

Genomics Gateway Plug-in Beta

User Manual

User manual for

Genomics Gateway Plug-in 1.0

Windows, Mac OS X and Linux

June 30, 2011

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**

2 Building a reference genome **6**

2.1 Define reference genome 6

2.2 Import tracks from file 8

3 Mapping reads to the reference genome and calling SNPs **10**

3.1 Using an existing mapping file to create a mapping track 11

3.2 SNP and DIP detection 11

4 Visualization: the genome browser **13**

4.1 Zooming and customizing the layout of the track 13

4.2 Hiding, adding and reordering tracks 15

4.3 Showing a track in a table 16

4.4 Finding a gene or a position on the genome 17

5 Refiners: annotate and filter tracks **18**

5.1 Amino acid changes 19

5.2 Exon number annotation 19

5.3 Flanking sequence 19

5.4 Gene and OMIM link annotation 19

5.5 Known variation filter 19

5.6 Overlap annotation and overlap filter 21

5.7 SNP haplotype compare filter 21

6 Future improvements and feedback **22**

7 Installation of the Genomics Gateway Plug-in **23**

8 Uninstall **25**

Index **25**

Chapter 1

Introduction to the Genomics Gateway Plug-in

The Genomics Gateway Plug-in 1.0 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 6 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.

Chapter 2

Building a reference genome

In later versions of the Genomics Gateway it will be possible to connect to 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. However, with the first beta version you need to import 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 two 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. 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).

Click **Next** to adjust output options (see figure 2.2):

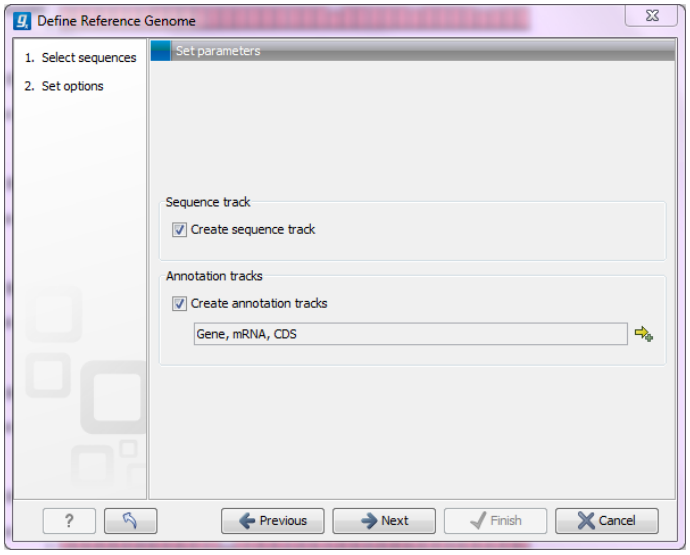


Figure 2.1: *Define the annotation types to be included in the reference genome.*

- Copy data to the new track set
- Refer to the original input data from the new track set

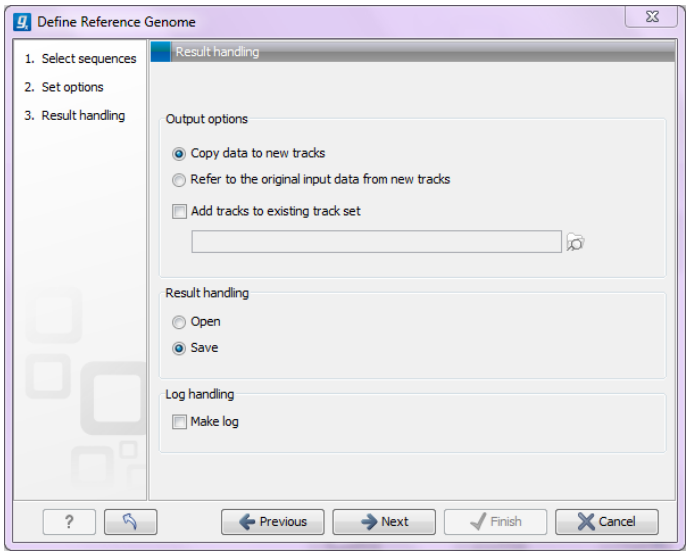


Figure 2.2: *The information can be copied and stored in the new track set, or it can be fetched from the original input file whenever it is needed for visualization or analysis.*

To make sense of these two options, a little background info is needed: a track set (the reference genome created by this tool is a track set) contains a number of tracks which can be displayed and analyzed together. The data of each track can be saved either in the track set itself, or it can be fetched from another file. This means that the same data source can be included in several track sets without duplicating the data, and it adds a great flexibility to managing the data sources. The main idea behind this is to be able to keep reference data stored in one or more files that are separate from the files storing experimental data but still both can be analyzed together.

