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 meto 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 meextraordinary experiences throughout the work. I owe her a great deal of thanks for taking meunder her guidance and allowing me to soak some of her knowledge and insight. She has notonly 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 throughoutthe research. And I my very grateful to my parents and my family whohave 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. MohitChetal 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