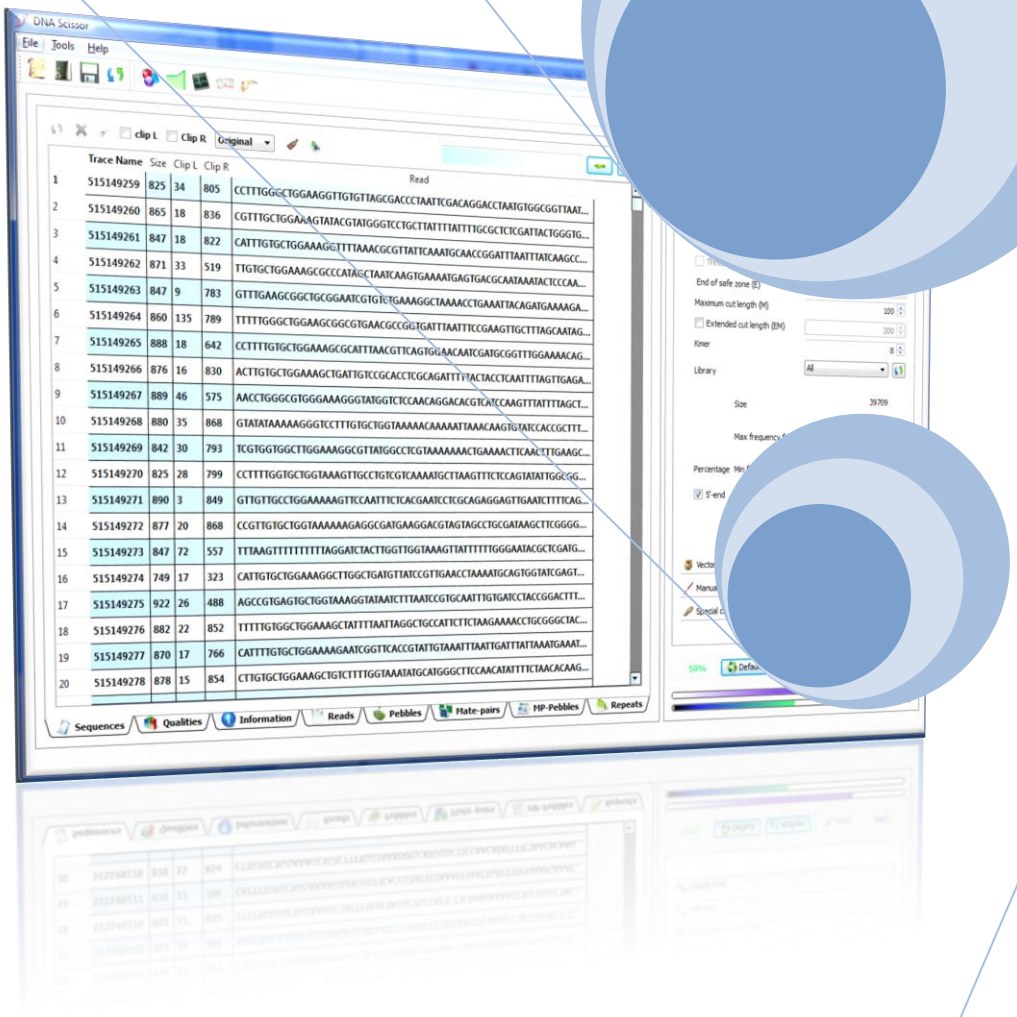


DNA Scissor Manual

As a part of Arapan Project

DNA Scissor is a multi-functional software which provides users with a set of easy-to-use graphical tools to perform several operations mainly: quality trimming, vector masking, vector-like contaminants removal, and repeats identification.

