StreamingTrim 1.0 Manual

"A powerful trimming supply for your DNA reads"

Giovanni Bacci

[Pick the date]

INDEX

1 INTRODUCTION

During the last few years, DNA and RNA sequencing have started to play an increasingly important role in biological and medical applications. This is mainly due to the growth of sequencing data produced from the new sequencing machines and to the enormous decrease in sequencing costs. One of the most important problems related to the production and to the utilization of DNA sequence reads is to analyze the quality of every reads present in a sequence file and to be able to trim the low quality segment without lose too much information.

Here, a reads trimming software is described, capable of analyzing the quality of every single base present in a reads file and to search for low quality zone in a very conservative way, in order to preserve most of the information.

The aim of this software is to provide a "simple to use" tool able to analyze and trim reads file independent from file size. This software reads and analyze sequences one by one from the input file without keeping anything in memory. Every trimmed sequence is saved in an output file provided by the user.

Next, the software print a simple statistics summary containing the mean and the standard deviation of length and quality of the whole sequence file. The tool is also capable of printing two different type of plots in order to give the user a more accurate description of the length and quality distribution.

2 GETTING STARTED

StremingTrim is a standalone Java software, built with Java 1.7. In order to install and use it, it is necessary to have Java installed on your system. In this chapter is described how to check your java installation and how to download and install the StreaminTrim software.

2.1 OBTAINING AND INSTALLING THE PROGRAM

As already said StremingTrim is a software build with Java 1.7, so you have to ensure that you have at least Java 1.7.0 version installed on your system. In order to do this you have to open you command windows (cmd.exe in Windows systems and terminal in OS systems) and type this:

```
# java -version
```

If you receive an error it means that you don't have java installed on your systems. Instead if you receive a message like this:

```
# java version "1.7.0_09"
# OpenJDK Runtime Environment (IcedTea7 2.3.4)
# OpenJDK 64-Bit Server VM (build 23.2-b09, mixed mode)
```

If the number between brackets is smaller than 1.7.0 it means that you have Java installed on your system but you have an old version of the software. In both cases you have to install an up to date *Java Runtime Environment*, you can download it from the oracle web site: http://www.java.com/en/download/ (if you have an old version of Java is recommended that you uninstall it before install the new version).

If your Java version is up to date you can proceed to download the software from the GitHub repository at https://github.com/GiBacci/StreamingTrim and save it in a folder of your choice.

2.2 FIRST RUN

Once you have downloaded the tool you can launch it double clicking one of the two launchers present in the software's folder. If you have a Microsoft Windows based system you have to use the **windowsLauncher.bat** file, while, if you have a Linux based system or a Mac OS based system, you can launch it with the **unixLauncher.sh** file (remember to allow executing file as an application). If everything has gone well you would be able to see the main window of StreamingTrim software. Otherwise, please report your problem in the StreamingTrim GitHub repository at https://github.com/GiBacci/StreamingTrim/issues.

2.3 LICENSING

This software is distributed under the BSD-2-Clause license, here is the content of the LICENSE.txt file:

3 STREAMINGTRIM OVERVIEW

StreamingTrim is a very conservative trimmer designed for fastq reads file. The software is able to decode ASCII quality from the most common encodings types (Illumina, Solexa, Sanger and Solid). It's not necessary to specify your file type since the program can detect it automatically. However, if you are sure about your file encoding type, you can force the program to use it by selecting one of the quality encodes present in the <Reads Type> menu.

As you may have noticed in this manual we use some Type-setting conventions. We use:

```
# this format
```

in order to refer to command line input or output, but also to refer to external text (for example a DNA sequence contained in a sequence file); when we want to indicate a program menu or function we use <this format>. If you see something like: <File → Open File> it means that we refer to the Open File item in the File menu.

3.1 TRIMMING ALGORITHM

StreamingTrim uses a "dynamic window" algorithm to cut away low quality segments of DNA sequences beginning from the end of the read. This approach is very useful because it allows users to set a more stringent quality cutoff without risking to loose too much information. Here, we describe the program algorithm in order to give the user all the information needed to run the trimmer properly.

First of all, if the user has not specify a quality cutoff (see section 5.1 below) the program performs a rapid statistical analysis and set the cutoff automatically, as the mean quality of all the reads in the sequence file minus one standard deviation (e.g.: if we have a file with a mean quality of 31.46 and a standard deviation of 6.54 the cutoff quality is set to 31.46 – 6.54 = 24.92 and approximated up to 25).

