Microbial Responses to Environmental Change in Canada's High Arctic

Graham Colby

Thesis submitted in partial fulfillment of the requirements for the degree of Masters in Biology - Bioinformatics Specialization [1]

Department of Biology
Faculty of Science
University of Ottawa

[1]The program is a joint program with Carleton University.

Abstract

The Arctic is undergoing a rapid environmental shift with increasing temperatures and precipitations expected to continue over the next century. Yet, little is known about how microbial communities and their underlying metabolic processes will respond to ongoing climatic changes. To address this question, we focused on Lake Hazen, NU, Canada. As the largest High Arctic lake by volume, it is a unique site to investigate microbial responses to environmental changes. Over the past decade, glacial coverage of the lake has declined. Increasing glacial runoff and sedimentation rates in the lake has resulted in differential influx of nutrients through spatial gradients. I used these spatial gradients to study how environmental changes might affect microbial community structure and functional capacity in Arctic lakes. I performed a metagenomic analysis of microbial communities from hydrological regimes representing high, low, and negligible influence of glacial runoff and compared the observed structure and function to the natural geochemical gradients. Genes and reconstructed genomes found in different abundances across these sites suggest that high-runoff regimes alter geochemical gradients, homogenise the microbial structure, and reduce genetic diversity. This work shows how a genome-centric metagenomics approach can be used to predict future microbial responses to a changing climate.

Résumé

L'Arctique subit actuellement des changements environnementaux rapides causés par l'augmentation des températures et des précipitations dont l'effet devrait continuer à croitre de façon exponentielle au cours du siècle prochain. Cependant, l'information concernant la manière dont les communautés microbiennes et leurs processus métaboliques sous-jacents réagiront aux changements climatiques est limitée. Pour répondre à cette question, nous nous sommes concentrés sur le Lac Hazen, NU, Canada. Le Lac Hazen est le plus grand lac par volume situé au nord du cercle arctique et représente un site unique pour étudier la réponse des microorganismes aux changements environnementaux. Notamment, au cours des dix dernières années, la couverture de glace à la surface du lac a diminué, entrainant une augmentation du ruissellement glaciaire et des taux de sédimentation dans certaines régions du lac. Ceci a résulté en un apport différentiel de nutriments, créant des gradients spatiaux au sein même du lac. Ici, nous avons utilisé ces gradients spatiaux pour étudier l'impact potentiel des changements environnementaux sur la structure et la capacité fonctionnelle des communautés microbiennes. Nous avons effectué une analyse métagénomique des communautés microbiennes au travers d'un gradient de régimes hydrologiques représentants une influence élevée, faible, et négligeable du ruissellement glaciaire en provenance du bassin versant. L'analyse des génomes assemblés à partir des données métagénomiques suggère que les régimes à fort ruissellement modifient les gradients géochimiques, homogénéisent la structure microbienne et réduisent

la diversité génétique. Ce travail montre qu'une approche métagénomique peut être utilisée pour prédire la réponse des microorganismes aux changements climatiques.

Acknowledgements

I would like to thank my thesis supervisors, Dr. Alexandre Poulain and Dr. Stéphane Aris-Brosou, for their guidance, feedback, and continued support during this process. Your sustained enthusiasm, even when mine was wavering, was the strongest motivational force during my research. I believe that you have both been responsible for my growth as a person and a scientist throughout this experience. To my committee members, Theodore Perkins and Alex Wong, your advice and insights were instrumental. Alex Wong, I can now say with certainty your course was one of my favourite graduate experiences.

To my lab members, thank you for the good company, necessary morale support, and mandatory fun. Mija, Michelle-Claire, Matti, and Galen, I appreciate all the conversations we had about data analysis. I am sure my thesis is better for it. I would also like to extend a thank you to all of the biology graduate students that made my experience here more memorable.

To Greg Slater and Ben Evans from McMaster University – my time spent in your respective labs as an undergraduate student led me to this point. I am certain that my interest in environmental science and bioinformatics stems from my conversations with you as a wide-eyed young scientist.

Finally, to my friends and family – graduate school was not always easy. Your support and encouragement through my times of stress made all of this possible.

Contents

List of Figures

List of Tables

List of Abbreviations

AMR: Antimicrobial resistance

CBB: Calvin-Benson-Bassham pathway

COG: Clusters of orthologous groups of proteins

DNRA: Dissimilarity nitrate reduction

DPCoA: Double principle coordinate analysis

GTDB: Genome taxonomy database

HMM: Hidden Markov model

KEGG: Kyoto encyclopedia of genes and genomes

MAG: Metagenome assembled genome

MIMAG: Minimum information about metagenome-assembled genome

NCBI: National Center for Biotechnology Information

NMDS: Non-metric multidimensional scaling

OC: Organic carbon

PAM: Partitioning around medoids

PCA: Principal component analysis

PCoA: Principal coordinates analysis

PFAM: Protein families database

TIGRFAM: The Institute for Genomic Research's Database of Protein Families

t-SNE: t-distributed stochastic neighbor embedding

Chapter 1

Metagenomic applications to Arctic microbial ecology

1.1 General Introduction

Microbial communities shape Earth's ecosystems (Cavalier-Smith et al., 2006), and drive geochemical cycles (Mason et al., 2009); however, their precise roles in ecosystem services are only beginning to be understood (Sogin et al., 2006). Shotgun metagenomic approaches have allowed microbial ecologists to overcome the need for cultivation when characterising and identifying microorganisms. Analysing genomic DNA content from entire microbial communities and reconstructing microbial genomes from diverse environments provides novel insights into community structure and function that were previously unnoticed in 16S rRNA gene studies (Nesme et al., 2016; Shendure et al., 2017; Klindworth et al., 2013; Tremblay et al., 2015). As Arctic conditions are rapidly changing with increased temperature and precipitation (Screen and Simmonds, 2010), it is increasingly important that we understand how terrestrial and aquatic microbial ecosystems will respond to these changing conditions. Understanding ecosystem functioning and predicting responses to global climate change, requires

greater knowledge of microbial processes and their environmental interactions. This introductory chapter discusses the application of metagenomics in an environmental context, explains the vulnerability of Arctic ecosystems, and outlines the current understanding of the Lake Hazen watershed.

1.2 Environmental Metagenomics

Advances in both sequencing and bioinformatic technologies have transformed our capacity to investigate the structural and functional dynamics of complex microbial communities, from human guts to deep ocean sediments (Knight et al., 2018). Identification of microorganisms and evaluation of their functional capacities has historically relied on laboratory cultivation of individual species. Culture-independent studies, using 16S rRNA genes, improved our understanding of microbial diversity, identifying previously unknown bacteria and archaea lineages (Pace, 1997; Hugenholtz et al., 1998). Full-length 16S rRNA gene sequences have become standard practice to study microbial diversity because of the genes ubiquity in prokaryotes, slow evolutionary change, and large size (1,500 bp) (Janda and Abbott, 2007).

The analysis of marker gene data begins with the construction of operational taxonomic units (OTUs): clusters of reads that differ by less than a fixed sequence dissimilarity, most often 3% (Callahan et al., 2017), as illustrated in Figure 1.1. OTUs are typically constructed through *de novo* methods where reads are grouped by pairwise sequence similarities, and a representative sequence is mapped to a sufficiently similar sequence in a reference database such as RDP (Maidak et al., 2000), Greengenes (DeSantis et al., 2006), or SILVA (Quast et al., 2012). However, delineation of OTUs is an emergent feature of datasets, dependent on the relative abundances of the sampled community (Callahan et al., 2017). Even with infinite sequencing ability, two different OTU datasets are challenging to compare (Edgar, 2018). Some have suggested abandoning the 97% threshold previously used to define OTUs for

100% sequence variants, where each distinct sequence defines a separate OTU, thus greatly improving reusability and reproducibility across datasets (Callahan et al., 2017; Edgar, 2018). Although 16S rRNA gene sequencing is highly useful in regards to bacterial classification, it has low phylogenetic power at the species level, as it is reliant on a single gene (Janda and Abbott, 2007).

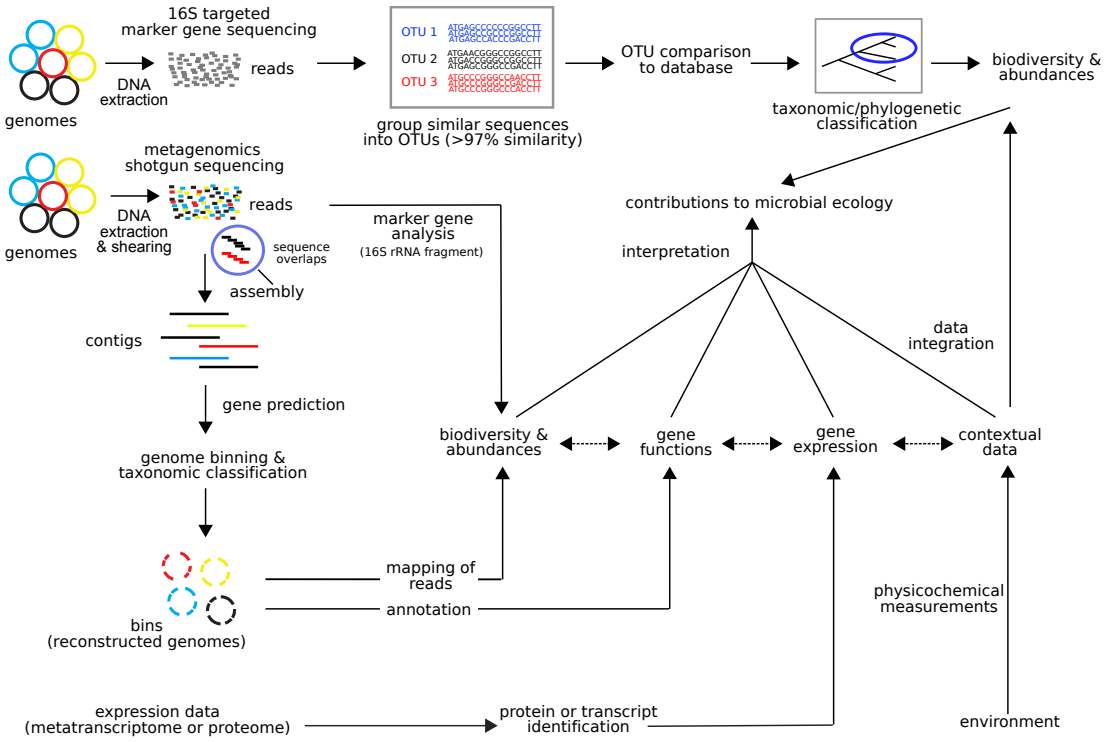

Figure 1.1: Comparison of methods for analysis of microbial communities. Both 16S rRNA gene sequencing analysis and shotgun metagenomics sequencing analysis are explained in the flow diagram. Adapted from Glockner and Teeling (2012).

Despite technological advances, microbial diversity is not completely described by 16S rRNA gene sequencing as a result of primer bias (Klindworth et al., 2013; Tremblay et al., 2015). In fact, nearly 10% of environmental sequences, most Patescibacteria (Candidate Phyla Radiation), and many archaea are missed by these classical techniques (Eloe-Fadrosh et al., 2016). Even more limiting is the fact that functional

interpretations of 16S rRNA gene sequences can only infer metabolic capacities, using a phylogenetic tree to link OTUs with gene content of similarly characterised organisms (Langille et al., 2013; Aßhauer et al., 2015). These procedures can be problematic when analysing microbial communities with large proportions of poorly characterised organisms (Glockner and Teeling, 2012; Jovel et al., 2016), as the accuracy the functional characterisation is tied to how well the genomes are represented in sequence databases (Thomas et al., 2012). Shotgun metagenomics (henceforth described as metagenomics) is an alternative approach that allows for the near-complete reconstruction of entire microbial genomes and evaluation of their functional capacities (Sharon and Banfield, 2013).

Metagenomic sequencing, like 16S rRNA gene sequencing, relies on the extraction of DNA; however, instead of targeting a specific genomic locus, all DNA is sheared into fragments and "universal adapters" are appended. Each small fragment is sequenced, resulting in reads that align to various genomic locations for the numerous genomes present (Thomas et al., 2012). In contrast to 16S rRNA gene sequencing that uses prokaryotic specific primers, metagenomic reads are more representative of the whole ecosystem capturing prokaryotic, viral, and eukaryotic DNA (Wooley et al., 2010). Metagenomic reads are generated from multiple genes that can be either taxonomically informative or encode functions, providing the opportunity to explore two aspects of the microbial community: who is there and what are they capable of doing? An added benefit of metagenomic sequencing is that it captures 16S rRNA genes and can be used to estimate microbial abundances and diversity (Pericard et al., 2017). Metagenomic sequencing typically generates read lengths 100-300 bp long and mostly over 99.9% accurate (Caporaso et al., 2012). This thesis used a Illumina HiSeq 2500 instrument to generate reads that were approximately 125 bp in length. However, it is worth mentioning newer technologies, such as PacBio's Sequel and Oxford Nanopore Technologies PromethION or MinION, that are capable of generating ultra-long reads approximately 800-900 kb in length (Jain et al., 2018). Ultra-long reads provide suf-

ficient length to capture entire prokaryotic genomes (Caporaso et al., 2012), which can improve sequence assembly and binning (van Dijk et al., 2018). Longer reads are particularly useful when trying to reconstruct regions containing highly repetitive transposable elements that exhibit high sequence identity, high copy number, or complex genomic arrangements (McCoy et al., 2014). As current binning methods rely heavily on using tetranucleotide frequencies as a distinguishing feature, it is challenging to group regions that appear as similar structural variations across all phylogenetic groups

A standard metagenomic workflow (Figure 1.1) involves the following steps: assembling reads into contigs, gene prediction of contigs, binning of contigs, taxonomic classification of bins, annotation of genes, and mapping of initial reads to the contigs. Assembling short reads into larger contigs is completed through *de novo* assemblers that combine reads on the basis of sequence overlaps (van der Walt et al., 2017). Contigs are scanned for open reading frames and gene locations are predicted (Hyatt et al., 2010). Binning typically uses tetranucleotide frequency and coverage across samples to group each contig into reconstructed genomes (also known as "bins" or "MAGs") (Alneberg et al., 2014; Wu et al., 2015). Binning is powerful process because it does not depend on a reference database or taxonomy to cluster contigs (Sedlar et al., 2017). High-quality reconstructed genomes obtained through binning provides detailed information on taxonomy, metabolic potential, and organism abundances in assemblages (Hug et al., 2018). Depending on the sequencing coverage (amount of reads generated) these reconstructed genomes can be taxonomically classified down to the species and even strain (Luo et al., 2015). The genes contained within each genome can be functionally annotated by searching a protein sequence against multiple databases using HMMER (Mistry et al., 2013) or BLAST (Altschul et al., 1997). The original sequences are mapped to the reconstructed genomes to approximate the biodiversity and abundance of each genome in the assemblage (Langmead and Salzberg, 2012).

Reconstructed genomes have indicated that microorganisms are heavily dependent on symbiotic and commensal relationships (Hug et al., 2018). This is particularly true for many of the unculturable organisms in the bacterial candidate phylum radiation (CPR) group and archaeal DPANN group (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) that engage in metabolic handoffs where multiple organisms perform sequential redox reactions (Brown et al., 2015; Castelle et al., 2015). Near-complete reconstructed genomes are suitable for phylogenetic and metabolic analyses, providing a key starting point to begin monitoring and accessing the environmental impacts on microbial communities (Anantharaman et al., 2016; Linz et al., 2018). Reconstructed genomes have dramatically shifted how we view the tree of life, uncovering novel phyla and species (Hug et al., 2016). The diversity of microorganisms, particularly those that are unculturable, has dramatically expanded with our increasing ability to reconstruct genomes from environmental metagenomic data (Parks et al., 2017). However, reconstructed genomes must be critically evaluated because they are typically incomplete (<100%) and potentially contain contigs from multiple strains or species due to assembly and binning processes (Parks et al., 2015). Reconstructed genomes form the basis from which we investigate microbes within this thesis.

To classify reconstructed genomes, the most common approach depends on NCBI taxonomy and subset of core ribosomal proteins (usually at least 16) that are found across all domains of life (Hug et al., 2013, 2016; Dombrowski et al., 2018; Probst et al., 2018). This method can be problematic as the NCBI taxonomy is often inconsistent with true evolutionary relationships, which is observed by the formation of polyphyletic groups (Parks et al., 2018). Recently, the expansion in sequenced and reconstructed genomes has made genome-based phylogenies possible improving the classification of metagenomic and uncultured bacterial datasets (Parks et al., 2017). This approach builds upon multilocus sequence alignment (MLSA), where sequences for multiple housekeeping genes were generated from isolated cultures, then concate-

nated and used to construct phylogenetic trees (Larsen et al., 2012). Genome-based phylogenies identifies prokaryotic organisms at a higher resolution than 16S rRNA gene-based studies, and can determine identity down to strains (Parks et al., 2018; Glaeser and Kämpfer, 2015). In this thesis, I used the Genome Taxonomy Database (GTDB), which relies on hundreds of ubiquitous single-copy proteins to classify taxonomy (Parks et al., 2018).

Previously, functional contributions of rare and uncultured bacteria would have remained elusive; however, recent metagenomics studies have characterised microbial functions in environmental systems (Wrighton et al., 2012; Rinke et al., 2013; Sunagawa et al., 2015; Anantharaman et al., 2016). Genes can be identified from reconstructed genomes (Hyatt et al., 2010). Once proteins have been predicted, genes and pathways can be annotated using databases such as TIGRFAM (Haft et al., 2003), PFAM (Finn et al., 2015), COG (Tatusov et al., 2003), KEGG (Kanehisa et al., 2015), or Metacyc (Caspi et al., 2007). These functional insights have recently been applied to understand the microbial role in biogeochemical cycles (Anantharaman et al., 2016; Linz et al., 2018; Probst et al., 2018). Within complex microbial systems few organisms can conduct sequential redox reactions (van Kessel et al., 2015), and most operate through interdependence of metabolism with other microorganisms (Anantharaman et al., 2016; Hug et al., 2018). However, growing evidence suggests that changing environmental conditions can change microbial assembleges and alter biogeochemical cycles (Hultman et al., 2015; Anantharaman et al., 2016). This has the potential to create a feedback loop wherein microbial communities respond to environmental conditions and their responses can fundamentally alter underlying biogeochemical processes. To the best of our knowledge, little work has been conducted investigating the microbial responses to environmental change, particularly in Canada's High Arctic.

1.3 Arctic Ecosystems

The Arctic is warming at unprecidents rates, with climate models predicting temperatures and precipitation increasing in the Arctic by as much as 8°C and 40%, respectively, by the year 2100 (IPCC, 2013). In fact, changes in near-surface air temperatures have been almost twice the global average in recent decades (Serreze and Francis, 2006). Reductions in sea ice coverage is largely implicated in this Arctic amplification warming process (Screen and Simmonds, 2010), and may also contribute to rising precipitation (Kopec et al., 2016). In aquatic ecosystems, there has been a reduction in the duration of lake ice coverage, increased nutrient delivery and primary productivity (Williamson et al., 2009). These environmental changes in the Arctic may have permanent detrimental effects on biodiversity, which has led to the formation of the Arctic Council – an intergovernmental forum investigating and establishing guidelines for the preservation of Arctic biodiversity (Meltofte et al., 2013).

While recent work has investigated species adaptation and loss due to ongoing environmental changes at mid latitudes (Pacifici et al., 2015; Kerr et al., 2015; Pecl et al., 2017), little is known about how complex microbial communities will respond. Microbial communities are commonly thought to be well adapted to climate change (McCalley et al., 2014; Hultman et al., 2015). Functions of individual microorganisms within communities are conventionally thought to be highly dynamic and collectively redundant, such that environmental changes will not impact overall function (Ramírez-Flandes et al., 2019). However, there is growing concern that as different assemblages of organisms are selected, linkages among the major biogeochemical cycles will change (Anantharaman et al., 2016). As microbial communities mediate most of the biogeochemical cycling in aquatic and terrestrial environments (Falkowski et al., 2008; Fuhrman, 2009), there is an urgent need to understand how these microbial ecosystems will change.

Arctic biodiversity is an irreplaceable ecological, scientific, economic, and cultural

asset; yet, the most significant organisms in maintaining the integrity and ecosystem services of the Arctic are the least understood. Microorganisms have fundamental roles in Arctic ecosystems where they are responsible for the production of methane (Woodcroft et al., 2018), mobilisation of metals (St. Pierre et al., 2018), and transformation of ammonia to nitrites (Alves et al., 2013), but they have been studied the least (Mohit et al., 2017; Hutchins and Fu, 2017). To date most of the work completed on Arctic microbial communities has leveraged 16S rRNA gene sequences to analyse the structure of microbial communities (Mohit et al., 2017; Comte et al., 2018; Ruuskanen et al., 2018). In particular, Arctic lakes are understudied, which may be due to logistical constraints of being ice covered 10-12 months of the year (Keatley et al., 2007). A few studies have used genome-centric approaches to investigate microbial communities in permafrost (Hultman et al., 2015; Woodcroft et al., 2018), finding that microbial diversity and functional capacity increases as warming temperature facilitate thawing soil (Hultman et al., 2015). In comparison, the knowledge of microbial communities in Arctic lake sediments is severely lacking. Yet, sedimentary ecosystems are highly interconnected networks serving as integrators of hydrological, atmospheric, and terrestrial processes (Williamson et al., 2008, 2009; Emmerton et al., 2016).

