Lake sediment microbial communities in the Anthropocene

Matti Olavi Ruuskanen

A thesis submitted in partial fulfillment of the requirements for the

Doctorate in Philosophy degree in Biology with

Specialization in Chemical and Environmental Toxicology

Department of Biology

Faculty of Science

University of Ottawa

Abstract

Since the Industrial Revolution at the end of the 18th century, anthropogenic changes in the environment have shifted from the local to the global scale. Even remote environments such as the high Arctic are vulnerable to the effects of climate change. Similarly, anthropogenic mercury (Hg) has had a global reach because of atmospheric transport and deposition far from emission point sources. Whereas some effects of climate change are visible through melting permafrost, or toxic effects of Hg at higher trophic levels, the often-invisible changes in microbial community structures and functions have received much less attention. With recent and drastic warming-related changes in Arctic watersheds, previously uncharacterized phylogenetic and functional diversity in the sediment communities might be lost forever. The main objectives of my thesis were to uncover how microbial community structure, functional potential and the evolution of mercury specific functions in lake sediments in northern latitudes (>54ºN) are affected by increasing temperatures and Hg deposition. To address these questions, I examined environmental DNA from sediment core samples and high-throughput sequencing to reconstruct the community composition, functional potential, and evolutionary responses to historical Hg loading. In my thesis I show that the microbial community in Lake Hazen (NU, Canada) sediments is structured by redox gradients and pH. Furthermore, the microbes in this phylogenetically diverse community contain genomic features which might represent adaptations to the cold and oligotrophic conditions. Finally, historical Hg pollution from anthropogenic sources has likely affected the evolution of microbial Hg resistance and this deposition can be tracked using sediment DNA on the Northern Hemisphere. My thesis underscores the importance of using culture-independent methods to reconstruct the structure, functional potential and evolution of environmental microbial communities.

Acknowledgements

First, I would like to thank both of you, Alex and Stéphane. You have been such great supervisors from the beginning and helped me grow as a scientist over these years. You have both pushed me to do better work, but also been always there for me and available when I needed assistance. While being co-supervised by two ambitious professors with sometimes diverging scientific focuses and views has not always been easy, it certainly has been very rewarding. Conducting my doctoral studies in Ottawa was overall a wonderful experience, and I am honored to have worked with you two during this time.

In Ottawa I shared my time between the Poulain and Aris-Brosou labs. I would like to thank my friends and colleagues for spending countless hours in the office and lab with me (depending on the group!), and our inspiring discussions: Dan, Jonathan, Martin, Ben, Mija, Galen, Graham, Emmanuel, Phil, Aaron, Neke, Danielle and Noémie. It has been a great pleasure and very inspiring working with you, and I hope to see you all again soon. I would also like to thank the members of my thesis committee, Jules Blais and Alex Wong for helping me polish my ideas and the original proposal for this thesis. My committee members together with Linda Bonen also made my comprehensive examination a pleasant and inspiring experience – contrary to its reputation! Furthermore, I would like to thank Jules Blais, Linda Bonen, Marina Cvetkovska and Jesse Shapiro for their thoughtful comments on this thesis that improved its quality and for the engaging discussion at my thesis defense. I feel privileged to have been able to study High Arctic environments in Canada in my thesis. Specifically, I would like to thank Vince and Kyra for this amazing opportunity and the great time we had when sampling at Lake Hazen.

At least once a year, I was also visiting back home in Finland. During these (sometimes long) visits I was fortunate enough to have a second academic home at the Virta lab at University of Helsinki in Viikki. Marko, Johanna, Christina, Kata and Windi, thank you so much for hosting me, and showing fresh perspectives and giving new ideas for my research in our discussions.

In addition to my lab mates, I also consider myself lucky to have met so many new friends in Ottawa. Special thanks to the Marginally Significant and other people in the PIZZA crew for making my studies in Ottawa thoroughly enjoyable. Maybe we can have another sittning sometime? Thanks also to my other friends from the international crowd in Ottawa for being so supportive and welcoming. A large part of my life in Canada were the Cape Breton Session and various other music groups in Ottawa. Music has always been an important part of my life and playing and singing with you really helped me to take my mind off the thesis. You have all left such a positive impression of Canada and Canadians to me.

Lastly, I would like to thank my family and friends in Finland. It has been hard living away from you for such a long time, and I have tried to make time to see you whenever I have been visiting back home. Thank you all for being so supportive over the years and believing in me. I will now make sure to see you all more often.

Ja viimeiset mutta ei vähäisimmät kiitokset. Kiitos jatkuvasta tuestanne ja ymmärryksestä, Äiti ja Isä.

