QIAGEN

Tutorial

Reference Genome and Annotation Tracks

June 27, 2019

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

Reference Genome and Annotation Tracks

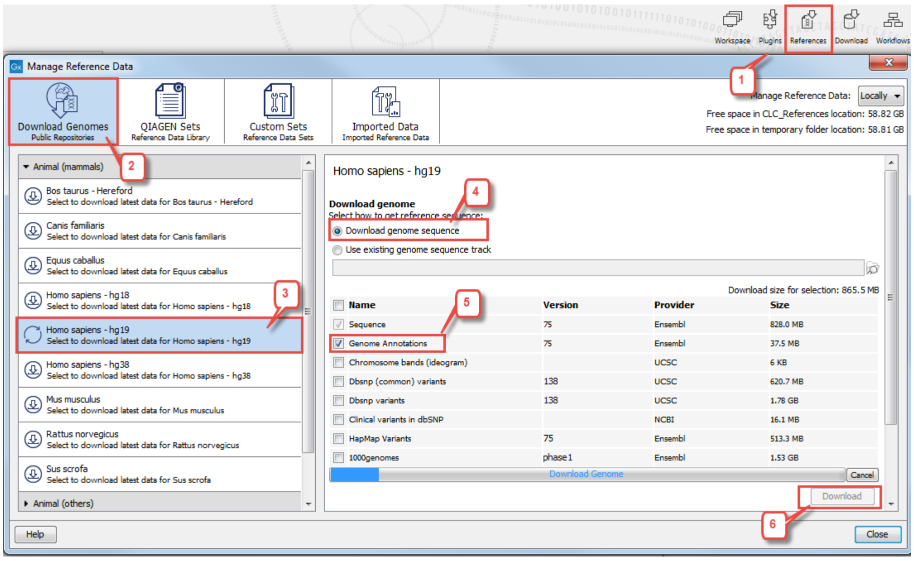
This tutorial introduces two ways to create reference genome and manage tracks lists in the *CLC Genomics Workbench*.

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

Before getting started you will need to download a reference genome and CDS track using the Reference Data Manager (1) found in the upper right corner of the Workbench (figure 1). Under the **Download Genome** tab (2), select the **Homo sapiens - hg19** data (3). Choose to "Download genome sequence" (4) and check the "Genome Annotations" item (5) to get annotation tracks for hg19. Click on **Download** (6). You can check the Download process at the bottom of the wizard.


Workspace
Plugins
References
Download
Workflows
1
Manage Reference Data
Download Genomes
Public Repositories
QIAGEN Sets
Reference Data Library
Custom Sets
Reference Data Sets
Imported Data
Imported Reference Data
2
Animal (mammals)
Bos taurus - Hereford
Select to download latest data for Bos taurus - Hereford
Canis familiaris
Select to download latest data for Canis familiaris
Equus caballus
Select to download latest data for Equus caballus
Homo sapiens - hg18
Select to download latest data for Homo sapiens - hg18
3
Homo sapiens - hg19
Select to download latest data for Homo sapiens - hg19
Homo sapiens - hg38
Select to download latest data for Homo sapiens - hg38
Mus musculus
Select to download latest data for Mus musculus
Rattus norvegicus
Select to download latest data for Rattus norvegicus
Sus scrofa
Select to download latest data for Sus scrofa
Animal (others)
4
5
Download Genome
Cancel
Download
6
Help
Close

Manage Reference Data: Locally
Free space in CLC_References location: 58.82 GB
Free space in temporary folder location: 58.81 GB

Homo sapiens - hg19

Download genome
Select how to get reference sequence:
● Download genome sequence
○ Use existing genome sequence track

Download size for selection: 865.5 MB

□ Name	Version	Provider	Size
☑ Sequence	75	Ensembl	828.0 MB
☑ Genome Annotations	75	Ensembl	37.5 MB
□ Chromosome bands (ideogram)		UCSC	6 KB
□ Dbsnp (common) variants	138	UCSC	620.7 MB
□ Dbsnp variants	138	UCSC	1.78 GB
□ Clinical variants in dbSNP		NCBI	16.1 MB
□ HapMap Variants	75	Ensembl	513.3 MB
□ 1000genomes	phase1	Ensembl	1.53 GB

Figure 1: *Open the Reference Data Manager to download the relevant reference databases.*

The CDS and reference sequence are now saved in the CLC_References | Genomes | Homo_sapiens_hg19 folder accessible from the Navigation Area.

