



Genomics Gateway Plug-in ^{Beta}

User Manual

User manual for Genomics Gateway Plug-in 2.0 beta

Windows, Mac OS X and Linux

February 13, 2012

This software is for research purposes only.

CLC bio
Finlandsgade 10-12
DK-8200 Aarhus N
Denmark



Contents

1	Introduction to the Genomics Gateway Plug-in	5
1.1	Basic data structure	5
1.2	Upgrading from version 1.X of the Genomics Gateway plug-in	6
2	Building a reference genome	7
2.1	Define reference genome	7
2.2	Download annotations from Ensembl	8
2.3	Import tracks from file	9
3	Mapping reads to the reference genome and calling SNPs	12
3.1	Using an existing mapping file to create a mapping track	13
3.2	SNP and DIP detection	14
3.3	New tools for mapping and variant detection	14
4	Visualization: the genome browser	15
4.1	Zooming and customizing the layout of the track	15
4.2	Adding, removing and reordering tracks	17
4.3	Showing a track in a table	18
4.4	Finding a gene or a position on the genome	18
5	Annotate and filter tracks	20
5.1	Annotate from overlapping annotations	20
5.2	Exon number annotation	20
5.3	Flanking sequence	20
5.4	Gene link annotation	20
5.5	Name filter	20

5.6	Variation frequency filter	21
5.7	Overlap filter	22
6	Comparison of variation data	23
6.1	Find common variations in group	23
6.2	Fisher exact test	23
6.3	Filter against control reads	24
6.4	Database variation filter	25
6.5	Variation haplotype compare filter	26
6.6	Copying and merging tracks	26
7	Functional consequences	27
7.1	Amino acid changes	27
7.2	Splice site effect prediction	27
7.3	GO enrichment analysis	28
7.4	Conservation score annotation	29
8	Future improvements and feedback	30
9	Installation of the Genomics Gateway Plug-in	31
10	Uninstall	33
	Index	33
	Bibliography	35

Chapter 1

Introduction to the Genomics Gateway Plug-in

The Genomics Gateway Plug-in 2.0 beta is a beta version of the Genomics Gateway which will become an integrated part of the *CLC Genomics Workbench* once it has evolved and stabilized. Section 8 lists the future improvements that we already know will become part of the coming development.

The idea behind the Genomics Gateway is to provide a visualization, comparison and analysis framework for genome-scale studies such as whole-genome or exome resequencing projects, transcriptome sequencing, ChIP-Seq etc.



This user manual will describe the basic concepts of the Genomics Gateway but will not go into detail about the specifics. Since this is in a beta stage, we expect that there will be a number of changes to the design before we reach the final release. The explanation of the features below will be focusing on a work flow like this:

- Define and build a reference genome
- Map reads to the reference genome
- Identify variants in the read mapping
- Compare the variants identified in the sample sequence to variants in public databases like dbSNP and COSMIC
- Filter the variants and compare with known annotations (e.g. gene annotations or regulatory regions)

In addition we will explain how to compare two or more sets of variants identified in different samples.

1.1 Basic data structure

All information in the Genomics Gateway is organized into tracks. All information that can be tied to a genomic coordinate is represented as tracks: a reference genome sequence, a set of genes, a coverage graph, a read mapping or variants from variant calling.

Tracks are saved as files in the **Navigation Area** and they have icons to represent their type, e.g. an annotation track (). In order to visualize several tracks together, you create a **Track List** ():

File | New | Track List ()

Alternatively, there is a button when you open a track that can be used for creating a track list. The track list does not contain any of the data which still resides in the individual tracks that are saved separately. This means that you can use the same track in many different contexts by creating separate track lists pointing to the data.

Once created, tracks can be added to the track list by simple dragging from the **Navigation Area** and into the list.

The visualization and management of the track list is described in section [4](#).

1.2 Upgrading from version 1.X of the Genomics Gateway plug-in

With version 2.0 of the Genomics Gateway plug-in, the data structure changed so that there is no longer track sets but only tracks and track lists (the latter only include references to tracks).

This means that any track sets created in previous versions need to be converted in order to be used with the new version. There is a special tool for doing this:

Toolbox | Genomics Gateway | Convert Old Genomics Gateway Data

Select a folder containing track sets from the previous version. All your track sets will then be converted into separate tracks that are saved in the folders where the track sets were located. For re-creating the visual representation of the tracks together in the track set, create a new track list (see section [1.1](#)).

Chapter 2


Building a reference genome

In later versions of the Genomics Gateway it will be possible to connect to more public genomic databases directly from the Workbench in order to create a reference genome with a few clicks without worrying about file formats, versions etc. This has partly been accomplished with the integration with Ensembl as described in section 2.2, but you will still in this beta version have to import sequence files or use already imported data to create a reference genome.

A reference genome is a collection of *tracks*. A track is the basic building block of all data in the Genomics Gateway. It can be a set of annotations (gene annotations, variants from dbSNP, experimentally derived SNPs etc.), it can be the genomic reference sequence, it can be a track containing reads that have been mapped to a reference, or it can be a graph track showing e.g. the percentage of non-specific matches for a track of mapped reads. Note that all these tracks have one thing in common: they are defined by having a position in the genome coordinate system (i.e. chromosome and position).


There are three tools that can be used for building a reference genome:

2.1 Define reference genome

This tool will take a sequence that has already been imported into the *CLC Genomics Workbench* and use that to build a reference genome. You can download genomic sequences from e.g. Ensembl (in Genbank format from <http://www.ensembl.org/info/data/ftp/index.html> or use the integrated **Search**  tool in the Workbench to download sequences from Genbank. For use with the built-in tool for annotating (see section 2.2), we recommend downloading the fasta files from the Ensembl ftp site. Once downloaded and imported, you can define these sequences as your reference genome:

Toolbox | Genomics Gateway | Define Reference Genome

In the first dialog shown, select the sequences for your genome (e.g. for humans you would select all the chromosomes) and click **Next**.

In the next dialog, choose which kind of annotation type you want to include in the reference genome by clicking the  button (see figure 2.1). Note that the sequence you imported has to include annotations in order to be able to select this (if you use a fasta file it will have no annotations - if you use a GenBank or EMBL file it will include annotations).

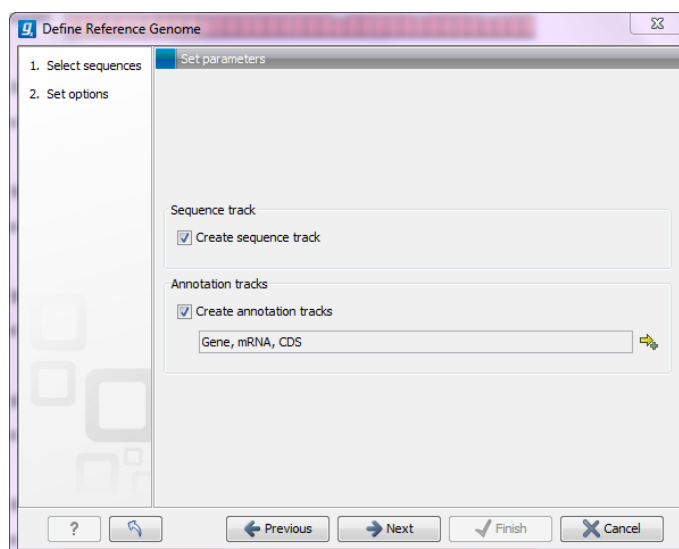


Figure 2.1: Define the annotation types to be included in the reference genome.

When you click **Finish**, a track is created for the reference sequence (if selected) and each of the selected annotation types.