By

Xxxxx Yyyyyy, Xxxxx Yyyyyy

5/16/2011

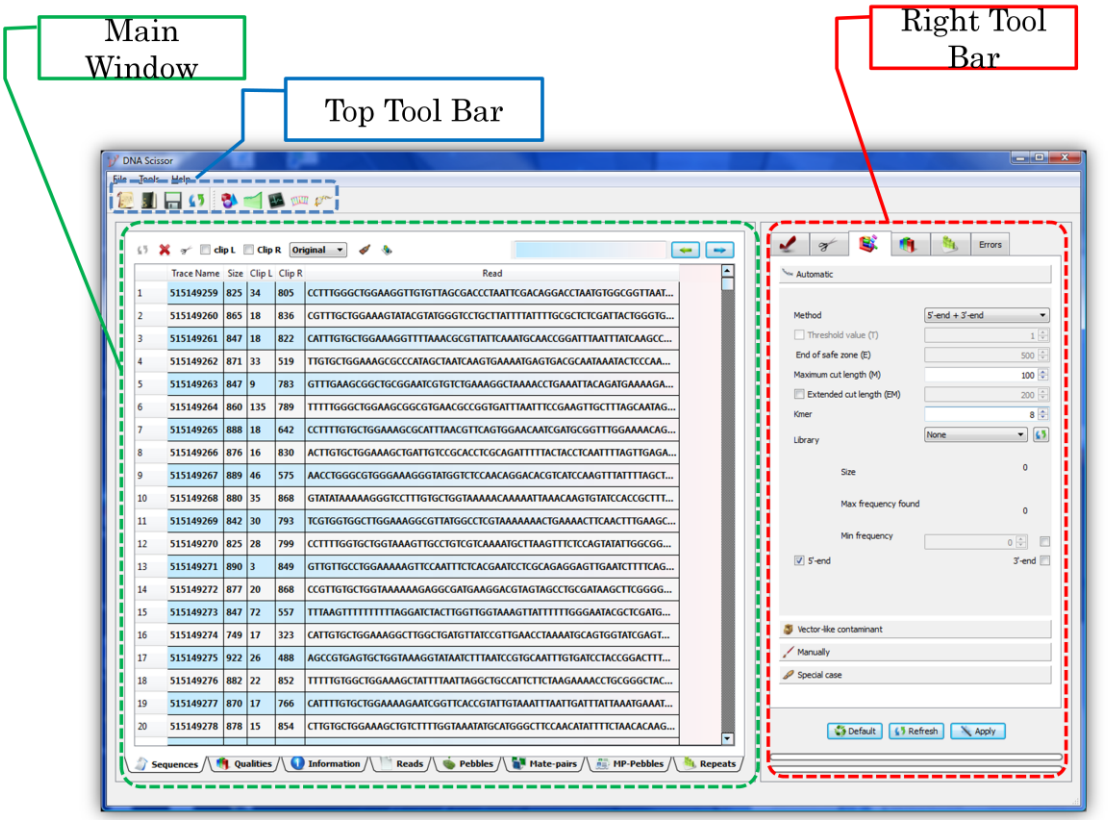
Manual

Contents

I. Main interface

1 Introduction

DNA Scissor is a multi-functional software which provides users with a set of easy-to-use graphical tools to perform several operations mainly: quality trimming, vector masking, vector-like contaminants removal, and repeats identification.

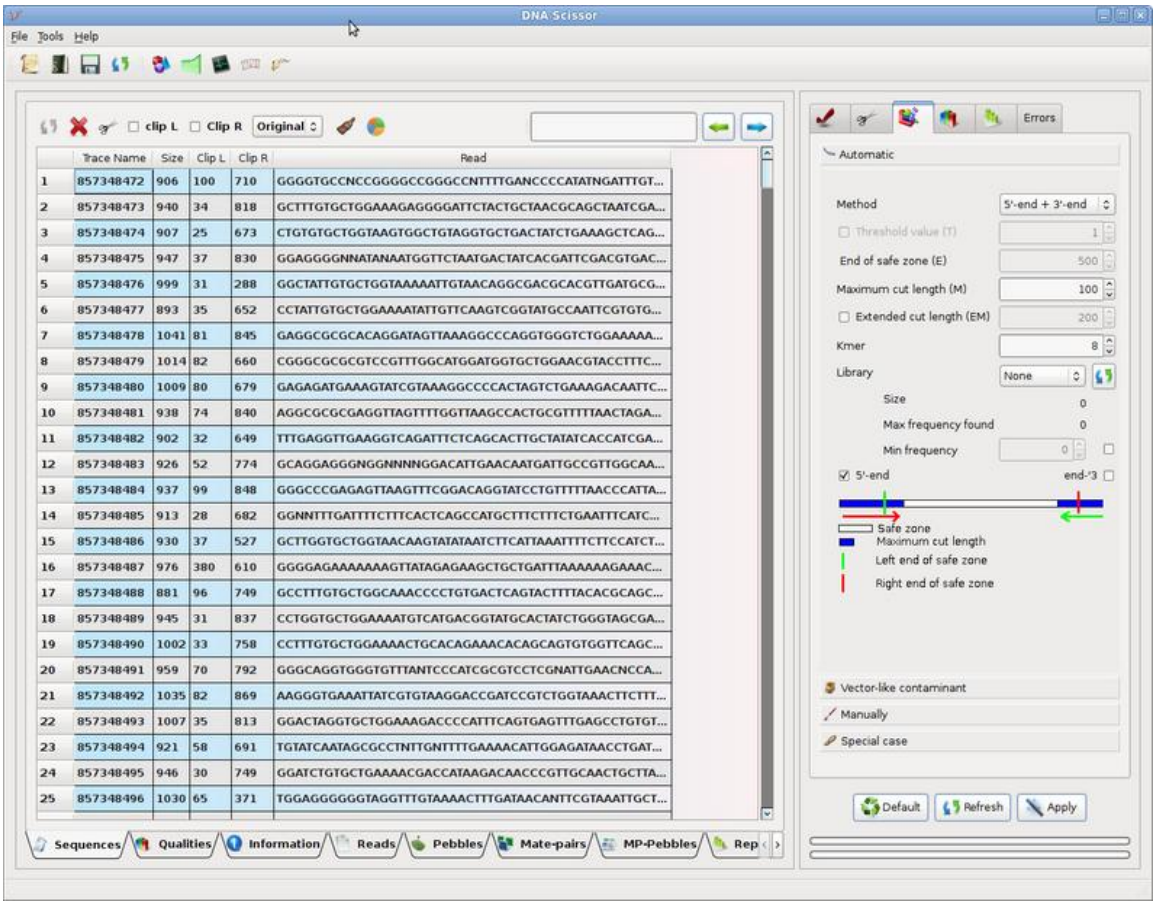


18

Figure 1: The main interface of DNA Scissor.

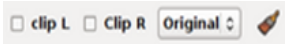
It is composed of two essential parts: the displaying interface and the panel control. We will describe each tab of the displaying interface independently.

2 Sequences



The loaded sequences can be shown in this tab along with the following essential information :

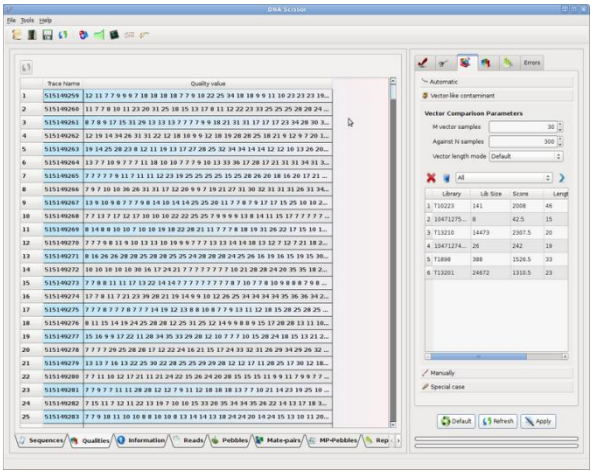
Trace Name	the sequence identifier.
Size	the length of the sequence
ClipL	the Left Clipping point of the sequence.
ClipR	the Right Clipping point of the sequence.
Read	DNA sequence.