The option to **Add tracks to existing track set** requires you to select an existing track set for the same reference genome (the same number and size of chromosomes). This will add your new

tracks to this track set rather than create a new track set. When you are building a new reference genome from scratch, just leave this option unchecked.

2.2 Import tracks from file

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

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 particular 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

Complete Genomics master var file This is the file format used by Complete Genomics for 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).

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 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 2.3) allows you to select the target genome that the annotations should be added to.

Click **Next** to select annotation file. Please note that if the file is zipped, it needs to be unzipped first.

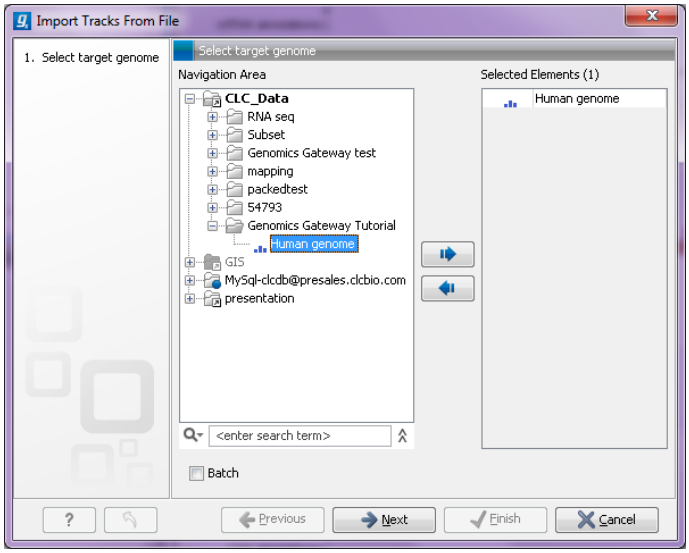


Figure 2.3: *Define the reference genome.*

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).

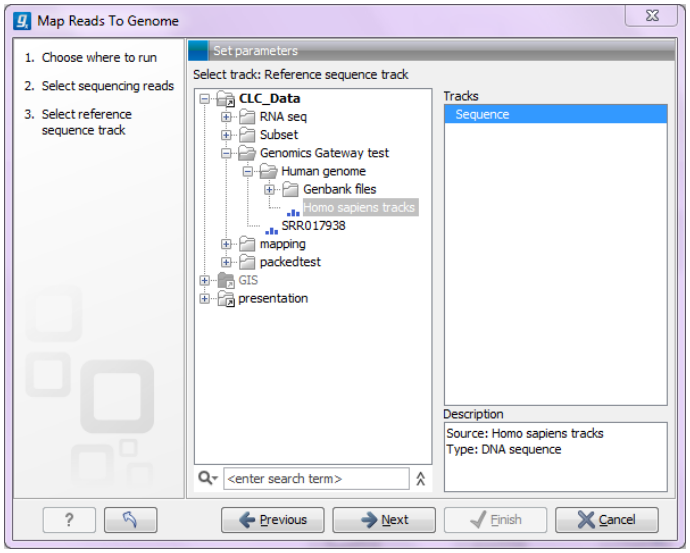


Figure 3.1: *Define the reference genome.*

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. This track can be saved independently of the reference genome, or you can choose to save it in a track set that you have previously created by clicking **Add tracks to**

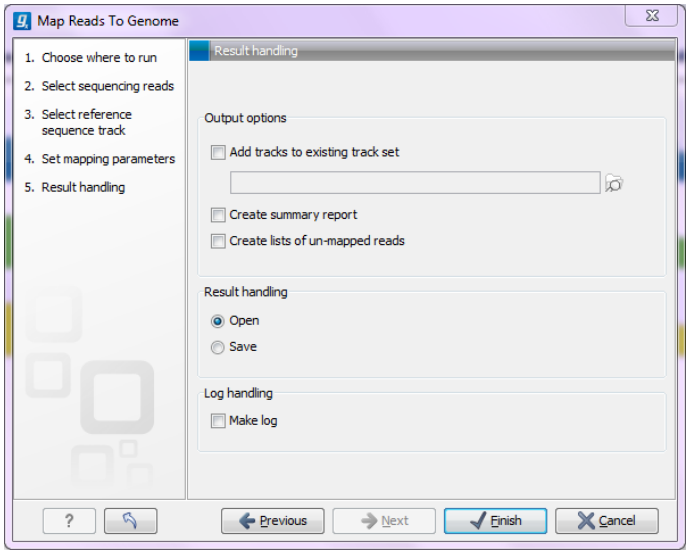


Figure 3.2: *Specify output options.*

existing track set. This will add the new track to the file that you select. This can be useful e.g. for saving the mappings of different samples in one file, allowing easy visualization and comparison of all samples. If you do not choose this option, a new file will be created containing the new track and references to the genome used as input.