2.2 Download annotations from Ensembl

Once you have converted a set of sequences to tracks, you can annotate this directly in the Workbench. We recommend using the fasta files on the Ensembl ftp sites to be sure the right name and version is used:

- Model organisms: <http://www.ensembl.org/info/data/ftp/index.html>
- Bacteria: <http://bacteria.ensembl.org/info/data/ftp/index.html>
- Fungi: <http://fungi.ensembl.org/info/data/ftp/index.html>
- Metazoa: <http://metazoa.ensembl.org/info/data/ftp/index.html>
- Plants: <http://plants.ensembl.org/info/data/ftp/index.html>

The types of annotation that can be retrieved from Ensembl depends on the organism you choose. The example below is based on *Homo sapiens* which is where you find the most elaborate information.

Toolbox | Genomics Gateway | Download Ensembl Annotations

This will display a dialog where you have to choose an existing track as shown in figure 2.2.

This has to be created using the define reference genome tool (see section 2.1).

Click **Next** and you can define organism and what kind of annotations you wish to download (see figure 2.3).

In the example of *Homo sapiens*, you can select both genes, transcripts, coding regions and variation annotations. If variation annotations is selected, clicking **Next** will display the choices that are available as shown in figure 2.4.

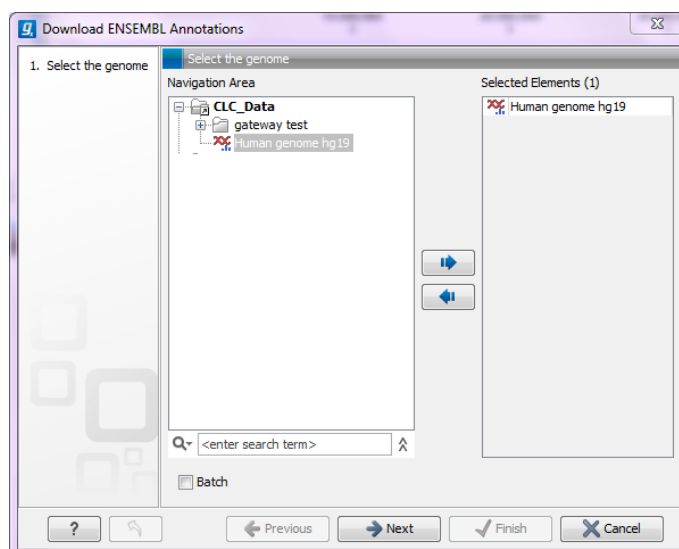


Figure 2.2: Define the reference genome.

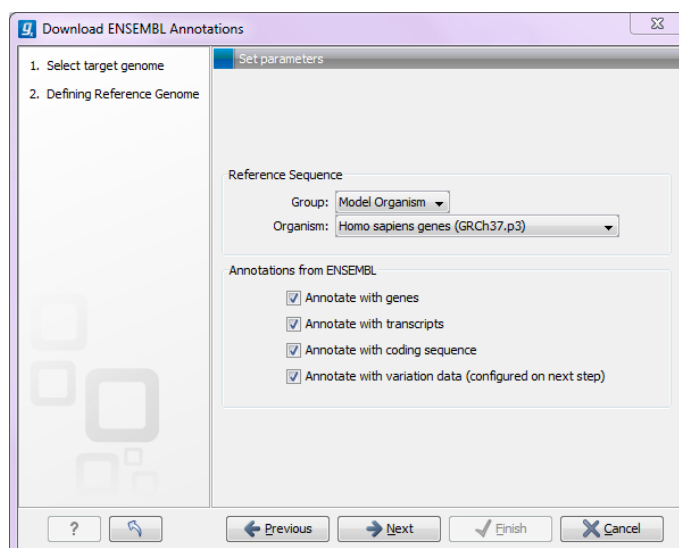


Figure 2.3: Select organism and types of annotations.

At the top, you can choose to get variation annotations from the COSMIC database [Forbes et al., 2008]. Below, you find different subsets of the dbSNP database [Sherry et al., 2001] which can be selected by pressing Ctrl (⌘ on Mac) while you select with the mouse.

Once you have clicked **Next** and **Finish**, the download process will start. This may take a while depending on the number of annotations (this also means that downloading the variation annotations takes significantly longer than genes, transcripts and coding genes due to more annotations). The results are stored in separate tracks for the different kinds of annotations.

2.3 Import tracks from file

In this first version of the Genomics Gateway you can also make use of information from public databases by downloading the data as raw data files to your computer and then import these files into the *CLC Genomics Workbench*.

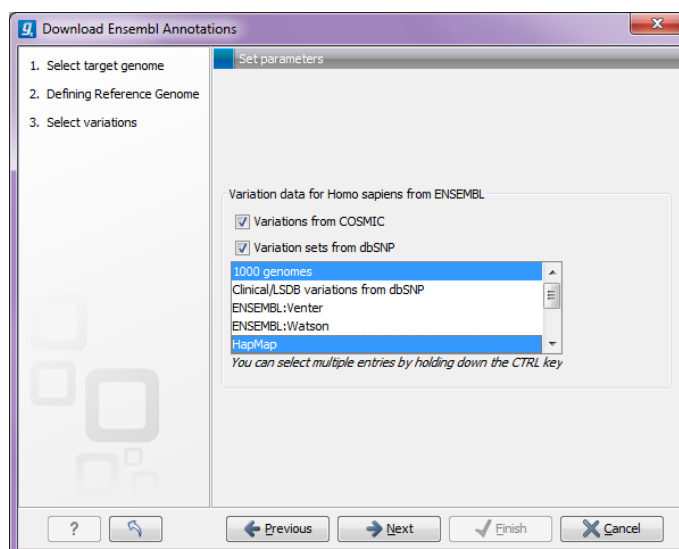


Figure 2.4: Select variation sources.

The formats currently accepted are:

GFF/GTF/GVF Annotations in `gff/gtf/gvf` formats. This is explained in detail in the user manual for another plug-in: <http://www.clcbio.com/annotate-with-gff>. In the context of the Genomics Gateway, this can be particularly useful for downloading gene and transcript annotations in `gtf` format and variation data in `gvf` format from Ensembl: <http://www.ensembl.org/info/data/ftp/index.html>.

VCF This is the file format used for variation by the 1000 Genomes Project. Read how to access data at <http://www.1000genomes.org/data#DataAccess>

BED Simple format for annotations. Read more at <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>.

Complete Genomics master var file This is the file format used by Complete Genomics for all kinds of variation data and can be used to analyze and visualize the variant calls made by Complete Genomics. Please note that you can import evidence files with the read alignments into the *CLC Genomics Workbench* as well (refer to the Complete Genomics import section of the Workbench user manual).

UCSC Variation database table dump This is mainly intended to allow you to import the popular *Common SNPs* variation set from UCSC. The file can be downloaded from the UCSC web site here: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/snp132Common.txt.gz>. Other sets of variation annotation can also be downloaded in this format. The files ending with `.txt.gz` on this list can be used: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/>.

Conservation scores This will accept files in fixed Wiggle format as they can be downloaded for example from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/primates/>.

All the data above is annotation data and if the file includes information about allele variations (like VCF, Complete Genomics and GVF), it will be combined into one **Variation** track that can be

used for finding known variants in your experimental data. When the data cannot be recognized as variation data, one track is created for each annotation type.

To import these annotation files, you need first to define your reference genome based on the chromosome sequences as explained in section 2.1. Once you have created the reference genome, you can add the additional annotation information:

Toolbox | Genomics Gateway | Import Tracks from File

The first dialog (shown in figure 2.5) allows you to select the target genome that the annotations should be added to.