Then, given a DNA sequence of length N, the algorithm starts from the sequence last nucleotide (the N nucleotide), using a window length (W) of 1 and checks if:

$$(Quality_N - Cutoff) > 0$$

If this is true, then the algorithm proceeds enlarging the window length by 1 (in this case putting $W = 2$), otherwise the N nucleotide is cut away and N is decreased by the number of nucleotide cut (in this case 1) and W is set to 1.

This process is repeated until the algorithm reaches the first nucleotide of the DNA sequence ($N = 1$), or if the length of the trimmed sequence goes below a minimum length value chosen by the user (see section 5.1 below). Here a formal description of the algorithm is shown:

$$\boldsymbol{N} = sequence\ length \mid \boldsymbol{W} = window\ length \mid \boldsymbol{M} = (N - W)$$

$$\boldsymbol{T} = \sum_{M<k\leq N} (Nucl_k - Cutoff)$$

If **T = true** $\rightarrow W = (W + 1)$ | If **T = false** $\rightarrow N = (N - W); W = 1$

Continue with the test T until $(N - W) \leq 0$ or $N < minimum\ length$ (for an explanation about the *minimum length* parameter see section 5.1 below)

The above-reported algorithm had been developed in order to be as much conservative as possible. A DNA sequence is eliminated only if all of its nucleotides are low quality nucleotides. If, in a sequence there is only a few "bad bases", the sequence is maintained in order to prevent the loss of information.

The high parsimony of this software is very useful especially in "amplicon based" analysis, like those commonly employed in bacterial or fungal community analyses (16S or 18S rRNA analyses), in which even the loss of one nucleotide can be crucial in order to retain the taxonomic information from the sequence.

3.2 PROGRAM WINDOWS

Here, all the windows of the StreamingTrim tool are described, in order to make an overview on the program function.

1. <Main Windows> is the principal window of the program. It is divided into four subsections and one progress bar in the bottom. Each section is described below:
    a. <Input File> this is an empty record, when the user opens a sequence file, the path to that file will be shown in this section.
    b. <Reads Properties> contains all the statistical parameters calculated from the programs related to the sequence file given as input.
    c. <Reads Type> here the user can select a particular type of quality encoding. If the user don't know the exact encoding of his reads he can select the <guess> option and the software will automatically detect the right encoding.
    d. <Controls> is the real controller of the software. Here all the buttons responsible for launching all software's command are present (for a list of all the function of StreamingTrim see the section 7).
    e. <Progress Bar> this bar, situated at the bottom of the main window, is useful in order to see the proper functioning of the program.
2. <Advanced Option> accessed thought <Window → Show advanced option>. In this window the user can set all the advanced option parameters. Next to each parameter there is a checkbox that allow the user to activate or deactivate the use of the specified parameter. If one of the checkboxes is selected, the user is able to write a specified value in the textbox on the right part of the windows.
   Important: if a value is wrote in a textbox, but the corresponding checkbox is not selected, the software will not use that parameter.
3. <Plot Window> accessed thought <Window → Show plot window>. In this window the quality and length distribution plots are displayed. All the interactive items in this window are disabled until the user opens a reads file. This window is composed of three different sections described below:
    a. <Plot Options> in this section the user can switch between quality and length distribution plots.
    b. <Raw data> in this panel the plots relating to the original reads are displayed.
    c. <Trimmed data> in this panel the plot relating to the trimmed reads is displayed.

4 OPEN AND ANALYZE SEQUENCES FILES

StreamingTrim is compatible with all fastq sequences file. Fastq format is a text based sequence format able to contain a very large number of sequence as text lines. Both DNA sequences and their quality are written using standard ASCII characters; for DNA sequences are used simple characters (ex: A for adenine, C for cytosine, T for thymine and G for guanine), while other special ASCII characters are used for quality. A typical fastq sequence should look like this:

```
# @FCC0H3RACXX:8:1101:2182:2178#ACTGTTCC/1
# TTACCGCGGCTGCTGGCACGGAGTTAGCCGGTGCTTCTTCTGTGGGTAACGTCAGGTCAAGGCGCT
# +
# bbbeeeeeggggiiiihiiifhghiiiihiaadbfghifebdege\b^_bbcccc]`b_bcaccc
```

For a more detailed description about fastq file format see the Wikipedia link: http://en.wikipedia.org/wiki/FASTQ_format.

