Annotate with GFF file Plugin

USER MANUAL

User manual for

Annotate Sequence with GFF File 2.3

Windows, Mac OS X and Linux

September 30, 2016

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
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 formats at http://www.sanger.ac.uk/resources/software/gff/spec.html and http://mblab.wustl.edu/GTF22.html.

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. In other words, most GFF3 files can be used to annotated sequences using this tool.

The GFF and GTF files can contain various types of annotations. In general, the Annotate Sequence with GFF File action adds the annotation in each of the lines in the file to the chosen sequence, at the position or region in which the file specifies that it should go, and with the annotation type, name, description etc. as given in the file. However, special treatment is given to annotations of the types CDS, exon, mRNA, transcript and gene. For these, the following applies:

- A gene annotation is generated for each gene_id. The region annotated extends from the leftmost to the rightmost positions of all annotations that have the gene_id (gtf-style).
- CDS annotations that have the same transcriptID are joined to one CDS annotation (gtf-style). Similarly, CDS annotations that have the same parent are joined to one CDS annotation (gff-style).
- If there are more than one exon annotation with the same transcriptID these are joined to one mRNA annotation. If there is only one exon annotation with a particular transcriptID, and no CDS with this transcriptID, a transcript annotation is added instead of the exon annotation (gtf-style).
- Exon annotations that have the same mRNA as parent are joined to one mRNA annotation. Similarly, exon annotations that have the same transcript as parent, are joined to one transcript annotation (gff-style).

Note that genes and transcripts are linked by name only (not by position, ID etc). 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*.

This manual will show two examples of how to use the plugin to annotate a genome for the purposes of RNA-Seq analysis in the *CLC Genomics Workbench* version 6.5.x and earlier.

If you are using the *CLC Genomics Workbench* and are interested in standard reference genomic data, please also take a look at the Download Genomes tool as described in the *CLC Genomics Workbench* manual at:

http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Download_reference_genome.html

Chapter 2

Annotating a reference genome with genes and transcripts

In this example we use the horse 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/equus_caballus/dna/`. The whole genome can be downloaded as a single file that ends with `.dna.toplevel.fa.gz`. Import () using Standard Import, check "Automatic Import", there's no need to unzip the file. Next, download the corresponding GTF file from `ftp://ftp.ensembl.org/pub/current_gtf/equus_caballus/`.

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

From the *CLC Main Workbench*:

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

Form the *CLC Genomics Workbench*:

> **Toolbox**| **Classical Sequence Analysis** ()| **General Sequence Analysis** ()| **Annotate with GFF/GTF File** ()

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

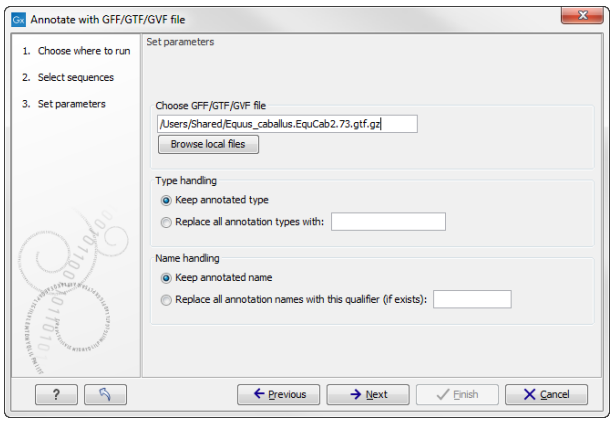

Figure 2.1: *Select the GTF file by clicking the browse icon.*

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.

Notes about gene annotations from the USCS. GTF-files downloaded from the UCSC genome browser are not compatible with choosing to run RNA-Seq Analysis on a annotated eukaryotic reference because the gene and transcript annotations cannot be matched. You may choose to use USCS gene annotations only for RNA-seq analysis: In the *CLC Genomics Workbench* version 7.x you can choose to only consider gene annotations by choosing the option "Genome annotated with genes only". For the *CLC Genomics Workbench* version 6.5.x and earlier, you can get the same effect by choosing to treat the reference as an annotated prokaryotic reference.

