Gen Aligner: Graphical user interface for reference based genomic data analysis

[1]Rakhee Raghoji, [2]S. Sahabudeen, [3]Shashi Kumar

[1,2]Department of Biotechnology Dayananda Sagar College of Engineering Bangalore, India
[3]GenEclat Technologies Bangalore, India

Abstract-Next-generation sequencing technologies are accelerating the biological research in areas of molecular biology, genetics, transcriptomics. The genome assembly is difficult as it is impossible to sequence a whole genome directly in one read using current sequencing technologies.The process of reconstructing a whole genome by joining these reads together up to the chromosomal level is genome assembly. In reference-based assembly, a reference genome from the same organism or a closely related species is used as a map to guide the assembly process by aligning the fragments being assembled. Different assemblers like BWA, bowtie, tophat are used for reference based sequence analysis, but all these tools don't have a graphical user interface. Operating of these command based tool will be difficult for users with less computer knowledge. Thus GenAligner, a platform independent tool is developed to make the operations of these assemblers accessible to users. This graphical user interface tool will integrates the functionality like checking the quality, trimming, mapping and visualization of the input genome data.

Keywords: next generation sequencing, genome assembly, reference based assembly, graphical user interface.

I. INTRODUCTION

Next-generation sequencing technologies (NGS) has revolutionized biology, as they reduced costs and increased the speed of data sequencing [1]. These sequencers are accelerating biological research in many areas such as genomics, transcriptomics, metagenomics, proteogenomics, gene expression analysis, noncoding RNA discovery, SNP detection, and the identification of protein binding sites [2]. But data analysis represents the bottleneck in their application. Users need to be familiar with computer terminal commands, the Linux environment, and various software tools and scripts. Analysis workflows have to be optimized and experimentally validated to extract biologically meaningful data [3].

The genome assembly is not easy assequence a whole genome directly in one read using current sequencing technologies is not possible.

Two types of genome assembly are reference-based assembly and the de novo assembly. In reference-based assembly [4], a reference genome from the same organism or a closely related species is used as a map to guide the assembly process by aligning the fragments being assembled where as in de novo assembly [4], no map or guidance is available for assembling the genome, so this approach represents assembly in the strict sense. So it is used to reconstruct genomes that are not similar to previously sequenced genomes. In this tool, reference based assembly approach is used wheresequence reads is aligned to a reference data using assembler bowtie. Different assemblers available are BWA [5], tophat[6], but these tools operates on linux environment where as bowtie [7] can operate on any operating system so platform independency is achieved.

The graphical user interface tools are available but these facilitate only mapping but GenAligner tool facilitates quality checking, trimming, mapping and visualization of genome data.

II. METHODOLOGY

A. Graphical user interface tool

All the tools used for assembly run through lengthy command lines composed of one or several parameters that influence the assembly results, which can be difficult for users with little computing experience [1]. So graphical user interface [8] tool, GenAligner: Graphical user interface for reference based genomic data analysis is developed. This graphical tool incorporates functionalities like building index file for reference genome data, checking the quality of target genome data which is to be aligned with reference index file, trimming the genome data and mapping the genome dataonto the reference genome data and visualization of data generated as shown in figure 1.

B. Implementation

GenAligner's Graphical User Interface (GUI) can be implemented and tested on windows and can run on any operating system since java program is used to achieve platform independency. The GUI was written in Java with help of eclipse Integrated Development Environment which has Java swing integrated in it and Java development toolkit (JDK).

C. Steps involved in genomic data analysis

1. Building reference index

The FASTA file of reference genome data is taken as input and its index files are built by executing bowtie2-build command. The user is provided with option to give a name for index file.

2. Preprocessing genome data

2.1 Check the quality of genome data

The FASTQ file of genome data is downloaded from Sequence Read Archive database. The quality of this genome data is checked using FastQC tool. FastQC report for genome data will be generated. If there are any abnormal sequences as indicated in report then it has be removed from FASTQ file.

