IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

A Comparative Study of Computational Tools for Biological Sequence Cleaning and Analysis

K.S.Mehta[1*], D.S.Mehta[2], V.Dahiya[3]
[1]Faculty of Computer Technology, GLS University, Ahmedabad, India.
[2]Faculty of Computer Technology, GLS University, Ahmedabad, India
[3]Institute of Information and Communication Technology, Indus University, Ahmedabad, India.

[*]Corresponding Author: krupa.mehta@glsuniversity.ac.in
Available online at: www.ijcseonline.org

Accepted: 10/Jul/2018, Published: 31/Jul/2018

***Abstract*—** The next generation sequencing(NGS) technology is playing an increasingly prominent role in capturing DNA and RNA sequencing by producing high-throughput sequences (HTS). The major challenge with HTS is the complexity and difficulty of data quality control (QC). Only a high quality data is capable for accurate diagnosis of the disease. For accurate diagnosis the data that needs to be analysed must be appropriate and correct. To fulfill this requirement, computer scientists have implemented the algorithms in easy to use manner that become convenient tools for biological research. The raw sequence generated by the NGS technologies is first cleaned and then moved further for clinical analysis. The step of cleaning includes removal of short sequences and trimming of inappropriate headers. This paper compares some popular, open source tools used for cleaning the captured sequences.

***Keywords*—** Illumina, FASTQ, FASTA, tag removal, single end, paired end.

I. INTRODUCTION

Due to the growth in the field of computer science, the analysis of biological sequences is becoming accurate and cost effective method for clinical diagnosis of rare or novel disease. New algorithms are emerging and existing algorithms are revised to analyse complex and novel biological sequence. This success presents many opportunities for many applications in clinical diagnosis. The advancement in genome sequencing technologies produces vast number of targeted sequencing. The first step towards perfect diagnosis is to capture the biological sequence. The captured sequence needed to be analysed to produce fruitful result.

The task of analysing the sequence is very important and crucial as the sequence produced are voluminous, complex and different from the sequences produced before. The field of computer science has the solution of this problem. Various algorithms are developed and revised to analyse the sequences, new algorithms are also being designed to serve the purpose.

To diagnose the disease accurately the sequence needs to be passed through various processing steps like: cleaning, mapping, variant calling and variant filtering and prioritization. Various algorithms are available and being developed to accomplish each step. Different algorithms for each processing step are available and it is converted into the easy to use tool by the computer scientists.

The first step in the bioinformatics workflow involves cleaning the sequence. It is necessary to trim a number of nucleotide reads off of the ends of sequences, because these end regions are more vulnerable to misreading than the interior regions [1]. Once the data has been cleaned, the next step of aligning the sequence is carried out. The cleaned data works as input for the alignment of the sequence which leads to the almost error free alignment. Thus, this cleaning step is very crucial in the process of accurate diagnosis of the disease. The process of cleaning raw sequences includes trimming of adapter, removal of duplicate reads, error correction, etc. Low-quality reads, PCR primers, adaptors, duplicates and other contaminants that can be found in raw sequencing data may compromise downstream analysis. Therefore, quality control (QC) is essential step in your analysis to understand some relevant properties of raw data, such as quality scores, GC content and base distribution, etc [2]. After performing the cleaning process, the sequencing data is ready to be processed for the actual step of sequence analysis.

The rest of the paper is organized as follows, section – I contains the introduction of the cleaning process of biological sequences, section – II contains the details of some of the open source tools available for cleaning and analysis of biological sequences, section – III compares the tools detailed in section – II and section – IV concludes the work.

II. COMPUTATIONAL TOOLS FOR SEQUENCE CLEANING

A. *Trimmomatic*

Trimmomatic is a fast, multithreaded 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. Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data [6]. Trimmomatic includes a variety of processing steps for read trimming and filtering, the main algorithmic innovations in trimmomatic are related to identification of adapter sequences and quality filtering [12].

Trimmomatic uses two approaches to detect technical sequences within the reads. The first, referred to as 'simple mode' which finds an approximate match between the read and the user supplied technical sequence. This mode has the advantage of working for all technical sequences, including adapters and polymerase chain reaction (PCR) primers, or fragments thereof. The second mode, referred to as 'palindrome mode', is specifically aimed at detecting this common 'adapter read-through' scenario, whereby the sequenced DNA fragment is shorter than the read length, and results in adapter contamination on the end of the reads [12].