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 ( ) ( ).

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

First, you have to choose whether you want to create a **copy** of the mapping data or just have the workbench **refer to the original** mapping file when visualizing and analyzing the data. This means that no data is duplicated (saving disk space) and it also means that any editing of the original data will cause the new track to be updated.

No matter what you choice is, you can choose to either create a new track set with your track or to **add it to an existing track set**.

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.

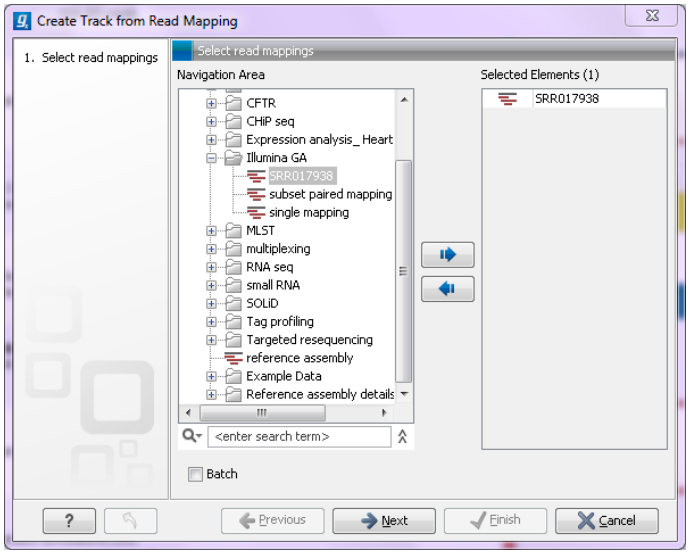


Figure 3.3: *Select a read mapping*

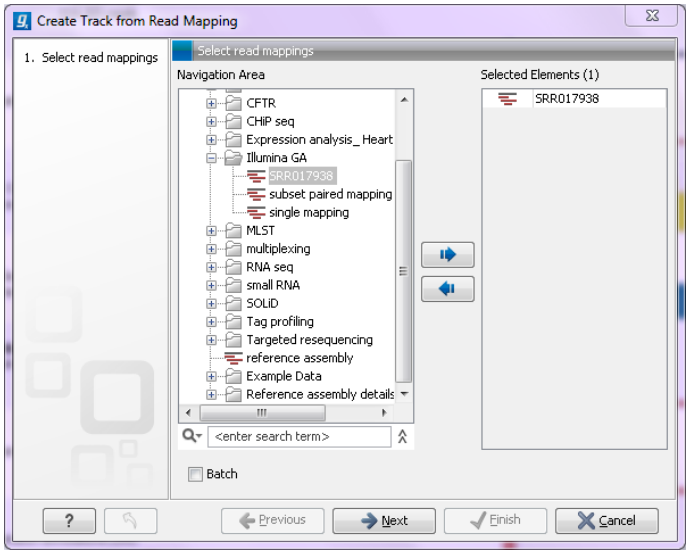


Figure 3.4: *Select output options*

There are three changes compared to the original versions: The first is that the input data is now a track set including a read mapping track that you select in the first dialog displayed. The second is that the output is a track that can either be saved as its own track set or it can be added to an existing track set. In most cases it makes sense to add to the track set containing the mapping track since you will then have all the information for this sample saved in one file.

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 refiners as explained in section 5.

Chapter 4

Visualization: the genome browser

Figure 4.1 shows an example of a track set 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.

