

# **Annotate with GFF file Plugin**

USER MANUAL

# User manual for Annotate with GFF File 22.0

Windows, macOS and Linux

January 4, 2022

**This software is for research purposes only.**

QIAGEN Aarhus  
Silkeborgvej 2  
Prismet  
DK-8000 Aarhus C  
Denmark



# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Annotating a reference genome with genes and transcripts</b>	<b>6</b>
<b>3</b>	<b>Naming of annotations</b>	<b>8</b>
<b>4</b>	<b>Troubleshooting</b>	<b>10</b>
<b>5</b>	<b>Online resources</b>	<b>11</b>
<b>6</b>	<b>Install and uninstall plugins</b>	<b>12</b>
6.1	Installation of plugins . . . . .	12
6.2	Uninstalling plugins . . . . .	13

# Chapter 1

## Introduction

Annotate 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 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://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Download\\_reference\\_genome\\_data.html](http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Download_reference_genome_data.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/](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/](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 (➡📁)**

From 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.

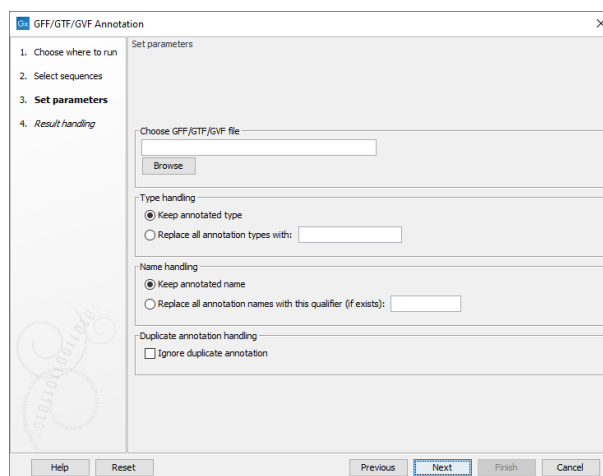


Figure 2.1: Select the GTF file by clicking the browse icon.

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 UCSC.** 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).

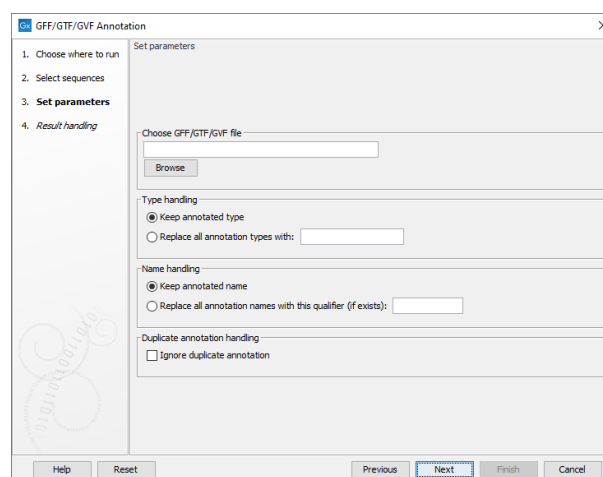


Figure 3.1: You can choose **Replace all annotation names with this qualifier** to specify your own naming convention.

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.

Finally, when **Ignore duplicate annotation** is checked, only one instance of a duplicate annotation will be retained on the annotated sequences.

## 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](http://wiki.geneontology.org/index.php/Reference_Genome_sequence_annotation)


## Chapter 6

# Install and uninstall plugins

Annotate with GFF File is installed as a plugin.

### 6.1 Installation of plugins

**Note:** In order to install plugins and modules, the *CLC Workbench* must be run in administrator mode. On Windows, you can do this by right-clicking the program shortcut and choosing "Run as Administrator". On Linux and Mac, it means you must launch the program such that it is run by an administrative user.

Plugins and modules are installed and uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins** (  ) button in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins...** (  )

The Plugin Manager has two tabs at the top:

- **Manage Plugins** An overview of your installed plugins and modules is provided under this tab.
- **Download Plugins** Plugins and modules available to download and install are listed in this tab.

To install a plugin, click on the **Download Plugins** tab (figure 6.1). Select a plugin. Information about it will be shown in the right hand panel. Click on the **Download and Install** button to install the plugin.

#### Accepting the license agreement

The End User License Agreement (EULA) must be read and accepted as part of the installation process. Please read the EULA text carefully, and if you agree to it, check the box next to the text **I accept these terms**. If further information is requested from you, please fill this in before clicking on the **Finish** button.

