“Taxonomic Sensitivity and Precision of Hypervariable Regions in Bacterial 16S rRNA genes”

A DISSERTATION

SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OT

THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

MASTERS OF TECHNOLOGY
BIOMEDICAL ENGINEERING

SUBMITTED BY
MONIKA GUBRELE

2K17/BME/03

Under the supervision of

Prof. Jai Gopal Sharma

Department of Biotechnology

Delhi Technological University, DELHI - 110042

(Formerly Delhi College of Engineering)

CANDIDATE'S DECLARATION

I, Monika Gubrele, 2K17/BME/03 student of M.Tech Biomedical Engineering, hereby declare that the project Dissertation titled "**Taxonomic Sensitivity and Precision of Hypervariable Regions in Bacterial 16S rRNA genes"** which is submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without paper citation. The work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place:

Date:

MONIKA GUBRELE

2K17/BME/03

PARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

CERTIFICAE

I hereby certify that the Project Dissertation titled "**Taxonomic Sensitivity and Precision of Hypervariable Regions in Bacterial 16S rRNA genes**" which is submitted by Monika Gubrele, 2K17/BME/03, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part of full for any Degree or Diploma to this University or elsewhere.

Place:

Date:

Prof. Jai Gopal Sharma

Supervisor and HOD

DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

ABSTRACT

RNA is the genetic material of bacteria except for the exceptions and out of that too 16S rna of the microbes courses of action be comprehensively use into regular microbial study and atomic developments dependable pined up for the ordered characterization and evolutionary investigation of organisms. Constrained within the currents aligning techniques, enormous sequence aligning of 16S rRNA genetic material amplicons surrounding the complete length of genesis not until now realistic. As a result, high-through put studies of microbial communities often do not sequence the entire16S rRNA gene. The test is to acquire dependable portrayal of bacterial networks through taxonomic classification of short 16S rRNA gene sequences known as hyper variable regions. However, the assortment of the most emerald hypervariable regions meant for phylogenetic estimate and taxonomic kind continues to be argued. Species explicit groupings inside a given hypervariable district comprise helpful final target destined in favor of problem-solving assay and other science related searches. Also, nix on its own region be able to distinguish among all microbes, consequently, organized study that think about the overall preferred position of every locale for explicit demonstrative objectives be desired. Here, first present an into silico pipeline for generating the 9 Hypervariable Regions from 16S rRNA sequences, using conserved regions and primers. The pipeline includes an error parameter taking into consideration in sertions, deletions and substitutions at some positions within the sequences. These hyper variable regions are then used to do a comparative study on the taxonomic sensitivity and accuracy of each of the 9 hyper variable regions. Each of the hypervariable regions are assigned taxonomy using QIIME (Quantitative Insights In to Microbial Ecology) and there sultant OUT table is used to generate abundance data. This abundance data is then used to decide the best hyper variable region for prediction at each level of the taxonomic hierarchy. In our study, we found that V2 region is best suited to assign phylum whereas V4 is better suited to decide deeper into the taxonomy, such as order and family. Also, certain examples are shown of 3 specific cases where a HV region has a bias against/towards a particular taxon, which can allow us money-spinning searching of change in microbial set of connections configuration together with the exceptional biosphere larger than space plus instant as well as can be functional straight away to initiative, similar to the Human Micro biome Project

ACKNOWLEDGEMENT

At the time of submission of my M. Tech Major Project Synopsis, I would first like to thank GOD for giving me patience, strength, capability, and willpower to complete my work. Apart from our efforts, the success of this project depends largely on the encouragement and guidelines of many others. I, therefore, take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

My initial thank is addressed to my mentor Prof. Jai Gopal Sharma, Department of Biotechnology, Delhi Technological University, who gave me this opportunity to work in a project under him. It was his enigmatic supervision, constant encouragement and expert guidance which have enabled me to complete this work. I humbly seize this opportunity to express my gratitude to him.

I would like to extend my sincere gratitude to Dr. Rajkumar Bidlan for her keen observation and continuous advice. Her thoughtful inputs on the project issues have been invaluable for the productive progress of the project. I extend my thanks to technical staff Mr. Jitender Singh and Mr. C.B. Singh who had been an aid whenever required. Lastly, I wish to extend my thanks to my family and friends who have supported me through the entire process.

MONIKA GUBRELE

2K17/BME/3

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DGGE	Denaturing gradient gel electrophoresis
FISH	Fluorescent in situ hybridization
T-RFLP	Terminal restriction Fragment length polymorphism
SSU	Smaller Sub-Unit
LSU	Larger Sub-Unit
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
HV	Hyper-Variable Region
OUT	Operational Taxonomic Unit
QIIME	Quantitative Insights Into Microbial Ecology
PAST	PAleontological STatistics

CHAPTER - 1

INTRODUCTIO

INTRODUCTION

RNA is the genetic material of bacteria's except for the exceptions and out of that too 16S rna of the microbes courses of action be comprehensively use into regular microbial study and atomic developments dependable pined up for the ordered characterization and evolutionary investigation of organisms. Constrained within the currents aligning techniques, enormous sequence aligning of 16S rRNA genetic material amplicons surrounding the complete length of genesis not until now realistic. As a result, high-through put studies of microbial communities often do not sequence the entire16S rRNA gene. The test is to acquire dependable portrayal of bacterial networks through taxonomic classification of short 16S rRNA gene sequences known as hyper variable regions. However, the assortment of the most emerald hypervariable regions meant for phylogenetic estimate and taxonomic kind continues to be argued. Species explicit groupings inside a given hypervariable district comprise helpful final target destined in favor of problem-solving assay and other science related searches. Also, nix on its own region be able to distinguish among all microbes, consequently, organized study that think about the overall preferred position of every locale for explicit demonstrative objectives be desired. Here, first present an into silico pipeline for generating the 9 Hypervariable Regions from 16S rRNA sequences, using conserved regions and primers. The pipeline includes an error parameter taking into consideration in sertions, deletions and substitutions at some positions within the sequences. These hyper variable regions are then used to do a comparative study on the taxonomic sensitivity and accuracy of each of the 9 hyper variable regions. Each of the hypervariable regions are assigned taxonomy using QIIME (Quantitative Insights In to Microbial Ecology) and there sultant OUT table is used to generate abundance data. This abundance data is then used to decide the best hyper variable region for prediction at each level of the taxonomic hierarchy. In our study, we found that V2 region is best suited to assign phylum whereas V4 is better suited to decide deeper into the taxonomy, such as order and family. Also, certain examples are shown of 3 specific cases where a HV region has a bias against/towards a particular taxon, which can allow us money-spinning searching of change in microbial set of connections configuration together with the exceptional biosphere larger than space plus instant as well as can be functional straight away to initiative, similar to the Human Micro biome Project.

