

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1 0
A A G G A 0 0 0 0 0
G A T G A 0 1 1 0 0
T A G A T 0 0 1 0 0
G A T G A 1 0 1 0 0
T G T A G 1 0 0 0 0
T A G T A 0 0 0 0 0
G A T A T 1 0 0 0 0
G A G T G 0 1 0 0 0
A G A T T 0 1 0 0 0
G A G T A 0 1 0 0 0
T G A T G 0 1 0 0 0
A T T A G 0 0 0 0 0
T A G A T 0 0 0 0 0
T A G T A 0 0 0 0 0
G A G A A 0 0 0 0 0
G T A T A 0 0 0 0 0
G A T A G 0 0 0 0 0
T A G A T 0 0 0 0 0
A G A A A 0 0 0 0 0
G A G A A 0 0 0 0 0
A A A A A 0 0 0 0 0

Tutorial

Tutorial: Resequencing Analysis using Tracks

February 15, 2013



Tutorial: Resequencing Analysis using Tracks

Introduction

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.

To run this tutorial, you must be working with the *CLC Genomics Workbench*, version 5.5 or higher.

Overview The analyses carried out in this tutorial include:

- Mapping reads to a reference
- 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 web site: <http://download.clcbio.com/testdata/chrM-tutorial-data.zip>.
2. Start the *CLC Genomics 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 three folders. One folder contains the reference genome track, which is the mitochondrial chromosome from the hg18 build of the human genome, along with Gene and CDS annotation tracks downloaded from the UCSC Genome Browser site¹, which were imported using the **Import Tracks** functionality of the *CLC Genomics Workbench*.

The other two folders contain sequence reads for two samples. The reads for each sample have been put in separate folders because we will use the batch functionality of the Workbench to run the analyses, and this will put the output of each analysis into the same folder as the input data. Thus, when using the batch functionality, having the input data in separate folders can make it easier to find the relevant results later.

¹If you want to learn how to create reference tracks, please refer to the tutorial called *Reference Genome Tracks*. To learn about importing annotations from external files, please refer to the tutorial *An Introduction to Annotation Tracks*

Mapping your sequences

In this section we map the reads to the reference sequence, making use of the batch functionality, which allows us to launch the mapping task for both sets of reads simultaneously.

Mapping functionality is described in detail in the manual, starting here:

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

Batch functionality is also described in the manual, starting here:

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

1. To begin the mapping, go to:

Toolbox | NGS Core Tools (🗄️) | Map Reads to Reference (🗺️)

Depending on your local setup, you may be asked where you wish to run the job - on your Workbench or on a Server. If you are presented with this window, choose the appropriate option for your work, and then click on the button labelled **Next**.

2. Click in the box near the bottom of the Wizard window, labelled **Batch**.
3. Click on the top folder called **chrM-tutorial-data** and then click on the right hand arrow icon (➡️) to move this folder into the Selected elements pane on the right hand side.

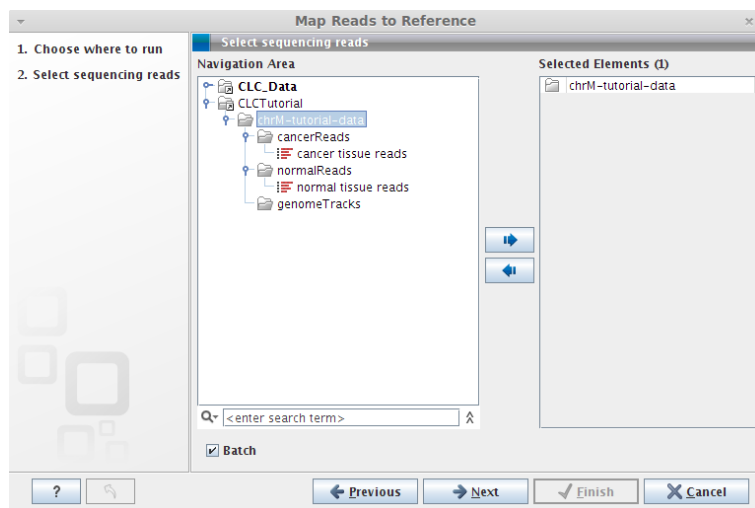


Figure 1: Select the top level folder, under which the reads folders containing the data are contained.

