

User manual for

Structural Variation 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 Structural Variation Plug-in **4**

2 The 'Structural Variation Detection' tool **6**

2.1 Introduction to the Structural Variation Detection Tool 6

2.2 Running the Structural Variation Detection Tool 8

2.3 Output 9

2.3.1 SV Table 10

2.3.2 SV Annotations 10

3 The 'Coverage Analysis Detection' tool **12**

3.1 Introduction to the Coverage Analysis Tool 12

3.2 Running the Coverage Analysis Tool 13

3.3 Output 14

3.3.1 Read Coverage Alteration Table 14

3.3.2 Read Coverage Alteration Annotations 15

4 Installation of the Structural Variation Plug-in **16**

5 Uninstall **18**

Chapter 1

Introduction to the Structural Variation Plug-in

The Structural Variation Plug-in is a group of tools designed to facilitate the identification of structural variants like insertions, deletions, duplications, copy number variations (CNVs) and inversions within an unknown chromosome or larger DNA sequence. Structural variations are a common event in all species (12% of the human genome include structural variations). They are important for the evolution of the species and they play an important part in the development of human diseases like cancer and Autism Spectrum Disorder (learn more about it at http://en.wikipedia.org/wiki/Structural_variation). At the moment the extension includes two tools. The 'Structural Variation Detection' tool enables the identification of insertions, deletions,inversions and interchromosomal variations based on paired-end reads mapped on reference chromosomes or larger contigs. The 'Coverage analysis' tool in comparison to that identifies regions on the reference chromosome in which significantly more or less reads have been mapped than expected. Regions with an increase in coverage are so called 'Amplifications'. Amplifications are regions which have been duplicated in the sequence under investigation (sample sequence). In comparison to that regions with a decrease in coverage are putative deleted regions in the sequence under investigation.

In order to make use of the Structural Variation Plug-in you need sequencing reads mapped on a reference genome. To get accurate results the reference sequence should be chosen as similar to the sample sequence as possible. The best results can be achieved if the reference sequence and the sequence reads are from the same organism. In case of the Structural Variation Detection Tool paired-end reads mapped on a reference sequence are required whereas the Coverage Analysis Tool can also deal with single reads mapped on a reference sequence.

The full work flow supported by *CLC Genomics Workbench* and Structural Variation Plug-in is this:

1. Download and import a reference genome into the *CLC Genomics Workbench* (e.g. from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz to get the latest human reference sequence assembly)
2. Optionally annotate reference sequence with known genes/transcripts from external databases like Ensembl or RefSeq (see user manual for the 'Annotate with GFF' plug-in)
3. Import sequencing data (using the High-Thoughput Sequencing Data import).

Remember: In case of the 'Structural Variation Detection' tool paired-end reads are required as input

4. Perform read mapping
5. If you have paired-end sequencing data, run the Structural Variation Detection Tool to identify structural variants (SVs). Otherwise, run the Coverage Analysis Tool to detect regions in reference genome having more (Amplifications) or less (Deletions) mapped reads than expected.
6. Run the SNP and DIP detection to detect variants in the sequencing data
7. Compare identified SVs, SNVs and DIPs to gene annotation using the View Area.

Check the user manual of *CLC Genomics Workbench* for more info on step 1,2,3,4,6 and 7.

Chapter 2

The 'Structural Variation Detection' tool

2.1 Introduction to the Structural Variation Detection Tool

The Structural Variation Detection Tool is designed to identify large structural variations like insertions, deletions, inversions and inter-chromosomal structural variations from a paired-end read mapping. The algorithm used in the plug-in looks for regions in the reference sequence with significant more anomalous read pairs than expected using a Poisson distribution. Anomalous read pairs have a shorter or longer distance between the mapped reads than some cutoff, one of the reads of a pair is reversed during the mapping or the reads are mapped to two different chromosomes.