CHAPTER - 2

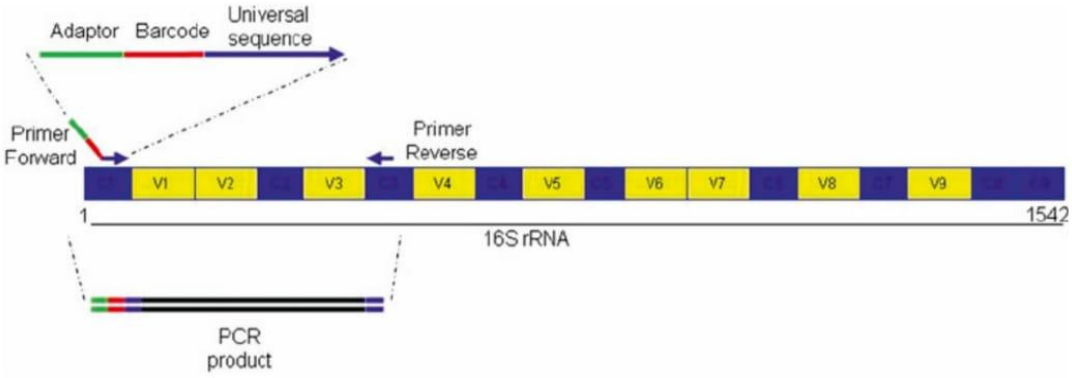
REVIEW OF LITERATURE

REVIEW OF LITERATURE

As the most important gamers in nearly each and every environment exploring like, bacteria make a contribution massively toward international electricity adaptation along with their cycling of count. As a consequence, profile of the microbial population is one of the main decisive duties for Microbiologists en route for see the sights various ecosystems. On the other hand, our knowledge of the empire Bacteria vestiges inadequate due to the fact maximum micro organism cannot be cultured or remote underneath laboratory situations. Here the times of yore very few decades, DGGE (Denaturing Gradient Gel Electrophoresis), T-RFLP (Terminal Restriction Fragment Length Polymorphism), FISH (Fluorescent in-situ Hybridization) furthermore gene chips be used as typical methods into investigations of bacterial networks moreover assorted variety in anticipation of the improvement of high-through situate sequencing advance. Lately, meta-genomic techniques furnished by next cohort aligning technology all along in the midst of Roche 454, Ion Torrent and illumina have facilitate a terrific extension of our indulgent on the subject of uncultured micro organism. The16S rRNA gene collection became first used in1985 for phylogenetic evaluation. Since all of it contains both very rationed areas for groundwork structure and hyper variable locales to recognize phylogenetic qualities of microorganisms, the16S rRNA quality succession turned into the most broadly utilized marked genetic material for profiling bacterial different populations. Complete end to end 16S rRNA quality arrangements comprise of the nine hyper variable districts that are isolated by means of nine exceedingly preserved areas. The bacterial 16S quality contains nine hyper variable areas (HVR1-HVR9) stretching out from around 29-99 base consolidates long that are related with the discretionary structure of the little ribosomal subunit. The level of protection shifts broadly between hyper variable locales, with increasingly moderated areas for Correlating to higher-level scientific categorization and less saved districts level to lower levels, for example, sort and species.

Despite the fact that the entire16S allows for assessment of all sequence hyper variable locales, at around 1500 base combines long it very well may be restrictively costly for studies looking to recognize or portray assorted bacterial networks.While16S hyper variable regions can be different vividly between bacteria, the16S quality all in all keeps up more prominent length homogeneity than its Eukaryotic partner, which can make arrangements simpler.

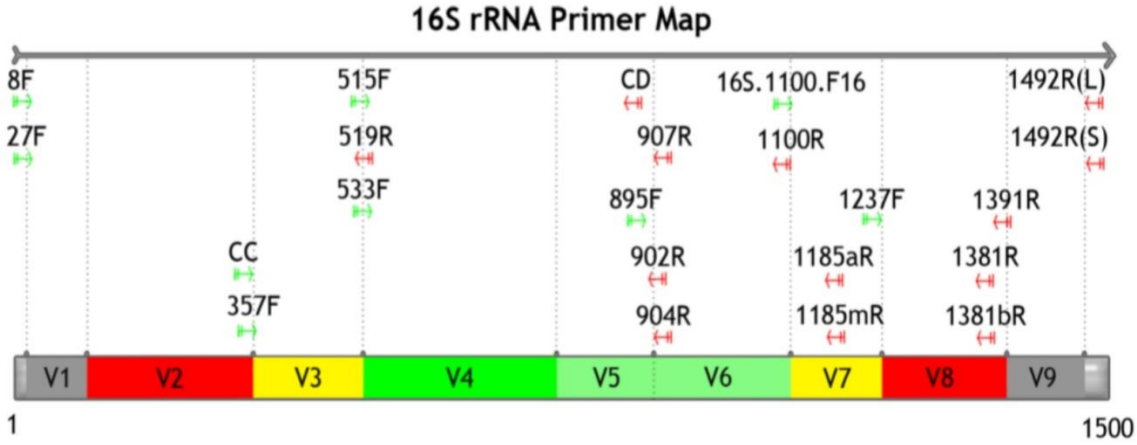
Moreover, the 16S quality contains exceptionally monitored arrangements between hyper variable areas, empowering the structure of widespread ground works that can dependably create similar segments of the16S succession crosswise over various taxa's. Albeit not a single hyper variable area can be precisely characterize all microorganisms from province to Species, a few can dependably foresee explicit ordered level. Many population studies select semi-conserved hyper variable regions like the HVR4 for this reason, as it can give goals as precisely as the full 16S quality. Whereas lesser-saved districts battle to group new species when higher request scientific classification are obscure, they are frequently used for the identify a nearness of explicit dangerous bacteria.



Conserved and hypervariable regions in the16S rRNA gene. The caught monitored areas C1–C9 are spoken to in blue, while the hypervariable districts are in yellow. An example of bar coded primer for amplification and pyro sequencing is shown

Constrained by sequencing innovation, the 16S rRNA quality groupings utilized in many examinations are fractional successions. Accordingly, the determination of legitimate preliminaries is basic to consider bacterial or microbial or micro-organism generation into different situations. Near the beginning examination have demonstrated that the utilization of various ground works may outcome within various DGGE design. Fresh and new studies utilize far above the ground through put expertise have also established that the use of sub most favorable primer pair outcome in the not smooth extension of definite species, cause whichever an under-or over evaluation of several species within a microbial vicinity. Albeit a few investigations have concentrated on ideal preliminary sets or, identically, ideal variable areas for the investigation of bacterial networks, they used engineered microbial networks and the taxa that were picked to direct those tests would generally impact the last outcomes. The proper introductions for a Polymerase and its basic Chain Reaction response are basic on the

grounds that an over-loosened up counterpart connecting a preliminary and that its objective prompts PCR disappointment. For 16S rDNAs, the ground installation (15–20 nucleotides (nt)) are to be found in the restrained areas that edge an objective district utilized for phylogenetic inspection. The main arrangements of ground works were planned by utilizing traditionalist districts of 16S rDNA groupings from various species and we renamed by their situations on Escherichia coli 16S rDNA, this has turned into the convention for ensuing preliminary structure. For instance, preface E685 compares to eubacterial P4 locale and groundwork A344 focuses on the archaeal H339 district. In the ongoing decades, more ground works have been intended for bacterial examinations with instruments, for example, ARB, as the quantity of known 16S rDNA arrangements increments. Additionally, ground works focusing on a predefined phylum have as of late been structured. Be that as it may, realize polymorphisms furthermore summative in the moderate areas, when countless 16S rDNA successions were created and saved in open databases, for example, the RDP database. Huge parallel sequencing systems permit uncommonly fast and efficient DNA sequencing. The problem of primers election is getting even more difficult and has attracted attention because of recent advances in meta genomic studies. Regarding one million successions of 400 nucleotides can be created by the Roche 454 FLX Titanium machine, permitting the profound aligning of the ecological DNA content of the bacterial micro-organism.



