Annotate Sequence with GFF File

User manual

User manual for

Annotate Sequence with GFF File

Windows, Mac OS X and Linux

April 15, 2010

This software is for research purposes only.

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

Contents

Chapter 1

Introduction to the Annotate Sequence with GFF File

The Annotate Sequence with GFF File makes it very easy to annotate a sequence with annotations from a GFF (Generic Feature Format) or GTF (Gene Transfer Format) file. A GFF/GTF file does not contain any sequence information, it only contains a list of annotations. You can read more about the format at http://www.sanger.ac.uk/Software/formats/GFF/.

There are many different versions of GFF and GTF. We support a big part of the GFF3 definition (see http://www.sequenceontology.org/gff3.shtml, and we also support GTF format as defined at http://mblab.wustl.edu/GTF22.html. We do not fully support GFF3, but for most purposes GFF3 will be annotated as expected.

For a comprehensive source of genomic annotation of genes and transcripts, we refer to the Ensembl web site at http://www.ensembl.org/info/data/ftp/index.html. On this page, you can download GTF files that can be used to annotate genomes for use in other analyses in the *CLC Genomics Workbench*, e.g. **RNA-Seq Analysis** ().

Note! GTF-files downloaded from the UCSC genome browser are not compatible with the RNA-Seq Analysis of *CLC Genomics Workbench* **because the gene and transcript annotations cannot be matched**.

This manual will show two examples of how to use the plug-in to annotate a genome. First example will prepare a reference genome for RNA-Seq analysis, and the second example will show how to annotate a reference genome with known SNPs that can be used to detect overlapping SNPs when running SNP detection in *CLC Genomics Workbench*.

Chapter 2

Annotating a reference genome with genes and transcript

In this example we use the mouse genome but the methods described here apply equally well for other genomes. First, we download the fasta files for the reference genome at Ensembl: `ftp://ftp.ensembl.org/pub/current_fasta/mus_musculus/dna/`. The downloaded files can then be imported ( ) into the Workbench.

Next, download the corresponding GTF file from `ftp://ftp.ensembl.org/pub/current_gtf/mus_musculus/`. The file needs to be unzipped before you can use it.

To annotate the reference with the genes and transcripts from the GTF file:

Toolbox | **General Sequence Analyses** ( ) | **Annotate with GFF/GTF File** ( )

Now, select all the mouse chromosomes and click **Next**. This opens the dialog shown in figure 2.1.

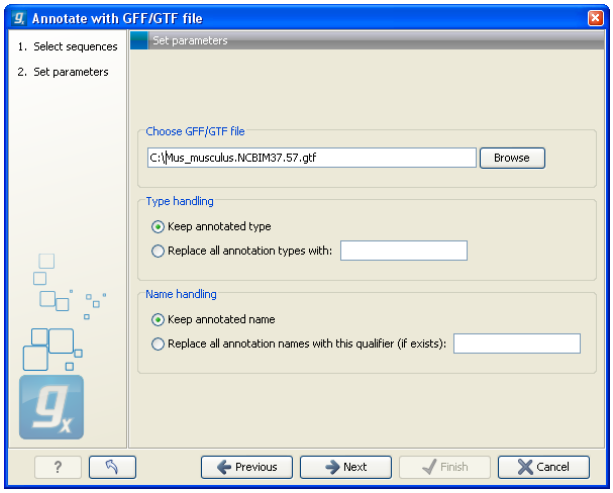

Figure 2.1: *Select the GTF file by clicking the browse icon.*

Click **Browse** to select the GFF/GTF file and click **Next**. Choose to **Save** the results and click **Finish**. This will add the annotations from the file to the sequences. Your reference genome is now ready for use in e.g. **RNA-Seq Analysis** ( ).

Chapter 3

Annotating a reference genome with known SNPs

In this example we annotate the human genome with known SNPs from the UCSC database. First, download the SNP GTF file from the UCSC table browser at `http://genome.ucsc.edu/cgi-bin/hgTables`. You can see an example query in figure 3.1.

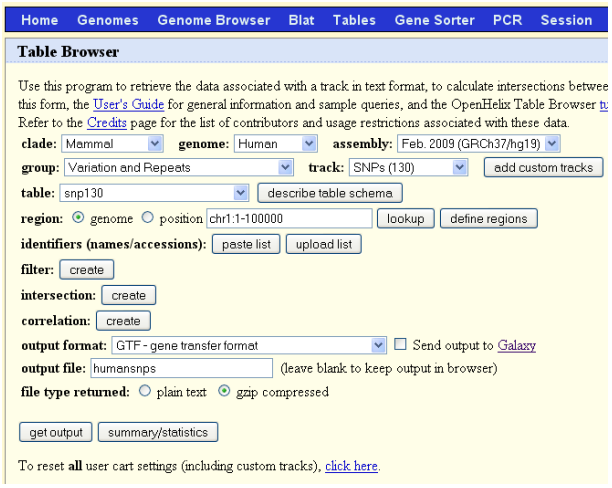

Figure 3.1: *Downloading an annotation file with known SNPs from UCSC.*

The file needs to be unzipped before you can use it, and you should add ".gtf" at the end of the file name.