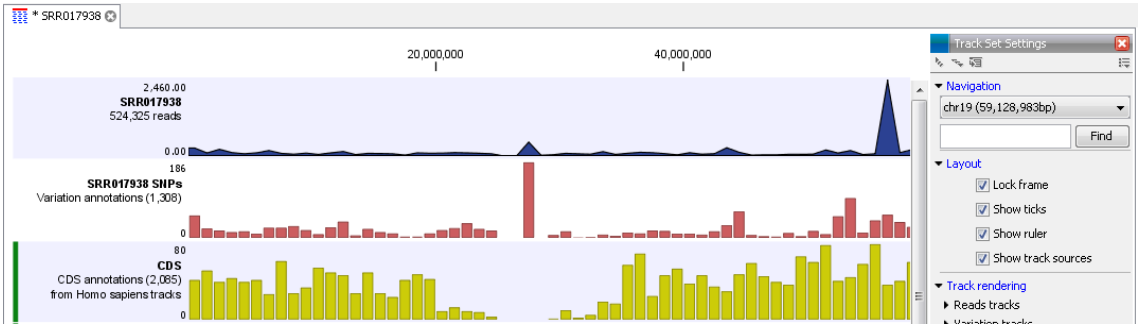


Figure 4.1: *Four tracks shown in the genome browser*

4.1 Zooming and customizing the layout of the track

Zooming in and out on 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.

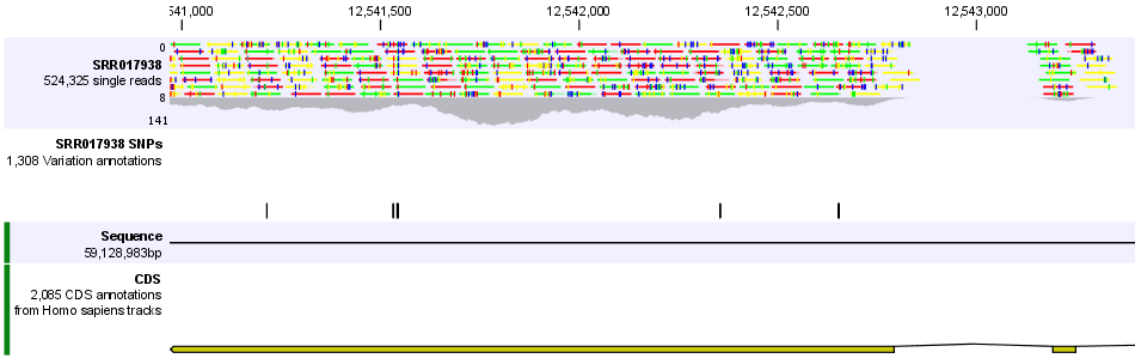


Figure 4.2: *Zooming in reveals more detail on each track*

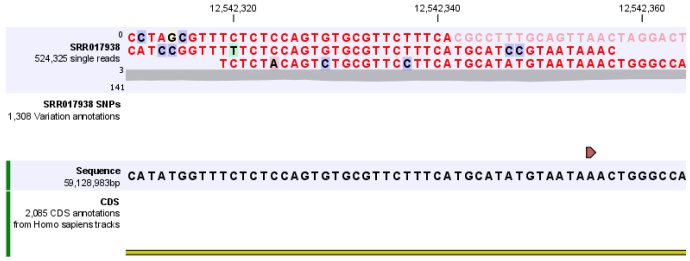


Figure 4.3: *Zooming all the way in shows the actual bases of the reads and the reference sequence.*

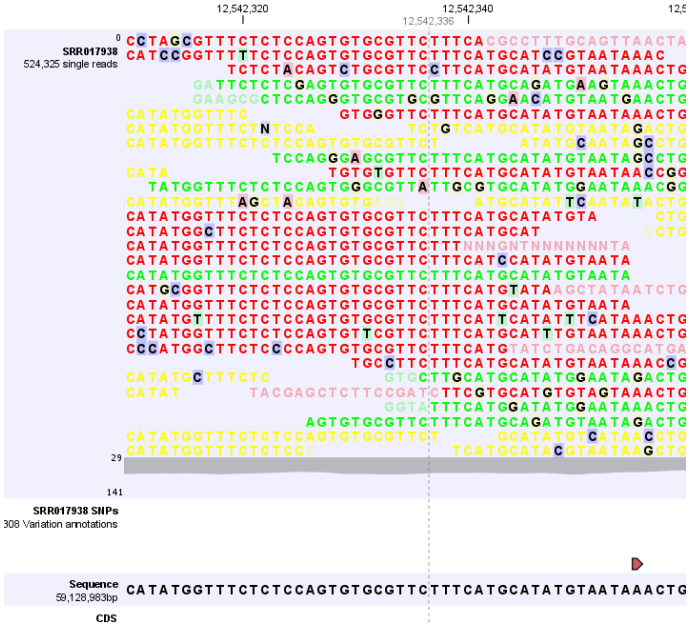


Figure 4.4: *Adjusting the height of the track.*

This can be done individually for each track and will be remembered when you open the track set again.