The accretion of recognized polymorphisms here in the sealed region means that the exposure rates of a quantity of primers are waning. This may reason troubles by means of widely established primers stipulation they be unsuccessful to make progress a high proportion of microbial species nun cultured environmental sample, as anticipated. By means of incorrect

primers resolve lead to disappointment to distinguish various bacterial species along with as a result unfinished survey in meta genomic studies. Prior studies encompass create that Archaea and Eubacteria precise primers cannot objective a band of domains. The acknowledged introduction meant for the Archaea are not continuously sensible for ornamenting the 16S rRNA amplicons for Korarchaeota or Nanoarchaeota. Utilizing the RDP which is used for the classification alongside the BLAST, Bakeret al. (2003) and Huwsetal. (2007) have research the species particularity and inclusion range of the known preliminaries. Nevertheless, the concluding study do not believe degeneracy's in these primers. In an ongoing work, the inclusion of a few realized preliminaries was reviewed utilizing a few arrangements of meta genomic information, and the groundworks with better execution were suggested for future work. Every one of these investigations utilized known preliminaries and gave brief data of their phylum particularity. Be that as it may, regardless we didn't have a positioning of the limits of the known preliminaries valuable for natural examples and a rundown of all hopeful introductions for bacterial 16S rRNA genes. Yong Wangetal. (2009) recognized conventional part of the genetic material fragment in16S rRNA genetic material from the RDP information base plus compile a record of aspirant primers. The predicted primers said on this observe comprise nearly a full- set of primers for prokaryotic 16S rRNA genetic material and in huge piece over lapped within the known primers, not considering of any transfer in position. The typical reporting lay the blame on of the primers is ninety six%, noticeably enhanced than that of dissimilar acknowledged primers. They also found the importance of their application, which will give the brightness on the topic of Meta – Genomic study. While 16S hyper variable region investigation is a dominant tool for bacteria ordered examinations, it battles to separate between firmly related species. In the families Enterobacteriaceae, Clostridiaceae and Pepto streptococcaceae, species can carve up upto 99% sequence resemblance from corner to corner the fullgrown16S RNA. Therefore, the HVR4 successions can vary by just a couple of nucleotides, leaving reference information bases unfit to dependably arrange these microorganisms at lower ordered levels. By constraining 16S investigation to choose hyper variable locales, these examinations can neglect to watch contrasts in firmly related taxa and gathering them into single ordered units, hence under evaluating the absolute decent variety of the example. Further increasingly, bacterial genomes can house multiple16S qualities, with the HVR1, HVR2, and HVR6 locales containing the best intra species decent variety.

Even as not the most exact strategy for grouping bacterial species, investigation of the hyper variable locales stays a standout amongst the most valuable devices accessible to bacterial

network ponders. Referring to the previous studies, 10 markers (conserved regions) situated in preserved region of the 16S rRNA genetic material were preferred to separate the complete extent align 16S rRNA gene sequence within the nine subregions. Each district begins with a monitored arrangement, and the remain derisa downstream factor succession. The tenth preserved district was one of the end marker, was expelled from the ninth subarea. The sub-regions were sequentially marked from HVR1 to HVR9. The cleaned dataset was separated keen on 9 files: nine for the V1-V9 subregions. Required knowledge on bacterial population organization can be compiled in the prevailing situation somewhere dissimilar communities which are represented as rows also the “species” as the columns, i.e., a community by the species template. Once relating bacterial population associations base on 16S rRNA genetic material sequences of the rna, each strand billed to the species, regularly term an operational taxonomic unit (OTU), by alignment base togetherness at a specific nucleotide aloofness, over and over again at a 97% uniqueness. The operational ordered unit (OTU) is an equipped characterization used to arrange gatherings of firmly associated people. Formerly it was introduced by the very know Robert R. Sokal and Peter H.A. Sneath with regards to Numerical scientific classification, someplace an "Operational Taxonomic Unit" is essentially the gathering of life forms at present being contemplated.

In such a logic, OTU is a realistic description to bunch folks by closeness, equal to however not really into accordance among traditional Linnaean scientific classification or present day evolution based scientific categorization of the taxonomic gene of the marker. In additional expressions, OUT’s are realistic proxy for mico-organism "species" at various ordered levels, without conventional frame works of natural arrangement as are accessible for plainly visible life forms. For quite a long while, OTUs comprise the most regularly utilized units of miro-organism assorted variety, particularly what time breaking down little subunit 16S or 18S rRNA marker quality grouping datasets. Thus, this population through OTU matrix, because it is mainly based on sepration between the smaller subunit genes of 16S rna found in OUT as together in row or column. This population by OUT table is used to see the differences between the mibrial populationsby the help of the plus or negative of the data in the matrix. All of these dissimilarities mixed in a distance matrix may be used for bacterial network comparisons by using ordination and clustering techniques. This method, termed “Taxonomy-unsupervised analysis,” originates from the distribution of 16S rRNA gene sequences into OTUs. When applying taxonomy unsupervised evaluation to very large numbers of sequences (>106) produced through the brand new sequencing technologies, a lot large computational

capacities are required to analyze the records. The alignment and clustering of sequences that require calculation of pair clever nucleotide distances is one bottle neck. Taxonomy unsubstantiated evaluation is constructive in that it consists of sequences that are not yet assignable to bacterial taxonomy, but the contemporary computational barriers make it a touch hard to practice.

As the real players in practically all situations investigated, microscopic organisms contribute monstrously to worldwide vitality change and there cycling of issue. Along these lines, profile making of the microbial complex system is a stick out amongst the most significant assignments for Microbiologists to investigate different environments. In any case, our comprehension of the kingdom Bacteria stays constrained on the grounds that most microscopic organisms can't be refined or secluded under research center conditions. After all these years of study, DGGE (Denaturing gradient gel electrophoresis), T-RFLP (Terminal restriction Fragment length polymorphism), FISH (fluorescent in situ hybridization) and genetic chips are used as main methodology in the researches of bacterial complex systems and an assortment of multiplicity in anticipation of the improvement of high-throughput sequencing modernization. Recently, meta-genomic methods provided by next generation aligning technology such as Roche 454, Ion Torrent and illumina have facilitate a extraordinary development of our infirmation about not growing in lab bacteria. The16S rRNA gene sequence become first used in1985 for phylogenetic evaluation. Since it carries each very moderated regions for preliminary plan and hyper variable locales to distinguish phylogenetic attributes of microorganisms, the16S rRNA eminence understanding become the maximum extensively useful marker for harmonizing the bacterial subgroups. Full-period 16S rRNA quality preparations add in of nine hyper variable regions which could be remote through 9 incredibly stored locales. The bacterial 16S quality contains nine hyper variable areas (HVR1-HVR9) running from around thirty to hundred base matches stretched that are engaged with the auxiliary organization of the little mRNA building machine called Ribosome that inclides smaller subunit too.

The level of protection fluctuates broadly between hyper variable locales, with increasingly rationed districts for the Correlating to higher-level scientific classification and less saved areas to lower levels, for example, sort and species. While the entire16S grouping takes into consideration correlation of all hyper variable districts, at around 1500 base matches long it very well may be restrictively costly for studies looking to distinguish or describe assorted

bacterial communities.While16S hyper variable locales can differ drastically between microbes, the16S quality all in all keeps up more prominent length homogeneity than its Eukaryotic partner, which can make arrangements simpler. Also, the 16S quality contains very monitored arrangements between hyper variable locales, empowering the structure of general preliminaries that can dependably create similar segments of the16S grouping crosswise over various taxa. Albeit no hyper variable locale can precisely order all microscopic organisms from province to Species, a few can dependably anticipate explicit ordered levels. Countless population studies select halfly-conserved hyper variable regions like the HVR4 for this reason, as it can endow with declaration as precisely as the jam packed 16S gene. Though lesser-preserved locales battle to crowd new species when higher request scientific classification is obscure, they are frequently used to distinguish the nearness of explicit gram positive or the dangerous or infectious micro-organisms. Biological cell is almost similar to the structural and functional unit of a bacterial cell. Micro-organism which are too many numbers in trillion are included in the realm of the Bacteria. Bacteria's are very small they are just hardly any micrometers in length, they have different different shapes. They are seen in rod shape, seen in round shape, seen in comma shape and in spiral shape. We know that the bacteria's are the first form of life that appeared on the not favorable for living conditions of earth and they are now more likely to be present everywhere you can imagine. They can be found present in the deep down crust of the earth, they can found in very acidic conditions, they are found in normal environment, they are found in water, in air, everywhere you can imagine. Bacterial cells are also found in wastes like radioactive. They are found in soil, sometimes they are very beneficial but the other times they can be very dangerous. There are only 25 percent of the species of the bacteria which are already classified in phylum and kingdom can be grown separately in the laboratory conditions. This is the percentage of bacterial that can be grown and imagine the vast species which do not grow in such condition in the laboratory. There are many more species which are yet to be discovered. If we think then the becteria's are the most important micro-organism on the earth, they are only the organisms who have the capability to produce the vitamin which is required by all the living organisms. The vitamin which these bacteria's and some other archebacterias produce cannot be produced by the other organism naturally. It is the vitamin B12 that is produced by them and is included in the food chain and thereby get into the system. It is the very important vitamin which is required for many important functioning of the cell of the whole body. This vitamin is reqred for the formation of pigment called myelin which is required for the proper functions and working of the neuronal system of the organism. This vitamin is required by

every cell of the body of the human or other organism. Bacteria's are found in relation with the host they can live within the organism parasitically and they can live in symbiosis also.