You should now see something like that shown in figure 1.

4. Click on the button labelled **Next**.

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.

5. Click on the folder called **cancerData**.

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 Exclude and Include fields at the bottom of this window to ensure that only data objects with names that fit the pattern you want will be included.

6. Click on the button labelled **Next**.
 7. Click on the browse icon (🔍) within the section labelled **References**.
 8. Select the reference track NC_012920 (Genome) and put it into the Selected elements pane by clicking on the right hand arrow icon (➡), or just by double clicking on the name of the object in the left hand pane.
 9. Click on the button labelled **OK** in the reference selection window.
- At this point, you should see what is shown in figure 2.

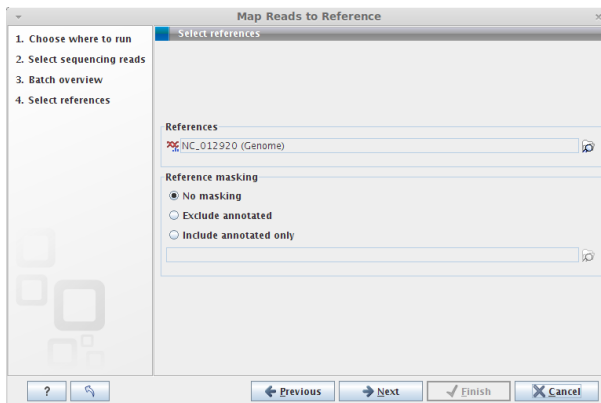
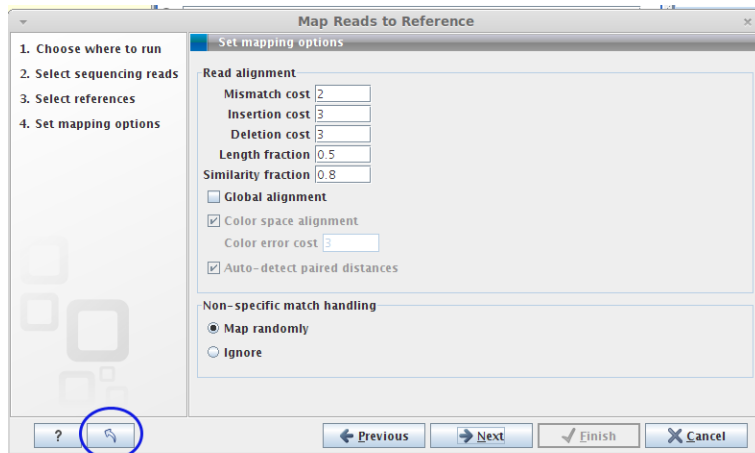


Figure 2: Specifying the reference sequence(s) to use and the masking to apply, if desired.

10. Click on the button labelled **Next**.
11. We will use the default mapping parameters as shown in figure 3. 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.



Click on this button to reset parameters for this wizard step to default values.

Figure 3: 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.

12. Click on the button labelled **Next**.

In this Wizard step, you choose what type of mapping output you wish to create.

13. Click in the radio button beside **Create read tracks**.

14. Click in the box beside **Create report** so there is a check mark in it.

15. Choose to **Save** the outputs of the mapping.

16. Click on the button labelled **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 4.

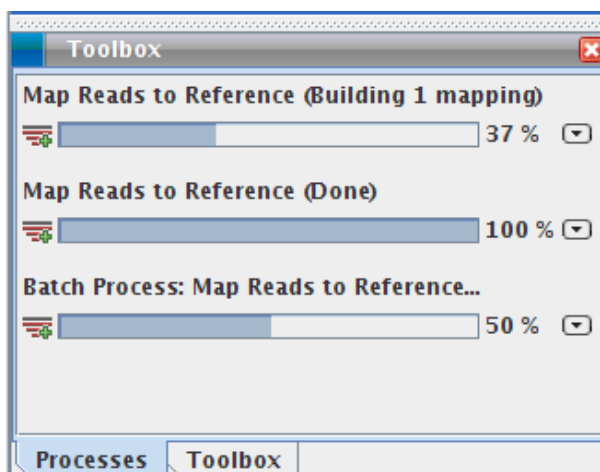


Figure 4: The progress of the jobs launched can be viewed in the Progress tab in the Workbench. Here, the batch process, and the two mapping jobs it launches, are listed.

