User manual

A quick manual for

123Fastq _{1.3}

Highlights:

- Revolutionary graphical user interface in the field of next-generation sequencing data analysis open source tools
- Quality check of sequencing reads in two modes
- Trimming by large amount of user defined parameters
- Adapter decontamination by two approaches
- FASTQ file error correction through trimming process (Format errors)
- Recommendations to trim a FASTQ file based on quality check results
- Interaction between quality check and trimmer tabs
- Fast5 to FASTQ converter
- SAM/BAM to FASTQ converter
- Operable in both Unix(Linux & MacOS) and Windows platforms using only one .jar file

Introduction

123NGS has great aspirations for fulfillment. The ease of next-generation sequencing (NGS) data analysis for everybody is its main goal. Now, at the first step, it introduces 123Fastq, a tool that tries to meet all the user's needs in the pre-processes of NGS reads.

Contents

How to download and run 123Fastq?

The latest version of 123Fastq is available on the 123NGS page at SourceForge:

https://sourceforge.net/projects/project-123ngs/

123Fastq is written in java and capable to launch in Linux, Windows or macOS by using only one .jar file. In other words, 123Fastq is cross-platform. The only requirement is to install java runtime environment (JRE) version 1.8 on your operating system.

To run 123Fastq, at the command line of your operating system, change the working directory to the extracted 123Fastq folder and then type below command and press Enter, simply.

   #   Java –jar 123Fastq.jar

Quality check

Each sequencing run should be assessed based on the quality and purity of output reads. 123Fastq performs quality checks in two ways: "Single-Mode QC" and "Comparative-Mode QC". Former makes a quality report for an imported file or a series of files, and later makes a quality report for two files or two series of files which is useful for Paired-end sequencing runs or comparison of a file before and after trimming. However, quality check modules of 123Fastq use the core of FASTQC code but it has added some improvements. In the following, these improvements have been discussed. For more information about FASTQC-like modules, please refer to the FASTQC manual.

- ✓ **A new interactive module: "Distributions of Mean Qualities per Reads Length"**
  In this interactive quality check module, 123Fastq presents distribution of reads based on mean quality and length through a bubble chart. This chart will be displayed only when imported reads have > 3 different mean quality and length.
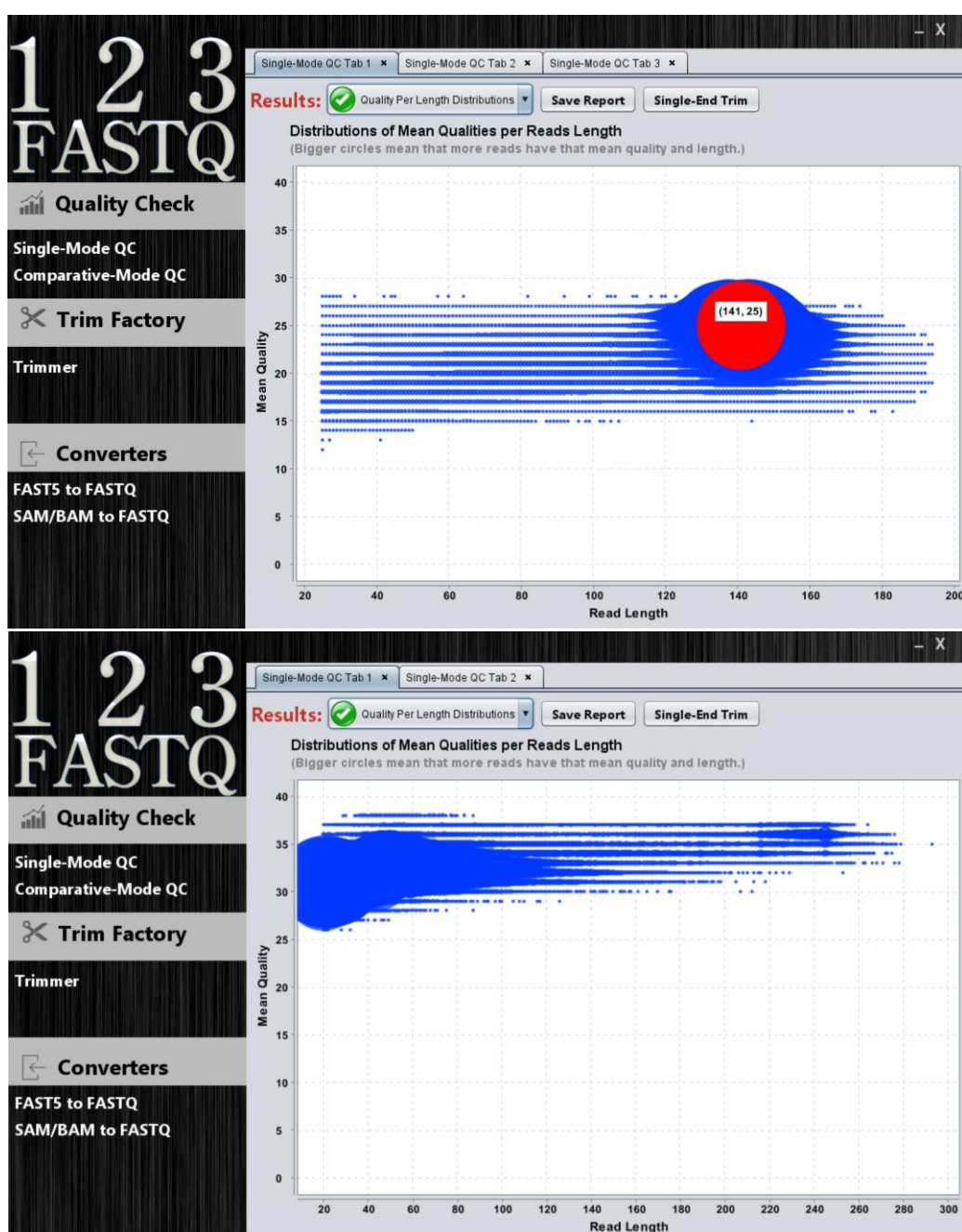
✓ **Graphical improvement of plots; made them more informative**