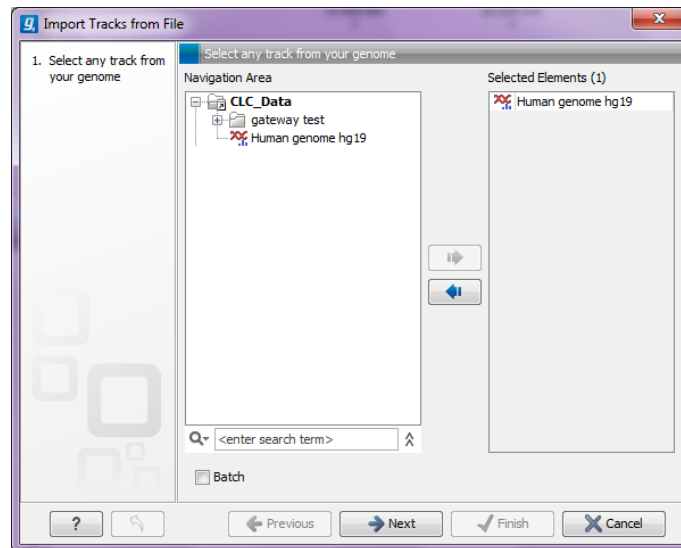


Figure 2.5: Define the reference genome.

Click **Next** to select annotation file. Zipped files are also supported.

Chapter 3

Mapping reads to the reference genome and calling SNPs

Once your reference genome has been set up, you can proceed to the next step which is to analyze your sequencing reads. The first step is to map them to your reference genome:

Toolbox | Genomics Gateway | Map Reads to Genome

This will open a dialog that will allow you to select your sequencing reads and click **Next**.

The next dialog lets you select the reference genome (see section 2 to see how to build a reference genome) as shown in figure 3.1).

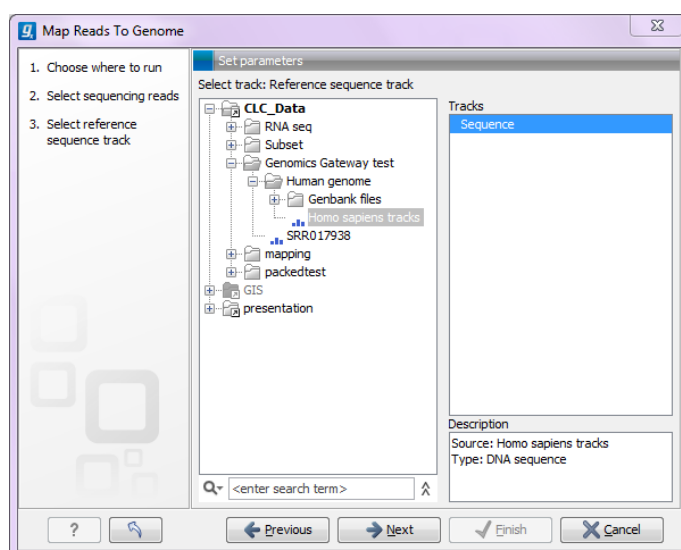


Figure 3.1: Define the reference genome.

Clicking **Next** shows the parameters for the mapping. These are described in the main user manual for the *CLC Genomics Workbench*. Clicking **Next** allows you to specify the output options as shown in figure 3.1.

The main result of this algorithm is a new track containing the reads that have been mapped to the reference genome.

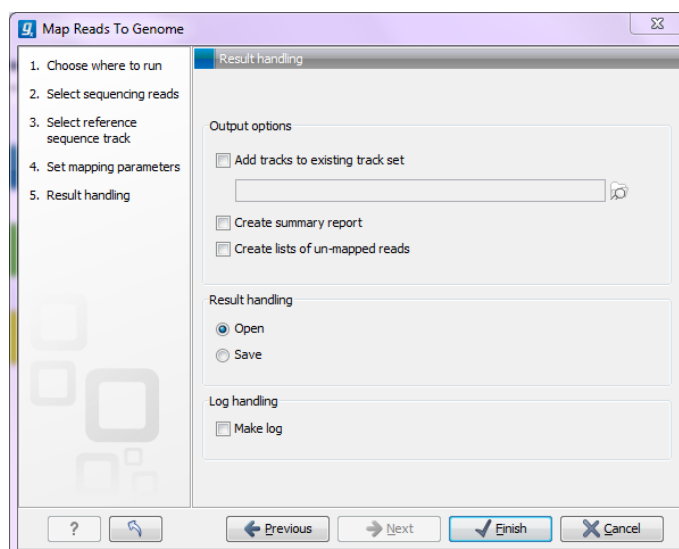


Figure 3.2: Specify output options.

Like in the standard read mapping in the workbench, you can create a **mapping report** and a list of **unmapped reads** as part of the output.

3.1 Using an existing mapping file to create a mapping track

In case you have data that has already been mapped with the standard Workbench mapping algorithm, you can convert this to a track:

Toolbox | Genomics Gateway | Create Tracks from Read Mapping

In the dialog shown (figure 3.3), select a read mapping (🇺🇸) (🇺🇸).

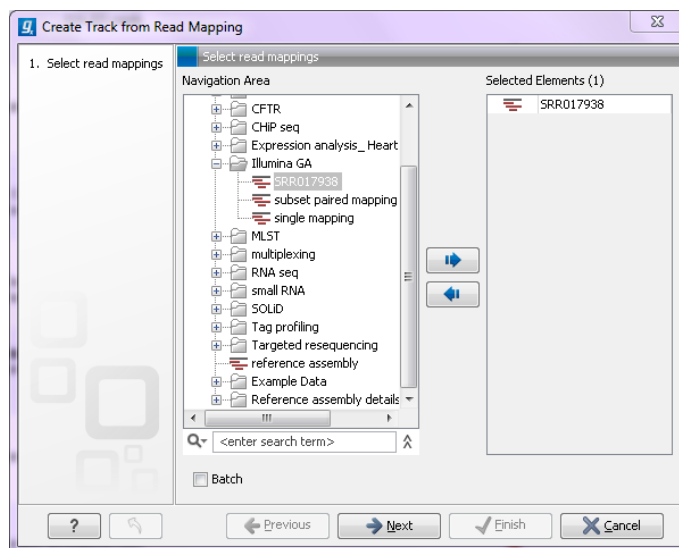


Figure 3.3: Select a read mapping

Clicking **Next** allows you to specify how the new track containing the mapped reads should be saved (see figure 3.4).

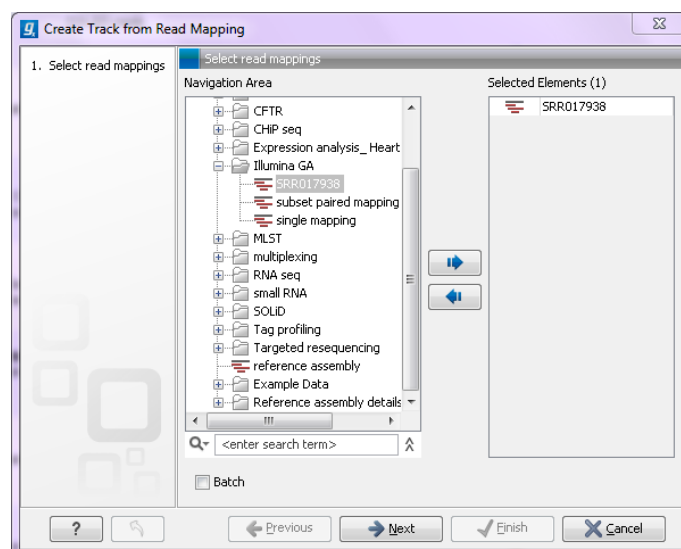


Figure 3.4: Select output options

3.2 SNP and DIP detection

The SNP and DIP detection are identical to the corresponding versions in the standard High-throughput Sequencing folder of the Toolbox. These are described in the main Workbench manual.

There are three changes compared to the original versions: The first is that the input data is now a read mapping track that you select in the first dialog displayed. The second is that the output is a track.

Please note that the information in the track is more simple than the SNP and DIP annotations of the standard workbench tables and annotations. To enrich this information, use the annotation tools as explained in section 5.

3.3 New tools for mapping and variant detection

We have several new and improved algorithms for both read mapping and variant detection that you can download as plug-ins. These include:

- A probabilistic variant caller to find SNVs and small InDels
- A structural variation detection tool to find structural variants
- A new read mapper adding a new level of speed to read mapping

All of them will have appropriate items in the Genomics Gateway part of the Toolbox once installed.

