INSIGHTS INTO GLACIAL METAGENOME AND SEQUENCE BIASES IN COMPARATIVE METAGENOMICS

by

SULBHA CHOUDHARI

A dissertation submitted to the

Graduate School – Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Andrey Grigoriev

And approved by

_____

Andrey Grigoriev

_____

Daniel H. Shain

_____

Jongmin Nam

_____

Roman Dial

Camden, New Jersey

May 2015

ABSTRACT OF THE DISSERTATION

Insights into Glacier Metagenomes and Sequence Biases in Comparative Metagenomics

by SULBHA CHOUDHARI

Dissertation Director:
Dr. Andrey Grigoriev

Of the land surface in the world, 25% is classified as a cold environment that is a large reservoir of microbial activity, such as glaciers and ice lakes. However, most of the resident organisms on glaciers are single celled and unculturable; therefore, the best way to gain insight into their community structure is by a metagenomics approach. Metagenomics by next generation sequencing has become an important tool for interrogating complex microbial communities, and has made it possible to study uncultured microbes. We analyzed the microbial diversity of an Alaskan glacier using 16S rRNA sequencing, and determined the functional potential of these communities by whole metagenomic sequencing. A rich and diverse microbial population of more than 2,500 species was revealed, including several species of *Archaea* that have been identified for the first time in the glaciers of the Northern hemisphere. A comparative analysis of the community composition and bacterial diversity present in Alaskan glacier with other environments showed a large overlap with an Arctic soil than with a high Arctic lake, indicating patterns of community exchange, and suggesting that these bacteria may play an important role in soil development. The metabolic potential of glacial ice metagenome showed a high versatility for different substrate at a low-nutrient environment. Numerous genes encoding for synthesis of unsaturated fatty acids and

cryoprotectants were detected, which are the characteristics for metabolic adaptations at sub zero temperatures. Also, many sequences showed similarities to genes for methane, nitrogen, and sulfur metabolism.

Though advancements in sequencing technology have made it possible to study metagenomes, they introduce different biases, which significantly affect the nucleotide distribution in a sequence. We formulated a method to detect sequence composition biases in the data generated by two different platforms, which efficiently detected sequencing based similarities and differences in the data. PCA analysis and phylogenetic heatmaps provided a compact visual image of the biases. It was found that the bias in the sequence is not only platform-specific, but other processes, like DNA-extraction protocols and experimental framework, also contribute to the differences. Therefore, caution should be exercised when interpreting the results of comparative metagenomics studies.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance and help of several people who, in one way or another, contributed and extended their valuable assistance in the preparation and completion of this study. First and foremost I offer my sincerest gratitude to my advisor, Professor Andrey Grigoriev, who has supported me throughout my dissertation with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my doctorate degree to his encouragement and effort and without him this dissertation, too, would not have been completed or written. Through his mentoring he has given me tremendous amounts of computational and analytical skills. I would like to thank my committee members, Professor Daniel Shain, Professor Roman Dial and Dr. Jognmin Nam for being in my dissertation committee and offering me their wisdom and expertise with utmost kindness and compassion. Their valuable suggestions and inputs guided my research projects, and also broaden my approach to scientific problems. I express my deep sense of gratitude to Professor Roman Dial for providing us with the samples from Alaska without which this study would not have been possible. This experience would not have been as valuable without the guidance, support and inspiration provided by each one of you.

This work was supported by grants and I thank the various sources including the Center for Computational and Integrative Biology at Rutgers-Camden, and the NSF grant DBI-1126052 granted to Professor Andrey Grigoriev.

Additionally, I would like to say thanks to all of the lab members, Sean, Ammar, Spyros, Karl and Joseph for providing a cheerful and friendly atmosphere to work, and Ruchi for her help in my initial analysis. I am also appreciative for the technical support

from Kevin. A special thanks to Geetika for being a wonderful friend and for her constant support and motivation throughout my difficult times. In my each and every problem I found her beside me. She believe in me more than myself.

Words cannot express my appreciation to my parents (Dr. R.S. Choudhari and Mrs. Sudha Choudhari), who have always supported, encouraged and believed in me throughout all of my endeavors. It is your unconditional love, support and patience that has inspired and driven me to tackle challenges head on. Very special thanks goes to my sister (Namrata) and brother (Prateek), who have always been a caring and supportive siblings and a very good friends who are always there for me in times of need, and to my adorable nephew (Neil) and niece (Navika) for bringing a smile to my face when times were tough.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ATP	adenosine-5’-triphosphate
A	adenosine
BLAST	Basic Local Alignment Search Tool
BLASTN	BLAST search using a nucleotide query
BLASTX	BLAST search using a translated nucleotide query
bp	base pairs
C	cytosine
°C	degree celsius
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
e. g.	exempli gratia, for example
et al.	et alii/alia, and others
Fig.	Figure
G	guanine
i. e.	id est, that is
Ion Torrent PGM	Ion Torrent Personal Genome Machine
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
MEGAN 4	MEtaGenome ANalyzer
MG-RAST	Metagenomic Rapid Annotations using Subsystems Technology
NCBI	National Center for Biotechnology Information

NGS	next generation sequencing
OTU	operational taxonomic unit
PCA	principal component analysis
PCR	polymerase chain reaction
pH	power of hydrogen
RDP	Ribosomal Database Project
rRNA	ribosomal ribonucleic acid
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SSU	RNA small subunit RNA
T	thymine
VAMPS	The Visualization and Analysis of Microbial Population Structure

CHAPTER 1

INTRODUCTION

1.1 Glacier retreat

More than 25% of the world's land surface could be classified as a cold environment that is incredibly varied in its nature (Choudhari et al., 2013). Permanently frozen environments, such as glaciers, snow ice, and ice lakes, are a gigantic reservoirs of microbial life, and are the locus of significant microbial activity. Due to extreme conditions at sub-zero temperatures, these glaciers had previously been considered to exist without life, or only as repositories for wind-transported microorganisms trapped in the ice (Cowan and Tow, 2004). However, in spite of extreme low temperatures, these permanently frozen environments have shown to harbor diverse microorganisms. A previous study reported a 9.61×10^{25} total number of bacterial cells in the Antarctic and Greenland glacial ice sheets that represented a reservoir of microbial diversity (Priscu and Christner, 2004). Glacial ice is also regarded as an environment equivalent to certain extraterrestrial cold habitats, such as Mars and Jupiter's moon Europa, and it is presumed that glacial ice is harboring the oldest prokaryotes on Earth (Willerslev et al., 2004). The reason behind organisms surviving in such cold temperatures could be the unique features in their proteins that have evolved with time.

During the past few decades, melting of glaciers across the globe has accelerated

dramatically, and 98% of Alaskan glaciers are shrinking and losing billions of tons of ice each year (Molnia, 2007). Recent changes in climate due to accumulation of anthropogenic sources of CO_2 and other gasses have led to substantial depletion of these glacial reservoirs. Glacier melting has global implications, as it contributes to sea level rise, flooding, global warming, and also gives life to microbial populations that thrive at such low temperatures. Similar to hot springs and other extreme environments, the icy biome, typified by temperatures <5 ºC, is dominated by microbial communities. Photosynthetic microbes use pigments that reduce albedo, thereby potentially increasing snowmelt. Increased snowmelt due to albedo reduction by photosynthetic microbes near the equilibrium line altitude could lead to more bare ice than without microorganisms. Because bare ice melt rates exceed snow covered ice melt rates, these microbes may be hastening glacial melt, thinning, and retreat. The biomass of microbial cells in and beneath the ice sheet may amount to more than 1,000 times the mass of all the humans on Earth, which contradicts the thought that cold environments would be too harsh and inhospitable to support any life. These microbial communities present at low temperatures play a significant role in soil development and other biochemical cycles (Schutte et al., 2010). The potentially important role these microbial communities play in accelerating loss of glacial ice and snow is currently likely underestimated, and the community composition is poorly understood. Thus, studying these microbial communities in Alaska is vital to understanding their key role in the glacier retreat, nutrient and other biochemical cycles.

1.2 Metagenomics- “beyond the genome”

“Once the diversity of the microbial world is catalogued, it will make astronomy look like a pitiful science.” - Julian Davies

This quote appears applicable as up to 99% of the microbes cannot be grown in laboratory cultures, which is a traditional way to study microorganisms, which makes it difficult to identify, classify, and research all microbes present. However, the advancements in the sequencing technology has improved and yielded researchers ways to study these microbial communities quickly and without culturing. When we study any microbial community, there are interactions of species within a community that cannot be possible while studying each single organism's genome separately. Therefore, there has to be a process or a technique that considers whole microbial community as a whole/single, which has been achieved by the scientific community with the advancements of next generation sequencing (NGS) in the form of Metagenomics. Metagenomics is the genomic analysis of microorganisms by direct extraction and sequencing of DNA from a community of organisms inhabiting a common environment (Handelsman et al., 1998). The current estimates of the number of different species found in the world point to between 10 million and 1 billion (Dykhuizen, 1998); such a huge number represents a significant problem to study global microbial diversity. There has been a rapid increase in number of studies focused on environmental microbiology and microbial ecology due to increase amount of data generated by NGS technologies using metagenomics approaches. In turn, this allows us to study whole microbial communities together; a sample can be

obtained from an environment and sequenced directly bypassing the cultivation steps. Metagenomics deliver the direct information as present in the nature: how genes are interacting with each other without any modification that could occur while culturing. This rapidly growing field provides us with information about the true diversity, and functional potential of microbial communities present in environments such as glaciers, soil, water, or microbiome of animals and humans. In the future, such studies will accelerate biotechnology and drug discovery by providing new genes with novel functions. A number of studies on the microbial diversity of frozen environment have been reported using metagenomics approaches (for e.g. Simon et al., 2009; Frank-Fahle et al., 2014). Along with the taxonomic composition, metagenomics can also get the relative abundance of genes, and biochemical and metabolic profiles of the microbial community. Currently, metagenomics is the most widely used quantitative approach for studying the microbial communities in an environment (von Mering et al., 2007; Simon et al., 2009).

1.3 Microbial diversity in glacier ice

"Where there is water, there is life." The first report of microbes found in glacier ice and other frozen environments is traced back in 1918 (Miteva, 2008). The discovery of microbes in Lake Vostok, one of the largest subglacial lakes of Antarctica, has received much attention lately (Price, 2000; Siegert et al., 2001). It was reported by Christner et al. (2003) that the ice contained 2-3x10^2 cells ml^{-1} (Karl et al., 1999). Additionally, isolates

were obtained from a 750,000-year-old ice core from the Guliya ice cap in Tibet (as cited in Miteva, 2008). *Actinobacteria* was found to be dominant phylum in the active layer soil in Arctic permafrost, followed by *Betaproteobacteria* (Yergeau et al., 2010). As reported by Johnson et al. (2007), the non-spore forming *Actinobacteria* dominated ancient permafrost due to their metabolic activity at low temperatures. Cyanobacterial communities were also reported from glacier in China based on 16S phylogeny (Segawa and Takeuchi, 2010). A recent study reported that the rich bacterial diversity of a glacier foreland in high Arctic was comparable to temperate and tropical soils (Schutte et al., 2010). Moreover, many bacterial groups were found in the remote site of Canadian high arctic (Cheng and Fogt, 2007; Skidmore et al., 2005), and 10 bacterial isolates were identified based on 16S gene from snow cover of an Arctic site (Amato et al., 2007).

The most common bacterial phylum found in cold environments are *Proteobacteria (Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria), Bacteroides,* and *Actinobacteria,* but *Archaea* are largely underrepresented (Yergeua et al., 2010; Lewin et al., 2013). In a recent study, while comparing bacterial communities in glacial cryoconite holes from Arctic and Antarctica, many bacterial groups were identified, proving that these extreme environments harbors diverse microbial communities (Cameron, 2012). While comparing the high Arctic snow and freshwater, it was revealed that highest diversity was seen in snow (Møller et al., 2013). Bacterial phylums such as *Proteobacteria, Bacteroidetes, Cyanobacteria, Actinobacteria, Firmicutes* and *Fusobacteria* were relatively abundant in snow while in freshwater,

Bacteroidetes, *Actinobacteria* and *Verrucomicrobia* were mostly found, and relatively few *Proteobacteria* and *Cyanobacteria* were reported (Møller et al., 2012). Furthermore, several studies from Antarctica (Kol, 1968), Greenland (Priscu et al., 1999; Miteva, 2008), and New Zealand (Sheridan et al., 2003) have investigated phylogeny of 16S rRNA genes. The phylogenetic analysis of a European glaciers revealed autotrophic lifestyles of such microbes due to low nutrient conditions (Simon et al., 2009). This clearly indicated that the cold environments like glacier, snow ice, and ice lakes harbor a massive reservoir of microbial life and communities. These specific microbes impact the dynamics of glacial ecosystem, and play a key role in soil formation and other biogeochemical cycles (Cheng and Fogt, 2007).

1.4 Metabolic and functional potential of glacier ice

Microbes play an essential role in biogeochemistry of earth and other system processes (Nazaries et al., 2013). The microbial diversity is high and microbes are ubiquitous in nature, as they possess tremendous metabolic and functional potential that are essential to all biogeochemical cycling processes (Prosser et al., 2007), such as oxygenic photosynthesis, nitrogen fixation, and carbon sequestration (Kasting and Siefert, 2002; Newman and Banfield, 2002). Such functions are not only performed in normal environmental conditions, but microbes survive and carry out the biological functions from everywhere, including even the most extreme environments (Reysenbach and Shock, 2002; Edwards et al., 2013). About 10% of Earth's terrestrial surface is

represented by the glacier ice containing 77% of the fresh water on the earth (Paterson 1994), which is a unique ecosystem preserving microbial life. Biological activity in these low-temperature habitats is generally believed to be restricted. Glaciers can be defined as simple, relatively closed ecosystems sustained by primary producers (e.g., photosynthetic bacteria and algae) in the snow and ice (Choudhari et al., 2013). There is need for better understanding of carbon and nitrogen cycles in these regions that determine total ecosystem carbon storage (Mack et al., 2004) and other functional and metabolic processes responsible for acclimation of these microbes at such low temperatures.

Recently, NGS approaches have been applied to study the functional genes present in a given environmental sample in the form of functional metagenomics. Such studies provide detailed knowledge of the metabolic and functional potential of the microbial community of a habitat, and is regarded as the most accurate quantitative approach at existing (Simon et al., 2009; von Mering et al., 2007).

Several studies report microbial diversity present on glaciers from Asia (Christner et al., 2003), Antartica (Priscu et al., 1999), Greenland (Miteva et al., 2004; Sheridan et al., 2003), Germany (Simon et al., 2009) and New Zealand (Foght et al., 2004). With the metagenomics approach, the relative abundances of all genes and metabolic capacity of the microbial communities of any environment can be determined. Recently, a comprehensive study of phylogenetic diversity and metabolic potential of European glacier has been conducted that reported high metabolic versatility of microbial communities (Simon et al., 2009). Another study performed a metagenomic and

metatrancriptomic analysis on a sub glacial lake from Antarctica (Rogers et al., 2013). Many sequences that encode for genes involved in the nitrogen cycle and carbon fixation were reported. Few methanogens and methanotrophs were detected in high Arctic permafrost during its functional analysis, and a limited number of genes involved in nitrogen fixation and ammonia oxidation were detected. Earlier, it was reported that Arctic terrestrial environments are generally nitrogen limited (e.g., Shaver & Chapin, 1980; Martineau et al., 2010). On the contrary, a recent study of functional diversity in an alpine glacier cryoconite ecosystem reported many genes linked to N, Fe, S and P cycling (Edwards et al., 2013). The genes involved in nitrogen fixation, ammonium oxidation, methane production, and methane oxidation were also reported from permafrost soils of NW Canadian Arctic (Frank-Fahle et al., 2014).

1.5 DNA sequencing and different sequencing technologies

DNA sequencing is the determination of the precise sequence of nucleotides in a sample of DNA. The basic method of DNA sequencing, also commonly known as 1st generation sequencing techniques, include Maxam-Gilbert (Maxam and Gilbert, 1977), and very commonly used Sanger sequencing (Sanger et al., 1977). These sequencing methods involve the insertion of labeled nucleotides on the template DNA during the process of DNA amplification. This is followed by obtaining the locations of the labeled nucleotides, which is done by separating the DNA fragments according to their lengths by electrophoresis in a polyacrylamide gel or capillary electrophoresis. The technology

advanced with years, and 2nd, or NGS techniques, were developed with an idea of “sequencing by synthesis” (Bentley et al., 2008). This technology paralleled the sequencing process, minimizing the need for the fragment-cloning methods and yielding more throughput. The rapid and cost-effective approach of such sequencing technologies was able to sequence large amount of DNA in relatively short amount of time. The fundamental aspects of three main NGS platforms for massively parallel DNA sequencing read production discussed here are used in the comparative metagenomics in this study.

1.5.1 Illumina (Genome Analyzer IIx, HiSeq, Miseq)

Illumina uses “sequencing-by-synthesis” technology that is similar to Sanger sequencing, except Illumina uses modified dNTPs containing a terminator. The fluorescent label terminator blocks further polymerization, and only a single base is added by enzyme to each growing DNA copy strand that can be detected by a camera. After the addition of dNTPs to the strand, the terminators are removed, and the images are recorded. This type of sequencing technology is based on reversible dye-terminator.

1.5.2 Roche/454 pyrosequencing

Another next generation technology is Roche 454, which is a technique based on pyrosequencing chemistry where library fragments are sequenced by synthesis. The short double-stranded DNA fragments are joined with an adaptor at either end with micro-sized

beads. This fragment-bead complex is then mixed with emulsion oil, which amplifies DNA inside water bubbles in an oil solution. Each bubble contains a single initial DNA molecule and a single primer-coated bead that the DNA can attach to and form a clonal colony (emulsion PCR). Basically, the light signal is captured as the product is amplified. Using the light signal from the CCD camera, Roche 454 generates a flowgram. One big advantage of 454 is that the read size is much longer (400 bases) than generated by other 2nd generation sequencing techniques.

1.5.3 Ion Torrent

Ion Torrent uses an entirely new approach of sequencing that enables a direct connection between chemical and digital information. During sequencing, when any of the four bases (A, T, G, and C) are incorporated into a strand of DNA by a polymerase, a hydrogen ion is released, which is detected by ion sensor. The technology uses an electrochemical detection system that detects hydrogen ions or protons as they are released by DNA polymerase during sequencing by DNA synthesis each time a nucleotide triphosphate is added. The proton release causes a slight pH shift that is detected by an ion sensor.

1.6 Sequence composition biases in metagenomes

SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and GS FLX Titanium/GS Junior from Roche typically

represent NGS systems (Liu et al., 2012). These platforms have been applied to metagenomics studies for the characterization of microbial communities in diverse environments, and have become the most widely used quantitative approaches for studying the uncultured microbes (Choudhari et al., 2014). One of the primary applications of metagenomics is 16S rRNA surveys, which are used to study the phylogeny and taxonomy of samples from complex microbial communities. Sequencing technologies utilize the hypervariable regions of 16S rRNA for the identification of different bacterial species. There are nine hypervariable regions (V1–V9) in 16S rRNA that exhibit sequence diversity in different species (Shah et al., 2011).

Although metagenomics and high-throughput sequencing have expanded our knowledge in the field of microbiology by directly accessing the microbial community genomes, there are still uncertainties in the data. For example, these different sequencing technologies introduce different biases in metagenomics data, which significantly affects the nucleotide distribution in a sequence, and thus interpretation of information from data. Sequencing technologies are vulnerable to multiple sources of bias, including protocol used for DNA isolation and library preparation, along with PCR amplification step before sequencing. PCR amplification prior to sequencing-by-synthesis methods introduces coverage bias related to GC content (Dohm et al., 2008; Aird et al., 2011; Sims et al., 2014). Moreover, DNA isolation and fragmentation steps can also introduce some kind of systematic bias (van Heesch et al., 2013). Various laboratory protocols have also been developed to reduce such kind of biases (Kozarewa et al., 2009), and current efforts

have been made to minimize such biases during downstream computational analyses in various NGS applications (Cheung et al., 2011; Davey et al., 2013; Wolf and Bryk, 2011; Szatkiewicz et al., 2013). Studies have also shown that the sequencing technologies have different biases depending on the approach adopted to obtain sequence data.

There are many challenges in analyzing metagenomic data generated by these platforms, such as the assessment of microbial abundance in environmental samples, which is based on the frequency of occurrence of an organism's DNA observed in sequencing reads (Morgan et al., 2010). In a recent study, the data generated by Illumina MiSeq, and Ion Torrent PGM platforms were compared for bacterial community profiling (Salipante et al., 2014). It has been shown that the relative frequencies of organisms depend significantly on the DNA extraction and sequencing protocol used. A few studies have investigated the bias imposed by various DNA extraction protocols using environmental samples (Morgan et al., 2010; Abusleme et al., 2014); moreover, comparison of two next generation technologies based on phylogenetic profiling of reads derived from sequencing has been reported (Salipante et al., 2014; Claesson et al., 2010). The differences in DNA sequencing protocol may introduce biases in the resulting sequences, as highlighted by a recent comparison of Illumina and Roche 454 platforms in an analysis of a freshwater planktonic community (Luo et al., 2012). The study revealed bias towards A’s and T’s over C’s and G’s in homopolymers sequence generated by Roche 454. With Illumina, these patterns were less common, and the errors were more randomly distributed. The evaluation of base-call error, frameshift frequency, and contig

length were also compared. The sequences produced by different platforms introduce systematic biases and unique patterns of biased sequence coverage (Harismendy et al., 2009). Furthermore, the approaches used for the taxonomic classification of metagenomic reads also introduced their own limitations (Mavromatis et al., 2007). However, a sequence composition-based comparison between different sequencing platforms has not been done yet.

1.7 Objectives

The goal of this study was to assess the taxonomic composition, relative gene abundance, and functional repertoire of the microbial communities present in the glacial ice of Alaska. Additionally, comparative metagenomics study was done to study the differences and similarities in the sequences generated by different sequencing technologies.