Clipping points can be gotten from the original data and updated by DNA Scissor clipping/trimming operations. In case the user wants to retrieve the original clipping points or reset them, the tool provides some useful operations 

so that the user can work conveniently.

3 Qualities

The quality values are shown in this tab along with its corresponding Trace Name.

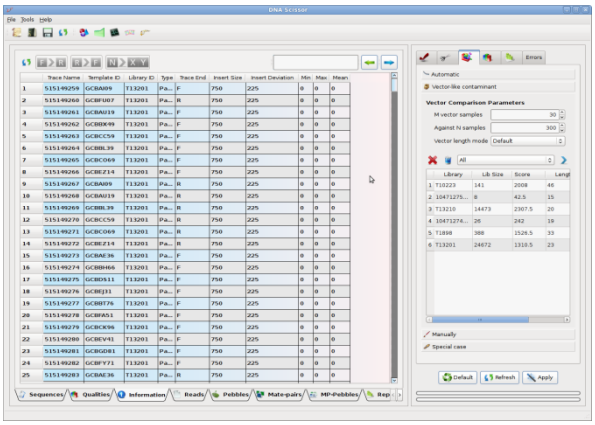


Remark: It is not suitable to display reads and quality values in case of dealing with large amount of data because of memory usage. DNA Scissor does not display them by default unless the user select this option in the preferences menu.

4 Information

The information concerning each sequence is shown here. It comprises from:

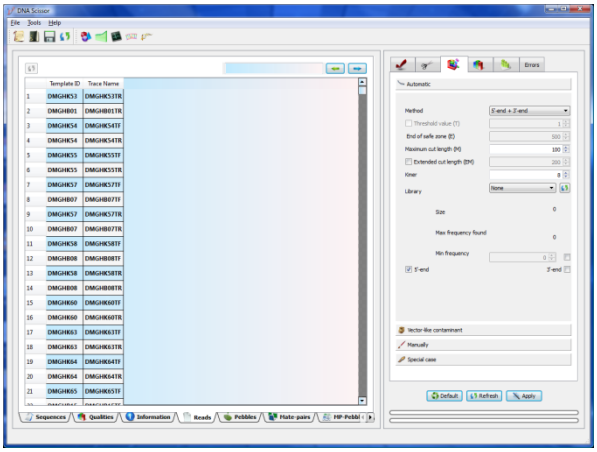
Trace Name, Temple ID, Library ID, Trace End, Insert Size, Insert Deviation, Insert Min, Insert Max and Insert Mean.



5 Reads

We define clean reads those which are not mate-pairs or pebbles.

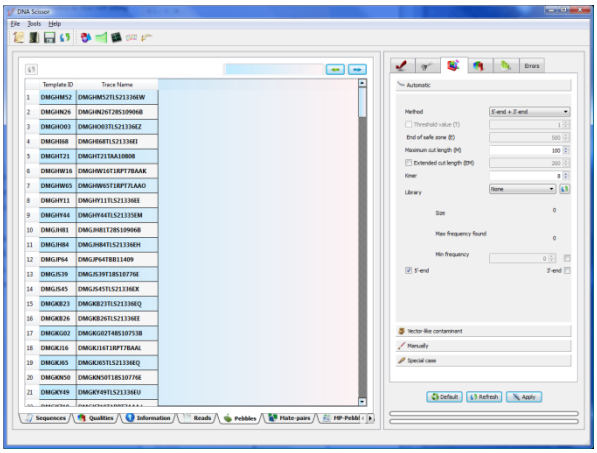
Each line contains a clean read ID along with its template ID.



6 Pebbles

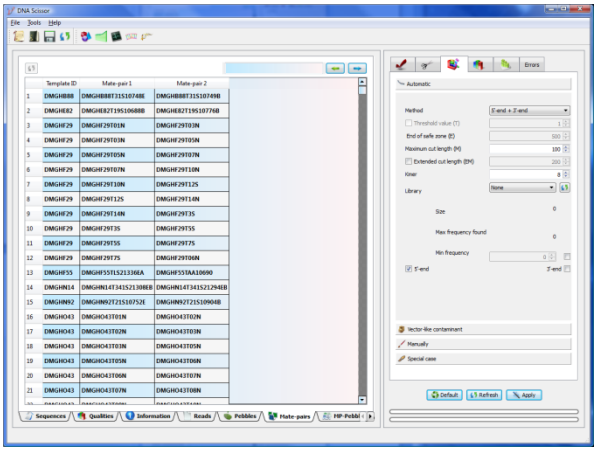
Pebble is a discarded read through the different cleaning processes. As a matter of fact, the discarded reads will be moved to the pebbles tab by the will of the user. Each line contains a read pebble ID along with its template ID.

Just to note that pebbles can be very useful in the finishing part of genome assembly process in order to get much more longer scaffolds.