Some of the plot's features like x-axis and y-axis units have become more informative (Specially, N-content and adapter content line charts). Also, some of the line charts have been presented as bar charts, now. Plots visual features have been improved, especially box plot of the "per base sequence quality".
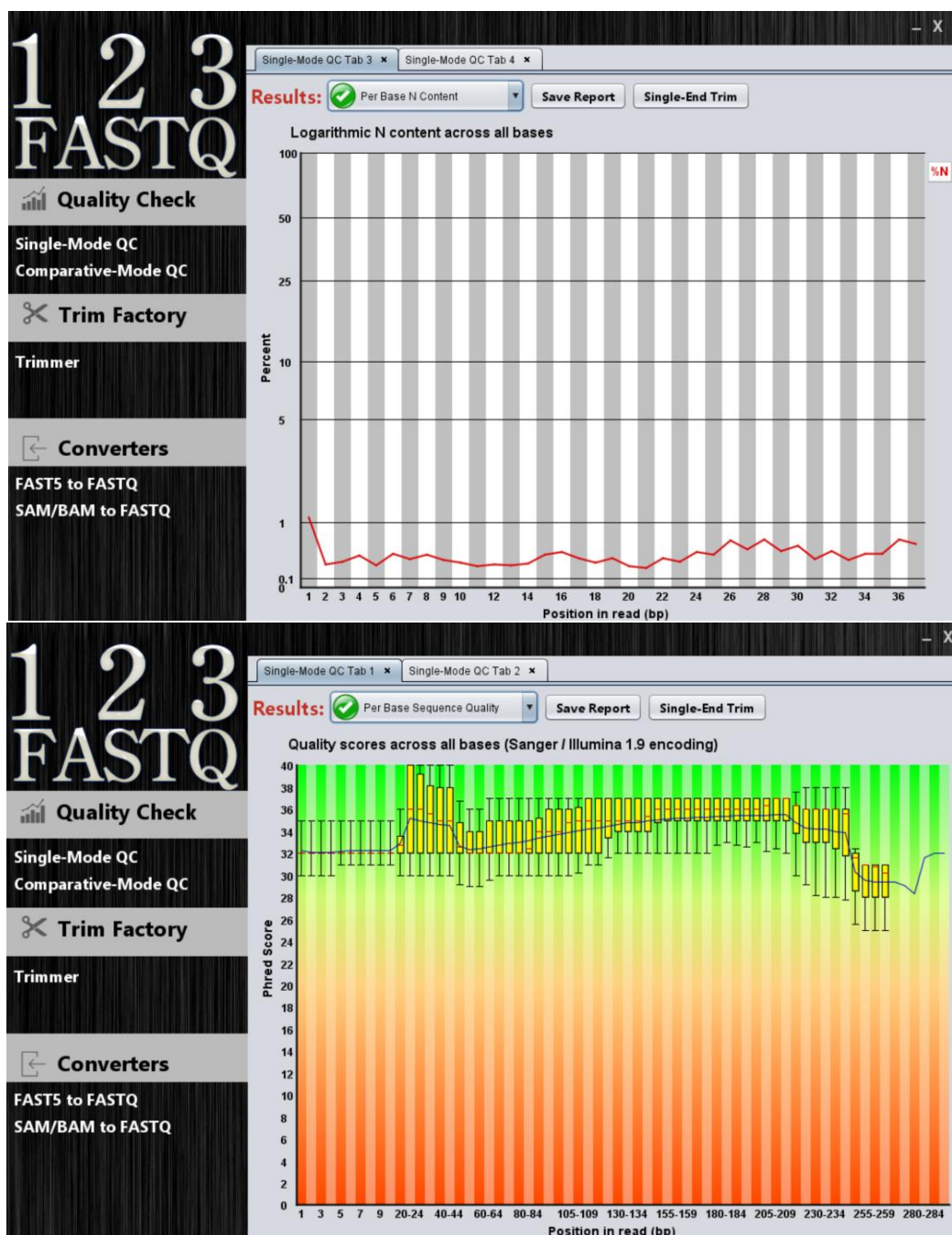
✓ **Recommendations to trim**

We believe trimming process should perform under the supervision of human intelligence, but this is not inconsistent with the usage of tools suggestions. Although the trimming strategy depends on the user's downstream recruited pipeline, 123Fastq can provide recommendations to improve imported FASTQ files as far as it can. 123Fastq generates recommendations based on mean qualities, most detected adapter sequence and occurred errors in the reads, simply.
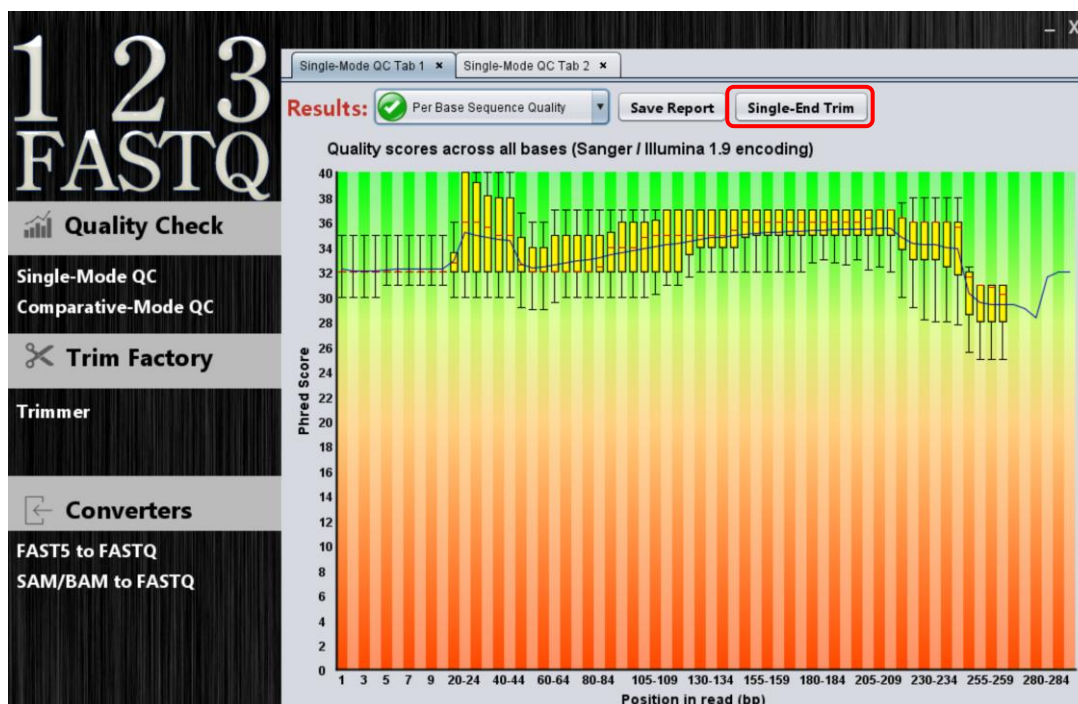
"Single-end trim" button is a gate to a new trim tab which its parameters completed based on imported file quality check results.