Specific objectives include the following:

1. **Phylogenetic analysis of the microbial community present in the glacier ice of Byron glacier.** Using next-generation sequencing, we performed a metagenomics study using 16S rRNA sequences for evaluating different bacterial and archaeal groups present in the glacier samples. This method enhanced our understanding of the microbial community structure of the ecosystem, and gave us information about enormous microbial diversity.
2. **Metabolic and functional competency of microbial assemblage present in a**

low temperature environment. We compared the metagenome data from glacier samples generated by MiSeq platform to databases, which integrate genomic, chemical, and functional information such as KEGG. We described the metabolic and functional potential of the microbial communities by analyzing their gene complements in the metagenomic data by comparing it with different environmental metagenomes.

3. **Sequence composition diversity in different metagenomes.** Comparative analysis of various metagenomes generated by different sequencing platforms was done based on di-,tri-,and tetra- nucleotide word frequencies using PCA. The similarities and differences were displayed using as different clusters of bacterial groups and phylogenetic heatmaps.

CHAPTER 2

COMPARATIVE METAGENOME ANALYSIS OF AN ALASKAN GLACIER

This work has been published as follows:

- Choudhari, S., S. Smith, S. Owens, J. A. Gilbert, D. H. Shain, R. J. Dial, and A. Grigoriev. 2013. Metagenome sequencing of prokaryotic microbiota collected from byron glacier, alaska. Genome Announcements 1 (2) (Mar 21): e0009913-13.

- Choudhari, S., R. Lohia, and A. Grigoriev. 2014. Comparative metagenome analysis of an alaskan glacier. Journal of Bioinformatics and Computational Biology 12 (2) (Apr): 1441003.

-

My contributions to this work include:

Conception and design of experiment, analysis of the data, and preparation of all the figures and/or tables except Fig. 2.3.

2.1 Introduction

In the present study, we have assessed the taxonomic composition of prokaryotic microbial communities present in the glacial ice of Alaska with the help of metagenomic analyses of environmental samples. To our knowledge, a taxonomic analysis of microbes present in a glacial habitat from Alaska has not been conducted so far. The microbial diversity of subglacial waters and sediment-laden ice based on 16S rRNA

phylogeny has been studied from Canadian high Arctic (Cheng and Fogt, 2007; Skidmore et al., 2005) but no Alaskan microbiome has yet been reported.

Previous work indicated a fairly small community in the glaciers. For example, 14 bacterial isolates were recovered from one of the oldest glacial ices in Tibet and classified based on 16S rDNA sequences (Christner et al., 2003). An early report listed 354 algal and cyanobacterial species, 77 fungal species, and 35 bacterial species that occur in snow (Kol, 1968). A direct metagenomic analysis of a glacier in the German Alps identified 72 bacterial operational taxonomic units (OTUs) at 97% identity cut-off (Simon et al., 2009). Several studies of microbial communities from Antarctica (Priscu et al., 1999), Greenland (Miteva, 2008; Sheridan et al., 2003), and New Zealand (Foght et al., 2004) have been performed based on phylogeny of 16S rRNA genes, and the bacterial diversity in a glacier foreland of high Arctic has also been sampled recently (Schütte et al., 2010). However, a metagenomic analysis of a glacial habitat from Alaska has not been conducted, and the present study allows for comparing and expanding the findings of the studies mentioned above. We compared the taxonomic distribution of our data with other metagenomic data available from high Arctic lake and soil and also from another previously studied Alpine glacier (Simon et al., 2009). Comparing these numbers to many metagenomics projects currently underway, which suffer from high complexity of the studied environments, it seems feasible to characterize and model the glacial ecosystem in its entirety (in contrast to, say, much richer soil samples).

2.2 Results

2.2.1 Phylogenetic distribution of the microbial community

The 7,728 unique sequences were used for the phylogenetic analysis that provided the identification of various groups. The percent distribution of phylogenetic groups of 16S rRNA sequences from the Alaskan glacier ice was generated (Table 1). These 16S rRNA gene sequences have been used to study the bacterial taxonomy and diversity present in the glacial ice. 16S rRNA are by far the most common housekeeping genetic marker used because of its omnipresence in bacteria and its variable V4 region has been found to be well-suited for metagenomic analyses (Ghyselinck et al., 2013). The sequences were assigned to the taxonomy outline using the MOTHUR (Schloss et al., 2009) software. Database sequence and taxonomy files from the SILVA (Pruesse et al., 2007) database for the reference sequences were utilized to assign the taxonomy. Taxonomy outlines and reference sequences were obtained for bacterial references (14,956 sequences), archaeal references (2,297 sequences), and eukaryotic references (1,238 sequences) from SILVA database.

Table 2.1. Percentage distribution of phylogenetic groups based on 16S rRNA sequences obtained from Byron glacier in Alaska and Northern Schneeferner in Germany. The taxonomic classification was based on a k-nearest neighbor consensus and Wang approach using the MOTHUR (Schloss et al., 2009) software. The groups accounting for less than 0.6% are classified under the artificial group, “Others".

Phylogenetic Group	Byron	Northern Schneeferner
Acidobacteria	2.99	0.26
Actinobacteria	9.25	9.52
Archaea	0.61	Not detected
Bacteroidetes	22.26	29.1
Cyanobacteria	4.68	1.06
Firmicutes	12.25	Not detected
Fusobacteria	0.7	Not detected
Planctomycetes	1.75	Not detected
Proteobacteria	40.55	57.67
Verrucomicrobia	2.72	0.26
Others	2.24	2.12

Our results showed that *Proteobacteria* is the most dominant phylum covering 40% of the population diversity present in the glacier; its dominance is also recorded from a previously studied Northern Schneeferner glacier in Germany (Simon et al., 2009). *Proteobacteria* account for more than 40% of the OTUs in the bacterial community in these two glaciers as it is one of the largest phyla in the domain *Bacteria* and is also one of the most widespread groups in the environment with 200 genera (Lee et al., 2005).

The next dominant group was *Bacteriodetes* (~20–30%) in both glaciers. *Firmicutes*, mostly gram-positive bacteria producing endospores that can survive in extreme conditions, was also one of the most common groups in our samples. However, this group was not reported from the glacier in Germany. When we examined a non-16S-based data set of pyrosequencing-derived genomic sequence fragments from Northern Schneeferner belonging to 468 distinct phylogenetic groups, we detected just three sequences corresponding to *Firmicutes* confirming significant underrepresentation of this group in the German glacier.

The group *Actinobacteria* also contributed to a major portion of the bacterial community present in the ice of both the glaciers. The nitrogen fixing actinobacteria belonging to genus *Frankia* is one of the most dominant *Acitnobacteria* (found in our data) that contributes to global nitrogen fixation (Weiss, 1983). Photosynthetic microorganisms like *Cyanobacteria*, some of which are cold-tolerant species that adapt their lives to the special conditions of glaciers, are commonly observed in most of the glaciers in the world.

The bacterial group *Verrucomicrobia* was found on both glaciers at different levels, however, it was more dominant in the Alaskan glacier (about 3% of the metagenome compared to only one species in the Northern Schneeferner using 16S rRNA gene). The reason for the low percentage distribution could be different sampling and sequencing techniques in the case of the European glacier as compared to the Alaskan

glacier. *Verrucomicrobia* were first observed in aquatic habitats but are now known to exist in many other habitats. They are found not only at moderate temperatures but also at cold temperatures in the deep sea and in Antarctica (Schlesner et al., 2006).

Two bacterial groups *Planctomycetes* and *Fusobacteria* were found in the present study, though at very low percentage distribution, were not reported from the glacier study in Germany using 16S rRNA. However, *Planctomycetes* from the same study were found when comparing pyrosequencing-derived sequences with the RDP II database (Cole et al., 2009).

About 41 sequences showed more than 90% sequence identity alignment to Archaeal 16S rRNA, with sequence similarity (> 98%) to other *Archaea* mostly from wet and cold regions (Table 2). As a control, no archaeal hits were found to 16S rRNA from several human microbiome samples. Rarefaction curve for assessment of microbial richness from results of sampling glacier demonstrate the number of OTUs observed as a function of the number of sequences sampled (Fig. 2.1). The curve reaches saturation at 20% distance level (phylum level) but not at 3% distance (species level). The steep slope on the left indicates that a large fraction of the species diversity remains to be discovered but they will belong to only a few additional phyla (Gotelli and Colwell, 2001). Thus, our current estimate of ~2,500 species is only a lower bound of the species richness in the Byron glacier metagenome.

We also attempted to estimate the relative cellular abundance of individual species

in our sample. Since species-specific identification was not possible for most of the reads, we employed the following indirect method. For each taxonomic group (e.g. *Proteobacteria*), we normalized the total number of reads (rather than non-redundant unique reads) belonging to that group by the number of OTUs found in it, and further normalized that by the mean number of rRNA clusters found in genomes of that group (Klappenbach et al., 2001). Note that such estimate does not take into account possible PCR biases. The results (Fig. 2.2) indicate that the most abundant species in terms of cellular counts inferred from sequence counts were not the representatives of *Proteobacteria*, but those of *Cyanobacteria, Actinobacteria and Planctomycetes.* This likely reflects the dependence of the ecosystem on the energy obtained through photosynthesis and close links with the microbial community of the soil.

Table 2.2. A subset of nearest neighbors of sequences (identity above 98.7% across complete read length) belonging to *Archaea* domain determined by a BLASTN search against the non-redundant database confirming that *Archaea* found in the datasets are not artifacts but show good similarity with *Archaea* found in wet and cold regions.

Name of sequence	**Closest GenBank match; Acc.No.**	**Region of Isolation**
6825-1,7266-1	DQ004709.1	Arable Europe soil
3840-1	EF022335.1	Trembling aspen soil
3020-1	EF022414.1	Trembling aspen soil
1682-2	EF023086.1	Trembling aspen soil
4020-1,4232-1	HQ268987.1	Potato field, wet anoxic conditions
2429-1	HQ269055.1	Soil crust, wet anoxic conditions, Negev Desert
5737-1,6425-1	JN002507.1	Low temperature, serpentinized dunite
3852-1	JN002684.1	Low temperature, serpentinized dunite
2569-1	JN205388.1	Iron-ore mines
2653-1	JN221262.1	Ice wedge core
4180-1	JN617429.1	Lake Taihu sediment
3948-1	JN863129.1	Iron-ore mines
3925-1	JN900444.1	Cinnamon soil

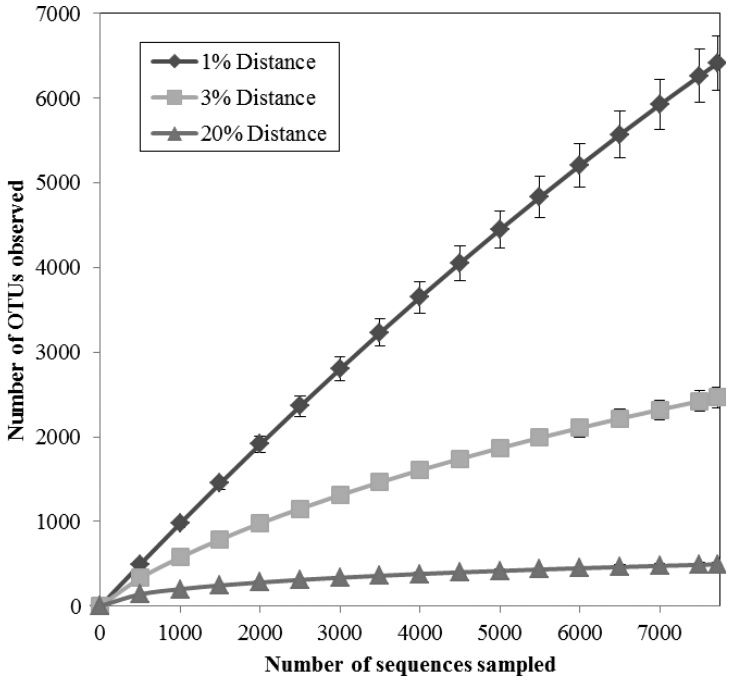


Fig. 2.1. Rarefaction curves representing the number of OTUs generated from 16S rRNA sequences sequenced from glacial ice. The curve was calculated using MOTHUR. OTUs are shown at genetic distances 1%, 3% and 20%. Error bars represent the 95% confidence interval.

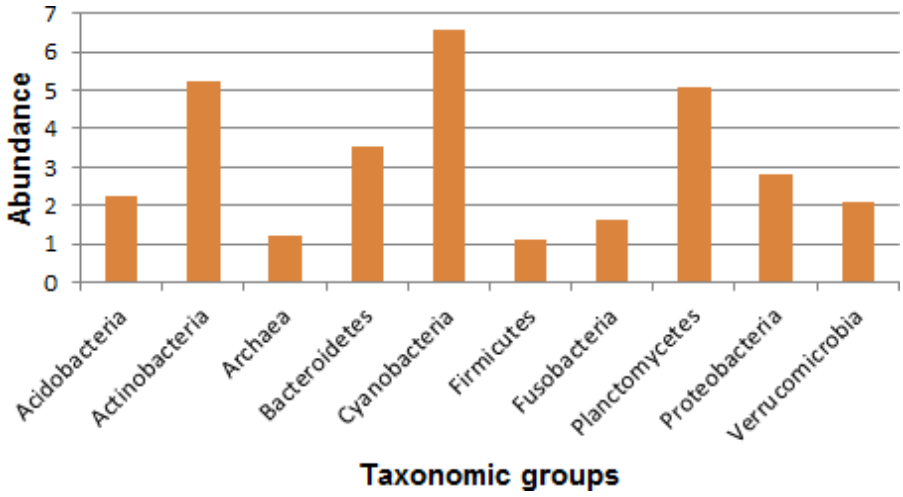


Fig. 2.2. Estimate of abundance of community members.

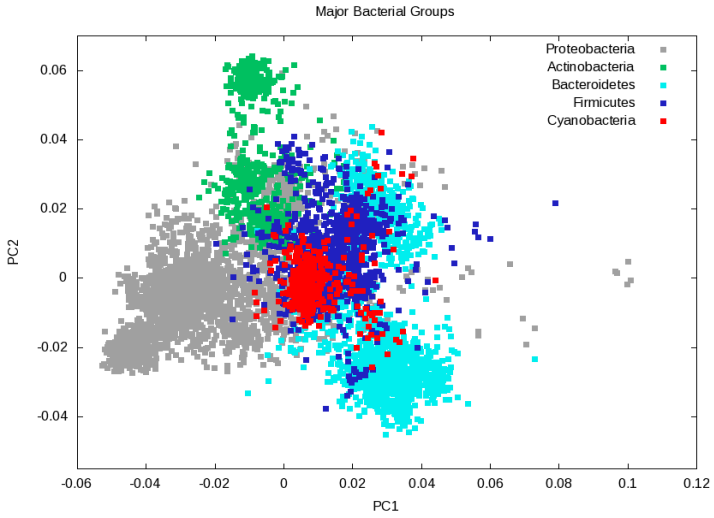
2.2.2 Nucleotide composition analysis and comparison with another glacier

We utilized PCA for finding nucleotide composition patterns in the sequenced reads and for comparisons with other metagenomes. Nucleotide word frequencies were derived from the sequences corresponding to the 16S rRNA V4 regions (151 bp), which were already pre-classified taxonomically in different groups as described above. We analyzed how taxonomic classification relates to the clustering/grouping of sequences on the basis of their nucleotide frequency (Fig. 2.3).

Fig. 2.3(a) shows the dominant groups present in glacial ice. All phyla form different clusters based on nucleotide word frequencies in their sequences. The largest cluster for the most dominant group, *Proteobacteria*, as well as the clusters for *Cyanobacteria* and *Firmicutes*, form relatively tight knots, while other groups show wider spread. Such differences are suggestive of lower conservation of sequence composition and may also indicate a certain degree of misclassification of the outliers. The comparison was also done with a European glacier (Simon et al., 2009); the 16S rRNA sequences from the Northern Schneenfener were plotted together with the Byron glacier from Alaska (Fig. 2.3(b)). Both data sets formed separate clusters suggesting that the microbial species present on the two geographically different sites have different sequence composition. The Northern Schneenfener study included full 16S sequences, thus this comparison was done using only the V4 hypervariable regions of 16S rRNA in both cases in order to avoid bias.

Our results show that there is a difference in the nucleotide word frequency between the two samples collected from different locations and they are more distant than samples collected from the same location. The same groups of bacteria as in Alaska were reported in the Northern Schneenfener glacier, and we expected them to be mixed together on the PCA plot. However, all sequences from the Northern Schneenfener clustered together in a small group, very distinct from the clusters formed by the Byron glacier metagenome. One reason may be that there is more diversity in the Byron (temperate) glacier samples. Another possible explanation for the observed composition difference could be different sampling and sequencing techniques used (plus biases in PCR, cloning, etc.). The samples were collected from different depth in the two cases, for the Byron glacier, the sample was collected at a depth of 2 m while for Northern Schneenfener glacier ice was collected from 0.5 m depth. Additionally, the first 30 cm of glacial core exposed to the glacier was removed and discarded in case of Northern Schneenfener. This could have resulted in the elimination of those extra groups that were detected in Byron glacier but not present in Northern Schneenfener (Table 2.1). The sequencing technique used also could have contributed to the biases in species representation. We used Illumina sequencing for 16S rRNA V4 amplicon analysis unlike the traditional methods involving PCR amplification of 16S rRNA genes and construction of clone libraries for sequencing as used in the Northern Schneenfener study.

a



b

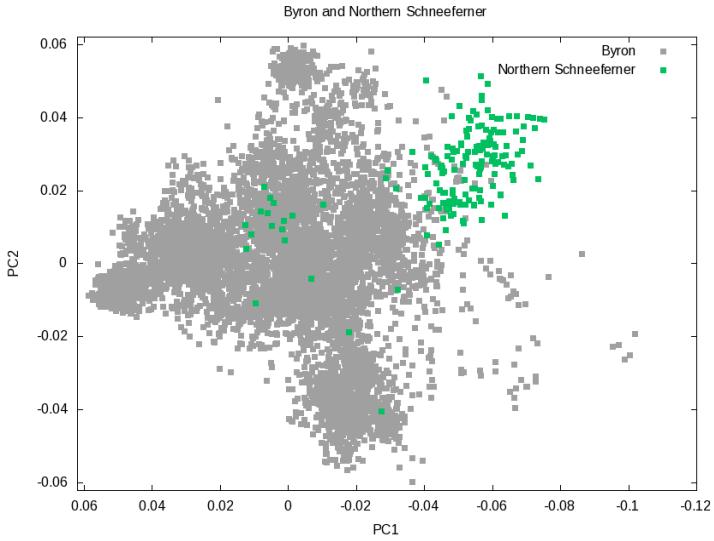


Fig. 2.3. Nucleotide word frequency principal component analysis (PCA) of the 16S rRNA sequences from samples collected from glaciers in Alaska and Germany. a. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = green, *Bacteroidetes* = turquoise, *Cyanobacteria* = red, *Firmicutes* = blue, *Proteobacteria* = grey), b. Metagenome sequence colored by location (Northern Scheenefener=green, Byron = grey). The first component, PC1 represents 58.65 % of the variance and second component i.e. PC2 represents 40.70% variance. The two components thus yield a model that covers 99.4% of variance.

2.2.3 Comparison of glacier and non-glacial environment

We compared the species distribution of the Byron glacier with other related ecosystems. The sequence data from high Arctic lake (Møller et al., 2013) and soil (Schütte et al., 2010) were taxonomically classified into different taxonomic groups using MOTHUR (Schloss et al., 2009) as described previously for the glacier data set. We compared the distribution of different taxonomic groups in three different habitats including glacier and visualized the relative diversity contribution of each group using a triangle plot (Fig. 2.4), where each apex represents a particular habitat.

Taxonomic analysis of all the data revealed that *Bacteroidetes* was present in all of the three habitats. Less abundant phyla *Firmicutes* and *Fusobacteria* were mainly reported from glacier while they were rare or not detected in soil and water. *Planctomycetes*, and *Verrucomicrobia*, also rare in most of the environment, were contributing to diversity mostly in high Arctic lake. *Archaea* showed predominance in water compared to glacier and was not a prominent contributor in soil. *Acidobacteria* and *Proteobacteria* were dominant in glacier and soil compared to lake, as was *Cyanobacteria*. *Actinobacteria* was reported from all of the three habitats, with highest relative diversity in Arctic lake.

We utilized the alignment results of these three microbial communities with the SILVA (Pruesse et al., 2007) database using MOTHUR (Schloss et al., 2009) to compare the species overlap between the different environments. The alignments with the

reference sequences showed that the soil metagenome is more diverse compared to freshwater and glacier with the highest number of soil-specific species, and freshwater has the least microbial diversity among the three (Fig. 2.5). A previous study has shown (Cole et al., 2009) that nitrogen and water content are the main factors affecting the microbial population. The high microbial abundance in glacier and soil can be related to the availability of organic matter and inorganic nutrients (with allochthonous material possibly better retained in a solid glacial environment compared to water of the ice-covered lake).

A large number of common species (442) found between the soil and glacier samples indicates that the glacier ecology and microbial community is closer to soil than to lake water and suggests species exchange. On the one hand, soil dust particles are likely to be deposited in the glacier snow by the wind. On the other hand, glacier microbiota stays in the soil following glacier retreat and thus is likely to strongly affect soil development.

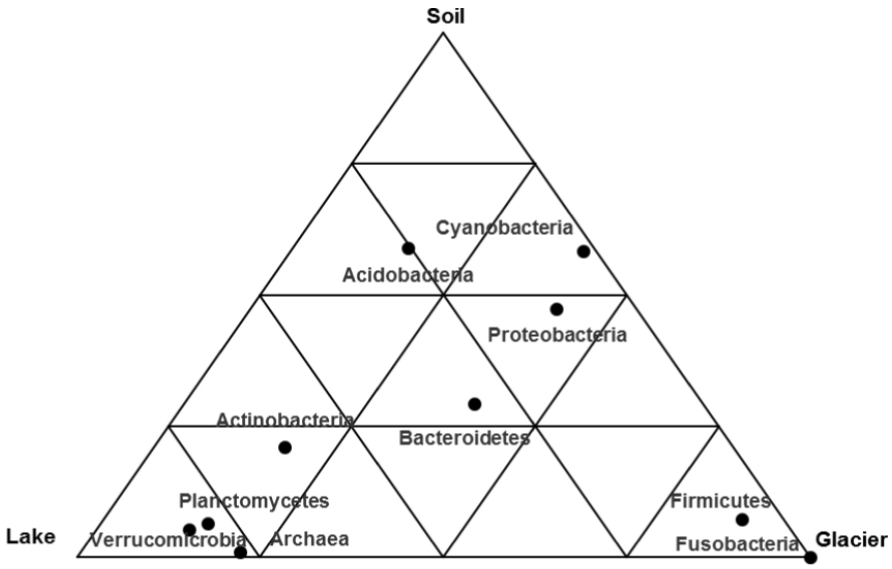


Fig. 2.4. Distribution of taxonomic groups in Byron glacier, high arctic lake and arctic soil. The percentage of various taxonomic groups associated with each habitat is pictured in a triangle plot. The location in the triangle indicates the relative abundance of each phylum among the three habitats.