1.4 Study Area: Lake Hazen Watershed

The Lake Hazen watershed is located on Northern Ellesmere Island (Nunavut, Canada) and is the largest Arctic Lake by volume (Köck et al., 2012), illustrated in Figure 1.2. Most of the High Arctic is a polar desert area with barren land surfaces, intense and persistent cold, and low amounts of precipitation (Surdu et al., 2016). The Lake Hazen watershed is a thermal Arctic oasis, characterised by atypical warmer microclimate than the surrounding polar desert (Surdu et al., 2016; Keatley et al., 2007). Inputs to the highly oligotrophic Lake Hazen are primarily limited to stream runoff

from glaciers during the late summer melt (June - August) (Surdu et al., 2016; Lehnherr et al., 2018). The hydrological regimes have been well described and stream runoff is known to enter from the glacial range on the NW half of the catchment and drain through Ruggles River, the primary outflow, to the SE coast of Ellesmere Island to the Arctic Ocean (Lehnherr et al., 2018; St. Pierre et al., 2019). From 2007-2012 increased glacial runoff raised water levels by 0.8 m and increased the annual outflow discharge by 370% (0.49 to 1.8 km^3), relative to 1996-2006 levels (Lehnherr et al., 2018).

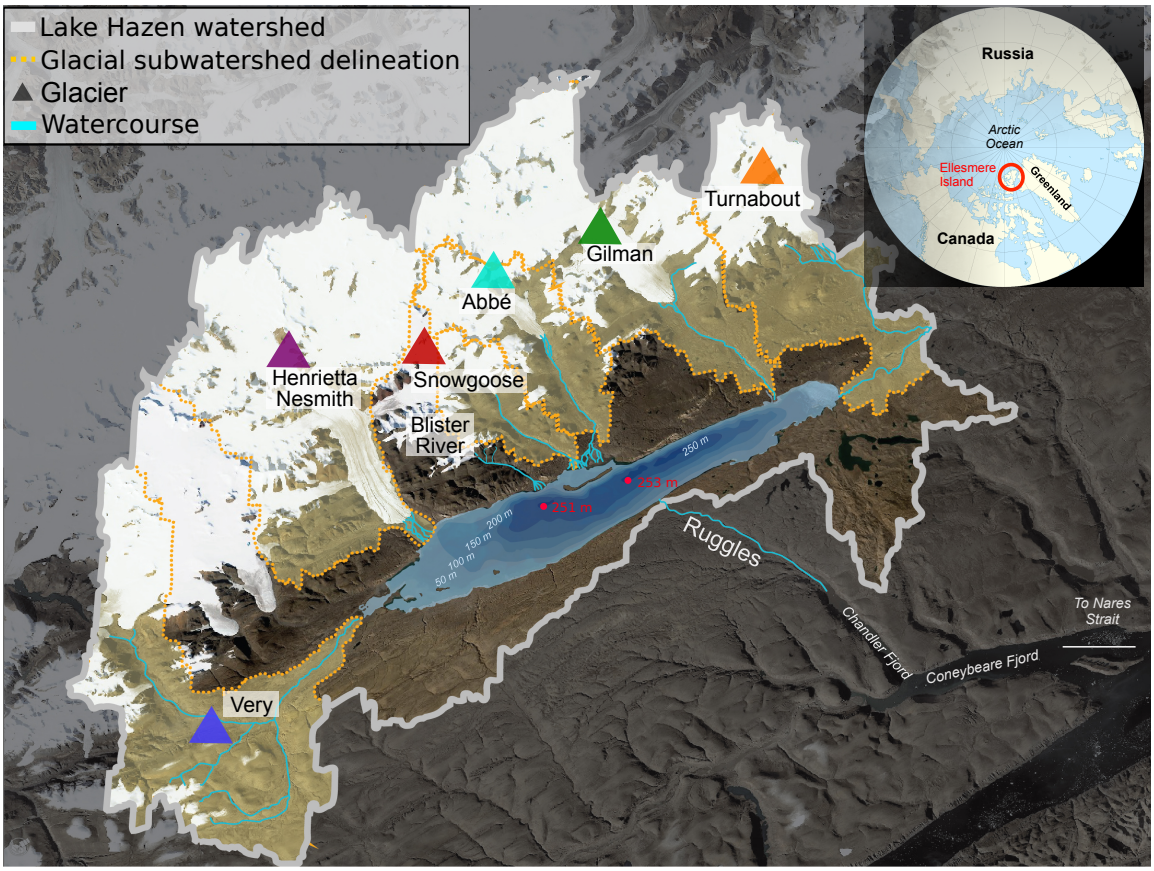

Figure 1.2: Map of Lake Hazen, Nunavut. High and low runoff regimes and deepest points are identified. Estimated runoff can be found in Supplementary Table A.1. Adapted from St. Pierre et al. (2019).

As a result of warming temperatures, these hydrological changes increased sedi-

ment and nutrient delivery to the lake (Williamson et al., 2009; Lehnherr et al., 2018), driving ecosystem shifts (Michelutti et al., 2003; Antoniades et al., 2007; Rühland et al., 2015); yet, the microbial responses to these changes have not been documented in this system. Although incredibly remote, it is a well stuided lake from both a hydrological (Smith, 1999; Surdu et al., 2016; St. Pierre et al., 2019) and biogeochemical perspective (Keatley et al., 2007; Emmerton et al., 2016; St. Pierre et al., 2018). The microbial work that has been completed in this environment is limited to a single 16S rRNA study (Ruuskanen et al., 2018), that described changes in community structure along pH and redox gradients within single sediment cores with limited spatial distribution. The subsequent work in Chapter 2 uses metagenomics to assess how the structure and metabolic potential changes through runoff regimes.

1.5 Objectives of Thesis

Within this thesis, I used spatially distributed sampling to capture changes in geochemical gradients that are a result of differential runoff caused by warming Arctic temperatures. Chapter 2 investigates the changes to the microbial community structure and metabolic capabilities as a result of these environmental changes observed in the Lake Hazen watershed. Chapter 3 summarises the contributions and limitations of this research, and directions for future work.

Chapter 2

Microbial structure and function responds to climate change in largest Arctic lake

Graham A. Colby[1], Matti O. Ruuskanen[1], Kyra A. St. Pierre[2], Vincent L. St. Louis[2], Alexandre J. Poulain[1], & Stéphane Aris-Brosou[1,3]

[1] University of Ottawa, Department of Biology, Ottawa, ON, Canada

[2] University of Alberta, Department of Biology, Edmonton, AB, Canada

[3] University of Ottawa, Department of Mathematics and Statistics, Ottawa, ON, Canada

Contributions

G.C. completed sampling, and laboratory analyses. G.C. and S.A.B. performed data analyses. G.C., A.P., and S.A.B. designed the study and wrote the manuscript. V.S.L. generated the chemical data and contributed to field design. G.C., M.R., K.S.P., V.S.L., A.P., and S.A.B. reviewed the manuscript.

2.1 Abstract

Climate change is occurring more rapidly in the Arctic than anywhere else on Earth. Arctic temperature and precipitation is expected to increase dramatically over the next century, yet little is known about how Arctic microbial communities in Arctic lakes will respond to these extensive changes. To address this question, we focused on Lake Hazen, NU in Canada's Arctic. As the largest High Arctic lake, by volume, it is a unique site to investigate microbial responses to environmental change. Over the past decade, glacial coverage of the lake has declined. Increasing runoff and sedimentation rates have result in a differential influx of nutrients through spatial gradients. Here, we used these spatial gradients to study how on-going environmental change may affect microbial community structure and function in Arctic lakes. We performed a metagenomic analysis of sites sampled from hydrological regimes representing high, low, and negligible influence of glacial runoff and compared the observed structure and function to the natural geochemical gradients. Both genes and reconstructed genomes were found in different abundances across these sites, suggesting that high runoff regimes alter geochemical gradients, homogenise the microbial structure and reduce genetic diversity. This work shows how a metagenomic approach can be used to predict future microbial responses to a changing climate.

2.2 Introduction

Climate change has dramatic impacts on the physical (Laudon et al., 2017; Bliss et al., 2014; O'Reilly et al., 2015), biogeochemical (Frey and McClelland, 2009; Lehnherr et al., 2018), and ecological (Smol et al., 2005; Wrona et al., 2016) processes in Arctic environments. In particular, these impacts are amplified in polar regions, where near-surface warming has been almost twice that of the global average in the last decade (Overpeck et al., 1997; Serreze and Francis, 2006; Screen and Simmonds, 2010). In fact, climate models predict temperature and precipitation increasing in the Arctic by as much as 8°C and 40%, respectively, by the year 2100 (IPCC, 2013). While recent work has investigated plant and animal adaptation and loss due to climatic changes at mid-latitudes (Pacifici et al., 2015; Kerr et al., 2015;

Pecl et al., 2017), little is known about how complex microbial communities will respond in Arctic environments. As microbial communities mediate most of the biogeochemical cycling in aquatic and terrestrial environments (Falkowski et al., 2008; Fuhrman, 2009), there is a growing need to understand how these ecosystems and the ecological services provided by microbes will change.

To address this outstanding question, we focused on the sediment and soil of the Lake Hazen watershed. Located in Quttinirpaaq National Park on northern Ellesmere Island, Nunavut, it is the world's largest High Arctic lake by volume (Köck et al., 2012). The Lake Hazen watershed is already experiencing the effects of climate change with increasing temperatures leading to more glacial melt, stream runoff, permafrost thaw, and less seasonal ice coverage (Lehnherr et al., 2018). Arctic lakes are typically low-nutrient ecosystems; yet, increasing runoff due to melt and thaw processes has the potential to introduce important quantities of sediment, nutrients, and organic carbon to the aquatic environment (Williamson et al., 2009). The consequences of these changes on the microbial sediment communities and their functions are currently unknown. In marine sedimentary ecosystems some functional groups, such as nitrogen fixers and denitrifiers, may actually benefit from changes in pH, temperature, and nutrient availability (Hutchins and Fu, 2017). In terrestrial soils, microbial communities have predictable and consistent responses to changes in temperature (Yergeau et al., 2012; Oliverio et al., 2017) and precipitation (Castro et al., 2010), which selects for distinct microbial communities. The outstanding question is how climate caused shifts in microbial assemblages may ultimately change functional capacity and alter biogeochemical processes. Physical changes like thawing soil, in permafrost systems are known to influence the structure and function of microbial communities and in turn, propagate changes to the geochemical environment (McCalley et al., 2014; Hultman et al., 2015; Mackelprang et al., 2016; Woodcroft et al., 2018). However, microbial responses to environmental changes in polar lake sediments are still poorly characterised. Sediments contain abundant and phylogenetically diverse microbial communities (Lozupone and Knight, 2007; Eisenhauer et al., 2012), which may be well positioned to withstand fluctuating environmental conditions, but this, to our knowledge, has not been tested *in situ*.

To date, most work performed on microbial communities in Arctic environments has been limited to descriptive studies using 16S rRNA gene sequencing (Stoeva et al., 2014; Thaler et al., 2017; Mohit et al., 2017; Ruuskanen et al., 2018). While providing rich insights into microbial community structures and their potential role in nutrient cycling, we are not aware of any prior work addressing the impact of climate change on microbial sediment communities in the Arctic. Lakes are relatively understudied in Arctic ecosystems, even though they are considered sentinels and regulators of climate change because they integrate physical, chemical, and biological changes happening throughout their watersheds (Williamson et al., 2009). Uniquely here, we make use of spatial runoff differences across the Lake Hazen watershed to mimic the temporal effects of climate change; differential runoff volume and sources lead to heterogeneous geochemical gradients that may affect microbial communities. Sediments and soils were sampled from three hydrological regimes representing high, low, and negligible influence of runoff. Based on this space-for-time substitution (Blois et al., 2013; Lester et al., 2014), we suggest that higher runoff regimes will be most similar to the future trajectory of aquatic inputs. For the first time in the sediments of a remote High Arctic lake, we apply shotgun metagenomics to evaluate the microbial community responses to changing environmental gradients, and show that climate change can lead to homogenisation of the microbial community structure and reduction in genetic diversity.

2.3 Methods

Sample Collection and Processing

Soil and sediment cores were collected from the Lake Hazen watershed (82°N, 71 W°: Figure 2.1A) during the start of the summer melt season, between the 10^{th} of May and the 10^{th} of June 2017. Sediment cores approximately 30 cm in depth were collected with an UWITEC (Mondsee, Austria) gravity corer from five locations: L1 (depth: 50 m), L2 (depth: 251 m), H1 (depth: 21 m), H2 (depth: 253 m), C (depth: 50 m). Three sediment cores per site were collected for genomic analysis approximately 3 m apart to account for site heterogeneity.

The top 5 cm of each core was sliced and placed into Whirlpack bags. These slices were homogenised by hand inside of the bags and stored in a −20°C freezer until shipment back to the University of Ottawa where they was stored at −80°C. An additional two sediment cores per site were collected for porewater, combustion, and microprobe analysis. Soil cores (L-Soil, H-Soil, and C-Soil) were collected from the landscape adjacent to the sediment core collection sites in triplicate approximately 1 m apart to account for site heterogeneity. After snow and surface vegetation was removed, cores approximately 10 cm in depth were recovered with a stainless steel hand coring device. For soil samples the entire depth of the core was place into a falcon tube, homogenised, and stored as above. Contamination of samples was minimised by sterilising the coring equipment between use with 10% bleach and 90% ethanol and by wearing non-powdered latex gloves during sample handling. As we were most interested in the entire community change through chemical gradients, we combined the top 5 cm of sediment and 10 cm of soil for sample extraction and subsequent sequencing.

For each sample, the triplicate cores were extracted and combined. Samples were thawed overnight and 250 mg–400 mg (wet weight) was then washed in a sterile salt buffer (10 mM EDTA, 50 mM Tris-HCl, 50 mM Na_2 HPO_4 $7H_2O$ at pH 8.0) to remove PCR inhibitors (Zhou et al., 1996; Poulain et al., 2015). All sample handling was conducted in a laminar flow hood (HEPA 100) stainless steel sterile cabinet that was treated with UVC radiation and bleach before each use. DNA extractions were performed in a laminar flow hood (HEPA 100), using the DNeasy PowerSoil Kit (MO BIO Laboratories Inc, Carlsbad, CA, USA), and following the kit guidelines, except that the final elution volume was 30 μL instead of 100 μL. The PowerSoil kit was used as it has been shown to perform constantly across different soil and sediment matrices (Mahmoudi et al., 2011; Vishnivetskaya et al., 2014; Ramírez et al., 2018). DNA integrity was validated with a NanoDrop Spectrometer and PCR combined with electrophoresis of the *glnA* gene (Figure A.4, Table A.6). Adequate DNA concentrations for sequencing were reached through combining triplicate extractions for a total volume of 45 μM and a concentration $\geq$ 50 ng/μL (Table A.7).

Chemical Analyses

Chemical features were measured for the sediment sites along with depth. Redox potential, pH, and dissolved O_2 concentration profiles were measured at 100 μM intervals in the field within an hour of collection, using Unisense (Aarhus, Denmark) microsensors connected to the Unisense Field Multimeter. Cores used for porewater chemistry analysis were sectioned in 1 cm intervals into 50 mL falcon tubes in the field, followed by flushing of any headspace with ultra-high-purity N_2 before capping. Sediment porewater was extracted following centrifugation at 4,000 rpm, after which the supernatant was filtered through 0.45 μm cellulose acetate filters into 15 mL tubes, which were then frozen until analysis. Concentrations of $NO_2^- + NO_3^-$, NH_3, Cl^-, Total Dissolved Phosphorus (TDP), and SO_4^{-2} were measured in the sediment porewater by ion chromatography. However, TDP was removed from data analysis because insufficient porewater was collected form TDP analysis at site C. All porewater analyses were performed at the Elemental Analysis and Stable Isotope Ratio Mass Spectrometry Laboratory (Department of Renewable Resources, University of Alberta). The centrifuge sediments were retained for percent $CaCO_3$ and percent Organic Carbon (OC) estimation through loss on ignition (Heiri et al., 2001). For soils collected temperature at sampling time was measured (Table A.3), but mass of samples were insufficient for further chemical analysis.

The chemical features were summarised into 1 cm intervals through the top 5 cm of the sediment cores. The geochemical properties of each sediment site were then visualised using a Principle Component Analysis (PCA) and clustered using Partitioning Around Medoids (PAM). The appropriate number of clusters were determined using the silcheck function from the hopach (van der Laan and Pollard, 2003) R package. Each feature was further visualised as a boxplot and a Dunn Test was used to perform multiple pairwise comparisons.

Sequencing and Data Processing

Metagenomic libraries were prepared and sequenced by Genome Quebec on an Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA) on a paired-end 125 bp configuration using Illumina TruSeq LT adapters (read 1: AGATCGGAAGAGCACACGTCT-

GAACTCCAGTCAC, and read 2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT). A total of 8 DNA samples (5 sediments, 3 soils) were sequenced generating over 150 Gb of data (see Figure A.5 for Illumina library validation). Read count summaries were tracked throughout each step of the pipeline for quality control (Table A.7). Low quality reads, sequencing adapters, unpaired reads, and low quality bases at the ends of reads were removed to generate quality controlled reads with Trimmomatic (v0.36) (Bolger et al., 2014) using the following paired-end commands: phred33, ILLUMINACLIP:TruSeq3-PE-2.fa:3:26:10, LEADING:3 TRAILING:3, SLIDINGWINDOW:4:20, MINLEN:36, CROP:120, HEADCROP:20, AVGQUAL:20. FASTQC (v0.11.8) (Andrews, 2010) was then used to confirm that the Illumina adapters were removed and that trimmed sequence length was 90 bp in length.

Metagenomics Assembly, Binning, and Annotation

To recover genomes, metagenomic quality controlled reads from all samples were coassembled using Megahit (Li et al., 2015) software with a k-mer size of 31 and "meta-large" setting (summary statistics in Table A.5). EukRep (West et al., 2018) was used to remove any eukaryotic DNA from the contigs prior to the formation of an Anvio (v5) (Eren et al., 2015) contig database. A contig database was generated removing contigs under 1000 bp and gene prediction was performed in the Anvio environment (Table A.5). Sequence coverage information was determined for each assembled scaffold by mapping reads from each sample to the assembled contig database using Bowtie2 (Langmead and Salzberg, 2012) with default parameters. The resulting SAM files were sorted and converted to a BAM files using samtools (v0.1.19) (Li et al., 2009). The BAM files were prepared for Anvio using the "anvi-init-bam" and contig database generated using "anvi-gen-contigs-database". The contig database and BAM mapping files were further used as input for "anvi-profile", which generated individual sample profiles for each contig over the minimum length of 2500 bp. These profiles were then combined using "anvi-merge" and summary statistics for abundance and coverage were generated with "anvi-summarize". Automated binning was performed using CONCOCT (Alneberg et al., 2014). Scaffolds were binned on the basis

of %GC content and differential coverage abundance patterns across all 8 DNA samples. Manual refinement was done using Anvio's refine option. Kaiju (Menzel et al., 2016) was used to classify taxonomy of the assembled contigs with "anvi-import-taxonomy-for-genes" and aided in the manual refinement process. Open reading frames and potential genes were identified with Prodigal (v2.6.3) (Hyatt et al., 2010). Anvio's custom HHMs were run, along with NCBIs cog annotation to identify genes. PFAM, TIGRFAM, GO terms, KEGG enzymes, KEGG pathways and Metacyc pathways were annotated with Interproscan (v5) (Jones et al., 2014). These gene annotations were then combined with the Anvio database with "anvi-import-functions".

Genome completeness and contamination was evaluated based on the presence of a core set of genes using CheckM (v1.0.5) "lineage_wf" (Parks et al., 2015). Only genomes that were at least 50% complete and with less than 10% contamination were further analysed (meeting the MIMAGs standard of medium or high quality genomes (Yilmaz et al., 2011)). A total of 300 high-quality genomes were recovered from this binning process. All of these genomes were used to calculate an average amino acid identity across all genomes using compareM (v0.0.23, function "aai_wf") (Parks, 2014). CheckM was used again to identify contigs that were not contained in any of the 300 high-quality genomes. These unbinned contigs were binned using an alternative binning algorithm MaxBin (v2.0) (Wu et al., 2015). An additional 481 genomes were recovered, but were not included in further analysis as only a 21 genomes were of average completion $> 65\%$ (Supplemental Data File 1: `https://github.com/colbyga/hazen_metagenome_publication/blob/master/maxbin2_unbinned_contigs_summary.csv`).

