

Targeted Resequencing Extension

User manual

User manual for

Targeted Resequencing Extension 1.0.5

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 targeted resequencing extension 4

2 Running the targeted resequencing analysis 5

3 Coverage report 7

4 Coverage table 9

5 Per base coverage table 11

6 Installation 12

7 Uninstall 14

Chapter 1

Introduction to the targeted resequencing extension

The targeted resequencing extension is designed to facilitate the quality control of targeted re-sequencing experiments using a variety of technologies such as RainDance, Agilent (SureSelect) and Nimblegen (sequence capture) and Febit in combination with high-throughput sequencing. This is also known as sequence enrichment, sequence capture or amplicon sequencing (learn more at http://en.wikipedia.org/wiki/Exome_sequencing#Target-enrichment_strategies).

In order to make use of the targeted resequencing extension you need to annotate your reference genome with the targeted regions. This can be done e.g. using a GFF file via the Annotate with GFF plug-in (see http://www.clcbio.com/annotate-with-gff). You can obtain the annotation file from the provider of the capture platform.

The full work flow supported by *CLC Genomics Workbench* and targeted resequencing extension is this:

1. Import or download reference genome (import the file from http://download.clcbio.com/testdata/human_refseq.zip into the *CLC Genomics Workbench* to get started with e.g. the human genome)
2. Annotate reference genome with targeted regions
3. Optionally annotate reference genome with known SNPs (see user manual for Annotate with GFF plug-in)
4. Import sequencing data (using the High-throughput Sequencing Data import)
5. Perform Read mapping
6. Perform Targeted Resequencing Analysis to get an overview of the quality of the results
7. Run SNP and DIP detection to detect variants in the sequencing data

Check the user manual of *CLC Genomics Workbench* for more info on step 1, 4, 5 and 7.

The targeted resequencing analysis will inspect the annotated regions and provide coverage statistics for each one.

Chapter 2

Running the targeted resequencing analysis

To run the targeted resequencing analysis:

Toolbox | **High-throughput Sequencing** ( ) | **Targeted Resequencing Analysis** ( )

This opens the dialog shown in figure 2.1.

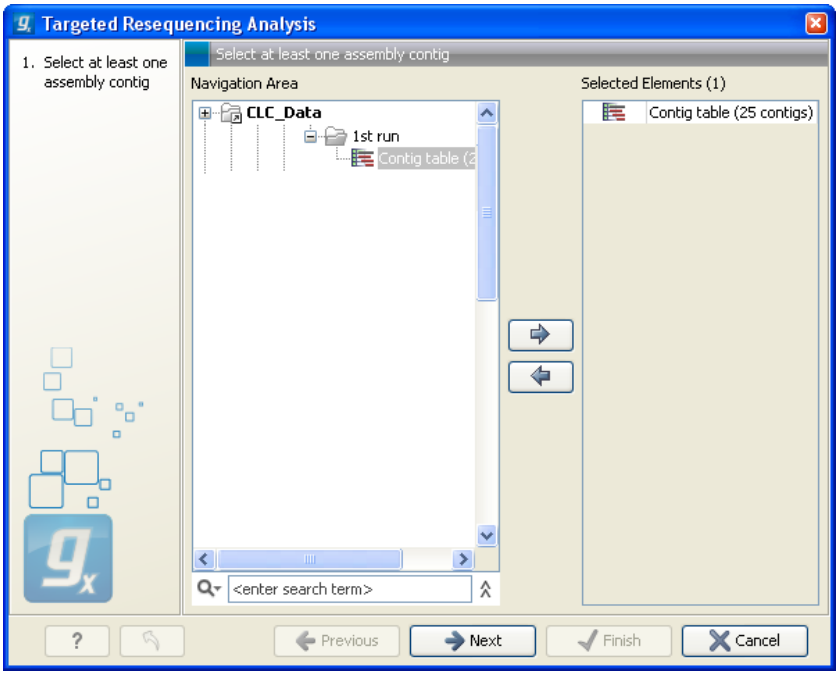


Figure 2.1: *Select mapping results.*

Select mapping results and click **Next**. This opens the dialog shown in figure 2.2.

Click the **Browse** ( ) icon to select an annotation type that defines the targeted regions of your reference genome. At the top, you can specify coverage boundaries for the coverage report (see below).

