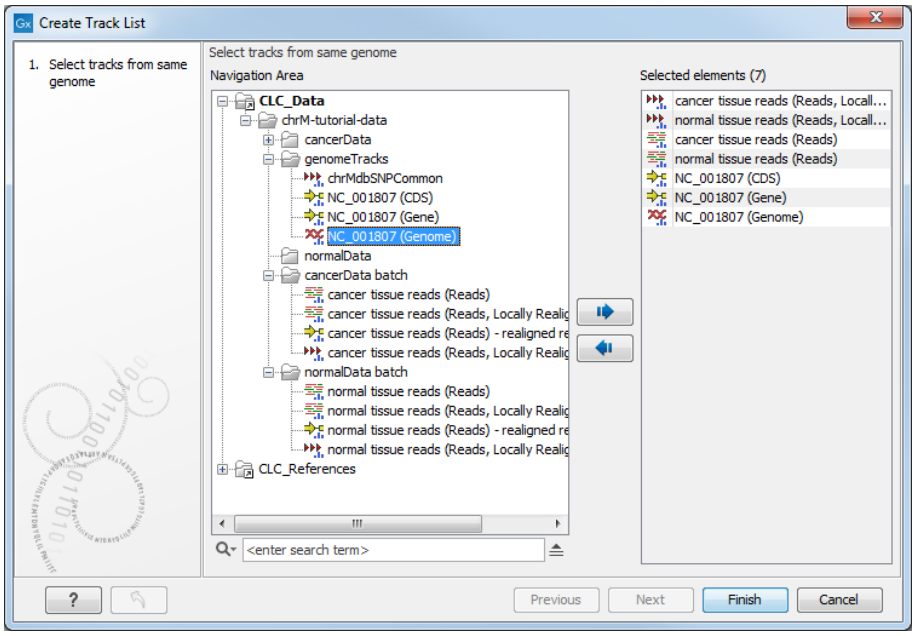


Figure 14: *Selecting tracks for a track list.*

1. Click on the name of a track object in the **Navigation Area**.
2. Either click again on the name or press the **F2** key on your keyboard or right-click on the name and choose **Rename**.
3. Edit the name of the data object to what you wish it to be.

Organizing your tracks You should now see a track list resembling what is shown in figure 15.

It is often useful to organize your tracks so that they appear in a different order within the viewer. For example, you may wish all your annotation tracks to be gathered near the top, or perhaps you wish to view one of your variant tracks right above the read mapping.

To rearrange tracks, click on a track in a track list using the left mouse button. Keeping the mouse button down, you can drag the track to the point in the list where you want it to be.

To increase and decrease the height of any given track, you position the mouse cursor in the left hand area of the track list in the view area, where the names of the tracks are, and moving it to the boundary between two tracks. The cursor should change to look like a double arrow. This indicates you that you can drag the boundary up or down, to change the width of the track. It can be particularly useful to increase and decrease the height of the read mapping tracks when looking at results. Note that read mappings can be stored as tracks, as we did in this tutorial, but also as stand-alone read mapping objects. The stand-alone objects provide more editing possibilities, as well as show a consensus sequence in the view. If you wish to look at the standard read mapping object for your read mappings, you can convert them from tracks using the Convert From Tracks tool.

Filter, annotate and compare your results

At this point, you can start filtering and comparing the data. There are many different filtering and refinement tools available in the **Resequencing Analyses** and **Track Tools** folders of the Toolbox.

Figure 15: *Track list view.*

The result of each filtering and refinement step is a track. Like other tracks, they can be opened and viewed individually, or added to track lists. Note that the default view for tracks is graphical, but if you click on the table icon ( ) at the bottom of the viewing area, you can view the tabular data.

After filtering, we will annotate variants that cause a change in the amino acid based the reference sequence CDS regions.

Filter for cancer-specific variants by comparing to normal reads Here we use the **Remove Variants Present in Control Reads** tool to filter the variants called in one sample against information contained directly in the mapped sequencing reads from another sample. We will compare the variants called for the cancer sample with the read mapping for the normal sample. At sites called as a variant in the cancer sample, there may be reads from the normal sample that have the same change relative to the reference, but where the evidence in the normal sample was not strong enough for a variant to be called. Using the **Remove Variants Present in Control Reads** tool, we can filter for potential cancer-specific variants, using all the data in the mapping of the normal sample, rather than just filtering against the sites that had strong enough evidence for a variant to have been called in the normal sample.