Table of Contents

List of Figures

List of Tables

Abbreviations

[x]	Concentration of compound x
bp	Base pair
BCE	Before current era
CE	Current era
CI	Confidence interval
CPR	Candidate phyla radiation
ddPCR	Droplet digital polymerase chain reaction
DNA	Deoxyribonucleic acid
DNRA	Dissimilatory reduction of nitrite to ammonia
DPCoA	Double principle component analysis
dw	Dry weight
FDR	False discovery rate
FE	Fixed effect
GTR + Γ	General time reversible plus gamma distribution (model of sequence evolution)
HMM	Hidden Markov model
kbp	Kilobase pairs (1000 base pairs)
MAG	Metagenome assembled genome
MSPE	Model standard prediction error
mV	Millivolts
N_e	Effective population size
NMDS	Non-metric multidimensional scaling
OOB	Out-of-bag (error)
OTU	Operational taxonomic unit
PCA	Principal component analysis
PCR	Polymerase chain reaction
RE	Random effect
rf	Random forest
rRNA	Ribosomal ribonucleic acid

RNA	Ribonucleic acid
Tb	Terabytes
t-SNE	*t*-distributed stochastic neighbor embedding
UHP	Ultra high purity
UPGMA	Unweighted pair group method with arithmetic mean (for tree construction)
ww	Wet weight
As	Arsenic
C	Carbon
Cl	Chloride
Cs	Cesium
Fe	Iron
Hg	Mercury
N	Nitrogen
O	Oxygen
P	Phosphorus
Pb	Lead
S	Sulfur
Sb	Antimony

Chapter 1: General introduction

This thesis addresses how sediment microbial communities in high latitude (54ºN – 82ºN) lake sediments are structured and adapted to the past and present environmental conditions. As these environments are being impacted by anthropogenic action and will continue to transform in the future, it is important to understand how the microbial communities react to these changes.

1.1 Anthropocene and its effects on environmental microbial communities

Environmental impacts of human civilizations caused by, for instance, agricultural practices and land use can be tracked back thousands of years in environmental archives (Dearing et al., 2006), but after the Industrial Revolution, the scale and geographical extent of these impacts have reached new levels (Dearing et al., 2006; Stearns, 2018). The signs of increased anthropogenic activity on Earth are so widespread and distinct, that we may already have entered a new human-dominated geological epoch, the Anthropocene (Lewis and Maslin, 2015). Markers of the Anthropocene have been discovered in a variety of environmental archives. These markers include toxic metals (Amos et al., 2013), radionuclides (Kansanen et al., 1991), effects of CO_2-related acidification of the oceans (Al-Rousan et al., 2004), and compounds resistant to microbial degradation like plastics (Zalasiewicz et al., 2016). This new epoch can be generalized as a period when human actions have a dominating influence on many key processes on Earth, ranging from major biogeochemical cycles to the evolution of life (Lewis and Maslin, 2015). Furthermore, before the Industrial Revolution, most of these effects were local, but subsequently they have reached a global scale (Stearns, 2018). Even environments distant from human settlements and industrial sites have been impacted by these changes, as many contaminants are transported by air and water (Iwata et al., 1993; Thomson Reuters, 2019). Furthermore, the

increased concentration of anthropogenic greenhouse gases in the atmosphere has led to increased solar heat retention and constantly climbing average temperatures (Hegerl et al., 2007). Changes caused by this warming will have serious global consequences in both human impacted and natural ecosystems (IPCC, 2018). For example, organic carbon accumulated over millennia in permafrost in the Arctic might become again available for microbial breakdown through thawing caused by higher average temperatures (Schuur et al., 2015). This could cause releases of large quantities of greenhouse gases from increased microbial metabolism.

Technology advancement is the cause of global environmental issues, but it also enabled the in-depth study of environmental microbial communities, which play key roles in the global carbon and other nutrient cycles. However, until the past few decades their in-depth study was mostly limited to chemically measuring the rates of processes catalyzed by complex communities, and to isolate organisms able to grow in the laboratory (Amann et al., 1995). While it was first shown nearly 90 years ago that only a fraction of microbes in environmental samples are easily cultivated in pure culture (Razumov, 1932), and the phenomenon received some subsequent attention (Jannasch and Jones, 1959), a thorough study of "the great plate count anomaly" showed that the cultured organisms represent only about 1% of species in the whole community (Staley and Konopka, 1985). The study of this uncultured majority of microbes was only enabled by the development of molecular technologies (Amann et al., 1995). The most important of these have been the amplification of nucleic acids from environmental samples (through Polymerase Chain Reaction, or PCR), and various sequencing methods. Information deduced directly from these nucleic acid sequences could then be used to answer the two most important questions in the study of microbial communities: "who is there?" and "what do they do?" Combining the new molecular tools with traditional culturing methods has recently enabled

us to even culture many previously unculturable microbes (Carini, 2019; Thrash, 2019). However, even with the advent of new culturing methods and biases of their own (e.g., Suzuki and Giovannoni, 1996), molecular methods still enable easy and robust estimations of the diversity and function of environmental microbial communities (Cordier et al., 2019). More information on environmental microbes is sorely needed to assess how they respond to the ongoing changes caused by past and present anthropogenic activity.