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.

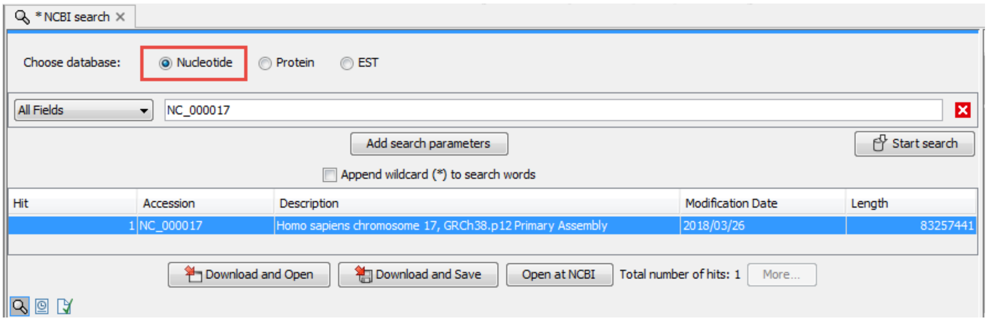
Method two: Using other sequence data sources

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. Make sure you have selected the correct database by selecting the radio button in front of "Nucleotide" (see red highlight in figure 2).

   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.
4. Click on the button labeled **Start search ()**.
5. You should find a single entry is returned to you - human chromosome 17. Select it and click **Download and Save**.



Choose database: ● Nucleotide ○ Protein ○ EST

All Fields: NC_000017

☐ Append wildcard (*) to search words

Hit	Accession	Description	Modification Date	Length
1	NC_000017	Homo sapiens chromosome 17, GRCh38.p12 Primary Assembly	2018/03/26	83257441

Total number of hits: 1

Figure 2: *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 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.

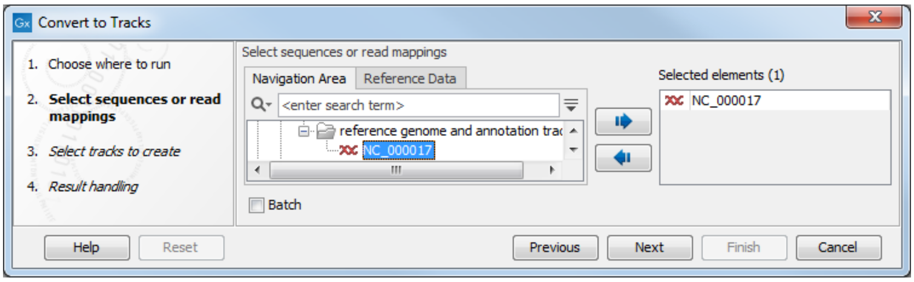
Creating tracks Once you have your sequence data available in track format, it can be used in a track list, which can then allow you to view sets of data that all refer to the same reference in a single view. Here, we will create a reference genome track from the chromosome sequence we just downloaded. 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** ( ) | **Track Conversion** ( ) | **Convert to Tracks** ( )

   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.

2. Choose the sequence you just downloaded `NC\_000017` (figure 3)



Convert to Tracks

1. Choose where to run
2. **Select sequences or read mappings**
3. *Select tracks to create*
4. *Result handling*

Select sequences or read mappings

Navigation Area | Reference Data

<enter search term>

reference genome and annotation trac

NC_000017

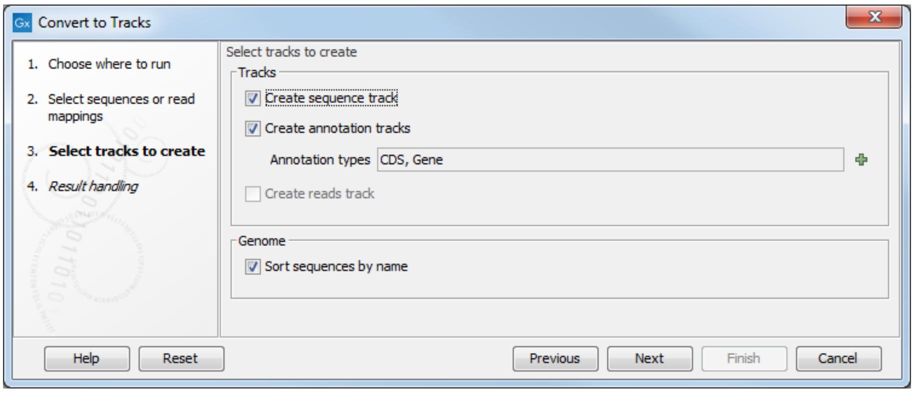
Selected elements (1)

NC_000017

□ Batch

Figure 3: *Select the human chromosome 17.*

3. In the next dialog, "Create sequence track" and "Create annotation tracks" are selected by default. Click on the green plus sign and choose the annotation types **CDS** and **Gene** (figure 4).