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 thus see something like figure 5 in the Navigation area of your Workbench.

Feel free to look at the mapping reports if you are interested. Just double click on the object in the Navigation area to open it in the Viewing area of the Workbench. This report, or the detailed mapping report that you can generate using another tool in the Workbench, can be very useful for checking your data before investing too much time in carrying out downstream analyses.

Variation detection

There are different tools in the *CLC Genomics Workbench* for variant detection. Here, we will run the **Probabilistic Variant Detection**. We recommend that you refer to the manual for further details on how this tools work before running them on your own datasets:

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

The variant detection tool will report Single Nucleotide Variations (SNVs), small insertions and deletions (InDels) and Multi Nucleotide Variations (MNVs).

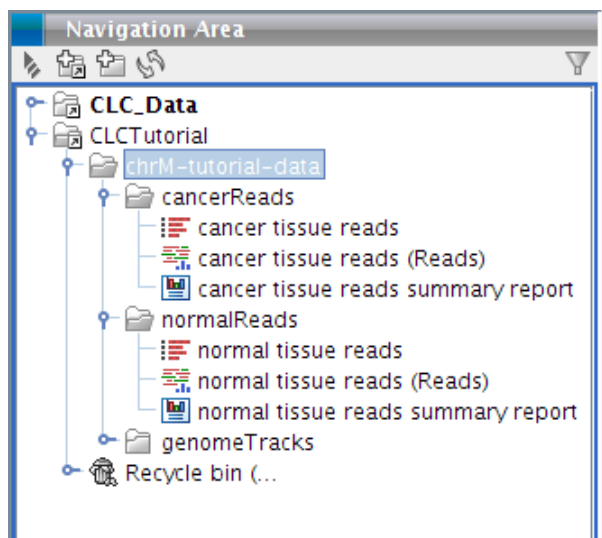


Figure 5: The mapping track and mapping report for each mapping are saved to the folder where the read set used for the mapping is.

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

Toolbox | Resequencing Analysis (🔧) | Probabilistic Variant Detection (TCA)

2. Click in the box near the bottom of the Wizard window, labelled **Batch**.
3. Click on the top folder called **chrM-tutorial-data** and then click on the right hand arrow icon (➡) to move this folder into the Selected elements pane on the right hand side.
4. Click on the button labelled **Next**.

In the next wizard window, we see the folders that contain data that can be used for variant detection.

5. Click on the folder called **cancerData**.
You then see all the data objects within this folder that will be used in this analysis. Here, only the read mapping tracks can be used for variant detection, so that is why you only see that one data object listed in the right hand pane in this case.
6. Click on the button labelled **Next**.
7. Click in the box labelled **Ignore non-specific matches**.
8. Click on the button labelled **Next**.
9. Choose to accept the default parameter values in this Wizard step, which is where you are defining the significant thresholds. If you aren't sure if the settings you have are the defaults, just click on the button in the bottom left with the (🔄) image to reset to the defaults.

For good variation detection analyses, you need to ensure that the settings you choose are relevant for your dataset and for your study. For example, if the **Minimum coverage** is set to 50 but you have a mapping with an average coverage of 20, a lot of potential SNPs will not be reported.

The Variant probability setting is asking for how certain you wish to be that a position in your sample is different than the position reported in the reference. Positions with this probability or higher of being different than the reference will be reported in the output. For example, the default value, 90%, indicates that in order to be included in the report, any given site must have at least a 90% probability of being different than the reference. For such sites, and in cases where there may be more than one variant type in the sample at that site, only the variant with the highest probability will be reported².

10. Click on the button labelled **Next**.
11. This Wizard page allows you to set parameters that affect the reporting of the variants. Set the **Maximum expected alleles** to **2**, and leave the **Genetic code** set to **Standard**.
12. Click on the button labelled **Next**.
13. Click in the box next to **Create track**, to generate track-based output. Uncheck the box next to **Create annotated table**.
14. Choose to **Save** your results.
15. Click on the button labelled **Finish**.

