

User manual for

Probabilistic Variant Detection plug-in 1.0 Beta

Windows, Mac OS X and Linux

February 24, 2012

This software is for research purposes only.

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

Contents

1 Introduction **4**

1.1 Statistical model 5

1.1.1 Genomic model 5

1.1.2 Error model 5

1.1.3 Combined model 5

1.2 Variant calling 5

1.2.1 Quality scores 6

2 Running the variant detection **7**

3 Results **9**

4 Installation of the *Probabilistic Variant Detection plug-in* **11**

5 Uninstall **13**

Chapter 1

Introduction

The *Probabilistic Variant Detection plug-in* is a new algorithm to find Single-Nucleotide Variations (SNVs), small InDels and Multi-Nucleotide Variations (MNPs). It will supplement the existing SNP and DIP detection tools of the *CLC Genomics Workbench*. Currently, the *Probabilistic Variant Detection plug-in* is in beta as a plug-in but it will eventually be part of the standard set of tools in the *CLC Genomics Workbench*.

The basic idea behind the new variant caller is to build a probabilistic model of the genomic sequence and of the sequencing errors. The parameters of the two models are estimated based on the data set being analyzed.

In the current version, the probabilistic variant detection is not designed to detect minor variants in a mixed sample (e.g. when sequencing a population and looking for rare alleles). If you have data like this, please use the existing SNP and DIP detection algorithms in *CLC Genomics Workbench*. When running the algorithm, you specify the maximum number of expected variations which should be set according to the ploidy of the genome that is sequenced.

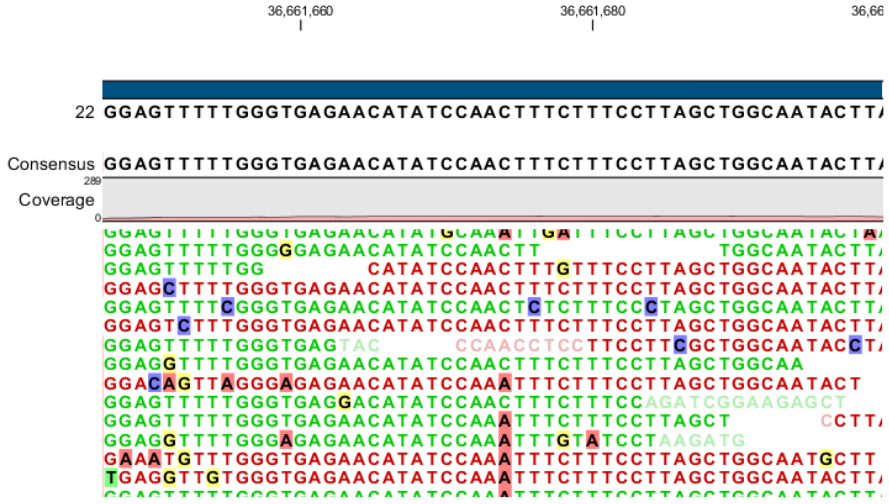


Figure 1.1: *An example of a heterozygous variation surrounded by a lot of noise from sequencing errors.*

1.1 Statistical model

1.1.1 Genomic model

The basic model of a diploid genome is that any site can be one of 15 different types -/-, -/A, -/C, -/G, -/T, A/A, A/C, ..., G/G, G/T, and T/T. The reason that it is 15 types and not 25 is that we cannot distinguish A/G and G/A. Each of these site types has a frequency. The same general approach is use for haploid, triploid, and quadroploid data.

However, the model is not as simple as this: In many cases, a low percentage of a variant is found in some sites. This can lead to false identification of heterozygote sites. This problem may arise for several reasons including weak systematic errors in sequencing, PCR errors and duplicate reads. To alleviate this problem, some extra site types are included. They consist of e.g. 95 % frequency of one nucleotide and 5 % frequency of another. If the main nucleotide is X and the other is Y, we can denote such a site as "X(Y)". Including gaps, this leads to another 20 new site types. Rather than report these sites as results, their probabilities are grouped with the sites that are homozygous for the main nucleotide. So the model for homozygous A, for example, is a combination of the types A/A, A(C), A(G), A(T), and A(-).

1.1.2 Error model

The basic sequencing error model consists of five types of sites. One for each nucleotide and one for the site between two nucleotides. This refers to what the true sequence is. For each site type, we then have an observed distribution of nucleotides in the read as determined by the sequencing. So if the true sequence is an A, we then have probabilities of observing nothing (a gap), A, C, G, T, and N. The same applies to the other site type including the site between two nucleotides where the typical observation will be nothing, but a nucleotide may be observed (leading to an insertion error in the read).