B. *TrimGlore*

TrimGlore is a wrapper script that makes use of the publically available adapter trimming tool Cutadapt and FastQC for optional quality control once the trimming process has completed. The main aim of TrimGlore is to eliminate the RRBS libraries or similar type of sequencing datasets and club the solution of potential problems in one convenient process. TrimGlore works for any type of high throughput dataset [15].

C. *AlienTrimmer*

AlienTrimmer has high sensitivity and speed in removing alien oligonuceotide sequences from short-insert paired-end reads [6]. AlienTrimmer is a tool that detects and removes contaminant sequences from both the ends of next-generation sequencing (NGS) reads. It performs trimming based on the decomposition of the specified alien sequences into nucleotide *k*-mers of fixed length *k*. AlienTrimmer is able to determine whether such alien *k*-mers are occurring in both read ends by using a simple polynomial algorithm. Therefore, AlienTrimmer can process typical NGS single or paired ends files with millions read within minutes with very low computer resources [12][7].

D. *Skewer*

The Skewer tool implements a novel dynamic programming algorithm dedicated to the task of adapter trimming and it is specially designed for processing Illumina paired-end sequences. It implements the algorithm that utilizes the equality of diagonal adjacent elements in the dynamic programming matrix. Skewers is used to detect and remove adapter sequences. It trims the sequences based on the quality score in both single end and paired end [8].

E. *Seqpurge*

Seqpurge is a highly-sensitive adapter trimmer that uses a probabilistic approach to detect the overlap between forward and reverse reads of Illumina sequencing data. It supports paired-end sequencing data that is based on a probabilistic approach. SeqPurge can detect very short adapter sequences, even if only one base long [9].

F. *SolexaQA*

"SolexaQA calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data. Originally developed for the Illumina system (historically known as "Solexa"), SolexaQA now also supports Ion Torrent and 454 data." [10]. The SolexaQA tool produces the summaries of data quality in graphical as well as tabular format. It aims to create standardized diagnostic information that helps to identify low-quality data rapidly and easily. This tool also provides a dynamic trimming function to manipulate sequence data at the level of individual reads. This tool is also capable to processes large files and produces trimmed datasets that yield significant improvements in downstream analyses, including SNP calling and de novo sequence assembly. SolexaQA is a user-friendly tool designed to generate detailed statistics and graphical representation of sequence data quality both quickly and in an automated fashion. It contains associated software to trim sequences dynamically using the quality scores of bases within individual reads [11].

G. *PRINSEQ*

PRINSEQ is used to efficiently check and prepare the datasets prior to downstream analysis. PRINSEQ comes in two versions, web interface and desktop application. The web interface is simple and user-friendly, and the desktop application allows offline analysis and integration into existing data processing pipelines. The results reveal whether the sequencing experiment has succeeded, whether the correct sample was sequenced and whether the sample contains any contamination from DNA preparation or host. It provides a computational resource which is capable to handle the amount of data that next generation sequencers are generating [12].

H. *Cutadapt*

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequences from the raw exome sequence. Cutadapt helps with the trimming task by finding the adapter or primer sequences in an error-

tolerant way. It can also modify and filter reads in various ways. Cutadapt also de-multiplex the input data, without removing adapter sequences at all [15].

I. *FASTQC*

FASTQC is used to clean FASTQ files and reports a wide range of information related to the quality profile of the reads. FastQC also assesses GC content, over-abundance of adaptors and over-represented sequence, from which an indication of PCR duplication rate may be inferred [15]. FastQC comes with two modes, one is standalone and the other is command line mode. The standalone interactive application is used for the immediate analysis of small numbers of FastQ files. The command line mode can be used when it would be suitable to integrate a larger analysis pipeline for the systematic processing of large numbers of files [15].

III. COMPARISON OF SEQUENCE CLEANING TOOLS

The method of predicting the disease is changing due to the accurate and low cost technology available in the field of biology to capture the biological sequences. The new technology is capturing DNA, RNA or protein sequences of living being and generating an enormous amount of high quality data. The field of biology can analyse the captured sequences but unable handle the large amount of complex data. The enormous data generated by the biological experiments are managed by the advance technology of computer science. The data of produced sequences are stored in the high speed and high capacity data storage devices which implements sophisticated algorithms to store and retrieve data efficiently.