✓ **Present results for comparison of two files or two series of files**

By using Comparative-Mode QC, users can apply one file or more as the forward file (first file) and also one file or more as the reverse file (second file). If users apply more than one file as the forward/reverse, 123Fastq assemble them to make a unified report for the forward/reverse (or the first/second file).

In the Comparative-Mode QC, if trimming is possible (<u>one</u> <u>FASTQ</u> file imported as forward <u>and</u> also reverse), "Paired-End Trim" button will appear in the result panel. This button is a gate to a new trim tab which its parameters completed by recommendations based on quality check results.
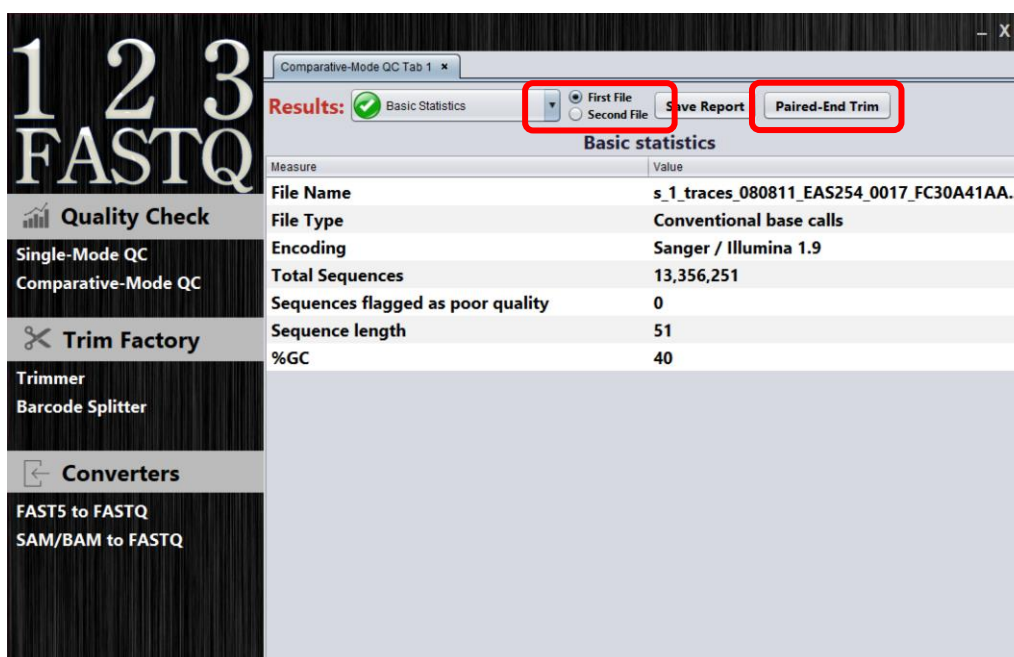
The left red rectangle shows radio buttons which can be used to switch between forward and reverse file(s) results, promptly.

Trimmer

123Fastq has provided wide options for trimming purposes. Trimmer works with FASTQ files (PHRED +33 or +64 encoded, Uncompressed or 'gzip' and 'bzip2' compressed).

To trim an imported FASTQ file, we have defined a logical trim cascade that has several steps (discussed below). Although it is recommended to trim based on this order, trimming can be better managed through running trimmer in several steps by user-desired order.

In the 123Fastq trimmer, some of trimmomatic functions have been included, but some options have been added to make trimming more flexible for the user.

Adapter decontaminators:

In the current version of 123Fastq trimmer, we have provided two approaches to remove adapters: "K-mer based" and "Trimmomatic".

K-mer based:

In the "Single-end mode", this approach clips the read once the user defined adapter sequence (K-mer) exactly found.

However, it's not recommended to use this approach in "Paired-end mode", both the user defined k-mer and its reverse-complement can be considered to search.