1. Go to:

   Resequencing Analysis ( ) | Variants Filtering ( ) | Remove Variants Present in Control Reads ( )

2. Choose the variant track called **cancer tissue reads (Reads, Locally Realigned, Variants)** that you generated earlier (figure 16).

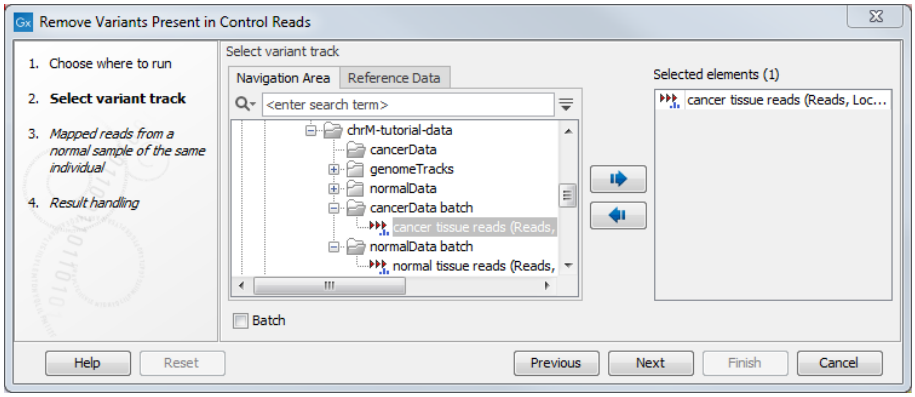


Figure 16: *Select the variant track to filter.*

3. Choose the **normal tissue reads (Reads)** read mapping track for the Control reads track, and to keep variants with control read count below **2** (figure 17).

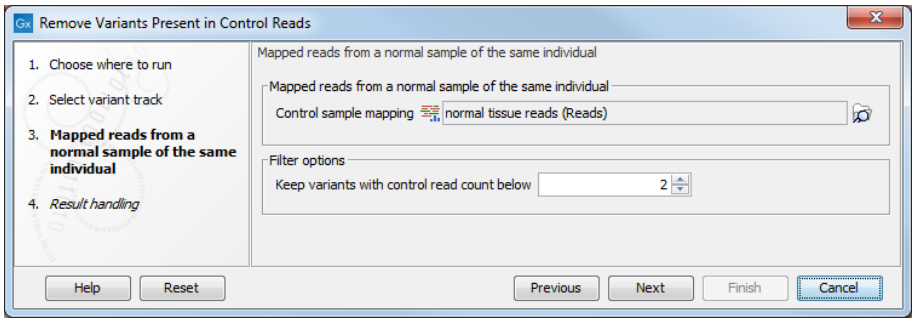


Figure 17: *Select the read mapping to filter the variant track against.*

4. Choose to **Save** the results. The name of the new track is the same as the input variation track, but with a (CTRL) added to the name.

Open this new track in the viewing area. Note that there were 50 variants in the cancer sample, and after comparison with the normal sample read set, there are 22 remaining variants.

Look for SNPs resulting in an amino acid change Here we filter the variations identified in the cancer data, but not the normal reads, and identify those that cause a change in the amino acid composition of a protein.

1. To do this go to:

   Resequencing Analysis ( ) | Functional Consequences ( ) | Amino Acid Changes ( )

2. Choose the track output you just created **cancer tissue reads (Reads, Locally Realigned, Variants, CTRL)**. Note that it can be a good idea to change the name of objects to be more meaningful, whenever possible. If you forget what you have done to create a particular data object, or you want to check the parameters you set for an analysis, you can click on the Show History button ( ) at the bottom of the viewing window when that data object is

open. In that view, you can see what version of the Workbench was used, what data was input, and what the parameters were. Click **Next**.

3. In the next wizard window (figure 18),

   - Select the NC_001807 (CDS) track as the CDS track.
   - Leave the mRNA track empty. The test data set does not contain an mRNA track as we do not need it for this demonstration.
   - Select the NC_001807 (Genome) track as the Genome track.
   - Leave the **Filter away synonymous variants** box unchecked and the **Filter CDS regions with no variants** checked as they are by default. The Genetic code set to 1 Standard.