1.1.1 Climate change

One of the most prominent global anthropogenic impacts is climate change caused by the release of greenhouse gases, mostly carbon dioxide, methane, nitrous oxide, and halocarbons (Hegerl et al., 2007). These releases have been increasing since the Industrial Revolution at the end of the 18th century (Solomon et al., 2007). The increased concentration of these gases and the related warming has caused acidification of the oceans, increased extreme weather events, melting of permafrost and glaciers, and rise of sea level (IPCC, 2018). These impacts have global reach, but the warming and its effects have been disproportionately amplified in the Arctic (Huang et al., 2017). Unfortunately, Arctic environments are also especially vulnerable to the warming, because key species in the simple trophic chains might not be able to react quickly enough to the drastic environmental changes (Bölter and Müller, 2016; Post et al., 2009). These changes can also impact microbial communities providing important ecosystem services, such as recycling of carbon and other nutrients. However, because these environments are difficult and expensive to sample, relatively little is currently known of the structure and function of the microbial communities. For example, only a handful of studies utilizing latest molecular methods exist of arctic environments (Emerson et al., 2015; Hauptmann et al., 2016; Mohit et al., 2017a; Schütte et al., 2016; Stoeva et al., 2014a; Thaler et al., 2017; Wang et al., 2016b). Furthermore, the

number of metagenomic studies on the functional diversity of arctic freshwater microbes is limited (*e.g,* Vigneron et al., 2018) and certain key environments, such as sediments, remain mostly uncharted by these methods. Hence, it would be timely to establish a baseline for the structuring and function of the communities, because they are responsible for cycling of carbon and other nutrients, which are crucial ecosystem services in lakes. Understanding the microbiology of the sediments may thus enable us to predict how the environments might react to warming-related changes, such as increased delivery of sediment and nutrients (St. Pierre et al., 2019b). As these microbes are rarely studied, they might also reveal novel phylogenetic and functional diversity and expand our knowledge on microbial adaptations to extreme environments. Anthropogenic activities have increased global temperatures, and this will likely cause microbes to adapt to new conditions. However, the same industrial processes have also increased releases of toxic metals into the environment, which might lead to changes in the structure and evolution of environmental microbial communities.