The species that were common between the glacier and the freshwater mostly consisted of uncultured bacteria belonging to various groups like *Acidobacteria*, *Proteobacteria*, and *Actinobacteria*. The 152 common species between lake and soil generally belong to *Proteobacteria*. The other groups which were common between lake and soil were from *Actinobacteria*, *Acidobacteria*, and *Veruucomicrobia*. Only 81 species, largely belonging to the *Proteobacteria* group, were common between the glacier, lake, and soil. Interestingly, *Proteobacteria* was the most dominant group found in the different samples belonging to different microbial communities. The glacier and lake (Møller et al., 2013)

share common species of *Archaea*, however, no common Archaeal species were identified between Byron glacier and foreland soil study (but see Table 2 for example of common Archaeal sequences with other soil samples).

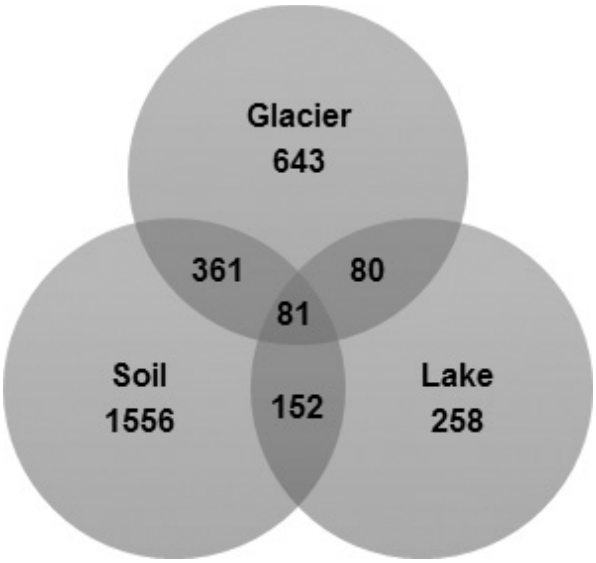


Fig. 2.5. Euler diagram for three habitats (Byron glacier, high arctic lake and arctic soil) from arctic regions. The numbers indicate the total number of species found in each microbiome. The numbers in the intersections are proportional to the number of shared species among the three microbial communities.

CHAPTER 3

FUNCTIONAL AND METABOLIC POTENTIAL OF GLACIAL ICE METAGENOME

My contributions to this work include:

Conception and design of experiment, analysis of the data, and preparation of all the figures and/or tables.

3.1 Introduction

Earlier studies have primarily focused on the taxonomic composition of the microbial community; however, more recent metagenomic studies aim to characterize the overall functional and metabolic profile of such communities. Different studies from cold habitats describe the functional aspects of microbial communities, and studies have shown that Arctic regions are nitrogen limited (Shaver, 1980), and active methanotrophic bacterial populations have been reported from Canadian high Arctic with low methane oxidation capacity (Martineau et al., 2010). These findings complement our results from phylogenetic study, where we found the presence of the group *Methanomicrobia*, belonging to *Archaea*, in our samples. Moreover, the metabolic capacity of glacial ice of the Northern Schneeferner has been examined previously, which exhibited dominance of

aerobic and facultative aerobic bacteria (Simon et al., 2009). A metagenomic snapshot of the functional potential of alpine glacier cryoconite was done that detected the presence of N, S, Fe and P cycling genes (Edwards et al., 2013), and the functional potential of permafrost-affected soils has also been studied (Yergeau et al., 2010). However, there is no current record of such studies from Alaskan glacier. Such an analysis would enlighten our knowledge of microbial life in frozen environments, and their adaptation to survive in extreme low temperatures.

Nonetheless, there have been several studies that have led to the successful discovery of complementary metabolic pathways from microbes that constitute a community. In this study, we analyzed the functional and metabolic potential of microbial communities present in snow/ice of Harding Icefield in Alaska. MEGAN (Huson et al., 2007) was used to compare the given reads against a KEGG (Kanehisa and Goto, 2000) classification reference, BLASTX (Altschul et al., 1990) search was performed against the non-redundant NCBI database (Sayers et al., 2008; Benson et al., 2009). Such sequence-based characterization allowed us to utilize the genome databases to parse the complexity of the complete ecosystem of a glacier. A comparative analysis of the different metabolic pathways of diverse metagenome, along with that of our glacier, was performed using MG-RAST (Meyer et al., 2008). The details of different metagenomes are described in the Methods.

3.2 Results

3.2.1 Analysis of the metabolic potential encoded by Alaskan glacier

Increasing temperatures in Arctic regions is likely increasing the microbial life in the glaciers of these regions. In the past, several studies have addressed the functional capacity of microbial assemblages at low temperature environment. In this study, we first time analyze the functional and metabolic potential of microbial communities living in Alaskan glacier. Here, we used a comparative metagenomic approach to analyze the frequency distribution of five different metagenomic sequences to elucidate the functional potential of biomes that include: (1) Alaskan glacier (our sample); (2) alpine glacier cryoconite from Austria (Edwards et al., 2013); (3) freshwater from a river in China; (4) marine water from the Mediterranean sea; and (5) grassland from Montana.

The sequences from the Alaskan glacier metagenome were compared to the KEGG platform (Kanehisa and Goto, 2000) employing MEGAN 4 (Kanehisa and Goto, 2000) and using the BLASTX algorithm (Altschul et al., 1990), and hits with an E-value of, $1e^{-05}$ were considered to be significant (Methods). A total of 54,252 sequences were significantly similar to functional genes within the KEGG (Fig. 3.1). The analysis of other four metagenomes from different biomes for comparative purposes were performed using MG-RAST database utilizing KEGG orthology. The breakdown of different KEGG functional categories from different metagenomes are displayed in the form of pie charts in Fig. 3.2.

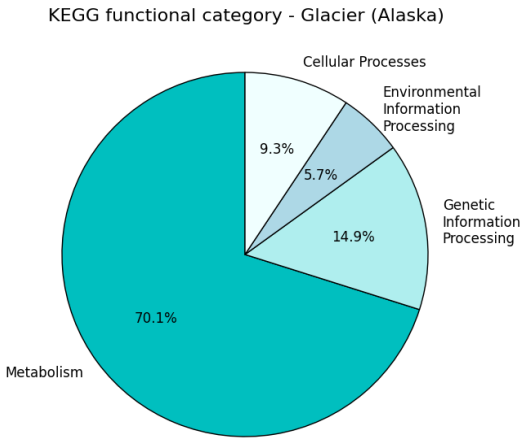


Fig. 3.1. Distribution of functional category metabolism in KEGG database in Alaskan glacier metagenome determined using the BLASTX and MEGAN 4. A total of 56% of the sequencing-derived data set (54,252) were functionally classified in glacial ice metagenome. Shown are the percentages of all classified sequences.

The sequences from the Alaskan glacier metagenome were compared to the KEGG platform (Kanehisa and Goto, 2000) employing MEGAN 4 (Kanehisa and Goto, 2000) and using the BLASTX algorithm (Altschul et al., 1990), and hits with an E-value of, $1e^{-05}$ were considered to be significant (Methods). A total of 54,252 sequences were significantly similar to functional genes within the KEGG (Fig. 3.1). The analysis of other four metagenomes from different biomes for comparative purposes were performed using MG-RAST database utilizing KEGG orthology. The breakdown of different KEGG functional categories from different metagenomes are displayed in the form of pie charts in Fig. 3.2.

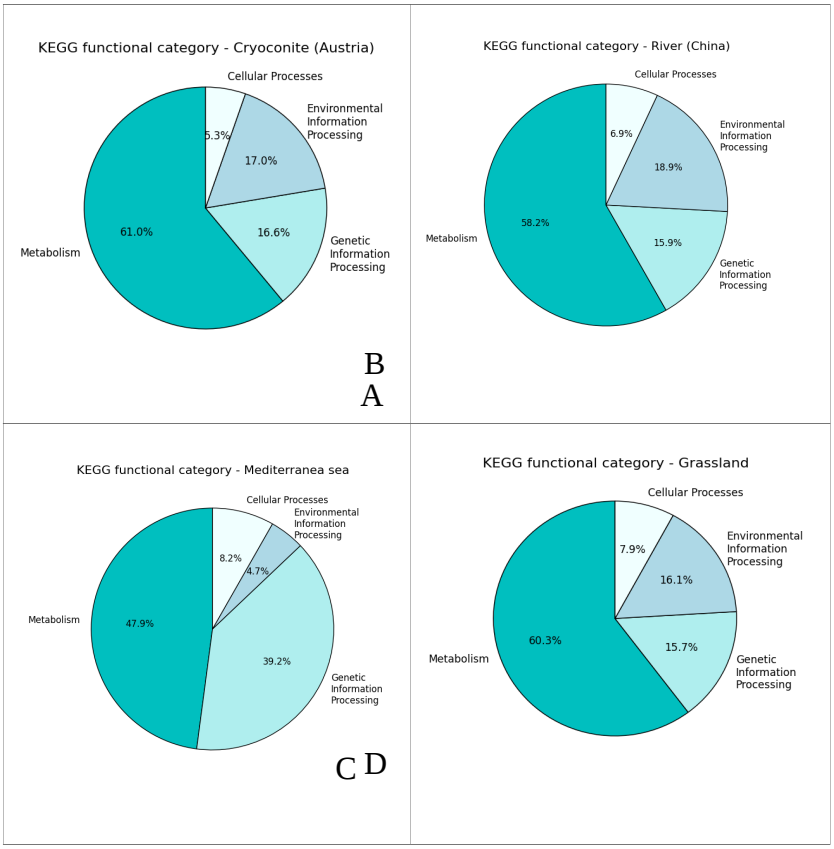


Fig. 3.2. Distribution of functional category metabolism in KEGG database in different metagenomes determined using the BLASTX and MEGAN 4. Shown are the percentages of all classified sequences. (A = alpine glacier cryoconite from Austria; B = freshwater from a river in China; C = marine water from the Mediterranean Sea; D = grassland from Montana).

The metabolism category of the KEGG database was represented by around 70% of all assigned reads for our sample. The other categories, such as environmental information processing, included 5.7%, genetic information processing yielded 15.8%, and cellular processes had 9.3% of all assigned sequences (Fig. 3.1). The results were comparable

with distribution from four other biomes, where the metabolism was taking the highest distribution (Fig. 3.2).

3.2.2 KEGG orthology

Several sequences fall into more than one subcategory within the KEGG categories, 30, 620 sequences occupied the metabolism category. Fig. 3.3 showed the different kinds of metabolism occurring in the microbial community of a glacier compared with four other metagenomes from different environmental niches. The majority of reads were related to carbohydrate and amino acid metabolism in all the five biomes, and the fraction of genes in carbohydrate, energy, lipid, and nucleotide metabolism was higher in our sample (i.e., glacial ice metagenome from Alaska) than in other metagenomes. The other functional category, xenobiotics biodegradation and metabolism, was also found higher in our samples than in metagenomes from other biomes with different temperature settings. The aforementioned breakdown of metabolic category indicated a high metabolic versatility of microbes in order to survive at different environmental conditions.

3.2.3 Carbohydrate metabolism

Here, we described the breakdown of the carbohydrate metabolism category into sub-categories from our samples only. In the sub-category of carbohydrate metabolism, most of the sequences shared homologies with the genes involved in glycolysis/gluconeogensis

(1188), citrate cycle (875), pyruvate (808), amino acid and nucleotide sugar (754), propanoate (683), and glyoxylate and dicarboxylate (642); (Table 1). A comparative analysis with metagenomes from the Austrian cryoconite, a river from China, Mediterranean sea, and grassland soil revealed similar trend in the functional categories in all the metagenomes. Moreover, the fraction of genes involved in processes, such as starch and sucrose (790), ascorbate and aldarate (176), propanoate (683), and butanoate (602) in carbohydrate metabolism were higher in glacial metagenome.

3.2.4 Energy metabolism

The functional category of energy metabolism accounted for 11.31% of the assigned reads. The sequences that occupied energy functions fall into the category of oxidative phosphorylation (774), carbon fixation pathways in prokaryotes (584), carbon fixation pathways in photosynthetic organisms (386), photosynthesis (217), nitrogen (398), methane (764), and sulfur metabolism (339); (Table 1). Moreover, only a few sequences from photosynthesis (217) showed similarities to genes encoding components of photosystems II, while most sequences were related to photosystems I and ATPase-like genes. The results were in accordance with previous findings, where only few genes involved in photosynthesis were observed in glacial ice metagenome of Alpine glacier (Simon et al., 2009). The reason for the low number of photosynthetic genes might be due to the fact that the ice core is covered by snow most of the year, and reachability of light is not sufficient.

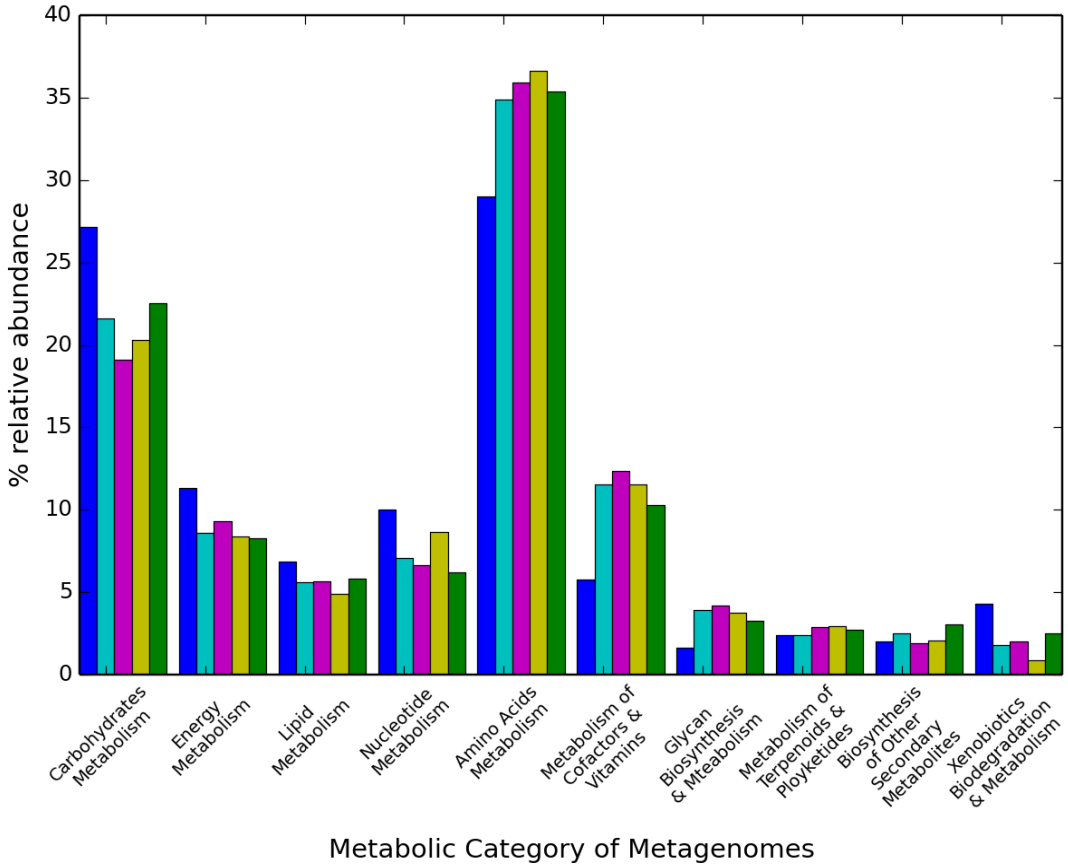


Fig. 3.3. Distribution of breakdown of different metabolic categories of KEGG database in different metagenomes. The relative abundance of genes falling in different metabolic categories is shown on the y-axis, and name of different categories on x-axis. The analysis of our samples (glacial ice metagenome from Alaska - blue) was determined using the BLASTX and MEGAN 4. The results from MG-RAST database were used for four different biomes used for the comparison in the study (alpine glacier cryoconite from Austria = cyan; freshwater from a river in China = magenta; marine water from the Mediterranean Sea = yellow; and grassland from Montana = green).

3.2.5 Lithotrophic metabolism

Lithotrophic bacteria uses an inorganic compound as a source of energy, and are mostly aerobic. The process of respiration is similar to an aerobic; instead, electrons are removed from a substrate and fed through an electron transport system to produce ATP by electron transport phosphorylation. There is still the electron transport chain involved, but instead of oxygen as the terminal electron acceptor, molecules such as nitrate (NO_3^-), sulfate (SO_4^{--}), sulfur (S), and methane (CH_4) are used as electron acceptors. The metabolism of inorganic nitrogen compounds plays many important physiological roles in microbes, and a number of genes in our samples were found encoding genes for methane metabolism (764). In our previous findings, we reported the presence of the group *Methanomicrobia* belonging to *Archaea* in our samples, and other studies have also reported that methane is accumulated in permafrost in Alaska and subsequently consumed by methanotrophic bacteria using CH_4 oxidation (Mackelprang et al., 2011). Few aerobic methanotrophs species were also reported from a lake in Washington (Beck et al., 2013), and these species were consuming C1 substrates, such as methylamine, methanol, and methane (Kalyuzhnaya et al., 2008). Such reports supported the finding of high number of sequences showing having similarities with genes involved in methane metabolism. Some bacteria use nitrate as a respiratory electron acceptor instead of oxygen, through a process called denitrification, in which nitrate is converted into nitrous oxide or dinitrogen gas. The genes (398) involved in dissimilatory nitrate reduction, nitrate to ammonia, alanine, aspartate, and glutamate metabolism were detected in

glacier. The presence of genes involved in nitrogen metabolism indicated that, in addition to organic sources, microbes living at low temperatures use inorganic nitrogen sources during respiration. Moreover, sequences associated with sulfur metabolism (339) were found in the ice metagenome that was related to assimilatory sulfate reduction, reducing sulfate to H_2S. This indicated the presence of bacteria using sulfur as a substrate, and the presence of sulfur in ice, which was also reported previously in glacial ice (Bottrell and Tranter, 2002). The nitrogen and sulfur metabolism was found higher in glacier compared to other metagenomes, indicating the presence of more sulfur reducing bacteria in the glacier compared to ice, or a higher amount of sulfur found in the glacier. The abundance of genes involved in methane (764) was higher than other inorganic compound metabolism.

Table 3.1. Number of sequences showing homologies to genes associated with KEGG pathways in the categories “carbohydrate metabolism” and “energy metabolism”

KEGG category	No. of matches
Carbohydrate metabolism	8317
Glycolysis/gluconeogenesis	1188
Citrate cycle (tricarboxylic acid cycle)	875
Pentose phosphate pathway	610
Inositol phosphate metabolism	204
Pentose and glucuronate interconversions	215
Fructose and mannose metabolism	390
Galactose metabolism	286
Ascorbate and aldarate metabolism	176
Starch and sucrose metabolism	790
Amino sugar and Nucleotide sugar metabolism	754
Pyruvate metabolism	808
Glyoxylate and dicarboxylate metabolism	642
Propanoate metabolism	683
Butanoate metabolism	602
C5-branched dibasic acid metabolism	94
Energy metabolism	3463
Oxidative phosphorylation	774
Photosynthesis	217
Carbon fixation pathways in prokaryotes	584
Carbon fixation in photosynthetic organinsms	386
Methane metabolism	764
Nitrogen metabolism	398
Sulfur metabolism	339

3.2.6 Maintenance of membrane fluidity

Studying the existence and survival of microbes at low temperatures provides insights into how the machinery of life functions at extreme environments. Such knowledge may also help to explore the possibilities of existence of life in the extraterrestrial environments. Microbes thriving at freezing temperatures have developed a number of strategies over many generations to adapt to specific environmental niches. These microbes have evolved several adaptive mechanisms at low temperature, such as the changes in the fatty acid profile of cell membrane to maintain optimum fluidity through the accumulation of polyunsaturated fatty acyl chains (Methe et al., 2005). A previous study showed that the enzyme desaturase is involved in the conversion of saturated fatty acids into unsaturated fatty acids (Suutari and Laakso, 1994), and induces the synthesis of short-chain fatty acids, branched chain fatty acids, and anteiso-fatty acids to long chain fatty acids, straight-chain fatty acids, and iso fatty acids, respectively, at low temperature (Suutari and Laakso 1994). Correspondingly, there were 7,042 sequences similar to this gene, when the BLASTX (Altschul et al. 1990) results were analyzed independently. It has been reported that desaturase plays a significant role in the growth of Antarctica cyanobacterial strain (Chintalapati et al 2004), and a large number of desaturases were found in the glacial ice from Alpine region (Simon et al., 2009). Moreover, the trans fatty acids are believed to facilitate the survival of bacteria at higher temperature by decreasing membrane fluidity. The conversion of cis fatty acids into trans fatty acids is facilitated by the enzyme cis-trans isomerase. A strain of *Pseudomonas* from

the Antarctic have shown a decrease in membrane fluidity with consequent increase in the amount of saturated and trans monounsaturated fatty acids (Kiran et al 2005). A number of sequences matching the *cis-trans isomerase* enzyme were detected in the BLASTX results from Alaskan glacier, suggesting the adaptive mechanism of microbes for surviving at low temperatures.

Furthermore, it is known that synthesis of branched chain fatty acids use intermediates formed during degradation of isoleucine and valine. Strikingly, the KEGG (Kanehisa and Goto, 2000) category amino acid metabolism contained 1, 141 sequences exhibiting similarities with valine, leucine, and isoleucine degradation. Hence, the results suggest that the cells of microbes at low temperature require fatty acids, for the maintenance of membrane fluidity and these microbes acquire these molecules both through anabolic and catabolic pathways.