The stored sequences are a great source of information for further investigation and research. The reason behind the success of next generation sequencing is the analysis of biological sequences that help in the identification of rare and novel diseases for which traditional disease identification methods fail. To carry out the analysis, first important step is to capture accurate sequence which is performed very well by high end instruments. Though the captured sequence is accurate, it needs to be passed from cleaning process. The cleaning process is necessary because the captured sequences contain some header information in the form of tags which needs to be removed for further analysis. For long reads, it is possible that the sequence captured contain some genotype error or noise which can affect the final result in the disease prediction. The cleaned sequence is capable of producing accurate results of the analysis. Cleaning a sequence is an important step, different algorithms are available to clean the sequences and new algorithms are emerging. The algorithms are implemented in easy to use format.

The computational tools available for cleaning the biological sequences are both licensed and open source. The comparison table [Table-1] presented considers only open source tools based on various parameters. The first benefit of using open source tools is it is available freely. Second and most important benefit is the availability of the source code, so the functionality can be altered as per the requirement. The tools are compared considering different parameters like operating system, input format, output format, platform, whether the tool provides functionality of tag removal, filtering and trimming.

The parameter Operating system indicates the operating systems on which the given tool can be installed and executed. As the tools selected are open source, all the tools are compatible with Linux. Few tools like SolexaQA runs on both Linux and Mac where as the tools like Trimmomatic, PRINSEQ, Cutadapt and FASTQC works for all types of operating systems. The input and output parameters indicates the file type supported for by the tools to perform input and output operations. FASTQ file format is the standard file format to store the biological sequences whereas SolexaQA, PRINSEQ, Cutadapt and FASTQC support other formats too.

The platform parameter indicates that the sequences generated using the listed platforms can be used for processing. Generally, the sequences produced by NGS are using illumina platform. All the tools listed here supports illumina platform. The next parameter, summary report indicates whether the tool is used to produce summary report or not. Summary report is very important part of any tool it is used to show the summary of the process done and the final output. Trimmomatic, TrimGlore, SolexaQA, PRINSEQ, Cutadapt and FASTQC are used to provide summary report. Tag removal is the process of eliminating additional known or unknown tag sequences from the raw sequence.

Trimmomatic and Cutadapt perform the process of tag removal. The process of filtering is used to remove genotyping errors from the raw sequence. Trimmomatic, TrimGlore, Skewer and PRINSEQ are capable to filter errors from the given sequence. The process of trimming aims to remove the reads that do not match the provided standard. All the tools compared in Table - 1 trims the raw sequence according to given standard, except one i.e. FASQC. Programming language parameter shows that the tool is developed using which language. PE and SE parameters give the idea about paired end and single end sequences are processed or not.

[Table – 1 Comparison of different open source trimming tools][OS = Operating System, I/P = Input, O/P = Output, PE = Paired End, SE = Single End]

	Parameters										
Tools	**OS**	**I/P**	**O/P**	**Platforms**	**Summary Report**	**Tag removal**	**Filtering**	**Trimming**	**Programming Language**	**PE**	**SE**
Trimmomatic [3, 4]	Lin, Mac, Win	FASTQ, FASTA	FASTQ, FASTA	Illumina	Yes	Yes	Yes	Yes	Java	Yes	Yes
TrimGlore [5]	Unix/Linux	FASTQ	FASTQ	Illumina	Yes	No	Yes	Yes	Perl	Yes	Yes
AlienTrimmer [6, 7]	Unix/Linux	FASTQ	FASTQ	Illumina	No	No	No	Yes	Java	Yes	Yes
Skewer[8, 1, 9]	Unix/Linux	FASTQ	FASTQ	Illumina	No	No	Yes	Yes	C++	Yes	Yes
Seqpurge[9]	Unix/Linux, Win	FASTQ	FASTQ	Illumina	No	No	No	Yes	C++	Yes	No
SolexaQA [11]	Lin, Mac	FASTQ	FASTQ, PNG	Illumina	Yes	No	No	Yes	Perl	Yes	Yes
PRINSEQ [12]	Lin, Mac, Win, Web interface	FASTA, FASTQ, QUAL FASTA	FASTA, FASTQ, QUAL FASTA, HTML	Illumina, 454	Yes	No	Yes	Yes	Perl	Yes	No
Cutadapt [13]	Lin, Mac, Win	FASTA, FASTQ	FASTA, FASTQ	454, Illumina, SOLID	Yes	Yes	No	Yes	Python	Yes	Yes
FASTQC [15]	Lin, Mac, Win	(CS) FASTQ, SAM, BAM	HTML	Illumina, ABI SOLiD	Yes	No	No	No	Java	Yes	Yes

IV. CONCLUSION

The biological sequences have been widely adopted for clinical diagnosis of rare and novel diseases. The advancement in computational technologies leads to capture the biological sequences easily and cost effectively. The analysis of such sequences is becoming very crucial. Before analysing the raw sequence, it must be processed to clean. Different algorithms are being developed and revised to clean the captured sequences which are implemented in the form of adaptable tools.