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

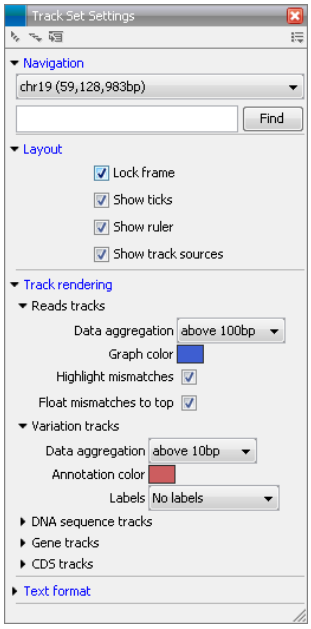


Figure 4.5: *The Side Panel makes it possible to adjust the aggregation level.*

zoomed in, and a high value means that you can see details even when zoomed out.

4.2 Hiding, adding and reordering tracks

You can organize your tracks by right-clicking on any of the tracks as shown in figure 4.6.

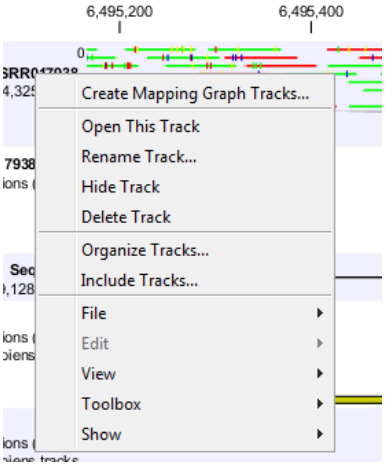


Figure 4.6: *Options to organize tracks.*

The options are explained below:

Create Mapping Graph Tracks This will allow you to create a new track from a mapping track with any of the following information:

- Read coverage
- Non-specific read coverage
- Non-perfect read coverage

- Paired read coverage
- Broken pair coverage
- Paired read distance

Open This Track This opens a table view of the track as described in section 4.3.

Rename Track Please note that this is not possible for tracks that are referred from the track set you currently have open. This is indicated by the color bar at the left-hand side of the track (the green bar in figure 4.1). Read more about referring to other tracks in section 2.1.

Hide Track This will hide the track from the current view. It can be displayed again by choosing **Organize Tracks** (see below).

Delete Track This will delete the track from the track set. The track can only be retrieved again by clicking **Undo** while the view is still open. Please note that this is not possible for tracks that are referred from the track set you currently have open. This is indicated by the color bar at the left-hand side of the track (the green bar in figure 4.1). Read more about referring to other tracks in section 2.1.

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.

Organize Tracks This allows you to add or remove tracks from the view as shown in figure 4.7. In the panel to the left you have all the available tracks organized by track source, track type or alphabetically. The track sources are the track sets that are available (to add more see the above). In this example it is the reference genome (Homo sapiens tracks) and the track set with the experimental data (SRR017938). The panel to the right shows the tracks that are currently shown in the view. You can use the horizontal arrow buttons to add or remove tracks from this list, and the vertical arrow buttons can be used to change the order of the tracks being displayed.

Edit Refinement This is available for a refined track to edit the refinement settings (see section 5).

Create New Refined Track This will allow you to create a new refined track based on the track you clicked (see section 5).

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 4.8 shows an example of a variation database track open 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).