Convert to Tracks

1. Choose where to run
2. Select sequences or read mappings
3. **Select tracks to create**
4. *Result handling*

Select tracks to create

Tracks

☑ Create sequence track

☑ Create annotation tracks

Annotation types: CDS, Gene

□ Create reads track

Genome

☑ Sort sequences by name

Figure 4: *When using the tool Convert to tracks you can choose which of the annotation types found should be converted to tracks.*

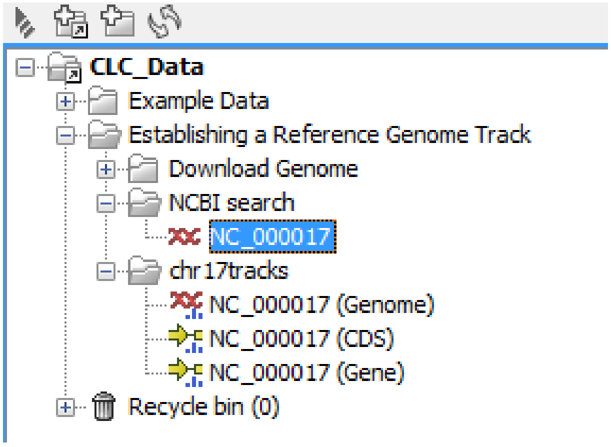
4. Choose to **Save** the results to a folder you can call **chr17tracks**. Click **Finished**.

You should now have three track objects (figure 5). 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.



CLC_Data
Example Data
Establishing a Reference Genome Track
Download Genome
NCBI search
NC_000017
chr17tracks
NC_000017 (Genome)
NC_000017 (CDS)
NC_000017 (Gene)
Recycle bin (0)

Figure 5: *The Navigation Area after converting the reference to tracks.*

Creating a track list

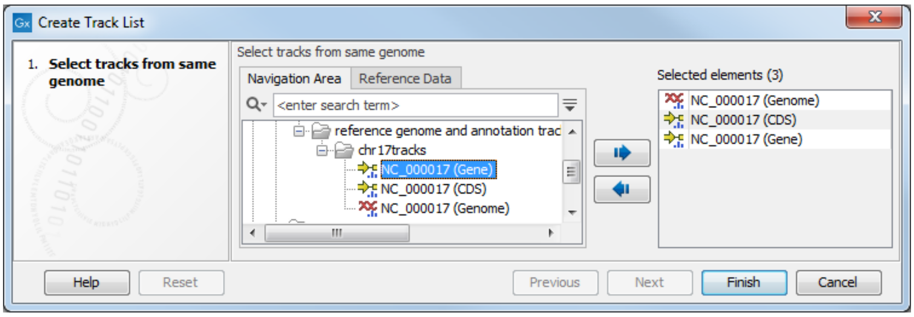
1. Go to:

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

2. Select the tracks you wish to include in your track list. There are two conditions that must be met:

   - A reference sequence track must be one of the members of the track list.
   - 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 6) here are for the data in the second part of the tutorial.



Create Track List
1. Select tracks from same genome
Select tracks from same genome
Navigation Area | Reference Data
<enter search term>
reference genome and annotation trac
chr17tracks
NC_000017 (Gene)
NC_000017 (CDS)
NC_000017 (Genome)
Selected elements (3)
NC_000017 (Genome)
NC_000017 (CDS)
NC_000017 (Gene)