4.1 OPEN A SEQUENCES FILE

To open a sequences file in the program the user can click on <File → Open File> or type the "Ctrl+o" shortcut on his keyboard. After that, the file open windows will appear on the screen and the user can select the file to open. Unfortunately, fastq format has not defined extensions, .fastq, .fq and .txt are the most used; if the user has a fastq file with another extension he must select the "All file" option in the extension menu in the <File Open> windows and then select the right file to open. Otherwise he will not be able to see and select his file. After selecting the file and press the <Open File> button the <Input File> section in the main windows will fill with the path to the selected file.

4.2 ANALYZE A SEQUENCES FILE

After had successfully opened a sequences file the user can analyze it in order to see the quality and length distribution of the DNA sequences present in the file. If the user has not opened a file yet, when he press the <Analyze> button, an <Open File> window will appear and he can select the interested file from here.

In order to analyze the file the user had to press the <Analyze> button in the <Controls> section of the <Main Window> (see 3.2 above). When the user press the button, the <Progress Bar> will begin to move and the file will be analyzed. After that, the <Reads Properties> window will display all the statistics related to the file. If the user wants a more accurate description of quality and length distribution, he can press the <Plot> button in the <Controls> section of the <Main Window>, <Plot Window> opens and the software begins to deeply analyze all the sequences in the file. When the program has finished analyzing data a plot will appear in the <Raw data> section of the <Plot Window> (see 3.2 above).

The user can now save all the plots by simply right clicking them and choose the "Save as" option in the popup menu.

4.3 CONVERT FASTQ FILE TO FASTA

StreamingTrim can also convert fastq file into a simple fasta file. In order to do that the user has to open a file (see 4.1 above) and then just click on the <FASTA> button in the <Controls> section of the <Main Window>. After done that a <Save FASTA file> window opens and here the user can choose the fasta file destination and the preferred fasta extension to add at the file.

During the conversion process the program transforms each fastq sequence id in a fasta id, by simply change the '@' character with a '>' character. The sequence is reported without changes and the quality is eliminated. If the user wants to find a specific read in the fasta converted file, he can search for its id without the first character '@'.

5 TRIMMING THE READS

The principal function of StreamingTrim is to cut low quality bases from each sequence in a DNA sequences file. In order to perform this operation, the software uses a conservative algorithm described in the section 3.1. The user can define some trimming parameters or he can let the software chooses the correct parameters automatically.

First of all, in order to start the trimming process, the user has to open a valid input file as described in section 4. Then, the user can proceeds to start the analysis pushing the <Trim> button in the <Main Window → Control Window> section. When the <Trim> button is pressed a <Save File> window appear and the user can choose the destination and the name of the file containing the trimmed reads. After the <Progress Bar> begins to move and the trimming process starts using the default trimming parameters or the user defined parameters (if previously specified).

If the user wants to convert the trimmed file to fasta format, he had to open it in the trimmer as input file and convert it by using the <FASTA> button in the <Main Window → Control Window> (see section 4.3 above). Otherwise, the user can select the <Trim to FASTA> check box before beginning the trimming process in the <Main Window → Control Window>, as discussed in the section 5.2 below.

5.1 ADVANCED OPTIONS

In the <Advanced Option> window (see 3.2 above) the user can specify some trimming parameters in order to adjust the trimming process to his will. Here, all the advanced options are described in order to understand the complete StreamingTrim functionality.

1. <Cutoff> this parameter represents the quality cutoff to be used by the software during the trimming process. Typically, the quality range of a fastq file goes from 0 to 40, representing hypothetical error probabilities of 100% and 0.01%, respectively. If this parameter is not selected, the trimmer chooses a cutoff based on the mean quality and the standard deviation of the reads in the given file, as described in chapter 3.1.
   The user can change this parameter in order to perform a more or less stringent trimming by using higher or lower cutoff values, respectively.
2. <Offset> this parameter indicates the number of bases to eliminate at the beginning of every reads. Setting a value higher than 0 is useful when the presence of adapters or some unwanted region at the beginning of each sequence is known. Otherwise it is recommended to leave this parameter unchecked.
3. <Minimum Length> here the user can specify a length cutoff (in bases). Sequences that, after the trimming process, have a length lower than this parameter are not saved in the output file. This parameter is very useful in amplicon based analysis, where reads that result too short after trimming are useless for following analyses (e.g. taxonomic identification).