Bacteria's are the biggest and the vastest phylum. They are innumerous in numbers they are about 40.5 million if we look in the small amount of the soil and they are present in so many numbers in few milliliters of the water. This show has much they can be in numbers if we all try to discover the new bacteria's we will be surprised with the numeracy of these micro-organisms. They are required for all kinds of reactions happen in the universe. They are for atmospheric reactions like fixations of the molecular nitrogen and that is very very important for the regulation of the molecular form of nitrogen in the environment. This is because they get associated with the root nodules of the plants and then they convert the molecular nitrogen to other forms and make the nitrogen available for other organisms in their own favorable taking form. The gobar gas the brightest invention for the poor people to use and that too very easily with the help of these micro-organisms. These micro-organisms especially the bacterial species. These bacterial species are well known for their decomposition. These micro-organisms break the very complex substances to their simpler form and there by getting them re circulated in the recycling form. Bacteria's help the bigger substances to reduce into smaller ones and then they get disappeared in the soil. As we have discussed above that bacteria's are goon in so many senses but nothing is too good when they also cause harm to health of our organisms very drastically. They can actually cause so many different diseases to human and other that even a very small or a single wrong bacterial to the health body of such a multi cellular organism can get very sick or even get to the position where it can never be fine again. People died in numbers when they bacteria's were at their worst activities. These bacterias can survive bad condition inside the body of the organism which they use as their defense mechanism for getting healthy gain.

Bacterias are also be used in laboratory for reactions and used in industries and in many big research institutes and universities. Sometimes the in small dis-handelling and not proper protection can be very dangerous for the environment. This happen although very often in the past that there has been the discharge of the industrial was in the nearby naturally occurring water body that polluted, they water of the city and this causes dangerous situation of the marine life and for the life outside the water. Such incidents can make every single person get in and can cause epidemic disease and this is not the easy situation for the city or the government of the city. Every person in such companies or industries should do their work

very carefully because many lives are dependent on them. But in the new development the scientist has found the cure of everything. There are so many equipments that are invented or discovered for the cleaning and detection of the harmful effluents discharged from the industries and companies.

To reduce the action of these harmful bacteria's all the developed countries uses the antidots that can kiil the bacteria inside the body of the host. These antidots or we can commonly say them as the anti-biotics kill the bacteria's by just removing the outer layer of their cells. When the cell wall of the bacteria is removed they cannot survive much longer inside the body. The host organism or the body of the human creates dangerous and uncomfortable environment for the bacteria's inside the body. They cannot survive in high temperature, cannot survive the acidic or alkaline ph of the plasma or the other fluids of the host body. Bacteria with the cell wall is very venerable, they do not have any double membrane structure in their cell body. They just have this capsule kind of double protecting layer on the covering of the cell. This cell wall covers the whole body of the bacteria and provide it a safe and protective environment. In the anti-biotics what we do is to get their cell wall busted or to prick the cell wall of the bacteria. Due to this the cell will get in the low osmolarity fluid hence the water or the body fluid gets inside the bacterial cell and then the bacterial cell will get busted because of intake of the low osmolarity fluid. Like other eukaryotes they have double membrane bound cell organelles and the controller of the single cell who we generally call the Nucleus. Nucleus control every single activity of the cell, help the cell to keep its authenticity for the DNA's and get the same authenticity every next generation. Where as in Prokaryotes or bacteria's this is not the case, the do not have any kind of power that controls the whole cell. They just have the single membrane bound cell organelles they do their work and the bacteria have the clustered RNA's or DNA's that is usually called as the plasmid. The bacteria does not even get divide like a normal cell. Bacteria does not have any stages of cell division. We all know how the cells get divided they go whole lot of a big nuclear and cellular changes before they finally become the two daughter cells. In Bacteria they do not follow any such method they are divided into two daughter cells very easy and very simply. That process through which the cell of the bacteria become two daughter cells is called the conjugation. In Conjugation two bacteria gets closer to each other and they share tube like canal, they develop the tube-like canal between them and then they transfer the plasmid of one cell to another and this happens until any minute change in the environment happen. The conjugation tube breaks when there is any very small or very minute change happens.

There are some exceptions in the species that comes under bacteria that do not follow the same rules as the main species of bacteria do. There are some bacteria which do not have any cell wall around them. They do not need it and hence they are not infectious. They are the smallest bacteria reported and can even be filtered from the bacterial filter. There are some other realms also that has the double membrane cell wall but does not have any organelles they are considered as the oldest bacteria present on earth.

There are some species of bacteria which are considered as the predators. These predator bacteria have so many ways to kill their target. There are some bacteria which can kill the target by making circular attack with the help of other bacteria of the same species and then kill their target. Generally, the target of such bacteria is the other small micro-organism only, they kill these small micro-organisms very easily just but ganging up on them. The other way of killing the small microbes are to get inside of them. There are some species of bacteria which gets inside of the cell in the cytosol of the bacteria they then imbibe all the nutrients of the cell and make the target shrunk to die. Many bacteria does not follow either of the practice to kill the target microbe, usually sole bacteria simply attach to the surface of the target and leech out the cell fluid of the target bacteria. This is how the many species of the bacteria are known to be as the predators. Some of these species of the bacteria helps to get the other dangerous species of the bacteria to vanish out and hence such bacterial species are very important for the research purpose and they are going to be used in future science to get the infectious and harmful bacteria to vanish out.

OBJECTIVES

1. Developing an in-silico pipeline to retrieve the 9 Hypervariable regions from a 16S rRNA sequence using conserved regions or primers already identified. Also, create such a data base of each hyper variable region from the given16S rRNA database.

2. Analysing the taxonomic sensitivity and accuracy of each hyper variable region in taxonomy assignment. This is to be done by a comparative analysis on the taxonomic abundance data from each hypervariable region, taking the primers into account, as well as without any primer effect.

3. Doing data analysis on the abundance data to find specific cases of bias towards/against a particular taxon for a HV region, and to create a pipeline to find the preference order of HV regions for a taxon.