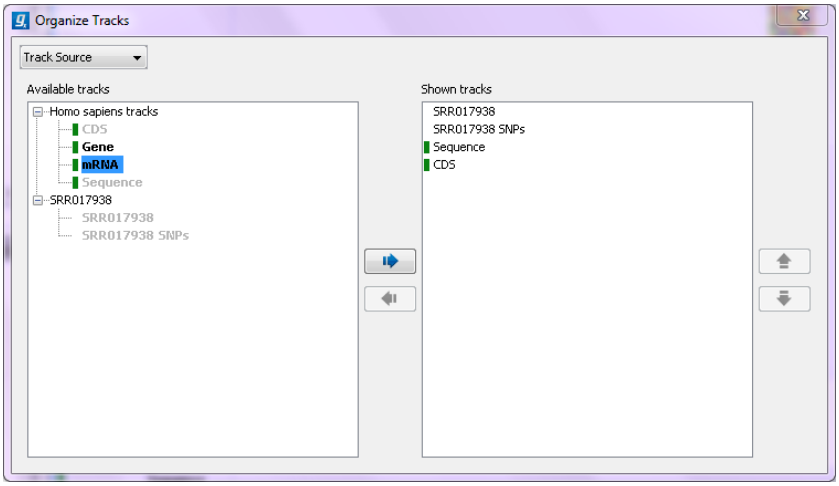


Figure 4.7: *Removing and adding tracks to the view.*

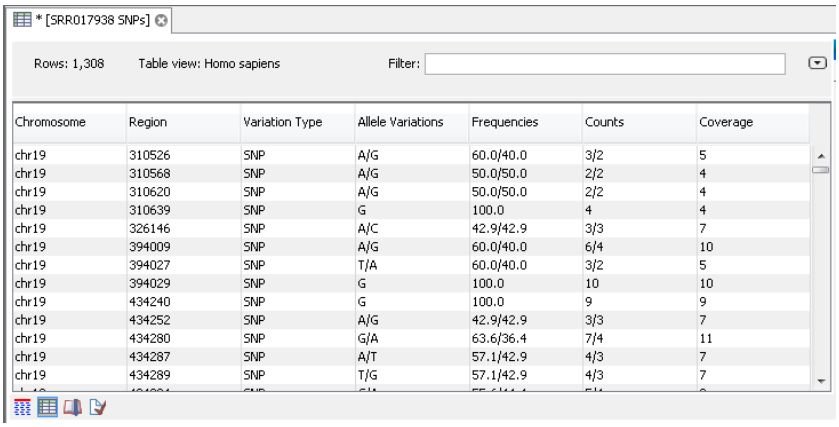


Figure 4.8: *Showing a variation track in a table.*

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**).

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

Refiners: annotate and filter tracks

The refiners are used to annotate and filter a track, typically by comparing it to another track. One example is to filter a variation track to only show the variations that are not known from variation databases. Another example is to add information about synonymous and non-synonymous substitutions. When you refine a track, the result is a new track that is added to the track set. Once the track has been created, you can edit the refinement settings (right-click the track label and choose **Refinements...**).

Please note that the collection of refiners available in the current version will be extended going forward to incorporate many more of the comparative analyses we wish to target with the Genomics Gateway.

To create a new refined track:

Toolbox | **Genomics Gateway** | **Refine Track** or **right-click th**

In the dialog shown, first select the relevant track set in the **Navigation Area** and second select the track within this track set that you want to refine. Click **Next**.

This will bring up a dialog (as shown in figure 5.1) that shows the list of available refiners.

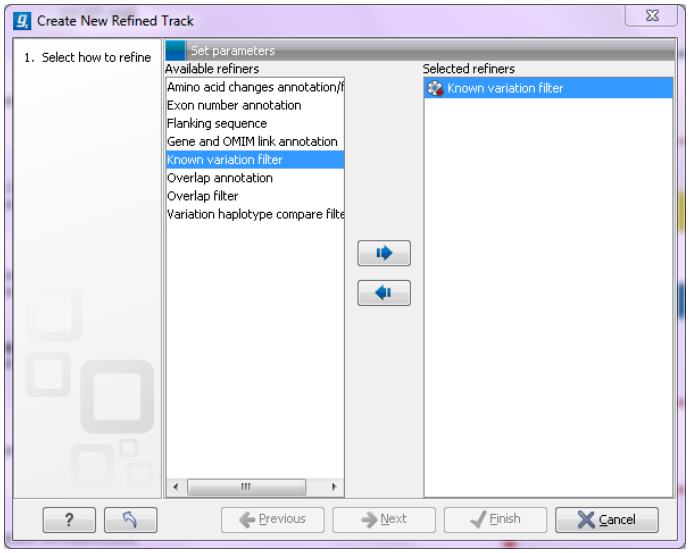


Figure 5.1: *The list of refiners.*

Add refiners from the panel to the left. Note that a refiner can be added several times. Once a refiner has been added, it will show a red icon if there are required parameters that need to be set and a green icon if the default parameters are OK. In any case, open the refiner's settings dialog by double-clicking the refiner in the list. In the following each refiner and its settings are described.

5.1 Amino acid changes

This refiner annotates variations with amino acid changes given a track with coding regions and a reference sequence (see figure 5.2).

