# Tutorial

## Reference Genome and Annotation Tracks

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

# Reference Genome and Annotation Tracks

This tutorial introduces two ways to create reference genome and manage tracks lists in the *CLC Genomics Workbench*.

The first method to create a reference genome is for those wishing to download model organism genome data and annotations related to those genomes. The second method is more general and involves downloading data from Genbank.

We will subsequently download an annotation file from an external source and import this as a track. Doing this allows you to add valuable information to an existing reference track in the *CLC Genomics Workbench*.

## Method one: Downloading model organism sequences and annotations

In this section (optional to the rest of the tutorial), we obtain the hg19 build of the human genome and associated annotations from public resources. Using the method in this section, you will end up with these data as tracks.

1. Click on the button marked **Download (** **)** in the top toolbar.
2. Choose the option **Download Reference Genome Data (** **)**
3. 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 labeled **Next**.
4. Select **Animals(mammals)** and **Homo sapiens (hg19)** from the drop down list of organisms available. Click on the button labeled **Next**.
5. In this window, you are offered the choice of using an existing reference track, or downloading the sequences. Here, we will download the sequences, and then also choose some of the available annotations to be added as tracks. If you were to come back later and wish to get additional annotation tracks, you could then choose to use the existing track you are creating now. Select to **Download genome sequence** in the next window and click on the button labeled **Next**.
6. Here you should see a list of annotations that can be downloaded as tracks. Apart from a description of the tracks, there is also a column showing the file size. Here we will just choose two of the smaller annotation sets to download: click on the box to the left of **Gene annotation** from Ensembl, and also check the box to the left of **Dbsnp (common) variants** from UCSC. Click on the button labeled **Next**.
7. Choose to **Save (** **)** the results and click **Next**. You may wish to click on the ( ) button to create a new folder to hold this data before clicking on the button labeled **Finished**.

You have just started to download the full human genome. This will take some time!

If you do not really wish to continue with this download, you can cancel it by going to the Processes tab in the bottom left of the Workbench. Click on the small triangle next to the download process and choose to stop the job.

There is data for a number of model organisms available to download using this method, and the annotation data available to download directly depends on the genome. For organisms not available for download, or for working with a subset of a particular reference data, see the section on method 2.

### Method two: Using other sequence data sources

We first look at downloading data into the *CLC Genomics Workbench* from the NCBI. We then look at making tracks for such data within the Workbench.

**Downloading data from the NCBI** Using this method in this section, you can search and download whatever you like from Genbank. In this tutorial, we will retrieve human chromosome 17.

1. Click on the button marked **Download (** **)** in the top toolbar.
2. Choose the option **Search for Sequences at NCBI (** **)**.
3. Type **NC_000017** in the box next to **All Fields**. This is the accession number for human chromosome 17[1]. Make sure you have selected the correct database by selecting the radio button in front of "Nucleotide" (see red highlight in figure 1).
4. Click on the button labeled **Start search (** **)**.

   You should find a single entry is returned to you - human chromosome 17.
5. Click on the entry returned to you to highlight it.
6. Click on the button labeled **Download and Save** (figure 1).

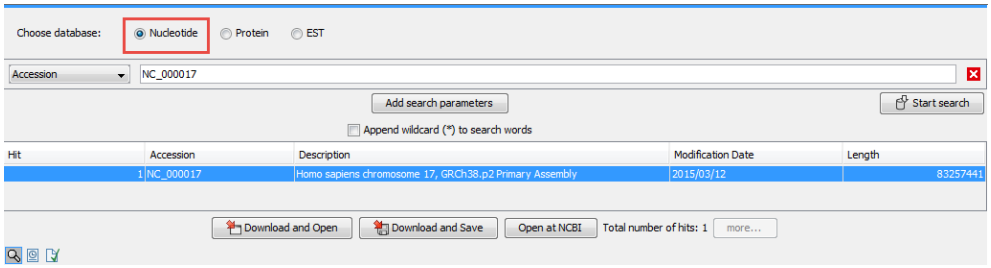


Figure 1: *NCBI search for human chromosome 17.*

Data held at Genbank often has annotations associated with the sequence data. This means that the size of files containing whole chromosomes can be quite large, and can take some time to download.