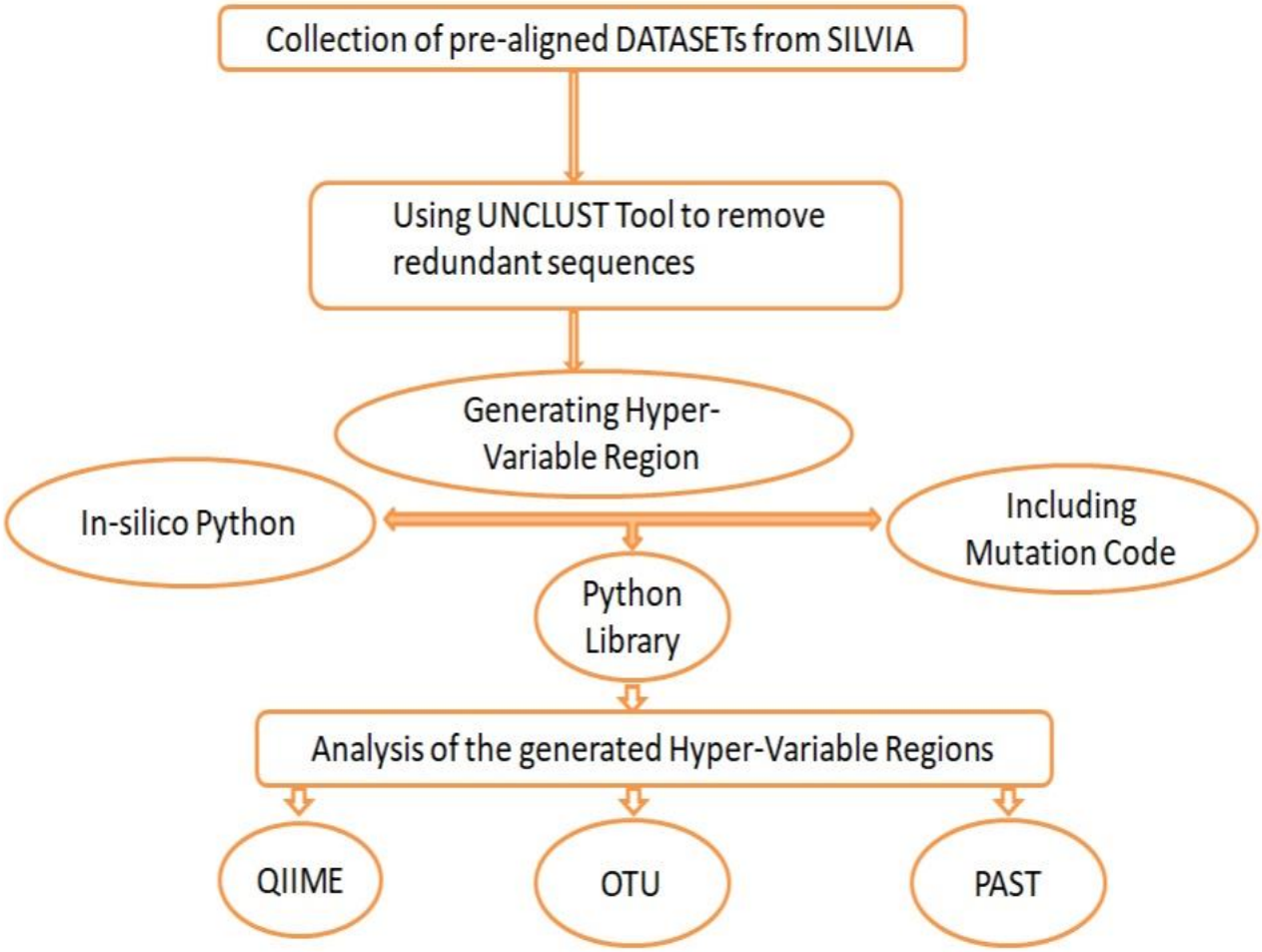
CHAPTER - 3

WORKFLOW

AND

MATERIALS AND METHOD

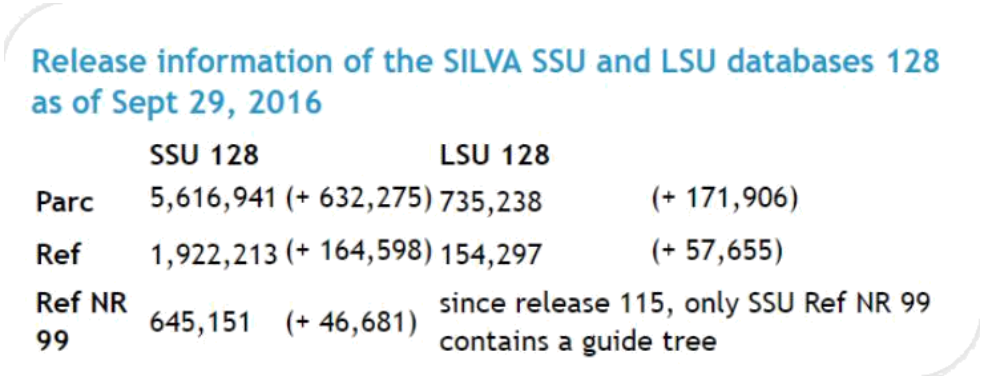
Work flow


Collection of pre-aligned DATASETs from SILVIA
Using UNCLUST Tool to remove redundant sequences
Generating Hyper-Variable Region
In-silico Python
Including Mutation Code
Python Library
Analysis of the generated Hyper-Variable Regions
QIIME
OTU
PAST

MATERIAL AND METHODS

DATASETS:

1. The pre-aligned and truncated SILVA Ref12816 SrRNA NR99 data set was arranged by downloading through the SILVA online data set as the first and foremost data set. It is based on the Ref records set with a ninety nine% criterion carried out to cast off redundant sequences the use of the UCLUT device. Sequences from cultivated species are preserved in all cases.



Release information of the SILVA SSU and LSU databases 128 as of Sept 29, 2016

	SSU 128		LSU 128	
Parc	5,616,941	(+ 632,275)	735,238	(+ 171,906)
Ref	1,922,213	(+ 164,598)	154,297	(+ 57,655)
Ref NR 99	645,151	(+ 46,681)	since release 115, only SSU Ref NR 99 contains a guide tree	

Figure5: Release information for Silva 128

2. The final data set contains 6,45,151 sequences and can be used as a representative data set for phylogenetic analysis and classification. Separate files were created for each hypervariable region, which were then used in the comparative analysis.

3. Due to primers chosen creating the main difference in the taxonomy assignment of archaea, our study is restricted only to Bacterial populations.

4. Conserved Sequences used to derive the hypervariable regions are taken from primers already being used as wel as predicted primers with good coverage rates. These predicted primers are taken from Yong Wang, Pei-Yuan Qian (PLOSone2009).The already existing primers along with conserved sequences (along with their coverage rates) are as shown below:

Primer*	Sequence (5'-3')	Target Group	Reference
8F	AGAGTTTGATCCTGGCTCAG	Universal	Turner et al. 1999
27F	AGAGTTTGATCMTGGCTCAG	Universal	Lane et al. 1991
CYA106F	CGGACGGGTGAGTAACGCGTGA	Cyanobacteria	Nübel et al. 1997
CC [F]	CCAGACTCCTACGGGAGGCAGC	Universal	Rudi et al. 1997
357F	CTCCTACGGGAGGCAGCAG	Universal	Turner et al. 1999
CYA359F	GGGGAATYTTCCGCAATGGG	Cyanobacteria	Nübel et al. 1997
515F	GTGCCAGCMGCCGCGGTAA	Universal	Turner et al. 1999
533F	GTGCCAGCAGCCGCGGTAA	Universal	Weisburg et al. 1991
16S.1100.F16	CAACGAGCGCAACCCT	Universal	Turner et al. 1999
1237F	GGGCTACACACGYGCWAC	Universal	Turner et al. 1999
519R	GWATTACCGCGGCKGCTG	Universal	Turner et al. 1999
CYA781R	GACTACWGGGGTATCTAATCCCWTT	Cyanobacteria	Nübel et al. 1997
CD [R]	CTTGTGCGGGCCCCCGTCAATTC	Universal	Rudi et al. 1997
1100R	AGGGTTGCGCTCGTTG	Bacteria	Turner et al. 1999
1391R	GACGGGCGGTGTGTRCA	Universal	Turner et al. 1999
1492R (l)	GGTTACCTTGTTACGACTT	Universal	Turner et al. 1999
1492R (s)	ACCTTGTTACGACTT	Universal	Lane et al. 1991