#### Installing a cpa file

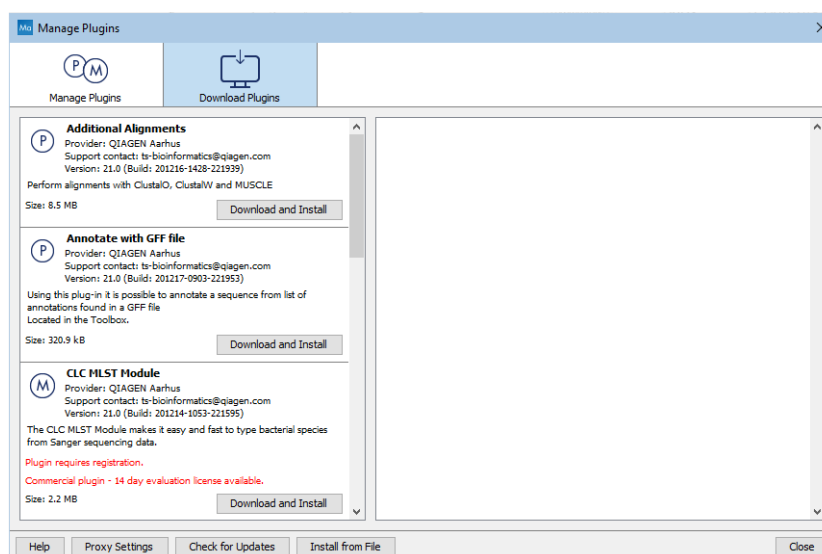


Figure 6.1: Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.

If you have a .cpa installer file for Annotate with GFF File, you can install it by clicking on the **Install from File** button at the bottom of the Plugin Manager.

If you are working on a system not connected to the internet, plugin and module .cpa files can be downloaded from <https://digitalinsights.qiagen.com/products-overview/plugins/> using a networked machine, and then transferred to the non-networked machine for installation.

### Restart to complete the installation

Newly installed plugins and modules will be available for use after restarting the software. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the CLC Workbench.

## 6.2 Uninstalling plugins

Plugins and modules are uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins (P)** button in the top Toolbar, or go to the menu option:

**Utilities | Manage Plugins... (P)**

This will open the Plugin Manager (figure 6.2). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

To uninstall a plugin or module, click on its entry in the list, and click on the **Uninstall** button.

Plugins and modules are not uninstalled until the Workbench is restarted. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the CLC Workbench.

### Disabling a plugin without uninstalling it

If you do not want a plugin to be loaded the next time you start the Workbench, select it in the

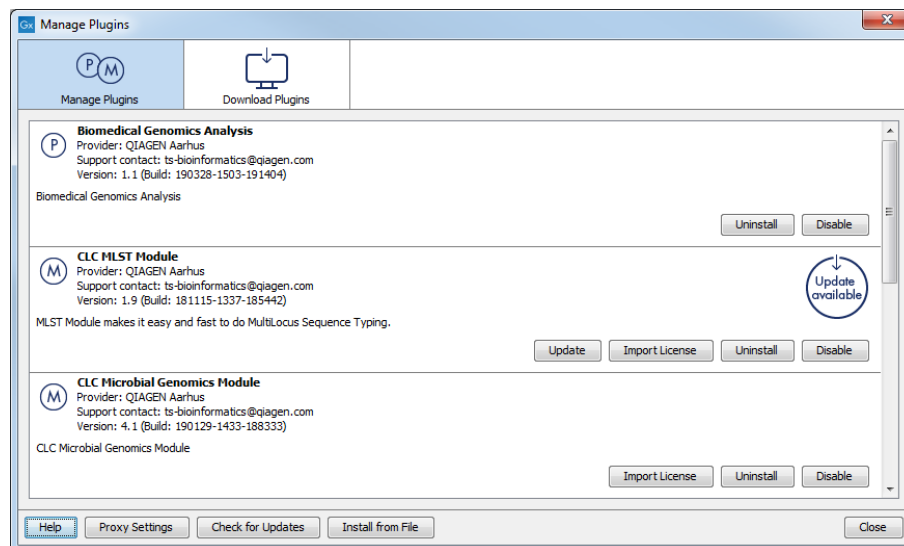


Figure 6.2: Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.

list under the Manage Plugins tab and click on the **Disable** button.