The data from glacial ice showed 20 sequences showing similarities with carotenoids biosynthesis. Carotenoids were associated with the regulation of membrane fluidity, and they have been also reported to be important pigments of cell membranes of psychrophilic bacteria (Chattopadhyay and Jagannadham, 2001; Feller and Gerday, 2003; Simon et al., 2009). Additionally, 4, 532 sequences showed similarities with gene encoding peptidyl-prolyl cis-trans isomerase that is known to be essential for maintaining protein-folding rates at low temperatures (D'Amicao et al., 2006).

3.2.7 Role of cryoprotectants

A cryoprotectant is a substance used to preserve living materials by protecting tissues and cells from freezing due to ice formation, and organisms present in cold regions create cryoprotectant in their bodies to minimize the freezing at sub-zero temperatures. Many genes detected encode the synthesis of well-known cryoprotectants are considered to prevent cold induced accumulation of proteins and maintain fluidity of membranes at low temperature. This mechanism for avoiding cell damage by formation of ice crystals is also a requirement for living at subzero temperatures. The genes encoding for known osmolyte, such as glycine and betaine, were found along with genes encoding choline, sarcosine, and glutamate in the data. The presence of such substances have been reported in previously conducted studies from microbes living at low temperatures (e.g., Chattopadhyay 2002; Cleland et al., 2004; Simon et al., 2009).

CHAPTER 4

SEQUENCE COMPOSITION DIVERSITY IN ALASKAN GLACIER AND OTHER METAGENOMES

This work has been in published as follows:

- Choudhari S, Dial RJ, Kumar D, Shain DH, Grigoriev A. (2014) Sequence composition diversity in Alaskan glacier and other metagenomes. PeerJ PrePrints 2:e734v1 https://dx.doi.org/10.7287/peerj.preprints.734v1

My contributions to this work include:

Conception and design of experiment, analysis of the data, and preparation of all the figures and/or tables.

4.1 Introduction

We performed a sequence-based comparison of two geographically distinct glacier metagenomes (Choudhari et al., 2014), that of an Alaskan glacier (Choudhari et al., 2013) and an Alpine glacier (Simon et al., 2009), and observed a striking difference between them. Not only have we seen a significant difference in the numbers of operating taxonomic units between the two samples, but a sequence composition of their reads was

very dissimilar (Choudhari et al., 2014). Although the same type of starting material (ice/snow) was used for the two samples, they were sequenced with different platforms. This motivated us to consider a broader picture of how metagenome sample composition may relate to the corresponding sequencing platform. The current study emphasizes the comparative analysis of metagenomic sequencing data from different platforms in order to understand the variance in the data generated by different sequencers. We compared sequence data generated by two different platforms from the same biological sample, as well as sequences of different samples generated by same platform. Additionally, here we examined the sequencing data of an environmental sample (glacier) generated via two different platforms, while keeping all other pre-sequencing steps similar. The general question is whether it is appropriate to use the information obtained via the two technologies for comparative purposes. The present study outlines how various sequencing tools generate differences in the distribution of nucleotides in a sequence. We compared the sequences generated by these platforms, and also classified the data by binning into distinct taxonomic groups. The approach used for the analysis of sequence data involved the generation of nucleotide word frequencies that represents the distribution of nucleotides in the reads produced by different technologies and was displayed using principal component analysis (PCA). We observed striking differences in the nucleotide composition of the reads generated by different sequencing platforms and related them to the PCA load factors and GC composition of the hypervariable regions of 16S rRNA.

4.2 Results

4.2.1 Deep sea and leech gut metagenome via 454

To examine the sequence data of different samples generated by single sequencing platform, we used data produced by 454 from deep sea metagenome and leech gut. Only the variable V6 region of the 16S rRNA was extracted from the reads and used for phylogenetic identification of species, and we analyzed nucleotide composition patterns using PCA. The sample sequences were also classified taxonomically into different bacterial groups. For this analysis, we selected the most dominant bacterial groups, including *Bacteroidetes* and *Firmicutes* (that constitute the vast majority of the dominant human gut microbiota (Arumugam et al., 2011)), as well as *Actinobacteria* and *Proteobacteria* (which dominate mostly all environmental metagenome (Lee et al., 200; Choudhari, et al., 2014). We found a good overlap of clusters between the two samples. The sequences cluster in separate taxonomic groups in the reduced sequence composition space represented by PCA, and the same bacterial groups from different samples formed overlapping clusters (Fig. 4.1). This was in agreement with our expectation that the sequences of same bacterial groups would be similar in composition, even for such very diverse sources as leech gut and deep sea metagenomes. This suggested that the origin of sample was not creating any kind of bias in sequences generated by the same sequencing platform.

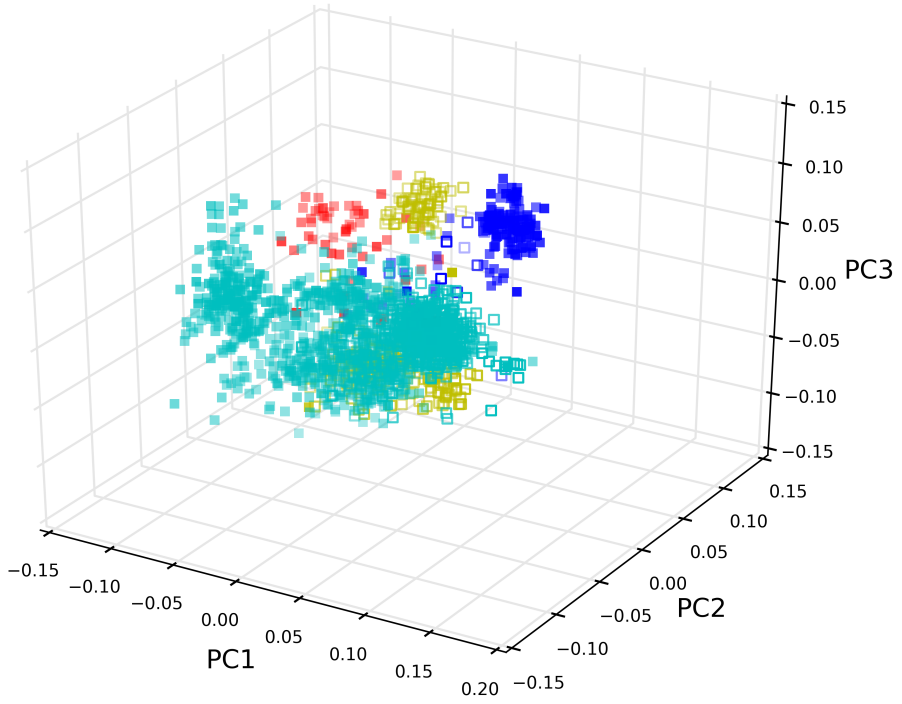


Figure 4.1. Deep sea metagenome (solid boxes) and Leech gut (empty boxes) data generated through 454: Nucleotide word frequency principal component analysis (PCA) of V6 hypervariable regions of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise).

One noticeable feature in the grouping of *Firmicutes* by both the metagenomes was observed: It formed two separate clusters in the 3d space of PCA plot, with both clusters containing sequences from both samples. This behavior was likely caused by a bimodal GC content distribution for the reads from this group in each sample (Fig. 4.2).

When we considered the PCA load factors, we saw that the main contributions in PC1 and PC3 were provided by dinucleotides that were either extremely GC-rich or extremely GC-poor (Table 1). By comparing the result of PCA plot, load factors, and GC content, it can be concluded that although the similar grouping effect for both the platforms was obtained, the separation of different bacterial groups was achieved for each platform, and the separation was due to difference in the GC-rich dinucleotides in different groups.

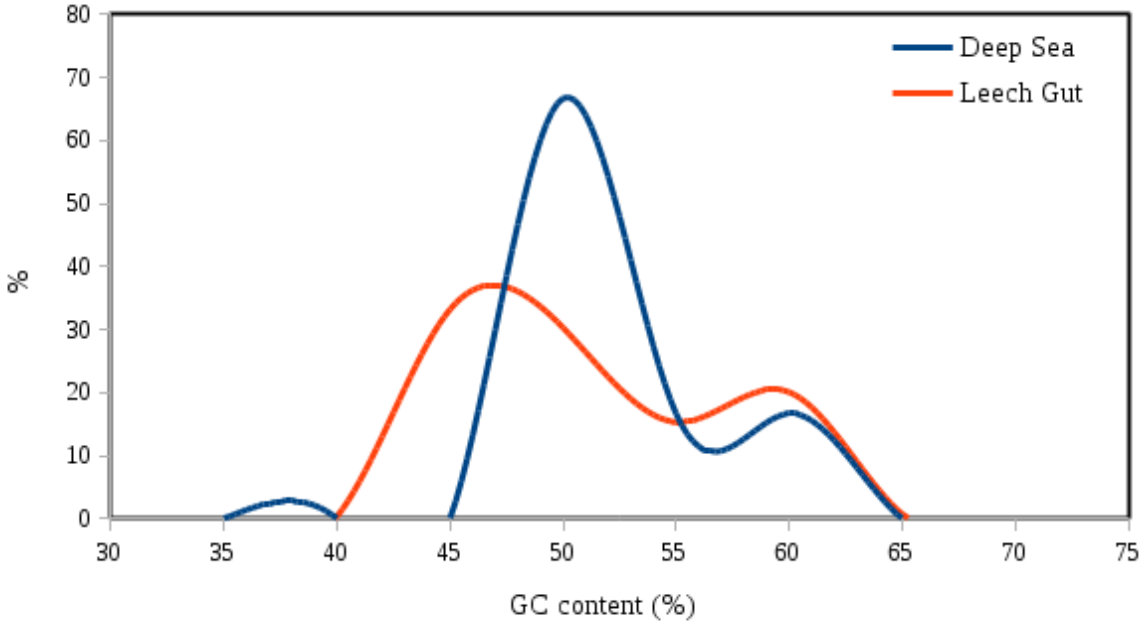


Figure 4.2. GC-curve of diverse metagenomic datasets: The GC% profile of bacterial group, *Firmicutes* of deep sea metagenome (blue) and leech gut metagenome (red).

Table 4.1. First, second and third principal components of dinucleotide word frequencies and their corresponding load factors

	PC1 Component	PC2 Component	PC3 Component
		Deep Sea – Leech Gut[454]	
AA	-0.07	**0.28**	0.00
AC	-0.09	-0.27	-0.17
AG	0.07	-0.28	-0.03
AT	-0.20	0.10	**0.32**
CA	-0.01	0.05	-0.03
CC	**0.40**	-0.24	-0.11
CG	-0.21	-0.54	0.11
CT	0.02	0.12	-0.49
GA	-0.14	-0.20	-0.32
GC	0.14	-0.21	-0.16
GG	**0.65**	-0.20	**0.37**
GT	0.02	0.02	0.32
TA	-0.05	0.28	0.21
TC	-0.21	-0.31	-0.11
TG	0.16	**0.29**	0.03
TT	-0.45	0.12	0.30

The grays with boldface colored boxes correspond to frequencies having top two highest coefficients of variables (load factors) and the light gray colored represent bottom two lowest coefficients of variables
454 – 454 Sequencing

4.2.2 Human Urine metagenome via MiSeq and Ion Torrent

Further, we considered sequence data from a single source, but generated by two different platforms in order to examine if platforms introduced some kind of bias in the sequences. We analyzed human urine sample sequenced by MiSeq and Ion Torrent and platforms. The V1-V2 hypervariable region of 16S rRNA (common for these samples)

was used for the calculation of dinucleotide word frequencies, and phylogenetic classification was performed as described earlier. In a PCA plot (Fig. 4.3) the two different technologies formed two distinct clusters. The same bacterial groups were considered as described earlier and all the groups clustered closer to their platform counterparts. Although the separation between the two platforms was small, there was no overlap between the two platforms, in contrast to the previous case. The first three principal components provided a good resolution, accounting for around 80 % of the variance. The top load factors again showed preference for word frequencies of extreme GC-rich or GC-poor dinucleotides (Table 2). The positive scatter region on PCA plot was occupied by Ion Torrent, indicating that this platform produced more GC-rich sequences compared to MiSeq. This was confirmed on the GC content plot (Fig. 4.4), where each individual bacterial group showed higher GC content when sequenced by Ion Torrent.

In general, these observations suggest that different sequencing technologies generate compositional bias in the sequences they produce. It is important to note that, in this section, significant bias in sequence composition of same sample was observed. This is in stark contrast to no bias observed when entirely different samples were sequenced using the same platform. However, GC content appeared to be the main contributor to the cluster separation in both cases (same platform or same sample).

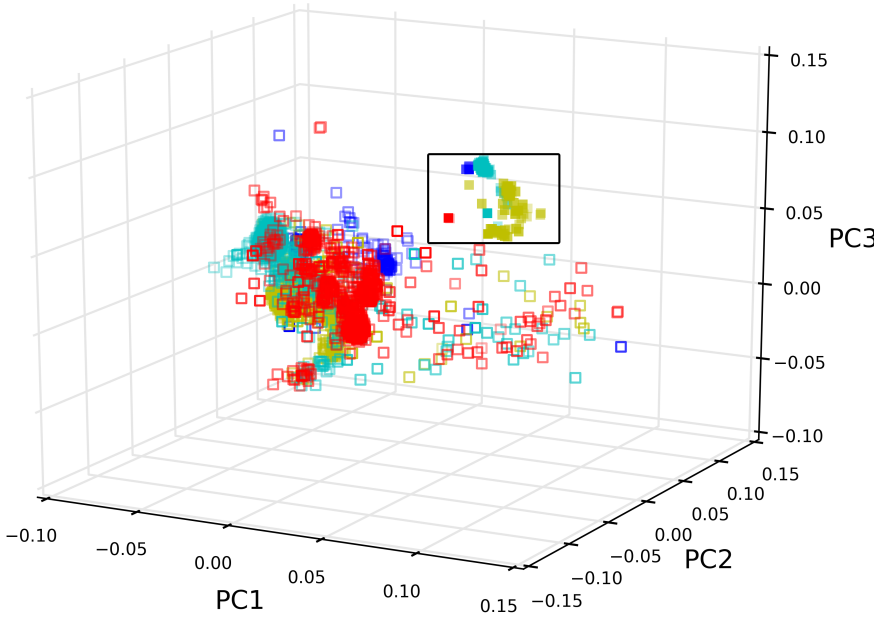


Figure 4.3. Human urine metagenome generated via Illumina MiSeq (empty boxes) and Ion Torrent (solid boxes): Dinucleotide word frequency principal component analysis (PCA) of V1-V2 hypervariable regions of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise). The selected area is drawn around tightly clustered Ion Torrent sequences to highlight their separation from the MiSeq.

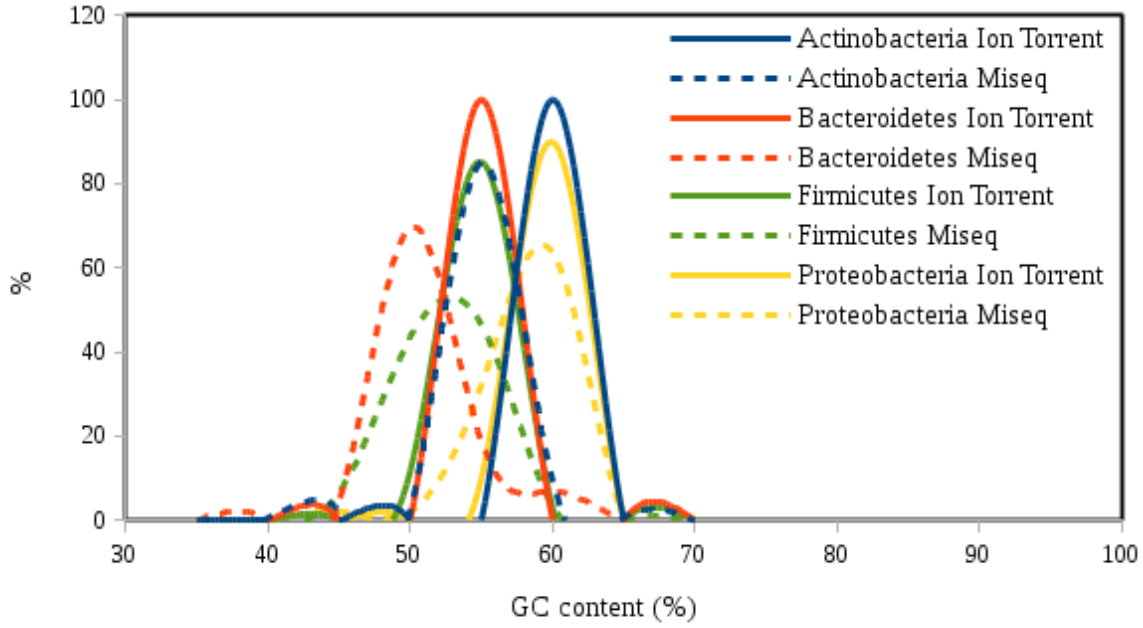


Figure 4.4. GC-curve of different bacterial groups: The GC% profile of bacterial group, *Actinobacteria* (blue), *Bacteroidetes* (red), *Firmicutes* (green) and *Proteobacteria* (yellow) sequenced via Ion Torrent (continuous) and Illumina MiSeq (dotted).

Table 4.2. First, second and third principal components of dinucleotide word frequencies and their corresponding load factors

	PC1 Component	PC2 Component	PC3 Component
		Human urine [M, IT]	
AA	-0.16	-0.55	-0.18
AC	-0.15	-0.08	0.09
AG	-0.17	-0.28	-0.07
AT	0.14	-0.14	-0.21
CA	0.11	-0.37	**0.41**
CC	0.14	0.05	**0.44**
CG	-0.06	0.09	0.20
CT	0.08	0.28	-0.15
GA	-0.23	-0.08	-0.33
GC	-0.10	0.10	0.10
GG	-0.57	**0.40**	0.23
GT	0.08	0.14	0.05
TA	-0.05	-0.08	-0.32
TC	**0.36**	-0.01	0.28
TG	-0.01	**0.36**	-0.28
TT	**0.58**	0.18	-0.24

The grays with boldface colored boxes correspond to frequencies having top two highest coefficients of variables (load factors) and the light grays colored represent bottom two lowest coefficients of variables.
M - Illumina Miseq
IT - Ion Torrent

4.2.3 Glacier Metagenome

As mentioned above, a motivation for this analysis was our earlier observation (Choudhari et al., 2014) of drastically different composition of hypervariable regions sequenced for two glaciers (in Alaska and Alps). We continued it in this study and compared the same snow samples from an Alaskan glacier using two sequencing platforms (Ion Proton and MiSeq). Although different sequencers were utilized, every

other step prior to sequencing (i.e. the method of DNA isolation and preparation of libraries) was kept the same. The difference with the previous examples was that the sequences taken into account were from shotgun metagenomics, as opposed to specific PCR-amplified hypervariable regions of 16S rRNA in previous examples. From this shotgun data set we selected V9 regions based on best match against refv9 database but such regions were of different length in the case of shotgun, possibly affecting the dinucleotide word frequencies and taxonomic assignments as compared to the previous two cases, where hypervariable 16S rRNA regions were specifically sequenced.

The same bacterial groups were studied as described earlier. In a PCA plot (Fig. 4.5) two clusters are observed, and we used a 2d picture as it better shows a mirror symmetry along the PC1 axis. This symmetry is due to the fact that both plus and minus strand BLAST hits were considered together and it clearly illustrates that the PCA is very sensitive to detecting difference between reverse complimentary sequences. However, no platform-specific separation is seen. Instead, one can observe closeness of clusters corresponding to the same bacterial groups, although the clusters are not as clearly separated as in the case of the 454 sequencing.

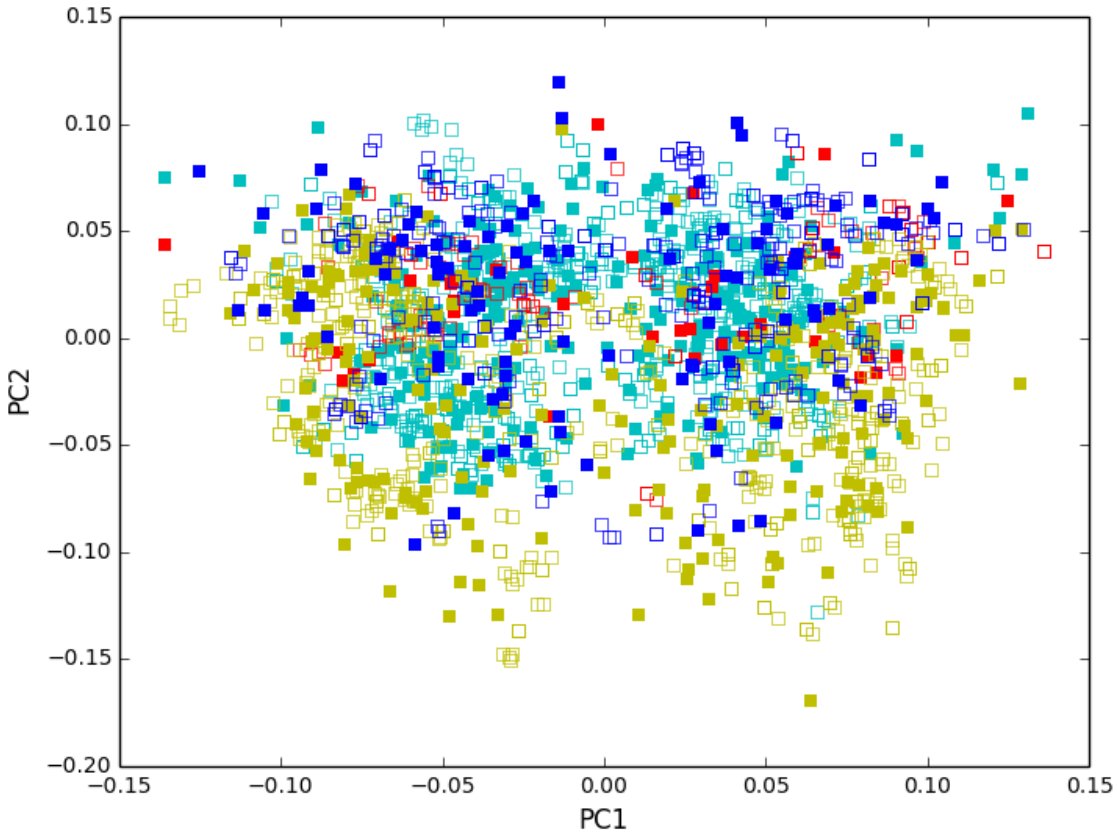


Figure 4.5. Glacier metagenome sequenced via Ion Proton (solid boxes) and MiSeq (empty boxes)- Dinucleotide word frequency principal component analysis (PCA) of V9 hyeprvariable region of 16S rRNA sequences. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise).