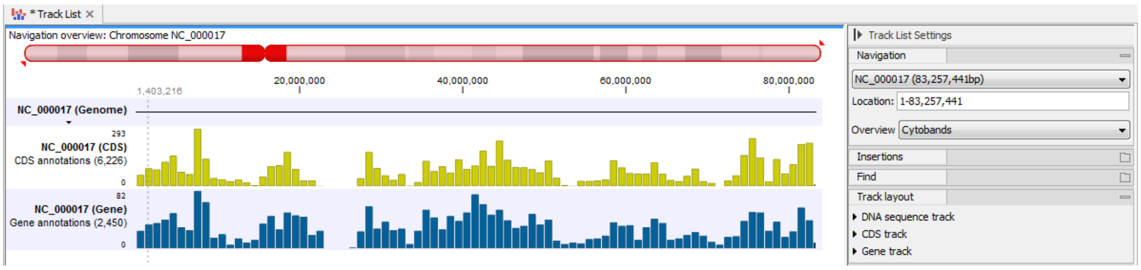
Figure 6: *Selecting tracks to be included in a track list.*

3. Click on **Finished**.

Your track list should then open up in the workspace, see figure 7.

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:

- Right click on the tab at the top of the new view, and choose **Save As...** from the menu that appears, or



NC_000017 (83,257,441bp)
Location: 1-83,257,441
Overview Cytobands

Figure 7: *Viewing the chromosome 17 track list.*

- Drag the tab over into a folder in the **Navigation Area** of the Workbench.

You can easily add additional annotation tracks to an existing track list by drag-and-drop the new track into the open Track List view.

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 for chromosome 17 as shown in figure 8.

    - 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. Click **get output** to download 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.

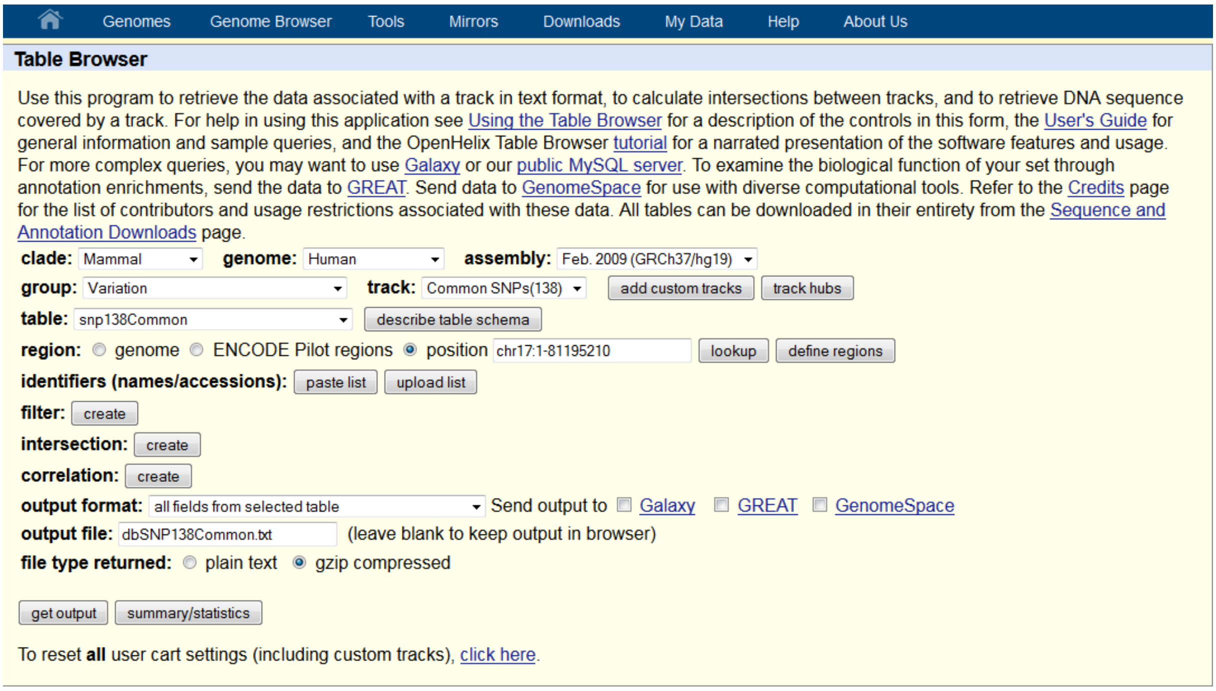


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Variation **track:** Common SNPs(138) add custom tracks track hubs

table: snp138Common describe table schema

region: ○ genome ○ ENCODE Pilot regions ● position chr17:1-81195210 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Send output to □ Galaxy □ GREAT □ GenomeSpace

output file: dbSNP138Common.txt (leave blank to keep output in browser)