Phylogenetic Analysis Using Ribosomal and Core Proteins

Phylogenetic analysis was performed using two different sets of markers for bacteria and archaea. The list of 120 bacterial and 122 archaeal marker genes have been previously used to standardise bacterial taxonomy (Parks et al., 2018) (full list of proteins available on github). The marker genes were extracted from each genome by matching the Pfam (v31) (Finn et al., 2015) and TIGRFAMs (v15.0) (Haft et al., 2003) annotations from GTDB

(v86) (Parks et al., 2018) to those found in the genomes. Marker genes from each of the 300 genomes were translated using seqinr (Charif and Lobry, 2007). As some genomes had multiple copies of a marker gene, duplicate copies were filtered keeping the most complete sequence with a custom R script. Additionally, if a genome was missing marker genes, the genome was still included, but as an empty sequence. The markers from each bin were combined with the individual GTDB trimmed multiple sequence alignments. MUSCLE (v3.8.31) (Edgar, 2004) was used to align the genomes to the GTDB reference sequences in R (v 3.5.1) (R Development Core Team, 2008). Archaeal sequences were removed from the bacterial alignment on the basis of results from CheckM (Parks et al., 2015) and independently verified using a custom list of archaea specific marker genes. Alignments were then refined using trimAI (Capella-Gutiérrez et al., 2009) and the "-gappyout" parameter. FastTree2 (Price et al., 2010) was used to infer approximately-maximum-likelihood phylogenetic trees from protein sequence alignments under the WAG model (Whelan and Goldman, 2001). The archaeal tree was rooted on Euryarchaeota and bacterial tree on Patescibacteria using APE (Paradis et al., 2004). Trees were visualised and coloured according to phylum using ggtree (Yu et al., 2017).

Relative Abundance and Community Composition Analysis

To determine the relative abundance of each genome in the eight samples, sample-specific genome abundances were normalised by sequencing depth [(reads mapped to a genome)/(total number of reads mapped)], making comparison across samples possible. Genome abundances were generated using the CheckM "profile" function (Parks et al., 2015). All further analyses were performed in RStudio (RStudio Team, 2015) using R (v 3.5.1) (R Development Core Team, 2008). To determine the average abundance of major taxonomic groups across (determined by the phylogenetic analysis described above) the abundances for genomes from the same taxonomic were summed and visualised using phyloseq (McMurdie and Holmes, 2013). These same abundance values were the basis for a community composition analysis. *t*-SNE plots were constructed by assigning each genome to a site based on where it was most abundant using Rtsne (Krijthe et al., 2018). Heatmaps were generated in gplots (Warnes

et al., 2009). A tanglegram connecting the genomes between cluster by abundance and phylogeny were generated in dendextend (Galili, 2015).

Metabolic Potential Analysis

As described previously gene prediction for individual genomes was performed using Prodigal (v2.6.3) (Hyatt et al., 2010) and annotated with Interproscan (v5) (Jones et al., 2014). To analyse functional marker genes in the metagenomes, we used a custom database of reference proteins sequences (KEGG (Kanehisa et al., 2015), TIGRFAM (Haft et al., 2003), COG (Tatusov et al., 2003), PFAM (Finn et al., 2015)) based of off the marker genes previously used in other studies (Anantharaman et al., 2016; Dombrowski et al., 2018) (Supplemental Data Files on GitHub). Pathways were also predicted using MinPath (Ye and Doak, 2009) to map all identified KEGG enzymes to the most parsimonious MetaCyc pathways (Caspi et al., 2007). MinPath presents only pathways represented by multiple genes. As most genomes were present even at low abundances across all sites, a cutoff value of $-\log_{10} \leq 0.25$ was set for a genome to be considered functionally relevant at any site. The high abundance genomes for each site were used as the basis of further examination. We aggregate marker genes and pathways by function and phylogenetic groups. We further grouped all taxa together at each site to tested for significant differences in major cycling processes (carbon, nitrogen, and sulphur) between sites using a hierarchical clustering with P-values (Suzuki and Shimodaira, 2006) and Fisher's exact test. Antibiotic resistance genes (AMR) were identified from each of the high quality genomes using the Comprehensive Antibiotic Resistance Database (CARD 2.0.3) (Jia et al., 2016). AMR genes were then plotted as a heatmap using CARD's Resistance Gene Identifier (RGI 4.2.0) (McArthur, 2018) and visualised in R.

Availability of Data

All data will be made accessible prior to manuscript submission. Scripts and supplemental data files can be accessed from `https://github.com/colbyga/hazen_metagenome_`

publication. Raw sequence reads of the shotgun metagenomic data will be submitted to the NCBI sequence read archive under Bioproject. Metagenome assembled genomes will be submitted to NCBI. The geochemical data will also be made available at the National Centers for Environmental Information online repository.

2.4 Results

Heterogeneous Runoffs Lead to Distinct Geochemical Gradients

We first characterised how geochemical properties of the sediments vary across a spatial gradient (Figure 2.1A). Samples were collected along two transects and can be separated into three hydrological regimes by seasonal runoff volume: high (H sites), low (L sites), and negligible runoff (C sites; Table A.1). Cores were taken at eight sampling locations along these transects. For sediment core sites, detailed chemical measurements were obtained through the top 5 cm of the cores and averaged over 1 cm intervals. Within the watershed, runoff flows from the outlet glaciers along the northwestern shoreline through poorly consolidated valleys, depositing sediments into Lake Hazen (H and L sites). The lake then drains into Ruggles River along its southeastern shoreline (C sites). The surrounding glaciers deliver different amounts of sediments, likely with different chemical compositions, and do so unevenly throughout the lake (Table A.2). The sediments then heterogeneously settle at the deepest parts of the lake with the top 5 cm in each sediment location not necessarily representing the same depositional dates. In the high runoff H2 site, the top 5 cm of sediments were deposited between 2011-2017; in contrast in the low runoff L2 site, the top 5 cm were deposited between 1987-2017 (Table A.2).

As it is the only substantial input to the lake ecosystem, glacial runoff transporting sediments is likely responsible for the clustering of sites into four distinct geochemical groups (Figure 2.1B). Higher concentrations of ammonia, sulphate and a greater percentage of calcium carbonate were present in the high runoff sites. However, higher concentrations

of oxygen, nitrate/nitrates, and greater redox potential were present in the low runoff and control sites. In the principal component analysis, PC1 explained 43% of the total variance, and differentiated the high and low runoff sites. Organic carbon, pH, and chlorine were more determinant when distinguishing between shallow and deep sites within a hydrological regime (Figures A.1, A.2).

The observed chemical changes between different sediment sites were more significant than the changes through depth of a single core (Figure 2.1BC). Rather than grouping spatially with the high runoff sites, the control site (C) was most chemically similar to the L1 site. Site C's chemical features showed little or no significant difference from L1 (Figures 2.1C, A.3). In measurements of pH and organic carbon, the shallow sites (L1 and H1) were not significantly different from each other, yet were both significantly different from the deeper sites (L2 and H2) suggesting that although most chemical features were similar within a hydrological regime, some were still influenced by their spatial proximity to the shoreline (Figure 2.1C).

Chemical Gradients, Not Phylogeography, Correlate with Site-Specific Phylogenetic Diversity

To assess the impacts of these geochemical constraints on biodiversity, we reconstructed a total of 300 (290 bacteria and 10 archaea) genomes (metagenome-assembled genomes) that were >50% complete and with <10% contamination from $1,455,655$ contigs (Tables A.4, A.5). Their phylogenetic placement was obtained by constructing two phylogenetic trees: one for bacteria and one for archaea based on ∼120 ubiquitous single copy proteins (Figure 2.2). While most major phyla were represented in the reconstructed genomes, no Firmicutes and only a small number of archaea were identified. In constrast, Gammaproteobacteria ($n = 50$), Actinobacteria ($n = 31$), Alphaprobacteria ($n = 24$), Chloroflexoata ($n = 30$), Planctomycetota ($n = 24$), and Acidobacteriota ($n = 19$) were the most commonly reconstructed taxa across the entire watershed. Uncultured bacteria comprised ∼ 11% of genomes identified in Lake Hazen, with representative genomes from multiple taxa: Eisenbacteria ($n = 12$), Patescibacteria ($n = 9$), Omnitrophica ($n = 5$), KSB1($n = 1$), Armatimonadota

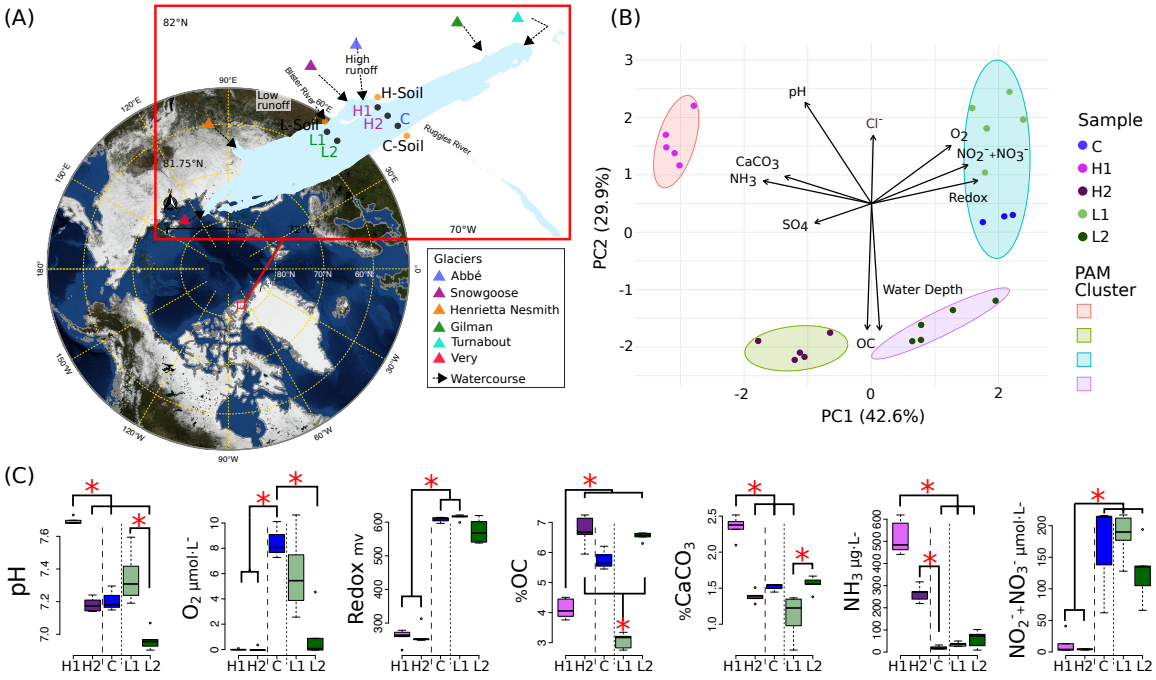

Figure 2.1: Lake Hazen sampling design and chemical composition. (A) Basemap: The location of Lake Hazen (red box). Inset map: soil (orange dots) and sediment (black dots) sample sites are separated into hydrological regimes of high (purple), low (green), and negligible/control (blue) runoff. Global map data from NASA Earth Observatory (Stöckli et al., 2005) and inset map data from NRCAN (NRCAN, 2018). (B) Principal component analysis (PCA) showing the differences in physical and chemical composition of the sediment sites. Vectors display the following environmental features: pH, dissolved oxygen (O_2), redox potential, nitrates and nitrites concentration ($NO2_2^- + NO3_2^-$), water depth, percent organic carbon (OC), percent calcium carbonate ($CaCO_3$), sulphate concentration (SO_4), and ammonia concentration (NH_3). Individual points represent the mean values using 1 cm intervals measured in the top 5 cm. Partitioning around medoids was used to identify clusters. (C) Distribution of chemical features for sediment sites. Branches and asterisks indicate significant differences between sites $P < 0.025$ (Dunn test). If branch tips form a dichotomy or trichotomy, the interactions within that group is not significant. Long dashes separate high runoff sites and dotted line separates low runoff sites. There was insufficient data to include soil sites in B and C.

($n = 1$), Lindowbacteria ($n = 1$), USBP1 ($n = 1$), UBP10 ($n = 1$), Zixibacteria ($n = 1$).

As these reconstructed genomes formed dense clusters, often branching closely to one another on the phylogenetic tree, we verified their uniqueness by comparing the average amino acid identity (AAI) of each genome. Most genomes had an AAI below 80% and only 29 were between 80-87%, which indicated that the genomes were indeed distinct, despite often being closely related. Few genomes branched basally (deep in the tree), providing little evidence for the identification of novel phyla: only six genomes (2%) were identified with branch length > 0.6, with all belonging to Chloroflexa, Plactomycetota or Lindowbacteria phyla. These long-branching lineages were placed within a larger clade, providing further confidence in the taxonomic assignment, except for Lindowbacteria, which only has two representative sequences in the GTDB (GCA_001782795.1, GCA_001784175.1).

In spite of this broad phylogenetic distribution, the reconstructed genomes were not represented evenly across all sites. When summarising the reads mapped to each phylum, we observed that the soil community structure was dominated by a few phyla, with Actinobacteria, Gammaproteobacteria, and Crenarchaeota comprising $\sim$ 75% of all mapped reads (Figures 2.2 inset, A.14). Unlike in soil, the sediment community structure was more diverse (Figure 2.2 inset). The reconstructed genomes tended to be more representative of the organisms found in sediment, on average 50.62% of reads from sediment samples mapped to reconstructed genomes whereas $\sim$ 80% of reads from the soil sites could not be mapped to any of the 300 reconstructed genomes (Figure A.14).

To understand the drivers of this site-specific phylogenetic diversity, we bi-clustered genomes by their normalised abundances (on a $-\log_{10}$ scale to reduce skew), and by sample site (Figure 2.3). This led to clustering of the sites based on geochemical features observed previously (Figure 2.1B). For both bacteria and archaea, the high runoff sites (H1 and H2) group separately from low runoff sites, and the soil sites were most similar to each other (Figures 2.3, A.12). For archaea, high runoff sites formed a clade with the soil sites, suggesting that soil sites shared more structural similarity in terms of abundant organisms within the high runoff regimes (Figure 2.3B). While site clustering followed the geochemical gradients identified above, the normalised abundances of these genomes showed no strong

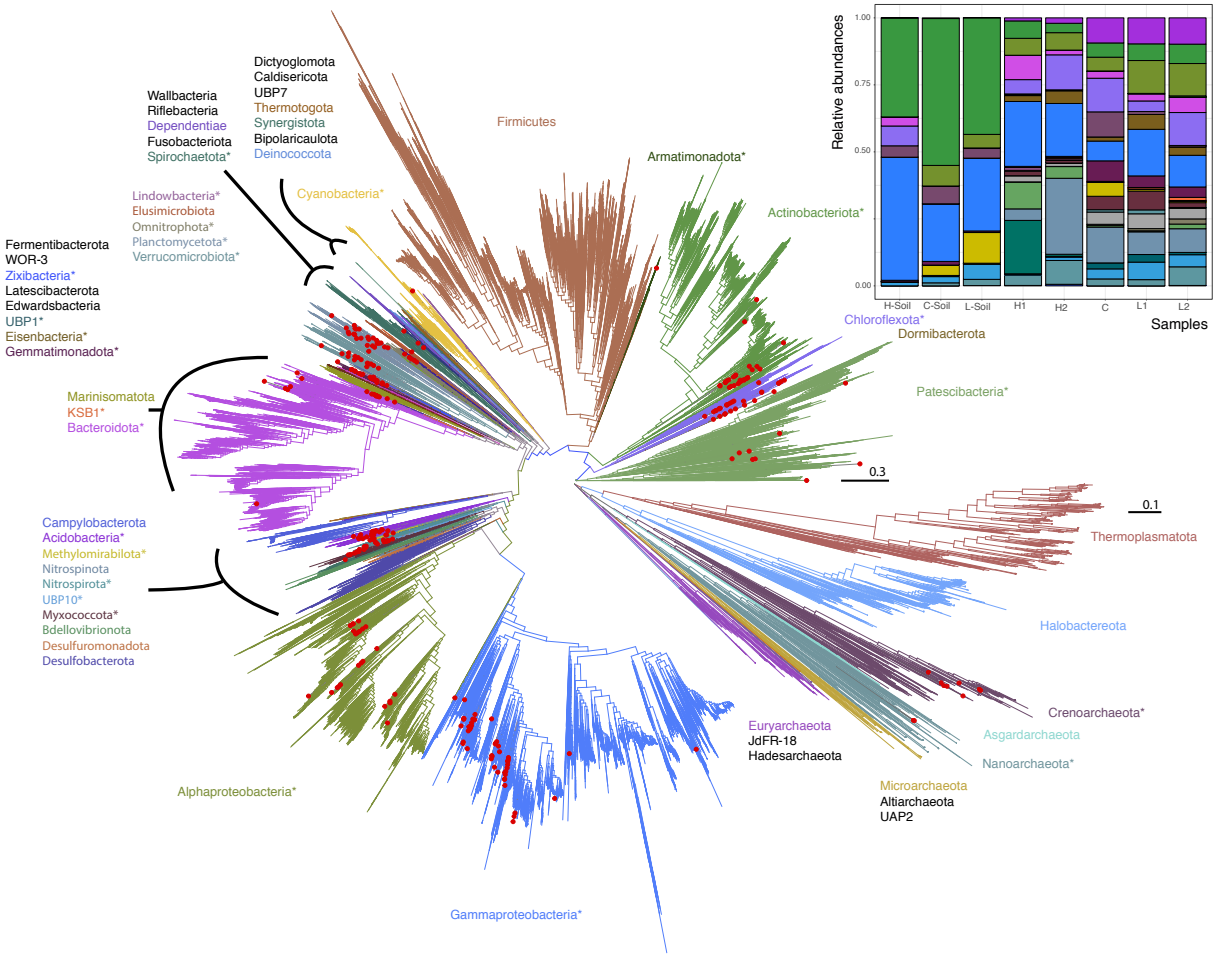

Figure 2.2: Maximum-likelihood phylogenetic trees of Lake Hazen genomes based on 120 concatenated bacteria and 122 concatenated archaea protein-coding genes. Red Dots: Lake Hazen genomes. Asterisks (*) indicate phyla that contain Lake Hazen genomes. Bacteria tree is rooted on Patescibacteria and Archaea tree is rooted on Euryarchaeota, JdFR-18, and Hadesarchaeota. See GitHub account for full taxonomy tree files and for original tree files (Supplemental Data File 2 and 3). Inset shows reconstructed genome abundance across sites, in the 300 high quality genomes for each sample normalised to 100%.

phylogenetic pattern (Figures 2.3, A.9). In spite of this absence of phylogeographic pattern, the tanglegram suggests that the β-diversity of high abundance genomes in the low runoff and control sites was greater than the high runoff sites, which was highlighted by the distribution of green lines connecting the phylogenetic tree to the clustered tree (Figure 2.3).

Structural Redundancy and Interconnectivity Between Aquatic and Terrestrial Environments

To further analyse the relationship between microbial communities and geochemical gradients, we determined the site where each genome was most abundant. Based on this information, an unsupervised clustering with t-SNE showed that the most abundant reconstructed genomes recapitulated the sites and their geochemical gradients (Figure 2.4A). As water flows from sites 1 to 2 in the L and H transects (Figure 2.4A, shaded arrows), the shift from L to H runoff was almost orthogonal in the t-SNE projection space, suggesting that a change in runoff regime leads to a dramatic reshuffling of microbial communities, or at least of their most abundant reconstructed phyla. The most affected phyla Planctomycetota, Zixibacteria, Cyanobacteria, Lindowbacteria, Patescibacteria, and Spirochaetota all increase in abundance from low to high runoff regime (Figure 2.4B).

Overlaying sediment chemistry to genome abundance in an NMDS ordination further showed that phyletic diversity within sites is partially controlled by chemistry (Figure 2.5). In the sediment sites, ammonia concentrations ($P = 0.03$), nitrate and nitrite concentrations ($P = 0.03$), and redox potential ($P = 0.03$) were significant in determining the distribution of reconstructed genomes (permutation test $P < 0.05$). In the sediment, we observed that the sites with the greatest biodiversity (L1, L2, and C) were also those with the greatest redox potential, oxygen concentrations, and nitrate/nitrite concentrations. In the low diversity sediment sites (H1 and H2), ammonia and sulphate concentrations were greater, with lower redox potential and access to substrates. The NMDS ordination (Figure 2.5) showed a similar grouping of sites as those indicated by the geochemical variables (Figure 2.1B) and was further supported by DPCoA and PCoA analysis (Figures A.11, A.13).