CHAPTER 5

DETECTING COMPOSITION BIASES IN METAGENOMES WITH PHYLOGENETIC HEATMAPS

This chapter forms the basis of a manuscript, to be submitted.

My contributions to this work include:

Conception and design of experiment, analysis of the data, and preparation of all the figures and/or tables.

5.1 Introduction

Although metagenomics and high-throughput sequencing have expanded our knowledge in the field of microbiology by directly accessing the microbial community genomes, there are still uncertainties in the data. For example, the sequence composition may vary while comparing sequences from a single environment generated by different platforms. Also, biases can be introduced while interpreting results and during sequencing. In a recent study, while assigning genes to specific organisms, differences in the results were observed when two different sequencers were used for sequencing the mock communities (Salipante et al., 2014). While comparing the results of glacier

shotgun metagenome generated by Miseq and Ion Proton, we did not observe any bias in the sequence composition in our previous study. However, biases were observed while comparing different hypervariable regions of 16S rRNA. Although the biases are not present in the shotgun metagenome, the variance in the data could be due to amplification of 16S rRNA genes (Ahn et al., 2012), DNA extraction (Morgan et al., 2010; Abusleme et al., 2014), and sequencing protocol utilized. Upstream processes, such as DNA extraction also impose certain degrees of bias in the sequences. For example, another recent study found that the species representation of mock communities varied when DNA was extracted using different protocols (Abusleme et al., 2014). Not only do technical steps like DNA extraction, PCR, and sequencing introduce biases in the data, but another study found that sample storage conditions can also generate variation in the data (Cardona et al., 2012). Results from another study, which compared fecal and soil bacterial communities, showed variability in sample profiles with different template concentrations (Kennedy et al., 2014).

Here we describe a method for detecting sequence-based bias in the metagenome sequence data generated by different sequencing platforms. To evaluate our method we examined the sequences from mock communities generated by single platform, but the DNA extraction was done using different protocols (Abusleme et al., 2014). Additionally, we compared two diverse metagenomes (human gut and soil) generated by the same sequencing platform. Furthermore, we analyzed another human gut metagenome, which was a part of human microbiome project generated by two different sequencing

platforms. Finally, we analyzed three different soil metagenomes (Frank-Fahleet al., 2014; Franzetti et al., 2013) generated by different platforms to study the sequence composition in an environmental sample.

The general question is whether it is appropriate to use the information obtained via two technologies for the comparative purposes. The study outlines how various sequencing tools generate differences in the distribution of nucleotides in a sequence. We compared the sequences generated by these platforms, and also classified the data by binning into distinct taxonomic groups. The approach used for the analysis of sequence data involved the generation of nucleotide word frequencies that represents the distribution of nucleotides in the reads produced by different technologies, and was displayed using principal component analysis (PCA) and phylogenetic heatmaps. We observed striking differences in the nucleotide composition of the reads generated by different sequencing platforms and related it to the PCA load factors and GC composition of the hypervariable regions of 16S rRNA. We also examined the load factors of principal component that gave the maximum variance to the data to identify which particular word nucleotide was responsible for the difference. Additionally, the GC-content of the data informed us about the GC-biased platforms. We determined that for a metagenome sequence composition of reads varied based on sequencing platform, and we also noticed a difference in the sequences of a gene from a single species involving different DNA extraction protocols prior to sequencing.

5.2 Results

We evaluated the method for detecting composition bias using data from a recent study that detected a difference in bacterial species representation in mock community while using different DNA extraction methods (Abusleme et al., 2014). The species included in mock community were *Streptococcus, Fusobacterium, Porphyromonas,* and *Veillonella*. This data was used as positive control where we observed that each species formed distinct cluster in PCA plot based on di-, tri-, and tetra nucleotide word frequencies (Fig. 5.1.A). Also, each species formed an overlapping cluster when DNA was extracted using different protocols. Additionally, a compact visual image was generated in the form of concentric heatmap representing the similarities and difference in the data. The V1-V2 hypervariable region of the 16S rRNA were used for analyzing the sequence composition and for phylogenetic classification of species.

The clusters of each species scattered distinctly (Fig. 5.1.A) along the PC1 axis, which explained around 45% of variability, and PC2, which explained about 31% of variability, but same species clusters were overlapping for each protocol. PC1 highlighted a separation between different species based on the nucleotide composition of the resulting sequences, and PC2 provided a separation between *Porphyromonas and Streptococcus* with the other two species, *Fusobacterium* and *Veillonella*. Each cluster represented specific gene (V1-V2 region) from a single species, and therefore, we expected a single point rather than a clusters of sequences. Ideally, all the sequences of a gene from a single species would be identical in sequence composition; instead we

observed variance in the data. Notably, we observed the most redundant sequences of species from each protocol centered in each species cluster (black solid boxes in Fig. 5.1.A), and the clusters were due to some sequencing errors.

Additionally, we investigated the effectiveness of clustering in PCA analysis by using phylogenetic composition heatmaps for the two samples. We calculated the Pearson's correlation coefficient between respective individual bacterial species sequences (Fig. 5.1.B) for all the nucleotide word frequencies from the two data sets, as detailed in the Methods. The mean of r were then converted into different shades of gray, with black corresponding to zero (low correlation) and plotted in a concentric heatmap. The wedges in outer circle of the heatmap represented the correlation between the bacterial species that were found closest to each other in a PCA plot within a sample, for example *Porphyromonas* (P) and *Streptococcus* (S) with upper case letters showing correlation within one DNA-extraction protocol, and lowercase letters within another (*Porphyromonas* (p) and *Streptococcus* (s)). The inner circle of heatmap showed the correlation between the same bacterial species found in two samples (e.g., *Porphyromonas* (P) vs *Porphyromonas* (p)). When there is no bias, the same species from different samples were expected to be closer in nucleotide composition, and in the PCA plot, to each other than to other species; thus, the inner circle will display higher correlation and hence should be lighter in shade than the outer circle. In present example, a lighter shade of the inner circle compared to the outer circle was an indicator of no bias in the sample, showing that same species in two samples were closer to each other in

terms of 16S nucleotide composition.

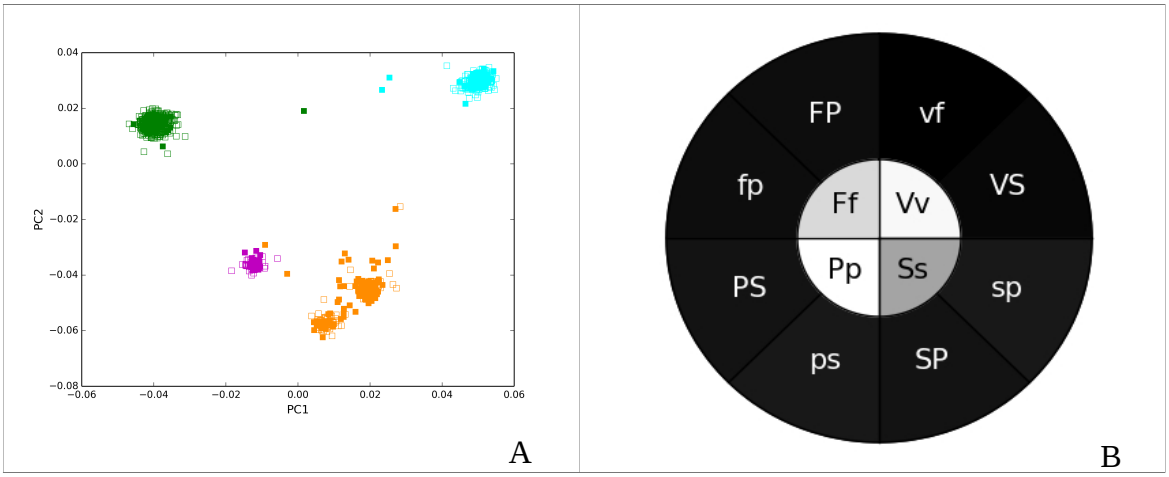


Fig. 5.1. V1-V2 hypervariable regions of 16S rRNA of bacterial species using different DNA-extraction protocols. **A:** Nucleotide word frequency PCA of sequence data from Q (solid boxes) and BB (empty boxes) protocols. PCA orientation with phylogenetic classification at the species-level (*Fusobacterium* = green, *Porphyromonas* = magenta, *Streptococcus* = orange, *Veillonella* = turquoise); **B:** Phylogenetic heatmap were displayed via heatmap, where the inner circle showed the mean of Pearson's correlation (Fisher transformation) computed between same bacterial group from Q and BB protocols, and the wedges in outer circle represent the correlation within one protocol (where uppercase letter indicated sample from protocol Q and lowercase letter from protocol BB).

We compared two diverse metagenomes generated by the same platform within a same experimental framework that in a recent study was aimed to understand Illumina sequencing bias (Kennedy et al., 2014). Here, we analyzed the human gut I and soil I metagenome sequenced by Illumina MiSeq, and observed a good overlap between both samples in PCA (Fig. 5.2.A). The V3 hypervariable regions from both data sets were utilized for analyzing the sequence composition and taxonomic classification. For this

study, we selected the most dominant bacterial groups, including *Bacteroidetes* and *Firmicutes* (which constitute the vast majority of the dominant human gut microbiota (Arumugam et al., 2011)), as well as *Actinobacteria* and *Proteobacteria* that are abundant in environmental samples (Lee et al., 2005). Different bacterial groups formed distinct

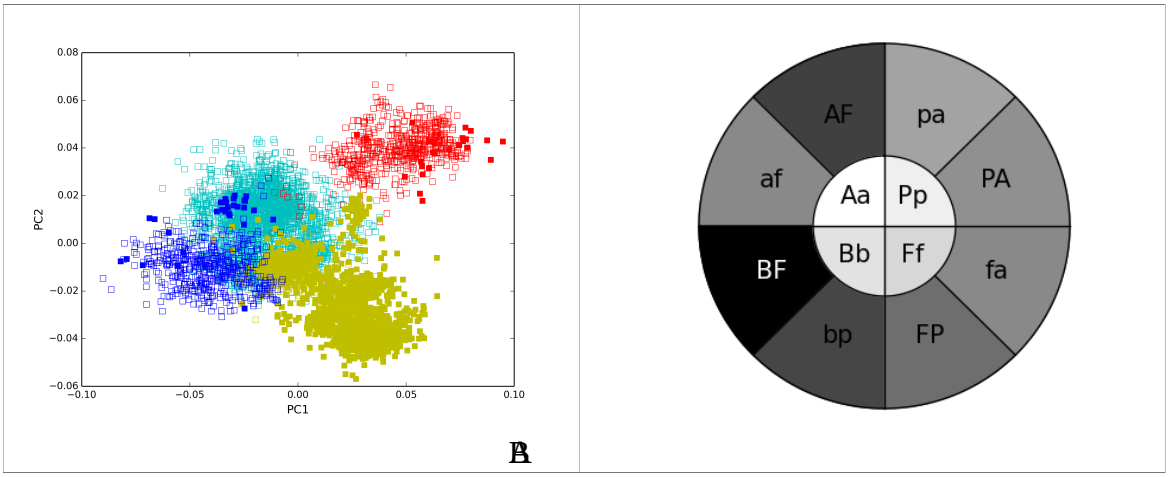


Fig. 5.2. V3 hypervariable regions of 16S rRNA of bacterial groups from two diverse metagenomes generated through Illumina MiSeq: A: Nucleotide word frequency PCA of sequence data from human gut (solid boxes) and soil (empty boxes) metagenome. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise); **B: Phylogenetic heatmap:** Phylogenetic heatmap were displayed via heatmap, where the inner circle showed the mean of Pearson's correlation (Fisher transformation) computed between same bacterial group from human gut and soil metagenome, and the wedges in outer circle represent the correlation within one metagenome (uppercase letter indicated phylum from human gut and lowercase letter represented phylum from soil metagenome).

clusters, but such clusters overlapped for the same bacterial group from different samples; thus, as expected, members of the same bacterial group showed similar composition trends. The effectiveness of clustering was determined by using phylogenetic composition heatmaps for the two samples. As described in previous example we calculated the correlation coefficient between respective individual bacterial phyla from the two data sets (Fig. 5.2.B). Here, the wedges in outer circle of the heatmap represented the correlation between the different bacterial groups within the gut sample e.g., *Bacteroidetes* (B) vs *Firmicutes* (F) with upper case letters, showing very low correlation and hence the color is dark. Similarly, a dark shade of gray wedge in the outer circle represented a low correlation between *Bacteroidetes* (b) vs *Proteobacteria* (p) with lower case letters within the soil sample. The inner circle of heatmap indicated the high correlation between the same bacterial group found in two samples (e.g., *Actinobacteria* (A) vs *Actinobacteria* (a)). Similar results were observed as in case of DNA protocols, same phylogenetic group from different samples had high correlation compared to each other than to other groups; thus, the inner circle displayed higher correlation and hence was lighter in shade than the outer circle. This case also showed a condition of no bias, a lighter shade of the inner circle compared to the outer circle, showing that same phylogenetic group in two samples was closer to each other than with other groups based on nucleotide composition within a sample.

By comparing the result of PCA plot, and phylogenetic heatmap in these examples, it can be concluded that our method found the differences and similarities in

the sequences based on nucleotide composition. These results indicated that while comparing two metagenomes with same laboratory handling, utilizing the same protocols, and generated by single platform, there is no bias generated in sequence composition.

5.2.1 Human gut II: Illumina vs 454

We performed comparative analyses on 16 human gut metagenome data sets sequenced by 454 and Illumina technologies. In the present study, the variable V4 region of the 16S rRNA was common between the samples, and this region was extracted from the reads of sequenced microbiomes for phylogenetic identification of species, and for analyzing nucleotide composition patterns using PCA. After selecting only the V4 regions followed by quality check only small fraction of remaining reads were used for analysis. PCA was applied to the di-, tri- and tetra- nucleotide word frequencies data in order to access the similarities and differences in the nucleotide composition of the sequences generated by two different platforms. The scatter plot of PC1 vs PC2, which accounted for 43% of variability, showed a clear separation between the two platforms along the PC2 axis (Fig. 5.3.A). Moreover, this suggested that the clustering of reads was based on word frequencies from the two platforms in the study; the two clusters were far apart, suggesting that frequencies with which different nucleotide words occur in sample sequences were specific to sequencing technology.

The sample sequences were also classified into taxonomic groups, and the same

bacterial group were analyzed as described above. The different bacterial groups were scattered along the PC1 axis, and grouped with their sequencing platform along the PC2 axis. We also observed that the clusters from Illumina occupied the positive PC1 axis whereas the 454 cluster were scattered along PC1 axis. Although phyla formed separate taxonomic groups, in the reduced sequence composition space the phyla remained in the vicinity of their sequencing platform counterparts (Fig. 5.3.A). On the other hand, sequences obtained by different platforms from the same bacterial group occupied different locations in the PCA plot, indicating clear differences in their sequence composition.

To investigate the reason for this difference, we compared the load factors of the first principal components to find the nucleotide word frequencies, which would explain the observed variance. Principal component load factors indicate the importance of each variable (in this case, frequency of a specific) in accounting for variability in the principal component. We plotted the top 20 positive and top 20 negative load factors as histograms for PC1 (Fig. 5.3.B), coloring the bars according to GC content of the nucleotide words (e.g., the color red indicates all-GC words, with 100% G or C; the color blue indicates non-GC words, with 0% of G or C; and the color green indicates all intermediate values).

We made two observations from this analysis. First, the sample/platform separation was very clearly provided by PC2 and one di-nucleotide, "GG", provided the largest contribution to the PC1. Second, the frequency distribution pattern of “GG” was same in both the platforms. This was clearly shown in Fig 5.3.A, where the clusters of

454 (except *Bacteroidetes*) were at the negative PC1 axis, and negative axis in Fig 5.3.B was dominated by GC-rich (red bars) nucleotide words. This indicated that the two technologies clearly differ in their ability to capture sequences with different GC content and "GG" being the main contributor.

The GC-content distributions of the sample sequences (Fig. 5.3.C) showed Illumina and 454 data sets follow the same distribution pattern with a sharper peak at 55% GC. PCA load factors provided a better insight into the contribution of specific GC-rich/poor nucleotide words, while GC-content plots showed more general trends. The phylogenetic heatmap showed a darker shade of gray in the inner circle compared to the outer circle, indicating a bias in the two samples. The sequences composition of a phylogenetic group was closer to other groups within its sample than to the same group in other sample. Ideally, same group should be closer to each other irrespective of sequencing platform used, and hence the inner circle should be lighter in shade as in the previous two cases.

A striking indicator of bias, for example, was the sequence composition of bacterial group *Actinobacteria* from the two different platforms. These two *Actinobacteria* clusters were further apart than sequences of *Actinobacteria* and other bacterial groups (*Bacteroidetes*, *Firmicutes* and *Proteobacteria*) generated by same platform. Contrary to the expectation that the sequences of same bacterial group would be similar in composition regardless of sequencing platform utilized for generating data we found that the sequences of unrelated bacterial groups were similar to each other if they were generated from same sequencing technology.

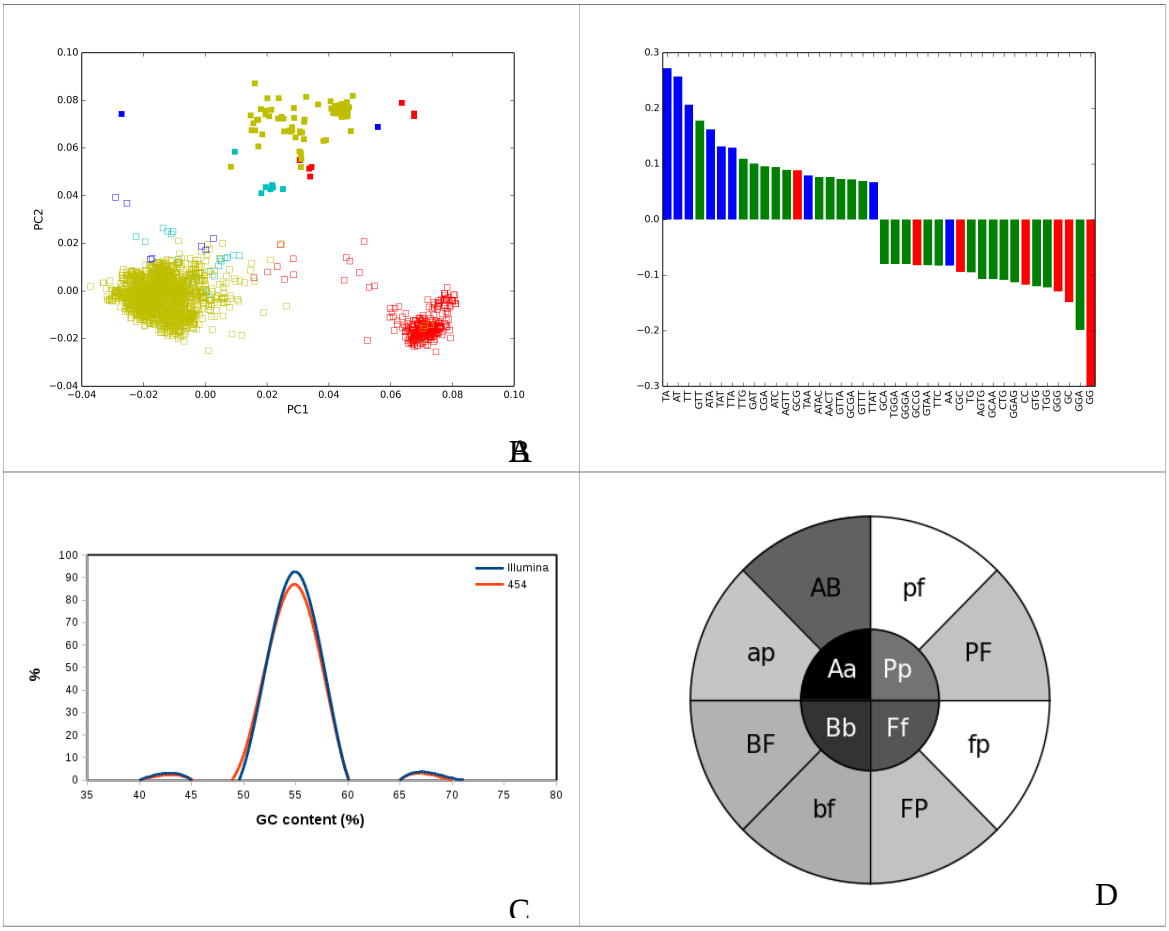


Fig. 5.3. V4 hypervariable regions of 16S rRNA of bacterial groups from two humn gut metagenomes generated through Illumina MiSeq and 454: A: Nucleotide word frequency PCA of sequence data from human gut metagenome generated via Illumina MiSeq (solid boxes) and 454 (empty boxes) platforms. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise); **B:** Histograms of the components of the first principal component with top 20 and bottom 20 load factors and their corresponding frequencies. The different colors indicate the percentage of 'G' or 'C' in the nucleotide frequency (red = 100% 'G' or 'C', green = less than 100% and more than 0% 'G' or 'C', blue = 0% 'G' or 'C') **C:** GC-content- GC% was calculated for each read and the percentage of the data set from each platform in intervals of bin width 5 is shown vs the percentage of the reads; **D:** Phylogenetic heatmap were displayed via heatmap, where the inner circle showed the mean of Pearson's correlation (Fisher transformation) computed between same bacterial group from both the platforms, and the wedges in outer circle represent the correlation within one platform (uppercase letter indicated phylum from Illumina and lowercase letter represented phylum from 454).