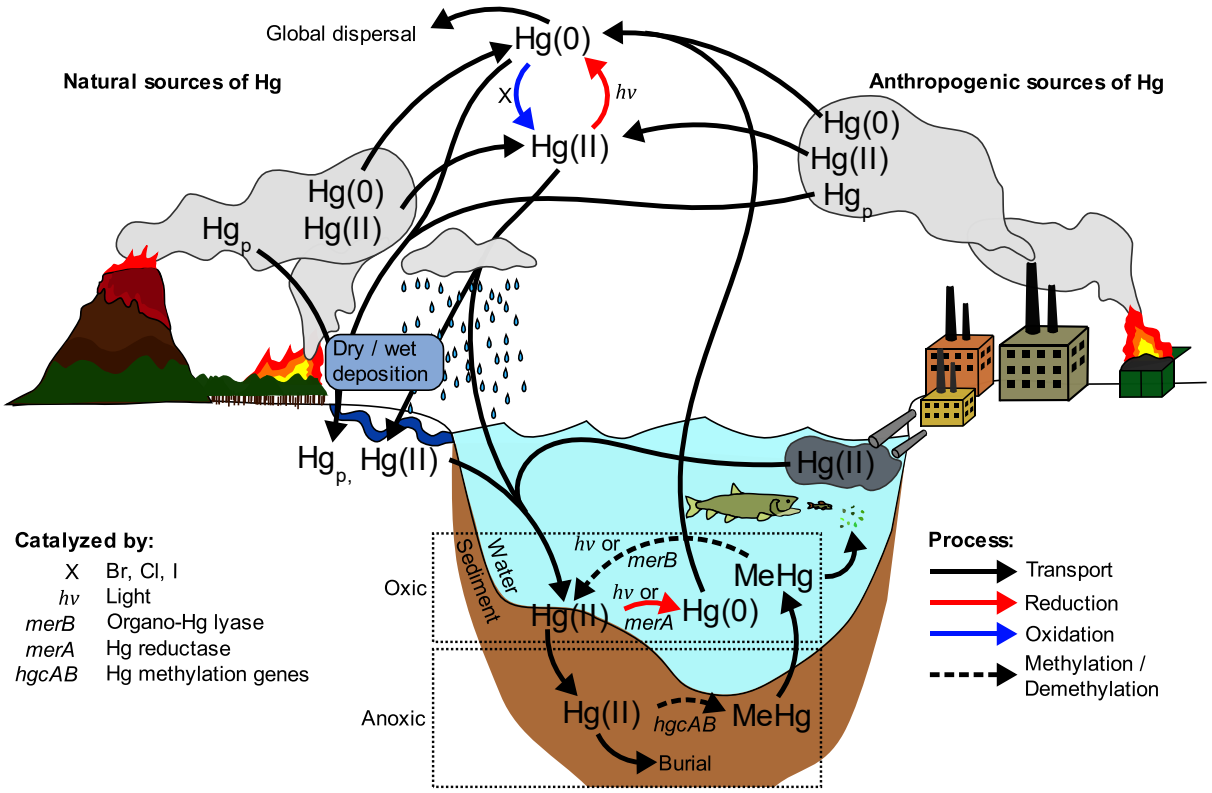

Figure 1.1: Overview of the biogeochemical cycling of mercury (Hg) with separation of the sources to natural and anthropogenic remobilization from its stable deposits. The catalyzers of processes other than physical transport are identified in the legend. Processes involving enzymes encoded by *merA*, *merB*, *hgcA* and *hgcB* occur intracellularly .

1.1.2 Anthropogenic impacts on the global mercury cycle

The remobilization of toxic metals from their natural deposits has increased since the Industrial Revolution. Mercury (Hg) is a top priority environmental contaminant, with a complex biogeochemical cycle. Organic forms of Hg can biomagnify in food webs and cause serious toxic effects, especially in animals (Driscoll et al., 2013). Hg is a naturally occurring element present in the Earth's crust and is released into the environment by geological activity and volcanism (**Figure 1.1**; see also Selin, 2009). However, remobilization of Hg from the natural deposits

through human activities has increased its amount in the global biogeochemical cycles by a factor of three to five (Selin, 2009). The concentration of volatile Hg in the atmosphere is currently 7.5 times higher than in the pre-industrial era (2000 BCE; Amos et al., 2013; Gworek et al., 2017). While mercury in particulate matter, Hg_p, and the water-soluble Hg^{2+} have residence times in the air from hours to days, the gaseous Hg^0 has a lifetime of up to one and half years in the atmosphere (Gworek et al., 2017; Schroeder and Munthe, 1998). Most of the Hg is deposited with rain and snow in the particulate and water-soluble forms, but Hg cycles between the different forms through atmospheric redox reactions (**Figure 1.1**). Thus, while most of Hg emissions are deposited close to the emission sources, they can reach even the most distant environments on the planet as Hg^0, where they are subsequently deposited as Hg^{2+}. To help curb the Hg emissions, the Minamata Convention on Mercury (UNEP, 2014), was designed to protect the environment and human health from Hg. This global treaty has introduced strict limitations to reduce the use and emissions of Hg, and after its ratification in 2017 the measures are currently being adopted by most countries in the world. However, because of re-emission of previously released Hg, these limitations are currently only preventing the present amount of Hg in the natural cycles from rising, and are only expected to lower it over a long period of time (Amos et al., 2013; Gworek et al., 2017).

Two critical parts in the Hg cycle are predominantly catalyzed by microbes. Firstly, while all forms of Hg are toxic, many anaerobic microbes (*e.g.*, sulfate and iron reducers, methanogens and fermenters; Gilmour et al., 2013) are capable of methylating Hg^{2+} to methyl mercury (MeHg), which has a higher toxicity and which bioaccumulates in food chains. The most serious effects of Hg toxicity in organisms high in the food chain, such as humans, are caused specifically by MeHg. The purpose and specific mechanisms of Hg methylation are not currently