When you download data using the above method, you create a data object in the Workbench: a sequence object if you downloaded a single sequence or a sequence list if you downloaded

[1]You can use standard Entrez search queries in the search boxes. For example, if you wrote **NC_00017[Accession] OR NC_012920[Accession]**, you would retrieve human chromosome 17 and a human mitochondrial genome. You can also search for multiple search terms of different types. For example, you could select **Organism** from the drop down list of fields to search in. You could then enter the text **human**. Then, you could add another search term by clicking on the button with a plus symbol. If you keep the **All Fields** option, and then wrote **NC_000*[Accession]**, you would find all the human chromosomes, but not the mitochondrial genome.

more than one sequence. For this tutorial, we are interested in creating tracks. Converting your sequence data into a reference track is covered in the next section.

**Creating tracks** Once you have your sequence data available in track format, it can be used in a track list, which can then allow you to view sets of data that all refer to the same reference in a single view. Here, we will create a reference genome track from the chromosome sequence we just downloaded. This is very straightforward, and the same process is taken if you are working on multiple sequences held in a sequence list.

1. Go to:

   **Toolbox** | **Track Tools (** ) | **Convert to Tracks (** )

2. 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 labeled **Next**.

3. Choose the sequence you just downloaded so that it appears in the Selected Elements column on the right hand side. If you have not changed the name, this will be called NC_000017. Click on the button labeled **Next**.

4. Click in the box next to **Create sequence track** to create a track from the sequence data only.

5. Click in the box next to **Create annotation tracks**. We wish to create tracks from the annotations for our sequence.

6. Now click on the green plus sign to the right of the box next to Annotation types. This opens up a window allowing you to choose which types of annotations you wish to make tracks from. A track will be created for each annotation type you choose. Choose the annotation types **CDS** and **Gene** by selecting them and clicking on the arrow to move them into the right hand pane (figure 2).

   Click on the button labelled **Done** when you are done selecting annotations, and click on the button labeled **Next**.

7. Choose to **Save** the results. Click on the button labeled **Next**.

8. Choose to add a new folder ( ) that you can call **chr17tracks**. Click on that new folder so the tracks will be saved there before clicking on the button **Finished**.

You should now have three track objects (figure 3). One contains your reference sequence track, and the other two are annotation tracks - one track with gene annotations and one track with cds annotations.

## Track lists

A track list is a collection of tracks. Gathering them in a list allows you to view multiple tracks for the same reference simultaneously.

This can be done at any time, as it is easy to add and remove tracks from existing track lists.

