

User manual for

Duplicate Read Removal Plug-in 1.0 Beta

Windows, Mac OS X and Linux

September 8, 2011

This software is for research purposes only.

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

Contents

1 Introduction to the Duplicate Read Removal Plug-in Beta **4**

1.1 Looking for neighbors 4

1.2 Sequencing errors in duplicates 5

1.3 Paired data 6

1.4 Known limitations 7

2 Running the Duplicate Read Removal Plug-in Beta **8**

3 Installation of the Duplicate Read Removal Plug-in Beta **10**

4 Uninstall **12**

Chapter 1

Introduction to the Duplicate Read Removal Plug-in Beta

The duplicate read removal tool is designed to filter out duplicate reads. This tool is specifically well-suited to handle duplicate reads coming from PCR amplification errors which can have a negative effect because a certain sequence is represented in artificially high numbers.

The purpose of the tool is to reduce the data set to include only one copy of the duplicate sequence. The challenge is to achieve this without removing identical or almost identical reads that would arise from high coverage of certain regions, e.g. repeat regions or highly expressed exons from transcriptome sequencing. The algorithm takes sequencing errors into account (see below).

The approach taken here is based on the raw sequencing data without any knowledge about how they map to a reference sequence. This means that this is well-suited for both de novo assembly and resequencing purposes.

1.1 Looking for neighbors

An example of a read duplication can be easily distinguished when mapping reads to a reference sequence as shown in figure 1.1.

The typical signature is a lot of reads starting at the same position, resulting in a sudden rise in coverage, and all reads have the same orientation (denoted by the color). In a normal data set, you will also see fluctuations in coverage as shown in figure 1.2, but they lack the two important features of duplicate reads: they do not all start at exactly the same position, and they are from different strands.

The duplicate reads tool works directly on the sequencing reads, so there is no need to map the data to a reference genome first (figures 1.2 and 1.2 show the reads mapped for illustration purposes). In short, the algorithm will look for "neighboring" reads (i.e. reads that share most of the read sequence but with a small offset) and use these to determine whether there is generally high coverage for this sequence. If this is not the case, the read in question will be marked as a duplicate.

For certain sequencing platforms such as 454, the reads will have varying lengths - and this is

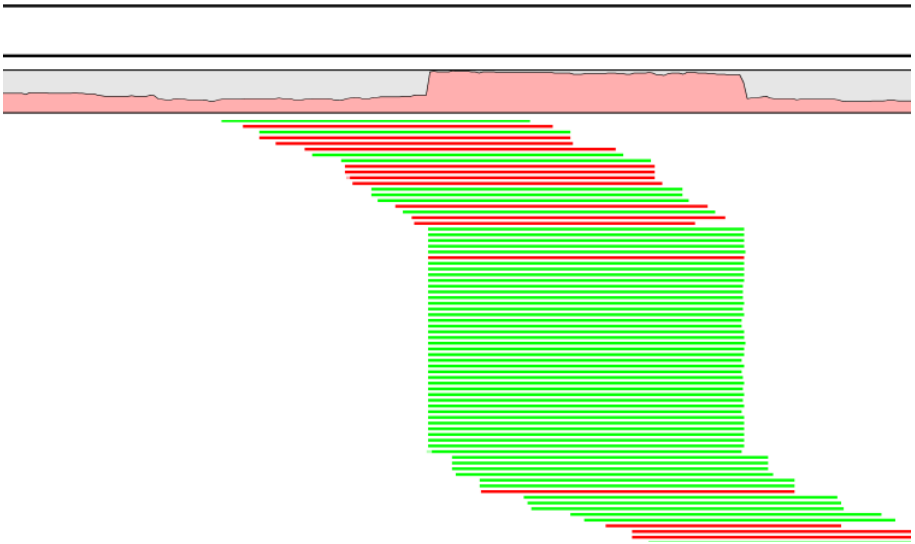


Figure 1.1: *Mapped reads with a set of duplicate reads, the colors denote the strand (green is forward and red is reverse).*