file type returned: ○ plain text ● gzip compressed

get output summary/statistics

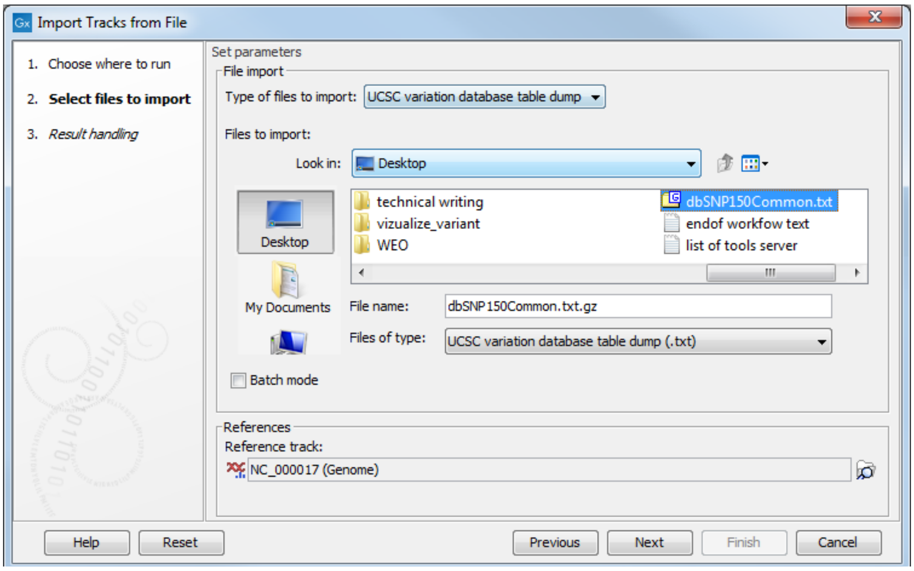
To reset **all** user cart settings (including custom tracks), click here.

Figure 8: *Search terms for the UCSC table browser window.*

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

   Import ( ) |Tracks ( )

2. Fill in the wizard window as seen in figure 9. The "Type of files to import" should be switched to "UCSC Variation database table dump". Find and select the annotation file that you downloaded and under "Reference track:" select the chromosome 17 track `NC\_000017 (Genome)`.



Set parameters

File import

Type of files to import: UCSC variation database table dump

Files to import:

Look in: Desktop

File name: dbSNP150Common.txt.gz

Files of type: UCSC variation database table dump (.txt)

□ Batch mode

References

Reference track: NC_000017 (Genome)

Figure 9: *Importing dbSNP common annotations.*

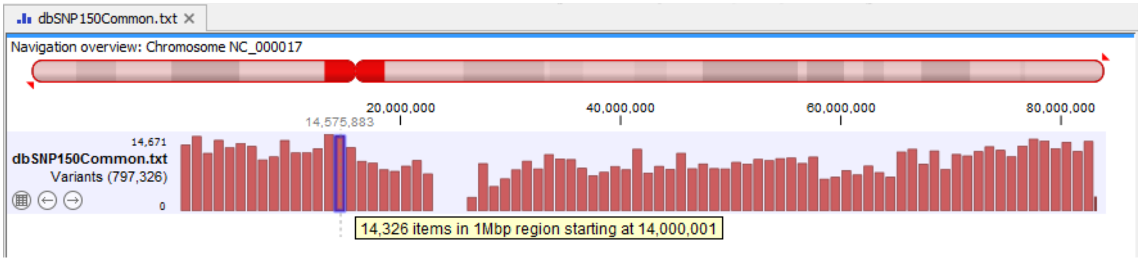
3. Choose to **save** the track and click **Finish**.

Importing can take some time, depending on the size of your annotation file. Once imported, consider to rename the file to a more appropriate name if needed.

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 10).