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

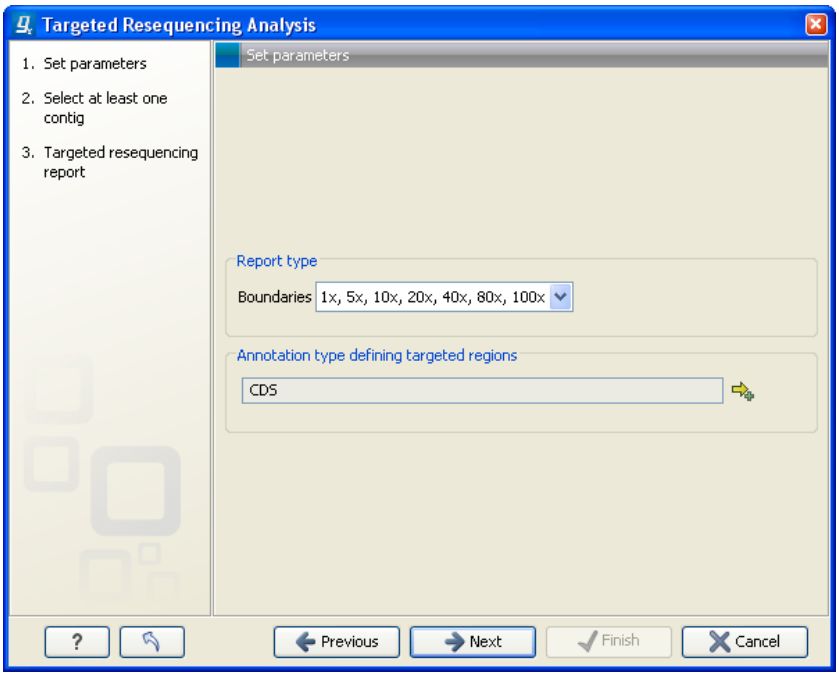


Figure 2.2: *Specifying the annotation type for the targeted regions.*

This plug-in will generate two kinds of output:

Coverage report The report shows various statistics on coverage of all the targeted regions. This can be used to get an overall impression of the results in terms of coverage.

Coverage table The table has one row per targeted region and shows details on each targeted region.

Per base coverage table (optional) The table has one row per position in the targeted regions (can potentially be quite large).

The output is described in the two sections below.

Note that paired reads are treated as two separate reads in the counting and coverage calculations.

Chapter 3

Coverage report

An example of a coverage report is shown in figure 3.1)

This figure show the top of the report. The full content is explained below:

Coverage summary This table shows overall coverage information.

Mean coverage The average coverage across all the targeted regions.

Total bases in targeted regions The sum of the size of all the targeted regions (this means it is calculated from the annotations alone and is not influenced by the reads).

Fold coverage This shows the percentage of the targeted regions that are covered by this many reads. The intervals can be specified in the dialog when running the analysis. Default is 1, 5, 10, 20, 40, 80, 100 times. In figure 3.1 this means that 35.83 % of the positions on the reference sequence in the regions of interest (i.e. the targeted regions) are covered by at least 5 reads.

Targeted regions overview For each reference sequence, the following information is displayed:

Reference The name of the reference sequence.

Total mapped reads The total number of mapped reads on the reference, including reads mapped outside the target regions.

Mapped reads in targeted region Total number of reads in the targeted regions. Note that if there are overlapping regions, reads covered by two regions will be counted twice. If a read is only partially inside a targeted region, it will still count as a full read.

Specificity The percentage of the total mapped reads that are in the targeted regions.

Base coverage This table and corresponding graph is similar to the table at the top of the report. It shows for each fold of the mean coverage (0.1 to 1.0) how many bases of the targeted regions are covered. Because this is based on mean coverage, the numbers can be used for cross-sample comparison of the quality of the experiment.

Mean coverage per position Three plots listing the mean coverage for each position of the targeted regions. The first plot shows only forward reads, the second plot shows only reverse reads and finally the third plot shows all reads. In each plot, The scale on the X axis is the raw nucleotide numbers with a max of 5000 bp. That is, the data point to the very left represents the average coverage of the first position of all targeted regions. The

1 Coverage by read depth

coverage	
Mean	8,84
Total bases	91.093.444
1x	43,17%
5x	35,83%
10x	29,24%
20x	18,05%
40x	3,90%
80x	0,32%
100x	0,18%