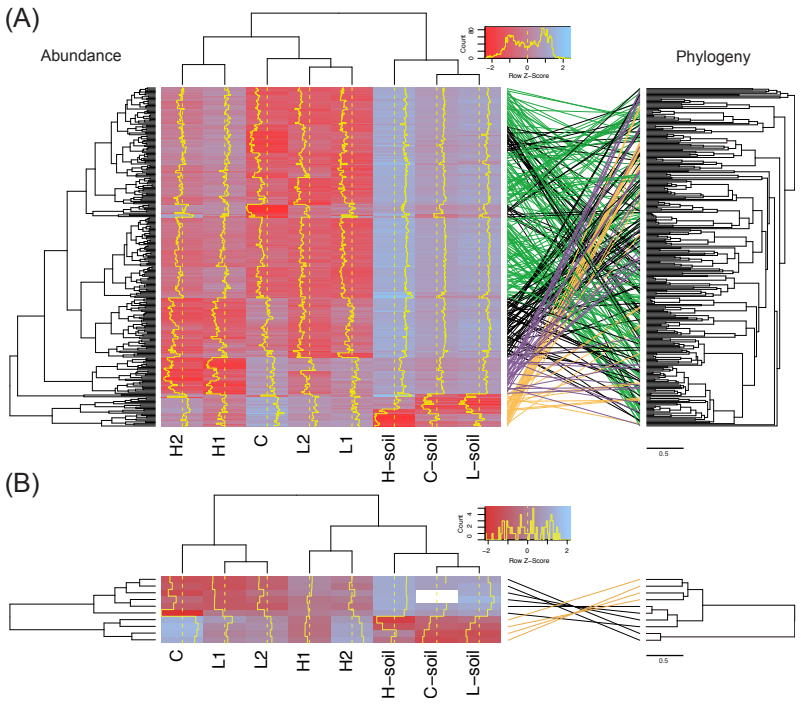

Figure 2.3: Genome abundance heatmap, and tanglegram separated by (A) bacterial genomes and (B) archaeal genomes. Left: The heatmap displays genome abundance per site normalised by amount of reads in each sample and transformed to a $-\log_{10}$ scale. Dotted yellow lines represent mean abundance values, and yellow traces represent the raw z-scores above (red) and below (blue) the mean. The abundance values are grouped both by sites (top dendrogram) and genomes (left dendrogram). Right: Tanglegram between dendrogram clustered by similar abundances and phylogenetic tree. Highlights in tanglegram: orange lines are genomes abundant in soil, green lines are genomes abundant in low runoff sediment, purple lines are genomes abundant in high runoff sediment, black lines are genomes shared in multiple environments.

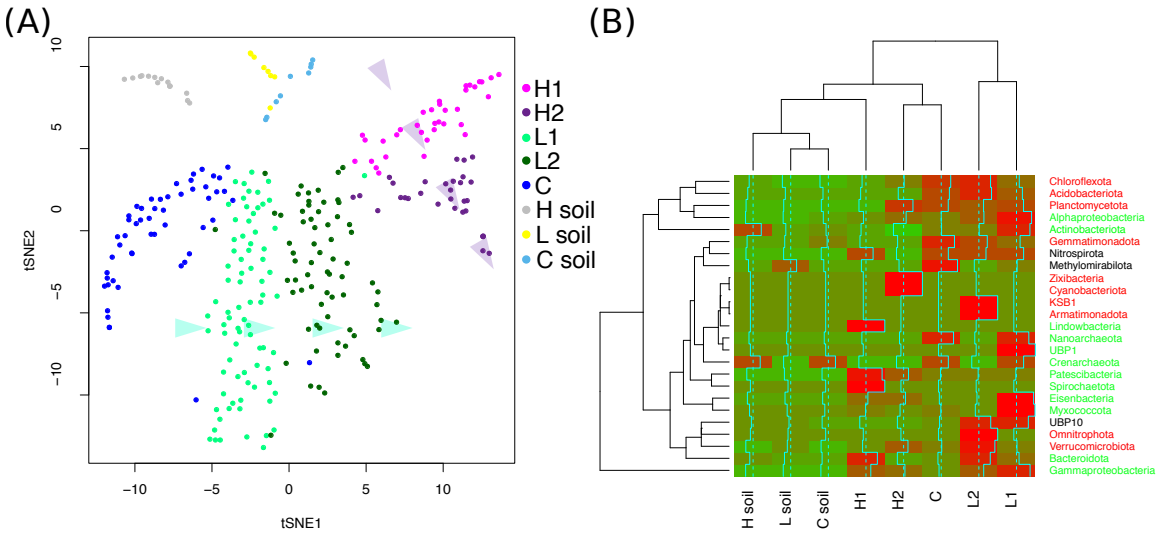

Figure 2.4: Reconstructed genome abundances across sites. (A) *t*-SNE analysis of genome abundance for each sediment sample. All 300 genomes are shown and each genome is assigned to sample where its has the greatest abundance. Shaded arrows display direction of water flow between the high runoff (green) and low runoff (purple) sites. (B) Heatmap of genome presence grouped by phylum. Red bars show high abundance phyla and green bars show low abundance phyla. Names in red increase in abundance from shallow to deep sites (from H1 to H2 or L1 to L2), while those in green decrease in abundance from shallow to deep sites (from H1 to H2 or L1 to L2).

Metabolic Functions Differ Between Runoff Regimes

Given that these genomes yielded such a large number of microbial lineages, we inferred potential physiological capabilities by assigning metabolic functions to proteins in each individual genome. Detailed genome-specific metabolic potential was determined by profiling all the genes against specific databases: KEGG (Kanehisa et al., 2015), TIGRFAM (Haft et al., 2003), COG (Tatusov et al., 2003), PFAM (Finn et al., 2015). We further used KEGG enzyme commission numbers to identify the fewest number of Metacyc (Caspi et al., 2007) pathways expressed in each genome. We specifically targeted genes and pathways involved in microbial metabolism with specific ecosystem functions such as carbon, nitrogen, and sulphur cycling. Using a stringent abundance cutoff value ($-\log_{10} \leq 0.25$), only the most abundant genomes per site were reported, and condensed by phylum (Figures 2.6, A.15,

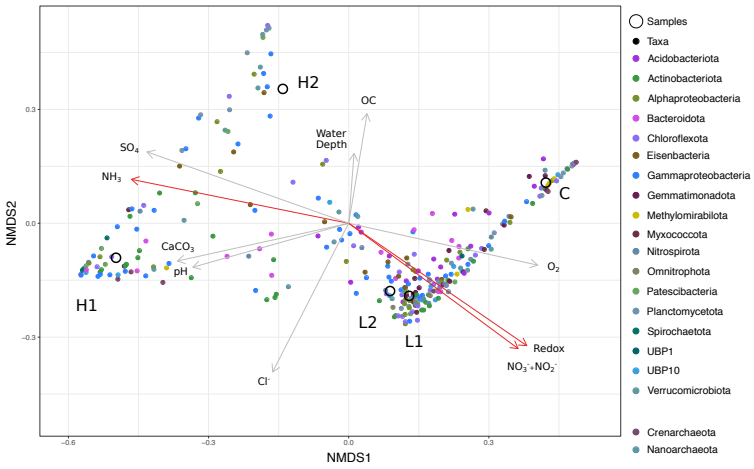

Figure 2.5: NMDS analysis of genomes from 20 most abundant phyla in sediment sites. Physical and chemical vectors are fitted to data. Red vectors are significant (permutation test: $P < 0.05$).

A.8). Considering the carbon, nitrogen, and sulphur metabolism capacity across sites, we identified a groups that conformed to both the geochemical and microbial community structures (Figure A.17), where sites group independently, except for the control site that was not significantly different from the low runoff sites. P-values from Fisher's exact test on the metabolic differences between regimes (Table A.8), indicated that the carbon, nitrogen, and sulphur metabolism in the high runoff sites was significantly different from all other sites, except for the comparison of nitrogen metabolism between high and low runoff.

Common carbon metabolisms were shared across all runoff regimes. We observed carbon oxidation and reduction reactions regulated through Wood-Ljungdahl pathway in the high runoff regime. Here, the Spirochaetota (acetogens) were likely responsible for anaerobic respiration and carbon fixation producing acetate (CH_3COO^-) as an end product. This metabolic potential was only present in the high runoff locations where sedimentary conditions were anoxic throughout the first 5 cm (Figures 2.6, A.1). However, carbon fixation through the Calvin-Benson-Bassham (CBB) pathway was present across all locations. We observed a greater abundance of reconstructed genomes with the potential to perform carbon monoxide oxidation reactions in the low and control sediments. Carbon monoxide oxidation was detected by the presence of *cox* genes and likely performed by chemolithoau-

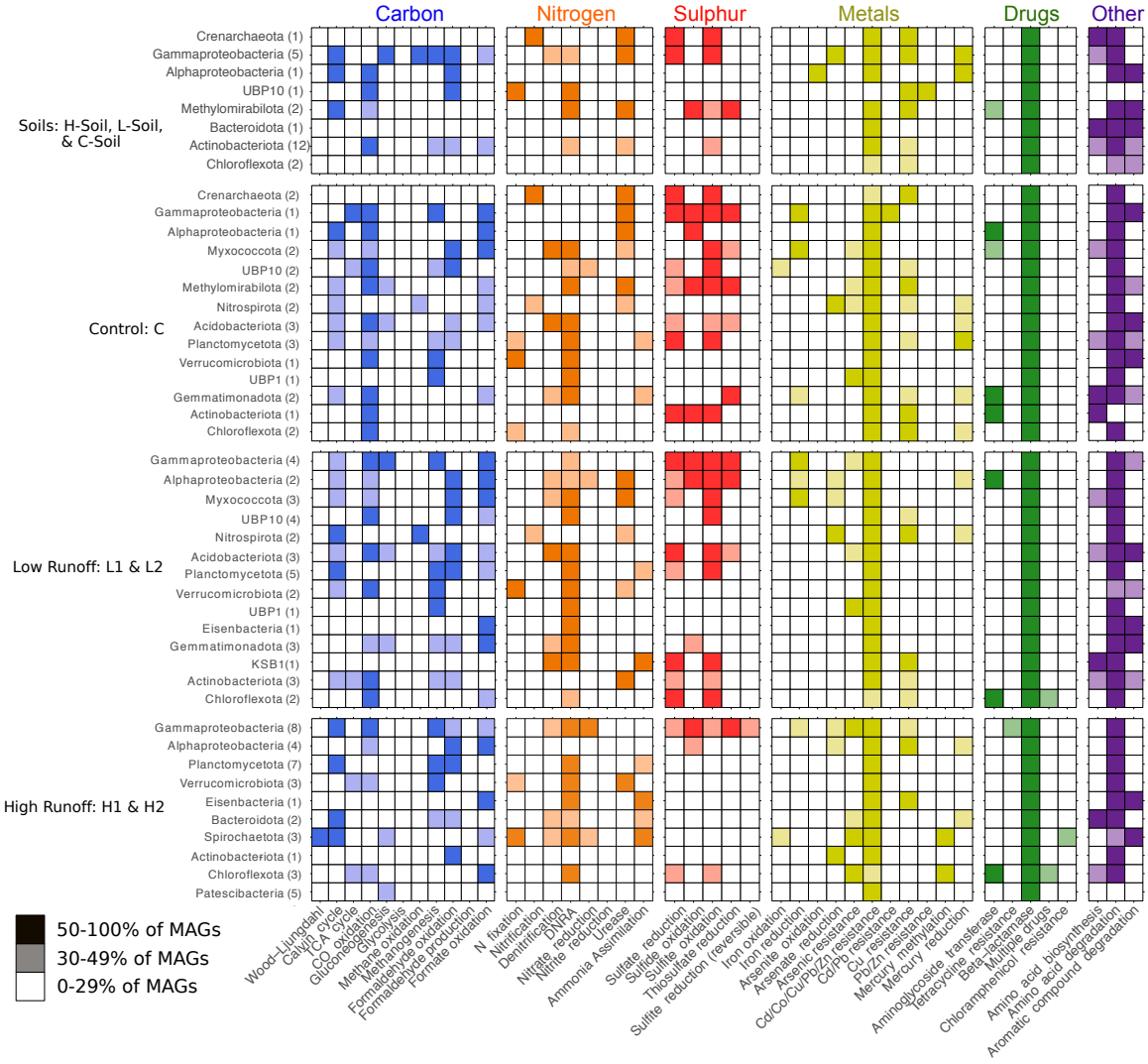

Figure 2.6: Metabolic capacity of genomes separated by hydrological regime. Genomes are only considered to contribute to a site if 0.56% ($-\log_{10} \leq 0.25$) of reads per sample mapped to a genome. Presence of core metabolic genes involved in carbon metabolism, nitrogen metabolism, sulphur metabolism, metal cycling, antibiotic resistance, and other metabolism are shown. Number of genomes for each taxa are shown in parentheses. No colours: genes predicting function are absent or in low abundance. Faded colours: genes predicting function are present in 30-50% of genomes per phylogenetic group. Dark colours: genes predicting function are present in >50% of genomes per phylogenetic group. If a column is blank it indicates no markers were present in the high abundance genomes that are based on the 0.56% cutoff value (See Figure A.8)

totrophs. We found that across all sites, microbes have the capacity for simple carbon metabolism. Simple carbon metabolism is an important metabolic process for freshwater microbes that were likely exposed to a greater variety of low-complexity carbon sources, such as carbohydrates, carboxylic acids, and C1 compounds. Methanogenesis pathways were present across all sites, but methane oxidation pathways were absent from high runoff sites, where oxygen is limited. Notably, two pathways, glycolysis and formaldehyde production, were absent from all microorganisms across all sites. The marker genes that are indicative of these pathways were absent in the high abundance genomes at each site (Figure A.8), which might be a result of true absence from the genomes, incomplete annotations, or incomplete reconstructed genomes.

Markers for nitrogen metabolism were found in different phylogenetic affiliations of the reconstructed genomes. Across all runoff regimes dissimilarity nitrate reduction (DNRA) was present in multiple genomes (Figure 2.6). In the high runoff regime there was both an absence of nitrification and a greater presence of markers for ammonia assimilation. Greater concentrations of ammonia, in the high runoff regimes, also suggests that N-containing organic matter was mineralised through ammonification (Figure 2.5). In contrast, urease markers were found more abundantly in low runoff regimes, where ammonia concentration was lower (Figure A.2).

The functional ability of microbes to cycle sulphur between oxidised and reduced forms was the most significant difference between the high and low runoff regimes (Table A.8). In the high runoff regime, Gammaproteobactiera were the only organisms with the metabolic capacity to expansively utilise sulphur, performing sulphide oxidation and thiosulphate reduction. Sulphate reduction was predominantly found in the soil, control, and low runoff regimes.

Aside from nutrient cycling, we also assessed the capacity of microbial communities to process metals and antibiotics. Metal resistance and cycling was mostly ubiquitous throughout all of the sites, regardless of runoff. Methyl mercury production, identified by the presence of both *hgcA* and *hgcB* genes (Parks et al., 2013), was only implicated in the high runoff sites, in Spirocheatoa and Chloroflexoata. However, genes conferring mercury

resistance involved in the conversion of inorganic Hg^{II} to the less toxic Hg^0 – were evenly distributed throughout the sites. There was a broad presence of metal tolerance that was indicated by genetic determinants related to heavy metal resistance of cadmium, cobalt, copper, lead and zinc (Figure 2.6). Furthermore, antimicrobial resistance genes specifically β-lactamases, were ubiquitous across all genomes at all sites based on the HMM annotations from the PFAM database (Figure 2.6). We further identified 90 genomes with antibiotic resistance genes using an exact sequence alignment (Figure A.10). Overall, antibiotic resistance was prevalent throughout all phyla, including one Creanarchaeota (archaeal) genome (Figs. 2.6, S A.10). Finally, we found that while amino acids were readily synthesised and degraded by most organisms (Figure 2.6), the degradation of polycyclic aromatic compounds appeared to be least prevalent in the high runoff sites.

2.5 Discussion

While recent work has focused on the structure of Arctic microbial communities (Ruuskanen et al., 2018; Comte et al., 2018), here we took a hypothesis-driven approach contrasting community structure and function between low and high runoff conditions, as a proxy for future environmental conditions under climate change scenarios. Previous work based on 16S rRNA gene sequencing data showed structural diversity and predicted functional redundancy in Lake Hazen sediments (Ruuskanen et al., 2018). The shotgun metagenomic approach presented in this study provides a less biased view of the sediment microbial community structure and function. Although it is possible that the microbes captured in the reconstructed genomes may not be metabolically active, a majority of the DNA reads recovered from sediment samples mapped to the genomes (Figure A.14), and our future work will aim to confirm the metabolic predictions using metatranscriptomic data. The genomes in these samples were similar to the taxa previously reported in other Arctic lake sediments (Ruuskanen et al., 2018; Carnevali et al., 2018; de Jong et al., 2018).

Notably, none of the reconstructed genomes at any site belonged to the Firmicutes phylum (Figure 2.3). However, the sequence-level analysis of the unbinned quality-controlled

reads indicated the presence of Firmicutes (Figure A.16). Their absence in the reconstructed genomes may be due to assembly bias in the form of divergence within Firmicute phylum (Chen et al., 2013; Sangwan et al., 2016) or incomplete binning (Sieber et al., 2018). Further complicating this identification issue is that two different taxonomy databases where used: GTDB for the reconstructed genomes and NCBI for sequence-level analysis. The reconstructed genomes were based on the placement of the genomes within a phylogenetic tree using the GTDB, that extensively reclassified the taxonomic placement of many genomes that were previously thought to be Firmicutes based on poorly supported monophyletic groups (Parks et al., 2018). In contrast, the sequence-level analysis was based exclusively on Diamond exact sequence matches (Menzel et al., 2016) providing a less reliable identification of individual reads. A possible solution to overcome this identification uncertainty in future studies is to use multiple binning algorithms, aggregate the results and only present the unique reconstructed genomes, which noticeably improves the number of genomes recovered (Sieber et al., 2018).

As expected, the microbial sediment communities of Lake Hazen were more phylogenetically diverse than the surrounding soils (Figure 2.2). Although geochemical measurements for the soil samples were absent, we still observed that genomes from the high runoff sites clustered most frequently with the soil sites, whether we looked at individual genome abundances (Figure 2.3) or assigned genomes to a site by abundance (Figure 2.5B). As the high runoff sites were most similar to the soil, this may suggest that the terrestrial environment contribute to the sediment community structure in Arctic watershed (Ruiz-González et al., 2015; Comte et al., 2018).

Microbial communities are spatially heterogeneous (Raynaud and Nunan, 2014) and minor differences (μm) in sampling location can change the community structure (Ramette and Tiedje, 2007). With this in mind, we collected triplicate cores at each site and pooled DNA extracts together, reducing any site-specific spatial variation. We observed that microbial assemblages differed across sites (Figures 2.2, 2.3), but a key remaining question is whether this spatial distribution pattern is a result of selection by present day environmental variables or historical contingencies. As others have previously noted, Arctic microbial

communities were quick to respond to changes in their geochemical environment (Castro et al., 2010; McCalley et al., 2014; Hultman et al., 2015). The clear environmental clustering of sites along geochemical gradients in this study (Figures 2.1, 2.3) leads us to suspect that the microbial community patterns observed here were most likely a result of the rapidly changing conditions.

The sediment sites clustered into four distinct groups based on the geochemical properties (Figure 2.1BC) and the differences in geochemical properties correlated with the community composition (Figures 2.4A, 2.5). This finding supports the microbial biogeography concept – that microbial distribution of biodiversity is influenced by environmental variables (Martiny et al., 2006). These structural differences across sites were not always translated into obvious changes in the functional capacity, suggesting some functional redundancy (Figure 2.6). However, some metabolic processes were partitioned with only a few taxa encoding for each marker gene or pathway. If those taxa or genomes were lost (i.e., below the abundance cutoff) the metabolic processes encoded would be lost too. We detected markers for carbon cycling and fermentation across all sites. Yet, the Wood-Ljungdahl pathway, which is the lowest energy cost carbon fixation pathway (Probst et al., 2018), was only represented in the high runoff sites.

The sulphur cycle fluctuated the most between sites (Figure 2.6). Few genomes encoded for the molecular mechanisms to metabolise sulphur in the high runoff site, except for Gammaproteobacteria. This metabolic plasticity – the ability to switch metabolic processes in response to environmental conditions, could explain the success of Gammaproteobacteria and Alphaproteobacteria across all sites. Although we cannot determine which processes were active, greater metabolic diversity may explain why these taxa were found at all sites.

There was a large variation in the abundance of genomes at each site within the phylogenetically diverse communities inhabiting the Lake Hazen ecosystem. The metabolic repertoire showed some redundancy, especially in carbon cycling where different phylogenetic groups encoded the same metabolic function (Figure A.15). However, this redundancy was less obvious when considering nitrogen and sulphur cycling. Whereas others have suggested that functional redundancy may stabilise the metabolic activity when community

composition varies (Dombrowski et al., 2018; Anantharaman et al., 2016; Linz et al., 2018; Louca et al., 2018), we observed that environmental perturbations were most likely responsible for changing the phylogenetic diversity and metabolic potential of these Arctic communities. This may indicate a more rapid evolution than previously thought.