You have now launched two Probabilistic Variant Detection tasks - one for the normal sample and one for the cancer sample. Again, the output here will be written to the same folder as the read mapping used in the analysis.

These jobs may take a few minutes to complete.

Viewing the data - working with track lists

Create a track list The output of each of the tasks above included a track object. Each individual track can be viewed by opening it in the viewing area. However, the power of track visualisation 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:

Toolbox | Track tools (📁) | Create Track List (📊)

2. Add the variant and mapping tracks you have created, as well as the genome and annotation tracks, into the Selected elements pane, as shown in figure 6
3. Click on the button labeled **Finish**.

Note that this just opens a track list in the Viewing area. You will need to explicitly save it if you wish to access it again later.

²The chances of the position being different than the reference is the sum of the probabilities of all the different types (homozygous and heterozygous combinations) that could be at that position. The reference call is assumed to be homozygous.

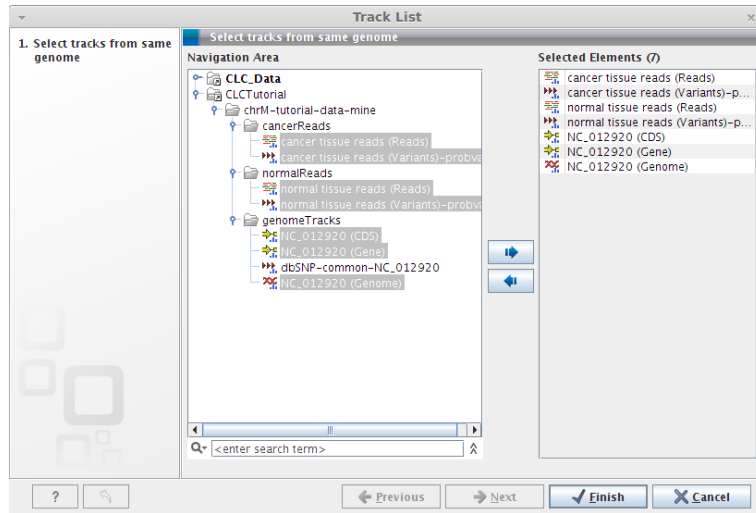


Figure 6: Selecting tracks for a track list.

4. Save the track list by clicking the view tab and dragging it into the navigation area.

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:

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.
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 something like that in figure 7



Figure 7: Track list view.

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 depressed, you can drag the track to the point in the list where you wish it to be.

You can also increase and decrease the height of any given track. You do this by positioning 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 an up arrow with a small horizontal bar about it. 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 to increase and decrease the height of the read mapping tracks when looking at results³.


Filter, annotate and compare your results

At this point, you can start filtering and comparing the data. For example, here we will take two routes to filtering to get a list of variants that may be specific to our cancer sample:



1. Filter out any variants identified by the Probabilistic Variant Detection tool in both the normal and cancer samples.
2. Compare the variants identified by the Probabilistic Variant Detection tool in the cancer sample to the reads from the normal sample.

There are many other different filtering and refinement tools available in the Workbench that are not touched on in this tutorial. You are welcome to investigate these as you wish. You will find most of them under subsections of the **Resequencing Analyses** section of the toolbox.

After doing the filtering mentioned above, we will take the output of the second filter and annotate those variants that would cause a change in the amino acid coded for in our sample compared to the reference.

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.

Filter for cancer-specific variants by removing variants called in the normal sample Here, we subtract out variants called in the normal sample from the list of variants called in the cancer sample. To do this:


1. Go to:
Toolbox | Resequencing Analysis () | Annotate and Filter Variants () | Filter against Known Variants ($\frac{T/C}{T/C}$)
2. Choose the variant track for the cancer sample that you created earlier.
3. Click on the button labeled **Next**.
4. Choose the variant track for the normal sample that you created earlier as the variant track for the other sample.