In this plug-in the cutoff is automatically calculated from paired-end reads, which have not been broken by the mapper. The formula is the following:

Upper cutoff: median(disAllPairs) + 3 * mad(disAllPairs)
Lower cutoff: median(disAllPairs) - 3 * mad(disAllPairs)

disAllPairs = All distances between paired reads in the dataset
mad() = Absolute deviation from median

Mean of the Poisson distribution is the mean number of anomalous read pairs spanning all regions on the reference sequence. For each of the following signatures a separate Poisson distribution is calculated:

- Insertion: Distance between paired reads is less than lower cutoff
- Deletion : Distance between paired reads is more than upper cutoff
- Inversion: On of the paired reads has been reversed during the mapping process
- Inter-chromosomal structural variations: Reads are mapped to two different chrmosomes or contigs

For each signature regions on the reference sequence, in which more anomalous read pairs spanning the region than expected (given the Poisson distribution and a user selected p-value cutoff) are annotated according to the signature.

In a next step deletions are verified based on the average read coverage in the region of the variation. Only deletions are reported, in which the read coverage in this region is significantly lower (p-value 0.05) than the average coverage of the complete chromosome.

Afterwards inter-chromosomal structural variations are discovered by identifying region on two chromosomes, which are connected by ad least two read-pairs and. Only regions are reported, which connect with less than two other regions. So many repeat regions are filtered out.

In the last step of the algorithm annotations are merged if they overlap or the distance between them is less than the median distance between all paired reads. For a complete list of annotations see chapter 2.3.2

In order to make use of the Structural Variation Detection Tool you need to have a paired-end read mapping.

The full work flow supported by *CLC Genomics Workbench* and Structural Variation Detection Tool is this:

1. Download and import a reference genome into the *CLC Genomics Workbench* (e.g. from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz` to get the latest human reference sequence assembly)
2. Optionally annotate reference sequence with known genes/transcripts from external databases like Ensembl or RefSeq (see user manual for the 'Annotate with GFF' plugin)
3. Import sequencing data (using the High-Thoughput Sequencing Data import).
4. Perform read mapping with very stringent parameter settings (0.9 similarity and 0.9 sequence overlap between read and reference sequence)
5. Run the Structural Variation Detection Toolto detect structural variations.
6. Compare identified structural variations to gene annotation to identify functional relevant genes/transcripts, which have been duplicated, deleted or are part of an inversion in the sample under investigation. Often also partial deleted genes, parts of the promotor region and deletions in intron are of interest, because these can cause changes in expression or function of the gene. Insertion of unknown DNA could be a hint, that DNA from a different chromosome (more common event) or a horizontal gene transfer had taken place in the sample sequence.

Check the user manual of *CLC Genomics Workbench* for more info on step 1,2,3,4 and 6.

2.2 Running the Structural Variation Detection Tool

To run the Structural Variation Detection Tool:

Toolbox | **High-throughput Sequencing** ( ) | **Structural Variation** ( ) | **Structural Variation Detection** ( )

This opens the dialog shown in figure 2.1.

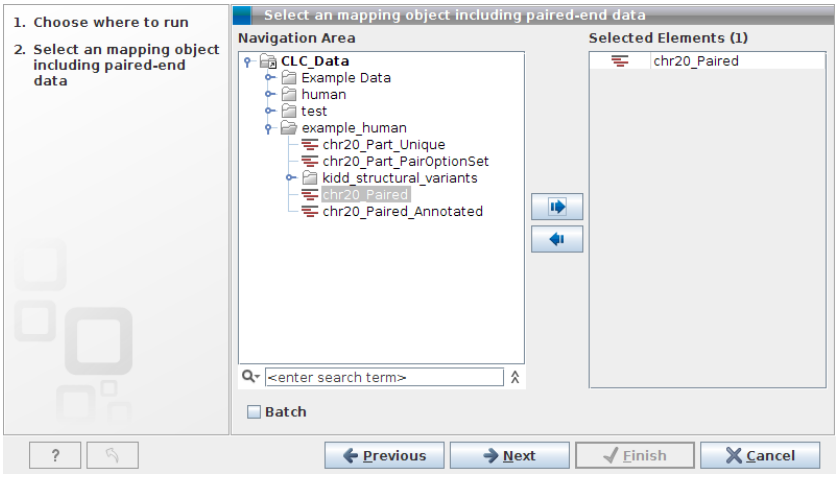


Figure 2.1: *Select paired-end read mapping results.*

Select a paired-end read mapping result and click **Next**. This opens the dialog shown in figure 2.2.