Figure 6: List of known primers along with their references

Bacteria	Position	Sequence	Average rate	Coverage rate
E	321–336	ACTGAGACACGGYCCA	95.7%	96.1%
E	329–343	ACGGYCCARACTCCT	95.3%	96.0%
E	338–358	ACTCCTACGGGAGGCAGCAGT	97.3%	96.3%
A	346–361	GGGGYGCAGCAGGCG	94.2%	94.3%
E	350–364	GGCAGCAGTRRGGAA	95.1%	95.5%
E	505–524	GGCTAACTHC GTGCCAGCAG	95.3%	95.1%
A	514–528	GGTGYCAGCCGCCGC	97.3%	98.5%
E	515–532	GTGCCAGCAGCCGCGGTA	92.6%	91.0%
U	515–532	GTGYCAGCMGCCGCGGTA	-	96.9%/96.9%
A	519–539	CAGCCGCCGCGGTAAHACCRC	96.7%	97.1%
E	683–700	GTGTAGMGGTGAAATKCG	92.6%	90.5%
E	783–797	CAGGATTAGATACCC	97.9%	97.9%
E	785–806	GGATTAGATACCCTGGTAGTCC	95.9%	94.6%
A	785–800	GGATTAGATACCCSGG	98.1%	98.4%
U	785–800	GGATTAGATACCCBGG	-	98.4%/97.1%
A	884–898	TGGGRAGTACGKHCG	97.1%	97.1%
A	899–913	CAAGDMTGAAACTTA	97.6%	97.6%
A	905–920	TGAAACTTAAAGGAA	98.3%	98.3%
A	921–936	TTGGCGGGGGAGCAC	98%	97%
E	909–926	ACTCAAAKGAATTGACGG	98.5%	97.9%
U	909–928	ACTYAAAKGAATTGRCGGGG	-	93.2%/92.1%
E	919–939	ATTGACGGGGRCCCGCACAAG	96.3%	96.1%
A	947–964	GCSTGCGGYTYAATTGGA	91.6%	90.5%
E	949–964	ATGTGGTTTAATTCGA	93.5%	93.5%
A	958–973	AATTGGABTCAACGCC	90.6%	93.5%
E	969–984	ACGCGARGAACCTTAC	97.4%	97.1%
A	1045–1059	GAGGWGGTGCATGGC	95.7%	97.4%
A	1052–1071	TGCATGGCCGYCGYCAGYTC	96.6%	95.1%
E	1052–1072	TGCATGGYTGTCGTCAGCTCG	97.1%	99.0%
U	1052–1071	TGCATGGYYGYCGYCAGYTC	-	95.1%/98.8%
E	1063–1081	CGTCAGCTCGTGYCGTGAG	99.2%	99.3%
E	1096–1114	CCCRYAACGAGCGCAACCC	96.8%	95.6%
E	1177–1193	GGAAGGYGGGGAYGACG	98.2%	98.2%
A	1226–1242	CACGCGSGCTRCAAWGG	93.8%	93.5%

Figure 7: List of the predicted primers by Yong Wang (2009) along with their coverage rates

METHOD:

Generating the Hypervariable Regions:

1. The 6,46,151 sequences from the final data set were first parsed using the Bio python library, and files were generated for each hypervariable region.

2. Regular expressions for each of the hypervariable regions were written using primers and conserved sequences researched from existing literature.

3. Specific code was written to include mutations in the primer sequences, which can be set to a particular tolerance level. Mutations can be insertions, deletions as well as substitutions.

4. Finally, the script was run on the entire data set of 6,46,151 sequences, generating a huge database of hypervariable regions which can be used for the comparative study.

5. The number of expressed hypervariable regions were counted and compared, thus giving an idea on the level of coverage of each HV region from the entire dataset.

6. An in-silico pipeline was developed in Python to do this entire task, which can update the computationally generated hypervariable regions periodically with updated 16Sr RNA datasets.

```
### Generating hypervariable regions ###

primers={1: 'AGAGTTTGATCATGGCTCA',2:'GGCG[GCA]ACGGGTGAGTAA',3:'CA[TC]TGG[GA]ACTGAGACACGG[TC]CC',4:'GGCTAACT[ACT]CGTGCCAGCAGC',
         5: 'CGAAAG[TC]GTGGG[GT]A[GT]C[GA]CAGG',6:'ACTCAAA[GT]GAATTGACGGGG[GA]C',7:'GTG[GC]TGCATGG[TC]TGTCGTCAGCT',8:'GGAAGG[TC]GGGGA[TC]GACGTCAA
         9:'TGTACACACCGCCCGTCACACAC',10:'AAGTCGTAACAAGGTAACCGTA'}

='[A-Z]*?(?b)(?:AGAGUUUGAUCAUGGCUCA)(e<=4)([A-Z]*?)(?b)(?:GGCG[GCA]ACGGGUGAGUAA)(e<=4)([A-Z]*?)(?b)(?:CA[UC]UGG[GA]ACUGAGACACGG[UC]CC)\
(e<=4)([A-Z]*?)(?b)(?:GGCUAACU[ACU]CGUGCCAGCAGC)(e<=4)([A-Z]*?)(?b)(?:CGAAAG[UC]GUGGG[GU]A[GU]C[GA]CAGG)(e<=4)([A-Z]*?)(?b)\
(?:ACUCAAA[GU]GAAUUGACGGGG[GA]C)(e<=4)([A-Z]*?)(?b)(?:GUG[GC]UGCAUGG[UC]UGUCGUCAGCU)(e<=4)([A-Z]*?)(?b)(?:GGAAGG[UC]GGGGA[UC]GACGUCAA)(e<=4)([A-Z]*?)
(?b)(?:UGUACACACCGCCCGUCACACAC)(e<=4)([A-Z]*?)(?b)(?:AAGUCGUAACAAGGUAACCGUA)(e<=4)[A-Z]*?'


r=['(?b)(?:'+primers[i]+')(e<=4)([A-Z]*?)(?b)(?:'+primers[i+1]+')(e<=4)' for i in range(1,10)]
v_regions={1:[],2:[],3:[],4:[],5:[],6:[],7:[],8:[],9:[]}
flag=0
for seq_record in SeqIO.parse("E:\\99_otus_16S_new_3.fasta", "fasta"):
    flag+=1
    seq=str(seq_record.seq)
    for i in range(9):
        new=seq_record.upper()
        t=regex.search(r[i],seq)
        if t:
            new.seq=Seq(t.group(1))
            new.description=''
            new.id='V'+str(i+1)+'_'+new.id
            hv_regions[i+1].append(new)
            #print 'Found %i' %(i)
    if flag%5000==0:
        print flag
        path='E:/Hypervariable//HV_'+str(flag)
        os.makedirs('E:/Hypervariable//HV_'+str(flag))
        for i in hv_regions:
            SeqIO.write(hv_regions[i], path+'//V'+str(i+1)+'_regions.fasta', "fasta")
```

Figure 8: A screen shot from the code for the in-silico pipeline

Analysis of the generated Hypervariable regions:

QIME: The generated hypervariable regions were separated into 9 distinct files for each region. QIIME software package was then used to assign.

Taxonomy to al the short HV sequences.