Chapter 4

Visualization: the genome browser

Figure 4.1 shows an example of a track list including a track with mapped reads at the top, followed by a SNP detection track. Below are two tracks from the reference genome: the genomic sequence and CDS annotations.

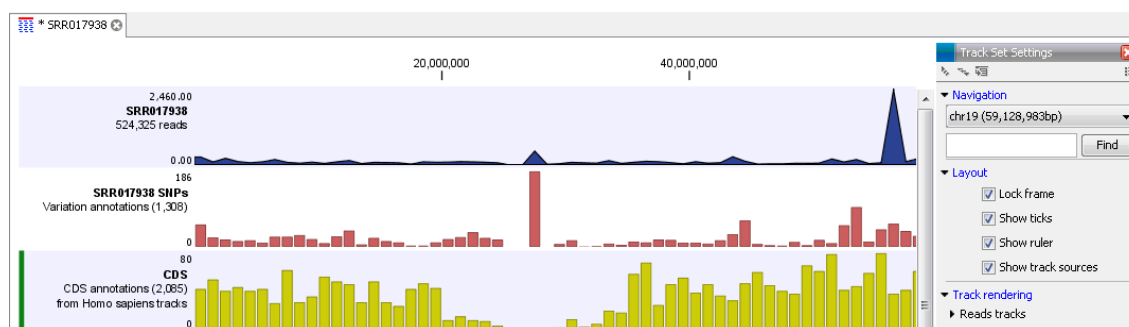


Figure 4.1: Four tracks shown in the genome browser

4.1 Zooming and customizing the layout of the track

Zooming in and out on the view shown in 4.1 is done either through the zoom tools in the right-hand corner of the Toolbar, using the + and – keys on the keyboard or by using a mouse scroll wheel or gesture while pressing the Ctrl (⌘ on Mac) key.

When zooming in and out you will see that the data is visualized in an aggregated format using a density bar plot or graph when zoomed out. This allows you to navigate the view more smoothly and get an overview of e.g. how many SNPs that are located in a certain region.

In figure 4.2 we have zoomed in on a specific region, and you can see that the read track at the top is now showing the individual reads and the CDS and SNP annotations are shown in full detail as well.

Zooming in even further will also display the alignment of the reads so you can see the reference sequence and the reads at a nucleotide-level resolution (see figure 4.3).

In this case, we can only see three reads so it makes sense to adjust the height of the reads track. This is done by simply dragging with the mouse at the bottom of the track. In that way you can see more reads as shown in figure 4.4.

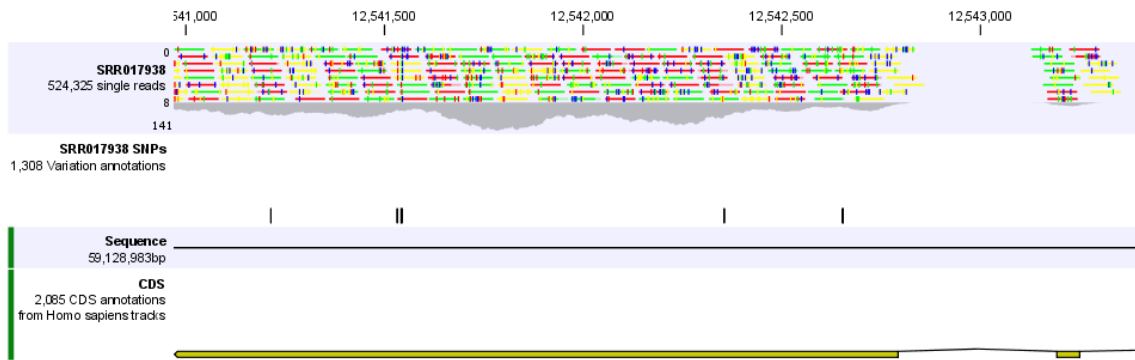


Figure 4.2: Zooming in reveals more detail on each track

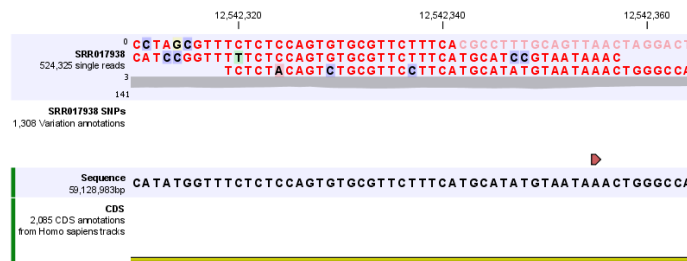


Figure 4.3: Zooming all the way in shows the actual bases of the reads and the reference sequence.

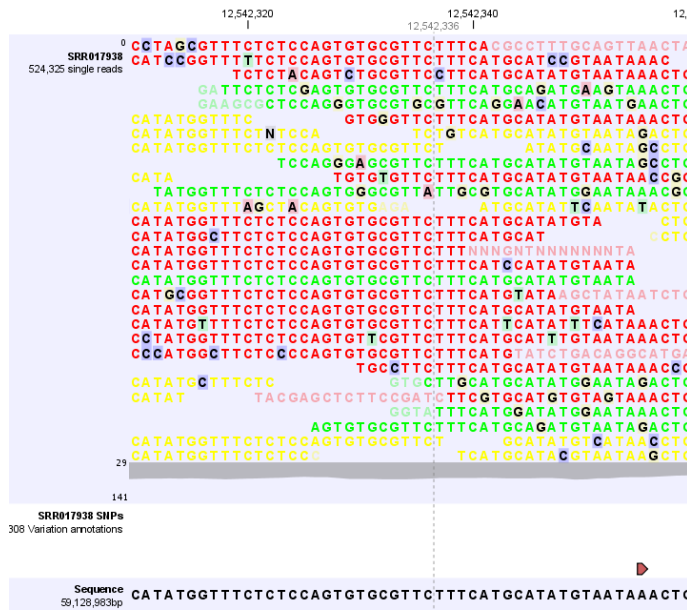


Figure 4.4: Adjusting the height of the track.

In the **Side Panel** you can adjust when the information in the track should be aggregated or when it should be displayed in detail. Figure 4.5 shows the options for a read track and an annotation track.

The aggregation setting can be adjusted: with a low value the details will only be visible when zoomed in, and a high value means that you can see details even when zoomed out.

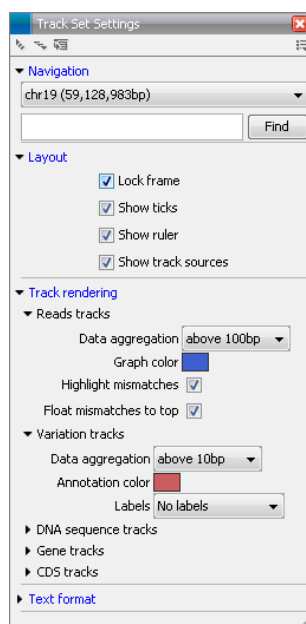


Figure 4.5: The Side Panel makes it possible to adjust the aggregation level.

4.2 Adding, removing and reordering tracks

You can organize your tracks by dragging them up and down, and right-clicking on any of the tracks as shown in figure 4.6 gives you several options:

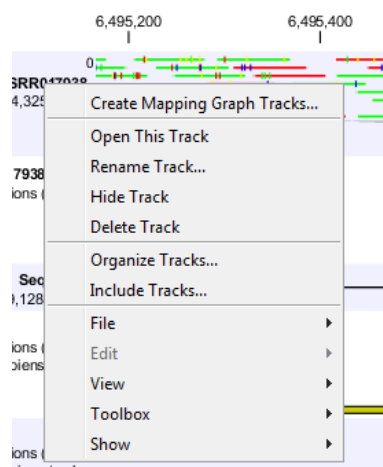


Figure 4.6: Options to organize tracks.

