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

User Interface for Metagenomics Data Analysis

Priyanka Patil[1], Naveen Kumar N [2], Dr. Shinomol George K [3]
[1] Department of Biotechnology, Dayananda Sagar College of Engineering,
[2] GenEclat Technologies, Bangalore,
[3]Department of Biotechnology, Dayananda Sagar College of Engineering,

Abstract: Metagenomics is the direct genetic analysis of genomes contained with an environmental sample. Metagenomics applies a suite of genomic technologies and bioinformatics tools to directly access the genetic content of entire communities of organisms. The field of metagenomics has been responsible for substantial advances in microbial ecology, evolution, and diversity over the past 5 to 10 years, and many research laboratories are actively engaged in it now. It provides access to the functional gene composition of microbial communities and thus gives a much broader description than phylogenetic surveys, which are often based only on the diversity of one gene. Performing metagenomics analysis is a tedious work using command prompt as it needs more computational knowledge. So it is better to have a GUI to perform this work in simple steps for sequence based metagenomics data analysis. This UI helps in measuring the concentration of the microbes like bacteria and fungus in water, soil and in clinical research. This User Interface is developed using java, which is platform independent.
Keywords: Metagenomics, Microbes, Clinical research, Microbial ecology, Data analysis.

I. INTRODUCTION

The advent of Next-Generation Sequencing (NGS) or high-throughput sequencing has revolutionized the field of microbial ecology and brought classical environmental studies to another level. This type of cutting-edge technology has led to the establishment of the field of "metagenomics", defined as the direct genetic analysis of genomes contained within an environmental sample without the prior need for cultivating clonal cultures. Initially, the term was only used for functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample, but currently it is also widely applied to studies performing polymerase chain reaction (PCR) amplification of certain genes of interest. The former can be referred to as "full shotgun metagenomics", and the latter as "marker gene amplification metagenomics" (ie, 16S ribosomal RNA gene) or "meta-genetics"[1].

Shotgun metagenomic DNA sequencing is a relatively new and powerful environmental sequencing approach that provides insight into community, biodiversity and function [5]. But, the analysis of metagenomic sequences is complicated due to the complex structure of the data. Fortunately, new tools and data resources have been developed to determine which microbes are present in the community and what they might be doing [7].

More generally, in the field of metagenomics there are few real datasets for which the correct results are known [5]. Recent metagenomic studies of global algal distribution and human microbiome have derived results that conflict from previous studies in the same environments [8]. The tools which are used for shotgun metagenomics are command based tools so we create a graphical user interface for preprocessing and sequence reconstruction of the shotgun metagenomics data.

II. METHODOLOGY

Next Generation Sequencing data is taken for the Shotgun metagenomic analysis. Performing metagenomics analysis is a tedious work using command prompt as it needs more computational

knowledge. So it is better to have a GUI tool to perform this work in simple steps for sequence based metagenomics data analysis. This user interface is developed using java, which is platform independent.

Here we perform two steps, one is Preprocessing [1] and another one is Sequence reconstruction and grouping [1]. In preprocessing there are three levels those are Quality check, screening and De-replication, we use different tool in different levels. In sequence reconstruction there are two levels one is Assembly and another one is Binning. Using these two steps here we developed user interface.

2.1 Preprocessing

2.1.1 Quality Check

In preprocessing first level is Quality Check, here we check the quality of the fastq files by using FastQC tool [1]. FASTQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which we can use to give a quick impression of whether your data has any problems of which we should be aware before doing any further analysis.

2.1.2 Screening

After checking quality of the data we should perform screening, in screening we remove the contamination present in the sequence by using Trimmomatic tool [9]. Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.

2.1.3 De-Replication

Next in De-replication We remove the duplicate reads by comparing the two sequence reads by using FastUniq tool [6]. It is a Fast De Novo Duplicates Removal tool for paired short reads. FastUniq identifies duplicates by comparing sequences between read pairs and does not require complete genome sequences as prerequisites [10].

2.2 Sequence Reconstruction and Grouping

2.2.1 Assembly

Next step is the sequence reconstruction and grouping, in this first we performed the aligning of the sequences i.e Assembly. The MetaVelvet tool [4] is used for the aligning the read sequences. A single-genome assembly [2] program (assembler) is not capable of resolving metagenome sequences, so assemblers designed specifically for metagenomics have been developed. MetaVelvet is an extension of the single-genome assembler Velvet.