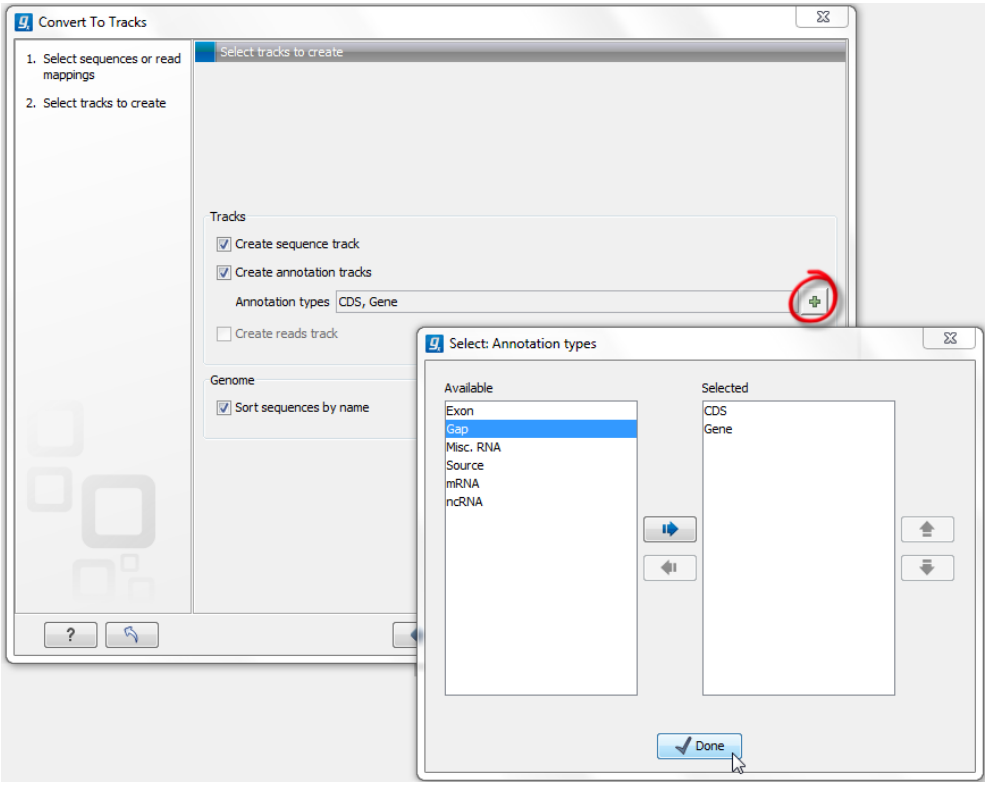


Figure 2: *When using the tool Convert to tracks you can choose which of the annotation types found should be converted to tracks.*

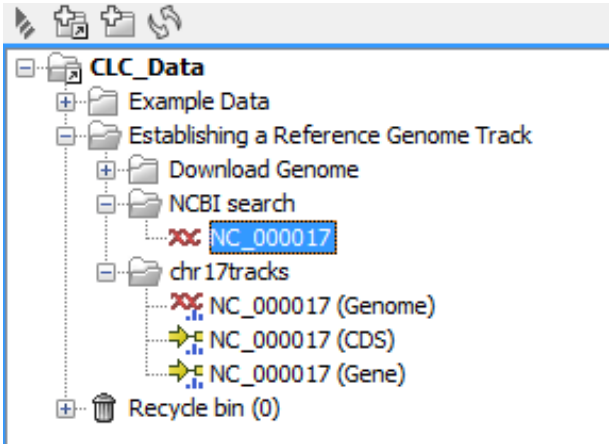


Figure 3: *The Navigation Area after converting the reference to tracks.*

### Creating a track list

1. Go to:

   **Toolbox** | **Track Tools (** ) | **Create Track List (** )

2. Select the tracks you wish to include in your track list. Click on the arrow to move these into the right hand pane. There are two conditions that must be met:

   (a) A reference sequence track must be one of the members of the track list.

   (b) The tracks you choose must be appropriate for that reference track.

For this tutorial, you should select all the tracks that you have from either the first section of the tutorial *or* the second part of the tutorial. The images (figure 4) here are for the data in the second part of the tutorial.

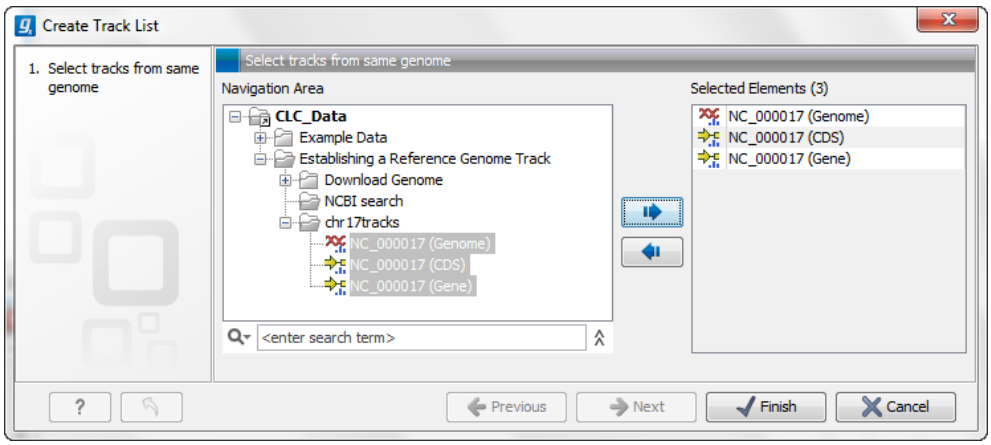


Figure 4: *Selecting tracks to be included in a track list.*

3. Click on **Finished**.

Your track list should then open up in the workspace, see figure 5.