A researcher can select any of the tools to perform experiments. Each tool has its own characteristics so according to the requirement, a tool can be selected.

REFERENCES

[1]https://twistbioscience.com/company/blog/twistbioscienceexomesequencing4dataanalysis
[2] White Paper: Exome Sequencing and Data Analysis. Scigenom.com
[3] Trimmomatic Manual: V0.32.
[4] Anthony M. Bolger, Marc Lohse and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. Vol. 30 no. 15 2014, pages 2114–2120.
[5] Taking appropriate QC measures for RRBS-type or other -Seq applications with Trim Galore!. Babraham Bioinformatics. September 03, 2013.
[6] Alexis Criscuolo and Sylvain Brisse. ALIENTRIMMER: A tool to quickly and accurately trim off multiple short
contaminant sequences from high-throughput sequencing reads Frontiers in Genetics. Vol. 5, Article 130, May 2014.
[7] Alexis Criscuolo. AlienTrimmer User Guide. [Version 0.2.1] *September 2012*

[8]Hongshan Jiang, Rong Lei, Shou-Wei Ding and Shuifang Zhu. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads.BMC Bioinformatics 2014, 15:182.
[9] Marc Sturm, Christopher Schroeder and Peter Bauer. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. Sturm et al. BMC Bioinformatics (2016) 17:208.
[10] https://genomics.sschmeier.com/ngs-qc/index.html
[11] Murray P Cox, Daniel A Peterson, Patrick J Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. Cox et al. BMC Bioinformatics 2010, 11:485
[12] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. Vol. 27 no. 6 2011, pages 863–864.
[13] Marcel Martin. cutadapt Documentation Release 1.16. Feb 21, 2018.
[14] Richard M. Leggett, Ricardo H. Ramirez-Gonzalez, Bernardo J. Clavijo, DarrenWaiteand Robert P. Davey. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Frontiers in Genetics. December 2013, Volume4, Article288.
[15] FASTQC Manual
[16]https://www.reddit.com/r/bioinformatics/comments/63nu1f/comparing_quality_trimming_and_adapter_removing/ [25 May, 2018]
[17] https://cutadapt.readthedocs.io/en/stable/ [25 May, 2018]
[18] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 2. 256-278, 21 January 2013.
[19] Hongshan Jiang. Skewer: A fast and accurate adapter trimmer for paired-end reads: User's Manual. Chinese Academy of Inspection and Quarantine May 12, 2015.

Authors Profile

Ms. K.S.Mehta pursued Bachelor of Computer Application from MK Bhavnagar University, Gujarat,India in the year 2004 and Master of Computer Application from the same university in the year 2007. She is currently pursuing Ph.D. from Indus University, Ahmedabad,Gujarat,India and working as Assistant Professor in Faculty of Computer Technology, GLS University, Ahmedabad, Gujarat, India. Her main research work focuses on the knowledge extraction from exome sequences using computational techniques and big data. She has 2 years of corporate experience, 8 years of teaching experience and 3 years of research experience.

Dr. Devarshi Mehta, pursed Bachelor of Science(Physics) from Gujarat University in 1995 and Master of Computer Application from Bhavnagar University in year 1998. She is in the field of teaching since 2000 and currently working as Associate Professor in Faculty of Computer Technology, GLS University from 2010. She also posses guide-ship for Ph.D. students in various universities like, Gujarat Technological University(GTU), Indus University and GLS University. She has published more than 18 research papers in international journals and more than 10 in national & International conferences. Her main research work focuses on Bioinformatics, Data Mining, Telemedicine and Information Extraction and Management. She has 18 years of teaching experience, 2 years of programming experience and 12 years of Research Experience.

Dr. V. Dahiya did Master of Computer Application from MDU Rohtak, Haryana in 2001 and Ph.D. from Sardar Patel University, VV Nagar, Gujarat in 2013. She is currently working as an Associate Professor in IICT, Indus University since 2010. She has published more than 30 research paper in referred journals. Her main area of research is Image Processing, Data Mining and Big data analytics. She has 18 years of teaching and 10 years of research experience.