2.2.2 Binning

After alignment next step is grouping of the sequences i.e. Binning by using BiMeta2 tool [3]. Binning, attempts to assign every metagenomic sequence to a taxonomic group. The algorithm consists of two phases. In the first phase of the algorithm, reads are grouped into groups based on overlap information between the reads. The second phase merges the groups by using an observation on *l*-mer frequency distribution of sets of non-overlapping reads [2].

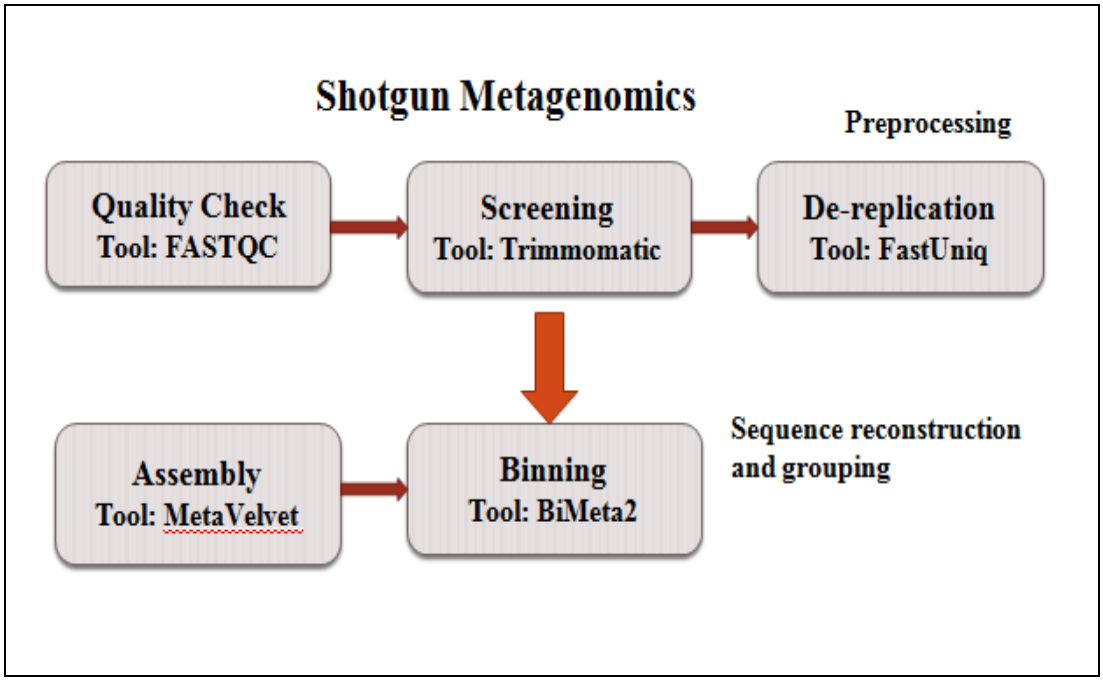

Figure 1: Workflow of Metagenomics Data Analysis

III. RESULTS

The Fig 2 shows the user intercae for MetSeq tool. Here we are dealing with De-novo sequencing so we are taking paired end data for the sequence analysis. First we should select the 2 fastq files by clicking on the browse button. Perform quality check by clicking on the QC button for the fastq files selected by using FastQC tool. The QC report will generated, it consist of many modules. We should check for the Per sequence base quality, it should be in a good condition, if it is in bad condition we will perform screening.

Based on the QC report we set the parameters for trimming. If the sequences are not present in the very good region we will trim by clicking on screening button using Trimmomatic tool. Here we screen the contaminated read sequences present in the fastq files. After trimming generated files are in .trimmed.fq format and per sequence based quality is in good condition. Those are taken as input for the next level. Next is the De-replication here we remove the duplicates data present in the sequences by using FastUniq. The output generated in this level are taken as input for the next level. Perform de-replication by pressing De-replication button.

In Assembly level we take de-replicated files as input to perform Metavelvet. In MetaVelvet we run three commands to perform aligment. First is run the velveth command it generate the hash tables which are used to generate a graph by running velvetg command. Files generated after running both velveth and velvetg are taken as input for the execution of the meta-velvetg command.

