International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

Next Generation Sequencing quality trimming (NGSQTRIM)

Danamma B.J[1], Naveen kumar[2], V.G Shanmuga priya[3]
[1]*M.Tech, Bioinformatics, KLEMSSCET, Belagavi*
[2]*Proprietor, GenEclat Technologies, Bengaluru*
[3]*Asst. Professor, Dept. of Biotechnology, KLEMSSCET,Belagavi*

Abstract-Next generation sequencing (NGS) technology has revolutionized genomic and genetic research. NGS refers to high-throughput sequencing that parallelizes the sequencing process, producing millions or billions of sequences concurrently.

NGS generates enormous amount of data, a single sequencing run could produce about one tear bytes (TB) of data which is difficult to maintain and the presence of poor quality or technical sequences such as adapters in NGS data can easily result in suboptimal downstream analysis so data needs to be pre-processed. Before analyzing data, quality checking and trimming of data is important. To do this Trimmomatic tool is used to check the quality of data generated from Illumina sequencing. Trimmomatic is a command prompt tool it requires more computational knowledge to perform, for a non-technician or for a biologist. To overcome this issue it is necessary to develop a GUI for a Trimmomatic tool, which helps biologist to perform quality check with less computational knowledge.

A transition from a Trimmomatic command prompt tool to Graphical User Interface (GUI) is done using java where user can do quality check in minimal button clicks and time saving.

Keywords-Next-generation sequencing, Adapter trimming, Data pre-processing, command line, Graphical User Interface.

I. INTRODUCTION

Next Generation Sequencing is often referred to as massively parallel sequencing, which means that millions of small fragments of DNA can be sequenced at the same time, creating a massive pool of data.

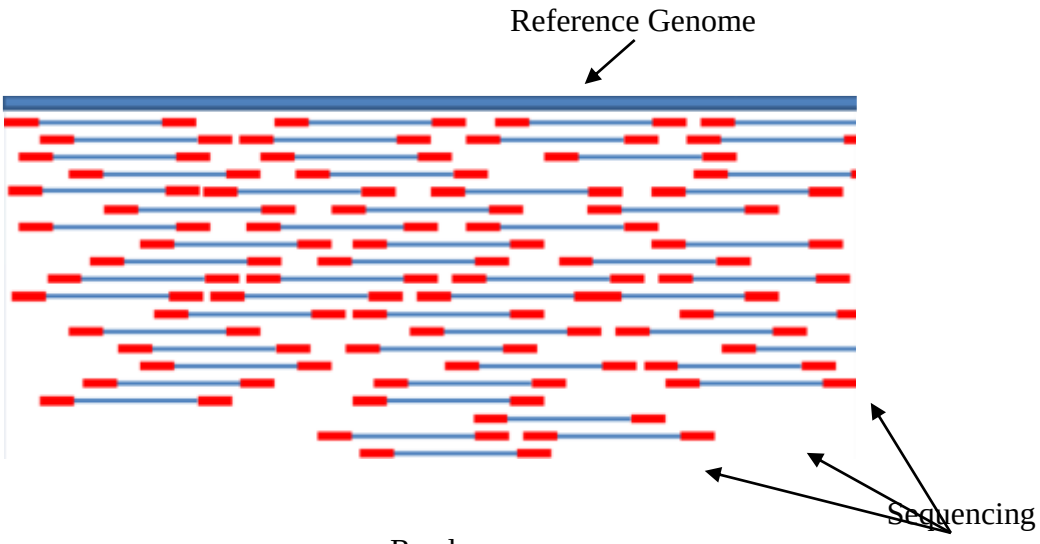

Figure 1. Thousands of Fragments being sequenced in parallel

This pool of data can reach tera/gigabytes in size, which is equivalent of 1 billion base pairs of DNA this leads to the problems like memory usage and execution time also the presence of poor quality or technical sequences such as adapters in next-generation sequencing (NGS) data can easily result in suboptimal downstream analysis so sequence quality filtering and trimming is must. [1]

There are many tools available to check quality and trimming of raw sequence but executing a series of tools in succession would involve the creation of intermediate files at each step, a non-trivial overhead given the data size involved, and would still require pair-awareness to be built into every tool used. These issues suggest that the typical approaches to achieve flexibility by combining multiple single-purpose tools are not optimal. As a result Trimmomatic was developed. [1]

Trimmomatic is a command line tool that can be used to trim and crop Illumina data as well as to remove adapters. These adapters can pose a real problem depending on the library preparation and downstream application. It evolved as more flexible, pair-aware and efficient pre-processing tool, optimized for Illumina NGS data.