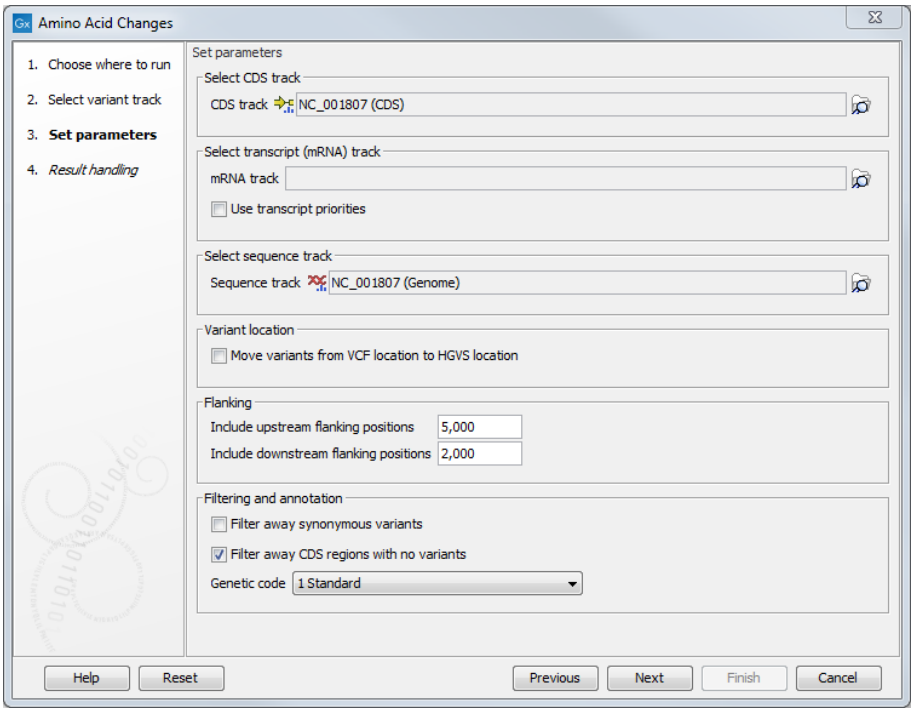


Figure 18: *Set reference and parameters for the Amino Acid Changes tool.*

4. Choose to **Save** the results and click **Finish**.

5. The tool generates 2 views: one called by default **cancer tissue reads (Reads, Locally Realigned, Variants, CTRL, AAC)**, and one called **NC_001807 (CDS) (Amino Acids)**. Click on the **table view** icon ( ) at the bottom of the view area of the first output.

6. Double click on the tab **Cancer tissue...**. Double clicking on the name in the tab maximizes the table to occupy the full space of the Genomics Workbench. This makes it easier to view the whole table at once.

   Notice there are six additional columns compared to the variant track you started with. These are called **Coding region change**, **Amino Acid Change**, **Amino Acid Change in longest transcript**, **Coding region change in longest transcript**, **Other variants within codon**, and **Non-synonymous**.

   All the annotation tools of this type report results by adding columns to the previous results and returning a variant track containing the original data plus these new columns.

Double click on the name in the tab in the Viewing area again to bring the whole Workbench back into view.

Annotate variants You may wish to filter against or annotate using information for known variants. Here, we will annotate the variants we already annotated with amino acid change information with information from dbSNP Commmon. A track of these variants is within the sample data provided.

1. Go to:

   Resequencing Analysis ( ) | **Variant Annotation** ( ) | **Annotate from Known Variants** ( )

2. Choose the variant track for the cancer sample that you used to annotate with amino acid changes. Unless you changed the name, the name of this track will end in (Variants, CTRL, AAC).

3. In the next dialog (figure 19), choose the track called chrMdbSNPCommon, which is in the folder called genomeTracks. You can choose to have the Auto join option to **Automatically join adjacent MNVs and SNVs** checked or unchecked.