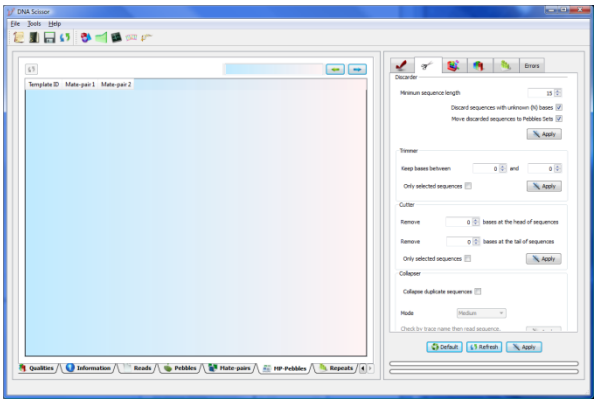
7 Mate-pairs

Sometimes, the data also include mate-pairs file. In such a case, mate-pairs will be downloaded to this tab in which each line contains two mate-pairs' ID along with its template ID.



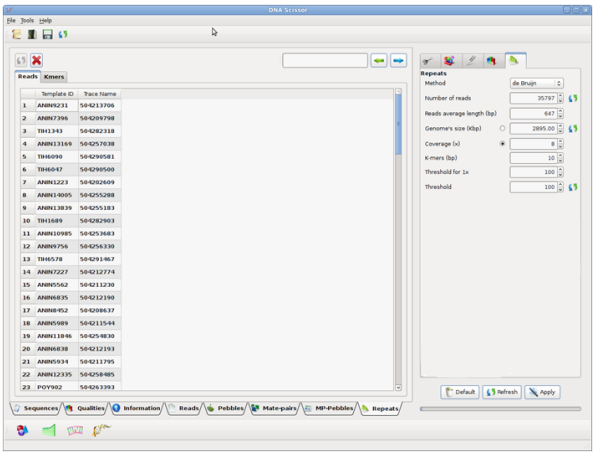
8 MP-Pebbles

The discarded mate-pairs through any cleaning process can be moved to this tab for further analyses. Each line contains two mate-pairs pebbles' ID with the template ID for each.



9 Repeats

DNA Scissor detects repetitive sequences by calculating the k-mer distribution for all reads. If some k-mers appears more frequently and exceeds a predetermined threshold, it may be originated from a repetitive sequence. The most frequent k-mers and the detected repeats can be displayed in this tab such that each line contains a repeat ID.



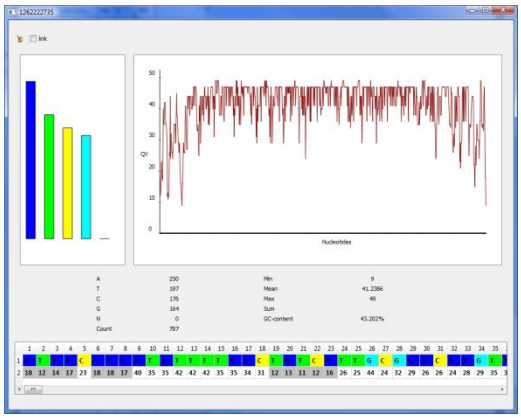
10 Refresh buttons 

Since we deal with huge data, the displaying process might slow the different operations if it was done automatically. For this reason, DNA Scissor does not automatically display the result in general, but it sends signals to refresh buttons so that the user can click on them in case s/he wants to see the changes.

II. Statistics

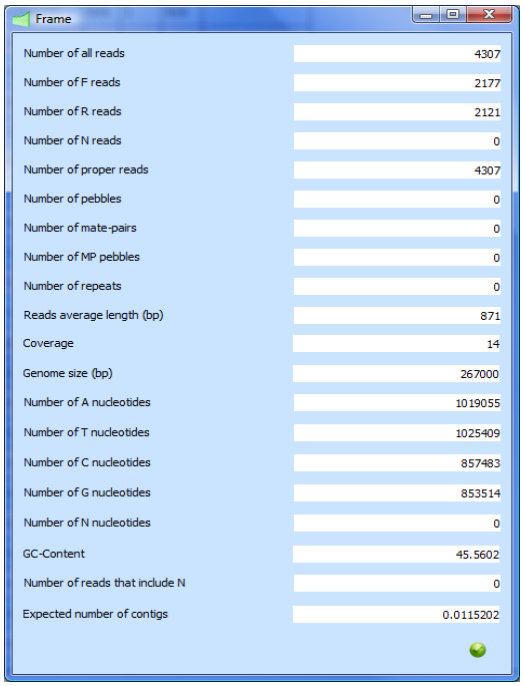
1 Sequence statistics

To show the statistics concerning a specific sequence, you can select the target sequence in the sequences tab and click the following button:



2 Whole statistics

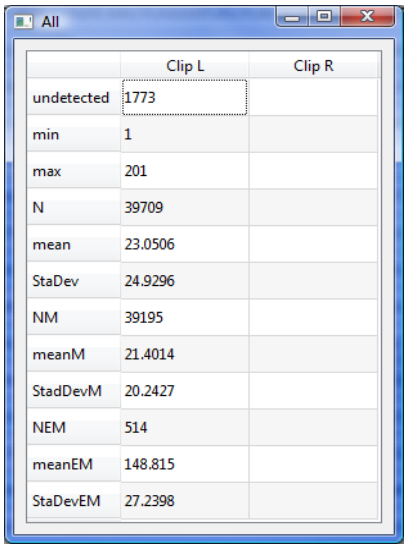
This window gives an overall statistics concerning the data being processed by clicking on:



3 Clipping points statistics

DNA Scissor can also give statistics on the left clipping points for the current version.