We would, however, generally recommend getting the annotations from a source where genes and transcripts are linked for the purposes of RNA-seq on eukaryotic genomes, such as from Ensembl.

Chapter 3

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 3.1).

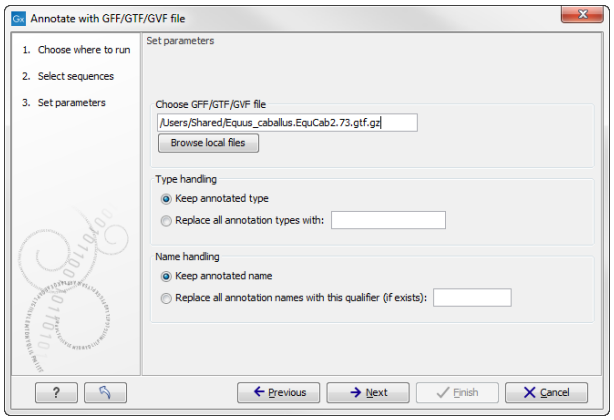

Figure 3.1: *You can choose Replace all annotation names with this qualifier to specify your own naming convention.*

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 4

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**. It is these identifiers, the one on your sequence and the ones in the first column of the GFF file, that are matched so that the system knows which sequence the annotation belongs to. You may need to change the name of your sequence objects to make them match the names used in the first column of the GFF/GTF file, or alternatively, change the identifiers used in the first column of your GFF/GTF file to ensure these match with the names of your sequence objects.

Chapter 5

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 6

Installation

The Annotate Sequence with GFF File is installed as a plugin. Plugins are installed using the plugin manager. In order to install plugins on Windows, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

Help in the Menu Bar | **Plugins... (** **)**

or **Plugins (** **) in the Toolbar**

The plugin manager has two tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins on CLC bio's server.

To install a plugin, click the **Download Plugins** tab. This will display an overview of the plugins that are available for download and installation (see figure 6.1).

Clicking a plugin 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 plugin is downloaded and installed.

If the Annotate Sequence with GFF File is not shown on the server, and you have it on your computer (for example if you have downloaded it from our website), 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 plugin. The plugin 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 plugin will not be ready for use until you have restarted.