2.2 Removing abnormal sequences from genome data

All the abnormal sequences present in FASTQ file are removed using Trimmomatic tool. After setting the parameter and performing the necessary operation it will produce .fastq output file.

The quality of .fastq file is checked to ensure that abnormal sequences have been removed.

3. Processing genome data

The output .fastq file produced by trimming is taken as input in processing step and it is mapped with the reference genome index file by bowtie2 tool. The bowtie tool will align the two files and generate a .sam file.

4. Visualization

The .sam (sequence alignment map) file generated from mapping of reference index file and t.fastq file is visualized using IGV (Integrative Genome Viewer) [10].

III. RESULTS

The figure2shows the front view of GenAligner tool. The view includes the menu bar and different buttons for performing different actions. The different buttons are created like import data to import reference genome file, Fastqc button to check the quality of genome FASTQ data, trimming button to remove abnormal sequences from the FASTQ data, mapping button to align the genome data onto the reference genome data and view button to visualize the output data.

First, the reference genome data is selected as shown in Figure 3. Theuser can either import reference genome data using browse button by specifying index file name, the index files will be built using bowtie2 tool or select the reference genome datawhose index file are already created like human genome, mouse, Campylobacterjejuni, klebsiella pneumonia. Here we have taken campylobacter jejuni for validating the tool.

The fastq file of Campylobacter jejuni, SRR3350893.fastq is downloaded from Sequence Read Archive (SRA) and its quality is checked byFastQC tool and the result is displayed as in Figure 4. The result shows that per base sequence quality status is Fail. It indicates that abnormal sequences are present in file. To remove these sequences trimmomatic tool is used and we set parmeter for trimming (removing) the abnormal sequence as shown in Figure 5 and execute trimmomatic tool, it will generate a SRR3350893_trim.fastq. All the abnormal sequences are removed from genome file. We check the quality of this trimmed file as show in Figure 6. The results shows that per base sequence quality status is pass indicating that all abnormal sequences are removed.

The outputSRR3350893_trim.fastq file obtained from trimming is taken as input for mapping by browsing it as show in figure 7. This target fastq file is aligned to reference genome index file that was selected from list of reference index in Figure 3. By clicking on Map button, the target genome and reference genome file will be mapped using bowtie2 tool and the result of alignment will be generated in is SRR3350893_trim.sam (sequence alignment map).

The SRR3350893_trim.sam file generated from mapping is taken as input for visualization and this file contains the sequence reads, which will be sorted by using sort option as in figure 8. This will generate SRR3350893_trim.sorted.sam file and index file will be created for this sorted file using index option. SRR3350893_trim.sorted.sam.sai will be generated by executing index command. The result will be visualized by clicking IGV button which opens IGV tool shown in fig 9.

Figure 9 shows the mapping of SRR3350893_trim.sorted.sam file onto the Campylobacter jejuni genome.the mapped sequences are represented grey color and unmapped regions are represented by while blank space. For each sequence, its name, its location, position where the alignment has started for a that sequence will be showed as in figure 9.