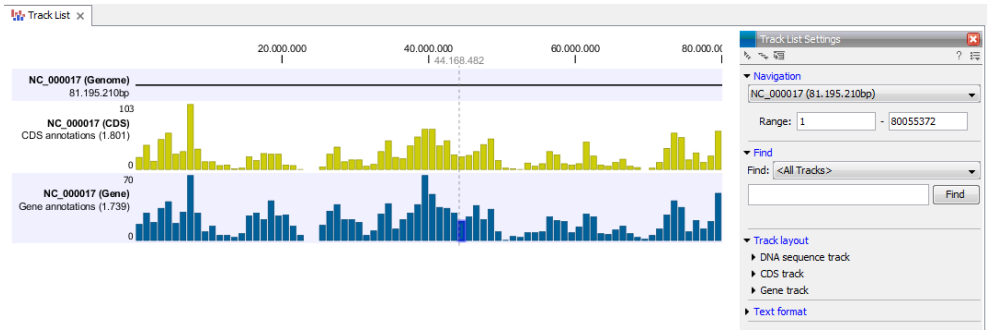


Figure 5: *Viewing the chromosome 17 track list.*

**Note** that this track list is not saved anywhere. You must save it if you wish to keep it. You can do this a number of ways. Two ways are:

1. Right click on the tab at the top of the new view, and choose **Save As...** from the menu that appears, or
2. Move the cursor over the tab at the top of the new view, press the left mouse button. Keeping the mouse button down, drag the tab over into a folder in the **Navigation Area** of the Workbench. The data will then be saved into that folder.

The track list, by default, is given a rather generic name. It is a good idea to change it if you have saved the track list. To do this, click once on the name of the track list in the **Navigation Area**. After a brief pause, click on the name again. This should show the name highlighted with a cursor. You can then change the name.

You can easily add additional annotation tracks to an existing track list.

1. Open the track list you wish to add a track to.
2. Locate the new track in the **Navigation Area** and drag-and-drop this file into the open track list view.
3. **Save (** **)** the updated track list.

### Downloading annotation data from external source

We will download *dbSNP(common)* annotations from the UCSC site. We are working with chromosome 17 of the human genome, so we wish to download annotations specific for this chromosome. If we download the dbSNP Common annotations via the **Download** function in the Workbench, we will get all chromosomes and to be able to download annotations for only chromosome 17 we have to do it directly from the UCSC site.

We need to download the annotations in a format that is recognized by the *CLC Genomics Workbench*. Here, we will choose txt format. To save space, we will choose to download it as a gzip compressed file.

The full list of annotation formats recognized by the Workbench can be found in the *CLC Genomics Workbench* usermanual.

1. Go to the UCSC table browser: `http://genome.ucsc.edu/cgi-bin/hgTables`.
2. Specify the search parameters as shown in figure 6. Key to downloading annotations only for chromosome 17 are:
   - Selecting **region: position**
   - Entering **chr17**
   - Clicking on the button **lookup**

   Note: To download dbSNP for the entire genome, you should simply select **region: genome** instead.

   Remember to give the output file the following name: "dbSNPCommon.txt", and to chose to download it as a gzip compressed. Click on "Get output".
3. Once all the parameters are as shown in figure 6, click **get output**.

If you have used the settings shown in figure 6, then the file you download will be called commondbSNP138chr17.txt.gz.

### Importing the annotation data

Now we have a compressed txt format file containing the dbSNP common variations. We now import this data into the Workbench.