5.2.3 Soil Metagenome II: 454, Illumina and Ion Torrent

We further considered platform-specific biases in environmental metagenomes using soil samples sequenced by 454, Illumina, and Ion Torrent platforms. We performed two pairwise comparisons: Illumina vs Ion Torrent (Fig. 5.4), and Ion Torrent vs 454 (Fig 5.5). The common V5 hypervariable region of 16S rRNA was used for the calculation of nucleotide frequencies, and phylogenetic classification was performed as described earlier. The same set of four bacterial groups were considered, and all these groups clustered closer to their platform counterparts in the first case that is Illumina vs Ion Torrent (Fig 5.4.A), whereas the same phylogenetic groups from different platforms overlapped with each other in the second case that is Ion Torrent vs 454 (Fig 5.5.A).

In the first case, one can again observe separation of the two technologies on the PCA plot (Fig 5.4.A), with further sub-clustering of different bacterial groups within each cluster. Unlike human gut, the sample separation from different platforms was very clearly represented by PC1 in this case. The GC-content of the sequences from the two samples was also much closer, with the peaks for each technology within 3% G+C, Fig. 5.4.C. By far, the top three load factors in the first component were all-GC nucleotide words (i.e., red bars; Fig 5.4.B). This finding was in agreement with the arrangement of the cluster in the PCA plot, where the Ion Torrent solid boxes were in the positive half of the PC1 axis, indicating that this platform produced more GC-rich sequences compared to Illumina. Although this may seem contrary to the slightly higher mean GC content of

Illumina vs Ion (Fig. 5.4.C), a more detailed investigation of the top load factors revealed that, strikingly, the distributions of all high load factors in both samples were bi- or multi-modal, with the top three all-GC nucleotide words favoring Ion Torrent compared to Illumina, and thus resulting in the separation of clusters along the PC1 axis (Fig 5.4.A).

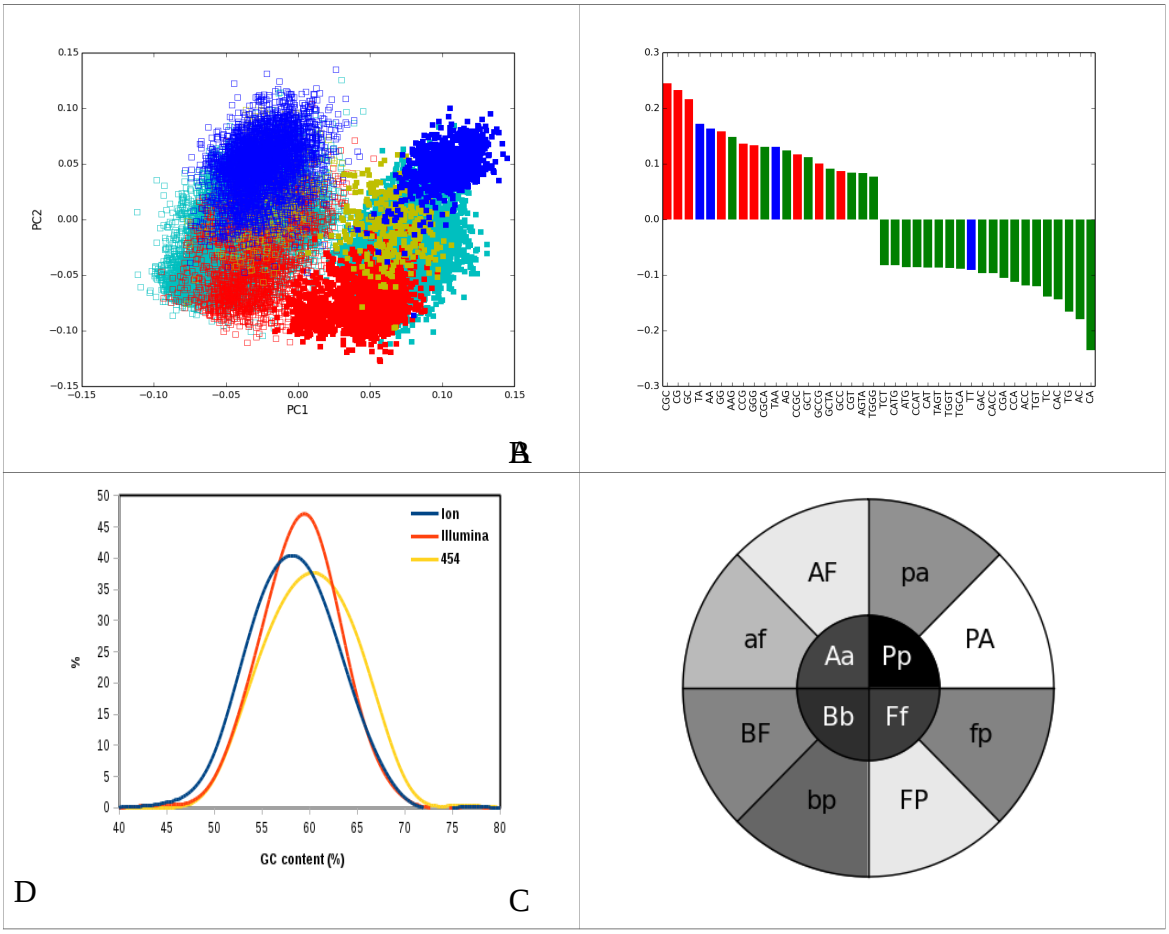


Fig. 5.4. V5 hypervariable regions of 16S rRNA of bacterial groups from two soil metagenomes generated through Ion Torrent and Illumina Genome Analyzer: A: Nucleotide word frequency PCA of sequence data from soil metagenome generated via Ion Torrent (solid boxes) and Illumina Genome Analyzer (empty boxes) platforms. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise); **B:** Histograms of

the first component; **C:** GC-content; **D:** Phylogenetic heatmap were displayed via heatmap, where the inner circle showed the mean of Pearson's correlation (Fisher transformation) computed between same bacterial group from both the platforms, and the wedges in outer circle represent the correlation within one platform (uppercase letter indicated phylum from Ion Torrent and lowercase letter represented phylum from Illumina).

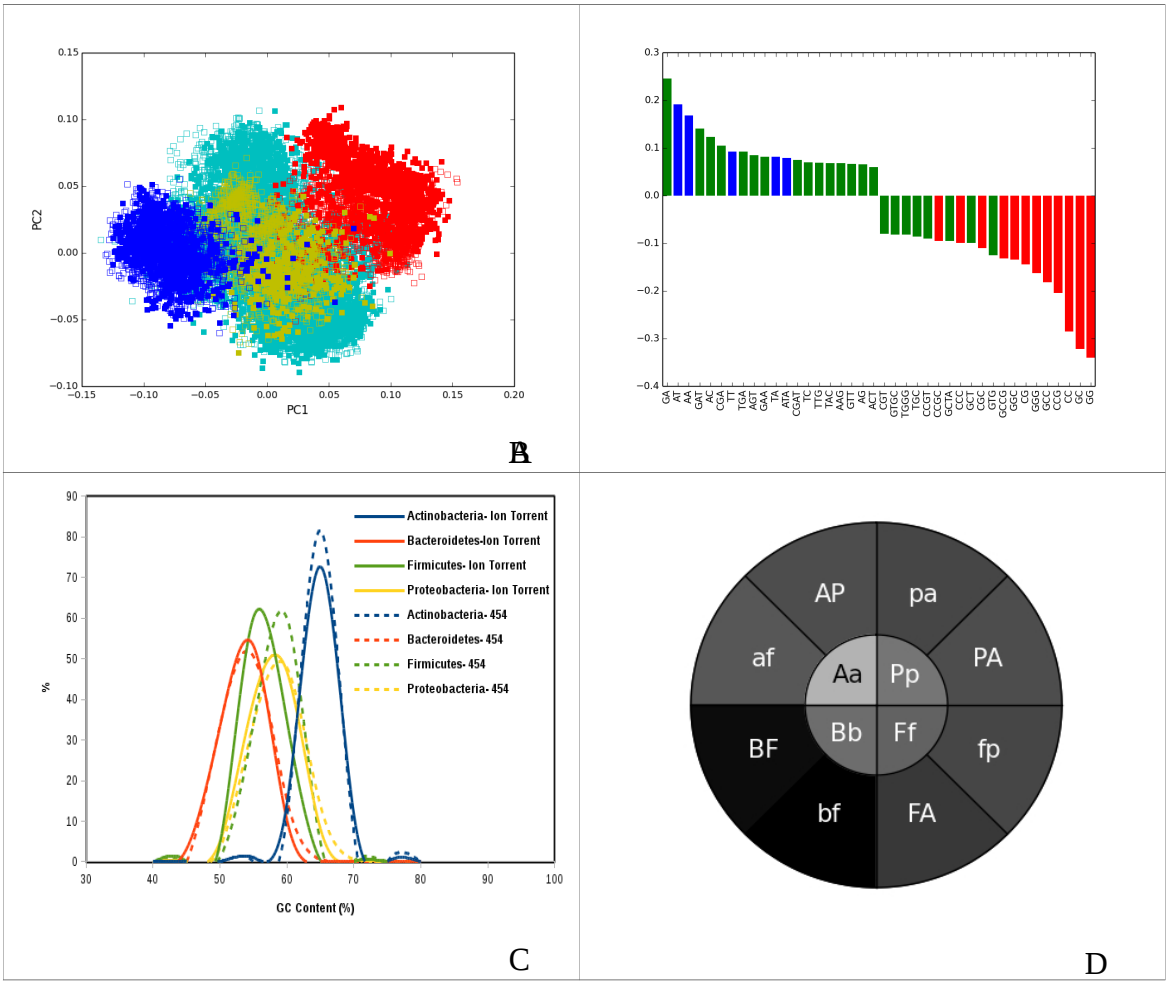


Fig. 5.5. V5 hypervariable regions of 16S rRNA of bacterial groups from two soil metagenomes generated through Ion Torrent and 454. A: Nucleotide word frequency PCA of sequence data from soil metagenome generated via Ion Torrent (solid boxes) and 454 (empty boxes) platforms. PCA orientation with phylogenetic classification at the phylum-level (*Actinobacteria* = blue, *Bacteroidetes* = red, *Firmicutes* = yellow, *Proteobacteria* = turquoise); **B:** Histograms of the first component; **C:** GC-content of all

bacterial groups; **D:** Phylogenetic heatmap were displayed via heatmap, where the inner circle showed the mean of Pearson's correlation (Fisher transformation) computed between same bacterial group from both the platforms, and the wedges in outer circle represent the correlation within one platform (the uppercase letter indicated phylum from Ion Torrent and lowercase letter represented phylum from 454).

When we analyzed the data from Ion Torrent and 454, an overlap of clusters from both technologies was observed (Fig. 5.5.A). The sequences dispersed along PC1 axis, which accounted for around 20% of the total variance, and clustered into separate taxonomic groups in the reduced sequence composition space of PCA, where the same bacterial group from different samples formed overlapping clusters (Fig. 5.5.A). When we considered the load factors of PC1 (as the separation was along this axis), we observed that the negative load factors were dominated by GC-rich nucleotide words (i.e., red bars). *Actinobacteria* occupied the negative domain on the PC1 axis, suggesting that this group was GC-rich. This finding was in agreement with the fact that *Actinobacteria* showed higher GC content on both platforms plot (Fig. 5.5.C). The GC% content distribution of 454 and Ion Torrent were closer, peaking at 60% GC and 58 % GC, respectively (Fig 5.4.C). However, the GC-content peak for the most distant clusters on PCA, *Actinobacteria*, and *Bacteroidetes* were 60% and 53%, respectively, in both platforms (Fig. 5.5.C). This finding was in agreement with these groups being GC-rich and GC-poor, respectively. *Proteobacteria* and *Firmicutes* overlapped, as they shared the similar GC–content in both platforms (58-60%). We can conclude here that the main reason for different clustering on PCA plot was due to difference in the GC-rich

nucleotides in different phylogenies, as similar grouping for both platforms was observed.

As with the human gut metagenome II, we observed bias in phylogenetic heatmaps when comparing the nucleotide composition of sequences generated from Illumina and Ion Torrent (Fig. 5.4.D). The clusters of the same bacterial group from different platforms (*Proteobacteria* from each platform) were further apart in the PCA plot, and were closer to other groups from the same platform (*Actinobacteria*, *Bacteroidetes* and *Firmicutes*). A similar pattern was clearly observed in the phylogenetic heatmap, where the inner circle was darker in color compared to the outer circle, indicating a variance in the data from two platforms. Further, this indicated that there is low correlation between sequence composition of a phylogenetic group from two platforms (inner circle darker means low correlation value) and different groups from same platforms were more similar (outer lighter shade of gray represented high correlation values) in same sample generated by same technology than with the same group in another sample generated by different technology.

Furthermore, the results were different when sequences generated via Ion Torrent and 454 were compared. The overlapping of the same bacterial group from different platforms was observed in PCA plot (Fig. 5.5.A), and the inner circle in the heatmap was lighter in shade compared to the outer one. Thus, this indicated that the correlation between same bacterial group in two samples was higher compared to the correlation in groups in a sample. This reflects a case of no sample bias, and hence inner circle was

little lighter compared to outer circle (Fig. 5.5.D). There was not much variation in the shades of the two circle, but still there was slight difference in the shades, the reason is due to very good overlapping and close clustering.

CHAPTER 6

DISCUSSION

Cold environments, such as glaciers, are large reservoirs of microbial life, and studying these unculturable microbes has been made possible through metagenomics, which grants a unique possibility to investigate diverse environments in great detail. Metagenomics by next generation sequencing has become an important tool for interrogating complex microbial community structure and function. About 25% of the land surface on earth is classified as a cold environment (Choudhari et al., 2013), and biological activity in these low-temperature habitats is generally believed to be restricted. Glaciers are simple and relatively closed ecosystems inhabiting primary producers (e.g., photosynthetic bacteria and algae) in the snow and ice.

In the present study we took a snapshot of the current microbial inhabitation of an Alaskan glacier (which can be considered as one of the simplest possible ecosystems) by using metagenomic sequencing of 16S rRNA recovered from ice/snow samples. The functional and metabolic potential of glacial ice metagenome was also analyzed, which provided important clues about functional and metabolic diversity, as well as variation with other metagenomes from different environmental niches. Furthermore, we analyzed several pairs of metagenomic samples obtained by different methods and detected biases, resulting in different nucleotide compositions of the sequenced reads. The pairwise sample comparison was based on the principal component analysis of word frequencies

in sequences obtained from different platforms.

6.1 Phylogenetic analysis of the microbial community present in the glacier ice of Byron glacier:

We analyzed 16S rRNA sequences recovered from the ice/snow samples of the Byron glacier in Alaska, unveiling an unexpectedly rich microbial community including > 2, 500 species of *Bacteria* and also *Archaea*, which had so far escaped detection in the glaciers of the northern hemisphere. Recently, other studies have reported Archaea from Greenland (Miteva et al., 2015) and Iceland (Lutz et al., 2015) that have corroborated our finding of Archaea. An analysis of taxonomic composition of the glacial ice based on 16S rRNA gene revealed that primary contributors in species diversity of the Byron glacier are *Proteobacteria*, *Bacteriodetes* and *Firmicutes*. In terms of relative cellular abundance at the level of individual species, representatives of *Cyanobacteria*, *Actinobacteria*, and *Planctomycetes* were the most numerous. This likely reflects the dependence of the ecosystem on the energy obtained through photosynthesis and close links with the microbial community of the soil. This study is limited to one site and time point, thus providing a snapshot of the glacier ecosystem in Alaska. However, the study also allowed the comparative analyses with other glacial and adjacent microbial communities, along with an expansion of previous findings. Our comparisons with non-glacial communities (i.e., glacier foreland and lake water in high Arctic) showed that soil and glacier habitats have most of the species in common. Contrarily, a comparison to 16S rRNA metagenome

sequencing of the European glacier indicated a notable difference in nucleotide composition. This little overlap between the two samples may be attributed to biogeographical differences, but also to likely biases in the observed community representation related to differences in sampling and sequencing. In this regard, the following rough estimate (ignoring potential biases of cloning, PCR, sequencing, etc.) supports a higher completeness of our sample. The European glacier study (Simon et al., 2009) recovered some 150 pyrosequenced non-unique 16S fragments per kg of ice, and about 40 from the cloned 16S approach. On the Byron glacier, the number of unique sequenced 16S fragments corresponds to 10 OTU ml^{-1}. With some 50 non-unique sequences per OTU in our case, it is much closer to the estimated density of bacteria in the glaciers of 10^4 cells ml^{-1} in melted glacial ice (Anesio et al., 2010). The observed differences suggest that one needs to exercise caution when interpreting the results of the comparative analysis of different metagenomes when only a subset of the community is sampled.

Thus, our results provided a comprehensive description of the microbial diversity of glacial ice metagenome in Alaska and characterized the microbial community structure at low-temperature. The phylogenetic analysis was based on rRNA approach, which is the most common approach to determine microbial diversity. However, due to lack of robustness of the deepest nodes, many 16S rRNA-based analyses provide incomplete taxonomic classification. Additionally, PCR also introduces some biases in the form of PCR artifacts, another disadvantage is the varying genomic copy number of the 16S. The

copy number varies from one to several in some species (Kembel et a., 2012), which can result in variation of the relative abundance of 16S genes in samples. This can results in over estimation of microbes due to variation in 16S copy number among those, and underestimation of microbes with low number of 16S rRNA genes (Sheridan et al., 2003). Also, it can lead to incorrect prediction of a phylogenetic group and gives incomplete taxonomy. Here, we utilized the copy number and abundance estimation approach (Fig. 2.2) that provided better estimation of different phyla in microbial communities.

6.2 Metabolic and functional competency of microbial assemblage present in a low temperature environment:

The functional analysis of the metagenome provided the functional and metabolic repertoires of the microbial community members inhabiting the glacial ice metagenome of Harding Icefield in Alaska by evaluation of large sequencing-derived data. The sequences were compared to KEGG orthology, and a wide metabolic diversity was established. Additionally, a comparative analysis was done with other metagenomes from different environmental niches, and the distribution of different metabolic categories was comparable in all the biomes. The high versatility of carbohydrate metabolism was detected, including a large degradative capacity and the ability to assimilate inorganic and organic compounds. Previous results showed the presence of the group *Methanomicrobia* belonging to *Archaea* and also methanogens and methanotrophs were detected in the active layer of soil and permafrost (Frank-Fahle et al., 2014). In our analysis, we found a

lot of genes involved in the carbon and nitrogen cycle, methane generation and oxidation, and organic matter decomposition that might have an important role in climate change or glacier degradation.

Until now not much has been explored about the metabolic functions of microbial communities of frozen environments. Previously, it was proposed that most of the organic matter originates from production sites elsewhere from atmospheric deposition on the glacier surface and may be allochthonous (Stibal et al., 2008). The presence of inorganic and organic forms of nitrogen, sulfur and phosphorus have been reported in the past from cold environments (Bottrell and Tranter, 2002; Hodson et al., 2005). In this study, evidence of genes for dissimilatory nitrate reduction, i.e., nitrate to ammonia, alanine, aspartate, and glutamate metabolism were detected. In the energy metabolism category, many sequences encoding genes for methane, nitrogen, and sulphur metabolism were detected suggesting microbes present in glacial ice exhibit anaerobic respiration. These heterotrophic bacteria use CH_4, NO_3^-, SO_4^{2-} as electron acceptors, and this kind of metabolism is referred as lithotrophic metabolism. In addition, genes for assimilatory sulfate reduction, i.e., reducing sulfate to H_2S was found as reported previously in glacial ice (Bottrell and Tranter, 2002). It can be concluded based on the metabolic pathways that the microbial communities present in the glacial ice are aerobic or facultative aerobic. Alternatively, they perform lithotrohpic metabolism by using CH_4, NO_3^-, SO_4^{2-} as electron acceptor in the absence of oxygen (Table 3.1).

In addition, several sequences showing high similarities with enzymes involved in

conversion of fatty acids to unsaturated fatty acids were found. The fatty acids are required by the cells of microbes at low temperature for the maintenance of membrane fluidity. Several metabolic pathways were present that were responsible for formation of compounds (cryoprotectants) essential for a psychrophilic lifestyle to prevent cold induced accumulation of proteins, and maintain fluidity of membranes at low temperature. The reported results increased our understanding of different strategies adapted by microbes living in a low-nutrient environment that helps in survival at subzero temperature. Such low temperature adaptation strategies of microbes arouses scientific interest that can not only unravels the life at sub-zero temperatures but also extraterrestrial environments. Such studies provide knowledge that is likely to be useful in controlling pathogenic bacteria, which survive and thrive in cold-stored food materials (Chattopadhyay, 2006). The key determinants of psychrophilic lifestyle involves the modification of enzyme functionality and stability and cell membrane fluidity (Bowman, 2008). The present metagenomic data provided essential information of genes encoding processes for low temperature adaptation of microbes and information of their functionality.

Thus, the diversity of microbes present in the glacial metagenome expressed a great variation in modes of energy generation and metabolism, and this feature allowed them to flourish in sub-zero temperature. The results elucidate several microbial strategies for cell survival in low temperature ecosystems.

6.3 Sequence composition diversity and detecting composition biases in metagenomes with phylogenetic heatmaps:

The advent of NGS along with metagenomics has enabled us to study the whole microbial community of an environment in a quick, cost-effective, and efficient way. Although technology has evolved to the point of making metagenome sequencing easy, one should exercise caution when analyzing NGS metagenomes, as different biases are introduced during sequencing and even prior to it. While interpreting information from sequencing data derived using different sequencing technologies, results may vary and one must be careful when interpreting these results.

We utilized PCA to show how the sequence composition changes with different sequencing technologies. We kept the same hypervariable regions of 16S rRNA while comparing two data sets in order to minimize any kind of regional bias. We compared V6 regions of 16S of two very different metagenomes: leech gut (Maltz et al., 2014) and deep-sea (Sogin et al., 2006), generated by 454 technology, and found the clustering of the data by phylogenetic group. We observed a good overlap of two metagenomes and the same bacterial groups from both samples clustered together. Thus, one should expect that if single sequencing technology is used, no bias is observed in the data regardless of the origin of the sample. Notably, we observed different clusters when the same type of human urine sample (Salipante et al., 2014) was sequenced via MiSeq and Ion Torrent. The V1-V2 variable regions of 16S were used for the analysis. Here, the bacterial groups clustered together with their platform counterparts, rather than grouping with same

bacterial groups. We found that the sequence composition changes when different sequencing platforms are used. The variability in the data appears to be due to GC-content of the reads, and it is also manifested in load factors for extreme GC-rich and GC-poor dinucleotides. Furthermore, with the exception of different sequencing platforms, we kept all aspects, such as the sample type, DNA extraction, and library preparation protocols, identical when comparing glacier metagenomes, and did not observe platform-specific separation. It appears that the platform bias is visible in the reads produced by amplifying hypervariable regions in 16S rRNA, but not in shotgun metagenome reads aligned to such hypervariable regions.