Though Trimmomatic is a flexible tool it is difficult to use as it is a command line some disadvantages of command line tool are, commands are to be typed precisely. If there is a spelling mistake then command will not respond, if user mistypes any instruction it is often necessary to start from scratch again and user can't guess what instruction might be and user can't just have a go so.

In this study we have developed a Graphical User Interface (GUI) for a Trimmomatic tool comprised of various easy-to-use standalone tools for quality check and filtering, trimming, of NGS Illumina data. The toolkit allows automatic and fast parallel processing of large amount of sequence data with user friendly options. Given the importance of Quality Checking (QC) of NGS data, we expect that this toolkit will be very useful for the sequencing based biological research. [3]

II. MATERIALS AND METHODS

GUI for a Trimmomatic tool has been developed using java programming language. Java is a general purpose, high-level programming language developed by Sun Microsystems. Today Java is a commonly used foundation for developing and delivering content on the Web.

Trimmomatic combines two approaches to detect technical sequences. The first, referred to as 'simple mode', conducts a local alignment of technical sequences against a read. f the alignment score exceeds a user-defined threshold, the portion of the read that aligns to the technical sequence plus the remainder of the read after the alignment (towards the 3' direction) are trimmed from the read.

Simple mode can detect any technical sequence at any location within the read; however, the user-defined threshold must be set sufficiently high to prevent false positives. Thus, 'simple mode' cannot detect the short overlaps between a read and technical sequence which often arise in case of adapter read through.

Trimmomatic second approach to technical sequence detection, referred to as "palindrome mode", Palindrome mode can only be used with paired-end data. When a read-through occurs, both reads in a pair will contain an equal number of valid bases followed by contaminating sequence from opposite adapters. The valid sequence in each of the pair's reads will be reverse complements.

Trimmomatic palindrome mode uses these characteristics to identify contaminating technical sequences arising from adapter read-through with high sensitivity and specificity. Operating in the palindrome mode, Trimmomatic prepends the Illumina adapter sequences to their respective reads in

the paired-end data. The resulting sequences are then globally aligned against one another. A high scoring alignment indicates that the first parts of each read are reverse complements of one another and the remaining parts of the reads match their respective adapters. Read bases matching the adapters are removed.

Trimmomatic works with Illumina FASTQ files using phred33 or phred64 quality scores. The appropriate setting depends on the Illumina pipeline used. The default is phred33, which matches modern Illumina pipelines.

Trimmomatic works with single ended as well as with paired end sequence. Paired end reads provides improved alignment of sequencing data and better detection of chromosomal rearrangements: insertions/deletions/translocations and fusions.

For single-ended data, a single input file is specified and single output file of trimmed reads will be generated.
For paired-end data, two input files (for forward and reverse reads) and 4 output files (for forward paired, forward unpaired, reverse paired and reverse unpaired reads) will be generated. [1]