1.1.3 Combined model

We combine the two models to give a full probability model of everything we observe. So a given genomic site may be any one of the 15 different genomic types and for each of these, a given read nucleotide may be from either haploid copy of the site and lead to the observed nucleotide determined by the sequencing and the read mapping. By making this combined model, we are able to determine the posterior probability distribution of genomic site type for each genome position.

All the parameters of the model are estimated for the data set being analyzed. This ensures that suitable error rates etc. are used in each analysis.

1.2 Variant calling

The sites reported are the ones where there is more than 95 % chance that they are different from the type defined by the reference sequence (this threshold can be adjusted in the dialog). In other words, if it is unlikely to be homozygous for the reference nucleotide, then it is reported. This is the only place where the reference sequence comes into play. Please note that variants in homopolymeric regions (e.g. AAAAAA or AAACAAA) are not reported as it is likely that these are sequencing errors.

1.2.1 Quality scores

The sequencing machines not only produce reads consisting of nucleotides, but also quality scores. These quality scores are taken into account in the error model by creating a distribution of observed nucleotides for each quality level. By doing this, the error rates are found for each individual quality score leading to better results.

Chapter 2

Running the variant detection

Once the *Probabilistic Variant Detection plug-in* is installed, there are two ways of starting the analysis: If you are working with Genomics Gateway (see http://www.clcbio.com/genomicsgateway/), there is a special version in the **Genomics Gateway** folder in the **Toolbox**. Otherwise, the standard version is available in:

Toolbox | **High-Throughput Sequencing** ( ) | **Probabilistic Variant Detection**

For the standard version, select a read mapping or read mapping table as input. For the Genomics Gateway version, select a reads track as input. Clicking **Next** allows you set parameters as displayed in figure 2.1.

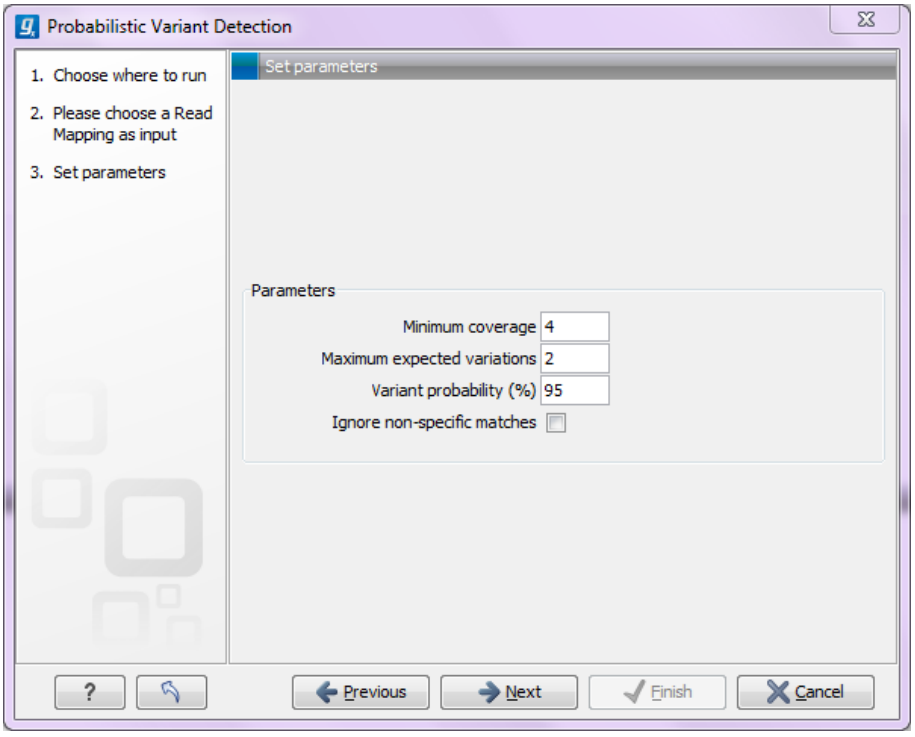


Figure 2.1: *Setting parameters for the variant detection.*

The follow parameters can be set:

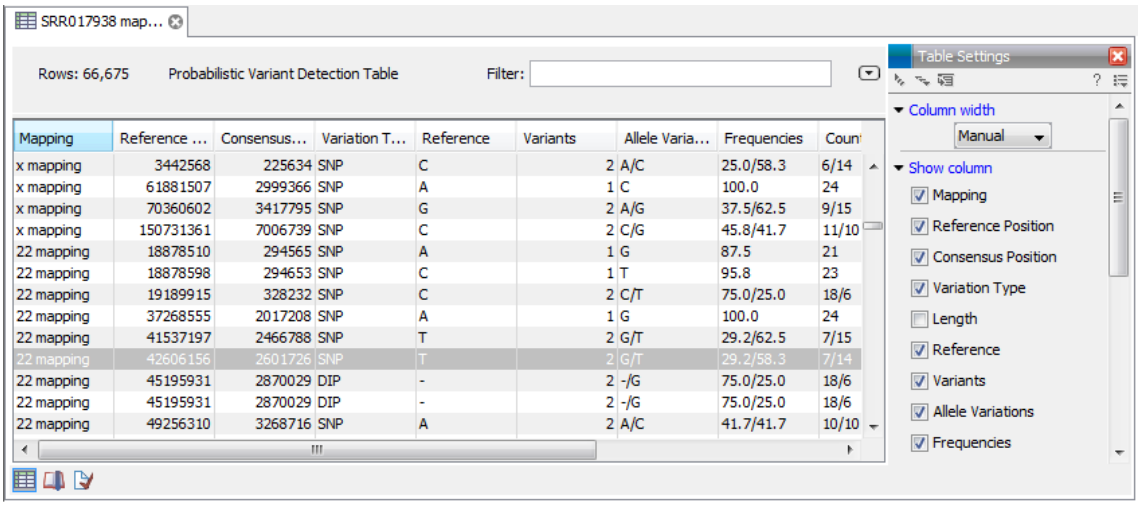
- **Minimum coverage** The minimum number of reads aligned to the site to be considered a potential variant.
- **Variant present in forward/reverse reads** Only variants will be reported, which are present in at least one forward and one reverse read.
- **Maximum expected variations** This is the number of alleles expected and the number of alleles that will be reported. If set to 1, only the majority allele is reported even if another allele is present as well.
- **Variant probability** This is the probability that the reported variant is different from the reference sequence.
- **Use only intact pairs** Only proper paired reads (displayed in blue in the workbench) will be taken in consideration.
- **Ignore non-specific matches**. Reads that map equally well on several places on the genome are marked as non-specific matches. They typically come from repeated regions. Since it is uncertain where they actually belong, you can decide to ignore these when calling variants.

Click **Next** to adjust output options. These are similar to the ones known from the existing SNP detection.

Chapter 3

Results

An example of a result is shown in a table in figure 3.1.



Mapping	Reference ...	Consensus...	Variation T...	Reference	Variants	Allele Varia...	Frequencies	Coun
x mapping	3442568	225634	SNP	C	2	A/C	25.0/58.3	6/14
x mapping	61881507	2999366	SNP	A	1	C	100.0	24
x mapping	70360602	3417795	SNP	G	2	A/G	37.5/62.5	9/15
x mapping	150731361	7006739	SNP	C	2	C/G	45.8/41.7	11/10
22 mapping	18878510	294565	SNP	A	1	G	87.5	21
22 mapping	18878598	294653	SNP	C	1	T	95.8	23
22 mapping	19189915	328232	SNP	C	2	C/T	75.0/25.0	18/6
22 mapping	37268555	2017208	SNP	A	1	G	100.0	24
22 mapping	41537197	2466788	SNP	T	2	G/T	29.2/62.5	7/15
22 mapping	42606156	2601726	SNP	T	2	G/T	29.2/58.3	7/14
22 mapping	45195931	2870029	DIP	-	2	-/G	75.0/25.0	18/6
22 mapping	45195931	2870029	DIP	-	2	-/G	75.0/25.0	18/6
22 mapping	49256310	3268716	SNP	A	2	A/C	41.7/41.7	10/10

Figure 3.1: *Results of the variant detection.*

Each variant will be reported including the following information:

- **Reference position**. The first position of the variant on the reference sequence.
- **Consensus position**. The first position of the variant on the consensus sequence.
- **Variation type**. Will be "DIP", "SNP" or "InDel/MNP". Note that when two SNP or DIP variations are found within the three bases it will also be reported as an "InDel/MNP" (for the Genomics Gateway version, this will be a separate track).
- **Reference**. The residues found in the reference sequence.
- **Variants**. The number of variants among the reads.
- **Allele variations**. The variations found in the reads.
- **Frequencies**. The frequencies of the variations, both absolute (counts) and relative (percentage of coverage).