1. Import the dbSNP annotations as a track by going to:

   **Import (** **)** **|Tracks (** **)**

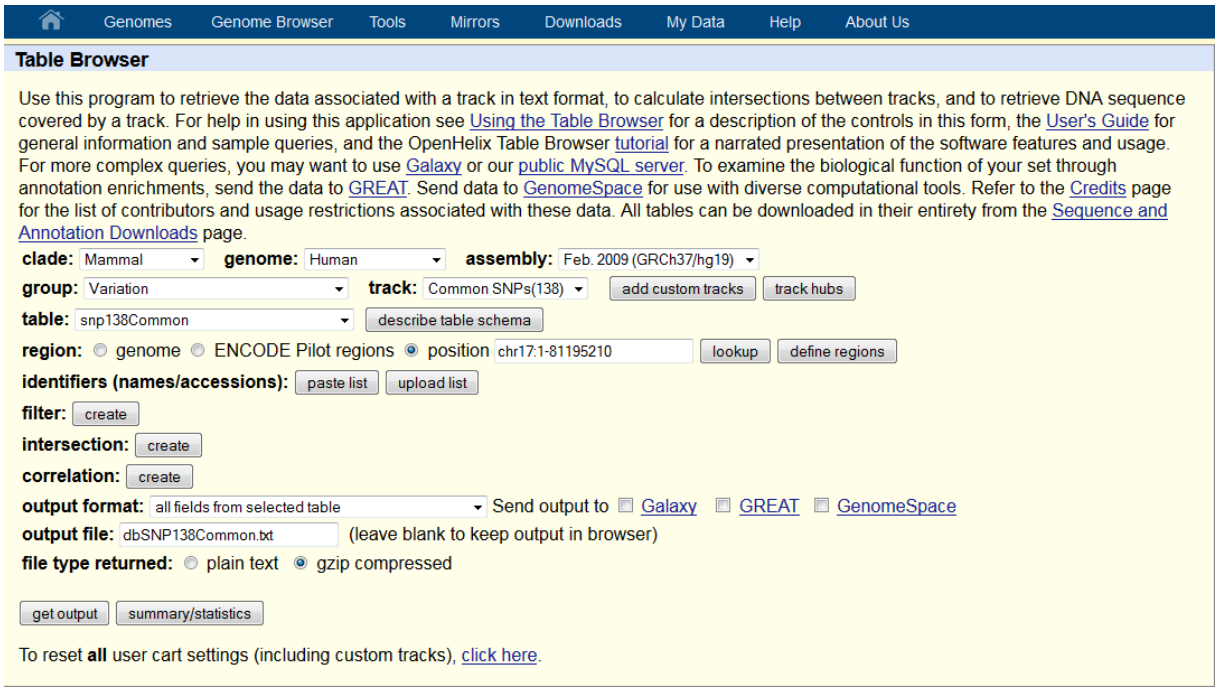


Figure 6: *Search terms for the UCSC table browser window.*

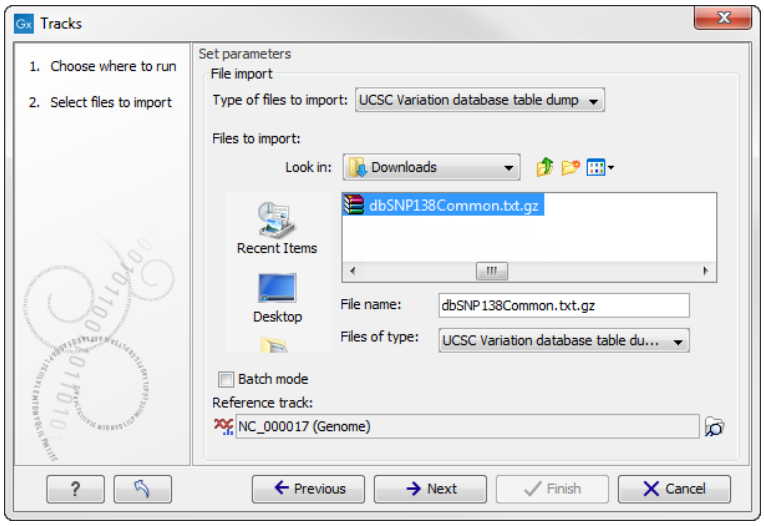


Figure 7: *When you import the dbSNP common annotations, the "Type of files to import" should be switched to "UCSC Variation database table dump". Find and select the annotation file that you have downloaded and under "Reference track:" select the chromosome 17 track. If you cannot se the chromosome 17 track in the folder in the Navigation Area that holds the data, it may be because you have not yet converted the chromosome 17 sequence to a track.*

2. Fill in the wizard window as seen in figure 7.

   - Under "Type of files to import" click on the drop-down list and select **UCSC Variation database table dump** and select the annotation file, which will be named **commondbSNP_chr17.txt.gz** if you used the settings from figure 6.
   - For **Reference track ( )** browse to locate the chromosome 17 reference track. For the data in figure 4, this would be the track called **NC_000017 (Genome)**.