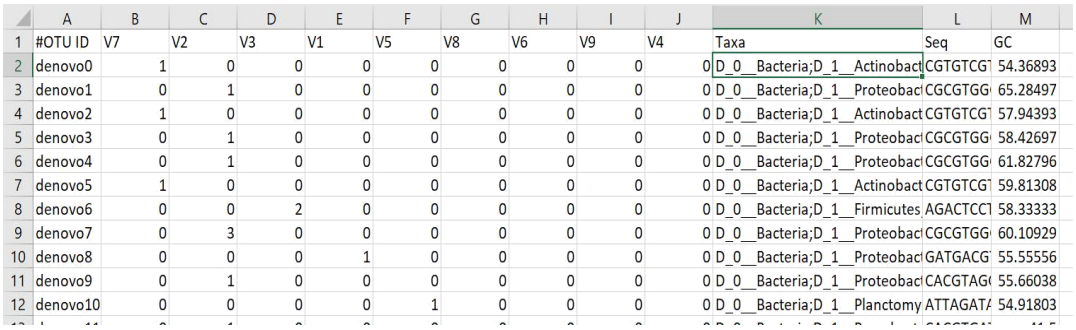
QIME is an open-source bioinformatics pipeline for performing micro-biome analysis from raw DNA sequencing data. QIME is designed to take users from raw sequencing data generated on the Illuminator other platforms in the course of periodical excellence graphics and statistics. This includes the considering more than two at a time and hence we call it doing the multi- plexing and eminence filtering, OTU selection, taxonomic project, and phylogenetic modernization, and multiplicity or assortment analyses and visualizations. QIME has been applied to studies based on billions of sequences from tens of thousands of samples. QIME has 3 OUT picking protocols:

- De-novo OUT picking: In a de-novo OUT choose manner, reads are bind against each other without any outside orientation sequence gathering. Pick de novo is the main interface for de-novo OTU selecting in QIME, and it is considered the important work of the taxonomy, checking and aligning of the strands and making of the tree.

- OUT checking in nearby references: The picking of the OUT file in the nearby references, in this the main read is done by the clustering of the many reads together and if they do not match with the reference read then there are excluded by the downstream picking of the OUT reads.

- Now let's talk about Open- checking of the OUT reading: In this all the reads are clustered together, and they are aligned or checked to the final OUT reading and they match we call it open reading of the OUT table and if they don't we can't call it so.

QIME returned out table, assigning taxonomy and generating out bins from the given sequences. Python scripts were written to break down the out table and get abundance data sorted by taxonomic ranks and hypervariable regions. The output abundance files were then compared with the actual abundance at a of the16S rRNA database used initially.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	#OTU ID	V7	V2	V3	V1	V5	V8	V6	V9	V4	Taxa	Seq	GC
2	denovo0	1	0	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Actinobact	CGTGTCG	54.36893
3	denovo1	0	1	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	CGCGTGG	65.28497
4	denovo2	1	0	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Actinobact	CGTGTCG	57.94393
5	denovo3	0	1	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	CGCGTGG	58.42697
6	denovo4	0	1	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	CGCGTGG	61.82796
7	denovo5	1	0	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Actinobact	CGTGTCG	59.81308
8	denovo6	0	0	2	0	0	0	0	0	0	D_0__Bacteria;D_1__Firmicutes	AGACTCC	58.33333
9	denovo7	0	3	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	CGCGTGG	60.10929
10	denovo8	0	0	0	1	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	GATGACG	55.55556
11	denovo9	0	1	0	0	0	0	0	0	0	D_0__Bacteria;D_1__Proteobact	CACGTAG	55.66038
12	denovo10	0	0	0	0	1	0	0	0	0	D_0__Bacteria;D_1__Planctomy	ATTAGAT	54.91803

Figure9: A screenshot of the generated OUT file

Since the comparison with the actual 16S rRNA database would also incorporate the bias created by the primers and the effect of the number of sequences from which we could successfully extract the hypervariable regions, another set of abundance data was created using python scripts. This was the actual taxonomic abundance data for each specific hypervariable region. This would remove any primer biases we have in the analysis.

PAST3.20: PAST is a comprehensive, software package for executing a range of standard numerical analysis and operations used in quantitative paleontology. The software, called PAST (PAleontological STatistics), runs on well known Windows computers and is to be had free of fee. PAST integrates spreadsheet-type data entry with univariate and multivariate statistics, curve fitting, time-series analysis, data plotting, and simple phylogenetic analysis. Many of the functions are specific to paleontology and ecology, and these functions are not found in standard, more extensive, statistical packages.

Consolidated abundance data for comparison was input in the PAST software and used for calculating Bray-Curtis distance matrices. Given in the form of similarity (1-Distance), the Bray-Curtis similarity of the output abundance data with the original 16S rRNA abundance issued a sour metric to decide the taxonomic sensitivity and accuracy of the hyper variable regions.

The Bray Curtis is similarity is used to quantify the differences in species populations between two different sites. Bray-Curtis is a popular similarity index for abundance data. It’s used primarily in ecology and biology, and can be calculated with the following formula.

PAST calculates Bray-Curtis similarity as follows:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} \qquad d_{jk} = 1 - \frac{\sum_i |x_{ji} - x_{ki}|}{\sum_i (x_{ji} + x_{ki})}.$$

The Bray-Curtis similarity between different hyper variable regions from the similarity matrix were also noted and analyzed. These are important since sequencing techniques can incorporate 2 or 3 contiguous hyper variable regions. Therefore, the similarity in the taxonomic abundance generated by different HV regions, in tandem with their similarity to the true 16S abundance, is also a major criterion in deciding which region of the 16S to sequence.

Jaccard index and Euclidean distance were 2 ot her metrics which were also evaluated to get a better idea of there sultan abundances. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It is popularly used in case of Presence/ Absence species data between two ecological samples.

Dendrograms were constructed for the visual representation of the closeness among the hyper variable regions. A dendrogram is a type of tree diagram showing hierarchical clustering—relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. The dendrograms constructed in our case were between the different hyper variable regions and the 16S data. Metric used for clustering was Bray-Curtis's similarity and algorithm used for tree

Species populations between two different sites. Bray-Curtis is a popular similarity index for abundance data. It's used primarily in ecology and biology, and can be calculated with the following formula:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

PAST calculates Bray-Curtis similarity as follows:

$$d_{jk} = 1 - \frac{\sum_i \left| x_{ji} - x_{ki} \right|}{\sum_i \left(x_{ji} + x_{ki} \right)}.$$

7. The Bray-Curtis similarity between different hyper variable regions from the similarity matrix were also noted and analyzed. These are important since sequencing techniques can incorporate 2 or 3 contiguous hyper variable regions. Therefore, the similarity in the taxonomic abundance generated by different HV regions, in tandem with their similarity to the true 16S abundance, is also a major criterion in deciding which region of the 16S to sequence.

8. Jaccard index and Euclid and distance were 2 ot her metrics which were also evaluated to get a better idea of the resultant abundances. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It is popularly used in case of Presence/ Absence species data between two ecological samples.

9. Dendrograms were constructed for the visual representation of the closeness among the hyper variable regions. A dendrogram is a form of tree diagram showing hierarchical clustering—relationships between comparable sets of facts. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. The dendrograms constructed in our case were between the different hyper variable regions and the 16S data. Metric used for clustering was Bray-Curtis's similarity and algorithm used for tree construction was the classical UPGMA clustering algorithm.

10. Different hyper variable regions were checked for any present biases towards any taxa. This was done using python scripts, and then the short listed cases were analyzed further using PAST. Bar charts were plotted and variance of abundance between the hyper variable regions was calculated.