A key question that arises from this work is whether changes in hydrological regimes will enhance or destabilise the microbial communities. Niche differentiation, where there is a reduction in competition through the use of different resources, may explain why the low runoff regime hosts a more diverse microbial community (Cordero and Polz, 2014). In the low runoff and control sites, there was likely more species competition for a limited supply of nutrients, leading to structural and functional diversification. We found evidence suggesting that the increased runoff is destabilising the community structure, potentially disrupting niche differentiation, and reducing the metabolic capacity of these organisms. While not within the scope of this study, we predict that in the high runoff regimes microbes are switching metabolic functions to fill the ecological niche created by low oxygen, low redox, and high ammonia concentrations.

The Lake Hazen watershed is rapidly changing in response to climate warming, and an ideal location to investigate the transformations to microbial community structure and functional capacity. Here we identified 300 unique bacterial and archaeal lineages that varied spatially in response to differential hydrological regimes. The differences in structure and function of the genomes across hydrological regimes is likely linked to the geochemical properties and energy sources at each site. We propose that a combination of increasing runoff and changing geochemical conditions homogenised the microbial community structure and reduced the metabolic capacity. Further studies of this kind are needed to examine the continued shifts in microbial structure and metabolism through both space and time, as these changes in Arctic ecosystems may impact key biogeochemical cycles. We find evidence that these highly diverse Arctic microbial communities may be vulnerable to alterations in flow regime and nutrient delivery expected with continued climate warming.

Chapter 3

Research synthesis

3.1 Summary

The overarching objective of my thesis research was to provide new insights into the microbial responses to climate change in the Arctic. Using metagenomic tools and geochemical analyses, I described novel transformations in microbial structure and function in a High Arctic environment. In Chapter 2, I identified distinct geochemical groups (Figure 2.1B) that conform to what is known about the hydrological regimes in the Lake Haze watershed (Table A.1). I sampled sites that belong to the following hydrological regimes: high, low, and negligible runoff. In this study, I tested the hypothesis that differential runoff, a consequence of climate change, produces unique geochemical environment that affects the microbial community structure and function of sediments. Using a space-for-time approach (Lester et al., 2014), I compared microbial communities along a runoff gradient to predict the ecological responses caused by transitions in runoff. I found that a combination of increased runoff and altered geochemical conditions homogenised the microbial community structure and reduced the metabolic capacity. Whether the sites were grouped by geochemical properties, abundance of each reconstructed genome, or metabolic capacity, the high runoff sites always clustered independently from the low-runoff sites (Figures 2.1, 2.3, 2.4). This suggests that contemporary disturbances, in the form of high runoff regimes, shift

the underlying microbial community structure and potential functions, which may disrupt major biogeochemical cycles (Anantharaman et al., 2016; Fuhrman, 2009).

3.2 Scientific Contributions

This work makes original contributions to the understanding of microbial responses to climate change in the Arctic. Here, I provided evidence that increased runoff, a condition that is expected to continue with ongoing climate change, disrupts the structure and function of microbial communities. Considering that Arctic environmental conditions are changing more rapidly than elsewhere on the planet (Laudon et al., 2017; Overpeck et al., 1997; Serreze and Francis, 2006), understanding the alterations to this environment and its impact on microbial communities is paramount to predicting ecosystem responses to climate change.

This study linked increased runoff with a reduction in diversity and functional potential of microbial genomes (Figures 2.2, 2.6). Others have found functional redundancy in microbial communities, such that changes in the environmental conditions do not have broader impacts on the community function (McCalley et al., 2014; Louca et al., 2018; Ramírez-Flandes et al., 2019). I found evidence to the contrary, particularly in the high runoff environments most effected by climate change – where few functions are encoded across multiple genomes. If these genomes are lost or even reduced to low abundance in the environment this could have broader consequences on ecosystem services (Allison and Martiny, 2008). This loss of functional potential has been demonstrated by others in microcosm studies (Delgado-Baquerizo et al., 2016), but here I demonstrate this concept *in situ*.

This study is unique because of the focus on the High Arctic, use of metagenomics, and high sequencing coverage. Very few studies have investigated changing microbial community structures in Arctic environments (Yergeau et al., 2012; Comte et al., 2018) and even fewer of these studies have looked at these changes at the polar extremes, above a latitude of 80°N. The Lake Hazen study site (Figure 2.1A) offers the advantage of being an intensely studied watershed with a wealth of previously collected hydrological (Smith, 1999; Köck

et al., 2012; St. Pierre et al., 2019), geochemical (Emmerton et al., 2016; St. Pierre et al., 2018; Lehnherr et al., 2018), and more recently microbial data (Ruuskanen et al., 2018). The use of shotgun metagenomics has become a common approach in microbial ecology, but likely because of cost, it is relatively unused in an Arctic setting (Hultman et al., 2015; Varin et al., 2012; Woodcroft et al., 2018; Yergeau et al., 2017; Nash et al., 2018). To our knowledge, this is the first study to use shotgun metagenomics on High Arctic sediments. Inherently, this work provides new insights into whole genome and gene abundance in this diverse ecosystem. Finally, the sequencing effort attempted to maximise the amount of reads generated per sample to confidently identify rare genomes from the environment (Howe et al., 2014; Quince et al., 2017).

This research addressed the need for more information on microbial responses to climate change in the Arctic. Understanding the relationship between microbial diversity and ecosystem functioning is essential for predicting transformations in ecosystem services and nutrient cycles (Allison and Martiny, 2008). I described essential baseline data on the genome abundance and functional potential, that will be important when comparing changes found in future studies. The findings support the concept that microbial composition is sensitive to environmental disturbances and that these disturbances may have direct impacts on ecosystem services.

3.3 Limitations and Future Work

In this study, there are two primary limitations that need to be acknowledged: the small number of samples and lack of replicates. Due do the timeline and expense, I used data gathered from a single field season. As I only characterised a single timepoint, I was not able to compare our observations to any seasonal variability. Ideally, data would have been collected over multiple seasons to capture long-term environmental shifts. Using a space-for-time ecological approach, where samples are spread across a spatial gradient, corrected for some of this uncertainty. This sampling effort still provided valuable information on microbial responses to changing environmental conditions and baseline data on microbial

diversity in Arctic lake sediments. Furthermore, as we only had eight samples the ability to interpret statistical significance in the distribution of genomes or functions with runoff regimes was challenging. A study with more samples, at least three per condition of high, low, and negligible runoff, would have provided greater statistical power, but was ultimately unfeasible. This study was also limited by a lack of technical and biological replication. Although, biological replicates were collected in the form of triplicate cores from each site, DNA was extracted from each and combined prior to sequencing. This reduced our statistical power, but allowed us to incorporate the spatial variability at each site. Ultimately, I generated high sequence coverage (approximately 130 million reads per sample), which should be sufficient in capturing rare community members (Howe et al., 2014). Technical replicates – extracting and sequencing each individual sample, was not feasible for the goals of this project.

Additionally, the reliance exclusively on metagenomic sequencing did not allow for the differentiation of live, dead, and dormant cells. I was able to characterise the microbial community; however, it is unlikely that all of these genomes are metabolically active (Albertsen et al., 2015). This could have wide implications on our interpretation of function, but does not fundamentally change what is observed as reduced functional potential in high runoff regimes. One of the more challenging aspects of using metagenomics is elucidating functional genes and pathways related to ecosystem services. I used both genes and pathways to characterise the functional potential; however, the presence of a gene or pathway does not always indicate that it is actively transcribed. Initially this project was supposed to include a metatranscriptomic analysis of the mRNA extracted from the whole community. This data has been generated, but yet to be analysed. Future work would benefit from incorporating a multi-omic approach to align the recovered transcripts back to the individual genomes (Hultman et al., 2015). A metatranscriptomic approach will allow for both the differentiation of live organisms from those that are dead and dormant, and provide greater details into the functional distinctions between the different geochemical and runoff regimes.

A fundamental issue that arises from metagenomic assemblies is that these recon-

structed genomes represent population-level genomes, and are more reflective of the average identity of genomes within a specific taxonomic classification (Sharon and Banfield, 2013; Parks et al., 2015). Although metagenomics captures the genomic content in uncultivated microorganisms (Wrighton et al., 2012), it may entirely miss rare, but important genetic variation that has a role in microbial composition and function (Brown et al., 2015). This issue may be improved by sequencing technologies that generate more sequencing coverage and longer reads (van Dijk et al., 2018; Jain et al., 2018).

Furthermore, our analysis does not present the diversity of the entire microbial community, as we excluded the analysis of both viruses and eukaryotes. Despite being the most abundant biological entity on Earth and having a influence on microbial ecology and evolution (Pan et al., 2017; Moniruzzaman et al., 2017), viruses are not considered in this study. Others have used metagenomics analysis to evaluate the structure of viral communities isolated from environmental samples (Culley et al., 2006), but few have incorporated prokaryotic and viral characterisations in a single study. Newer bioinformatic tools, such as FastViromeExplorer, VirMAP, and virMINE, have been developed to search prokaryotic metagenomic datasets for viral sequences (Tithi et al., 2018; Ajami et al., 2018; Garretto et al., 2019). We did not actively remove viral sequences, as such, it is possible they are included in our reconstructed genomes and or large proportion of unmapped reads in each sample (Figure A.14). In contrast eukaryotic contigs were actively removed from the assembly process prior to binning with Eukrep (West et al., 2018). Using these isolated eukaryotic sequences others have been able to reconstruct eukaryotic genomes (West et al., 2018). However, reconstructing eukaryotic genomes will likely remain challenging because of the low relative abundance of eukaryotic DNA to prokaryotic DNA found in environmental samples (Daniel, 2005). Future work should investigate viral and eukaryotic contributions as it would provide a more unified view of this ecosystem.

Finally, the future of metagenomics work will require moving studies from the field back into the laboratory. Metagenomic analyses have produced the common findings that microbial communities are reliant on other community members to drive geochemical and nutrient cycles (Hug et al., 2018). These sequence based methods have provided glimpses at indi-

vidual microorganisms functional roles (Anantharaman et al., 2016). However, enrichment-based methods are required to test predictions derived from environmental sequence data. Enrichment cultivation, where a microbial community can be provided specific nutrients, energy sources, and environmental variables allows for further hypothesis generation and testing (Luo et al., 2016). Characterisation of enriched microbial communities, will ultimately allow for a more complete mechanistic understanding of functional genes and pathways.

Appendix A

Supplemental information for Chapter 2

A.1 Supplemental Figures

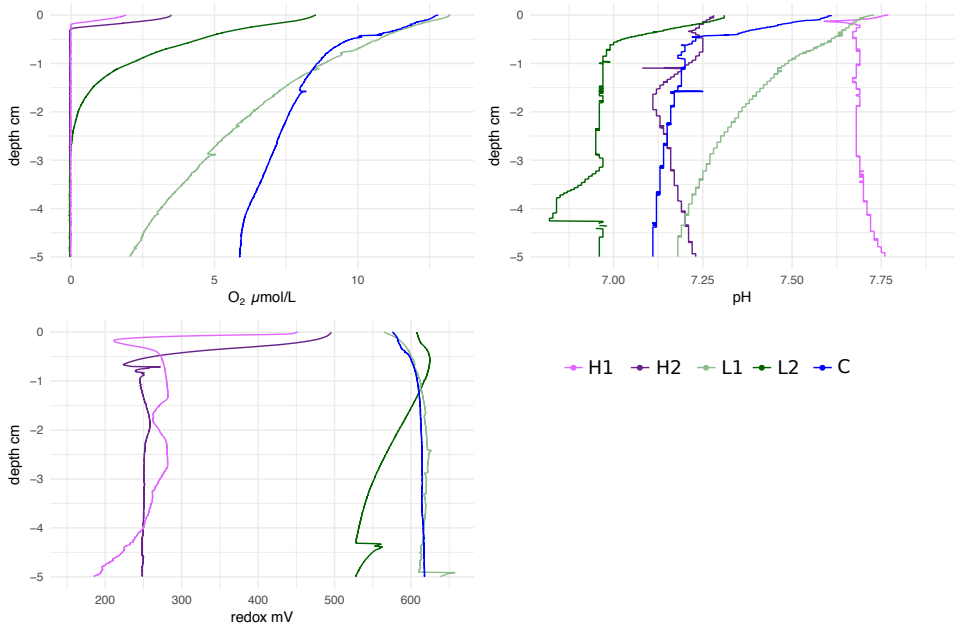

Figure A.1: Sediment microprobe profiles for oxygen, pH, and redox measured in 100 μm intervals. Legend (bottom right) indicates the colour of each sediment site in the profiles.

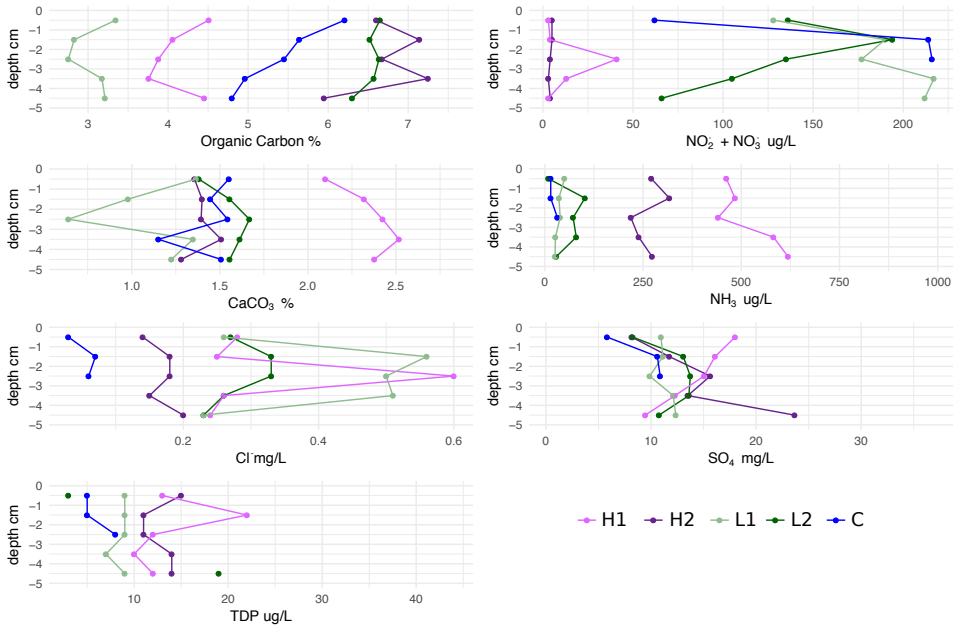

Figure A.2: Sediment porewater profiles for organic carbon, nitrite and nitrates, calcium carbonate, ammonia, chlorine, sulphate, and total dissolved phosphorus (TDP). TDP was removed when producing PCA and boxplots in Figure 2.1 because of incomplete measurements in C and L2. Legend (bottom right) indicates the colour of each sediment site in the profiles.

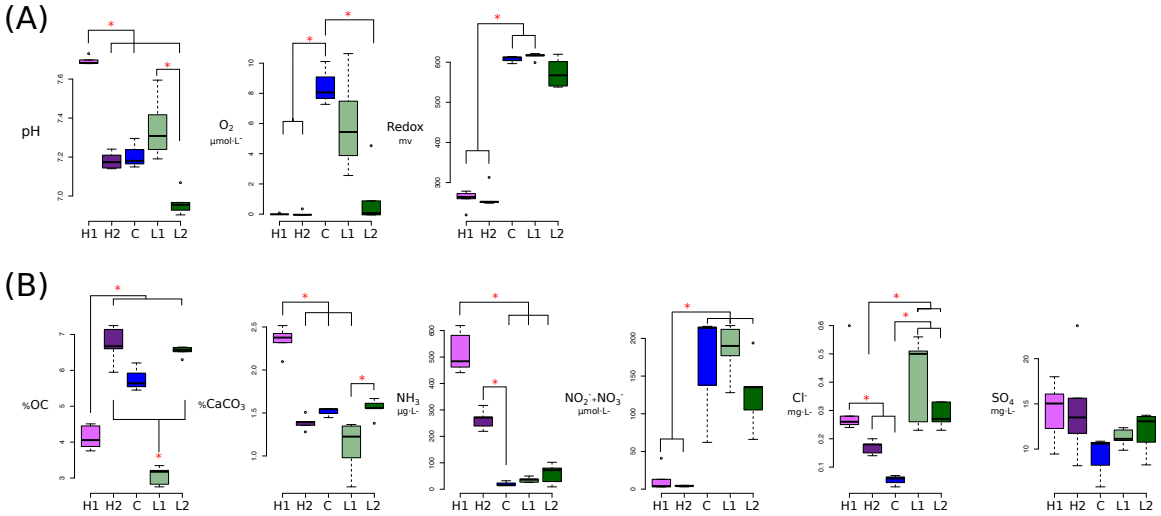

Figure A.3: Distribution of all chemical features for sediment sites. (A) Microprobe measurements collected at every 100 μm. (B) Porewater measurements collected from bulk 1cm intervals. Includes chlorine and sulphate measurements absent from Figure 2.1. Branches and asterisks indicate significant differences between sites $P < 0.025$ (Dunn Test). If branches form a dichotomy or trichotomy, the interactions within that group is not significant.

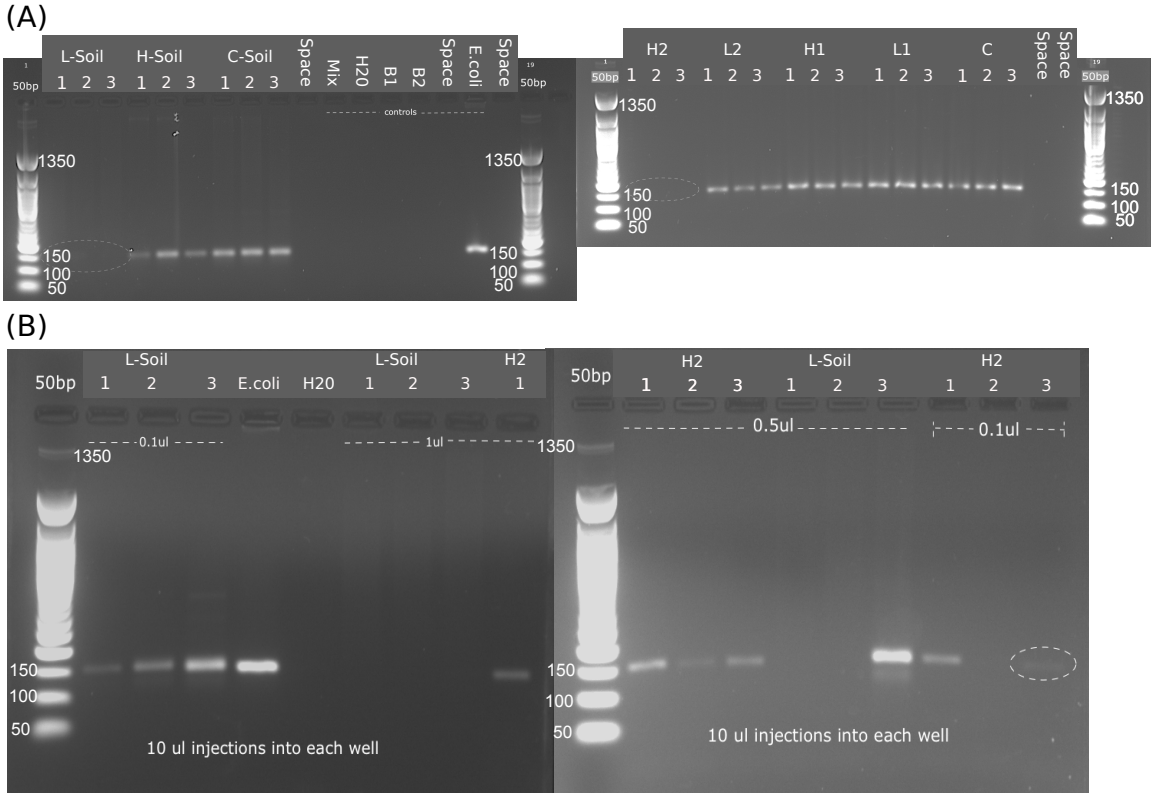

Figure A.4: Gel electrophoresis images of the *glnA* gene for each sample extracted in triplicate. (A) Initial gel results using 1 μL of DNA for each PCR reaction. Every sample contains *glnA* except L-Soil and H2. (B) Repeated PCR reaction for L-Soil and H2. Diluted DNA concentrations for PCR from 1 μL to 0.1 μL and 0.5 μL. 50 bp ladder values are labelled along the sides of each gel.