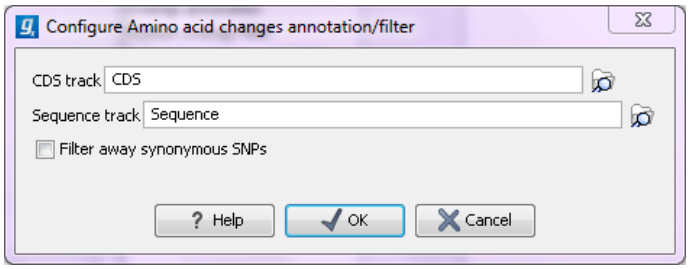


Figure 5.2: *The amino acid changes refiner.*

In addition, this refiner can be used to filter away all SNPs that do not cause an amino acid changes (i.e. only include non-synonymous SNPs in the new refined track).

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 5.3).

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.

5.4 Gene and OMIM link annotation

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

5.5 Known variation filter

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

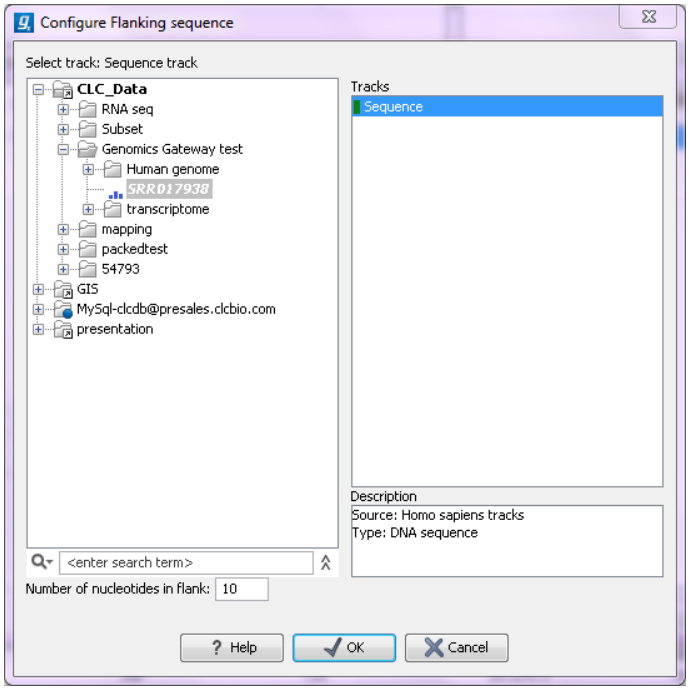


Figure 5.3: *The flanking regions refiner.*

figure 5.4).

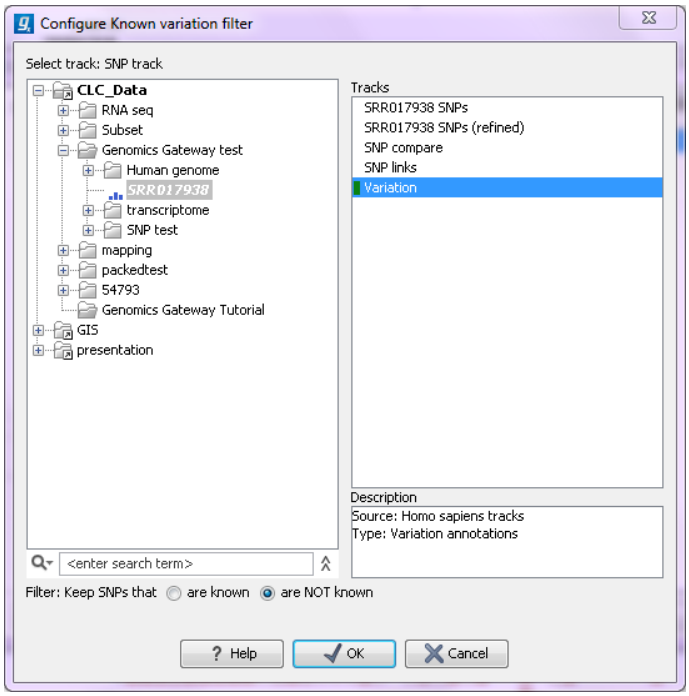


Figure 5.4: *Filtering known variations.*

In order to do this, you will have to import a file that is recognized as a variation file (see section 2.2). [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:

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

- 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.
- 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.

At the bottom of the dialog, you can filter away either known or private variations.

5.6 Overlap annotation and overlap filter

The overlap annotation will add information about overlapping annotations and the overlap filter can be used for filtering based on overlapping annotations. 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 5.7) or the Known variation filter (section 5.5) should be used instead.

5.7 SNP haplotype compare filter

This refiner is very similar to the **Known variation filter** described in section 5.5 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 **Known 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.