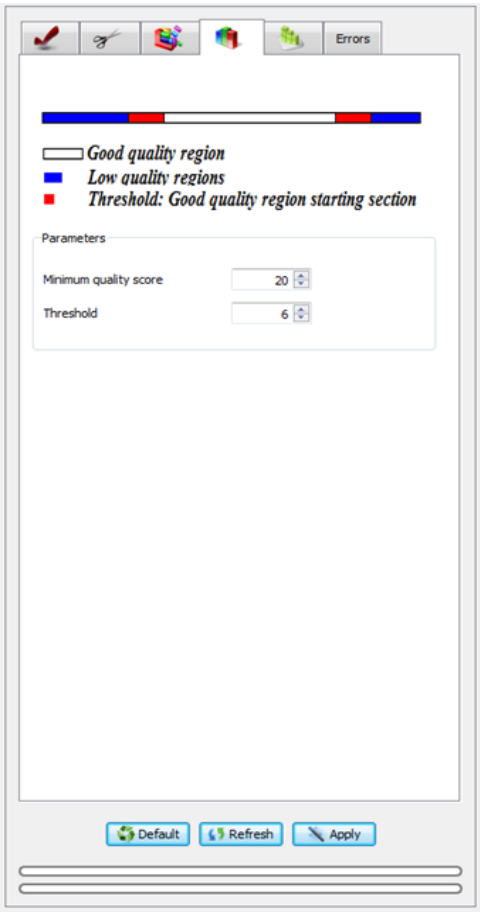
Concerning the meaning of different rows names please consult the part concerning vector trimming.



III. Quality values trimming

DNA Scissor can detect the longest low quality regions from the 5'-end and 3'-end of reads depending only on two parameters:

- Q the minimum quality score (20 by default),
- T the threshold (6 by default).



If DNA Scissor detects T consecutive nucleotides whose quality values are greater than Q, it stops exploring and marks the trimming points since it considers the T nucleotides as the starting section of the good quality region.

The case when there are some low quality regions within the good quality region, DNA Scissor does not detect them. However, it can deal with reads which contain some unknown 'N' nucleotides such that the user can then discard or mark them as 'pebbles' for further utilization during the genome assembly process.

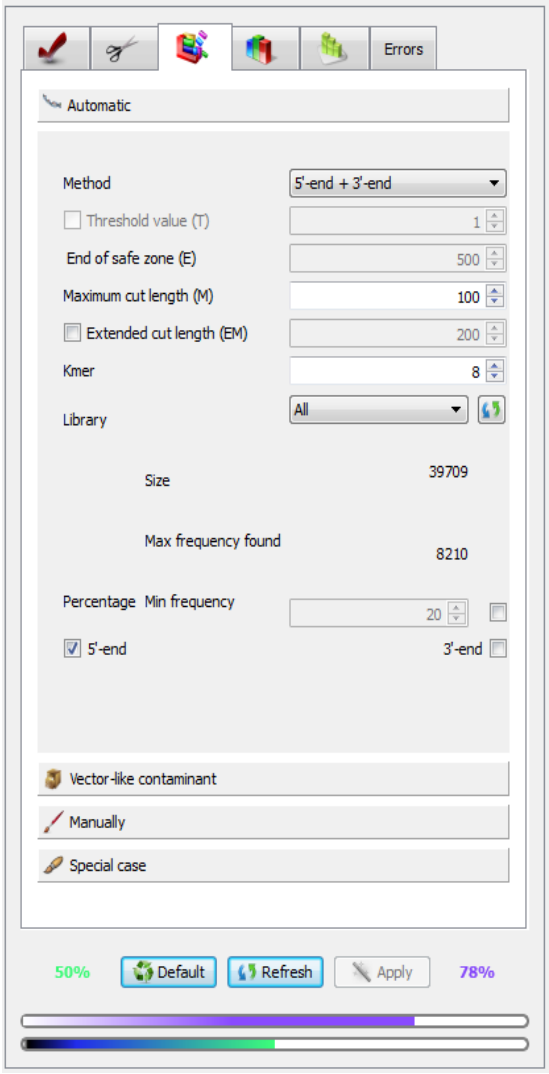
Remark 1 : Note that clipping points get the

maximum of the index position of new calculated points and the provided clip information (by NCBI Trace Archive). In case the user does not want to incorporate the provided clipping information, our tool can simply reset them before the process starting.

Remark 2 : It is preferably to run quality trimming after vector sequences trimming.

IV. Vector sequences clipping

1 Automatic



The program is able to detect the endmers (the k-mer that represents the end of vector) of vector sequences at the 5'-end of reads. In case short libraries are used, it can be forced to detect vector sequences at the 3'-end as well. For getting more accurate clipping points, DNA Scissor can detect libraries automatically. It is preferable to let the software detect the clipping points for each library in order to get more accurate results.

DNA Scissor can detect vector sequences endmers without prior knowledge of cloning vectors.

The algorithm is uses two main parameters:

- M : The maximum cut length (100 by default)
- EM : The extended maximum cut length (2*M by default).
- Kmer: the length of the short read (seed).
-

However, only M the maximum cut length parameter which should be defined by the user. DNA Scissor adjust other parameters automatically.

To give more flexibility to the algorithm implemented in DNA Scissor, the Kmer length is automatically changed according to the maximum cut length parameter. The minimum k-mer length is 4 (resp. 8) when the maximum cut length is very short (resp. very long). Table 1 shows different provided lengths.