1.1 Contig overview

Cluster	total mapped reads	mapped reads in targeted region	specificity
NC_000024 mapping	60.377	5.527	0,092
NC_000023 mapping	764.193	412.665	0,540
NC_000022 mapping	377.983	141.716	0,375
NC_000021 mapping	290.643	95.196	0,328
NC_000020 mapping	466.923	201.006	0,430
NC_000019 mapping	885.683	545.388	0,616
NC_000018 mapping	400.239	136.736	0,342
NC_000017 mapping	874.022	447.984	0,513
NC_000016 mapping	759.846	283.888	0,374
NC_000015 mapping	597.797	280.205	0,469

Figure 3.1: *The report with overviews of mapped reads.*

next data point is then for the second position of all the targeted regions and so on. These plots can be used to evaluate whether there is a general tendency towards lower coverage at the end of the targeted region, and whether there is a bias in terms of forward and reverse reads coverage. So, for example, one could evaluate biases in the 5' end of reads generated using a forward primer using the first plot, and use the second plot to consider biases in the 3' end of reads generated from the reverse primer.

Read count per %GC The plot shows the GC content of the reference sequence on the X-axis and the number of mapped reads on the Y-axis. This plot will show if there is a basis caused by higher GC-content in the sequence.

Chapter 4

Coverage table

This plug-in will generate two kinds of output: the following (an example is shown in figure 4.1):

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

Type The type of annotation (is relevant if several types were selected).

Name The annotation name derived from the annotation on the reference.

Length The length of the region.

Read count Number of reads that cover this region. Note that reads that only cover the region partially are also included.

Region The targeted region.

Minimum coverage The lowest coverage in the region.

Maximum coverage The highest coverage in the region.

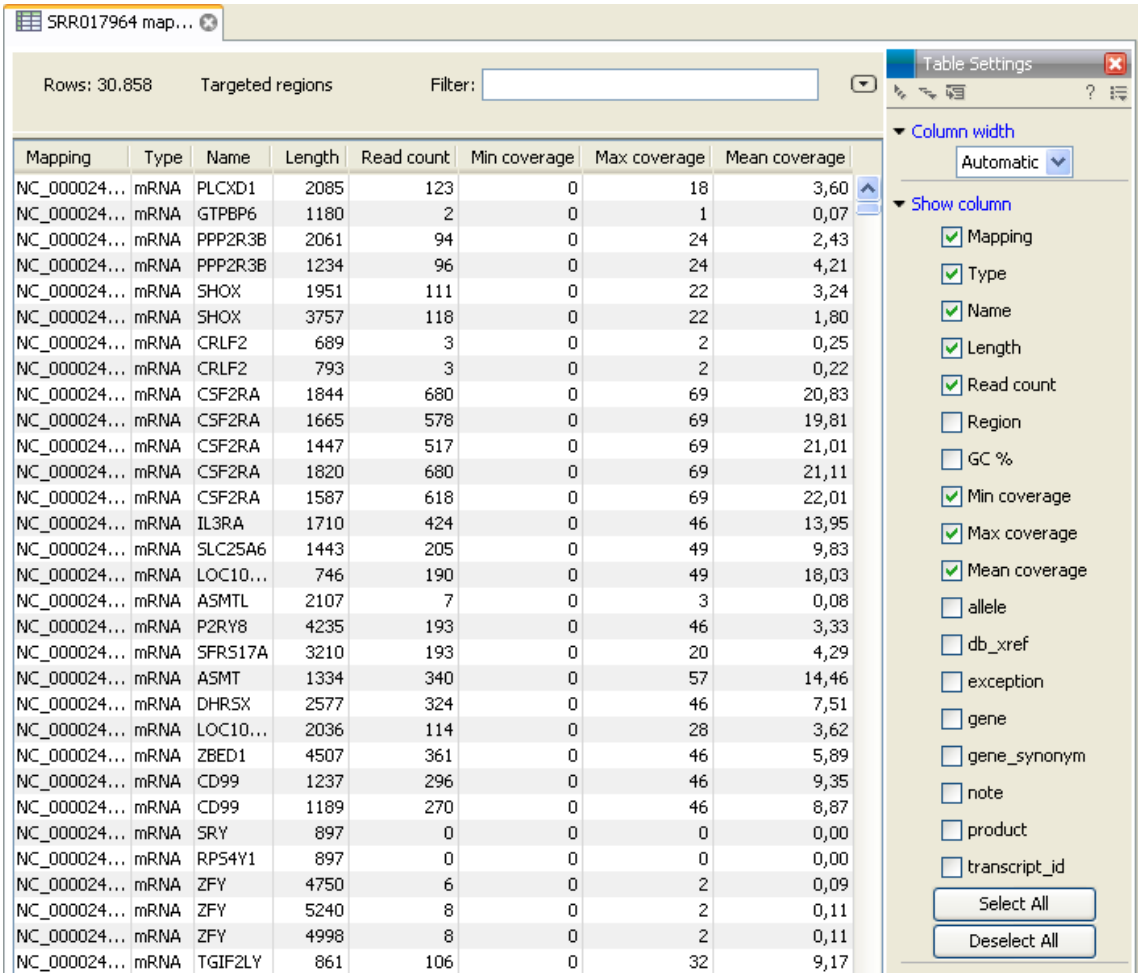
Mean coverage The average coverage in the region.