CHAPTER - 4

RESULTS

Results

1. Hyper variable region from the entire dataset were retrieved and classified into 9 separate files. An in-silico pipeline was created to get the 9 Hyper variable regions from an unknown 16S rRNA sequence whenever required. A sample of the resultant file is shown below:

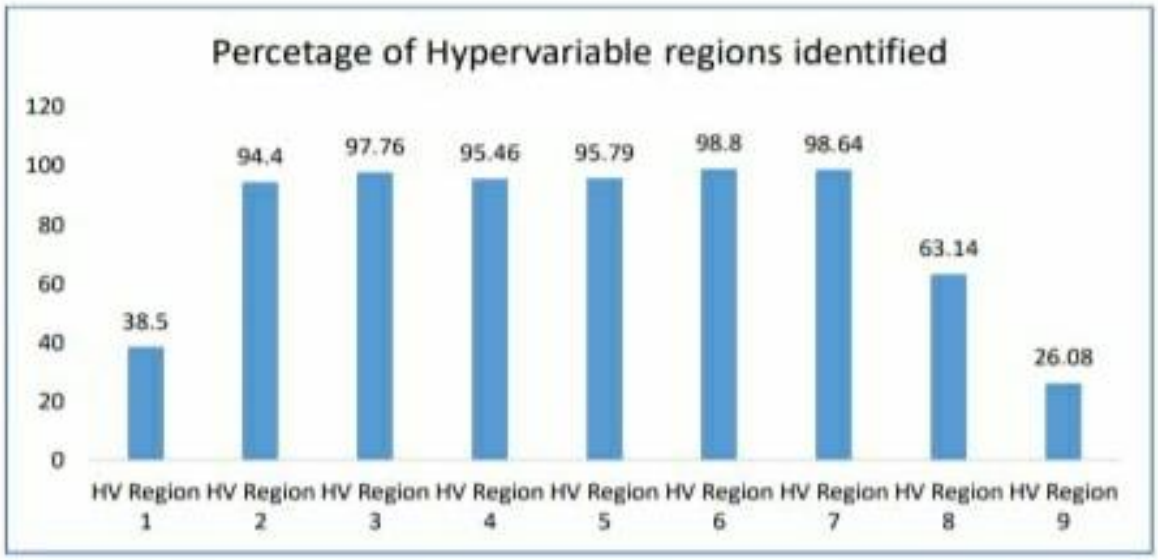
A Screen Shot of generated Hyper Variable Region

2. A sample of the starting and end position on E. Coli of each HV regions is shown below:

Region	Start Position	End Position
V1	8	96
V2	97	306
V3	307	487
V4	488	746
V5	747	885
V6	886	1029
V7	1030	1180
V8	1181	1372
V9	1373	1468

Tabular representation of the start points and the end point of the 9 Hyper Variable Regions on the 16S rRNA.

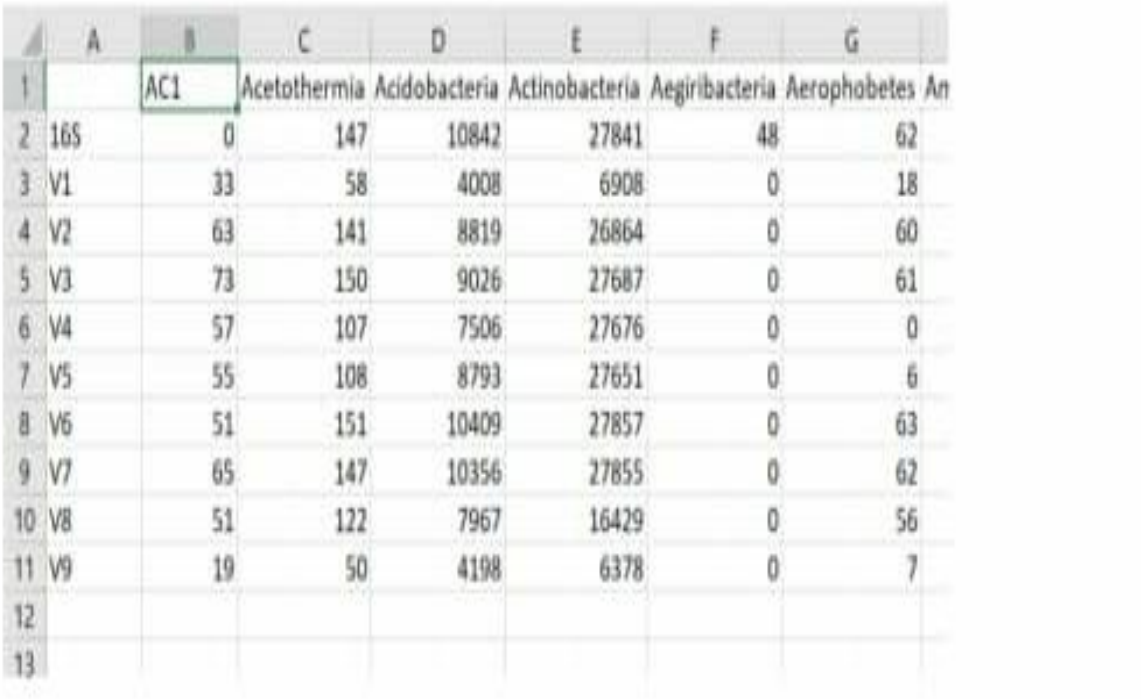
3. The percentage of Hyper Variable regions retrieved from all the sequences for each hyper variable region is as represented in the following bar chart:



Percentage of Sequences from which our in-silico pipeline could extract the corresponding Hyper Variable Region using the Primer Sequence.

Therefore, we can see that the current primers or identified conserved sequences are not able to fully cover all the 16S rRNA sequences. Moreover, since the percentage of sequences extracted from hyper variable regions 1, 8 and 9 were too low, they were not considered statistically significant in any of the following analysis.

4. The QIIME generated OUT file was classified according to the taxonomic ranks and clustered together for all hyper variable regions. Along with this another set of abundance files were generated having the corresponding actual abundances for each hyper variable region rather than the entire 16S rRNA sequence.



	A	B	C	D	E	F	G	
1		AC1	Acetothermia	Acidobacteria	Actinobacteria	Aegiribacteria	Aerophobetes	An
2	16S	0	147	10842	27841	48	62	
3	V1	33	58	4008	6908	0	18	
4	V2	63	141	8819	26864	0	60	
5	V3	73	150	9026	27687	0	61	
6	V4	57	107	7506	27676	0	0	
7	V5	55	108	8793	27651	0	6	
8	V6	51	151	10409	27857	0	63	
9	V7	65	147	10356	27855	0	62	
10	V8	51	122	7967	16429	0	56	
11	V9	19	50	4198	6378	0	7	
12								
13								

A screen Shot of generated output abundance file.
This particular screenshot is for the rank phylum.

REFRENCES

References

1. Heuer, H., Hartung, K., Wieland, G., Kramer, I., & Smalla, K. (1999). Polynucleotide probes that target a hypervariable region of 16S rRNA genes to identify bacterial isolates corresponding to bands of community fingerprints. Applied and Environmental Microbiology, 65(3), 1045-1049.

2. Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology, 44(4), 846-849.

3. Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods, 69(2), 330-339.

4. Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods, 69(2), 330-339.

5. Yu, Z., Garcia-Gonzalez, R., Schanbacher, F. L., & Morrison, M. (2008). Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. Applied and environmental microbiology, 74(3), 889-893.

6. Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261-5267.

7. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and environmental microbiology, 71(12), 7724-7736.*

8. Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., & Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS genetics, 4(11), e1000255.

9. Tsetsarkin, Konstantin A., Dana L. Vanlandingham, Charles E. McGee, and Stephen Higgs. "A single mutation in chikungunya virus affects vector specificity and epidemic potential." *PLoS pathogens* 3, no. 12 (2007): e201. Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., ... & Methé, B. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research.

10. Susan M., Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, and Mitchell L. Sogin. "Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing." PLoS genetics 4, no. 11 (2008): e1000255.

11. Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., ... & Won, S. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International journal of *systematic and evolutionary microbiology, 62(3), 716-721*

12. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72(7), 5069-5072.

13. Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K., & Lim, Y. W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. International journal of systematic and evolutionary microbiology, 57(10), 2259-2261.

14. Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., ... & Beiko, R. G. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology, 31(9), 814.