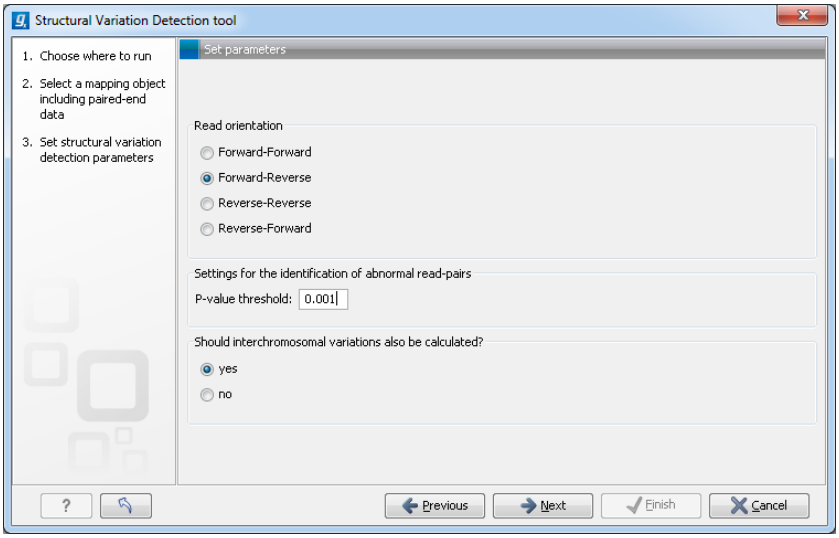


Figure 2.2: *Specify the original orientation of the paired-reads and select a p-value cutoff.*

Please select the original orientation of your paired-end reads (you can get it from the experimental protocol used for library construction).

- FF: Orientation of both read is in forward direction (on '+' strand)
- BB: Orientation of both read is in backward direction (on '-' strand)
- FB: Orientation of one read is in forward direction and the other read is in backward direction

- BF: Orientation of one read is in backward direction and the other read is in forward direction

Furthermore, please choose a p-value cutoff. The p-value is the probability that more anomalous reads spanning a specific region than actual discovered. If you like to be more stringent use a lower p-value cutoff.

As a new feature you can also select now if you like to calculate inter-chromosomal variations. Please note, that true consumes more memory and takes more time.

Click **Next** and **Finish**.

2.3 Output

This plugin will generate two kinds of output:

SV Table The table has one row for each identified strutural variation inclusive location and significance p-value.

SV Annotations For the visual inspection and the comparison with known gene/transcript annotations, all identified structural variations are also annotated on the reference sequence.

The output is described in the two sections below.

2.3.1 SV Table

The table includes the following columns:

Sequence name The name is taken from the reference sequence used for mapping.

Feature type The type of annotation

Start position The start position of this region on the reference sequence.

End position The end position of this region on the reference sequence.

Length The length of the region.