%GC The GC content of the region. (This is not shown per default but can be checked in the **Side Panel**).

Note that there is additional information that derive from the qualifiers on the annotations. In the figure above, there is information like db_xref and locus_tag which can be displayed in the table as well.

If you make a split view of the table and the mapping, you can click in the table to navigate the mapping view to the targeted regions. You can also use the filter at the top of the table to make easy interpretations on the data (e.g. the number of regions having less than 10x mean coverage).

The table can be exported () in Excel format.



Mapping	Type	Name	Length	Read count	Min coverage	Max coverage	Mean coverage
NC_000024...	mRNA	PLCXD1	2085	123	0	18	3,60
NC_000024...	mRNA	GTPBP6	1180	2	0	1	0,07
NC_000024...	mRNA	PPP2R3B	2061	94	0	24	2,43
NC_000024...	mRNA	PPP2R3B	1234	96	0	24	4,21
NC_000024...	mRNA	SHOX	1951	111	0	22	3,24
NC_000024...	mRNA	SHOX	3757	118	0	22	1,80
NC_000024...	mRNA	CRLF2	689	3	0	2	0,25
NC_000024...	mRNA	CRLF2	793	3	0	2	0,22
NC_000024...	mRNA	CSF2RA	1844	680	0	69	20,83
NC_000024...	mRNA	CSF2RA	1665	578	0	69	19,81
NC_000024...	mRNA	CSF2RA	1447	517	0	69	21,01
NC_000024...	mRNA	CSF2RA	1820	680	0	69	21,11
NC_000024...	mRNA	CSF2RA	1587	618	0	69	22,01
NC_000024...	mRNA	IL3RA	1710	424	0	46	13,95
NC_000024...	mRNA	SLC25A6	1443	205	0	49	9,83
NC_000024...	mRNA	LOC10...	746	190	0	49	18,03
NC_000024...	mRNA	ASMTL	2107	7	0	3	0,08
NC_000024...	mRNA	P2RY8	4235	193	0	46	3,33
NC_000024...	mRNA	SFRS17A	3210	193	0	20	4,29
NC_000024...	mRNA	ASMT	1334	340	0	57	14,46
NC_000024...	mRNA	DHRSX	2577	324	0	46	7,51
NC_000024...	mRNA	LOC10...	2036	114	0	28	3,62
NC_000024...	mRNA	ZBED1	4507	361	0	46	5,89
NC_000024...	mRNA	CD99	1237	296	0	46	9,35
NC_000024...	mRNA	CD99	1189	270	0	46	8,87
NC_000024...	mRNA	SRY	897	0	0	0	0,00
NC_000024...	mRNA	RPS4Y1	897	0	0	0	0,00
NC_000024...	mRNA	ZFY	4750	6	0	2	0,09
NC_000024...	mRNA	ZFY	5240	8	0	2	0,11
NC_000024...	mRNA	ZFY	4998	8	0	2	0,11
NC_000024...	mRNA	TGIF2LY	861	106	0	32	9,17

Figure 4.1: *The table output with detailed information on each region.*

Chapter 5

Per base coverage table

This table is optional since it can be quite big and time-consuming to create. It shows information on each reference base position in the targeted regions:

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

Reference position The type of annotation (is relevant if several types were selected).

Name The annotation name derived from the annotation on the reference.

Target region position The position in the targeted region, starting from 1.

Coverage Coverage at this position (i.e. the number of reads aligned to this position on the reference

Reference base The base on the reference sequence.

Consensus base The base of the consensus sequence. In the dialog where you choose to create the per base coverage table, you can also define a coverage threshold. For positions with coverage lower than this number, the consensus will always be reported as N.