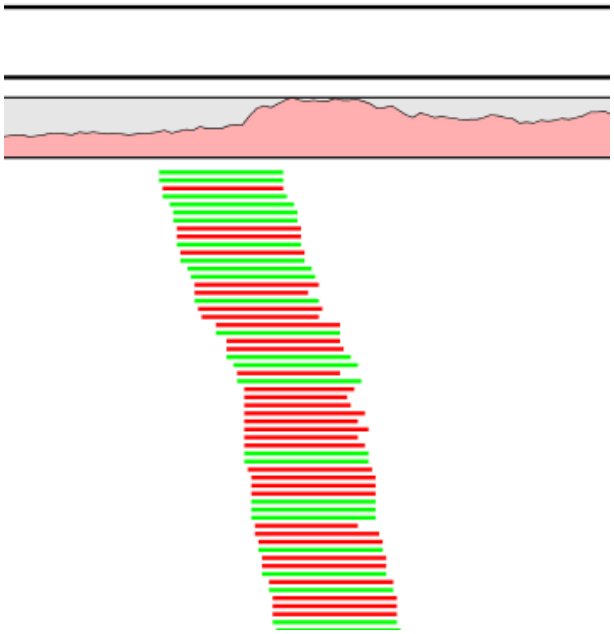


Figure 1.2: *Rise in coverage.*

taken into account by the algorithm as well.

SOLiD data is also supported, taking into account the dual base coding features of SOLiD data (color space).

1.2 Sequencing errors in duplicates

It is important to take sequencing errors into account when filtering duplicate reads. Imagine an example with 100 duplicates of a read of 100 bp. If there is a random 0.1 % probability of a

sequencing error, it means that 10 of these reads have an error. If the algorithm only removed the 90 identical reads, there will be 10 reads left with sequencing errors. This is a big problem since the correct sequence is only represented once.

To address this issue, the duplicate read removal program accounts for sequencing errors when it identifies duplicate reads. Specifically, reads are considered duplicates if:

- they share a common sequence of at least 20 bases[1] in the beginning, or at any of four other regions distributed evenly across the read, and
- the rest of the read has an alignment score above 80 % of the optimal score, where the optimal score is what a read would get if it aligned perfectly to the consensus for a group of duplicates.

Please note that these thresholds for similarity are not enough for reads to be marked as duplicates - they just define how different reads are allowed to be and still be considered duplicates. Rather, the duplicates are identified as explained in section 1.1.

1.3 Paired data

For paired data, the assumption is made that if both parts of the pair share the same sequence, they are duplicates, and only one copy of the pair is left in the output. Figure 1.3 shows an example of a paired read duplicate.