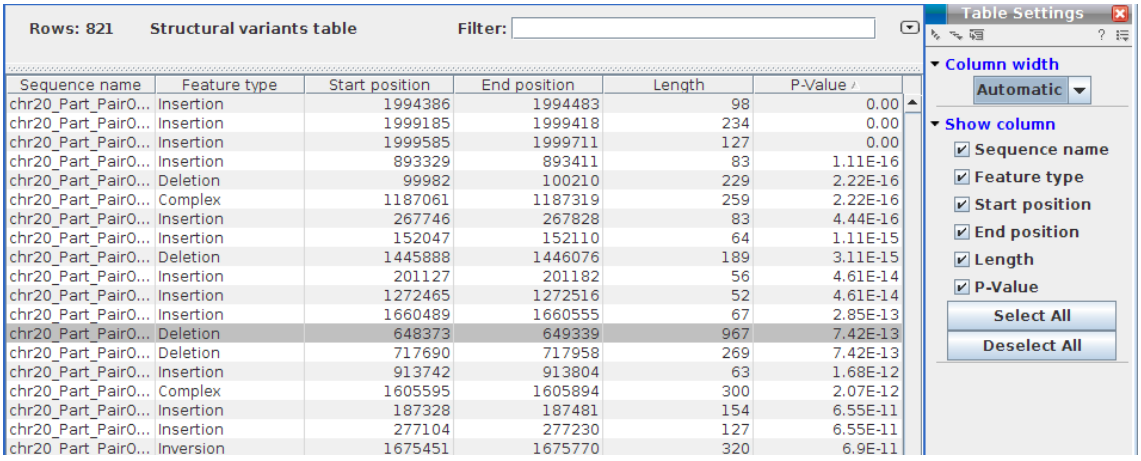
Mate chr. In case of an inter-chromosomal variations the other chromosome connected to this variation via read-pairs.

Mate start. In case of an inter-chromosomal variation the chromosome start point connected to this variation via read-pairs.

Mate stop. In case of an inter-chromosomal variation the chromosome end point connected to this variation via read-pairs.

P-Value The calculated significance p-value for this region.

Here is an example for a SV table (No inter-chromosomal variations found):



Sequence name	Feature type	Start position	End position	Length	P-Value
chr20_Part_Pair0...	Insertion	1994386	1994483	98	0.00
chr20_Part_Pair0...	Insertion	1999185	1999418	234	0.00
chr20_Part_Pair0...	Insertion	1999585	1999711	127	0.00
chr20_Part_Pair0...	Insertion	893329	893411	83	1.11E-16
chr20_Part_Pair0...	Deletion	99982	100210	229	2.22E-16
chr20_Part_Pair0...	Complex	1187061	1187319	259	2.22E-16
chr20_Part_Pair0...	Insertion	267746	267828	83	4.44E-16
chr20_Part_Pair0...	Insertion	152047	152110	64	1.11E-15
chr20_Part_Pair0...	Deletion	1445888	1446076	189	3.11E-15
chr20_Part_Pair0...	Insertion	201127	201182	56	4.61E-14
chr20_Part_Pair0...	Insertion	1272465	1272516	52	4.61E-14
chr20_Part_Pair0...	Insertion	1660489	1660555	67	2.85E-13
chr20_Part_Pair0...	Deletion	648373	649339	967	7.42E-13
chr20_Part_Pair0...	Deletion	717690	717958	269	7.42E-13
chr20_Part_Pair0...	Insertion	913742	913804	63	1.68E-12
chr20_Part_Pair0...	Complex	1605595	1605894	300	2.07E-12
chr20_Part_Pair0...	Insertion	187328	187481	154	6.55E-11
chr20_Part_Pair0...	Insertion	277104	277230	127	6.55E-11
chr20_Part_Pair0...	Inversion	1675451	1675770	320	6.9E-11

Figure 2.3: *The table output with detailed information on each region.*

The table can be exported ( ) in Excel format.

For a description of the annotations see section 2.3.2

2.3.2 SV Annotations

For the visual inspection and comparison to known gene/transcripts or other kind of annotations, all region are also annotated on the reference sequence.
After running the plugin you can see them as colored and labeled boxes above the reference sequence in the View Area or as seperate track in the GenomicsGateway.

The following annotations can be discovered:

- Insertion: Insertion of DNA at the start or end of this annotation
- Deletion: This part of the chromosome has been deleted in the sample sequence.
- Inversion: This part of the reference sequence has been inverted in the sample sequence
- Interchromosomal Variations: Translocations or duplications over two chromosomes.