A. Figures

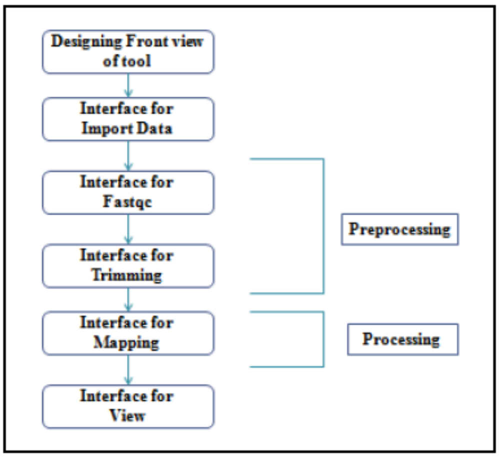


Figure1: workflow of GenAligner tool

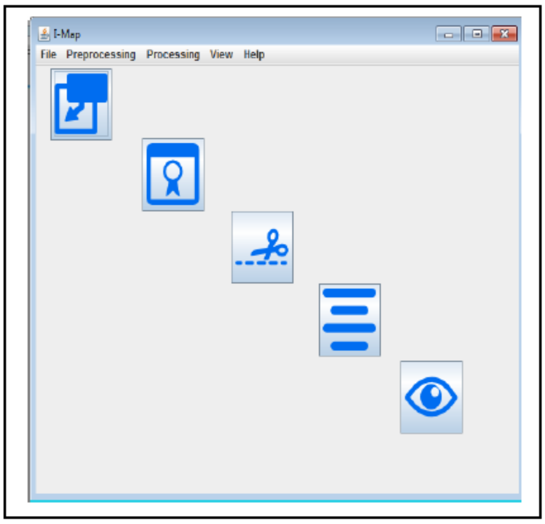


Figure 2: Front view of GenAligner tool.

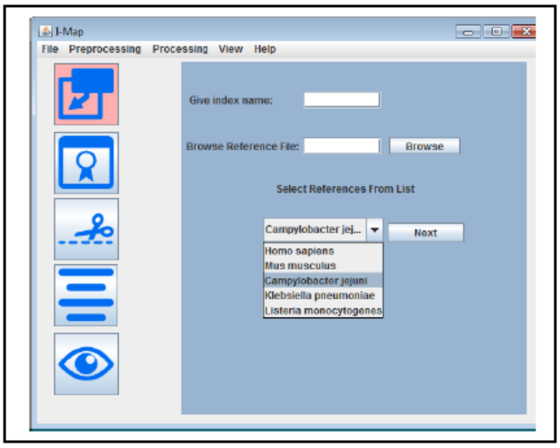
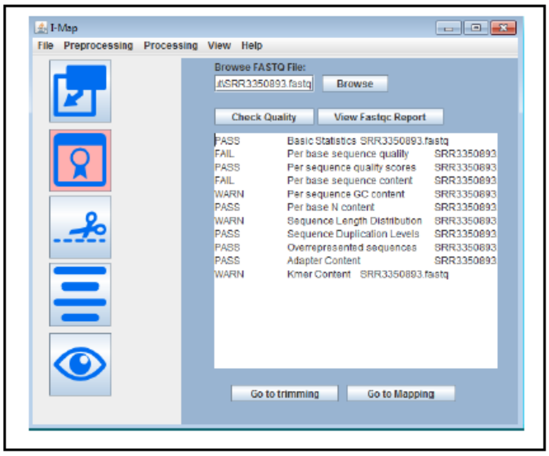


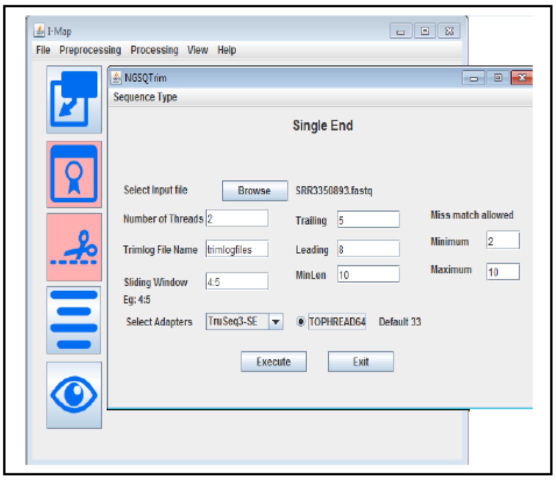
Figure 3: Selecting Reference genome data.



Browse FASTQ File: .t\SRR3350893.fastq

```
PASS        Basic Statistics SRR3350893.fastq
FAIL        Per base sequence quality       SRR3350893
PASS        Per sequence quality scores     SRR3350893
FAIL        Per base sequence content       SRR3350893
WARN        Per sequence GC content         SRR3350893
PASS        Per base N content              SRR3350893
WARN        Sequence Length Distribution    SRR3350893
PASS        Sequence Duplication Levels     SRR3350893
PASS        Overrepresented sequences       SRR3350893
PASS        Adapter Content                 SRR3350893
WARN        Kmer Content   SRR3350893.fastq
```

Figure 4: FastQc result for SRR3350893.fastq.