The table can be exported () in Excel and csv format.

Chapter 6

Installation

The targeted resequencing extension 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 6.1).

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 targeted resequencing extension and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the targeted resequencing extension 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.

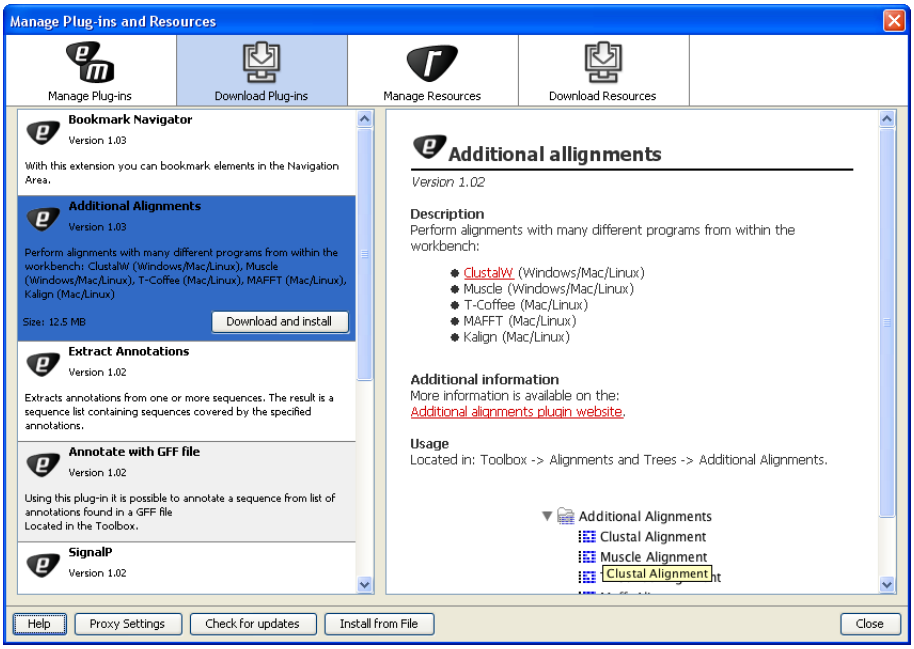


Figure 6.1: *The plug-ins that are available for download.*

Chapter 7

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 7.1.

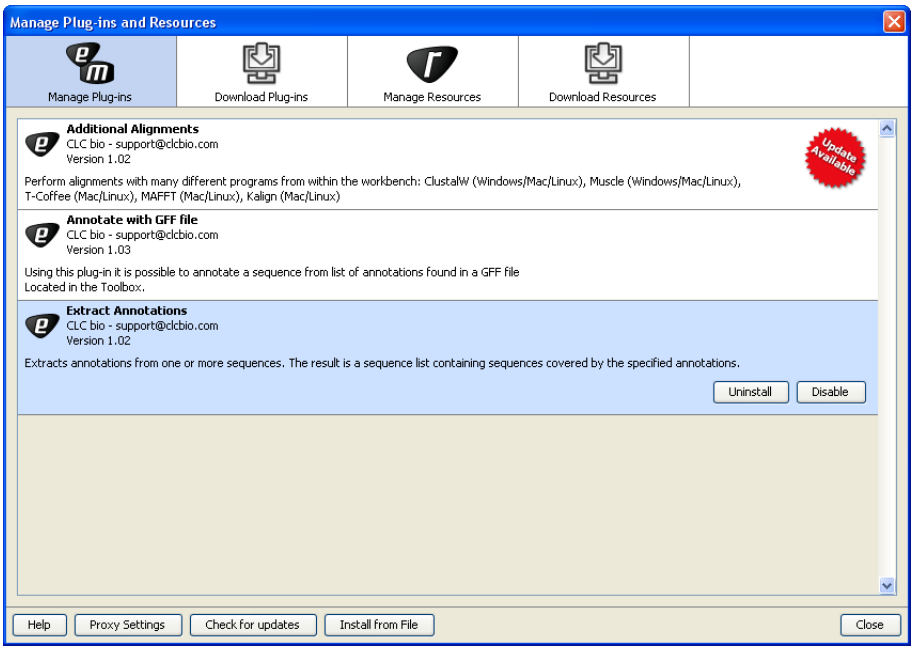


Figure 7.1: *The plug-in manager with plug-ins installed.*

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

Click the targeted resequencing extension | **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.