Here is an example for SV annotations:

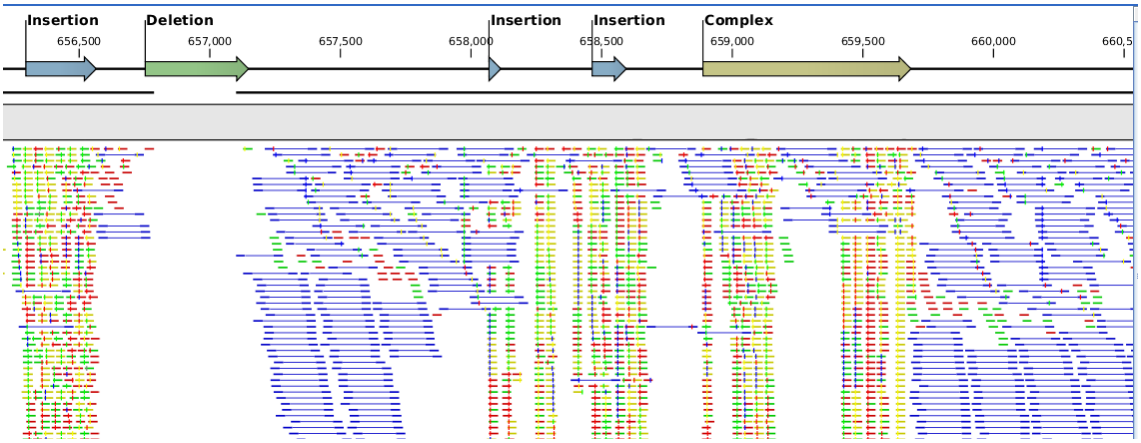


Figure 2.4: *For each identified structural variation an annotation is created on top of the reference sequence.*

Chapter 3

The 'Coverage Analysis Detection' tool

3.1 Introduction to the Coverage Analysis Tool

The Coverage Analysis Tool is designed to detect changes in read coverage like amplifications (read number is larger than expected) or deletions (read number is smaller than expected) from a read mapping. This could give hints to copy number variations.
Therefore the algorithm uses a Poisson distribution calculated from the mean read coverage over the whole chromosome/contig and a user given p-value cutoff to identify regions in which significant (p-value from Poisson distribution is smaller than p-value from the user) more reads have been mapped to the reference sequence than expected from the distribution. In case of amplifications the p-value is the probability that more reads covering this region. In the case of deletions the p-value defines the probability to see less reads than discovered. An annotation is added if the region discovered is minimum 100bp in size.
At the end of the algorithm annotations are merged if they have a distance to each other of less than 100bp.

In order to make use of the Coverage Analysis Tool you need to have a read mapping.

The full work flow supported by *CLC Genomics Workbench* and Coverage Analysis Tool is this:

1. Download and import a reference genome into the *CLC Genomics Workbench* (e.g. from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz` to get the latest human reference sequence assembly)
2. Optionally annotate reference sequence with known genes/transcripts from external databases like Ensembl or RefSeq (see user manual for the 'Annotate with GFF' plugin)
3. Import sequencing data (using the High-Thoughput Sequencing Data import).
4. Perform read mapping
5. Run the Coverage Analysis Tool to detect regions in reference genome having more (Amplifications) or less (Deletions) mapped reads than expected.
6. Compare identified amplifications and deletions to gene annotation to identify functional relevant genes/transcripts, which have been duplicated (amplification) or deleted in the sample under investigation. Often also partial deleted genes, parts of the promotor region

and deletions in intron are of interest, because these can cause changes in expression or function of the gene.

Check the user manual of *CLC Genomics Workbench* for more info on step 1,2,3,4 and 6.

3.2 Running the Coverage Analysis Tool

To run the Coverage Analysis Tool analysis:

Toolbox | **High-throughput Sequencing** ( ) | **Structural Variation** ( ) | **Coverage Analysis Tool** ( )

This opens the dialog shown in figure 3.1.