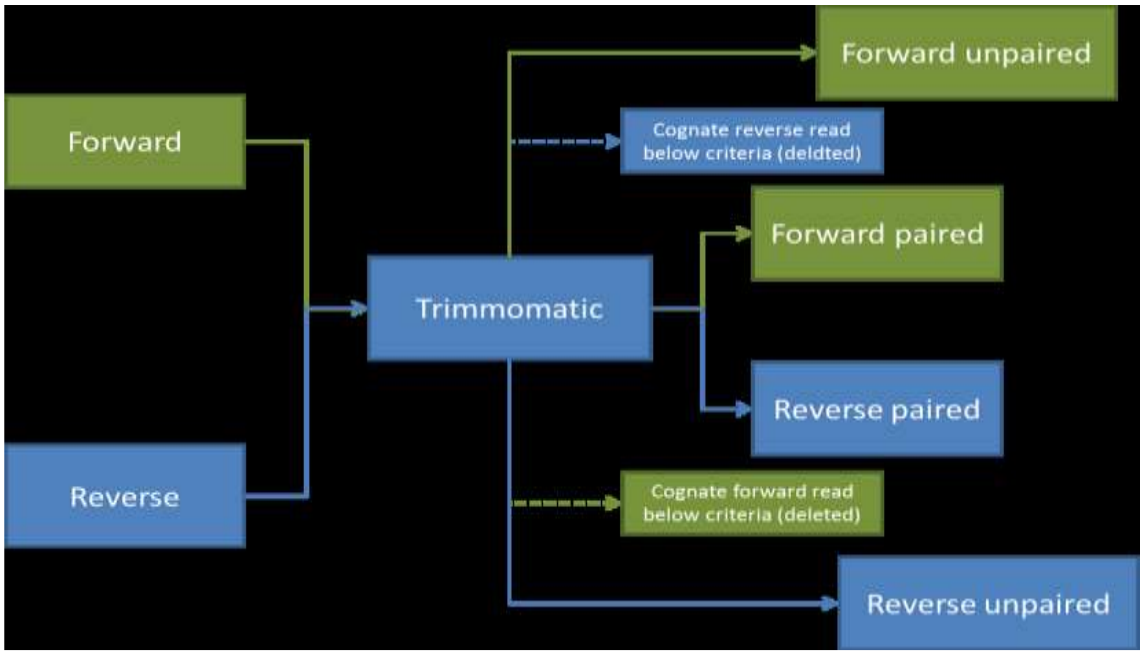

Figure 2: Flow of reads in Trimmomatic Paired End mode

Parameters specified by Trimmomatic command line tool have been implemented in GUI also. One of them is quality checking of raw sequence to do this FastQC is used.

FastQC used for quality assessment of the raw reads, includes an analysis of overrepresented sequences. When conducting this analysis, FastQC also checks to see whether any overrepresented sequences correspond to known Illumina adapter and primer sequences. If the resulting Overrepresented Sequences report flags matches with known adapter or primer sequences, these can be removed by using the adapter removal step.

To improve the performance of trimmomatic threads are implemented in the tool. Thread is a concept in java which improves performance of a program on multi-core computers. The number of threads can be specified by user.

Sliding Window quality filtering (e.g. 4:6) The Sliding Window uses a relatively standard approach. This works by scanning from the 5 end of the read, and removes the 3 end of the read when the average quality of a group of bases drops below a specified threshold. This prevents a single weak base causing the removal of subsequent high-quality data, while still ensuring that a consecutive series of poor-quality bases will trigger trimming. [2]
In e.g. 4 is the window size and 6 is the average quality of a group of bases.

Leading this step will remove low quality bases from the beginning of the reads, As long as a base has a quality value below the specified threshold the base is removed and the next base will be investigated. This step corresponds to the LEADING operation and the threshold parameter represents a phred score. A low value of 3 can be used to remove only special Illumina 'low quality regions' (marked with a score of 2), while a value in the range of 10-15 can be used for a deeper quality-based trimming (15 being more conservative in terms of required quality). The table below translates phred quality scores (ranging from 10 to 15) to base call error probabilities.

Phred quality scores to base call error probabilities.

Phred Quality score	Base call error probability
10	0.1
11	0.08
12	0.06
13	0.05
14	0.04
15	0.03

Table 1: Phred quality scores

Trailing this step will remove low quality bases from the end. As long as a base has a value below the specified threshold the base is removed and the next base (which as trimmomatic is starting from the 3‟ prime end would be base preceding the just removed base) will be investigated. This approach can be used removing the special Illumina “low quality segment” regions (which are marked with quality score of 2) [2]

Crop this step removes bases regardless of quality from the end of the read, so that the read has maximally the specified length after this step has been performed.

III. RESULTS

Graphical User Interface has been developed with the same parameters implemented by trimmomatic command line tool for single end and paired end and it is very easy to use