Single End

Select Input file: SRR3350893.fastq

Number of Threads: 2

Trimlog File Name: trimlogfiles

Sliding Window: 4:5

Eg: 4:5

Select Adapters: TruSeq3-SE

● TOPHREAD64 Default 33

Trailing: 5

Leading: 8

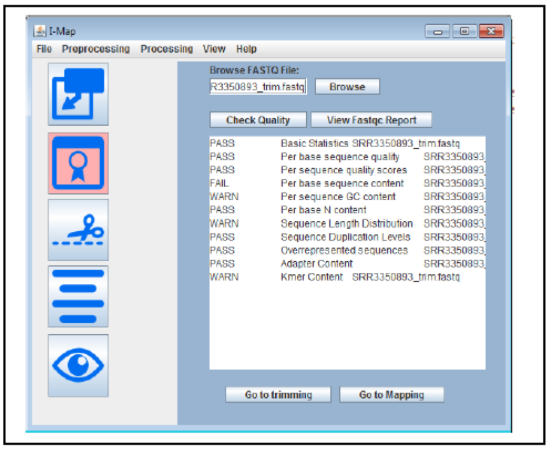
MinLen: 10

Miss match allowed

Minimum: 2

Maximum: 10

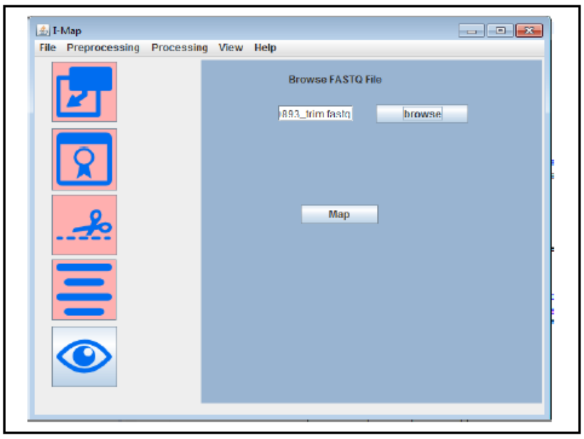
Figure 5: Setting parameters for Trimmomatic.



Browse FASTQ File: R3350893_trim.fastq

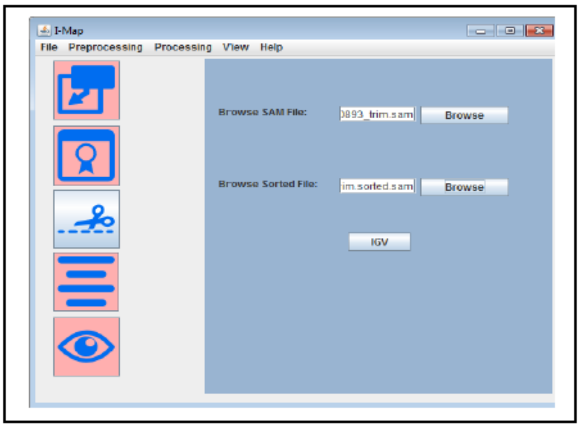
```
PASS        Basic Statistics SRR3350893_trim.fastq
PASS        Per base sequence quality       SRR3350893_
PASS        Per sequence quality scores     SRR3350893_
FAIL        Per base sequence content       SRR3350893_
WARN        Per sequence GC content         SRR3350893_
PASS        Per base N content              SRR3350893_
WARN        Sequence Length Distribution    SRR3350893_
PASS        Sequence Duplication Levels     SRR3350893_
PASS        Overrepresented sequences       SRR3350893_
PASS        Adapter Content                 SRR3350893_
WARN        Kmer Content   SRR3350893_trim.fastq
```

Figure 6: FastQC result for SRR3350893_trim.fastq.