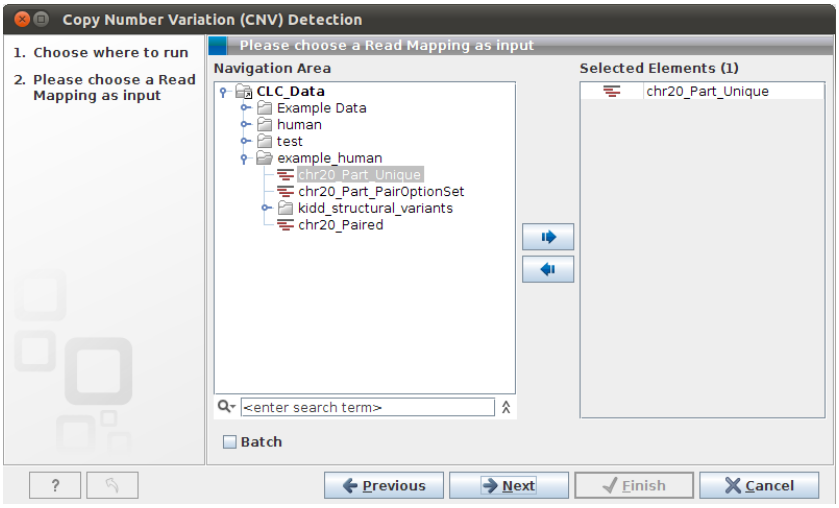


Figure 3.1: *Select read mapping results.*

Select a read mapping results and click **Next**. This opens the dialog shown in figure 3.2.

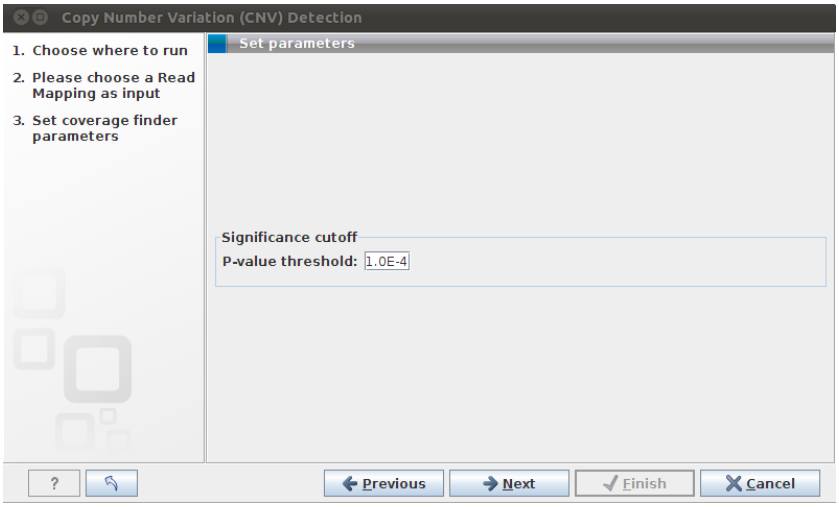


Figure 3.2: *Specify the p-value cutoff.*

Please choose the p-value cutoff. Only regions with a significance below that p-value will be reported at the end. In case bigger changes in read coverage are expected a very small p-value

should be choosen to get accurate results.

Click **Next** and **Finish**.

3.3 Output

This plugin will generate two kinds of output:

Read Coverage Alteration Table The table has one row for each identified read coverage alteration inclusive location and significance p-value.

Read Coverage Alteration Annotations For the visual inspection and the comparison with known gene/transcript annotations, all identified read coverage alterations are also annotated on the reference sequence.

The output is described in the two sections below.

3.3.1 Read Coverage Alteration Table

The table includes the following columns:

Sequence name The name is taken from the reference sequence used for mapping.