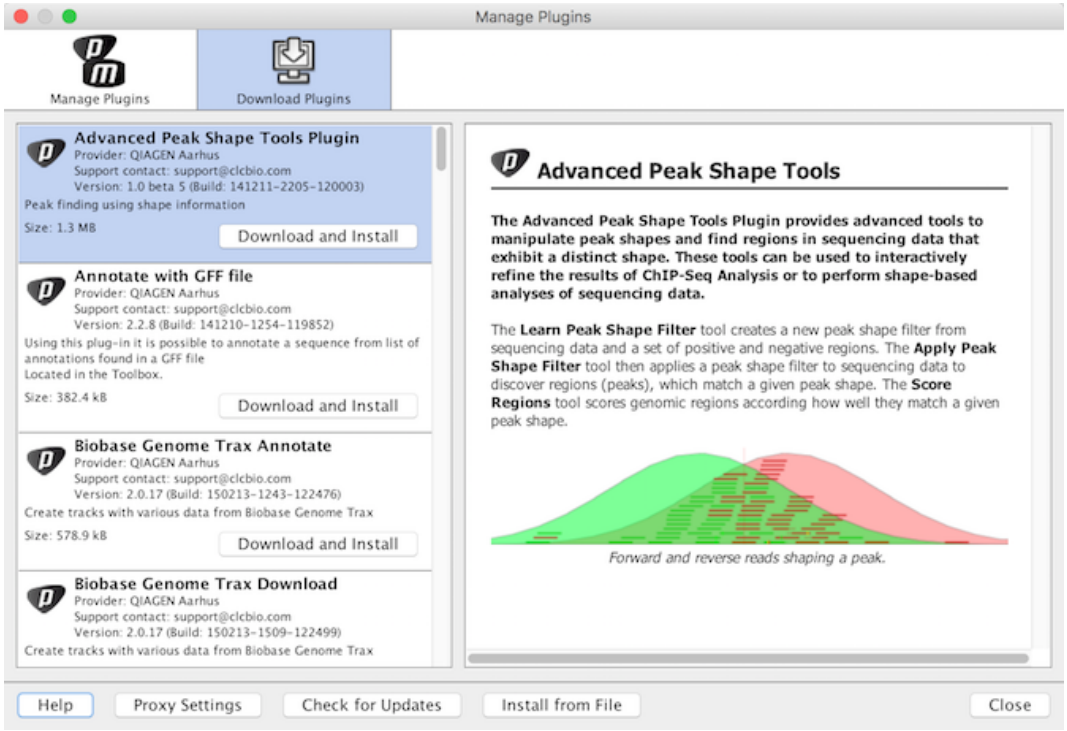

Figure 6.1: *The plugins that are available for download.*

Chapter 7

Uninstall

Plugins are uninstalled using the plugin manager:

Help in the Menu Bar | Plugins... ( )

or **Plugins ( ) in the Toolbar**

This will open the dialog shown in figure 7.1.

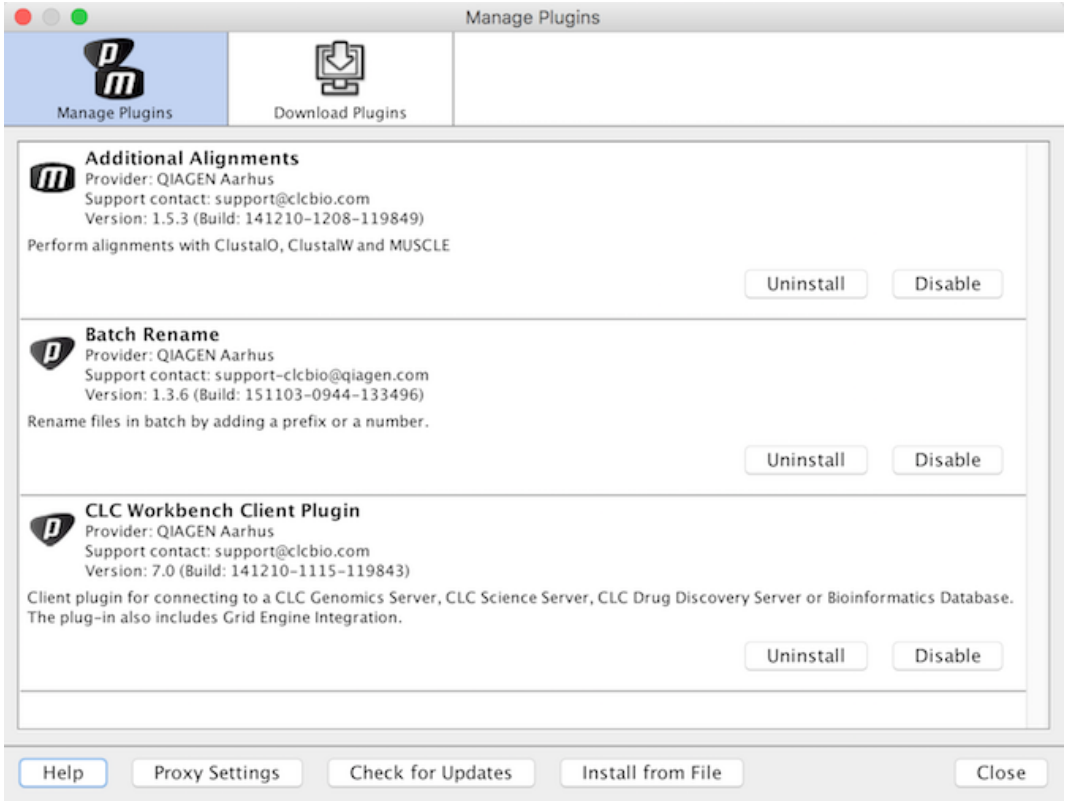

Figure 7.1: *The plugin manager with plugins installed.*

The installed plugins are shown in this dialog. To uninstall:

Click the Annotate Sequence with GFF File | Uninstall

If you do not wish to completely uninstall the plugin 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 plugin will not be uninstalled until the workbench is restarted.