Create Mapping Graph Tracks This will allow you to create a new track from a mapping track with any of the following information (note this is only available when you right-click a read mapping track):

- Read coverage
- Non-specific read coverage
- Non-perfect read coverage
- Paired read coverage
- Broken pair coverage

- Paired read distance

Find in Navigation Area . This will select the track in the Navigation Area.

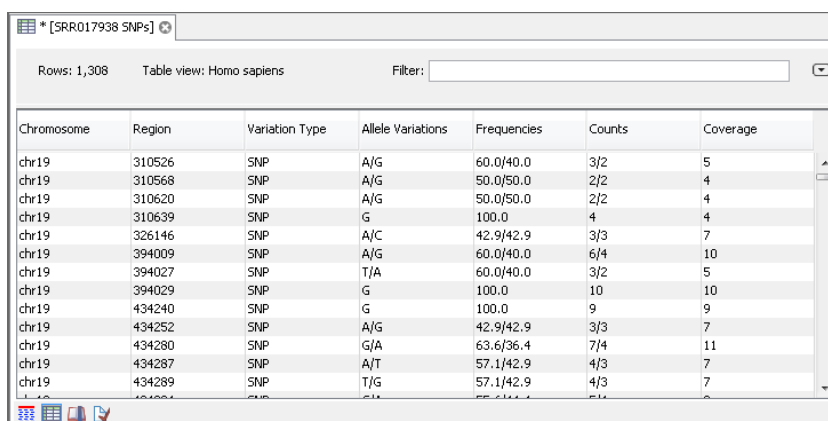
Open This Track This opens a table view of the track as described in section 4.3. This only applies to annotation tracks.

Remove Track This will remove the track from the current view. You can add it again by dragging it from the **Navigation Area** into the track list view or by pressing **Undo** (↶).

Include More Tracks This will allow you to add other track sets to your current track set. Please note that the information in the track will still be stored in its original track set. This means that including a track in this way is adding a reference to this track in another track set. This could be for example to include the SNP track from another sample to your current analysis. Once added, use the **Organize Tracks** option described below to select which of the tracks in the track set to show.

4.3 Showing a track in a table

All tracks containing annotations can be opened in a table. This is done either by double-clicking the label of the track or by right-clicking the track and choosing **Open This Track**. The table will have one row for each annotation, and the columns will reflect its information content. Figure 4.7 shows an example of a variation database track open in a table.



Chromosome	Region	Variation Type	Allele Variations	Frequencies	Counts	Coverage
chr19	310526	SNP	A/G	60.0/40.0	3/2	5
chr19	310568	SNP	A/G	50.0/50.0	2/2	4
chr19	310620	SNP	A/G	50.0/50.0	2/2	4
chr19	310639	SNP	G	100.0	4	4
chr19	326146	SNP	A/C	42.9/42.9	3/3	7
chr19	394009	SNP	A/G	60.0/40.0	6/4	10
chr19	394027	SNP	T/A	60.0/40.0	3/2	5
chr19	394029	SNP	G	100.0	10	10
chr19	434240	SNP	G	100.0	9	9
chr19	434252	SNP	A/G	42.9/42.9	3/3	7
chr19	434280	SNP	G/A	63.6/36.4	7/4	11
chr19	434287	SNP	A/T	57.1/42.9	4/3	7
chr19	434289	SNP	T/G	57.1/42.9	4/3	7

Figure 4.7: Showing a variation track in a table.

You can use the table to sort, filter and select annotations. Selecting a row in the table will cause the graphical view to jump to this position on the genome. Please note that the table filter only affects the table. The track itself keeps all the annotations. If you want to filter your track also in the graphical view, use the refiners instead (see the next chapter).

At the bottom of the table, there is a button to **Create Track from Selection**. By first selecting rows in the table, you can use this option to create a new track only including the subset of the annotations that you have selected. This is particularly useful in combination with the filter.

4.4 Finding a gene or a position on the genome

In the **Side Panel** at the top, there is a search field that will take you to the position on the genome that you are looking for. You can enter:

A position You can enter the position like this: `chr13:4550..10000`. This will lead you to the corresponding region on chromosome 13. If you just enter the position `4550..10000` it will find the position on the chromosome currently shown. If you want to find just a single nucleotide, simply enter `4550` or `chr13:4550`.

An annotation name This can be a gene name or any other annotation name, e.g. `DNM2` to find the `DNM2` gene and `rs78931249` will find the dbSNP annotation with this name. Note that only tracks currently shown will be searched and that only the name of the annotations will be searched (the name is what you can also display next to the annotation via an option in the **Side Panel**). Press `return` to search the next occurrence, if the first hit is not the right one.

The search will run through the genome and stop when it finds the first hit. Press `Enter` again to find the next hit.

Chapter 5

Annotate and filter tracks

This chapter lists a lot of simple annotation and filtering tools that can be applied on annotation tracks, typically for variants.

5.1 Annotate from overlapping annotations

This will create a copy of the track used as input and add information from overlapping annotations.

5.2 Exon number annotation

Given a track with mRNA annotations, a new track will be created in which variations are annotated with the numbering of the corresponding exon with numbered exons based on the transcript annotations in the input track.

5.3 Flanking sequence

This will add flanking sequence of both sides of an annotation. The user can decide the number of nucleotides to include (see figure 5.1).

You will also need to provide a sequence track that should be used for inferring the flanking sequence. Please note that the central position of the flanking sequence is taken from the reference sequence, not incorporating any of the variant alleles.

5.4 Gene link annotation

This will add information about gene names and hyper links to the corresponding GenBank and OMIM web sites.

5.5 Name filter

The name filter allows you to input a list of names to create a new track only with these names.

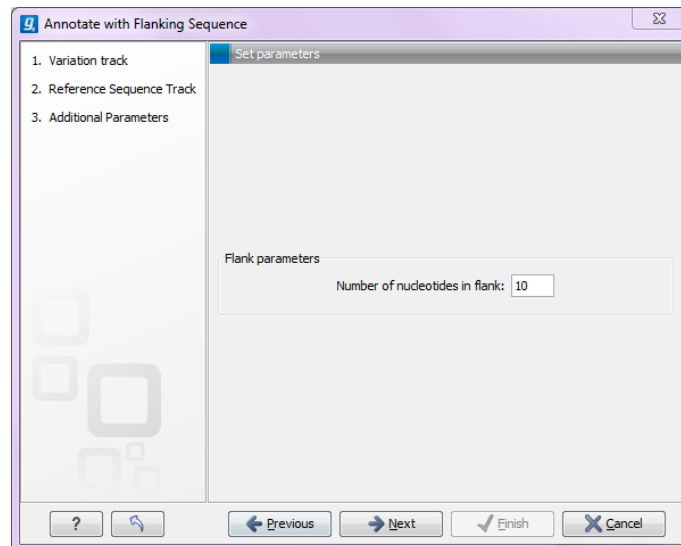


Figure 5.1: The flanking regions.

The intention is to support situations where only a subset of genes are to be analyzed. In this case, first create a new gene track by filtering the original gene track using the Name filter supplied with a list of the target genes. Next, use the overlap annotation filter on your variation tracks to exclude all variations that fall outside the target genes.

5.6 Variation frequency filter

This allows you to filter a variation track, so that only the variants that have a frequency above a user-defined threshold remain (see figure 5.2).