Furthermore, we presented a method that is able to detect a bias in data generated by different sequencing technologies. The method is based on the nucleotide distribution of a sequence in the form of word frequencies (di-, tri-, and tetra), and gives a very distinctive display of these patterns in the form of phylogenetic heatmaps and PCA. The phylogenetic heatmap is a very compact way of visualization, which displays the bias introduced by different sequencing platforms.

To validate our method, we compared mock communities of different bacteria generated by the same platform but utilizing different DNA extraction protocols (Abusleme et al., 2014). We analyzed human gut and soil metagenome (some of the richest metagenomes currently studied) through a pairwise comparison of human gut and soil metagenome sequenced by the same platform. In order to study the bias in the same metagenome, another human gut metagenome generated by different technologies was

analyzed. Additionally, we analyzed three different soil metagenomes generated by different sequencing platforms. The method proved efficient in detecting the sequencing bias in different data sets. Different DNA extraction protocols (Abusleme et al., 2014) may have escaped such biases, but the sequencing noise was clearly observed, and our method was able to detect the variance in the sequences of the same gene from single species. We found no bias when very diverse metagenomes, such as human gut and soil, sequenced by single platforms were compared. Contrarily, we observed a clear difference in another human gut metagenome sequenced by Illumina and 454; the two platforms formed separate clusters, and the outer circle in the heatmap was lighter in color (indicating a high correlation between different groups) compared to the inner circle (indicating a low correlation between same group). A similar kind of distant clustering and lighter shade of outer circle in phylogenetic heatmap were observed between two platforms in soil metagenome generated by Illumina (Franzetti et al., 2013) and Ion Torrent (Frank-Fahle et al., 2014). In both the cases, the bacterial group from different technologies formed clusters with other groups from the same technology. We expected that the sequences of the same bacterial group, regardless of sequencing platform, should be similar in composition, and should thus group together. On the contrary, a very good overlap of two platforms was observed while comparing soil metagenome sequenced via Ion Torrent (Frank-Fahle et al., 2014) and 454, and a lighter shade of inner circle in heatmap was observed, suggesting that same bacterial groups showed more correlation from different platforms. Such results were expected while comparing two metagenome

data sets, where sequences of the same bacterial groups should be similar in composition irrespective of technology used (although Ion Torrent and 454 represent close technologies).

In order to minimize additional bias generated due to different hypervariable regions of 16S rRNA, we used the same hypervariable regions for each comparison of two platforms. We observed that the differences in clustering patterns and different shades of gray in heatmap were due to different proportions of GC-rich and GC-poor nucleotide words in the sequences. The present study found that the Ion Torrent platform produced more GC-biased sequences compared to the other two platforms. Additionally, few bacterial groups, such as *Actinobacteria* and *Bacteroidetes*, were found to be GC-rich and GC-poor, respectively.

We conclude that the sequence composition of similar metagenome varied depending on the sequencing technologies used, and our method is efficient in detecting such kinds of differences and similarities. Not only are the biases platform-specific, but other upstream processes, such as DNA extraction (Morgan et al., 2010; Abusleme et al., 2014) and PCR amplification (Ahn et al., 2012) can also introduce some biases in the results. Additionally, if data generation is done with the same experimental framework, then there is no bias in the sequences regardless of the diversity of the metagenomes. If the same laboratory handling, primers, and sequencers are used, we conclude that there is minimal chance of bias. One important conclusion from this study is that sequence composition of metagenomic data varied depending on the sequencing platforms used.

Thus, one needs to be sensitive to such differences while comparing and combining metagenomic data produced by different technologies.

6.4 Future Directions:

Metagenomics is one of the fastest advancing fields in biology, expanding our knowledge of the diversity, ecology, and functioning of the complex microbial communities. It not only allows identification of the uncultured microbes and their functions potential, but it also leads to the emergence of new applications in many different areas. The present study was limited to one time point and one location, and therefore it delivered only a snapshot of the microbial and functional diversity. The same microbial communities sampled at different times or from different locations may vary significantly. Therefore, further studies that involve the collection of snow samples at different seasons and locations will provide us more comprehensive and detailed understanding of glacial ecosystem. The time-series sequencing data will reveal patterns of change of enzymatic activity and energy metabolism as microbial diversity can occur in different seasons.

In addition to DNA-based metagenomics, gene expression and protein production of microbial communities needs to be explored. In recent years, metatranscriptomics and metaproteomics have emerged, which allows us to understand the functional dynamics of microbial communities (Simon et al., 2009). Metatranscriptomics provides knowledge of how microbial communities respond to changes in their environment by investigating the

actively transcribed ribosomal and messenger RNA from a community (Gilbert and Hughes, 2011). Moreover, metaproteomics provides information on the protein molecular data obtained from complex communities using proteomics techniques. The functional potentials and gene expression of microbial communities present in glacial ice metagenome will shed light on ecosystem functions of microbial communities and evolutionary processes.

The continuous and dynamic development of faster and cheaper sequencing technologies are extending our capacity for the analysis of microbial communities from an unlimited variety of habitats and environments. However, in order to cope with the exponentially increasing amount of complex environmental sequence data sets, there is need for development of efficient computational approaches to determine the taxonomic and metabolic diversity of microbial communities. Therefore, in future in order to benefit from this large pool of data/information, we need to be equipped with improved analysis tools that minimize any kind of bias in comparative studies of data sets from different sequencing platforms.

CHAPTER 7

MATERIALS AND METHODS

This material and methods for each chapter are described as following**.**

7.1 Chapter 2: Material and methods

7.1.1 Sampling and sequencing

The sample was collected from Byron Glacier in Alaska (60.762003N, 148.846545 W). The sample that contained surface ice and snow was collected close to sea level (depth, 2 m; elevation, 154 m). The ice/snow was slowly melted on-site in a water bath whose temperature was maintained at 0–4 ºC followed by pre-filtering to remove metazoans and wind-blown material, then filtered twice. The first filtering at 4–5 μm was performed to remove wind-blown material, and it likely also removed many single-celled eukaryotes (particularly algae and fungi). The second filter at 0.2–0.4 μm allowed us to collect available prokaryotes. These filters were preserved in ethanol, and sequencing was performed through the Earth Microbiome Project (http://www.earthmicrobiome.org) (Gilbert et al., 2010).

7.1.2 Sequence analysis of the 16S rRNA gene sequences

Using the EMP standard protocols (http://www.earthmicrobiome.org/emp-

standard-protocols) (Caporaso et al., 2012) the hypervariable V4 region of the 16S rRNA gene was amplified with bacterial/archaeal primers 515F and 806R and sequenced on an Illumina HiSeq platform as described previously (Choudhari et al., 2013). The data obtained after sequencing were demultiplexed which yielded 136,579 reads. The sequences were screened for quality, and only sequences which had Phred scores greater than 20 and did not contain any `N's were retained. After quality check, we obtained 25,018 total reads of 151 bp, and after removal of exact duplicates, 7,728 unique sequences remained.

Taxonomic classification was performed by MOTHUR (Schloss et al., 2009) using the SILVA (Pruesse et al., 2007) database for eukaryotic, bacterial, and archael 16S rRNA sequences. Alignments were performed using the BLAST tool of the NCBI database (Altschul at al., 1990). The furthest-neighbor method of MOTHUR (Schloss et al., 2009) was used to determine the OTUs at sequence similarity levels of 99, 97, and 80%. Rarefaction curves were created to assess species richness from the results of sampling, plotting the unique OTUs count as a function of the number of sequences sampled.

PCA for finding nucleotide composition patterns in the sequenced reads and for comparisons with other metagenomes was performed using di-, tri- and tetranucleotide word frequencies in the subsequences corresponding to the 16S rRNA V4 variable region. Projections on the first two principal components were plotted highlighting relevant

taxonomical subsets of the sequences.

The sequences from a previously studied glacier from Germany (Simon et al., 2009) (EU978474 to EU978633, EU978636 to EU978652, and EU978654 to EU978854) were obtained from NCBI GenBank. Likewise sequences from studies involving high Arctic freshwater lake (Møller et al., 2013) (SRR066819) and Arctic glacier foreland soil (Schütte et al.,2010) (SRR036794) were obtained from the NCBI SRA. These data sets contained sequences from variable 16S regions other than V4, thus limiting possible sequence direct comparisons. Comparisons of the species distributions were performed after taxonomical classification.

7.1.3 Nucleotide sequence accession numbers

The 16S rRNA gene sequences derived from high-throughput metagenomic sequencing have been deposited in the NCBI SRA under accession number SRP018522.

7.2 Chapter 3: Material and methods

7.2.1 Sample collection

The glacier samples were collected from Harding Icefield in Alaska in August 2013. Snow was scooped from accumulation zone of the glacier when the temperature was around 4ºC, with wind 16 m/s, and rain at 2.3 mm/hr. After the samples melted at room temperature, the pre-filtering of three liters of melted water was done with coffee

filters to remove unwanted materials. Further filtering was done using a hand pump with 1.6 µm filter, in order to filter out eukaryotes. An electric suction pump was used with three durapore membrane filter of 0.22 µm in parallel for final separation of prokaryotes from the filtered (at 1.6 µm) water.

7.2..2 DNA extraction and sequencing

DNA extraction was carried out using bead beating method plus column filtration according to the manufacturer's instructions (MO BIO Laboratories, Inc). After isolation, DNA was fragmented using Covaris S2 and divided into two tubes for Illumina MiSeq and Ion Proton library preparation. Final library was enriched by running 15 PCR cycles prior to sequencing. In the current study data from Illumina MiSeq was used for functional analysis.

7.2.3 Data processing

Glacier: Illumina MiSeq-

To access the functional and metabolic capacity of microbial communities present in glacial ice, the reads from the data were compared against non-redundant NCBI database (Benson et al. 2005) using BLASTX approach (Altschul et al. 1990). Only 3,810,507 reads that passed the quality filtering of Phred scores greater than 20 and did not contain any 'N's were retained, and reads greater than 50 bp in length were only taken for the comparison. The results from BLASTX was used with MEGAN 4 database

for comparing the given reads against a reference database. MEGAN 4 takes the file of reads and resulting BLAST file and automatically provides a taxonomy and functional classification of the reads. For the functional classification KEGG classification was used. Sequence-based characterization allowed us to utilize the genome databases to parse the complexity of the complete ecosystem. Such insights and outcomes provided us the information about the functional capacity of the microbes surviving at sub-zero temperature.

For comparative purpose different metagenomes from MG-RAST (Meyer et al., 2008) database were used, which included Rotmoosferner glacier cryoconite from Austria (MG-RAST ID 4491734.3), Freshwater microcystis bloom metagenomes in China (MG-RAST ID – 4467058.3), marine water from the Mediterranean sea (MG-RAST ID – 4614521.3), and Loma Ridge grassland in Montana (MG-RAST ID – 4511137.3).

7.2.4 Analysis of Alaskan glacier

The data from Alaskan glacier metagenome was compared against NCBI non-redundant (Benson et al. 2009) database, BLASTX (Altschul et al. 1990) was used to query the NCBI-NR at a cutoff evalue of 10^{-5}. To filter out the low-complexity regions in the sequences, the SEG parameter was used. Low-complexity regions have an unusual composition that might give high scores to these regions creating difficulty in sequence similarity searching means, and thereby confuse the program to find the actual significant sequences in the database, so they should be filtered out. The results from BLASTX were

then used with MEGAN 4, which integrates the taxonomic and functional features of environmental sequence data. MEGAN 4 uses the KEGG classification (Kanehisa and Goto, 2000) where genes are mapped to KEGG orthology groups, which are mapped to enzymes present in different pathways. The data of metabolic potential of metagenomes from different environments for comparative analysis were obtained from MG-RAST database.

7.3 Chapter 4: Material and methods

7.3.1 Data sets

The glacier samples were collected from Harding Icefield in Alaska in August 2013. Three liters of snow were scooped from accumulation zone of the glacier when the temperature was around 4ºC, with wind 16 m/s, and rain at 2.3 mm/hr. After the samples melted at room temperature, the pre-filtering of samples with coffee filters was done to remove unwanted materials. Further filtering was done using a hand pump with 1.6 µm filter, in order to filter out eukaryotes. An electric suction pump was used with three durapore membrane filter of 0.22 µm in parallel for final separation of prokaryotes from the filtered (at 1.6 µm) water. DNA extraction was carried out using bead beating method plus column filtration according to the manufacturer's instructions (MO BIO Laboratories, Inc). After isolation, DNA was fragmented using Covaris S2 and divided into two tubes for Illumina MiSeq and Ion Proton library preparation. Final library was

enriched by running 15 PCR cycles prior to sequencing.

The data used in the analysis of different metagenome and same sequencer was taken from two very different metagenomes, leech gut (Maltz et al., 2014) and deep sea (Sogin et al., 2006), sequenced by the 454 to examine if there was difference in the different metagenomes sequenced by the same platform. The content of the intestinum and intraluminal fluid of the crop from leech fed one meal of heparinized sheep blood was sequenced via 454 (SRR1157610-11). The sea DNA samples for the analysis were collected from Atlantic Deep Water at a depth of 4 m; the already trimmed data was taken as described in the paper (Sogin et al., 2006). Moreover, the same human urine sample, which was sequenced, using two different platforms (MiSeq and Ion Torrent), was also studied (Salipante et al., 2014). The accession numbers of the data for the human urine samples were SRR1204944 (MiSeq) and SRR1205111 (Ion Torrent).

7.3.2 Data processing

Illumina MiSeq-

The data for glacier used in the study was shotgun metagenomic data, and not the 16S rRNA sequences. To access the taxonomic classification, reads from the data were compared against database comprised of V9 hypervariable sequences (refv9). Only 3,810,507 reads that passed the quality filtering and greater than 50 bp in length were taken for the comparison. The reads from data served as query to identify its closest match in self made reference database containing only V9 regions of 16S rRNA. 745

reads were found to match the V9 hypervariable region. The reference sequences were downloaded from The Visualization and Analysis of Microbial Population Structures (VAMPS) (Huse et al., 2014).

The data for the human urine metagenome was generated via Illumina MiSeq platform containing the V1-V2 hypervariable regions of the 16S rRNA gene. After quality filtering, only 14,912 V1-V2 hypervariable regions of 16S were extracted from 887,900 reads for human urine samples. The sequences that passed the Phred scores greater than 20 and did not contain any ‘N’s were retained.

Ion Torrent PGM/Ion Proton-

The glacier metagenomic data comparison was done in similar manner as the MiSeq data. Out of 12,052,104 quality filtered reads greater than 50 bp in length, only 2,430 V9 hypervariable regions remained.

643,464 sequences were downloaded from SRA generated through Ion Torrent platform of human urine samples. The 518 unique V1-V2 regions of 16S rRNA amplicons remained after quality filtering.

454/Roche GS-FLX-

The medicinal leech gut microbiota included content of intestinum. After quality filtering, 2,195 V6 hypervariable region of the SSU rRNA gene was extracted from 30,938 reads. Out of 9,282 trimmed reads from the lower deep water, only 5,279 unique V6 hypervariable regions were selected for the analysis.

7.3.3 Principal component analysis (PCA)

The key task when analyzing any metagenomic dataset is the assignment of anonymous metagenomic sequences to a diverse microbial population. In the current study we grouped the reads (representing hypervariable regions and sequenced by different sequencing platforms) based on dinucleotide patterns, also referred as dinucleotide word frequency. This study focused on how the nucleotide composition of sequences varies when different sequencing platforms are used for metagenome analysis. A K-dimensional feature vector (K=16) represented each DNA fragment, where each element in a vector encodes the dimer occurring in the fragment. To reduce the high dimensionality of feature space, the PCA was utilized which is an orthogonal linear transformation to highlight the differences and similarities among the data. With the PCA, the original data was transformed to a new set of variables, also known as principal components, ordered according to their corresponding variances, and we retained the three principal components that contribute most to the variance. Projections on these three principal components were plotted highlighting clustering of the sequences. We also analyzed the load factors of the first, second, and third principal components in order to see which frequencies were mainly contributing to the variance.

MOTHUR (Schloss et al., 2009) was utilized to describe the taxonomy of extracted hypervariable regions of 16S rRNA from different platforms. The taxonomy outline from SILVA (Pruesse et al., 2007) database was used to classify sequences into specific reference taxonomies. The k-nearest neighbor consensus and BLAST (Altschul

et al., 1990) approach was used for taxonomic classification of the reads.

7.4 Chapter 5: Material and methods

7.4.1 Data sets

The data was taken from SRA that was generated by the same sequencer however different DNA-extraction protocols were used for sample preparation. Out of four protocols analyzed in the study, only two methods were considered here for the analysis, methods Q and BB (SRP039007). The DNA was extracted from mock communities of seven representative oral bacteria, and sequenced via 454 (Abusleme et al., 2014). Next, we analyzed two very diverse metagenomes, human gut (Kennedy et al., 2014) and soil metagenomes (Kennedy et al., 2014), sequenced by the Illumina Miseq to study the sequence diversity of two different metagenomes sequenced by the same platform. Stool samples for gut metagenome have been studied previously in context of bacterial geography of human gastrointestinal tract (Stearns et al., 2011). A temperate deciduous forest (TDF) soil sample was selected from the Canadian MetaMicrobiome Library (Kennedy et al., 2014). In another study, both the gut and soil metagenome data have been used for evaluating bias of Illumina-based bacterial 16S rRNA gene profiles (Kennedy et al., 2014). The V3 region of 16S rRNA of human gut (ERR567417) and soil metagenome (ERR567426) was downloaded from SRA. In the study, the human gut data was referred as “human gut I” and soil data as “soil I”.

The other human gut data used in the analysis was taken from different studies that investigated the human gut microbiome using 454 GS FLX Titanium and Illumina MiSeq sequencing technologies. The hypervariable regions of 16S rRNA gene were sequenced using DNA extracted from various human gut samples. For each platform, sets of eight sequence data sets, SRR1029510-17 (454) and SRR1029468 -75 (Illumina) were downloaded from SRA database (Bioproject: PRJNA46315). This human gut was called as “human gut II” in the study. The different soil metagenomes data for the environmental sample used for analysis was generated from 454 GS FLX, Ion Torrent PGM, and Illumina Genome Analyzer IIx platforms. The accession numbers of the soil data downloaded from SRA were: SRX404651 (454), SRX481936 (Ion Torrent), and ERX093708 (Illumina) and in the study these soil metagenomes was identified as “soil II”.

These data sets contained sequences from various hypervariable regions of 16S rRNA and sequence composition of each region is different, which minimizes the possibility of direct comparisons. To overcome the limitation, we used the same hyper-variable region for the comparisons in each data to minimize any kind of additional bias in the analysis.

7.4.2 Data processing

All reads from the different platforms were quality filtered, and the sequences that passed the Phred scores greater than 20 and did not contain any ‘N’s were retained. Only

the unique hypervariable regions of the 16S rRNA gene from filtered reads were extracted, excluding the barcodes and primer sequences in order to compare different platforms based on unique hypervariable region. Table 7.4.1 represents a summary of the data screening details of each platform used in the study.

Table 7.4.1. Data processing details of NGS platforms used in this study

Platform	Illumina			454			Ion Torrent		
	Raw Reads	Quality Filter	HV	Raw Reads	Quality Filter	HV	Raw Reads	Quality Filter	HV
Human Gut I	44360[M]	19677[M]	V3[M]	-	-	-	-	-	-
Soil I	47685[M]	14604[M]	V3[M]	-	-	-	-	-	-
Human Gut II	358,773[M]	124[M]	V4[M]	154,374	3215	V4	-	-	-
Soil II	42,864[GA]	26,385[GA]	V5[GA]	729,514	11,349	V5	514,848	11,052	V5
DNA Extraction Q	-	-	-	12,317	1,050	V1-V2	-	-	-
DNA Extraction BB	-	-	-	13,724	3,243	V1-V2	-	-	-

GA - Illumina Genome Analyzer IIx
M - Illumina Miseq
HV - hypervariable regions of the SSU rRNA gene

7.4.3 Principal component analysis (PCA)

Oligonucleotide word frequencies are frequently used in metagenomic

applications (e.g., phylogenetic tree construction (Amann et al., 1995) metagenome binning (Deschavanne et at., 1999)). Here, we adopted a similar technique to analyze the sequence composition of hypervariable regions of 16S rRNA fragments using nucleotide word frequencies. We grouped different hypervariable regions of the reads sequenced from different sequencing platforms based on di-, tri and tetra nucleotide patterns. Subsequently, PCA was applied on the nucleotide frequencies by reducing the feature dimensionality to two while still retaining the most informative features. Each DNA fragment or read is represented by a K-dimensional feature vector (K=336), where each factor in a vector encodes the frequency of a particular di-, tri- or tetra-mer occurring in a read. The data are then transformed into principal components, which are ordered according to their corresponding variances. To observe similarities and differences, the projections on the two principal components are plotted, highlighting clustering of the sequences. The coefficients of variables (i.e., load factors), which in this case are the 336 word frequencies of the first principal components provided information about which nucleotide word frequencies were giving the maximum variance to the data. A histogram was generated with top 20 and bottom 20 load factors and their corresponding frequencies, and the GC-content was also calculated for each dataset.

The taxonomic classification of extracted hyper-variable regions of 16S rRNA was performed by MOTHUR (Schloss et al., 2009) using the taxonomy outline from SILVA database (Pruesse et al., 2007). The k-nearest neighbor consensus and BLAST (Altschul et al., 1990) approach was used for taxonomic classification of the reads.

7.4.4 Phylogenetic Heatmaps

Since PCA has no stochasticity constraints, the effectiveness of the clustering of different phylogenetic taxa in PCA was estimated by calculating the Pearson's correlation coefficient (r) of all the sequences with each other and taking the average correlation coefficient between the different groups within a sample and across sample. Correlation coefficients are not additive, therefore we transformed r values to Fisher Z transformation

$$Z = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$$

to make it into additive quantities. Once r values were converted into Z scores and an arithmetic mean score $\bar{Z}$ was computed, Fisher mean value $\bar{r}$ was calculated using the following equation:

$$\bar{r} = \frac{e^{\bar{Z}} - e^{-\bar{Z}}}{e^{\bar{Z}} + e^{-\bar{Z}}}$$

Furthermore, the arithmetic mean r for all group pairs were represented in a phylogenetic heatmap. A concentric heatmap was generated where the inner circle represented the correlation mean of same groups from different samples and the outer circle contained values of groups from same samples. For the outer circle only those group pairs were selected that were closer to each other in a PCA.