Trimmomatic:

This approach is completely derived from trimmomatic and we haven't make any change in their codes and terminology.
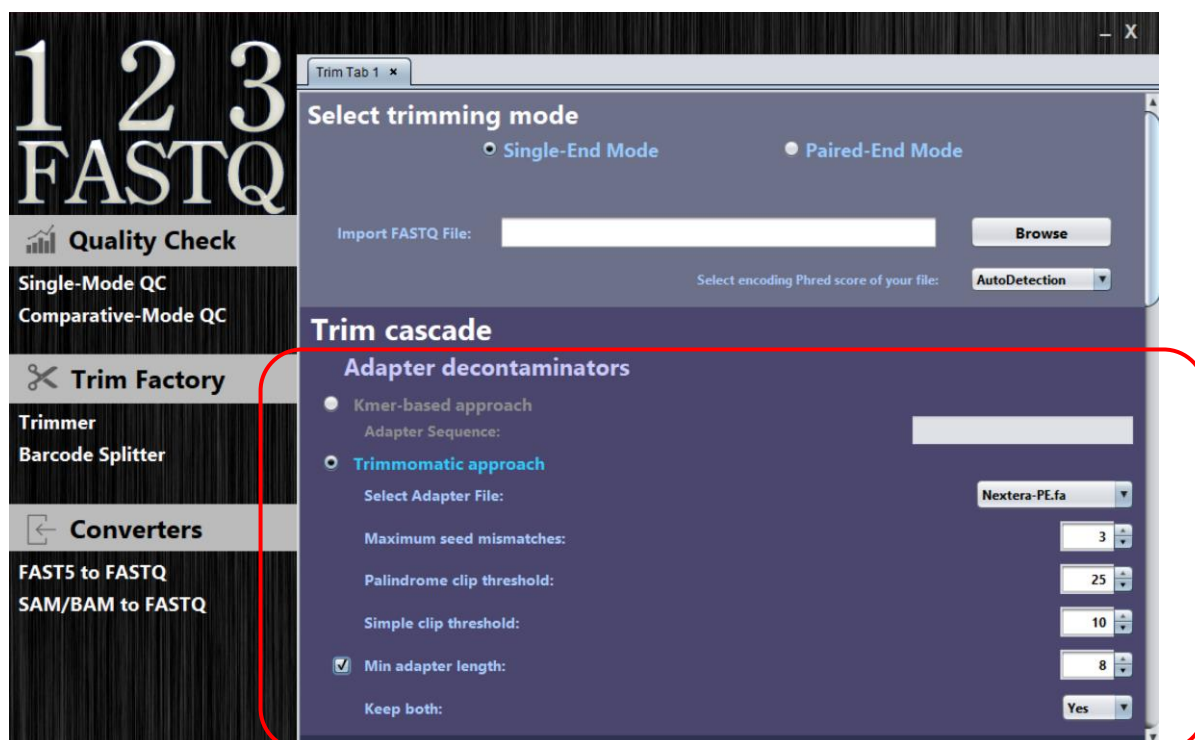
Adapter decontaminators (inactivated by default)

Trim options (Start options):