Feature type The type of annotation (Amplification or Deletion)

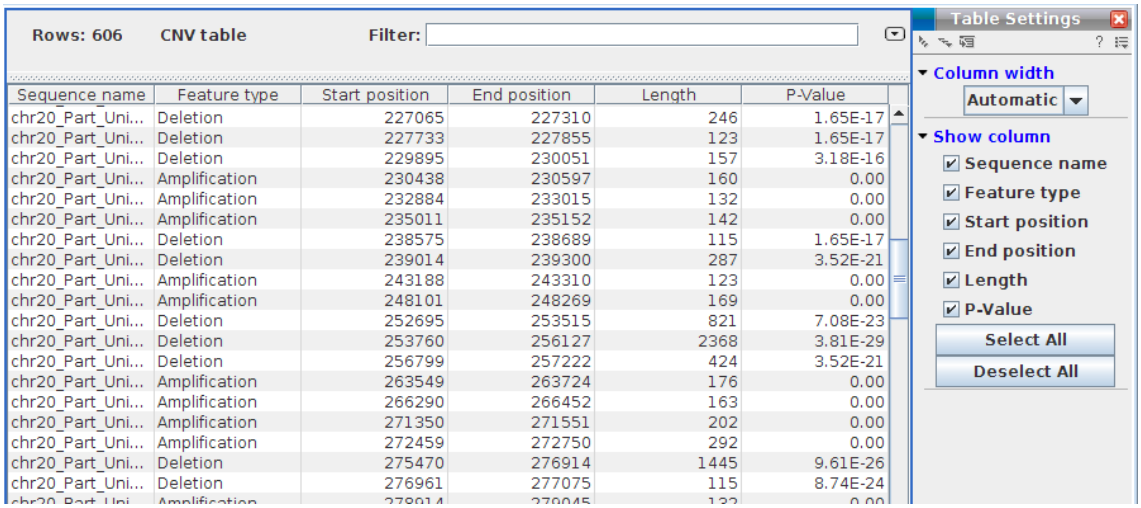
Start position The start position of this region on the reference sequence.

End position The end position of this region on the reference sequence.

Length The length of the region.

P-Value The calculated significance p-value for this region.

Here is an example for a Read Coverage Alteration table:



Sequence name	Feature type	Start position	End position	Length	P-Value
chr20_Part_Uni...	Deletion	227065	227310	246	1.65E-17
chr20_Part_Uni...	Deletion	227733	227855	123	1.65E-17
chr20_Part_Uni...	Deletion	229895	230051	157	3.18E-16
chr20_Part_Uni...	Amplification	230438	230597	160	0.00
chr20_Part_Uni...	Amplification	232884	233015	132	0.00
chr20_Part_Uni...	Amplification	235011	235152	142	0.00
chr20_Part_Uni...	Deletion	238575	238689	115	1.65E-17
chr20_Part_Uni...	Deletion	239014	239300	287	3.52E-21
chr20_Part_Uni...	Amplification	243188	243310	123	0.00
chr20_Part_Uni...	Amplification	248101	248269	169	0.00
chr20_Part_Uni...	Deletion	252695	253515	821	7.08E-23
chr20_Part_Uni...	Deletion	253760	256127	2368	3.81E-29
chr20_Part_Uni...	Deletion	256799	257222	424	3.52E-21
chr20_Part_Uni...	Amplification	263549	263724	176	0.00
chr20_Part_Uni...	Amplification	266290	266452	163	0.00
chr20_Part_Uni...	Amplification	271350	271551	202	0.00
chr20_Part_Uni...	Amplification	272459	272750	292	0.00
chr20_Part_Uni...	Deletion	275470	276914	1445	9.61E-26
chr20_Part_Uni...	Deletion	276961	277075	115	8.74E-24

Figure 3.3: *The table output with detailed information on each region.*

The table can be exported ( ) in Excel format.

3.3.2 Read Coverage Alteration Annotations

For the visual inspection and comparison to known gene/transcripts or other kind of annotations, all region are also annotated on the reference sequence.
After running the plugin you can see them as colored and labeled boxes above the reference sequence in the View Area.

Here is an example for Read Coverage Alteration annotations:



Figure 3.4: *For each identified region an annotation is created on top of the reference sequence.*

Chapter 4

Installation of the Structural Variation Plug-in

The Structural Variation 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 4.1).

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

If the Structural Variation 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.