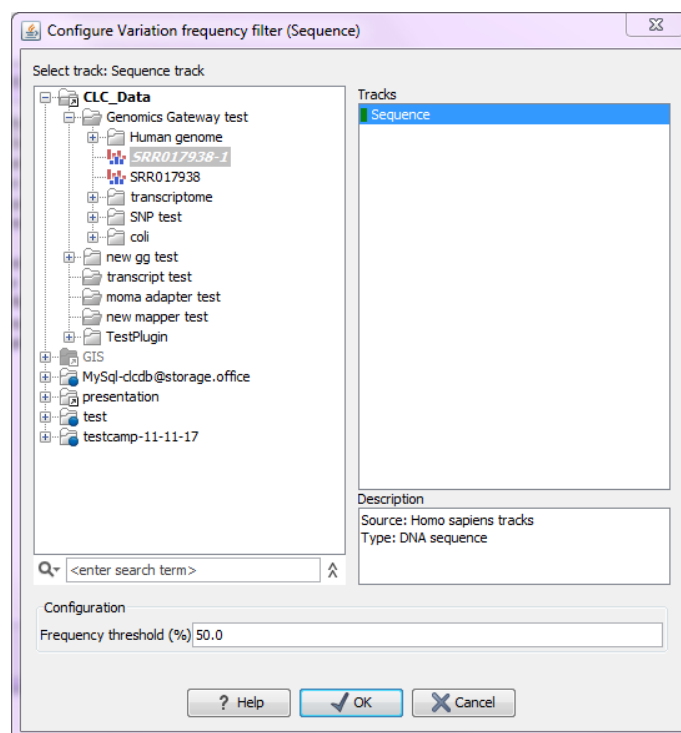


Figure 5.2: The variation frequency refiner.

Note that the filter only applies to the frequency of non-reference alleles.

5.7 Overlap filter

The overlap filter will be used for filtering an annotation track based on overlap with another annotation track. This can be used to e.g. only show variants that fall within genes or regulatory regions. Please note that for comparing variation tracks, the SNP haplotype compare filter (section 6.5) or the Known variation filter (section 6.4) should be used instead.

Chapter 6

Comparison of variation data

6.1 Find common variations in group

This tool should be used if you are interested in finding common (frequent) variants in a group of samples. For example one use case could be that you have 50 unrelated patients with the same disease and like to identify variations, which are present in at least 70% of all patients.

Furthermore, you have to specify a frequency threshold, which is the percentage of samples that should at least have the variant. Only variations over this threshold will be part of the output. Please note that each variant/allele is considered individually. Heterozygote variations are split into their alleles. Alleles equal to the reference sequence are not considered. The output is a variation file, which includes variations/alleles over this threshold with information about in how many samples and which samples they were found.

6.2 Fisher exact test

This tool should be used if you have a case-control study. This could be patients with a disease (case) and healthy patients (control). The idea is to identify variations which are more common in the case samples than in the control samples.

In the first step of the dialog, you select the case variant tracks. Clicking **Next** shows the dialog in figure 6.1.

Besides selecting a reference sequence track, this is also where the variation tracks from the control group should be added. Furthermore, you have to set a threshold for the p-value (default is 0.05). Only variations having a p-value below this threshold will be reported.

Each allele from each variation is considered separately. The Fisher exact test is applied on the number of occurrences of each variation/allele in the case and the control data set. Variations with a low p-value are potential candidates for variations playing a role in the disease/phenotype. Please note that a low p-value can only be reached if the number of samples in the data set is high.

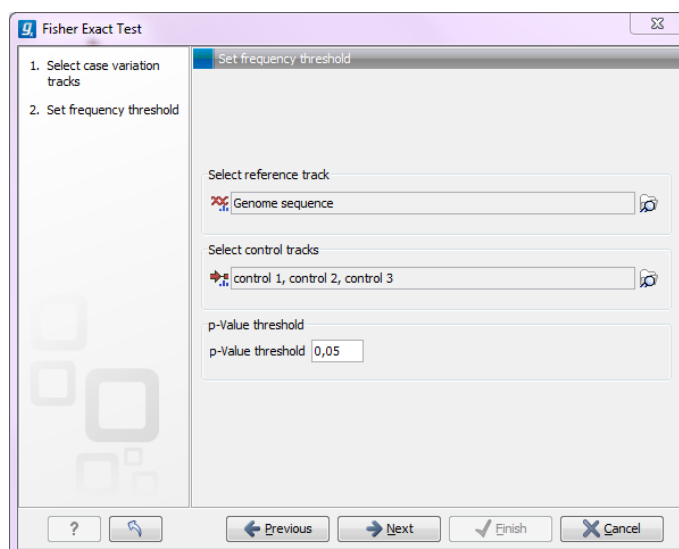


Figure 6.1: *The fisher exact test settings.*

6.3 Filter against control reads

The **Variation haplotype compare filter** described in section 6.5 can be used to filter down the number of variants in a two-sample case versus control experimental set-up. The remaining variants would be the ones only found in the case sample.

However, sometimes there will be false negatives variants in the control sample. This is often due to lack of coverage of the variant allele. In order to test if this is the case, the **Filter against control reads** tool can be used to verify that the variants are indeed negative in the control data set. This means that for this particular scenario, the variation haplotype compare filter does not need to be used.

The Filter against control reads need the variation track from the case sample as input and when you click **Next** you will need to provide the read track from the control data set (see figure 6.2).

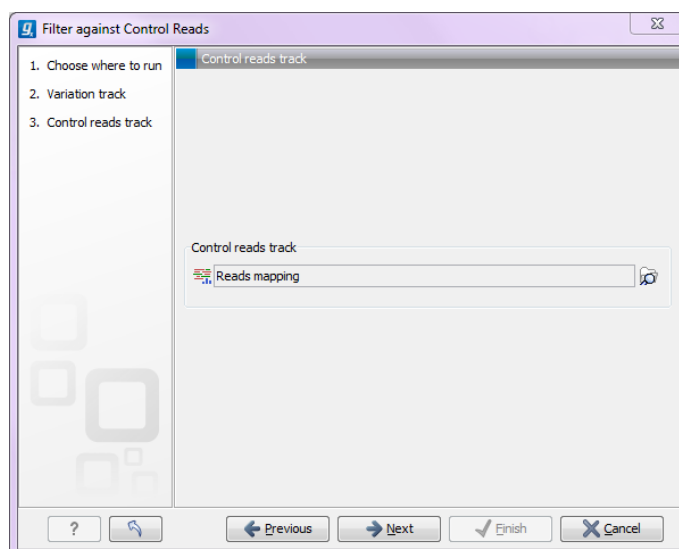


Figure 6.2: *The control reads data set.*

When clicking **Next**, you are asked to supply the number of reads in the control data set that should have the variant allele in order to include it as a match. All the variants where at least this number of control reads show the particular allele will be filtered away in the result track.

Please note that also variations, which have no coverage in the mapped control reads will be reported.

6.4 Database variation filter

This tool will allow you to select a variation track and add information about known variations from databases like dbSNP and COSMIC. You can use this to filter your experimental variations (see figure 6.3).

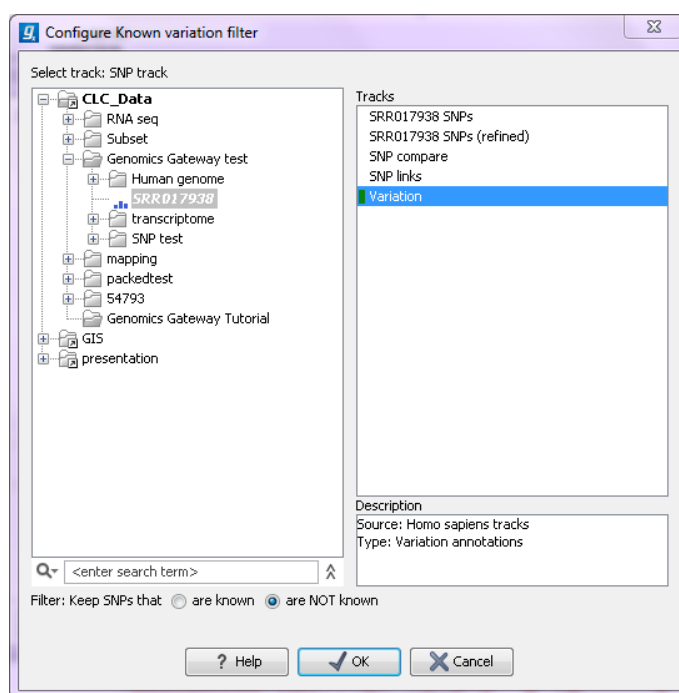


Figure 6.3: Filtering known variations.