Browse FASTQ File: 893_trim.fastq

Figure 7: Mapping of SRR3350893_trim.fastq on campylobacter_jejuni reference genome.



Browse SAM File: 0893_trim.sam

Browse Sorted File: im.sorted.sam

Figure 8: Sorting and indexing of SRR3350893_trim.sam

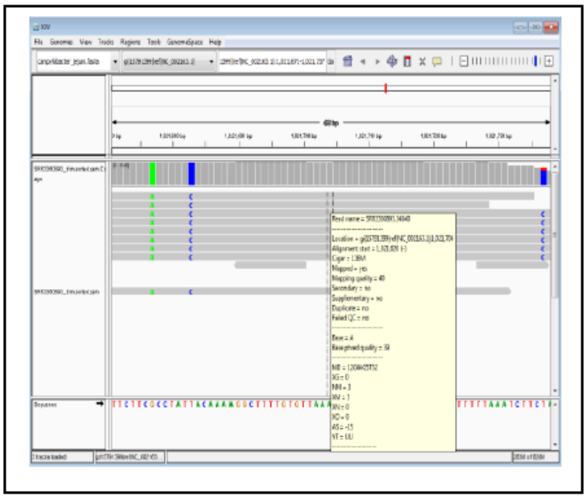


Figure 9: visualization of SRR3350893_trim.sorted.sam alignment.

IV. CONCLUSION

Aplatform independent tool is developed to make the operations of command based assembler and tool accessible to users. This graphical user interface tool will integrates the functionality like checking the quality, trimming, mapping and visualization of the input genome data. Thus GenAligner tool will be helpful to both biologists with less computational knowledge and bioinformaticians as a simple, timesaving tool for mapping of the genome sample data onto reference data and visualization of generated sam file.

REFERENCES

[1] F. Sciences, A. Veras, P. Sá, V. Azevedo, A. Silva, and R. Ramos, "AutoAssemblyD: a graphical user interface system for several genome assemblers.,"Bioinformation, vol. 9, no. 16, pp. 840–1, 2013.

[2] F. Torri, I. Dinov, A. Zamanyan, S. Hobel, A. Genco, P. Petrosyan, A. Clark, Z. Liu, P. Eggert, J. Pierce, J. Knowles, J. Ames, C. Kesselman, A. Toga, S. Potkin, M. Vawter, and F. Macciardi, "Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows,"Genes, vol. 3, no. 3, pp. 545–575, 2012.

[3] D. Velmeshev, P. Lally, M. Magistri, and M. Faghihi, "CANEapp: a user-friendly application for automated next generation transcriptomic data analysis," Bmc Genomics, vol. 17, no. 1, p. 49, 2016.

[4] S. El-Metwally, T. Hamza, M. Zakaria, and M. Helmy, "Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges," PlosComputBiol, vol. 9, no. 12, p. e1003345, 2013.

[5] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows–Wheeler transform,"MethodBiochem Anal, vol. 26, no. 5, pp. 589–595, 2010.

[6] D. Kim and S. Salzberg, "TopHat-Fusion: an algorithm for discovery of novel fusiontranscripts,"GenomeBiol, vol. 12, no. 8, p. R72, 2011.

[7] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," Genome Biol, vol. 10, no. 3, p. R25, 2009.

[8] Y. Ohtsubo, W. Ikeda-Ohtsubo, Y. Nagata, and M. Tsuda, "GenomeMatcher: A graphical user interface for DNA sequence comparison," Bmc Bioinformatics, vol. 9, no. 1, p. 376, 2008.

[9] A. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data,"MethodBiochem Anal, vol. 30, no. 15, p. btu170, 2014.

[10] J. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. Lander, G. Getz, and J. Mesirov, "Integrative genomics viewer," Nat Biotechnol, vol. 29, no. 1, pp. 24–26, 2011.

◈◈◈