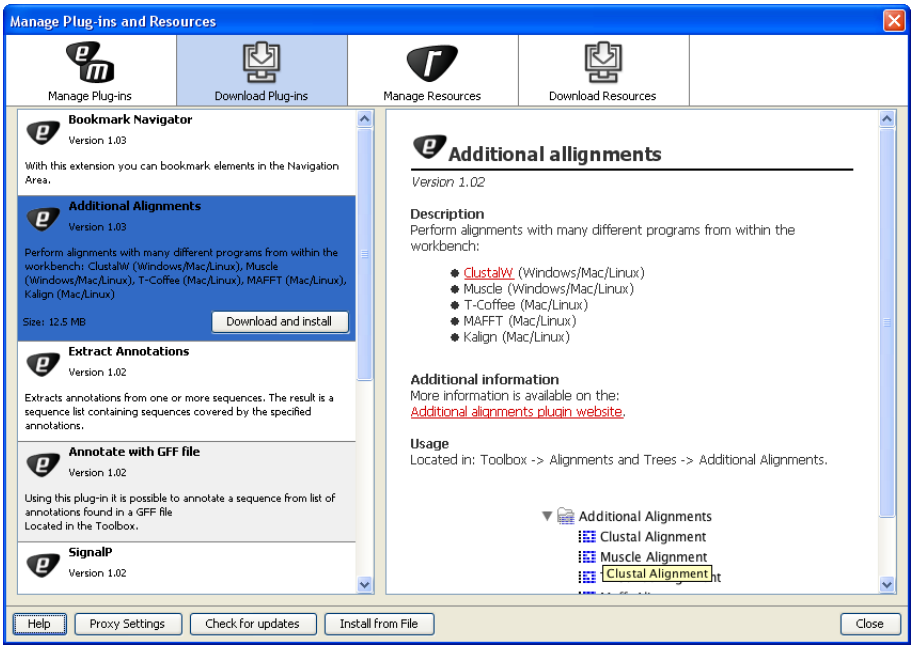


Figure 4.1: *The plug-ins that are available for download.*

Chapter 5

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 5.1.

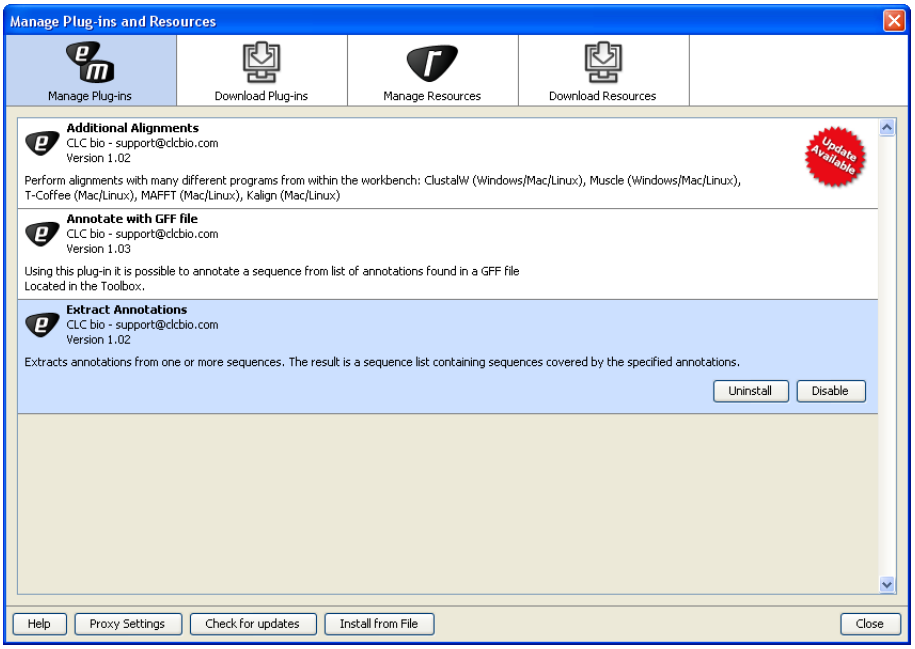


Figure 5.1: *The plug-in manager with plug-ins installed.*

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

Click the Structural Variation 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.