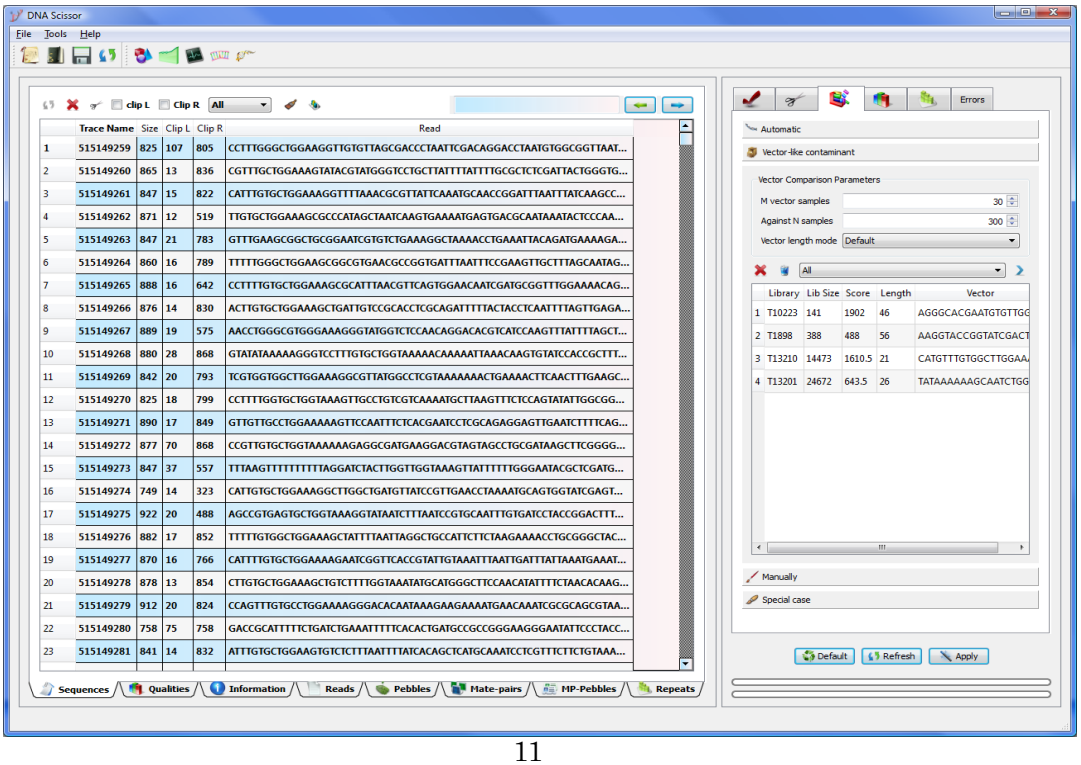
Range	*k*-mer length
$90 \leq M$	8
$50 \leq M < 90$	7
$30 \leq M < 50$	6
$20 \leq M < 30$	5
$M < 20$	4

Table . Maximum cut length ranges with the corresponding *k*-mer lengths.

In fact, some shotgun raw data (from NCBI Trace Archive) have very short vector sequence or adapter (e.g. of length 5). Such situations lured us into the idea of changing the *k*-mer length according to the maximum cut length. Nonetheless, the user can change the *k*-mer length whenever it is possible.

2 Vector-like contaminant

Since vector sequences do not only appear at the 5'-end of reads, but also may occur within the reads. The good point of DNA Scissor is the ability to detect a very reliable vector sequence automatically for each library without prior knowledge of the vector. As a result, despite the luck of vectors, user can get a very reliable vector sequence for each library and deal with contaminated reads. Note that when a vector sequence appears



beyond the 5'-end or represents the whole read, it is considered as 'contaminant' or in a specific term 'vector-like contaminant'.

DNA Scissor requires two parameters in this case:

- *p = number of random competitive vector sequences (30 by default)*
- *q = another number of random vector sequences (300 by default)*

After that, just click on  to start the process.

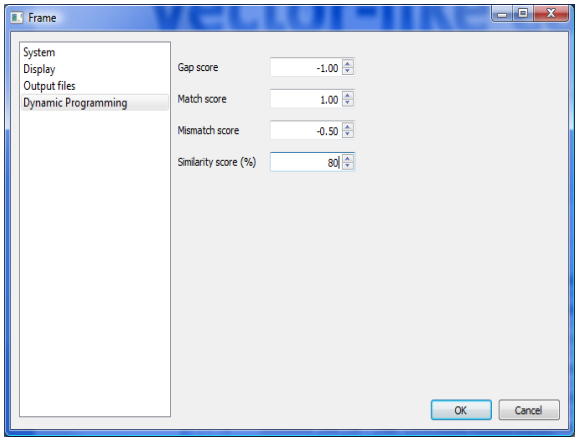
Remark: This task should be done immediately after detecting the vector sequences clipping points. Please do not trim data or run quality trimming process.

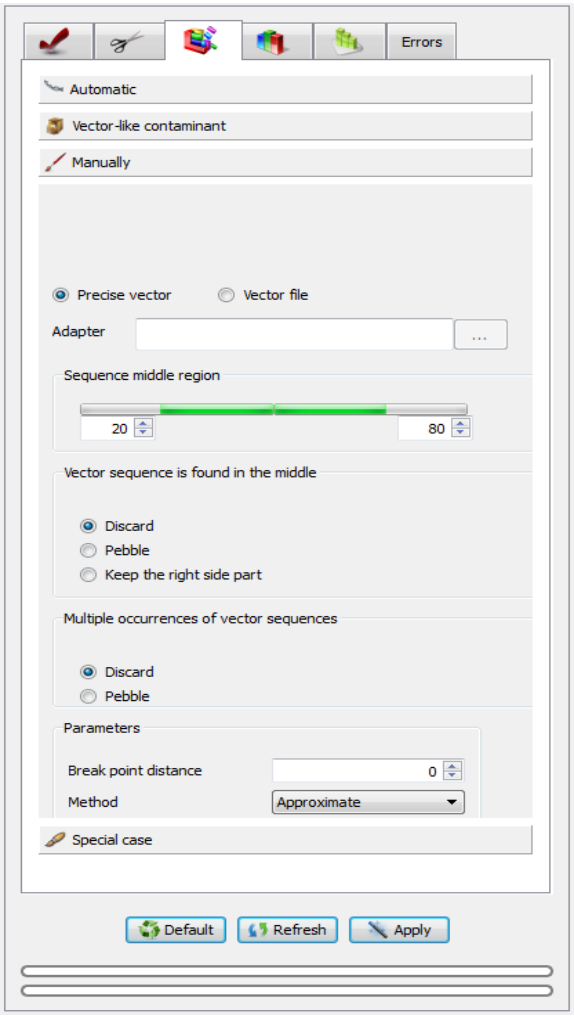
The obtained vector sequences can be treated as explained in the following.