It is used for the assembly purpose. The meta-vevelg.contig,fa file is taken as input for the binning level. Biining is performed by clicking on Binning button. In binning we group the sequence reads by using BiMeta2 tool. In binning we should create 2 separate folders one is input and one is output. In input folder input file is saved and in output folder the output files generated after performing binning will be created.

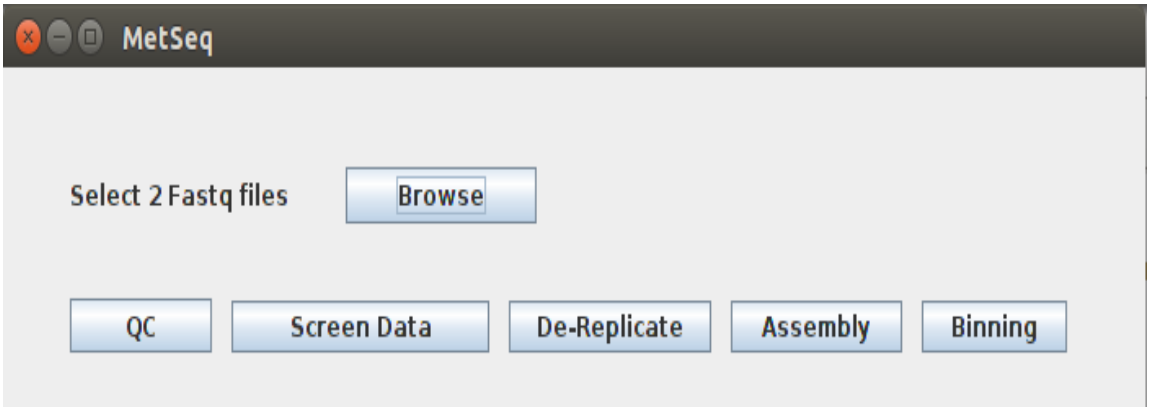

Figure 2: User Interface For Metagenomic Data Analysis

IV. CONCLUSION

User Interface for metagenomics data analysis is a platform independent and is developed to make the operations of command based assembler and easily accessible to users. This UI helps in measuring the concentration of the microbes like bacteria and fungus in water, soil and in clinical research. Here we perform quality check for the data, trimming of the unwanted data and screening of the data, alignment of the sequence and grouping of the data. It is also capable of handling huge data in a memory efficient way. This user interface will be helpful for the users with less computer knowledge.

REFERENCES

1. A. Oulas, C. Pavloudi, P. Polymenakou, Georgios A, N. Papanikolaou, G. Kotoulas et al, "Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies," *Bioinformatics and Biology Insights*, vol. 9 ,pp. 75-88, 2015
2. R. Dickson and G. Gloor, "Protein Sequence Alignment Analysis by Local Covariation: Coevolution Statistics Detect Benchmark Alignment Errors,"*Plos One*, vol. 7, no. 6, p. e37645, 2012.
3. L. Vinh, T. Lang, L. Binh, and T. Hoai, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithm Mol Biol*, vol. 10, no. 1, p. 2, 2015.
4. T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads," *Nucleic Acids Res*, vol. 40, no. 20, pp. e155–e155, 2012.
5. S. Johnson, B. Trost, J. Long, V. Pittet, and A. Kusalik, "A better sequence-read simulator program for metagenomics," *Bmc Bioinformatics*, vol. 15, no. Suppl 9, p. S14, 2014.
6. H. Xu, X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen, "FastUniq: a fast de novo duplicates removal tool for paired short reads.," *Plos One*, vol. 7, no. 12, p. e52249, 2012.
7. T. Sharpton, "An introduction to the analysis of shotgun metagenomic data," *Front Plant Sci*, vol. 5, p. 209, 2014.
8. M. Kim, K.-H. Lee, S.-W. Yoon, B.-S. Kim, J. Chun, and H. Yi, "Analytical tools and databases for metagenomics in the next-generation sequencing era.," *Genom Informatics*, vol. 11, no. 3, pp. 102–13, 2013.
9. A. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Method Biochem Anal*, vol. 30, no. 15, p. btu170, 2014.
10. A. Mitchell, F. Bucchini, G. Cochrane, H. Denise, P. Hoopen, M. Fraser, S. Pesseat, S. Potter, M. Scheremetjew, P. Sterk, and R. Finn, "EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data," *Nucleic Acids Res*, vol. 44, no. D1, pp. D595–D603, 2016.

