A Project Report

On

**Pre-processing of NGS data**

BY

**SARTHAK SOMANI**

**2021B1A70793H**

Under the supervision of

**JAYAPRAKASH K S**

**SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS OF**

**BIO F266: STUDY PROJECT**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)**

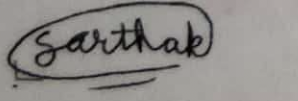
**HYDERABAD CAMPUS**

**(MARCH 2023**)

**ACKNOWLEDGMENTS**

**I would like to express my special thanks to my mentor Jayaprakash sir for his time and efforts he provided throughout the project. Your useful advice and suggestions were really helpful to me during the project’s completion. In this aspect, I am eternally grateful to you.**

**I would like to acknowledge that this project was completed entirely by me.**



**Signature**

**Sarthak Somani**

****

**Birla Institute of Technology and Science-Pilani,**

**Hyderabad Campus**

**Certificate**

This is to certify that the project report entitled “**Pre-processing of NGS data”** submitted by Mr. SARTHAK SOMANI (ID No. 2021B1A70793H) in partial fulfillment of the requirements of the course BIO F266, Study Project Course, embodies the work done by him under my supervision and guidance.

**Date: (JAYAPRAKASH K S)**

BITS- Pilani, Hyderabad Campus

**ABSTRACT**

**The project is about processing of Next Generation Sequencing data. It is taken from an article already published in a journal. The article talks about designing of effective primers for 8 genes of influenza virus so that each gene present in different strains is expressed in equal quantity and error like amplication of gene from an empty droplets is also reduced. Two strains of influenza virus H3N2 and H1N1 are taken. It starts with learning the basics of NGS and UNIX commands from a website followed by quality control with help of FastQC software. Quality control involves interpretation of graphs and report generated from FastQC software. Next step is mapping of data. STAR aligner was used for this. GTF files were made from fasta files of reference genome using plastid package and converted into ensemble format using pygtftk package. Genome index was generated with help of STAR which was used for mapping paired end reads. Data in log files (output of aligner) was used to make graph of no. of reads v/s gene name so as to check if each gene in different strains are produced in same quantity. Bulk sequencing is completed before midsem. After midsem single cell sequencing and downstream analysis will be taken into account.**

**CONTENTS**

Title page……………………………………………………………….1

Acknowledgements……………………………………………………..2

Certificate……………………………………………………………....3

Abstract………………………………………………………………....4

1.Basics of NGS……………………………………………….…………....6

2.Quality Control………………………………………………………...7

3.Mapping…………………………………………………………………8

Conclusion……………………………………………………………………..9

References……………………………………………………………………..10

Appendix………………………………………………………………………11

# **Chapter 1 : Basics of NGS**

## NGS data is generated with help of Illumina Sequencer in our case. A library is made with help of primers. The above process involves 2 cycles of addition of primers and then breaking the whole segment into small segments. Illumina Sequencer can sequence a read of maximum length of 300 bp.

## The small sequences are amplified using bridge amplification. Adapters are added to small sequence on both sides. There are fluorescent dNTPs which could generate a signal as process of base pairing is done. After data is generated we get a processed file in fastq format. This are the read files.

## 

## 

## Fastq file has a four line format. First line contains SeqID, second line “+”, third line sequence, fourth line quality score. Quality score is calculated from the formula Q=-log10 P where P is probability of incorrect base pairing.

## Next step is the quality control step in which we check the quality of our read file. Once this step is validated the next step is mapping step for which we have to generate GTF files so as to generate genome indexes and then finally SAM/BAM files as output of mapping step.After mapping downstream analysis is carried out which depends on the purpose for which we want to use the data.

# **Chapter 2 : Quality Control Step**

## In this step we provide fastq file as input to FastQC software which gives us a report as a output. This report includes various graphs, filename, filetype, encoding, total sequences, sequence length. Per base sequence quality graph gives us information about quality score of bases which is average for all bases in different read at that position. Per tile sequence quality graph show average quality score for each tile. The colors are on cold to hot scale. A graph with hotter colors indicate tile having worst quality scores. Per sequence quality scores graph show sequences of a particular quality against the no. of sequences having same quality. Per base sequence content graph shows proportion of each base in a file. A file having equal proportion of all bases is recommended for sequencing as it increases efficiency and accuracy. Per sequence N content graph shows N content across all bases. It presence indicate that we have sequences of shorter length than expected. Sequence length distribution graph shows number of sequences of a particular length against length. Sequence duplication levels graph show different degree of duplication of relative number of sequences. Overrepresented sequences graph shows sequences which are presented more than 0.1% of total reads. Adaptor content graph shows content of adaptor in our sequences which should be zero in ideal case. If it is present then adaptor trimming is done.

## 

Name of fastq file

graph

title

Various parameters

# **Chapter 3 : Mapping**

## Once quality control test is passed by data it is used for mapping. In this step reads are aligned to reference genome. STAR aligner is used. Reference genome in form of fasta file, GTF file are provided as input for generating genome indexes which are further used with fastq files to generate a BAM file. BAM file is a large file, therefore it is interpreted through featureCounts package. GTF file can be downloaded from EMBL or generated through packages like plastid and pygtftk.

## After above step I got statstics of mapping:

## Uniquely mapped reads % | 80.07%

## % of reads mapped to multiple loci | 7.33%

## % of reads unmapped: too short | 4.87%

## % of reads unmapped: other | 7.73%

## 

## 

# **Conclusion**

## From above histogram we can conclude that primers for PB2,PB1\_comm2,HAa, HAb, NP\_comm, M\_comm are efficient as this genes are getting expressed in equal numbers in different strains of influenza.

# **References**

## <https://bioinformaticsworkbook.org/>

## <https://github.com/ONeillMB1/>

## <https://github.com/alexdobin/STAR/blob/master/>

# **Appendix**

## Code for generating genome indexes

## STAR --runMode genomeGenerate --runThread 4 --genomeDir /path to genome directory where index will be stored --genomeFastaFiles /path to reference genome in fasta format --genomeSAindexNbases 5 --sjdbGTFfile /path to GTF files

## Code for mapping

## STAR --runMode alignReads --genomeDir /path to genome directory where index is stored --readFilesIn /path to fastq files --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 30 --outSAMtype BAM SortedByCoordinate --limitBAMsortRAM 1010005903