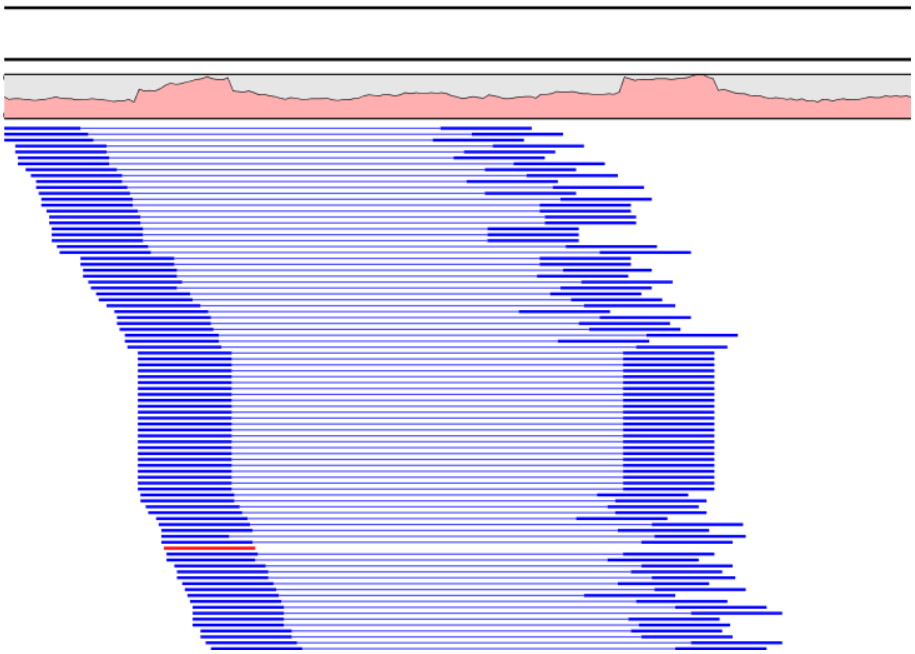


Figure 1.3: *Paired reads with identical starting positions.*

The algorithm also takes sequencing errors into account when filtering out paired data.

[1] For paired reads, this is only 10 bases.

1.4 Known limitations

In its current version, the duplicate read removal has a limitation when there are duplicate reads that contain several alleles. The algorithm will identify that there are duplicate reads to be removed, but it is not able to distinguish between sequencing errors and true variation in the reads. So if you have a heterozygous SNP in such an area, you may risk that only one of the alleles are preserved. We are working on improving the algorithm to handle this.

Chapter 2

Running the Duplicate Read Removal Plug-in Beta

To run the duplicate read removal:

Toolbox | **High-throughput Sequencing** | **Remove Duplicate Reads**

This will show a dialog letting you select the input data as shown in figure 2.1.

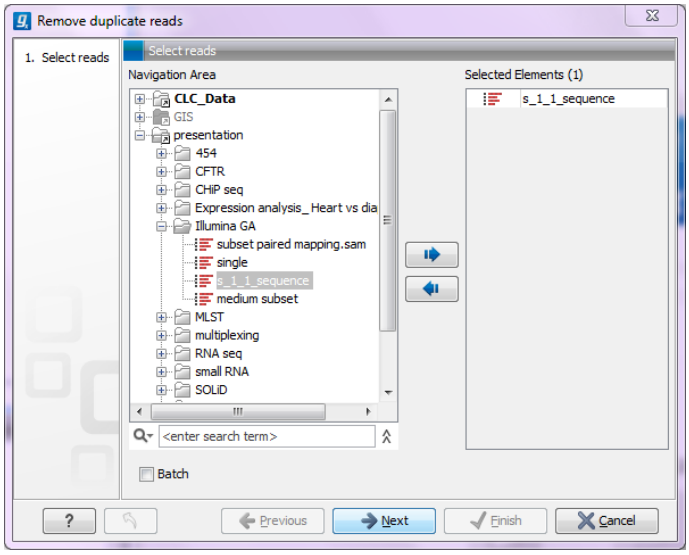


Figure 2.1: *Filtering known variations.*

As explained in section 1, the tool will take sequence lists as input, so it has to be run **prior to mapping or assembly** of the reads.

If you select multiple sequence lists, each list will be run separately through the algorithm.

Clicking **Next** will reveal the output options. The main output is a list of the reads that remain after the duplicates have been removed. In addition, you can get the following output:

List of duplicate sequences These are the sequences that have been removed.

Report This is a brief summary report with the number of reads that have been removed (see an

example in figure 2.2.

1 Summary

Name	Input sequences	Duplicates	Remaining sequences
s_1_1_sequence	5244764	55436	5189328

Figure 2.2: *Summary statistics on the duplicate reads.*

Chapter 3

Installation of the Duplicate Read Removal Plug-in Beta

The Duplicate Read Removal Plug-in Beta 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 3.1).

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 Duplicate Read Removal Plug-in Beta and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the Duplicate Read Removal Plug-in Beta 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.