3 Manually

Whether the user generated reliable vector sequences from the previous section or prefers to use his/her own vectors. In this case, we provide a set of tools to remove all vector sequences manually. Users have a full control of vector management. The affected DNA reads by vectors in the center may be discarded, cut at specific location or marked as pebbles.

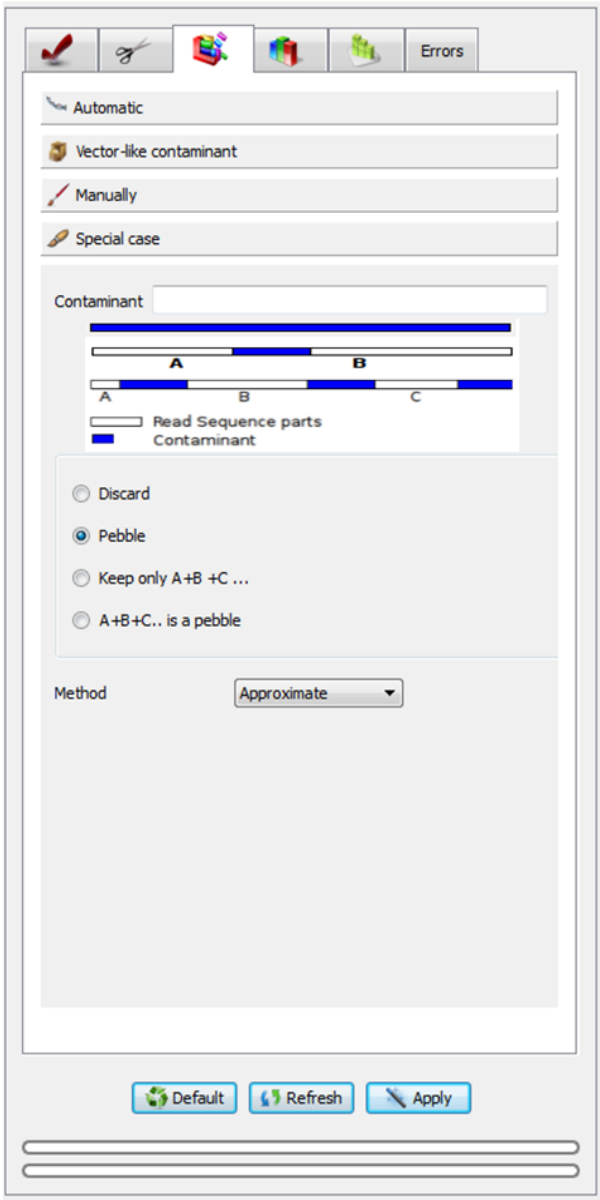
User should define his/her desired parameters of the dynamic programming by clicking on the submenu called "Preferences".





4 Special case

The vector sequences do not only appear in the 5'-end of reads, but also may represent the whole read or appear elsewhere in the read. DNA Scissor can also deal such situations. Before running this section the dynamic programming parameters should be defined.



V. Other utilities

Discarder

Discard sequences of minimum length and those which include unknown 'N' nucleotides.

Trimmer

The user can keep a specific part of sequences.

Cutter

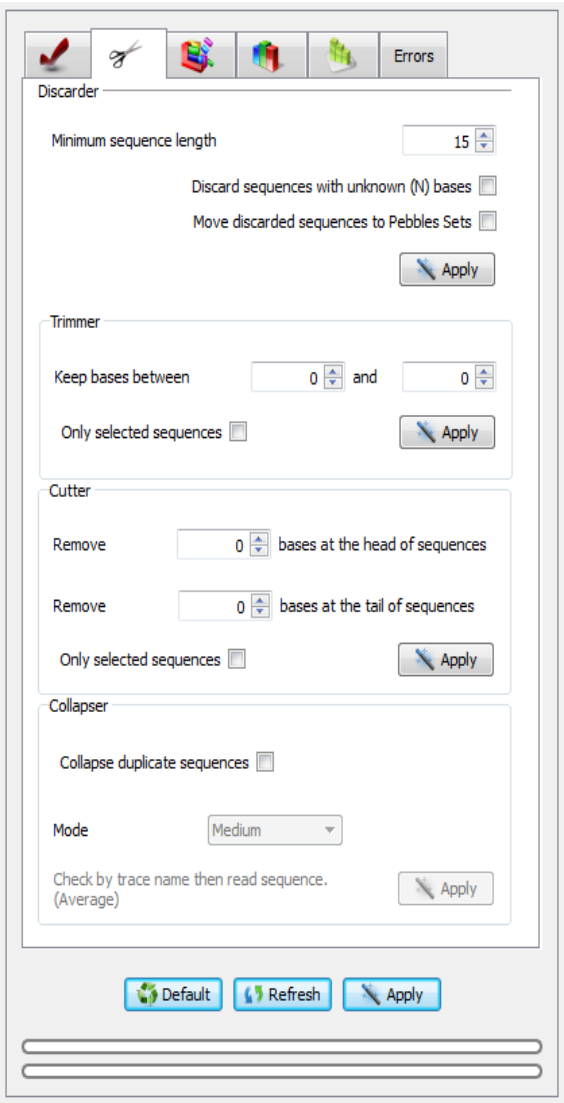
The user can remove a number of nucleotides at the head and the tail of sequences.