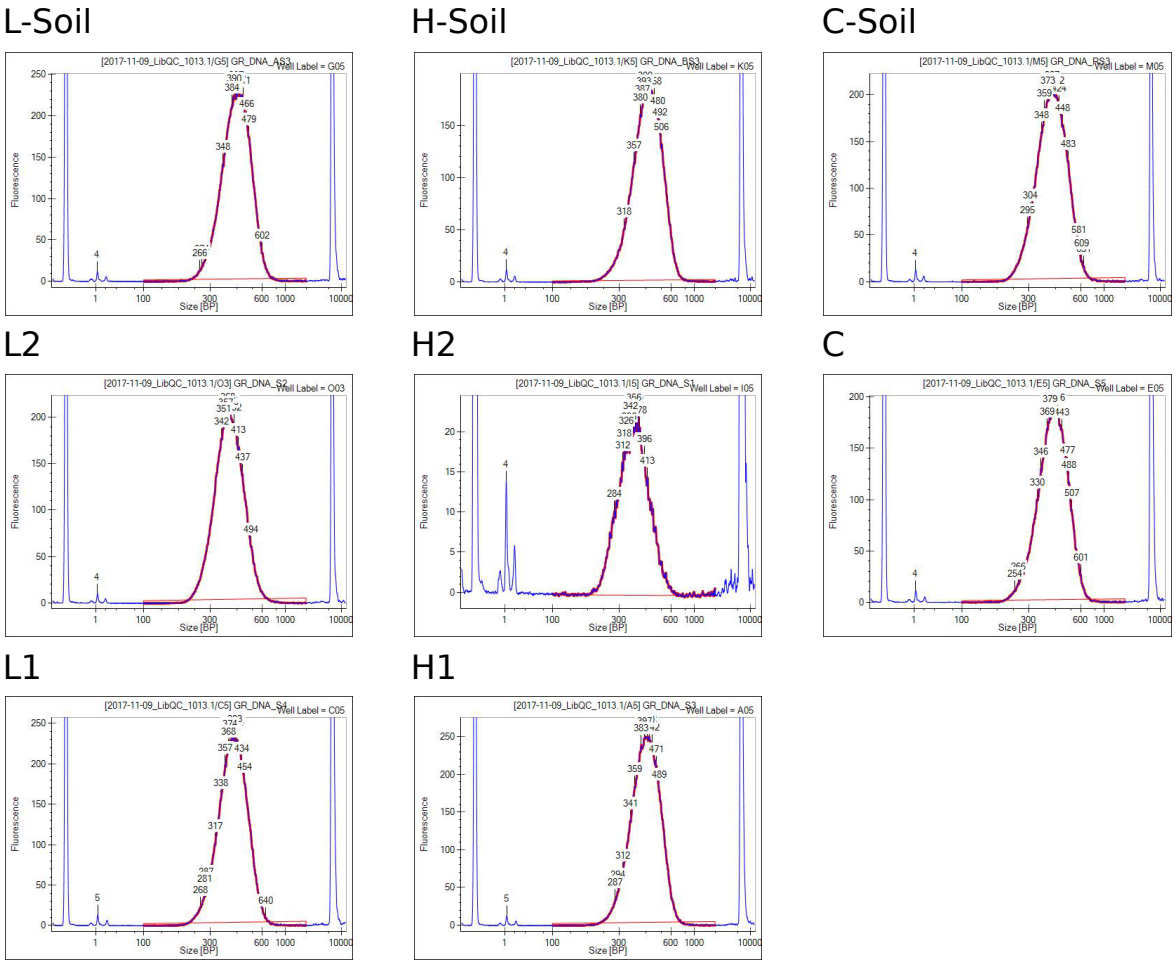

Figure A.5: Illumina HiSeq DNA sequencing library validation from an Agilent Technology 2100 Bioanalyzer. All libraries were constructed using Illumina TruSeq DNA PCR-Free Library Prep. The library results for each of the 8 samples are shown and labelled above the panel: L-Soil, H-Soil, C-Soil, L2, H2, C, L1, H1. Peaks between 300-600 bp indicate the presence of DNA. Peaks below 1 bp and above 1000 bp are internal standards.

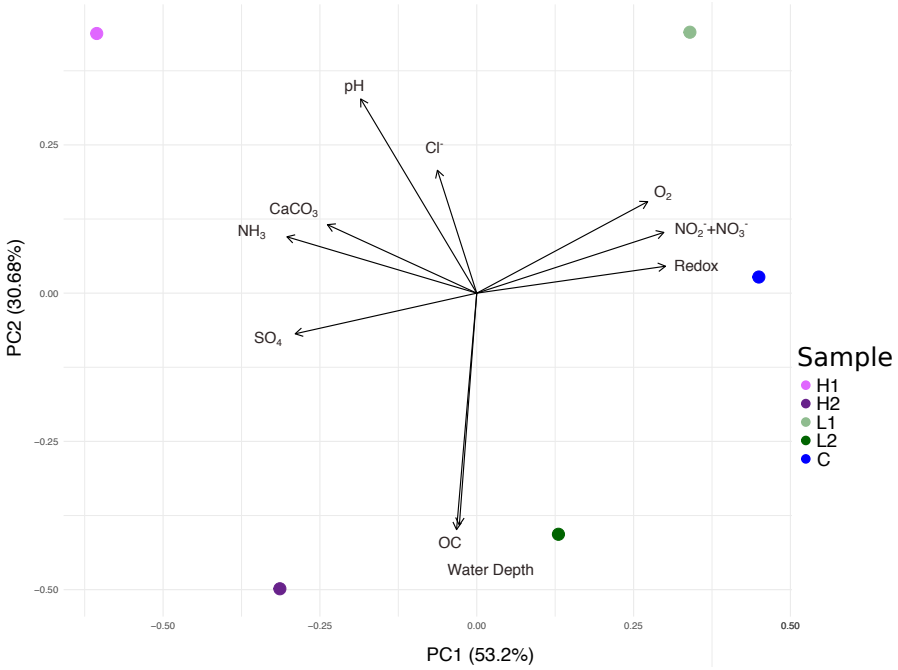

Figure A.6: Principal Component Analysis (PCA) showing the differences in chemical composition of the sediment sites. Individual points represent the sample locations based on a distance matrix of the mean values measured from 0-5 cm sediment depth. Sample legend shown.

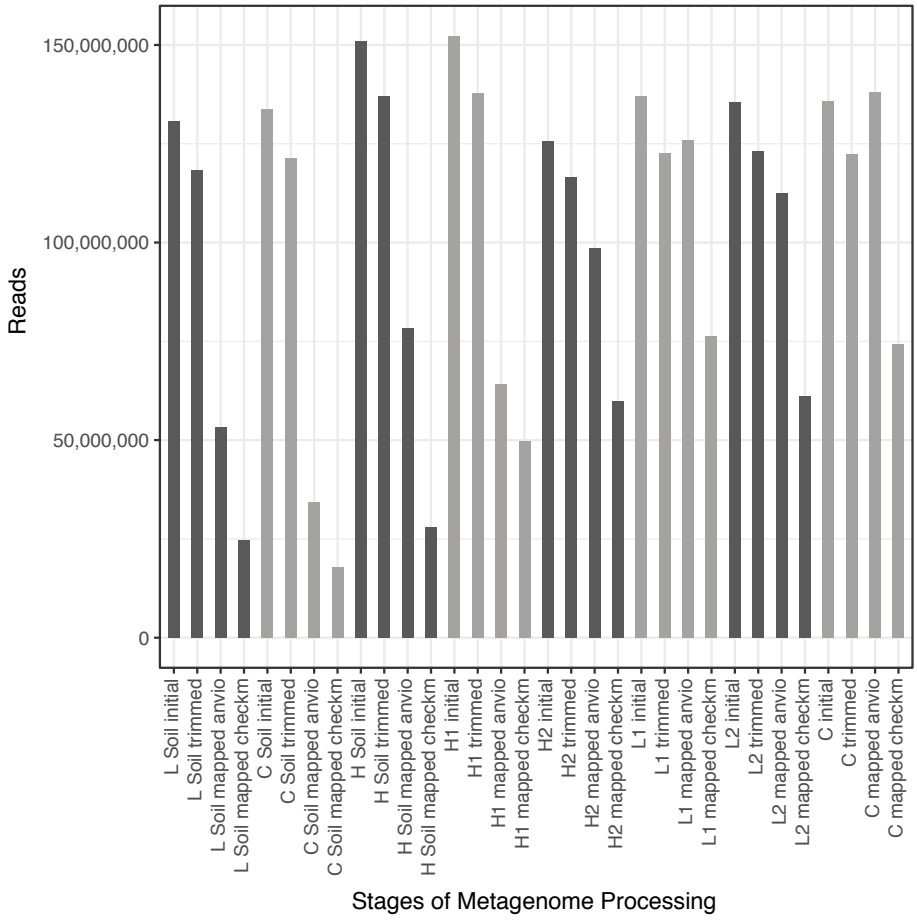

Figure A.7: Number of reads throughout analysis. Reads are successively decreasing, except for two instances, "C mapped anvio" and "L1 mapped anvio", related to the Anvio contig (Eren et al., 2015) database that have higher counts. As each one of the paired reads can align discordantly using Bowtie2 (Langmead and Salzberg, 2012) to more than one locus, they are reported as two mapping instances. Ultimately, the checkM (Parks et al., 2015) mapped read counts were used to generate abundance values and in all further analysis. The reads mapped using checkM include only reads that map to the 300 high quality reconstructed genomes, opposed to all reconstructed genomes in the Anvio mapping.

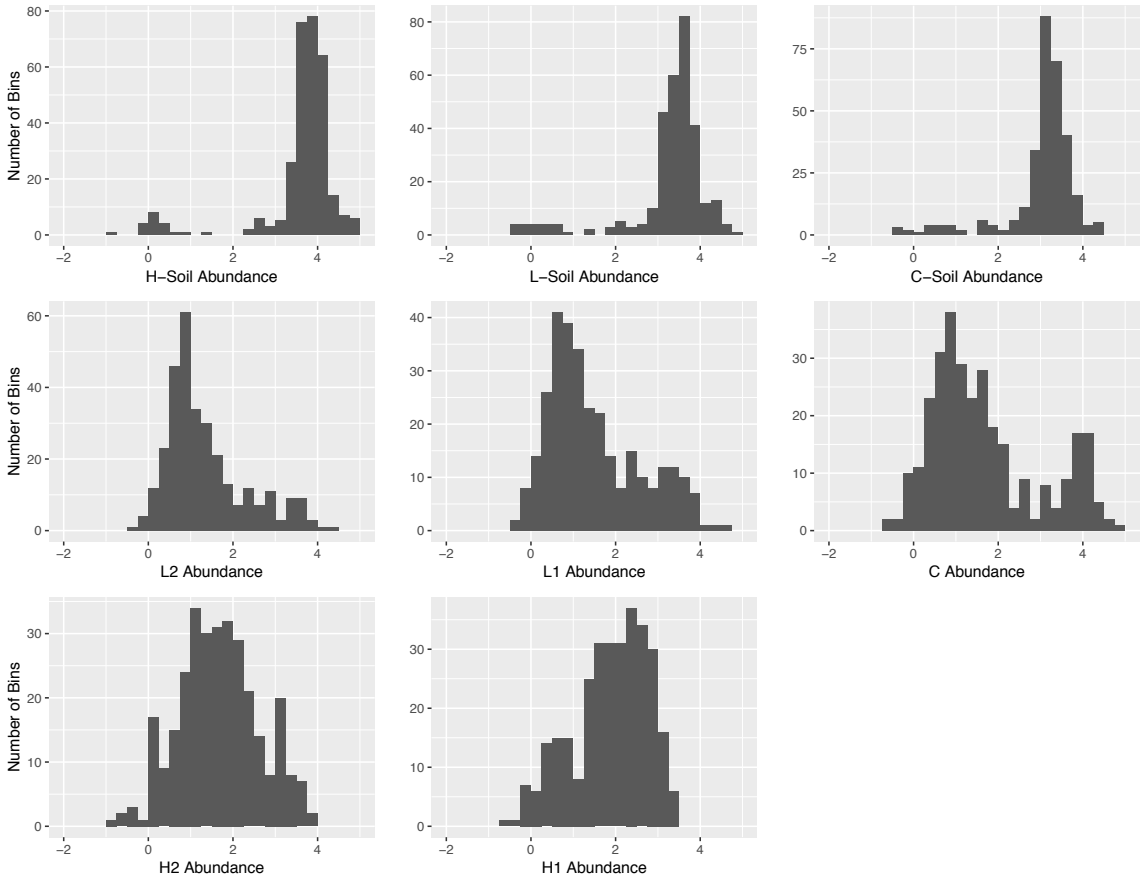

Figure A.8: Genome abundance per sample ($-\log_{10}$ scale). More negative genomes are found in greater abundance at each site. For further functional analysis genomes are only considered to contribute to a site if $\leq 0.56\%$ ($-\log_{10} \leq 0.25$) of reads per sample mapped to a genome.

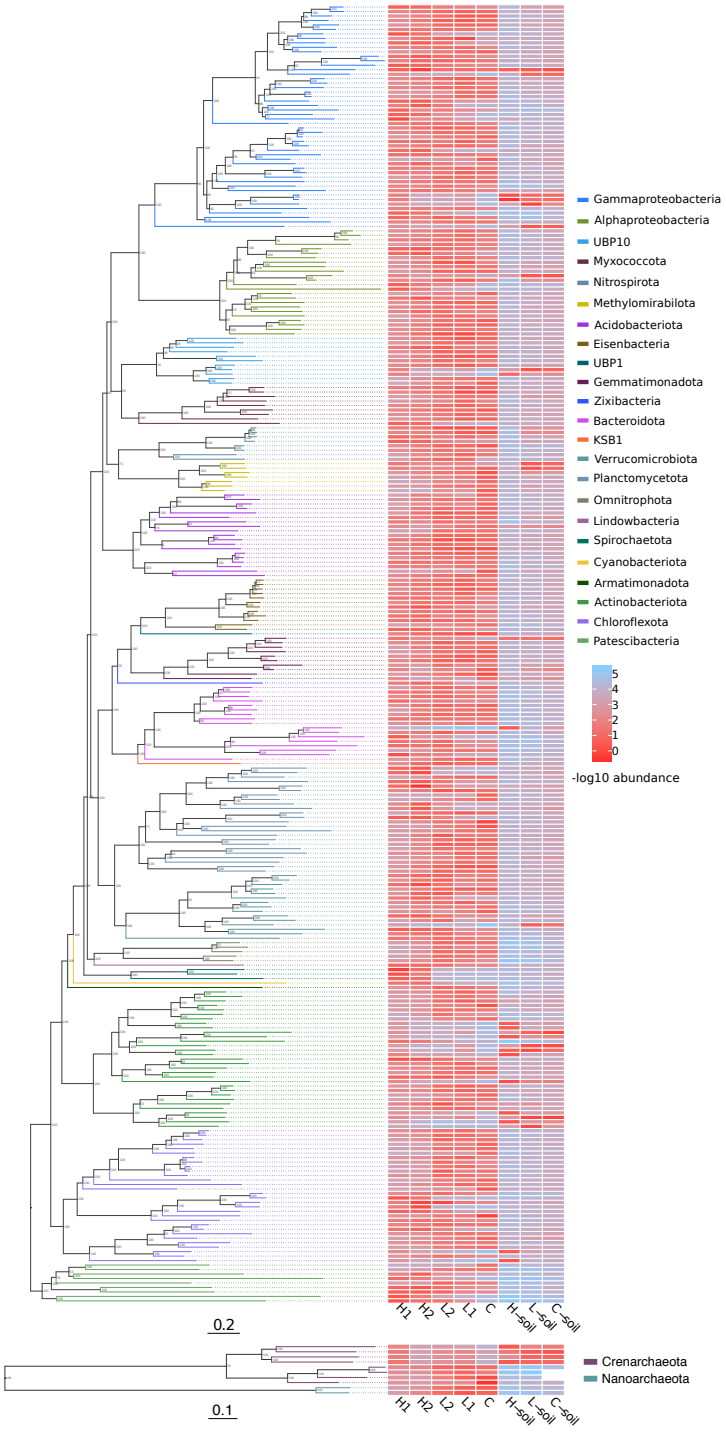

Figure A.9: Abundance of genomes are shown on a $-\log_{10}$ scale were more negative values (red) are more abundant than positive values (blue). The abundance values correspond by row with the genomes phylogenetic assignment. Support values (zoom to view) for phylogenetic tree are shown at each node.

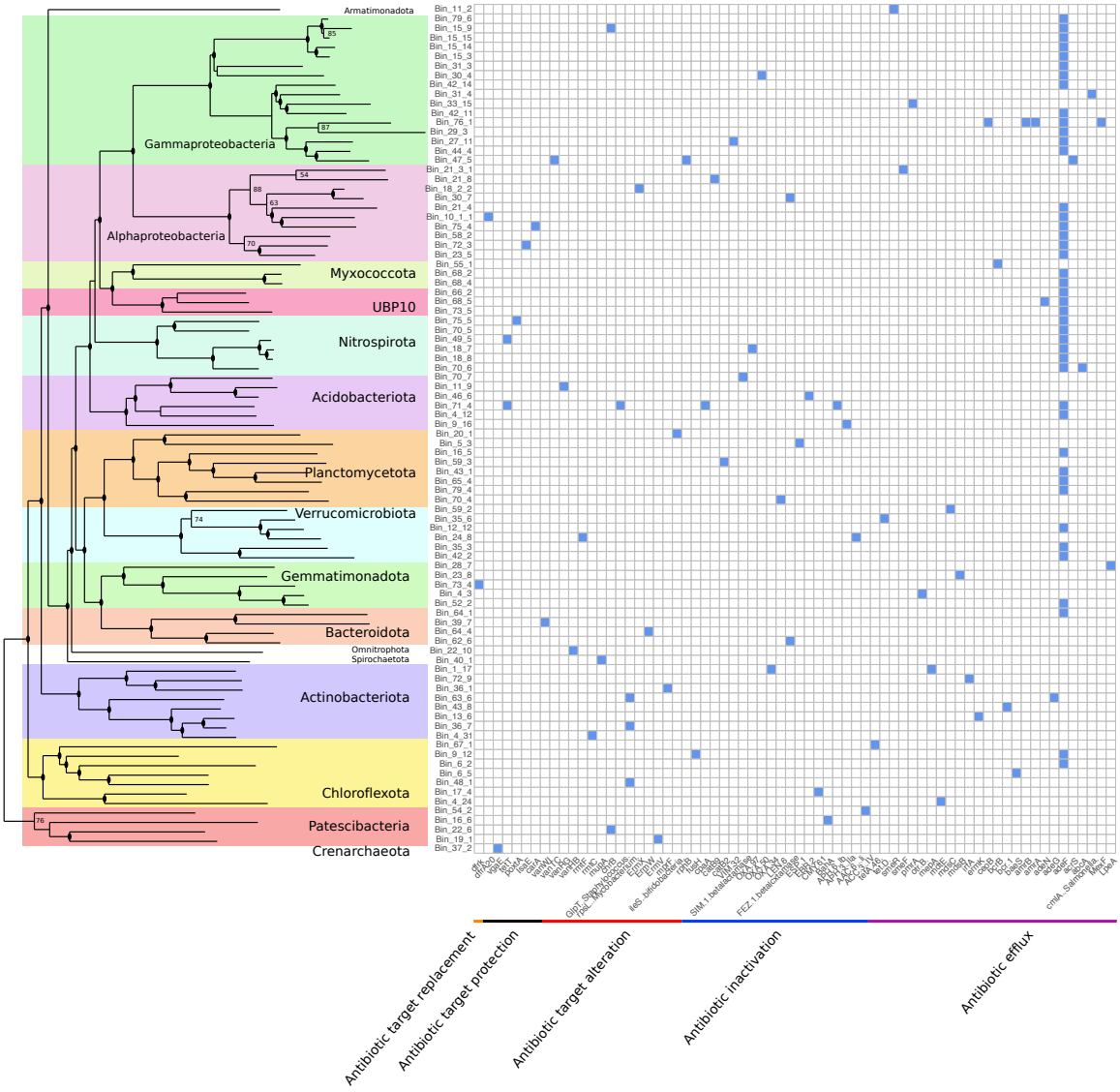

Figure A.10: Antibiotic resistance genes (AMR) identified from each of the high quality genomes using the Comprehensive Antibiotic Resistance Database (CARD 2.0.3). AMR genes were identified using the Resistance Gene Identifier (RGI 4.2.0). AMR genes are cluster by mechanism. "Strict" matches (within detection model cut-offs) to resistance genes (blue) and no matches (white) are shown. The RGI tool uses either BLAST or DIAMOND to align genes from the reconstructed genomes to AMR genes in the CARD database. We chose to use DIAMOND to perform an exact sequence alignment to identify AMR genes. Reconstructed genomes are placed vertically within a phylogenetic group from GTDB tree. Support values are shown, for confidence ≥ 90 are a solid dot and values < 90 are shown numerically. Only the 90 reconstructed genomes that contain an AMR gene are shown.