In order to do this, you will have to import a file that is recognized as a variation file (see section 2.3).¹ The refiner will then compare the variations provided in the input track with the ones reported in the database track and evaluate whether it is known:

- If the input data has one variant and the database track has the same variant on this position, it is a known variation
- If the input data has one variant and the database track has a variant on this position but with a different nucleotide, it is not marked as a known variant.
- If the input data has two variants and the database track has one variant, it is marked as a known variation if one of the variants of the input data is identical to the database variant and the other is identical to the reference sequence. If none of these conditions are fulfilled, it is not classified as a known variation.

¹The minimum requirements for such a file is that it for each variation states what the allelic variation is.

- If the input data has one variants and the database track has two variants, it will be classified as a known variation if the variant of the input data is identical to any of the database variants.

6.5 Variation haplotype compare filter

This refiner is very similar to the **Database variation filter** described in section 6.4 with an important difference: it will filter variants whose haplotypes are identical in the two variation tracks. With the **Known variation filter** the input variant does *not* need to be *identical* with the variant found in the database, it just has to be *included* in the set of allelic variations of the database variant. The rationale is that the **Database variation filter** can be used to compare to a database track that includes the sum of variants reported in several studies whereas the **Variation haplotype compare filter** is intended for direct comparison of variation tracks from two single samples.

6.6 Copying and merging tracks

In some situations it is necessary to merge two tracks. This can be accomplished using the **Merge Tracks** tool. You select the two tracks and then decide whether the resulting merged track should merge the duplicate annotations into one.

Chapter 7

Functional consequences

The tools for working with functional consequences all take a variant track as input and will predict or classify the functional impact of the variant.

7.1 Amino acid changes

This tool annotates variations with amino acid changes given a track with coding regions and a reference sequence (see figure 7.1).

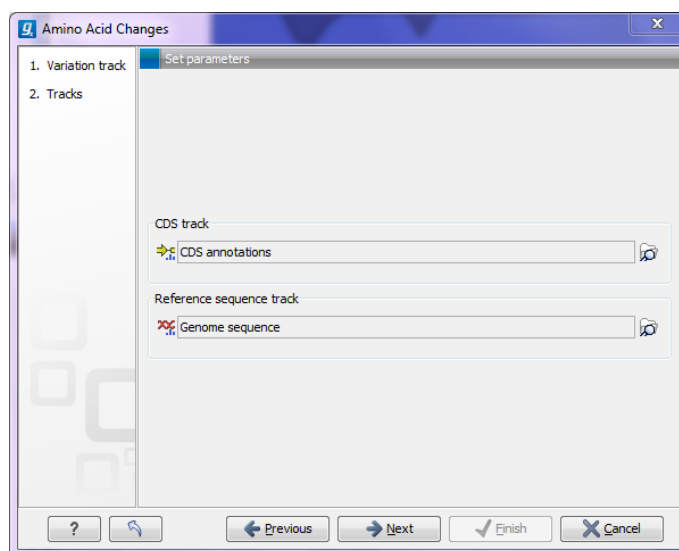


Figure 7.1: The amino acid changes annotation tool.

The result is a new track where each variant has information about the effect on the amino acid sequence of the corresponding protein.

7.2 Splice site effect prediction

This refiner will analyze a variation track to determine whether the variations fall within splice sites. A transcript track has to be selected as shown in figure 7.2.

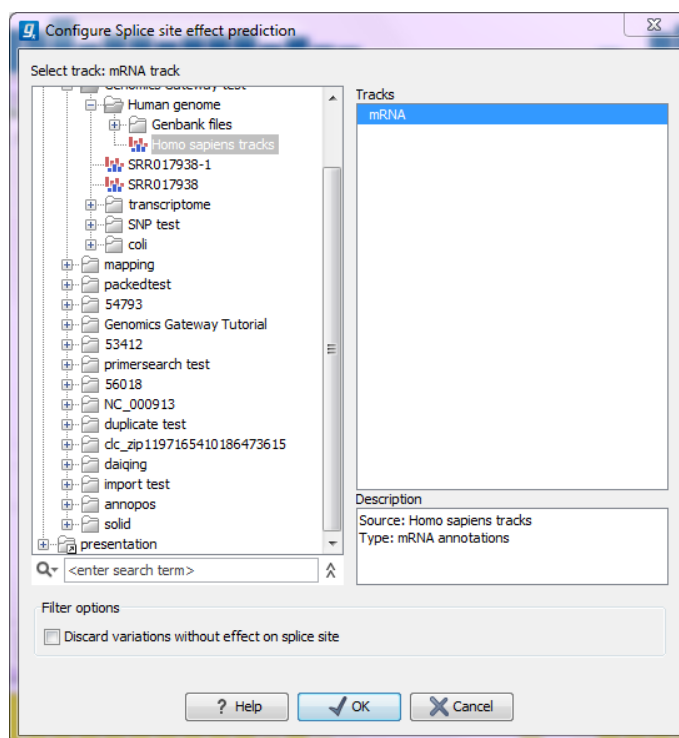


Figure 7.2: The splice site annotation.

If a variation falls within two base pairs of an intron-exon boundary, it will be annotated as a possible splice site disruption. As part of the refiner you can choose to exclude all variations that do not fall with splice sites.

7.3 GO enrichment analysis

This tool can be used to investigate candidate variations or better their corresponding altered genes for a common functional role. For example if you would like to know what is interesting in the zebu cattle in comparison to bison and taurine cattle, you can use this tool. For that approach, first filter all found variations in zebu for zebu-specific variations and afterwards run the GO enrichment test for biological process to see that more variations than expected are in immune response genes. These can then be further investigated.

For this, you need a GO association file, which includes gene names and associated Gene Ontology terms. You can download that from the Gene Ontology web site for different species (<http://www.geneontology.org/GO.downloads.annotations.shtml>). However, it is better to use a file with only the top-level GO terms annotated. For some species you can get that directly or you can create one on your own via the QuickGO tool (<http://www.ebi.ac.uk/QuickGO/GMultiTerm>).

When you run the GO Enrichment Analysis, you have to specify both the annotation association file, a gene track and finally which ontology (cellular component, biological process or molecular function) you like to test for (see figure 7.3).

The analysis starts by associating all of the variants from the input track with genes in the gene track, based on overlap with the gene annotations. Next, the Workbench tries to match gene names from the gene track with the gene names in the GO association file. Please be aware that

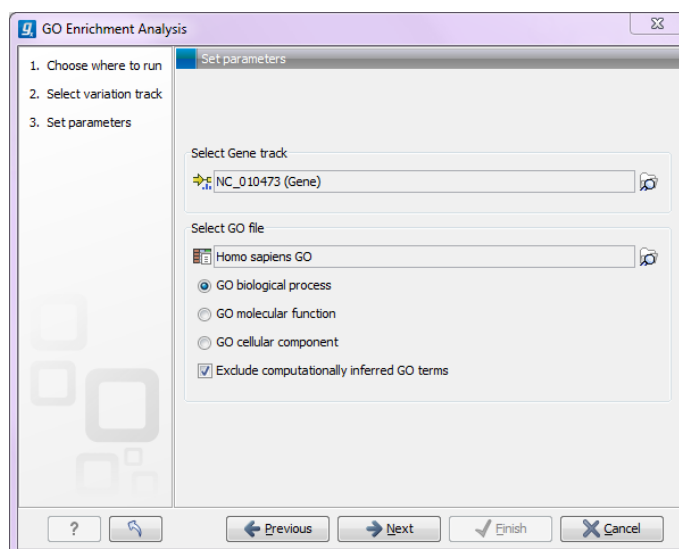


Figure 7.3: The GO enrichment settings.

the same gene name definition should be used in both files.