As the red rectangle shows, several simple self-explanatory trim tasks have been provided. To find out how they trim your files, take a look at the example below.
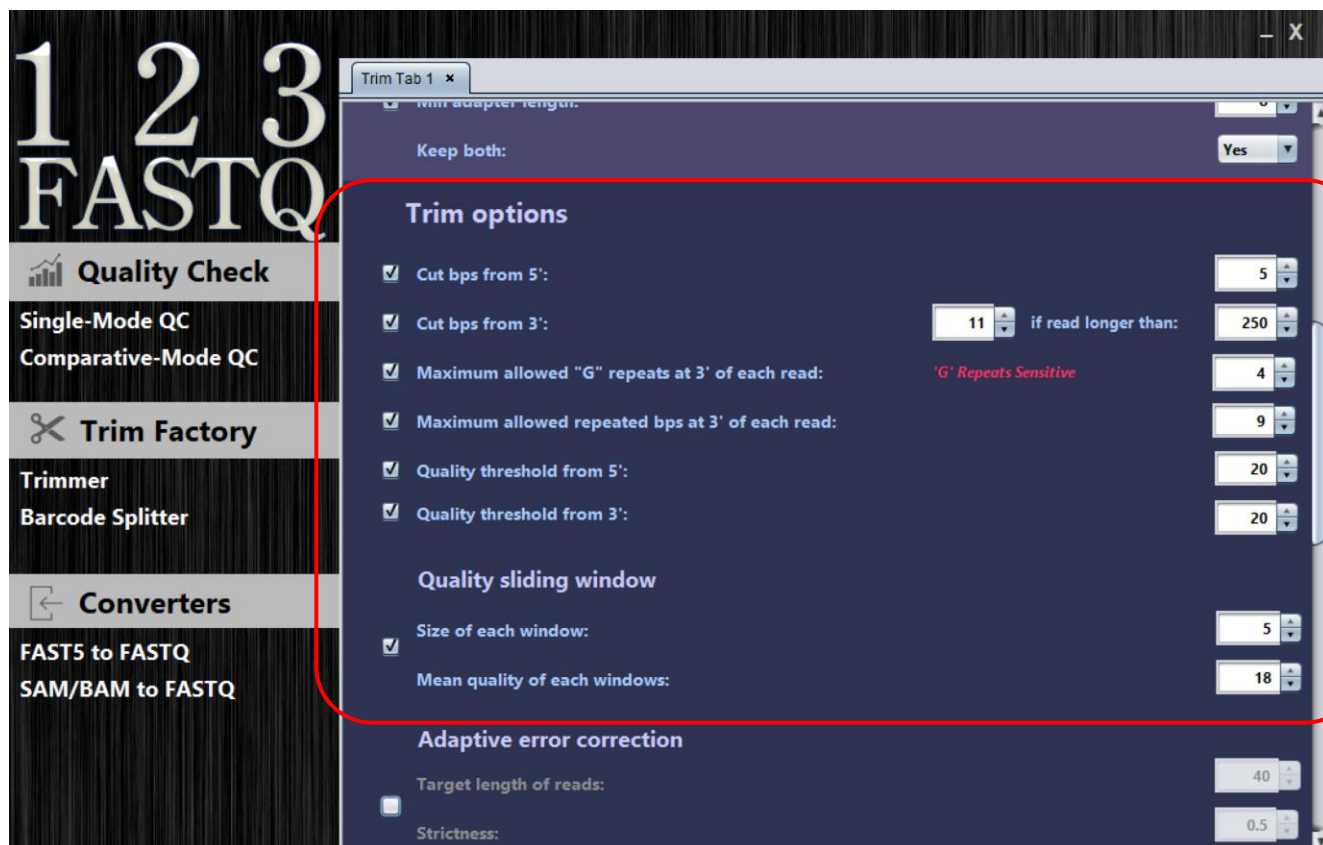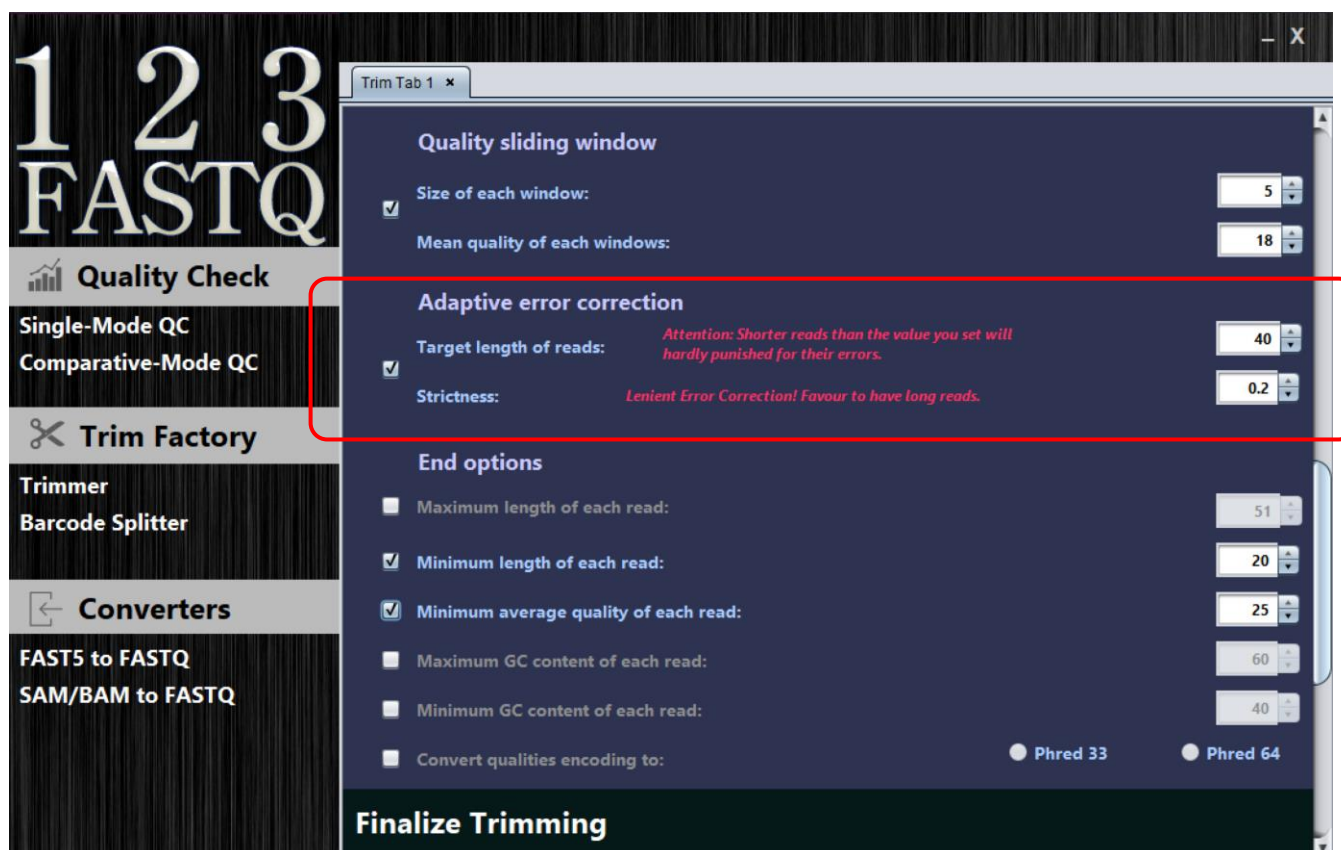

By using these parameters, the trim cascade steps are:
*(**It's not recommended to use all the cascade steps, simultaneously.**)*
-5 bases from the start of the reads will crop.
-11 bases from the end of the reads will crop if they are longer than 250 bases.
-At this step, if >= 4 "G" repeats exist at the end of the read, all the "G" repeats will crop.
-At this step, if >= 9 repeats exist at the end of the read, all of them will crop.
-All the bases of the start of the read (at this step) which their quality scores are below 20 will remove.
-All the bases of the end of the read (at this step) which their quality scores are below 20 will remove.
-Now, a 5 bases quality sliding window moves from start to end of the read and if average quality within the window falls below 18, the rest of the read will remove.

Adaptive error correction:

An adaptive quality trimmer which balances read length and error rate to maximize the value of each read. Actually, this step is the MAXINFO function of trimmomatic. Users should provide two values for this step, "target length" and "strictness". Shorter reads than the user-defined target length will be punished for their errors, hardly. Strictness, which should be set between 0 and 1, specifies the degree of error correction. Higher values favour to remove errors and lower values favour to have longer reads (for more details, please refer to trimmomatic manual).