- **Coverage**. The number of reads covering the variation site.
- **Overlapping annotations**. Says if the variation is covered, in part or in whole, by an annotation (this is not part of the Genomics Gateway version where the overlap annotation refiner can be used for this purpose)

Chapter 4

Installation of the *Probabilistic Variant Detection plug-in*

The *Probabilistic Variant Detection 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 *Probabilistic Variant Detection plug-in* and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the *Probabilistic Variant Detection 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 *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.

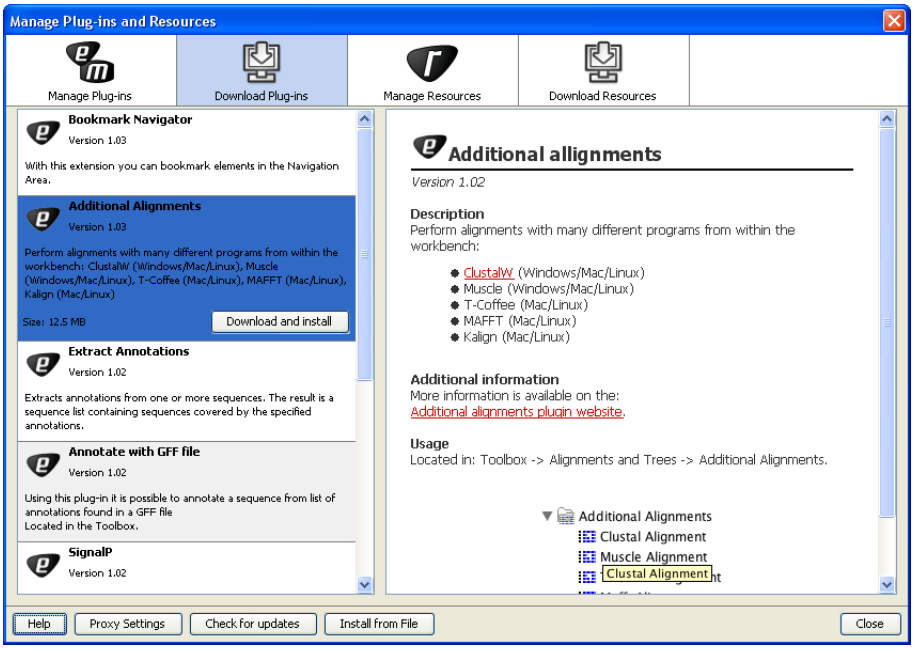


Figure 4.1: *The plug-ins that are available for download.*

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.

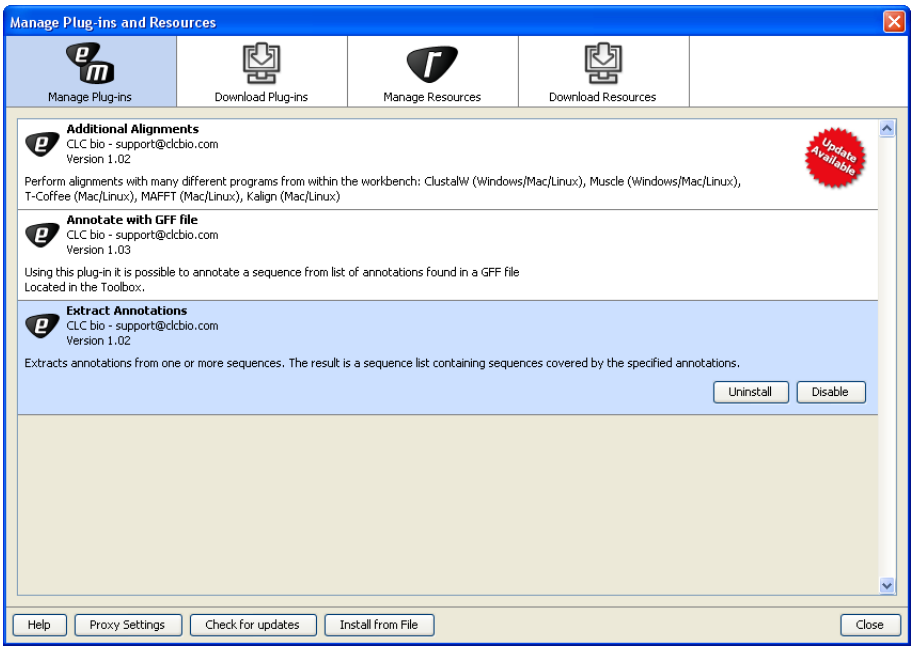


Figure 5.1: *The plug-in manager with plug-ins installed.*

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

Click the *Probabilistic Variant Detection 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.