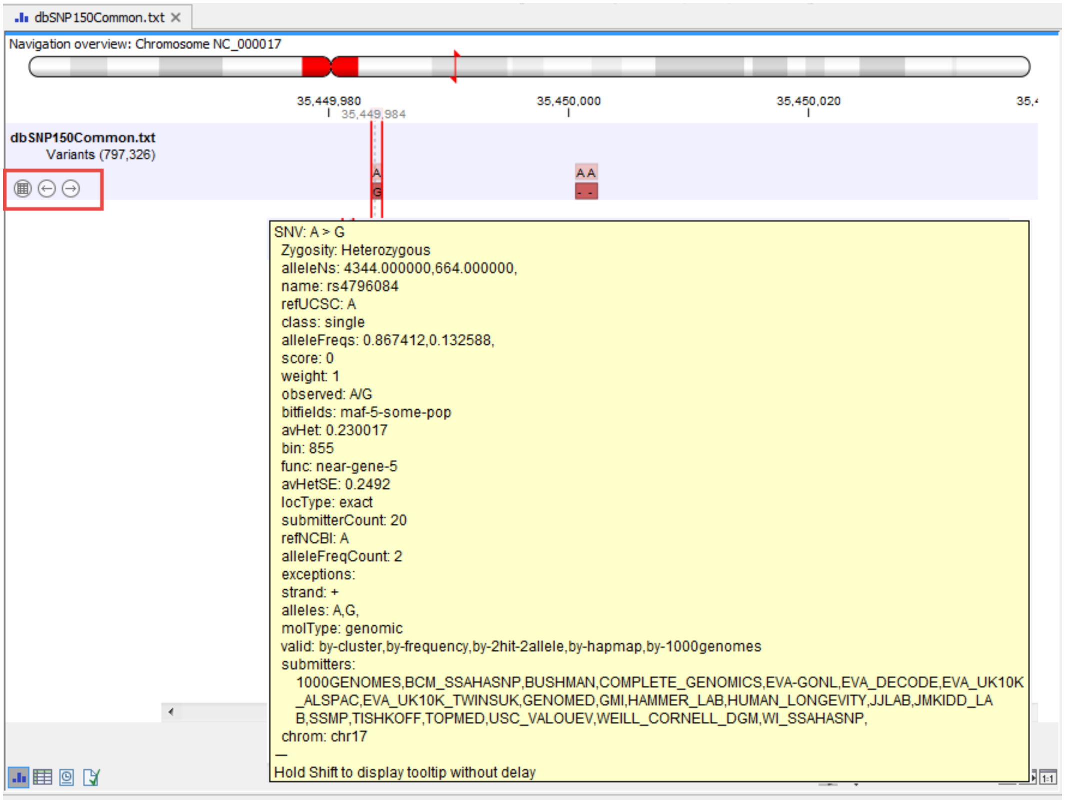


14,326 items in 1Mbp region starting at 14,000,001

Figure 10: *Open the commondbSNP_chr17 track to have a look.*

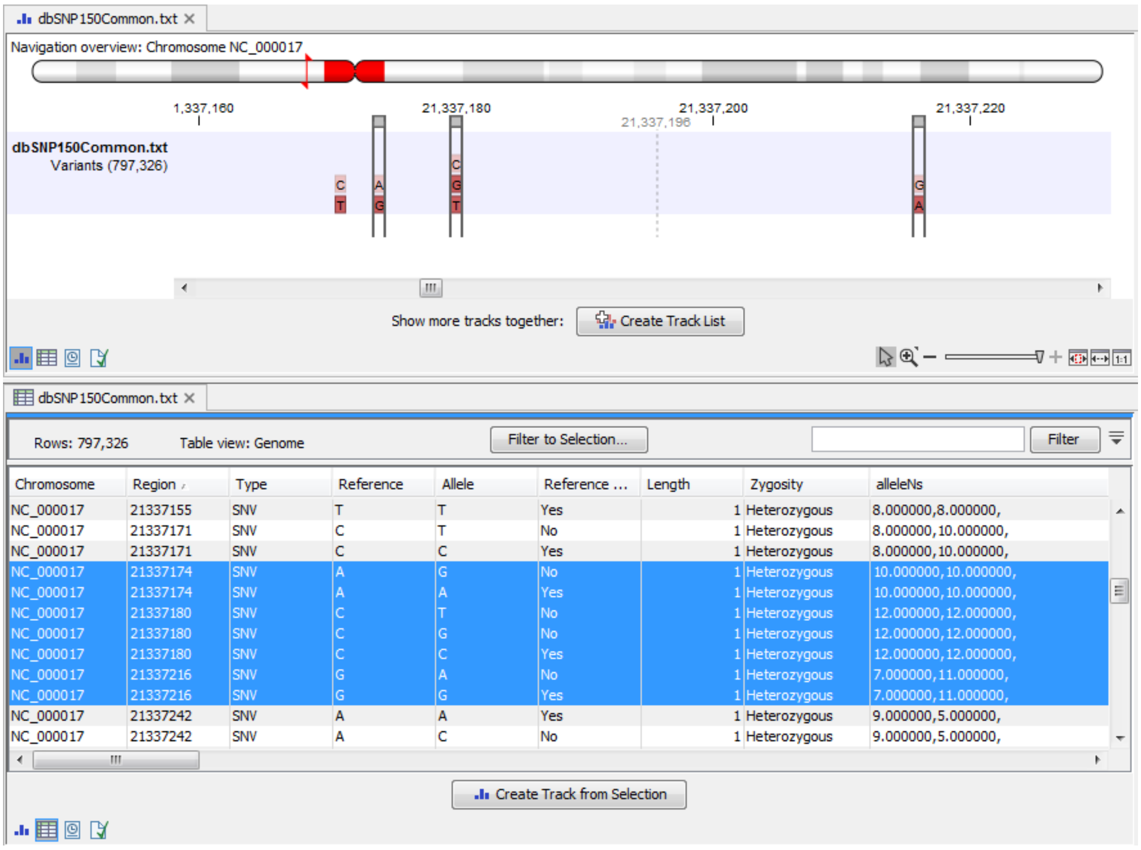
3. Zoom in until each SNP is visible at the nucleotide level. You can move from one SNP to the next with the
4. Mouse over one such bar. Now the pop-up window will display the information available for that specific SNP (see figure 11). You can move from one annotation to the next using the arrow under the track name.
5. You can also view the annotation track in table view clicking on the table icon under the track name. Track and table are linked, which means that selecting a row in the table highlights the specific variant in the track view with a vertical line (figure 12).