Adaptive error correction (inactivated by default)

End options:

At the end of trimming cascade, another simple self-explanatory tasks can be used.
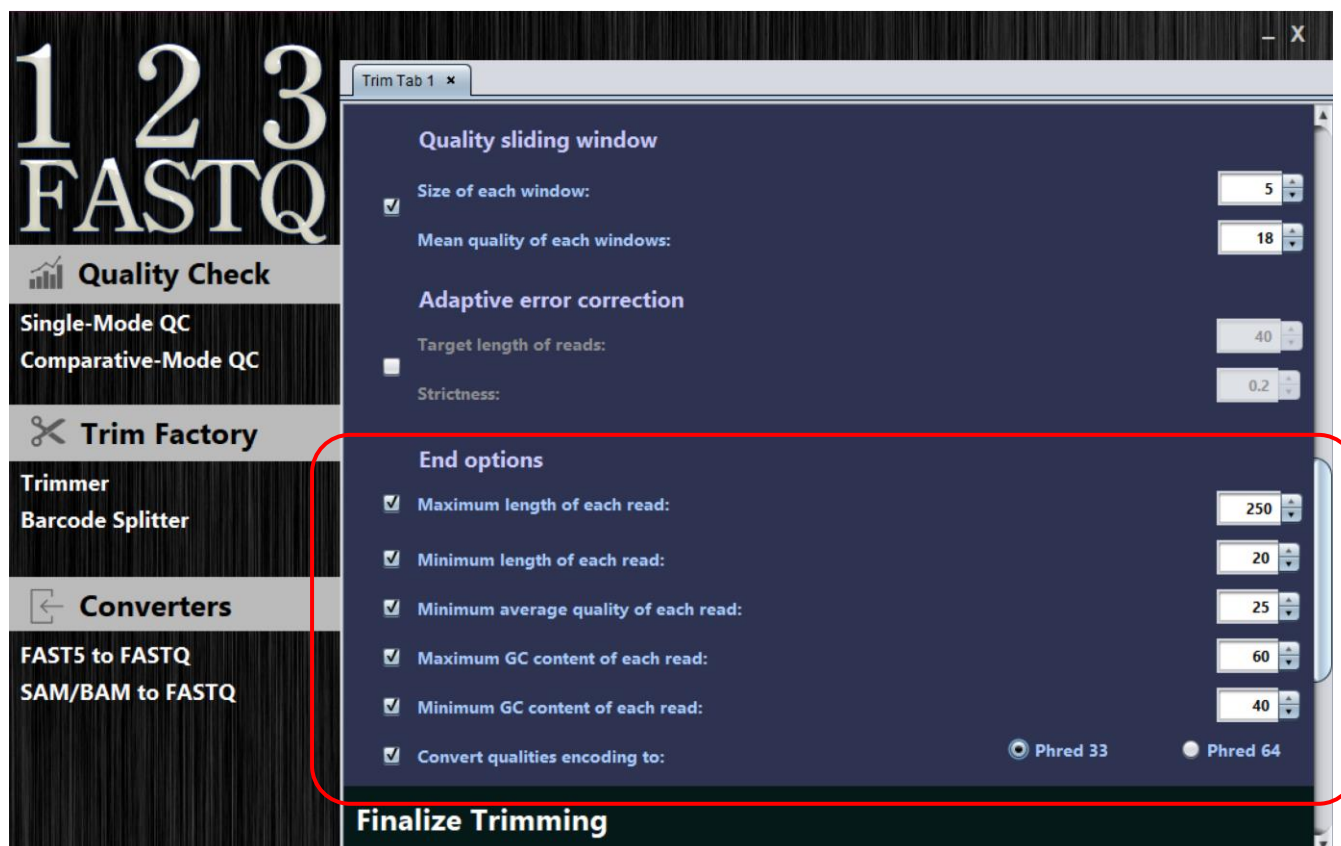
*By using these parameters, the trim cascade end steps are:*

*(**This is an example to describe how parameters work, it's not recommended to use all these steps, simultaneously.**)*

-At this step, the read will be thrown away if it has more than 250 bases.

-At this step, the read will be thrown away if it has fewer than 20 bases.

-At this step, the read will be thrown away if its average quality is fewer than the 25.

-At this step, the read will be thrown away if its GC content percent is upper than 60%.

-At this step, the read will be thrown away if its GC content percent is lower than 40%.

-At the end of the trimming cascade, if the read was not discarded, quality encoding format will be converted to PHRED 33, if it is anything else.

It should be noted, at each step of trimming cascade, if the read is completely discarded for any reason, the cascade stops and the next read trimming process begins at the beginning of the desired cascade.

Finalize trimming:

In this panel, firstly, the user should browse the output path. For the "single-end mode", users can define the output file name, but in the "Paired-end mode", users can only browse the output path. Accordingly, the 4 output file names will be address using the input files. As the trimmer of 123Fastq is multi-threaded, tasks can be done faster by using this option. Additionally, it's provided to remove comments of inputted reads to reduce the output file size, significantly. The last option which is inactivated by default is to create a text-based report sheet in the output path. Finally, by clicking "Let's trim" the trimming process will start.
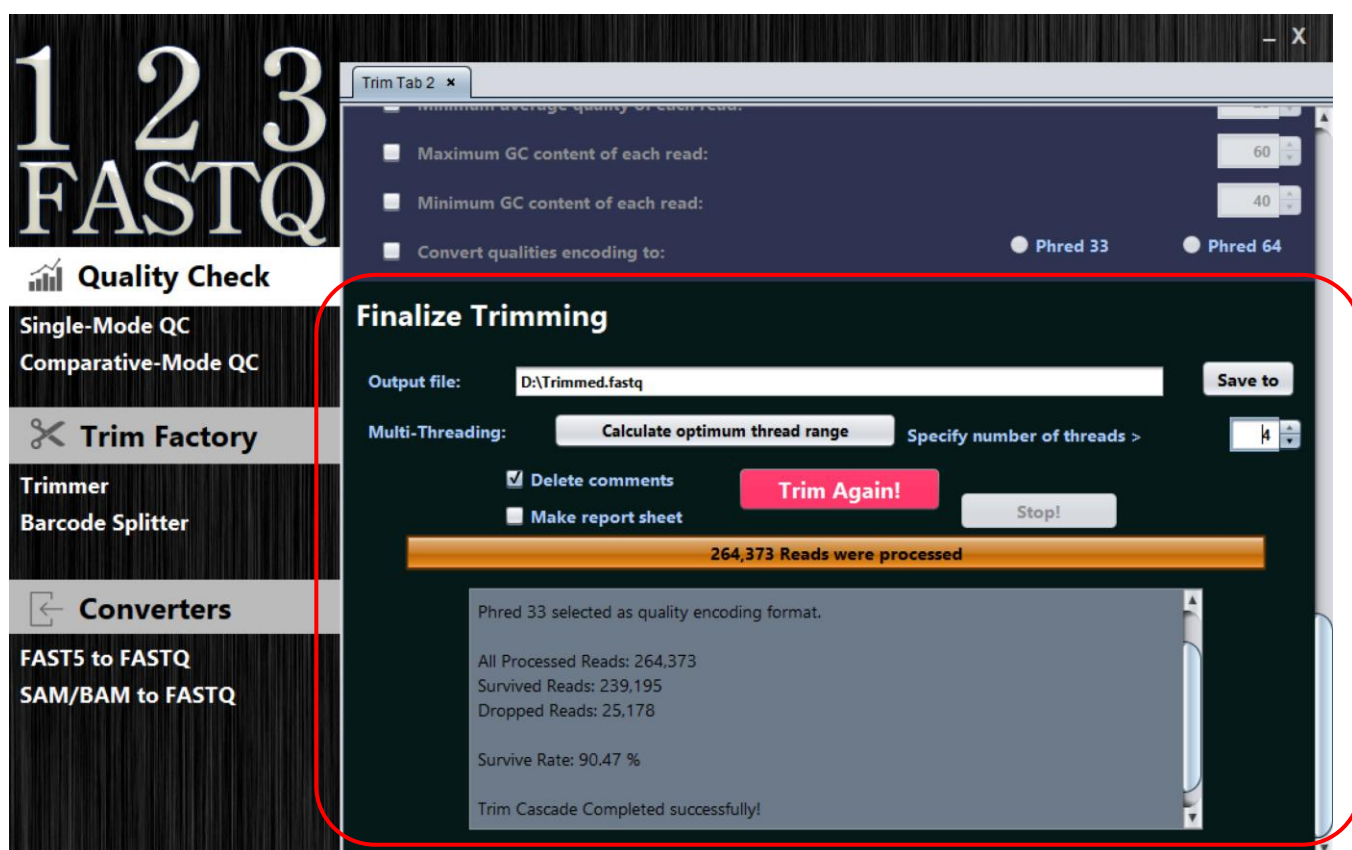
Another useful feature of 123Fastq trimmer is the FASTQ format corrector. If an imported FASTQ file has corrupted reads that reads will skip to reach next read. In other words, if a read has a problem like identifier line of a read starts without "@", comment line starts without "+" or one or more lines of a read were lost, that corrupted read will be skipped.

Barcode Splitter

The Barcode splitter module was removed in this version.

Converters

In the current version of 123Fastq, two format converters have been provided: "Fast5 to FASTQ" and "SAM or BAM to FASTQ".
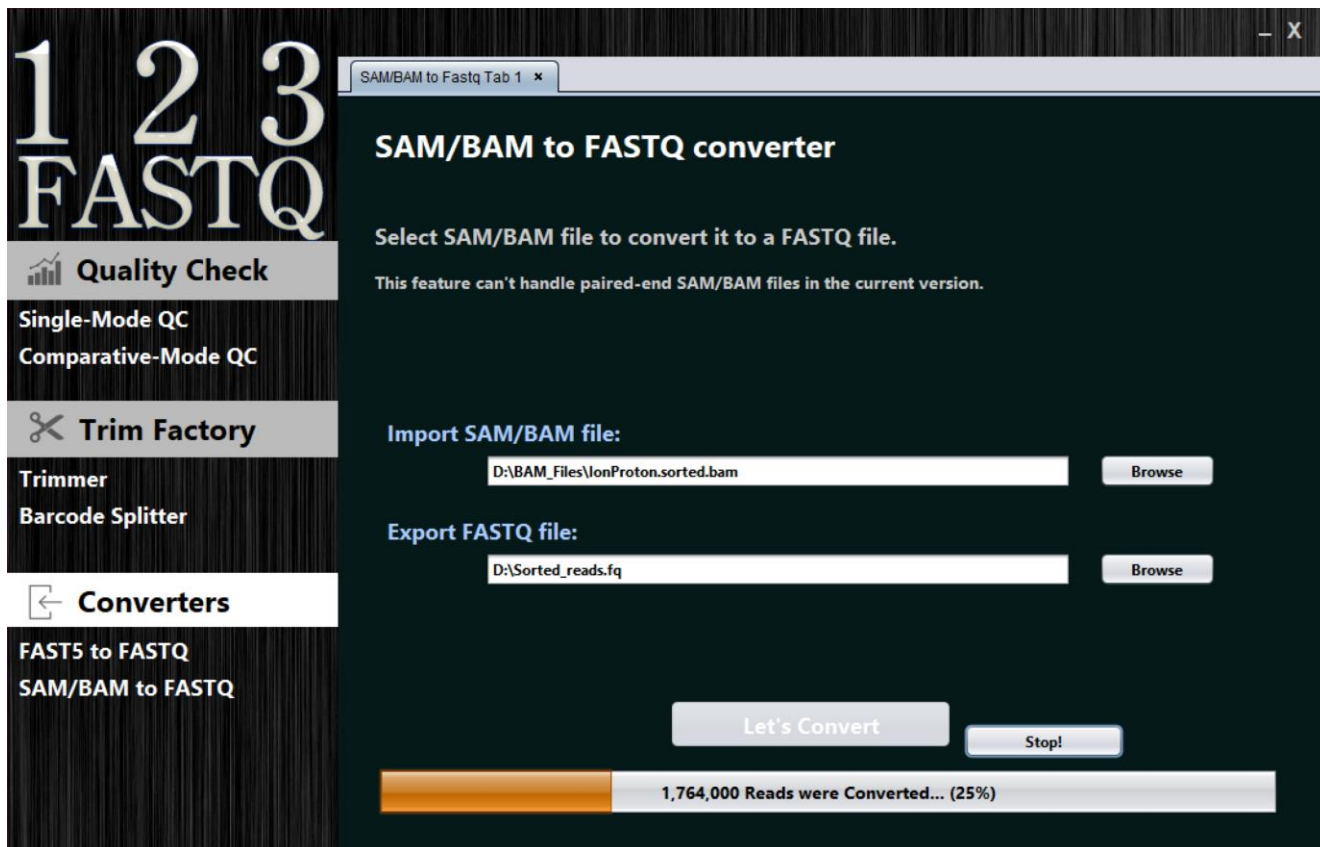
Fast5 to FASTQ:

In the third generation sequencing, oxford nanopore technology sequencer outputs are Fast5 files. Each Fast5 file contains information about only one read. Trimming in this format isn't possible by the current version of 123Fastq trimmer. Users can import one or more Fast5 files to convert them to a FASTQ file. In the following updates, MultiFast5 files will be supported in the quality check section and also here.
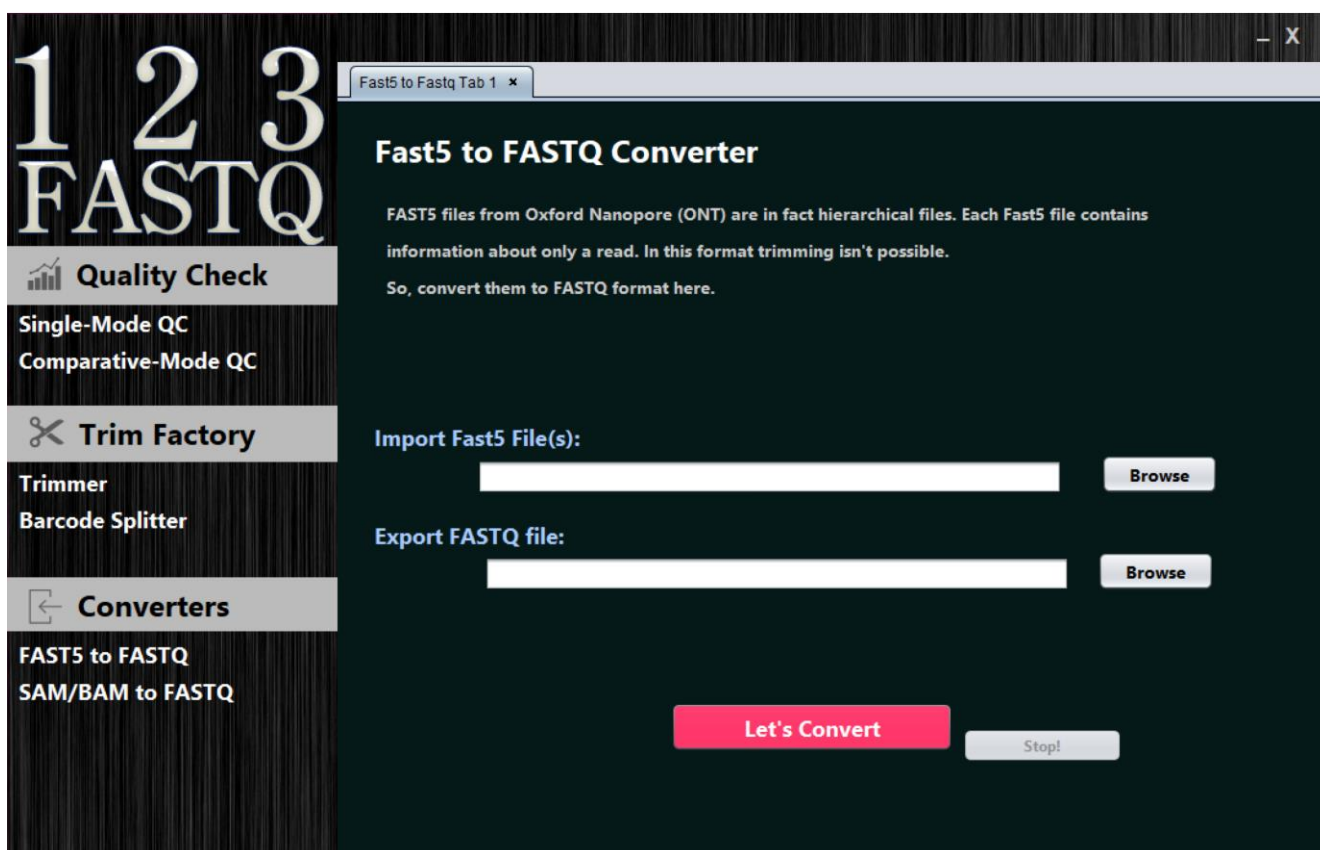
SAM or BAM to FASTQ:

In some conditions, an analyst may want to make FASTQ file from aligned reads. 123Fastq in this section provided a simple convertor that makes this conversion possible.

Conclusion

In a nutshell, up to now, 123NGS collects some of the best tools in the pre-process of next-generation sequencing reads and adds some improvements to them in a highly interactive graphical user interface and this way will continue…

To help and encourage us, give us your feedback here:

https://sourceforge.net/p/project-123ngs/questionsabout123fastq/general/thread/c0aa7dbc04/#6e43

Yours Sincerely,

Milad Eidi

Changes

- Version 1.1

  - Improvement in the normalization of bubble sizes in "Distributions of Qualities per Reads Length" module

- Version 1.2
  - Bug fixes
  - The Barcode splitter module was removed.

- Version 1.3
  - N-content and Adapter content plots in the quality check section are on **logarithmic** scale, now. So they are more informative.
  - Paired-end trimmer output files are named based on the input file names, now.