REFERENCES

Abusleme, L., B. Y. Hong, A. K. Dupuy, L. D. Strausbaugh, and P. I. Diaz. 2014. Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. Journal of Oral Microbiology 6 (Apr 23): 10.3402/jom.v6.23990. ECollection 2014.

Ahn, J. H., B. Y. Kim, J. Song, and H. Y. Weon. 2012. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. Journal of Microbiology (Seoul, Korea) 50 (6) (Dec): 1071-4.

Aird, D., M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. Genome Biology 12 (2): R18,2011-12-2-r18. Epub 2011 Feb 21.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215 (3) (Oct 5): 403-10.

Amato, P., R. Hennebelle, O. Magand, M. Sancelme, A. M. Delort, C. Barbante, C. Boutron, and C. Ferrari. 2007. Bacterial characterization of the snow cover at spitzberg, svalbard. FEMS Microbiology Ecology 59 (2) (Feb): 255-64.

Anesio, A. M., B. Sattler, C. Foreman, J. Telling, A. Hodson, M. Tranter, and R. Psenner. 2010. Carbon fuxes through bacterial communities on glacier surfaces. Annals of Glaciology 51 (56): 32-40.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, et al. 2011. Enterotypes of the human gut microbiome. Nature 473 (7346) (May 12): 174-80.

Beck, D. A., M. G. Kalyuzhnaya, S. Malfatti, S. G. Tringe, T. Glavina Del Rio, N. Ivanova, M. E. Lidstrom, and L. Chistoserdova. 2013. A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the methylococcaceae and the methylophilaceae. Peerj 1 (Feb 19): e23.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2009. Genbank. Nucleic Acids Research 37 (Database issue) (Jan): D26-31.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 (7218) (Nov 6): 53-9.

Bottrell, S. H., and M. Tranter. 2002. Sulphide oxidation under partially anoxic conditions at the bed of the haut glacier d'arolla, switzerland. Hydrological Processes 16 (12): 2363-8.

Bowman, J. P. Genomic analysis of psychrophilic prokaryotes. In psychrophiles: From biodiversity to biotechnology. *Margesin, R., Schinner, F., Marx, J.-C., Gerday, C. (Ed). Heidelberg: Springer-Verlag*: 265-84.

Cameron, K. A., A. J. Hodson, and A. M. Osborn. 2012. Structure and diversity of bacterial, eukaryotic and archaeal communities in glacial cryoconite holes from the arctic and the antarctic. FEMS Microbiology Ecology 82 (2) (Nov): 254-67.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, et al. 2012. Ultra-high-throughput microbial community analysis on the illumina HiSeq and MiSeq platforms. The ISME Journal 6 (8) (Aug): 1621-4.

Cardona, S., A. Eck, M. Cassellas, M. Gallart, C. Alastrue, J. Dore, F. Azpiroz, J. Roca, F. Guarner, and C. Manichanh. 2012. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC Microbiology 12 (Jul 30): 158,2180-12-158.

Chattopadhyay, M. K. 2002. The cryoprotective effects of glycine betaine on bacteria. Trends in Microbiology 10 (7): 311-.

Chattopadhyay, M. K. 2006. Mechanism of bacterial adaptation to low temperature. Journal of Biosciences 31 (1) (Mar): 157-65.

Chattopadhyay, M., and M. Jagannadham. 2001. Maintenance of membrane fluidity in antarctic bacteria. Polar Biology 24 (5): 386-8.

Cheng, S. M., and J. M. Foght. 2007. Cultivation-independent and -dependent characterization of bacteria resident beneath john evans glacier. FEMS Microbiology Ecology 59 (2) (Feb): 318-30.

Cheung, M. S., T. A. Down, I. Latorre, and J. Ahringer. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Research 39 (15) (Aug): e103.

Chintalapati, S., M. D. Kiran, and S. Shivaji. 2004. Role of membrane lipid fatty acids in cold adaptation. Cellular and Molecular Biology (Noisy-Le-Grand, France) 50 (5) (Jul): 631-42.

Choudhari, S., R. Lohia, and A. Grigoriev. 2014. Comparative metagenome analysis of an

alaskan glacier. Journal of Bioinformatics and Computational Biology 12 (2) (Apr): 1441003.

Choudhari, S., S. Smith, S. Owens, J. A. Gilbert, D. H. Shain, R. J. Dial, and A. Grigoriev. 2013. Metagenome sequencing of prokaryotic microbiota collected from byron glacier, alaska. Genome Announcements 1 (2) (Mar 21): e0009913-13.

Christner, B. C., E. Mosley-Thompson, L. G. Thompson, and J. N. Reeve. 2003. Bacterial recovery from ancient glacial ice. Environmental Microbiology 5 (5) (May): 433-6.

Claesson, M. J., Q. Wang, O. O'Sullivan, R. Greene-Diniz, J. R. Cole, R. P. Ross, and P. W. O'Toole. 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Research 38 (22) (Dec): e200.

Cleland, D., P. Krader, C. McCree, J. Tang, and D. Emerson. 2004. Glycine betaine as a cryoprotectant for prokaryotes. Journal of Microbiological Methods 58 (1) (Jul): 31-8.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, et al. 2009. The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* 37 (Database issue) (Jan): D141-5.

Cowan, D. A., and L. A. Tow. 2004. Endangered antarctic environments. Annual Review of Microbiology 58 : 649-90.

D'Amico, S., T. Collins, J. C. Marx, G. Feller, and C. Gerday. 2006. Psychrophilic microorganisms: Challenges for life. EMBO Reports 7 (4) (Apr): 385-9.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013. Special features of RAD sequencing data: Implications for genotyping. Molecular Ecology 22 (11) (Jun): 3151-64.

Deschavanne, P. J., A. Giron, J. Vilain, G. Fagot, and B. Fertil. 1999. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution* 16 (10) (Oct): 1391-9.

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36 (16) (Sep): e105.

Dykhuizen, D. E. 1998. Santa rosalia revisited: Why are there so many species of bacteria?

Antonie Van Leeuwenhoek 73 (1) (Jan): 25-33.

Edwards, A., JA Pachebat, M. Swain, M. Hegarty, AJ Hodson, TDL Irvine-Fynn, SME Rassner, and B. Sattler. 2013. A metagenomic snapshot of taxonomic and functional diversity in an alpine glacier cryoconite ecosystem. Environmental Research Letters 8 (3).

Feller, G., and C. Gerday. 2003. Psychrophilic enzymes: Hot topics in cold adaptation. Nature Reviews.Microbiology 1 (3) (Dec): 200-8.

Foght, J., J. Aislabie, S. Turner, C. E. Brown, J. Ryburn, D. J. Saul, and W. Lawson. 2004. Culturable bacteria in subglacial sediments and ice from two southern hemisphere glaciers. Microbial Ecology 47 (4) (May): 329-40.

Frank-Fahle, B. A., E. Yergeau, C. W. Greer, H. Lantuit, and D. Wagner. 2014. Microbial functional potential and community composition in permafrost-affected soils of the NW canadian arctic. PloS One 9 (1) (Jan 8): e84761.

Franzetti, A., V. Tatangelo, I. Gandolfi, V. Bertolini, G. Bestetti, G. Diolaiuti, C. D'Agata, C. Mihalcea, C. Smiraglia, and R. Ambrosini. 2013. Bacterial community structure on two alpine debris-covered glaciers and biogeography of polaromonas phylotypes. *The ISME Journal* 7 (8) (Aug): 1483-92.

Ghyselinck, J., S. Pfeiffer, K. Heylen, A. Sessitsch, and P. De Vos. 2013. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. PloS One 8 (8) (Aug 19): e71360.

Gilbert, J. A., and M. Hughes. 2011. Gene expression profiling: Metatranscriptomics. Methods in Molecular Biology (Clifton, N.J.) 733 : 195-205.

Gilbert, J. A., F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, et al. 2010. Meeting report: The terabase metagenomics workshop and the vision of an earth microbiome project. *Standards in Genomic Sciences* 3 (3) (Dec 25): 243-8.

Gotelli, N., and R. Colwell. 2001. Quantifying biodiversity: Procedures and pitfalls in measurement and comparison of species richness. *Ecology Letters* 4 (181): 184.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry & Biology 5 (10) (Oct): R245-9.

Harismendy, O., P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10 (3): R32,2009-10-3-r32. Epub 2009 Mar 27.

Huse, S. M., D. B. Mark Welch, A. Voorhis, A. Shipunova, H. G. Morrison, A. M. Eren, and M. L. Sogin. 2014. VAMPS: A website for visualization and analysis of microbial population structures. BMC Bioinformatics 15 (Feb 5): 41,2105-15-41.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. 2007. MEGAN analysis of metagenomic data. Genome Research 17 (3) (Mar): 377-86.

Johnson, S. S., M. B. Hebsgaard, T. R. Christensen, M. Mastepanov, R. Nielsen, K. Munch, T. Brand, et al. 2007. Ancient bacteria show evidence of DNA repair. Proceedings of the National Academy of Sciences of the United States of America 104 (36) (Sep 4): 14401-5.

Kalyuzhnaya, M. G., A. Lapidus, N. Ivanova, A. C. Copeland, A. C. McHardy, E. Szeto, A. Salamov, et al. 2008. High-resolution metagenomics targets specific functional types in complex microbial communities. Nature Biotechnology 26 (9) (Sep): 1029-34.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28 (1) (Jan 1): 27-30.

Karl, D. M., D. F. Bird, K. Bjorkman, T. Houlihan, R. Shackelford, and L. Tupas. 1999. Microorganisms in the accreted ice of lake vostok, antarctica. Science (New York, N.Y.) 286 (5447) (Dec 10): 2144-7.

Kasting, J. F., and J. L. Siefert. 2002. Life and the evolution of earth's atmosphere. Science (New York, N.Y.) 296 (5570) (May 10): 1066-8.

Kembel, S. W., M. Wu, J. A. Eisen, and J. L. Green. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology* 8 (10): e1002743.

Kennedy, K., M. W. Hall, M. D. Lynch, G. Moreno-Hagelsieb, and J. D. Neufeld. 2014. Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. Applied and Environmental Microbiology 80 (18) (Sep): 5717-22.

Kiran, M. D., S. Annapoorni, I. Suzuki, N. Murata, and S. Shivaji. 2005. Cis-trans isomerase gene in psychrophilic pseudomonas syringae is constitutively expressed during growth

and under conditions of temperature and solvent stress. Extremophiles : Life Under Extreme Conditions 9 (2) (Apr): 117-25.

Klappenbach, J. A., P. R. Saxman, J. R. Cole, and T. M. Schmidt. 2001. Rrndb: The ribosomal RNA operon copy number database. Nucleic Acids Research 29 (1) (Jan 1): 181-4.

Kol E. 1968. Kryobiologie: Biologie und Limnologie des Schnees und Eises. 1Kryovegetation 24: Schweizerbart'sche (Nagele und Obermiller), Stuttgart.

Kozarewa, I., Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. 2009. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods 6 (4) (Apr): 291-5.

Lee, K. B., C. T. Liu, Y. Anzai, H. Kim, T. Aono, and H. Oyaizu. 2005. The hierarchical system of the 'alphaproteobacteria': Description of hyphomonadaceae fam. nov., xanthobacteraceae fam. nov. and erythrobacteraceae fam. nov. International Journal of Systematic and Evolutionary Microbiology 55 (Pt 5) (Sep): 1907-19.

Leinonen, R., H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Research* 39 (Database issue) (Jan): D19-21.

Lewin, A., A. Wentzel, and S. Valla. 2013. Metagenomics of microbial life in extreme temperature environments. Current Opinion in Biotechnology 24 (3) (Jun): 516-25.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. 2012. Comparison of next-generation sequencing systems. Journal of Biomedicine & Biotechnology 2012 : 251364.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis. 2012. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community DNA sample. PloS One 7 (2): e30087.

Lutz, S., A. M., A. Edwards, and L. G. Benning. 2015. Microbial diversity on glaciers and ice caps. Frontiers in Microbiology 6 (307).

Mack, M. C., E. A. Schuur, M. S. Bret-Harte, G. R. Shaver, and F. S. Chapin. 2004. Ecosystem carbon storage in arctic tundra reduced by long-term nutrient fertilization. Nature 431 (7007) (Sep 23): 440-3.

Mackelprang, R., M. P. Waldrop, K. M. DeAngelis, M. M. David, K. L. Chavarria, S. J. Blazewicz, E. M. Rubin, and J. K. Jansson. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480 (7377) (Nov 6): 368-71.

Maltz, M. A., L. Bomar, P. Lapierre, H. G. Morrison, E. A. McClure, M. L. Sogin, and J. Graf. 2014. Metagenomic analysis of the medicinal leech gut microbiota. Frontiers in Microbiology 5 (Apr 17): 151.

Martineau, C., L. G. Whyte, and C. W. Greer. 2010. Stable isotope probing analysis of the diversity and activity of methanotrophic bacteria in soils from the canadian high arctic. Applied and Environmental Microbiology 76 (17) (Sep): 5773-84.

Mavromatis, K., N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods 4 (6) (Jun): 495-500.

Maxam, A. M., and W. Gilbert. 1992. A new method for sequencing DNA. 1977. Biotechnology (Reading, Mass.) 24 : 99-103.

Methe, B. A., K. E. Nelson, J. W. Deming, B. Momen, E. Melamud, X. Zhang, J. Moult, et al. 2005. The psychrophilic lifestyle as revealed by the genome sequence of colwellia psychrerythraea 34H through genomic and proteomic analyses. Proceedings of the National Academy of Sciences of the United States of America 102 (31) (Aug 2): 10913-8.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9 (Sep 19): 386,2105-9-386.

Miteva, V. 2008. Bacteria in snow and glacier ice. In . Vol. 2, 31-50.

Miteva, V. I., P. P. Sheridan, and J. E. Brenchley. 2004. Phylogenetic and physiological diversity of microorganisms isolated from a deep greenland glacier ice core. Applied and Environmental Microbiology 70 (1) (Jan): 202-13.

Miteva, V., K. , T. Sowers, A. Sebastian, and J. Brenchley. 2015. Abundance, viability and diversity of the indigenous microbial populations at different depths of the NEEM ice core. Polar Research

Møller, A., D. Søborg, W. Al-Soudm, S. Sørensen, and N. Kroer. 2013. Bacterial community structure in high-arctic snow and freshwater as revealed by pyrosequencing of 16S rRNA genes and cultivation. Polar Research, North America 32 : 17390,.

Molnia BF, Late nineteenth to early twenty first century behavior of Alaskan glaciers as indicators of changing regional climate, Global and Planet Change 56:23–56, 2007

Morgan, J. L., A. E. Darling, and J. A. Eisen. 2010. Metagenomic sequencing of an in vitro-simulated microbial community. PloS One 5 (4) (Apr 16): e10209.

Nazaries, L., Y. Pan, L. Bodrossy, E. M. Baggs, P. Millard, J. C. Murrell, and B. K. Singh. 2013. Evidence of microbial regulation of biogeochemical cycles from a study on methane flux and land use change. Applied and Environmental Microbiology 79 (13) (Jul): 4031-40.

Newman, D. K., and J. F. Banfield. 2002. Geomicrobiology: How molecular-scale interactions underpin biogeochemical systems. Science (New York, N.Y.) 296 (5570) (May 10): 1071-7.

Paterson, W. S. 1994. The physics of glaciers; . Tarrytown, New York, Pergamon/Elsevier Science, Inc. 3 : 480.

Price, P. B. 2000. A habitat for psychrophiles in deep antarctic ice. *Proceedings of the National Academy of Sciences of the United States of America* 97 (3) (Feb 1): 1247-51.

Priscu JC, Christner BC. 2004. Earth's icy biosphere. in: Bull AT (ed) microbial diversity and bioprospecting. ASM Press, Washington, DC, Pp: 130-45.

Priscu, J. C., E. E. Adams, W. B. Lyons, M. A. Voytek, D. W. Mogk, R. L. Brown, C. P. McKay, et al. 1999. Geomicrobiology of subglacial ice above lake vostok, antarctica. Science (New York, N.Y.) 286 (5447) (Dec 10): 2141-4.

Prosser, J. I., B. J. Bohannan, T. P. Curtis, R. J. Ellis, M. K. Firestone, R. P. Freckleton, J. L. Green, et al. 2007. The role of ecological theory in microbial ecology. Nature Reviews.Microbiology 5 (5) (May): 384-92.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35 (21): 7188-96.

Reysenbach, A. L., and E. Shock. 2002. Merging genomes with geochemistry in hydrothermal ecosystems. Science (New York, N.Y.) 296 (5570) (May 10): 1077-82.

Rogers, S. O., Y. M. Shtarkman, Z. A. Kocer, R. Edgar, R. Veerapaneni, and T. D'Elia. 2013. Ecology of subglacial lake vostok (antarctica), based on metagenomic/metatranscriptomic analyses of accretion ice. Biology 2 (2) (Mar 28): 629-50.

Salipante, S. J., T. Kawashima, C. Rosenthal, D. R. Hoogestraat, L. A. Cummings, D. J. Sengupta, T. T. Harkins, B. T. Cookson, and N. G. Hoffman. 2014. Performance comparison of illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Applied and Environmental Microbiology 80 (24) (Dec): 7583-91.

Sanger, F., S. Nicklen, and A. R. Coulson. 1992. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology (Reading, Mass.) 24 : 104-8.

Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, et al. 2009. Database resources of the national center for biotechnology information. Nucleic Acids Research 37 (Database issue) (Jan): D5-15.

Schlesner, H., C. Jenkins, and J. T. Staley. 2006. The phylum verrucomicrobia: A phylogenetically heterogeneous bacterial group. Prokaryotes 7 : 881-96.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75 (23) (Dec): 7537-41.

Schutte, U. M., Z. Abdo, J. Foster, J. Ravel, J. Bunge, B. Solheim, and L. J. Forney. 2010. Bacterial diversity in a glacier foreland of the high arctic. Molecular Ecology 19 Suppl 1 (Mar): 54-66.

Segawa, T., and N. Takeuchi. 2010. Cyanobacterial communities on qiyi glacier, qilian shan, china. Annals of Glaciology 51 (56): 153-62.

Shah, N., H. Tang, T. G. Doak, and Y. Ye. 2011. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing: 165-76.

Shaver, G.R., and Chapin III, F. S. 1980. Response to Fertilization by Various Plant Growth

Forms in an Alaskan Tundra: Nutrient Accumulation and Growth. Ecology 61:662–675. http://dx.doi.org/10.2307/1937432.

Sheridan, P. P., V. I. Miteva, and J. E. Brenchley. 2003. Phylogenetic analysis of anaerobic psychrophilic enrichment cultures obtained from a greenland glacier ice core. Applied and Environmental Microbiology 69 (4) (Apr): 2153-60.

Siegert, M. J., J. C. Ellis-Evans, M. Tranter, C. Mayer, J. R. Petit, A. Salamatin, and J. C. Priscu. 2001. Physical, chemical and biological processes in lake vostok and other antarctic subglacial lakes. Nature 414 (6864) (Dec 6): 603-9.

Simon, C., A. Wiezer, A. W. Strittmatter, and R. Daniel. 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. Applied and Environmental Microbiology 75 (23) (Dec): 7519-26.

Simon, C., and R. Daniel. 2011. Metagenomic analyses: Past and future trends. Applied and Environmental Microbiology 77 (4) (Feb): 1153-61.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews.Genetics 15 (2) (Feb): 121-32.

Skidmore, M., S. P. Anderson, M. Sharp, J. Foght, and B. D. Lanoil. 2005. Comparison of microbial community compositions of two subglacial environments reveals a possible role for microbes in chemical weathering processes. Applied and Environmental Microbiology 71 (11) (Nov): 6986-97.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America 103 (32) (Aug 8): 12115-20.

Stearns, J. C., M. D. Lynch, D. B. Senadheera, H. C. Tenenbaum, M. B. Goldberg, D. G. Cvitkovitch, K. Croitoru, G. Moreno-Hagelsieb, and J. D. Neufeld. 2011. Bacterial biogeography of the human digestive tract. Scientific Reports 1 : 170.

Stibal, M., M. Tranter, L. G. Benning, and J. Rehak. 2008. Microbial primary production on an arctic glacier is insignificant in comparison with allochthonous organic carbon input. *Environmental Microbiology* 10 (8) (Aug): 2172-8.

Suutari, M., and S. Laakso. 1994. Microbial fatty acids and thermal adaptation. Critical

Reviews in Microbiology 20 (4): 285-328.

Szatkiewicz, J. P., W. Wang, P. F. Sullivan, W. Wang, and W. Sun. 2013. Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. Nucleic Acids Research 41 (3) (Feb 1): 1519-32.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28 (1) (Jan 1): 33-6.

van Heesch, S., M. Mokry, V. Boskova, W. Junker, R. Mehon, P. Toonen, E. de Bruijn, et al. 2013. Systematic biases in DNA copy number originate from isolation procedures. Genome Biology 14 (4) (Apr 24): R33,2013-14-4-r33.

von Mering, C., P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science (New York, N.Y.) 315 (5815) (Feb 23): 1126-30.

Weiss, R. L. 1983. Fine structure of the snow alga (chlamydomonas nivalis) and associated bacteria. Journal of Phycology 19: 200-4.

Willerslev, E., A. J. Hansen, and H. N. Poinar. 2004. Isolation of nucleic acids and cultures from fossil ice and permafrost. Trends in Ecology & Evolution 19 (3) (Mar): 141-7.

Wolf, J. B., and J. Bryk. 2011. General lack of global dosage compensation in ZZ/ZW systems? broadening the perspective with RNA-seq. BMC Genomics 12 (Feb 1): 91,2164-12-91.

Yergeau, E., H. Hogues, L. G. Whyte, and C. W. Greer. 2010. The functional potential of high arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. The ISME Journal 4 (9) (Sep): 1206-14.