Collapser

DNA Scissor is able to delete the duplicate sequences such as the case of PCR duplicates which often appear in the NGS (Next-Generation Sequencer) data of PCR-amplified DNA sequences. The tool provides three modes to delete duplicates (Light, Medium, Heavy). The Light mode is the fastest since it only allows DNA Scissor1 to delete sequence whose have the same sequence ID. The Medium mode lets the software to check sequences IDs then the similarity of sequences. The slowest mode is the Heavy mode since it neglects checking sequences IDs and focus only on the similarity of sequences without taking into account their IDs, with which we can remove PCR duplicates.

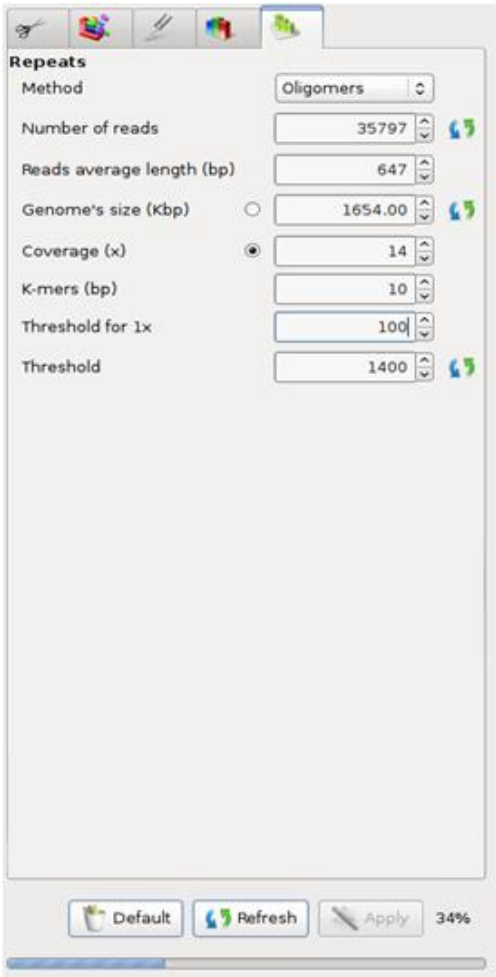
Pebbles

The discarded sequences can be marked as pebbles in case the user wants to use them in the finishing part of the genome assembly project.



VI. Repeats detection

This part is heavily parameterizable. The user should know at least the value of one parameter: the genome size or the coverage value. If one of them is known, the other can be calculated automatically. We also facilitated calculating other parameters depends on the genome size or the coverage value. We provide two methods: 'Oligomers' and 'de Bruijn k-mers'. Oligomers method is faster but it is not as efficient as de Bruijn k-mer based method. The algorithm detects repetitive sequences by calculating the k-mer distribution for all reads. If some k-mers appears more frequently and exceeds a predetermined threshold, it may be originated from a repetitive sequence.

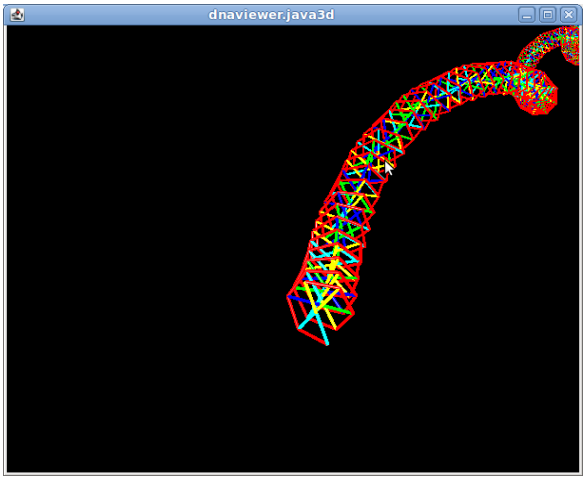


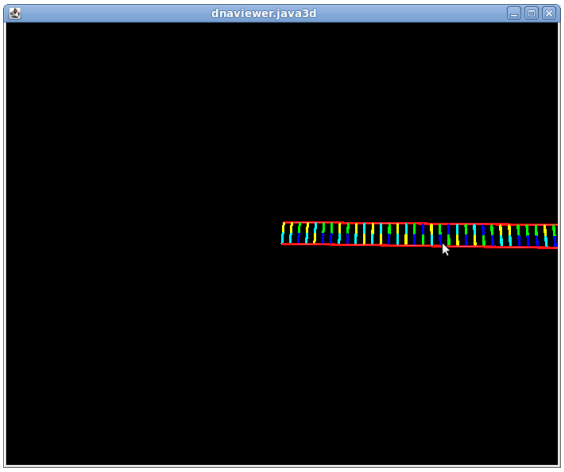
The critical parameters of this part is the coverage and threshold for 1x. They should be tuned carefully. However, in case the repeats were mistakenly detected. The user can delete them and tunes the parameters again.

For updating the other parameters please click on the buttons  before you start applying the process.

VII. DNA Viewer

The selected sequence can be seen in 3D mode whether in helix format or straight view as shown below. Java virtual machine must be installed in order to launch this feature.





The viewer is controlled by mouse buttons (the left and the right buttons).

VIII. Contact

For further questions, please contact us at dnascissor@hgc.jp.