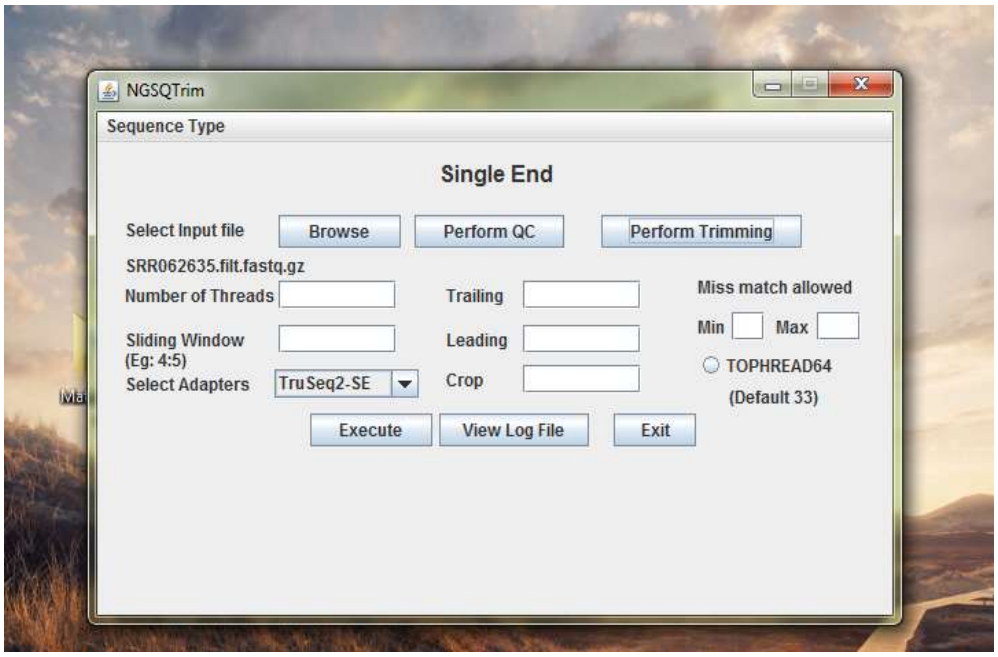

Figure 3: User Interface of Single end

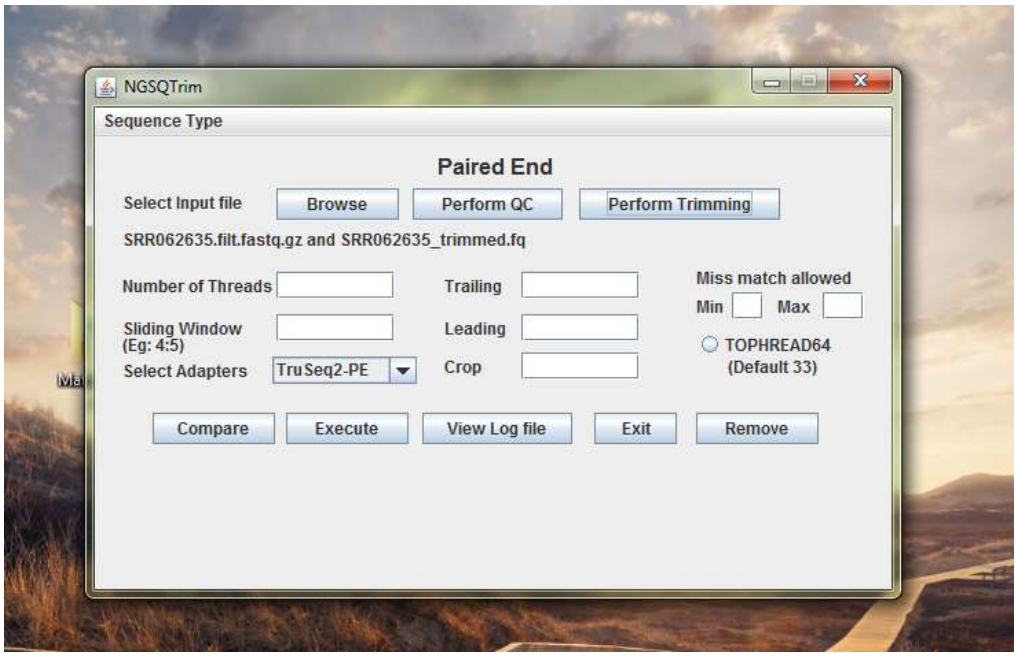

Figure 4: User Interface of paired end

REFERENCES

1. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30.15 (2014): 2114–2120
2. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic Manual: V0.32" *Bioinformatics* 30.15 (2014): 2114–2120
3. Patel, Ravi K., and Mukesh Jain. "NGS QC Toolkit: A Toolkit for Quality Control Of Next Generation Sequencing Data." Ed. Zhanjiang Liu. *PLoS ONE*7.2 (2012):e30619.
4. Lindgreen, Stinus. "AdapterRemoval: Easy Cleaning of next-Generation Sequencing Reads." *BMC Research Notes* 5 (2012): 337. *PMC*. Web. 30 Apr. 2016.
5. Zhou, Qian et al. "QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data." Ed. Zhang Zhang. *PLoS ONE* 8.4 (2013): e60234. *PMC*. Web. 30 Apr. 2016.