Based on this, the Workbench finds GO terms that are over-represented in the list. To find out which GO terms are over-represented, a hypergeometric test is used applied on the number of altered genes having GO term X in comparison to the number all genes in the GO association file having the same GO term.

The result is a table with GO terms and the calculated p-value for the candidate variations and a new variation file with annotated GO terms and the corresponding p-value. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed. That means how significant (trustworthy) a result is. In case of a small p-value the chance achieving the same result by chance with the same test statistic is very small.

Go-term	Description	Occurrences in all genes	Occurrences in overlap	pValues
0006950	response to stress	533	46	3.81E-3
0002376	immune system process	250	24	9.44E-3
0006412	translation	79	10	0.01

Figure 7.4: The GO enrichment results.

7.4 Conservation score annotation

Chapter 8

Future improvements and feedback

As stated in the beginning of this manual, this is a beta version of the Genomics Gateway, so we plan to enhance and add to the functionality during the coming months. The following list shows in brief what we have already planned.

- Direct integration with genomic databases to allow very simple selection of reference genome and annotations.
- Integrating more existing tools with the Genomics Gateway:
 - RNA-Seq
 - ChIP-Seq
 - Structural Variation (currently also in beta)
- Better support for sample comparisons
- Adding support for exporting tracks into formats like vcf, gvf etc.

Please note that this is not a prioritized list and that additional points will be added as we receive feedback from users. If you have tried the Genomics Gateway plug-in we will appreciate if you would take five minutes to give us some feedback at <http://www.clcbio.com/genomicsgateway/>.

Chapter 9

Installation of the Genomics Gateway Plug-in

The Genomics Gateway Plug-in is installed as a plugin. Plug-ins are installed using the plug-in manager¹:

Help in the Menu Bar | Plug-ins and Resources... (🔧)

or **Plug-ins (🔧) in the Toolbar**

The plug-in manager has four tabs at the top:

- **Manage Plug-ins.** This is an overview of plug-ins that are installed.
- **Download Plug-ins.** This is an overview of available plug-ins on CLC bio's server.
- **Manage Resources.** This is an overview of resources that are installed.
- **Download Resources.** This is an overview of available resources on CLC bio's server.

To install a plug-in, click the **Download Plug-ins** tab. This will display an overview of the plug-ins that are available for download and installation (see figure 9.1).

Clicking a plug-in will display additional information at the right side of the dialog. This will also display a button: **Download and Install**.

Click the Genomics Gateway Plug-in and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the Genomics Gateway Plug-in is not shown on the server, and you have it on your computer (e.g. if you have downloaded it from our web-site), you can install it by clicking the **Install from File** button at the bottom of the dialog. This will open a dialog where you can browse for the plug-in. The plug-in file should be a file of the type ".cpa".

When you close the dialog, you will be asked whether you wish to restart the CLC Workbench. The plug-in will not be ready for use before you have restarted.

¹In order to install plug-ins on Windows Vista, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

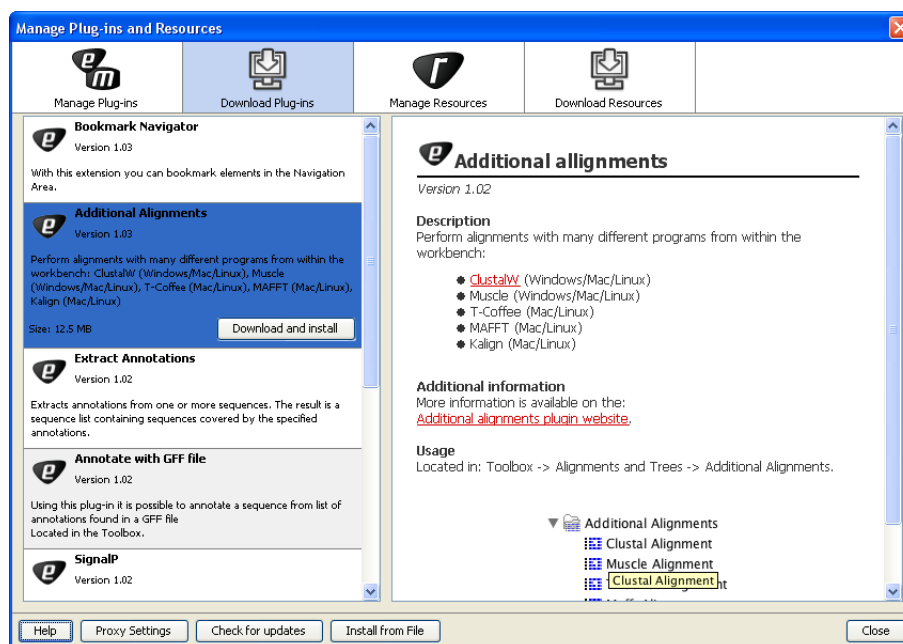


Figure 9.1: The plug-ins that are available for download.

Chapter 10

Uninstall

Plug-ins are uninstalled using the plug-in manager:

Help in the Menu Bar | Plug-ins and Resources... (📁)

or **Plug-ins** (📁) in the Toolbar

This will open the dialog shown in figure 10.1.

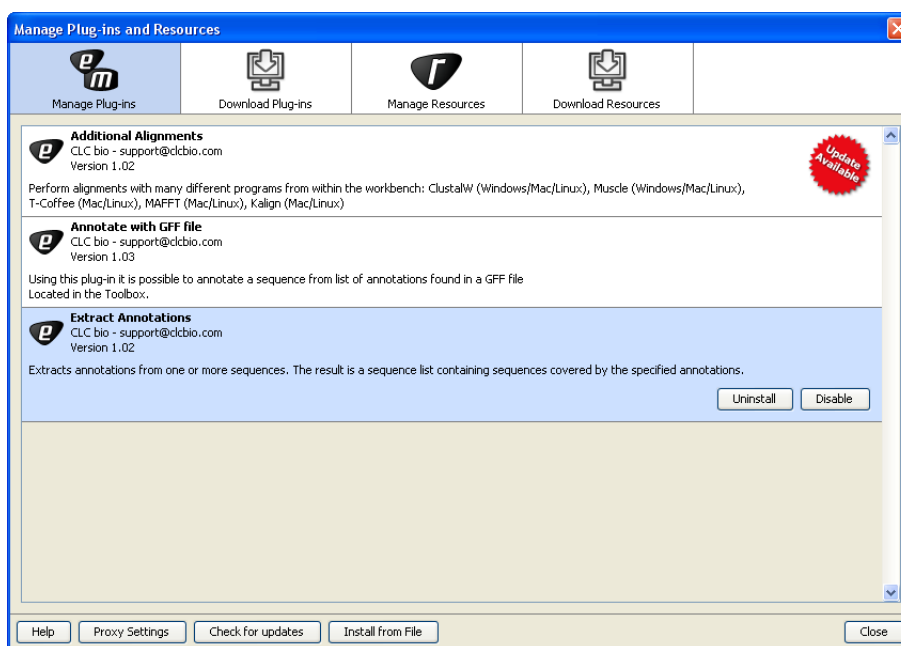


Figure 10.1: The plug-in manager with plug-ins installed.

The installed plug-ins are shown in this dialog. To uninstall:

Click the Genomics Gateway Plug-in | Uninstall

If you do not wish to completely uninstall the plug-in but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plug-in will not be uninstalled before the workbench is restarted.

Index

Bibliography, 35

GFF, 10

References, 35

Bibliography

- [Forbes et al., 2008] Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A., and Stratton, M. R. (2008). The catalogue of somatic mutations in cancer (cosmic). *Curr Protoc Hum Genet*, Chapter 10:Unit 10.11.
- [Sherry et al., 2001] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.