3. Click on the button labeled **Next**, choose to **Open** the track and click **Finish**.

Importing can take some time, depending on the size of your annotation file.

As the name of the imported data is "Variants", it may be a good idea to rename to a more appropriate name.

### Viewing the annotation track

Once the import has completed, you can open the track in the viewing area of the Workbench and investigate the annotations themselves.

1. Open the annotation track so it is visible in the viewing area of the Workbench. By default the track view is zoomed all the way out. That is, you see all the annotations for the whole chromosome. The annotations are represented by vertical bars.
2. Mouse over one of the bars. A pop-up window will provide information on the number of SNPs (items) in the specified range of bases (see figure 8).

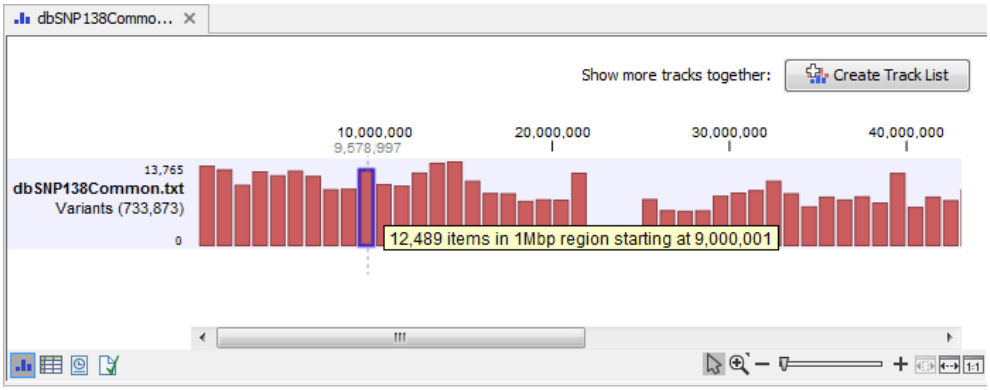


Figure 8: *Open the commondbSNP_chr17 track to have a look.*

3. Zoom in until each SNP is represented by its own small bar.
4. Mouse over one such bar. Now the pop-up window will display the information available for that specific SNP (see figure 9).

   You can also view the annotation track in table view. To open this in a split view:

5. Press and hold the Ctrl-button on the keyboard (⌘ on Mac) and click Show Table ( ) at the bottom of the view.

   This brings up a split view showing the track and table view at the same time. Notice that the two views are linked. Selecting a row in the table highlights the specific variant in the track view with a vertical line (figure 10).

   The table can be filtered in different ways. The most simple way is to click on the heading of a row. The other option is to use the filter found in the upper right corner of the table area.