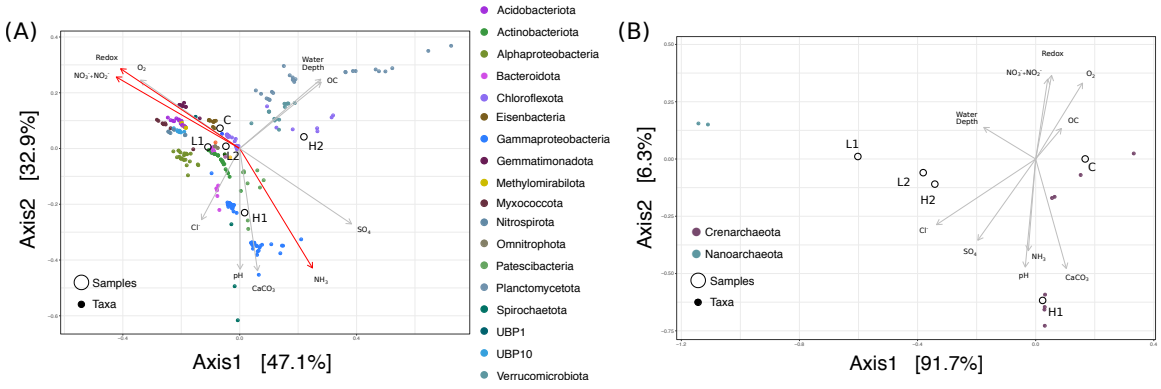

Figure A.11: Double Principle Coordinate Analysis (DPCoA) on reconstructed genome abundances for sediment sites with geochemical vectors. (A) DPCoA of bacterial genome abundances. (B) DPCoA of archaeal genome abundances. Phylum level genome classifications are shown.

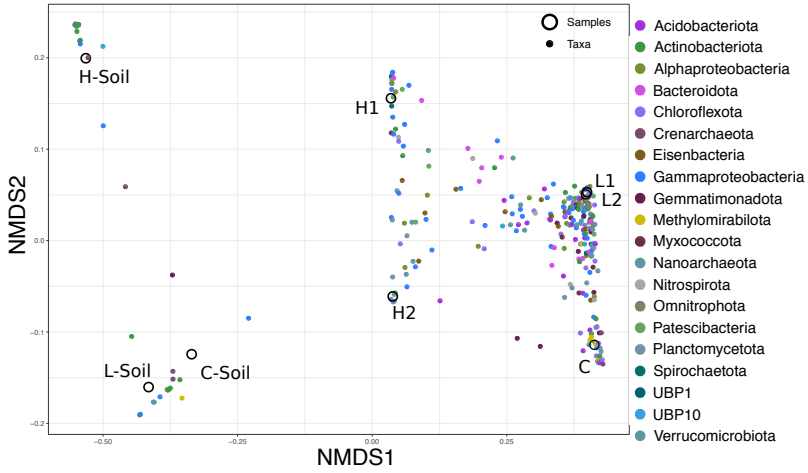

Figure A.12: Non-metric multidimensional scaling (NMDS) on reconstructed genome abundances for all sites.

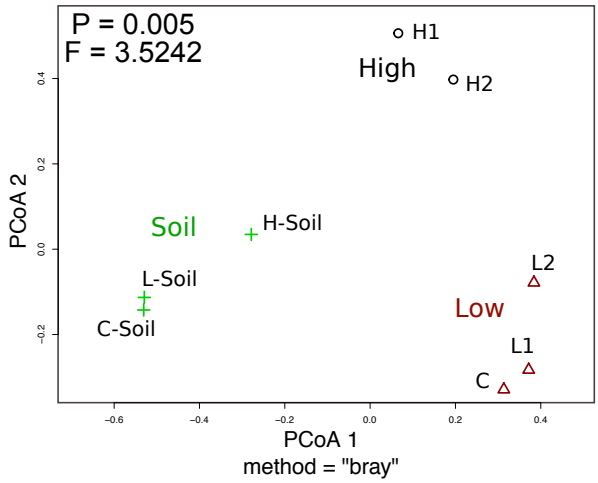

Figure A.13: Principal coordinate analysis (PCoA) on reconstructed genome abundances for all sites. Grouping the sites into soil, high, and low runoff was proven to be significant (PERMANOVA test: $F = 3.52$, $P = 0.005$).

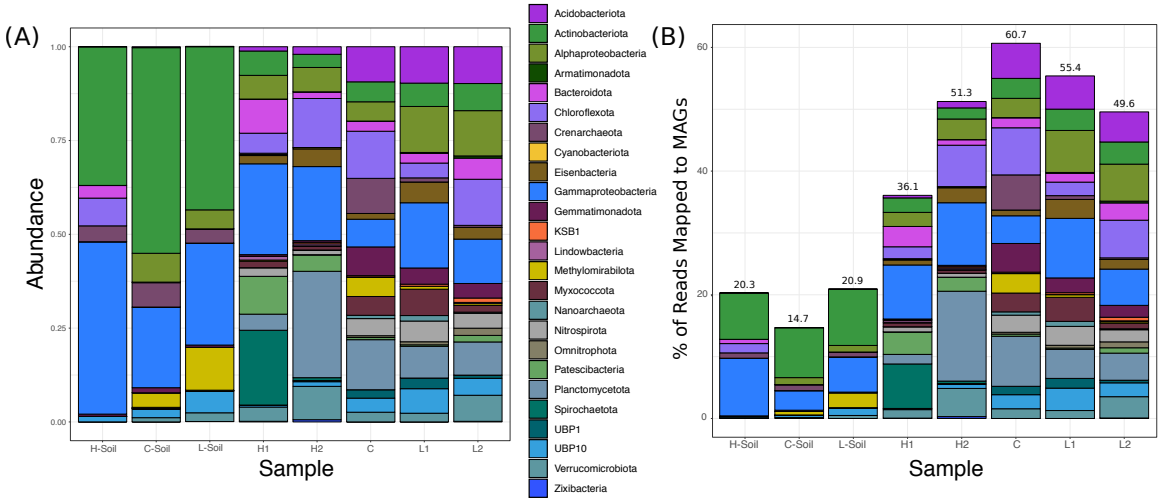

Figure A.14: Reconstructed genome abundance across sites. (A) Amount of reads mapped to the 300 high quality genomes for each sample normalised to 100%. Only reads that were mapped to genomes are shown and all unmapped reads have been excluded. (B) Amount of reads mapped to the 300 high quality genomes for each sample.

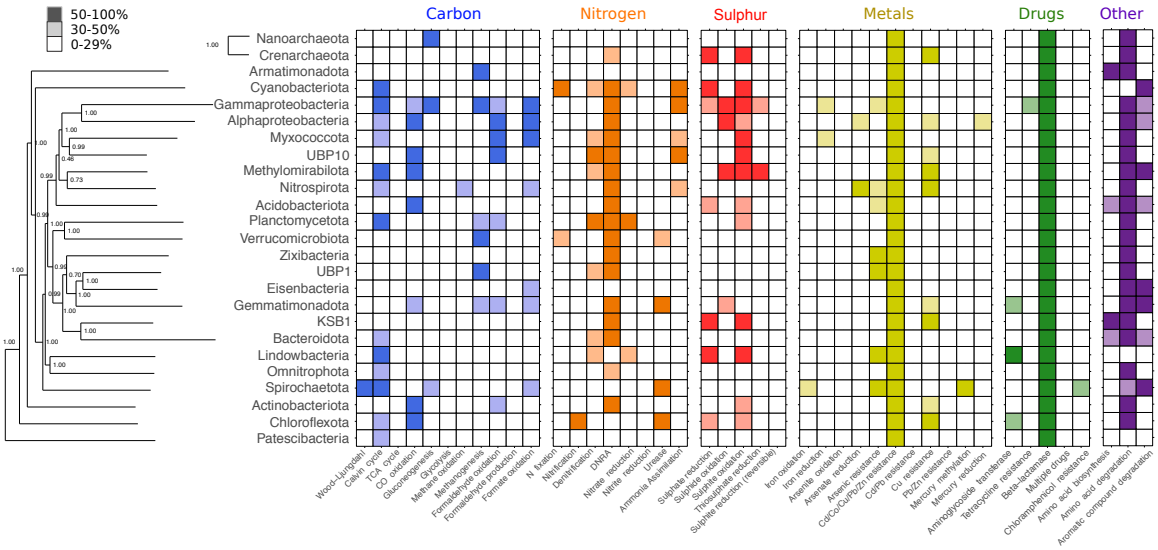

Figure A.15: Core metabolism detected in all of the Lake Hazen genomes. Presence of core metabolic genes involved in carbon metabolism, nitrogen metabolism, sulphur metabolism, metal cycling, antibiotic resistance, and other metabolism. No colours: genes predicting function are absent or low abundance. Shaded colours: genes predicting function are present in 30-50% of genomes per phylogenetic group. Solid colours: genes predicting function are present in >50% of genomes per phylogenetic group. Condensed phylogenetic trees are shown with support values shown.

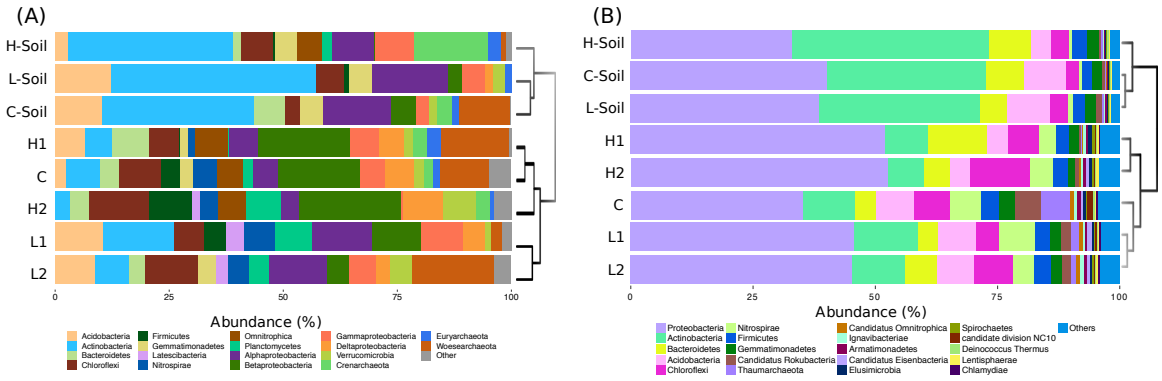

Figure A.16: Sequence level taxonomic classification. (A) Taxonomic abundance on 16S rRNA genes using MATAM Pericard et al. (2017). Recovered 16S rRNA genes are classified SILVA SSU 16S sequence database version 128. (B) Taxonomic abundance on quality controlled sequences using Kaiju Menzel et al. (2016). Sequences are classified using NCBI BLAST non redundant database accessed on 2018-05-16.

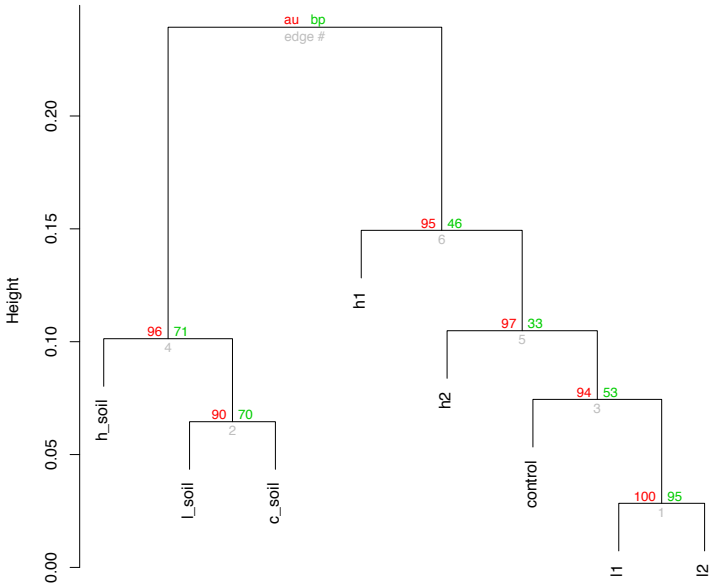

Figure A.17: Metabolic capacity for carbon, sulphur, and nitrogen cycles clustered with P-values via multiscale bootstrap resampling. AU (approximately unbiased) P-values are shown in red, with any value > 0.95 (significance level 0.05) being significant. BP (bootstrap probability) values are shown in green. Hierarchical clustering was completed using "correlation" as distance method and "average" as clustering method.

A.2 Supplemental Tables

Table A.1: Glacial runoff to the Lake Hazen Watershed. Lengths of rivers in km are shown in parentheses. Mass balance modelled runoff for years 2015 and 2016. Sampling dates in 2017 were prior to the summer runoff. Data adapted from St. Pierre et al. (2019).

	Surface area (km^2)			**Runoff volume (km^3)**	
River	**Catchment**	**Glacier**	**River**	**2015**	**2016**
High Runoff: Abbé (AB)	390	204	7.9 (21)	0.061	0.015
Low Runoff: Blister (BR)	n/a	6	2.5 (11)	0.002	<0.001
Gilman (GL)	992	708	5.1 (22)	0.192	0.08
Henrietta Nesmith (HN)	1274	1041	9.6 (4.6)	0.291	0.075
Snowgoose (SG)	222	87	5.6 (17)	0.026	0.006
Turnabout (TN)	678	259	13.4 (42)	0.082	0.024
Very (VR)	1035	269	32.9 (39)	0.165	0.08
Watershed total	7516	3078	91.2	0.979	0.291

Table A.2: Sediment deposition dates and rates for the two deep sites, H2 and L2, is based on ^{210}Pb constant rate of supply (CRS) dating model. Analysis was completed on 0.5 cm core intervals. Data adapted from St. Pierre et al. (2019).

H2 (Abbe Deepsite)			L2 (Blister Deep Site)		
Interval	Midpoint CRS date (CRS year)	Sedimentation rate (g/cm^2/yr)	Interval	Midpoint CRS date (CRS year)	Sedimentation rate (g/cm^2/yr)
0-0.5	2017.1	0.349	0-0.5	2016.5	0.0724
0.5-1	2016.6	0.292	0.5-1	2014.0	0.0701
1-1.5	2015.9	0.111	1-1.5	2011.3	0.1163
1.5-2	2014.1	0.671	1.5-2	2009.2	0.1759
2-2.5	2012.7	0.671	2-2.5	2007.3	0.1232
2.5-3	2012.7	0.227	2.5-3	2004.7	0.0909
3-3.5	2012.0	5.563	3-3.5	2001.9	0.0967
3.5-4	2011.3	5.563	3.5-4	1998.8	0.0854
4-4.5	2011.3	5.563	4-4.5	1994.0	0.0513
4.5-5	2011.3	5.563	4.5-5	1987.2	0.0375
5-5.5	2011.3	5.563	5-5.5	1976.7	0.0220
5.5-6	2011.3	5.563	5.5-6	1966.51	0.0515
6-6.5	2011.3	5.563	6-6.5	1957.66	0.0339
6.5-7	2011.3	5.563	6.5-7	1949.00	0.0537
7-7.5	2011.3	5.563	7-7.5	1940.41	0.0451
7.5-8	2011.3	5.563	7.5-8	1931.79	0.0668
8-8.5	2011.3	5.563	8-8.5	1925.07	0.0553
8.5-9	2011.3	0.441	8.5-9	1918.10	0.0553
9-9.5	2010.9	0.224	9-9.5	1911.08	0.0553
9.5-10	2009.8	0.123	9.5-10	1904.42	0.0553

Table A.3: Sample location, temperature at sampling time, and date.

Sample	Location	Temperature (°C)	Date
H-Soil	81° 84' 840" N; 70° 83' 849" W	-5	June 3, 2017
L-Soil	81° 80' 332" N; 71° 54' 239" W	0	June 7, 2017
C-Soil	81° 79' 382" N; 70° 44' 486" W	-2	June 3, 2017
H1	81° 84' 150" N; 70° 85' 175" W	not measured	May 24, 2017
H2	81° 82' 493" N; 70° 71' 498" W	not measured	May 29, 2017
C	81° 80' 343" N; 70° 50' 447" W	not measured	May 27, 2017
L1	80° 80' 521" N; 70° 52' 699" W	not measured	June 1, 2017
L2	81° 79' 171" N; 71° 46' 926" W	not measured	June 2, 2017

Table A.4: Number of bins recovered at each assembly step pre and post-manual refinement with Anvio.

	Anvio-Output (CONCOCT binning)	**CheckM**	**Anvio-Output (manually refined)**	**Final CheckM**
Total Bins	850	850	877	877
Completion, Redundancy				
C: > 90, R: <10	52	74	52	72
C: >70, R: <10	180	199	178	198
C: >50, R: <10	321	324	309	300
C: <50, R: any	529	526	568	577
Contamination				
<1	233	261	255	286
<5	544	632	590	676
>10	8	51	7	27
Good bins but contaminated				
C: >50, R: >10	2	27	0	2
C: >70, R: >10	1	6	0	0
C: >90, R: >10	0	2	0	0

Table A.5: Contigs assembled with Megahit and used in Anvio database. NA: MegaHit contigs did not have gene predictions.

	Contigs	**Total Length (bp)**	**Min Contig Length (bp)**	**Max Contig Length (bp)**	**Avg Contig Length (bp)**	**N50 (bp)**	**# of genes prodigal**
MegaHit	12026467	8477069127	200	792468	705	756	NA
Anvio	1455655	3414974759	1000	792468	1183	285024	4254625

Table A.6: DNA extraction masses. DNA was extracted in triplicate for each sample and then combined prior to sequencing.

Lake Hazen	Sample Location	Tube I.D.	Wet Weight (g)				PCR with *glnA*
			1	2	3	Total (grams)	
Sediment	H2: Deephole	S1	0.416	0.455	0.499	1.370	yes (diluted)
	L2: Blister Deep	S2	0.497	0.455	0.389	1.341	yes
	H1: Abbe	S3	0.469	0.514	0.552	1.535	yes
	L1: Blister Shallow	S4	0.429	0.402	0.361	1.192	yes
	C: Ruggles	S5	0.518	0.447	0.331	1.296	yes
Soil	L-Soil: Blister Soil	BS3	0.527	0.535	0.561	1.623	yes (diluted)
	H-Soil: Abbe Soil	AS3	0.443	0.537	0.417	1.397	yes
	C-Soil: Ruggles Soil	RS3	0.447	0.500	0.447	1.394	yes

Table A.7: DNA fluorescence assay quantification for each sample. Note: H2 sequencing required using full extraction volume of 65 μL to reach an appropriate concentration.

Sample	Volume (μL)	Concentration (ng/ul)	Total DNA (ng)	NanoDrop (concentration ng/ul)
H2	65	0.08	3.36	4.6
L2	42	2.55	107.1	7.6
H1	42	4.33	181.86	10.2
L2	42	17.98	755.16	32.8
C	42	21.2	890.4	36.4
H-Soil	42	38.14	1601.88	65.2
L-Soil	47	69.88	3284.36	149.0
C-Soil	47	74.38	3495.86	110.0

Table A.8: *P*-values for marker gene and pathway distribution between sites. A Fisher's exact test was used in the place of chi-square test for small count numbers. *P*-values of less than 0.05 are considered significant (labelled with an asterisk).

Functional Marker	High vs Low	High vs Control	High vs Soil	Low vs Control	Low vs Soil	Control vs Soil
Carbon Cycle	0.03298 *	0.0004998 *	0.0004998 *	0.07746	0.1384	0.001499 *
Nitrogen Cycle	0.07596	0.001999 *	0.0004998 *	0.7006	0.03448 *	0.1859
Sulphur Cycle	0.001999 *	0.002999 *	0.0004998 *	0.4913	0.02999 *	0.3823

Bibliography

Ajami NJ, Wong MC, Ross MC, Lloyd RE, Petrosino JF. 2018. Maximal viral information recovery from sequence data using virmap. Nature communications. 9(1):3205.

Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. 2015. Back to basics–the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One. 10(7):e0132783.

Allison SD, Martiny JB. 2008. Resistance, resilience, and redundancy in microbial communities. Proceedings of the National Academy of Sciences. 105(Supplement 1):11512–11519.

Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods. 11(11):1144.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25(17):3389–3402.

Alves RJE, Wanek W, Zappe A, Richter A, Svenning MM, Schleper C, Urich T. 2013. Nitrification rates in arctic soils are associated with functionally distinct populations of ammonia-oxidizing archaea. The ISME Journal. 7(8):1620.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U et al.. 2016. Thousands of microbial genomes shed light on

interconnected biogeochemical processes in an aquifer system. Nature Communications. 7:13219.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)`.

Antoniades D, Crawley C, Douglas MS, Pienitz R, Andersen D, Doran PT, Hawes I, Pollard W, Vincent WF. 2007. Abrupt environmental change in Canada's northernmost lake inferred from fossil diatom and pigment stratigraphy. Geophysical Research Letters. 34(18):L18708.

Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 31(17):2882–2884.

Bliss A, Hock R, Radić V. 2014. Global response of glacier runoff to twenty-first century climate change. Journal of Geophysical Research: Earth Surface. 119(4):717–730.