   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.



SNV: A > G
Zygosity: Heterozygous
alleleNs: 4344.000000,664.000000,
name: rs4796084
refUCSC: A
class: single
alleleFreqs: 0.867412,0.132588,
score: 0
weight: 1
observed: A/G
bitfields: maf-5-some-pop
avHet: 0.230017
bin: 855
func: near-gene-5
avHetSE: 0.2492
locType: exact
submitterCount: 20
refNCBI: A
alleleFreqCount: 2
exceptions:
strand: +
alleles: A,G,
molType: genomic
valid: by-cluster,by-frequency,by-2hit-2allele,by-hapmap,by-1000genomes
submitters:
1000GENOMES,BCM_SSAHASNP,BUSHMAN,COMPLETE_GENOMICS,EVA-GONL,EVA_DECODE,EVA_UK10K_ALSPAC,EVA_UK10K_TWINSUK,GENOMED,GMI,HAMMER_LAB,HUMAN_LONGEVITY,JJLAB,JMKIDD_LAB,SSMP,TISHKOFF,TOPMED,USC_VALOUEV,WEILL_CORNELL_DGM,WI_SSAHASNP,
chrom: chr17
—
Hold Shift to display tooltip without delay

Figure 11: *Zooming in, individual SNPs are represented as bars. Mousing over results in a pop-up box with information about the SNP.*



Rows: 797,326   Table view: Genome

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	alleleNs
NC_000017	21337155	SNV	T	T	Yes	1	Heterozygous	8.000000,8.000000,
NC_000017	21337171	SNV	C	T	No	1	Heterozygous	8.000000,10.000000,
NC_000017	21337171	SNV	C	C	Yes	1	Heterozygous	8.000000,10.000000,
NC_000017	21337174	SNV	A	G	No	1	Heterozygous	10.000000,10.000000,
NC_000017	21337174	SNV	A	A	Yes	1	Heterozygous	10.000000,10.000000,
NC_000017	21337180	SNV	C	T	No	1	Heterozygous	12.000000,12.000000,
NC_000017	21337180	SNV	C	G	No	1	Heterozygous	12.000000,12.000000,
NC_000017	21337180	SNV	C	C	Yes	1	Heterozygous	12.000000,12.000000,
NC_000017	21337216	SNV	G	A	No	1	Heterozygous	7.000000,11.000000,
NC_000017	21337216	SNV	G	G	Yes	1	Heterozygous	7.000000,11.000000,
NC_000017	21337242	SNV	A	A	Yes	1	Heterozygous	9.000000,5.000000,
NC_000017	21337242	SNV	A	C	No	1	Heterozygous	9.000000,5.000000,

Figure 12: *Track and table split view of the annotations. The two views are linked.*