³The standard read mapping object gives more viewing possibilities, as well as providing a consensus sequence in the view. If you wish to look at the standard read mapping object for your read mappings, you can covert them from tracks using the **Covert From Tracks tool** in the **Track Tools** section of the Toolbox.

5. Choose to **Keep variants with no match found in the track of known variants**.

6. Click on the button labeled **Next**.

7. Choose to **Save** the results.

You may wish to create a new folder to save the results of your comparison to, just to keep things tidy. If you wish to do this, just click on the **Add Folder** button  at the top of the Save window and then choose to save to the new folder you create.

The name of the new track is the same as the input variation track, but with a (HAPLO) added to the name.

8. Open this new track in the viewing area.

You started with 46 variants called in the cancer sample, but after subtracting those variants that were also called in the normal sample, you have 32 remaining variants.

Filter for cancer-specific variants by comparing to normal reads Here, instead of filtering one set of variants against another set of variants, we instead use the **Filter against Control Reads** tool to filter the variants called in one sample, against information contained directly in the mapped sequencing reads from another sample. We expect this tool to allow for a more stringent narrowing down of the list of variants specific to the first sample.

For example, here, we 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, in the normal sample the evidence was not strong enough for a variant to be called. Using the **Filter against 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:

Toolbox | Resequencing Analysis  | **Compare Variants**  | **Filter against Control Reads...** 

2. Choose the variant track for the cancer sample that you created earlier.

3. Click on the button labeled **Next**.

4. Choose the normal tissue reads track for the Control reads track.

5. Choose to keep variants with control read count below **2**.

6. Click on the button labeled **Next**.

7. 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.

8. Open this new track in the viewing area.

Here, there were 46 variants in the cancer sample, and after comparison with the normal sample read set, there are 29 remaining remaining variants.

Using the **Filter against Known Variants** tool, we had a list with 32 variants that appeared only in the cancer sample. So we can see that three additional variants were removed from the cancer-specific variant list when comparing the data to the normal sample read mapping, compared to when we filtered against a list of called variants from the normal sample.

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:

Toolbox | Resequencing Analysis (📁) | Functional Consequences (📁) | Amino Acid Change (🌿)

2. Choose the track output containing variants called in the cancer sample, and not supported by evidence in the normal read set. The default output name for this will be something like **cancer tissue reads mapping (Variants, CTRL)**. (It is a good idea to change the name of objects to be more meaningful, whenever possible.)⁴
3. Select the NC_012920 (CDS) track as the CDS track.
4. Select the NC_012920 (Genome) track as the CDS track.
5. Leave the Filter synonymous box unchecked, and the Genetic code set to Standard.
6. Click on the button labeled **Next**.
7. Choose to **Save** the results.
8. Open the results in the viewing area.
9. Click on the table view icon (📄) at the bottom of the viewing area.
10. Double click on the name in the tab in the Viewing area, which probably looks something like **cancer tissue...**
Double clicking on the name in the tab shows you that view in the full space of the Genomics Workbench. This makes it easier to view the whole table at once.
Notice there are three new columns, compared to the variant track you started with here. These are called **Coding region change**, **Amino Acid Change** and **Non-synonymous**.
All the annotation tools of this type work similarly in that they add columns to the previous results.
11. Open up the detailed filtering options for the table by clicking on the small triangle in the top right hand side. See figure 8 for one suggested filter you could apply.
12. Double click on the name in the tab in the Viewing area again.
This brings the whole Workbench back into view.

⁴If 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.

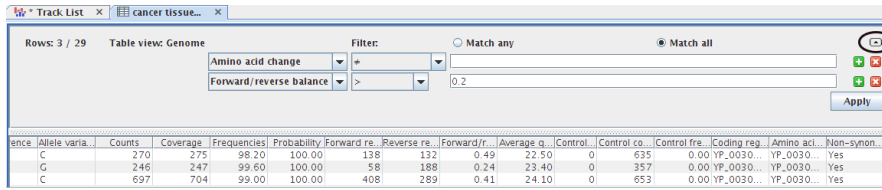


Figure 8: Complex filters can be set up for any table to help refine results.

Other things to note

Adding and removing tracks in tracklists 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's gone 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.