Blois JL, Williams JW, Fitzpatrick MC, Jackson ST, Ferrier S. 2013. Space can substitute for time in predicting climate-change effects on biodiversity. Proceedings of the National Academy of Sciences. 110(23):9374–9379.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. Nature. 523(7559):208.

Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 11(12):2639.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15):1972–1973.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M et al.. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal. 6(8):1621.

Carnevali PBM, Herbold CW, Hand KP, Priscu JC, Murray AE. 2018. Distinct microbial assemblage structure and archaeal diversity in sediments of Arctic thermokarst lakes differing in methane sources. Frontiers in Microbiology. 9:1192.

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C et al.. 2007. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research. 36(suppl_1):D623–D631.

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM et al.. 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Current Biology. 25(6):690–701.

Castro HF, Classen AT, Austin EE, Norby RJ, Schadt CW. 2010. Soil microbial community responses to multiple experimental climate change drivers. Applied and Environmental Microbiology. 76(4):999–1007.

Cavalier-Smith T, Brasier M, Embley TM. 2006. Introduction: how and when did microbes change the world? Philosophical Transactions of the Royal Society of London B: Biological Sciences. 361(1470):845–850.

Charif D, Lobry J. 2007. SeqinR 1.0.2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. New York: Springer Verlag. ISBN : 978-3-540-35305-8.

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. 2013. Effects of gc bias in next-generation-sequencing data on *de novo* genome assembly. PLoS One. 8(4):e62856.

Comte J, Culley AI, Lovejoy C, Vincent WF. 2018. Microbial connectivity and sorting in a High Arctic watershed. The ISME Journal. 12:2988–3000.

Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology. 12(4):263.

Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. Science. 312(5781):1795–1798.

Daniel R. 2005. The metagenomics of soil. Nature reviews microbiology. 3(6):470.

de Jong A, Zandt MH, Meisel OH, Jetten M, Dean JF, Rasigraf O, Welte CU. 2018. Increases in temperature and nutrient availability positively affect methane-cycling microorganisms in Arctic thermokarst lake sediments. Environmental Microbiology. 20(12):4314–4327.

Delgado-Baquerizo M, Giaramida L, Reich PB, Khachane AN, Hamonts K, Edwards C, Lawton LA, Singh BK. 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. Journal of Ecology. 104(4):936–946.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. 72(7):5069–5072.

Dombrowski N, Teske AP, Baker BJ. 2018. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. Nature Communications. 9(1):4999.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32(5):1792–1797.

Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics. 1:5.

Eisenhauer N, Scheu S, Jousset A. 2012. Bacterial diversity stabilizes community productivity. PLoS One. 7(3):e34517.

Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. Nature Microbiology. 1(4):15032.

Emmerton CA, Louis VLS, Lehnherr I, Graydon JA, Kirk JL, Rondeau KJ. 2016. The importance of freshwater systems to the net atmospheric exchange of carbon dioxide and methane with a rapidly changing high Arctic watershed. Biogeosciences. 13(20):5849–5863.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvio: an advanced analysis and visualization platform for omics data. PeerJ. 3:1319.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science. 320(5879):1034–1039.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A et al.. 2015. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 44(D1):D279–D285.

Frey KE, McClelland JW. 2009. Impacts of permafrost degradation on Arctic river biogeochemistry. Hydrological Processes: An International Journal. 23(1):169–182.

Fuhrman JA. 2009. Microbial community structure and its functional implications. Nature. 459(7244):193.

Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 31(22):3718–3720.

Garretto A, Hatzopoulos T, Putonti C. 2019. virmine: automated detection of viral sequences from complex metagenomic samples. PeerJ. 7:e6695.

Glaeser SP, Kämpfer P. 2015. Multilocus sequence analysis (mlsa) in prokaryotic taxonomy. Systematic and Applied Microbiology. 38(4):237–245.

Glockner FO, Teeling H. 2012. Current opportunities and challenges in microbial metagenome analysis: a bioinformatic perspective. Briefings in Bioinformatics. 13(6):728–742.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Research. 31(1):371–373.

Heiri O, Lotter AF, Lemcke G. 2001. Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results. Journal of Paleolimnology. 25(1):101–110.

Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proceedings of the National Academy of Sciences. 111(13):4904–4909.

Hug LA et al.. 2018. It takes a village: Microbial communities thrive through interactions and metabolic handoffs. MSystems. 3(2):e00152–17.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K et al.. 2016. A new view of the tree of life. Nature Microbiology. 1(5):16048.

Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. 2013. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome. 1(1):22.

Hugenholtz P, Goebel BM, Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology. 180(18):4765–4774.

Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB et al.. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 521(7551):208.

Hutchins DA, Fu F. 2017. Microorganisms and ocean global change. Nature Microbiology. 2(6):17058.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11(1):119.

IPCC. 2013. Summary for Policymakers, Book section SPM, p. 130. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et al.. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology. 36(4):338.

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of Clinical Microbiology. 45(9):2761–2764.

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN et al.. 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research. 45(D1):D566–D573.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al.. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236–1240.

Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL et al.. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. Frontiers in Microbiology. 7:459.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2015. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 44(D1):D457–D462.

Keatley BE, Douglas MS, Smol JP. 2007. Limnological characteristics of a High Arctic oasis and comparisons across northern Ellesmere Island. Arctic.:294–308.

Kerr JT, Pindar A, Galpern P, Packer L, Potts SG, Roberts SM, Rasmont P, Schweiger O, Colla SR, Richardson LL et al.. 2015. Climate change impacts on bumblebees converge across continents. Science. 349(6244):177–180.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research. 41(1):e1–e1.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D et al.. 2018. Best practices for analysing microbiomes. Nature Reviews Microbiology. 16:1.

Köck G, Muir D, Yang F, Wang X, Talbot C, Gantner N, Moser D. 2012. Bathymetry and sediment geochemistry of Lake Hazen (Quttinirpaaq National Park, Ellesmere Island, Nunavut). Arctic. 65(1):56–66.

Kopec BG, Feng X, Michel FA, Posmentier ES. 2016. Influence of sea ice on Arctic precipitation. Proceedings of the National Academy of Sciences. 113(1):46–51.

Krijthe J, van der Maaten L, Krijthe MJ. 2018. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.15.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R et al.. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 31(9):814.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods. 9(4):357.

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM et al.. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. Journal of Clinical Microbiology. 50(4):1355–1361.

Laudon H, Spence C, Buttle J, Carey SK, McDonnell JJ, McNamara JP, Soulsby C, Tetzlaff D. 2017. Save northern high-latitude catchments. Nature Geoscience. 10(5):324.

Lehnherr I, Louis VLS, Sharp M, Gardner AS, Smol JP, Schiff SL, Muir DC, Mortimer CA, Michelutti N, Tarnocai C et al.. 2018. The world's largest High Arctic lake responds rapidly to climate warming. Nature Communications. 9(1):1290.

Lester RE, Close PG, Barton JL, Pope AJ, Brown SC. 2014. Predicting the likely response of data-poor ecosystems to climate change using space-for-time substitution across domains. Global Change Biology. 20(11):3471–3481.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics. 31(10):1674–1676.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25(16):2078–2079.

Linz AM, He S, Stevens SL, Anantharaman K, Rohwer RR, Malmstrom RR, Bertilsson S, McMahon KD. 2018. Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. PeerJ. 6:e6075.

Louca S, Polz MF, Mazel F, Albright MB, Huber JA, O'Connor MI, Ackermann M, Hahn AS, Srivastava DS, Crowe SA et al.. 2018. Function and functional redundancy in microbial systems. Nature Ecology & Evolution. 2:936–943.

Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences. 104(27):11436–11440.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. Constrains identifies microbial strains in metagenomic datasets. Nature Biotechnology. 33(10):1045.

Luo F, Devine CE, Edwards EA. 2016. Cultivating microbial dark matter in benzene-degrading methanogenic consortia. Environmental Microbiology. 18(9):2923–2936.

Mackelprang R, Saleska SR, Jacobsen CS, Jansson JK, Taş N. 2016. Permafrost meta-omics and climate change. Annual Review of Earth and Planetary Sciences. 44:439–462.

Mahmoudi N, Slater GF, Fulthorpe RR. 2011. Comparison of commercial DNA extraction kits for isolation and purification of bacterial and eukaryotic DNA from PAH-contaminated soils. Canadian Journal of Microbiology. 57(8):623–628.

Maidak BL, Cole JR, Lilburn TG, Parker Jr CT, Saxman PR, Stredwick JM, Garrity GM, Li B, Olsen GJ, Pramanik S et al.. 2000. The RDP (ribosomal database project) continues. Nucleic Acids Research. 28(1):173–174.

Martiny JBH, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR et al.. 2006. Microbial biogeography: putting microorganisms on the map. Nature Reviews Microbiology. 4(2):102.

Mason OU, Di Meo-Savoie CA, Van Nostrand JD, Zhou J, Fisk MR, Giovannoni SJ. 2009. Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. The ISME Journal. 3(2):231.

McArthur A. 2018. Resistance gene identifier (RGI). `https://github.com/arpcard/rgi`.

McCalley CK, Woodcroft BJ, Hodgkins SB, Wehr RA, Kim EH, Mondav R, Crill PM, Chanton JP, Rich VI, Tyson GW et al.. 2014. Methane dynamics regulated by microbial community response to permafrost thaw. Nature. 514(7523):478.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina truseq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. PLoS One. 9(9):e106689.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 8(4):e61217.

Meltofte H, Barry T, Berteaux D, Bültmann H, Christiansen JS, Cook JA, Dahlberg A, Daniëls FJ, Ehrich D, Fjeldså J et al.. 2013. Arctic Biodiversity Assessment Synthesis. Conservation of Arctic Flora and Fauna (CAFF).

Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications. 7:11257.

Michelutti N, Douglas MS, Smol JP. 2003. Diatom response to recent climatic change in a high Arctic lake (Char Lake, Cornwallis Island, Nunavut). Global and Planetary Change. 38(3):257 – 271.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research. 41(12):e121–e121.

Mohit V, Culley A, Lovejoy C, Bouchard F, Vincent WF. 2017. Hidden biofilms in a far northern lake and implications for the changing Arctic. Biofilms and Microbiomes. 3(1):17.

Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW. 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nature Communications. 8:16054.

Nash MV, Anesio AM, Barker G, Tranter M, Varliero G, Eloe-Fadrosh EA, Nielsen T, Turpin-Jelfs T, Benning LG, Sánchez-Baracaldo P. 2018. Metagenomic insights into diazotrophic communities across Arctic glacier forefields. FEMS Microbiology Ecology. 94(9):fiy114.

Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, Frostegård Å, Heulin T, Jansson JK, Jurkevitch E et al.. 2016. Back to the future of soil metagenomics. Frontiers in Microbiology. 7:73.

NRCAN. 2018. CanVec geospatial data extraction series: drainage area 10VK001. `https://maps.canada.ca/czs/index-en.html`.

Oliverio AM, Bradford MA, Fierer N. 2017. Identifying the microbial taxa that consistently respond to soil warming across time and space. Global Change Biology. 23(5):2117–2129.

O'Reilly CM, Sharma S, Gray DK, Hampton SE, Read JS, Rowley RJ, Schneider P, Lenters JD, McIntyre PB, Kraemer BM et al.. 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophysical Research Letters. 42(24):10–773.

Overpeck J, Hughen K, Hardy D, Bradley R, Case R, Douglas M, Finney B, Gajewski K, Jacoby G, Jennings A et al.. 1997. Arctic environmental change of the last four centuries. Science. 278(5341):1251–1256.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science. 276(5313):734–740.

Pacifici M, Foden WB, Visconti P, Watson JE, Butchart SH, Kovacs KM, Scheffers BR, Hole DG, Martin TG, Akcakaya HR et al.. 2015. Assessing species vulnerability to climate change. Nature Climate Change. 5(3):215.

Pan D, Nolan J, Williams KH, Robbins MJ, Weber KA. 2017. Abundance and distribution of microbial cells and viruses in an alluvial aquifer. Frontiers in Microbiology. 8:1199.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 20(2):289–290.

Parks D. 2014. CompareM. `https://github.com/dparks1134/CompareM`.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology. 36:996–1004.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 25(9):1043–1055.

Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology. 2(11):1533.

Parks JM, Johs A, Podar M, Bridou R, Hurt RA, Smith SD, Tomanicek SJ, Qian Y, Brown SD, Brandt CC et al.. 2013. The genetic basis for bacterial mercury methylation. Science. 339(6125):1332–1335.

Pecl GT, Araújo MB, Bell JD, Blanchard J, Bonebrake TC, Chen IC, Clark TD, Colwell RK, Danielsen F, Evengård B et al.. 2017. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. Science. 355(6332):1–9.

Pericard P, Dufresne Y, Couderc L, Blanquart S, Touzet H. 2017. MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. Bioinformatics. 34(4):585–591.

Poulain AJ, Aris-Brosou S, Blais JM, Brazeau M, Keller WB, Paterson AM. 2015. Microbial DNA records historical delivery of anthropogenic mercury. The ISME Journal. 9(12):2541.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2: approximately maximum-likelihood trees for large alignments. PLoS One. 5(3):e9490.

Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CM, Emerson JB, Anantharaman K, Thomas BC, Malmstrom RR, Stieglmeier M et al.. 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nature Microbiology. 3(3):328.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 41(D1):D590–D596.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology. 35(9):833.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ramette A, Tiedje JM. 2007. Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. Proceedings of the National Academy of Sciences. 104(8):2761–2766.

Ramírez GA, Graham D, D'Hondt S. 2018. Influence of commercial DNA extraction kit choice on prokaryotic community metrics in marine sediment. Limnology and Oceanography: Methods. 16(9):525–536.

Ramírez-Flandes S, González B, Ulloa O. 2019. Redox traits characterize the organization of global microbial communities. Proceedings of the National Academy of Sciences. 116(9):201817554.

Raynaud X, Nunan N. 2014. Spatial ecology of bacteria at the microscale in soil. PLoS One. 9(1):e87217.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA et al.. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 499(7459):431.

RStudio Team. 2015. RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA.

Rühland KM, Paterson AM, Smol JP. 2015. Lake diatom responses to warming: reviewing the evidence. Journal of Paleolimnology. 54(1):1–35.

Ruiz-González C, Niño-García JP, del Giorgio PA. 2015. Terrestrial origin of bacterial communities in complex boreal freshwater networks. Ecology letters. 18(11):1198–1206.

Ruuskanen MO, St. Pierre KA, St. Louis VL, Aris-Brosou S, Poulain AJ. 2018. Physico-chemical drivers of microbial community structure in sediments of Lake Hazen, Nunavut, Canada. Frontiers in Microbiology. 9:1138.

Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 4(1):8.

Screen JA, Simmonds I. 2010. The central role of diminishing sea ice in recent Arctic temperature amplification. Nature. 464(7293):1334.

Sedlar K, Kupkova K, Provaznik I. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and Structural Biotechnology Journal. 15:48–55.

Serreze MC, Francis JA. 2006. The Arctic amplification debate. Climatic change. 76(3-4):241–264.

Sharon I, Banfield JF. 2013. Genomes from metagenomics. Science. 342(6162):1057–1058.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature. 550(7676):345.

Sieber CM, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology. 3(7):836.

Smith IR. 1999. Late quaternary glacial history of Lake Hazen basin and eastern Hazen Plateau, northern Ellesmere Island, Nunavut, Canada. Canadian Journal of Earth Sciences. 36(9):1547–1565.

Smol JP, Wolfe AP, Birks HJB, Douglas MS, Jones VJ, Korhola A, Pienitz R, Rühland K, Sorvari S, Antoniades D et al.. 2005. Climate-driven regime shifts in the biological communities of Arctic lakes. Proceedings of the National Academy of Sciences. 102(12):4397–4402.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. Proceedings of the National Academy of Sciences. 103(32):12115–12120.

St. Pierre K, St. Louis V, Lehnherr I, Schiff S, Muir D, Poulain A, Smol J, Talbot C, Ma M, Findlay D, Arnott S. 2019. Contemporary limnology of the rapidly changing glacierized watershed of the world's largest High Arctic lake. Scientific Reports. 9(4447):1–15.

St. Pierre KA, St. Louis VL, Lehnherr I, Gardner A, Serbu J, Mortimer CA, Muir D, Wiklund JA, Lemire D, Szostek L et al.. 2018. Drivers of mercury cycling in the rapidly changing glacierized watershed of the High Arctic's largest lake by volume (Lake Hazen, Nunavut, Canada). Environmental science & technology. 53(3):1175–1185.

Stöckli R, Vermote E, Saleous N, Simmon R, D H. 2005. The blue marble next generation - a true color earth dataset including seasonal dynamics from MODIS. `https://visibleearth.nasa.gov/view.php?id=73751`.

Stoeva MK, Aris-Brosou S, Chételat J, Hintelmann H, Pelletier P, Poulain AJ. 2014. Microbial community structure in lake and wetland sediments from a High Arctic polar desert revealed by targeted transcriptomics. PLoS One. 9(3):e89531.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A et al.. 2015. Structure and function of the global ocean microbiome. Science. 348(6237):1261359.

Surdu CM, Duguay CR, Fernández Prieto D. 2016. Evidence of recent changes in the ice regime of lakes in the Canadian High Arctic from spaceborne satellite observations. The Cryosphere. 10(3):941–960.

Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 22(12):1540–1542.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM,

Mazumder R, Mekhedov SL, Nikolskaya AN et al.. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4(1):41.

Thaler M, Vincent WF, Lionard M, Hamilton AK, Lovejoy C. 2017. Microbial community structure and interannual change in the last epishelf lake ecosystem in the north polar region. Frontiers in Marine Science. 3:275.

Thomas T, Gilbert J, Meyer F. 2012. Metagenomics-a guide from sampling to data analysis. Microbial Informatics and Experimentation. 2(1):3.

Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. Fastviromeexplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ. 6:e4227.

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in Microbiology. 6:771.

van der Laan MJ, Pollard KS. 2003. Hybrid clustering of gene expression data with visualization and the bootstrap. Journal of Statistical Planning and Inference. 117:275–303.

van der Walt AJ, van Goethem MW, Ramond JB, Makhalanyane TP, Reva O, Cowan DA. 2017. Assembling metagenomes, one community at a time. BMC Genomics. 18(1):521.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The third revolution in sequencing technology. Trends in Genetics. 34(9):666–681.

van Kessel MA, Speth DR, Albertsen M, Nielsen PH, den Camp HJO, Kartal B, Jetten MS, Lücker S. 2015. Complete nitrification by a single microorganism. Nature. 528(7583):555.

Varin T, Lovejoy C, Jungblut AD, Vincent WF, Corbeil J. 2012. Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. Applied Environmental Microbiology. 78(2):549–559.

Vishnivetskaya TA, Layton AC, Lau MC, Chauhan A, Cheng KR, Meyers AJ, Murphy JR, Rogers AW, Saarunya GS, Williams DE et al.. 2014. Commercial DNA extraction

kits impact observed microbial community composition in permafrost samples. FEMS Microbiology Ecology. 87(1):217–230.

Warnes MGR, Bolker B, Bonebakker L, Gentleman R. 2009. gplots: Various R programming tools for plotting data. R package version.

West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Research. 28(4):569–580.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 18(5):691–699.

Williamson CE, Dodds W, Kratz TK, Palmer MA. 2008. Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. Frontiers in Ecology and the Environment. 6(5):247–254.

Williamson CE, Saros JE, Vincent WF, Smol JP. 2009. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. Limnology and Oceanography. 54(6part2):2273–2282.

Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AA, Hoelzle RD, Lamberton TO, McCalley CK, Hodgkins SB et al.. 2018. Genome-centric view of carbon processing in thawing permafrost. Nature. 560(7716):49.

Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. PLoS Computational Biology. 6(2):e1000667.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH et al.. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science. 337(6102):1661–1665.

Wrona FJ, Johansson M, Culp JM, Jenkins A, Mård J, Myers-Smith IH, Prowse TD, Vincent WF, Wookey PA. 2016. Transitions in Arctic ecosystems: Ecological implications

of a changing hydrological regime. Journal of Geophysical Research: Biogeosciences. 121(3):650–674.

Wu YW, Simmons BA, Singer SW. 2015. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 32(4):605–607.

Ye Y, Doak T. 2009. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Computational Biology. 5(8):e1000465.

Yergeau E, Bokhorst S, Kang S, Zhou J, Greer CW, Aerts R, Kowalchuk GA. 2012. Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments. The ISME Journal. 6(3):692.

Yergeau E, Michel C, Tremblay J, Niemi A, King TL, Wyglinski J, Lee K, Greer CW. 2017. Metagenomic survey of the taxonomic and functional microbial communities of seawater and sea ice from the Canadian Arctic. Scientific Reports. 7:42242.

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G et al.. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology. 29(5):415.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 8(1):28–36.

Zhou J, Bruns MA, Tiedje JM. 1996. DNA recovery from soils of diverse composition. Applied and Environmental Microbiology. 62(2):316–322.