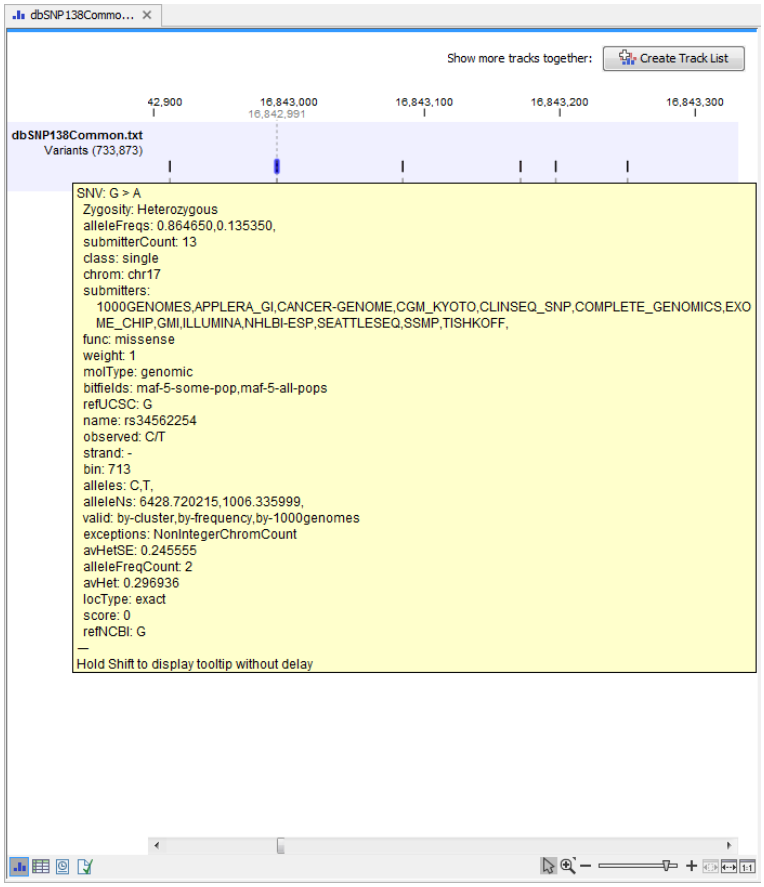


Figure 9: *Zooming in, individual SNPs are represented as bars. Mousing over results in a pop-up box with information about the SNP.*

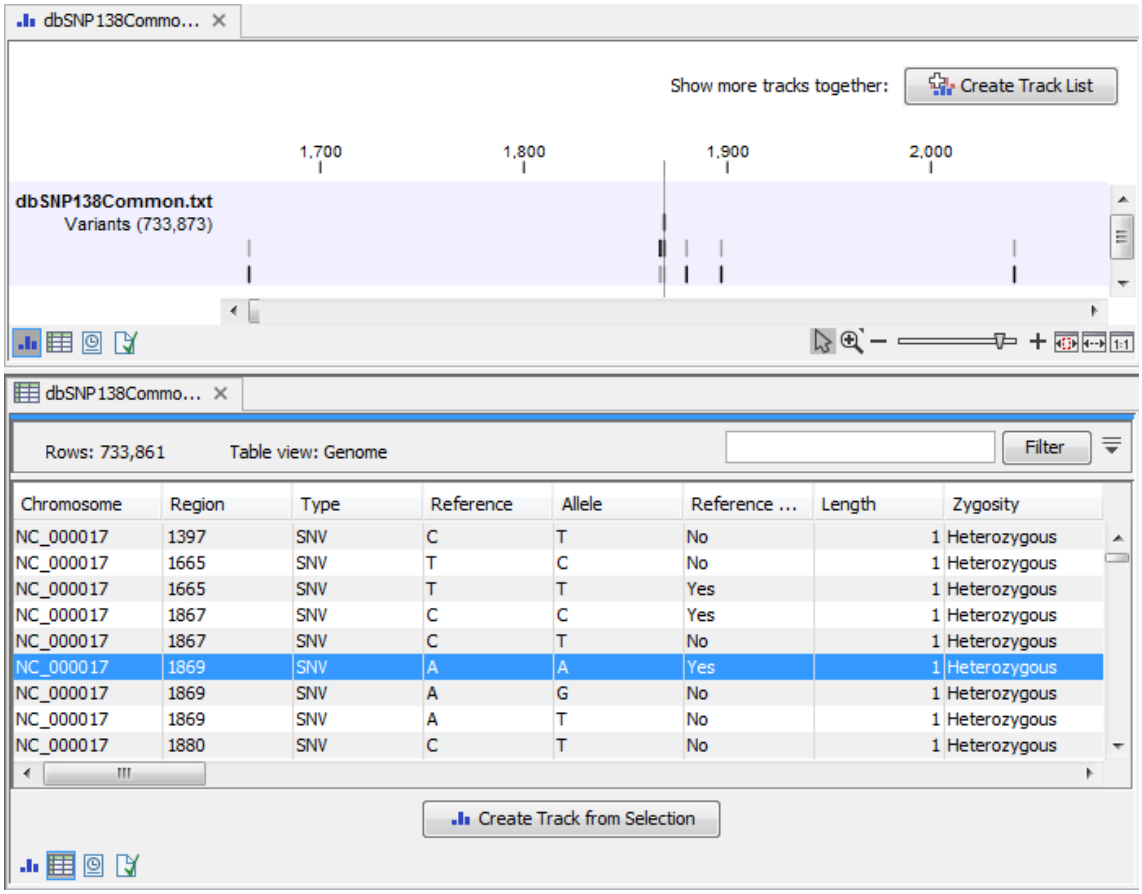


Figure 10: *Track and table split view of the annotations. The two views are linked.*