

# **Quality control approaches for Metagenomic data analysis.**

A Thesis

Submitted by

**Gaurav Chetal (S15014)**

For the award of the degree of

**MS by Research**



**School of Basic Sciences**

**Indian Institute of Technology Mandi,**

**Himachal Pradesh, India**

**September 2017 (Revised Final Copy)**

## **Declaration by the Research Scholar**

This is to certify that the thesis titled “**Quality control approaches for Metagenomic data analysis**”, submitted by me to the Indian Institute of Technology Mandi, for the award of the degree of Master of Science (By Research) is a bonafide record of the research work carried out by me under the supervision of **Dr. Tulika Prakash Srivastava**. The content of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Place: Kamand

Signature of the Research Scholar

Date:

## **THESIS CERTIFICATE**

This is to certify that the thesis titled “**Quality control approaches for Metagenomic data analysis**”, submitted by **Gaurav Chetal**, to the Indian Institute of Technology Mandi, for the award of the degree of Master of Science (By Research), is a bonafide record of the research work done by him under my supervision. The content of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

**Dr. Tulika Prakash Srivastava**

Supervisor

Assistant Professor

School of Basic Sciences

IIT Mandi

Kamand-175005

## ACKNOWLEDGEMENT

First and foremost, I extend my sincere thanks to the Almighty God, for bestowing His cherished blessings upon me.

I would like to record gratitude to my guide Dr. Tulika Prakash Srivastava for her supervision, advice and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. I owe her a great deal of thanks for taking me under her guidance and allowing me to soak some of her knowledge and insight. She has not only made me to work but guided me to orient towards research.

I am extremely thankful to the School of Basic Sciences, India Institute of Technology Mandi, for its support and valuable suggestions.

I would also like to thank all my batch mates for their constant help and support throughout the research. And I am very grateful to my parents and my family who have been a constant support for me throughout my life.

Last but not the least, I dedicate this thesis work to my younger brother Late. Er. Mohit Chetal who always had trust in me and gave me the mental support of doing this work.

Gaurav Chetal

Date:

## Abstract

Since the beginning of the Human genome project on October 1, 1990 genomics has unraveled the fundamental make-up of a being i.e. complete set of DNA within a single cell of an organism. Many –omics branches such as metagenomics and transcriptomics have emerged since then and these –omics approaches have brought together a challenge of the quality control of the data that is generated during production of sequencing reads from different Next Generation Sequencing platforms. The main focus of current study was the analysis of the quality filtering measures in the genomic and metagenomic datasets. This work discusses that the adapter filtration analysis on metagenomic dataset and the comparison of the filtered data with raw dataset. A comparison was also done between the Ap1 immuno compromised mice dataset, a non immuno compromised mice and a reference human dataset. Then, two quality control tools i.e. PRINSEQ and FaQCs were compared and selected features of these tools were integrated into a pipeline. This pipeline was further tested on genomic and metagenomic datasets for validation of the pipeline.

# Table of Contents

Abstract		i
List of Figures		iv
List of Tables		viii
List of Abbreviations		ix
1	Literature review	1
1.1	Explanation of the features Selected for the Comparison of different QC tools.	3
2	Comparison of Raw and Adapter-filtered metagenomic datasets.	12
2.1	Introduction	12
2.2	Method	13
2.3	Results and Discussion	15
2.3.1	Analysis Statistics of the compared datasets.	15
2.3.2	DRISEE	19
2.3.3	Functional Category Hits Distribution	21
2.3.4	Taxonomic Hit Distribution	24
2.3.5	Best Hit Rarefaction	27
2.4	Conclusion	28
3	The three Dataset Comparison (Ap1, Mouse and Human dataset)	29
3.1	Introduction	29
3.2	Method	30
3.3	Results and Discussion	31
3.3.1	Analysis Statistics of the three datasets taken for comparison	31
3.3.2	DRISEE	37
3.3.3	Functional Category Hits Distribution	39
3.3.4	Taxonomic Hits Distribution	43
3.3.5	Rarefaction Analysis	48
3.3.6	Sequence GC Distribution	50
3.4	Conclusion	51
4	Comparison of PRINSEQ and FaQCs	52
4.1	Introduction	52
4.2	Method	53
4.3	Results and Discussion	54
4.3.1	GC Content Based Trimming	54
4.3.2	POLY A/T Tail Trimming	55
4.3.3	Adapter and phiX Trimming	56
4.3.4	Ambiguous Bases Trimming	57
4.3.5	Homopolymer Bases Trimming	59

4.4	Conclusion	60
5	Applying integrated pipeline on genomic and metagenomic datasets	61
5.1	Introduction	61
5.2	Method	62
5.3	Results and Discussion	62
5.3.1	Genomic Validation of the Pipeline	62
5.3.1.1	GC Content Based Trimming	63
5.3.1.2	POLY A/T Tail Trimming	64
5.3.1.3	Adapter and phiX Trimming	65
5.3.1.4	Ambiguous Bases Trimming	66
5.3.1.5	Homopolymer Bases Trimming	67
5.3.1.6	Quality Report	68
5.3.2	Metagenomic Validation of the Pipeline	69
5.3.2.1	GC Content Based Trimming	70
5.3.2.2	POLY A/T Tail Trimming	70
5.3.2.3	Adapter and phiX Trimming	71
5.3.2.4	Ambiguous Bases Trimming	73
5.3.2.5	Homopolymer Bases Trimming	74
5.3.2.6	Quality Report	75
5.4	Conclusion	76
6	Summary	77
7	References	79
8	Appendix	82