To annotate your reference with the SNP annotations:

Toolbox | **General Sequence Analyses** () | **Annotate with GFF/GTF File** ()

Now, select all the chromosome sequences. Note that the chromosomes need to be named in the same way as in the gtf file which is simple chr1, chr2, chr3, etc. Clicking **Next** opens the dialog shown in figure 3.2.

Click **Browse** to select the GFF/GTF file and enter a type for the SNP annotations. This is because the UCSC file uses the annotation type `exon` for the SNPs. An example of the gtf file is shown below:

```
chr21 hg19_snp130 exon 10697933 10697933 0.000000 + . gene_id "rs55981545"; transcript_id "rs55981545";
```

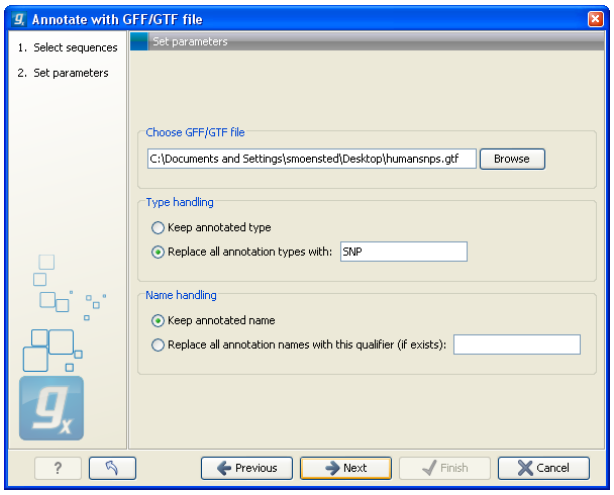

Figure 3.2: *Specifying an annotation file and setting an annotation type.*

```
chr21 hg19_snp130 exon 10697993 10697993 0.000000 + . gene_id "rs73327798"; transcript_id "rs73327798";
chr21 hg19_snp130 exon 10698080 10698080 0.000000 + . gene_id "rs73327799"; transcript_id "rs73327799";
chr21 hg19_snp130 exon 10698102 10698102 0.000000 + . gene_id "rs71245703"; transcript_id "rs71245703";
```

By specifying a different annotation type, the Workbench will convert the annotations to this type.

Click **Next** and Choose to **Save** the results and click **Finish**. This will add the annotations from the file to the sequences.

When you subsequently use contigs based on this reference for SNP detection, the overlapping SNP annotations will be reported in the overlapping annotations column in the SNP table.

Chapter 4

Naming of annotations

Annotations are named in the following, prioritized way:

1. If one of the following qualifiers are present, it will be used for naming (prioritized):
   (a) Name
   (b) Gene_name
   (c) Gene_ID
   (d) Locus_tag
   (e) ID
2. If none of these are found, the annotation type will be used as name

You can overrule this naming convention by choosing **Replace all annotation names with this qualifier** and specifying another qualifier (see figure 4.1.

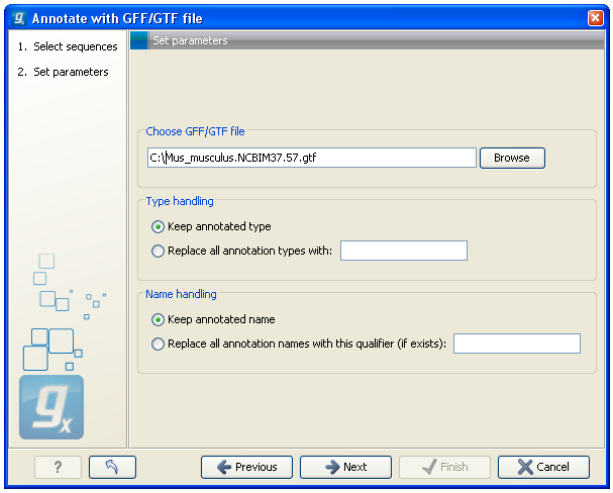

Figure 4.1: *You can choose Replace all annotation names with this qualifier to specify your own naming convention.*

Note that you have to type in the exact same qualifier as in the annotation file. This feature is recommended for advanced users only.

Note that transcript annotations are handled separately, since they inherit the name from the gene annotation.

Chapter 5

Troubleshooting

If you do not get the result you want when annotating with a GFF/GTF file, click the **Make log** checkbox. This will show you more information about the number of annotations that were found and if there are any that are not matched.

Typically, the problem is that the name of the file in the Workbench and the sequence identifier in the GFF/GTF file (the first column) have to be identical.