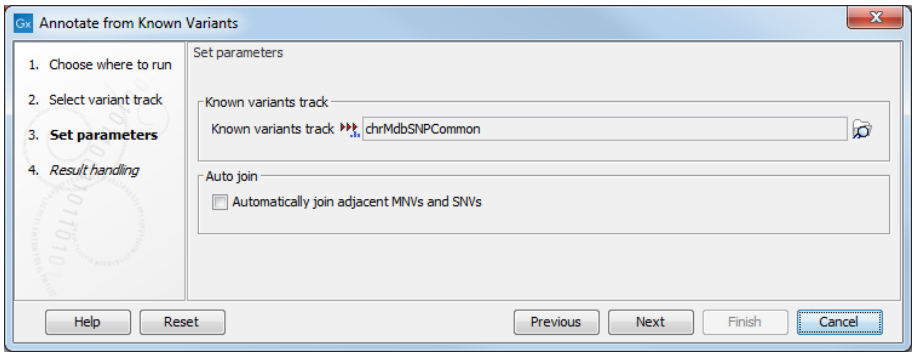


Figure 19: *Set reference and parameters for the Amino Acid Changes tool.*

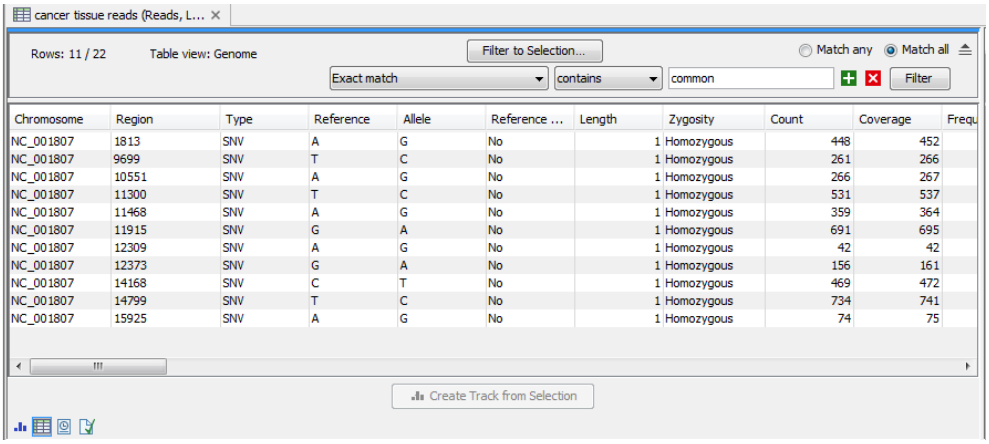
4. Choose to **Save** the results.

The name of the new track is the same as the input variation track, but with a KNOWN added to words in brackets at the end of the name. Open this new track in the viewing area. Switch to table view by clicking on the table icon ( ) found in the lower left corner of the **View Area**.

Apply the advanced filter again by clicking on the small arrowhead in the upper right corner of the table and apply the filter as shown in figure 20.

You can now see that 11 of the variants that were identified in the cancer sample for which evidence was not present in the normal sample have been annotated using information from dbSNP Common.

As for all tables in the Workbench, you can select and deselect the columns you wish to view in the right hand pane of Table Settings as well as filtering and sorting based on the information in particular columns.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequ
NC_001807	1813	SNV	A	G	No	1	Homozygous	448	452	
NC_001807	9699	SNV	T	C	No	1	Homozygous	261	266	
NC_001807	10551	SNV	A	G	No	1	Homozygous	266	267	
NC_001807	11300	SNV	T	C	No	1	Homozygous	531	537	
NC_001807	11468	SNV	A	G	No	1	Homozygous	359	364	
NC_001807	11915	SNV	G	A	No	1	Homozygous	691	695	
NC_001807	12309	SNV	A	G	No	1	Homozygous	42	42	
NC_001807	12373	SNV	G	A	No	1	Homozygous	156	161	
NC_001807	14168	SNV	C	T	No	1	Homozygous	469	472	
NC_001807	14799	SNV	T	C	No	1	Homozygous	734	741	
NC_001807	15925	SNV	A	G	No	1	Homozygous	74	75	

Figure 20: *With the advanced filter function you can easily see how many of the detected cancer variants that are already known in the dbSNP Common database.*

Tips

Adding and removing tracks in track lists Add more tracks to a track list by dragging and dropping track objects from the **Navigation Area** into the opened track list.

Remove tracks from the track list, by right clicking on the track in the track list you wish to remove. Then select **Remove Track** from the menu that pops up.

Saving changes If the name of a data object in the **Navigation Area** appears in bold, italicized text, it means your changes are not yet saved.

Two ways to save data objects open in a view are:

1. Right click on the tab at the top of the unsaved view, and choose **Save As...** from the menu that appears, or
2. Click on the tab at the top of the unsaved view and press **Ctrl-S** on the keyboard.

Once saved, the name of the data object should now appear in standard font in the **Navigation Area**.

History - check what went on before Every opened track in the Viewing area has a history, which you can see by clicking on the History view button ( ) at the bottom of the pane. This is also available for the track list. You will see in the history a full history of all the things you have done to create a track or all changes you have made. This is a good way of double checking what parameters you have used for analyses and what source data you have used for a given track.