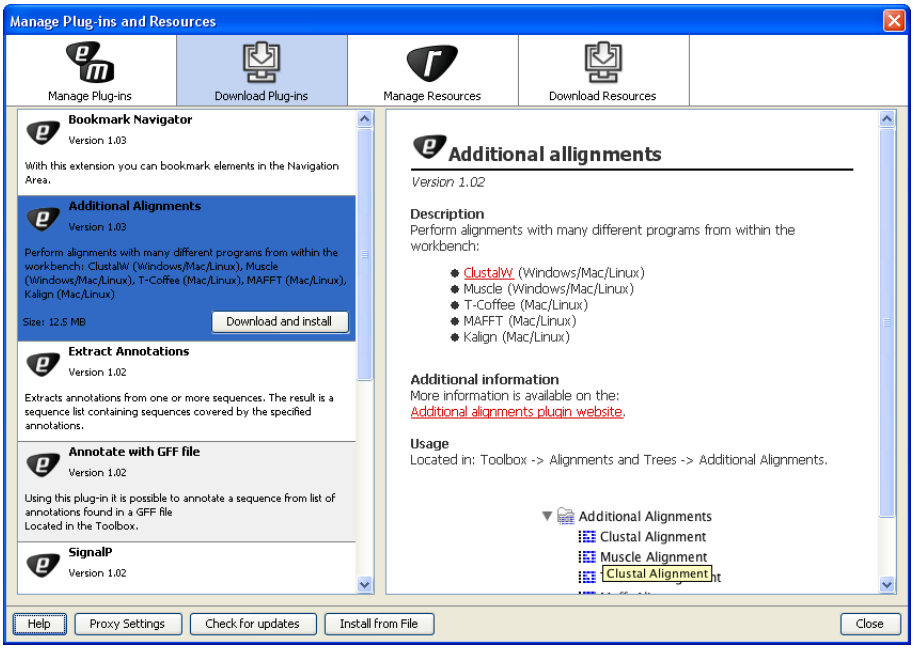


Figure 3.1: *The plug-ins that are available for download.*

Chapter 4

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 4.1.

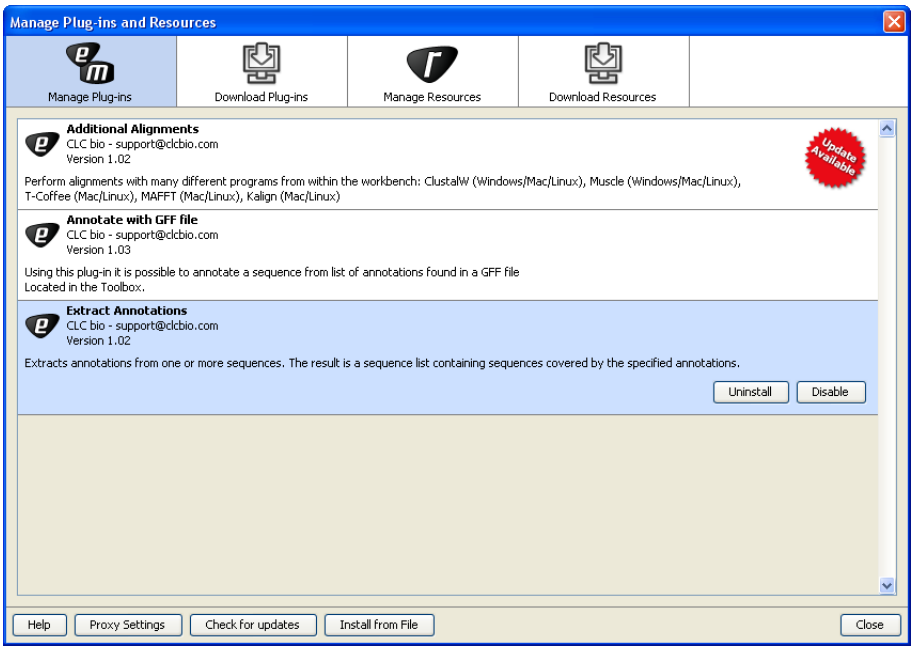


Figure 4.1: *The plug-in manager with plug-ins installed.*

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

Click the Duplicate Read Removal Plug-in Beta | 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.