Chapter 6

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)
- Adding support for exporting tracks into formats like vcf, gvf etc.
- Adding more refiners (e.g. to detect splice-site mutations) and making the refiners more intuitive to use

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 7

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 manager[1]:

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 7.1).

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.

[1] In 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. When you start the Workbench after installing the plug-in, it should also be run in administrator mode.

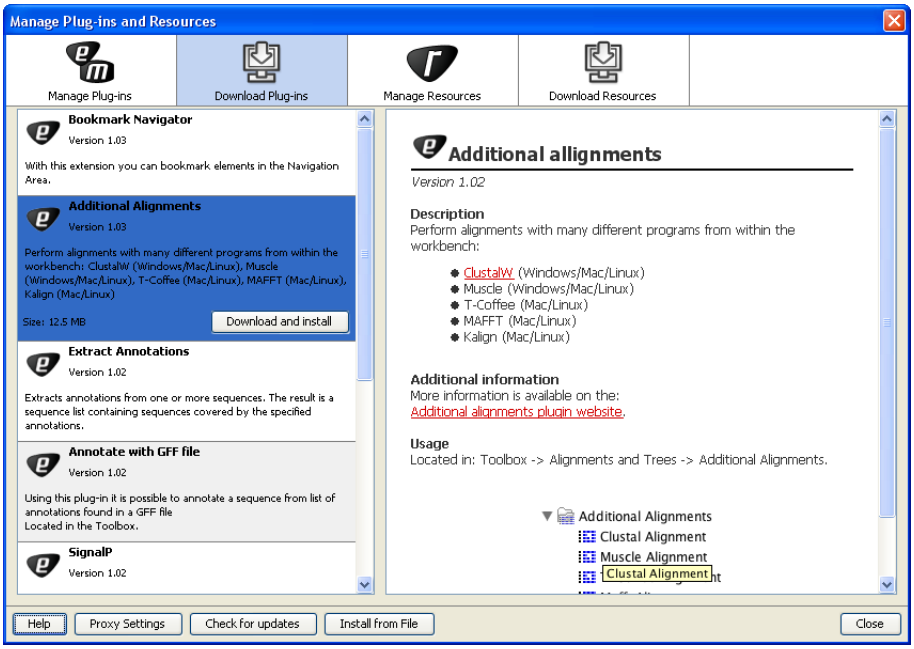


Figure 7.1: *The plug-ins that are available for download.*

Chapter 8

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 8.1.

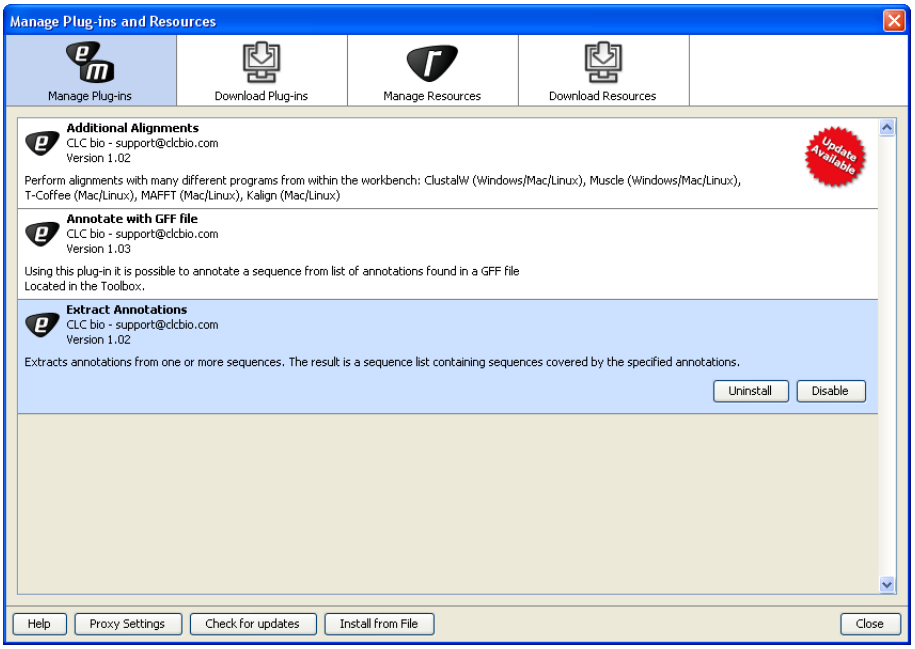


Figure 8.1: *The plug-in manager with plug-ins installed.*

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

GFF, 8