5.2 THE <TRIM TO FASTA> FUNCTION

StreamingTrim can convert a trimmed file into fasta format while the trimming process goes on. If the checkbox <Trim to FASTA> in the <Main Window → Control Window> is selected, when the user starts the trimming process the software simultaneously converts the output file to fasta format.

When the checkbox is selected from the user, a <Save FASTA file> window opens and the user can chose the directory and the file name he prefers.

This function is very useful if there is a need to trim more than one file with the same parameters, without analyzing them each time. In this way the trimming and conversion process are speeded up.

6 PLOTTING RESULTS

StreamingTrim supports a plotting function that allow the user to plot a graphical representation of the quality and length distribution of the sequences pre- and post- trimming process. As you may have just see in the section 3.2 the software contains a plot window accessible through <Window → Show plot window>; in this window the program draws the plot related to the raw sequences file and/or the trimmed sequences file.

If the user has open a file without trimming it and press the <Main Window → Control Window → Plot> button, the program automatically begins to analyze the raw file and draw plots related to that file. Otherwise, if the user have opened a file and trimmed it, when he clicks on the <Plot> button, the software begins to analyze the two files and draws plots related to each file (each plot in a separate window, as described in 3.2).

6.1 QUALITY PLOT

If the <Plot Window → Plot Options → Quality distribution> radio button is selected, the software displays plots related to the quality distribution along the sequences in the input file. There are two possible types of plot that can be selected:

1. <Deviation Plot> is a representation of the DNA bases quality distribution along each sequence. In the x-axis the length of the sequences is reported. If there are sequences with different length, then the length of this axis is the length of the longest sequence. In the y-axis the quality values from 0 to 40 are reported. The mean quality is represented as a bold line while the range between maximum quality value and minimum quality value is represented as a blue surface. In this way the user can see the distribution of every base quality, and not only the mean or the standard deviation.
2. <Box Plot> this is a standard box plot representation of the quality distribution for each sequence in the sequences file. If you have reads longer than 200 nt, this type of visualization can be very difficult to read, otherwise if you have short reads (about 100 – 150 nt) this plot can be very useful since also the median and the first and third quartile (as a normal boxplot) are reported.

6.2 LENGTH PLOT

If the <Plot Window → Plot Options → Length distribution> button is selected, StreamingTrim displays the plot of the reads length distribution. Here, only one type of plot is possible, where in the x-axis the sequence length values (they can change by changing the input file) are reported and in the y-axis the number of reads in the file that has the corresponding length value is shown.

The user can zoom anywhere in the plot, by simple clicking and dragging with the mouse the part of the plot that he wants to zoom. In the bottom of the plot there is the number of reads that are found in the plotted file.

7 COMMAND LIST

Here, all the StreamingTrim commands are reported as a list with a brief description. The commands are divided in three section, one for each software's window. Some commands are followed by a cross-reference to the section in which are explained in more detail.

7.1 MAIN WINDOW

- <File → Open File> - opens the input file (4.1).
- <File → Exit> - close StreamingTrim.
- <Window → Show plot window> - show the plot window (3.2).
- <Window → Show advanced option> - show the advanced option panel (3.2).
- <? → info> - show the info window.
- <Input File> - the path to the input file (3.2).
- <Reads Properties → Input File> - show input file quality mean and standard deviation (3.2).
- <Reads Properties → Trimmed File> - show trimmed file quality mean and standard deviation (3.2).
- <Reads Type> - switch between encoding type (3.2).
- <Controls → Analyze> - analyze the input file quality distribution (4.2).
- <Controls → FASTA> - convert the input file in a fasta file (4.3).
- <Controls → Trim> - starts the trimming algorithm and saves a trimmed file (3.1).
- <Controls → Plot> - opens the plot window and plots quality and length distributions (6).
- <Trim to FASTA> - saves a fasta file while the trimming algorithm goes on (5.2).

7.2 ADVANCED OPTION WINDOW

- <Cutoff> - sets a user defined cutoff value (5.1).
- <Offset> - sets a user defined offset value (5.1).
- <Minimum length> - sets a user defined minimum length value (5.1).

7.3 PLOT WINDOW

- <Plot Options → Quality Distribution → Deviation Plot> - show the deviation plot of the reads quality distribution (6.1).

- <Plot Options → Quality Distribution → Box Plot> - show the box plot of the reads quality distribution (6.1).
- <Plot Options → Length Distribution> - show the bar chart of the reads length distribution (6.2).