Chapter 6

Online resources

Online resources about GFF and GTF:

- Definition of GTF format: http://mblab.wustl.edu/GTF22.html
- Definition of GFF3 format: http://www.sequenceontology.org/gff3.shtml
- Annotation resources at Ensembl http://www.ensembl.org/info/data/ftp/index.html
- Annotation resources at UCSC: http://genome.ucsc.edu/cgi-bin/hgTables
- Links to annotation resources for various model organisms: http://wiki.geneontology.org/index.php/Reference_Genome_sequence_annotation

Chapter 7

Installation

The Annotate Sequence with GFF File is installed as a plug-in. 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 Annotate Sequence with GFF File and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the Annotate Sequence with GFF File 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.

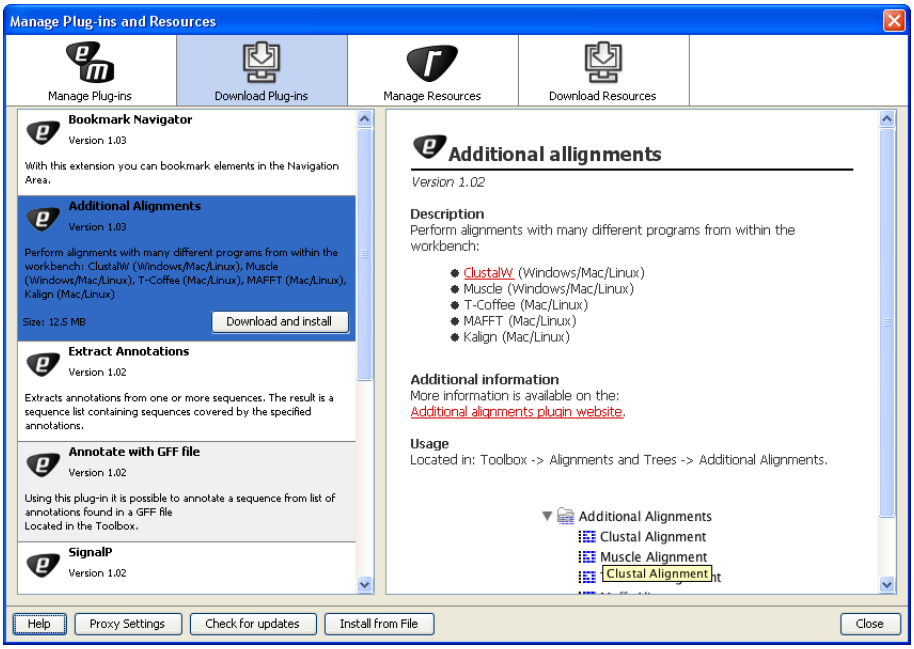

Figure 7.1: *The plug-ins that are available for download.*

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.

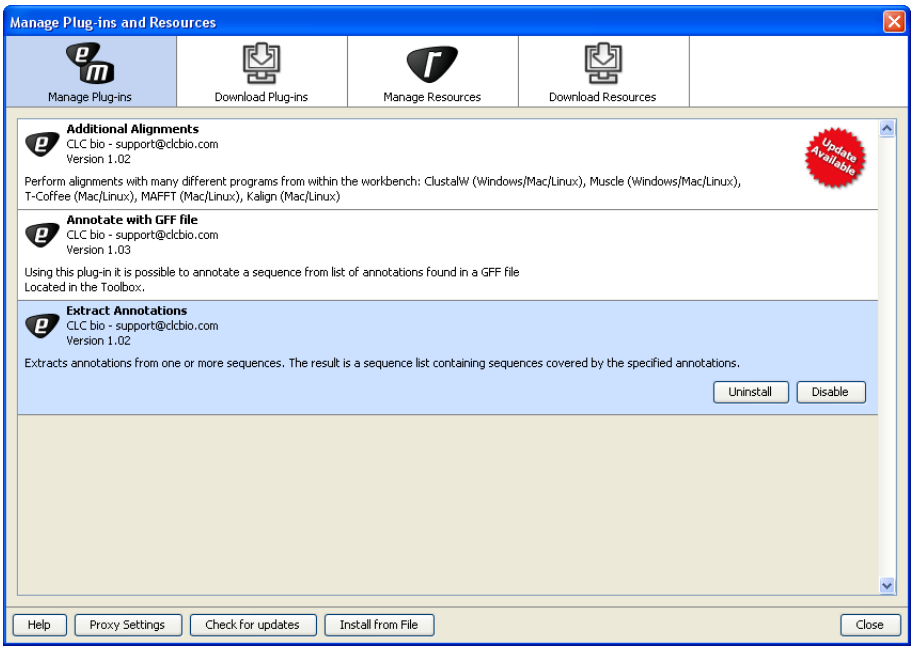

Figure 8.1: *The plug-in manager with plug-ins installed.*

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

Click the Annotate Sequence with GFF File | 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.