known well. However, microbial production of MeHg requires both HgcA and HgcB proteins (Parks et al., 2013), bioavailable Hg^{2+} likely actively transported into the cell (Hsu-Kim et al., 2013), and possibly syntrophic interactions to render its production energetically favorable (Yu et al., 2018). Secondly, because Hg can be naturally present in the environment at high concentrations and is highly toxic, specific resistance mechanisms to Hg have evolved in microbes. The best known of these mechanisms is the *mer*-operon, which efficiently detects, transports and reduces various forms of Hg to the volatile Hg^0 (Barkay and Wagner-Dobler, 2005; Mathema et al., 2011a). The produced Hg^0 is chemically unreactive and evades back to the atmosphere, rendering it harmless to the organism. This resistance pathway evolved over millions of years under geogenic Hg pressures (Boyd and Barkay, 2012a), but its evolutionary trajectory since the Industrial Revolution might be controlled by anthropogenic Hg deposition (Poulain et al., 2015). This has implications on the global Hg cycling, as changes in microbial mercury cycling might impact the rates of Hg reduction and shows that human actions have changed ecosystems in numerous ways, sometimes on a global scale. Tracking these changes over time in environmental archives is important to understand how the ecosystems have reacted to perturbations in the past (Dearing et al., 2008; Zolitschka et al., 2003). Knowledge of the past effects could then be used to improve tracking and modeling of such interactions, which would improve our predictions on the future trajectories of ecosystems. One of the most studied environmental archives are freshwater lake sediments (Dearing et al., 2006). These are important sites for nutrient and toxic metal cycling, and are known to preserve a historical record of their surrounding environment in the material deposited in them over time.

1.2 Microbial ecology of lake sediments and their use as environmental archives

Lake sediments are deposits of inorganic and organic material accumulating at the bottom of lakes. This material can originate from the lake itself, from the atmospheric deposition, or it can be transported from the watershed of the lake. The sediment composition and accumulation rates depend on the local conditions, such as trophic state of the lake, topography of the lake and its watershed, climate, and vegetation cover (Cohen, 2003), which might all be also interconnected. Through this accumulation of material they also contain a record of past environmental conditions, which can be then reconstructed from chemical and biological signals (Dearing et al., 2006). Similar reconstructions are usually much more difficult from soil horizons, because of the prevalence of mixing and degradation processes such as bioturbation, freeze-thaw cycling, erosion, and leaching (Kibblewhite et al., 2015). In addition to sediments, peat profiles can preserve the deposition history of, for example, toxic metals (Allan et al., 2013) and plant DNA (Parducci et al., 2015), which can be used in reconstructions of the past state of the environment. However, compared to sediments, peatlands are often more subject to short-term climatic variation like drying and freeze-thaw cycling, which might affect their stratigraphy (Kalnina et al., 2015).

1.2.1 Physicochemical conditions in lake sediments

Because sediments often contain large amounts of organic material, and thus chemical energy, they are also perfect habitats for microbes that recycle carbon and other nutrients, which are key processes in the lake ecosystems (Nealson, 1997). Such ecosystem services provided by sediment microbes are, for example, breakdown of Carbon (C), Nitrogen (N), Phosphorus (P) and Sulphur (S) containing organic compounds into inorganic forms. Depending on the conditions in the sediment, primarily the redox potential and pH, the energetics of different

processes are favored (Thamdrup and Canfield, 2000). This has both implications on which type of microbes inhabit sediments at different sites and if nutrients are buried, released into the atmosphere, or recycled back into the lake ecosystem as soluble forms. Many of the microbial processes also influence the physicochemical conditions in the sediment. For example, bacterial metabolism with contributions from different organisms can determine the redox potential of the sediment (Hunting and Kampfraath, 2013).

Both the chemical and biological processes in the sediments have complex interactions between each other. However, most often the diagenesis of sediments tends to establish reproducible gradients in similar environments, such as freshwater lakes (Nealson, 1997). Because oxygen and deposited material containing easily degradable organic carbon are delivered to the sediment from the top, the sediments closest to the surface have the highest energy availability. Oxygen diffusing from the water phase is first used to break down organic material in the topmost sediment horizon. Below this zone anaerobic respiration occurs with electron acceptors other than oxygen, such as iron (Fe^{3+}), nitrate (NO_3^-) or sulfate (SO_4^{2-}), but they all have smaller reduction potentials and thus release less energy per oxidized molecule. The most reactive fractions of organic carbon are also consumed at the top of the sediment and its reactivity decreases with depth (Katsev and Crowe, 2015). Thus, deep in the sediment both labile carbon and electron acceptors except for carbon dioxide are depleted, and very slow growing methanogenic organisms usually dominate. These patterns in microbial metabolism lead to the establishment of a redox gradient with higher redox potential at the top and lower redox potential deeper in the sediment. Also, many chemical gradients are established in the sediment through microbial metabolism. Such gradients include (among others), oxygen, nitrate, nitrite, sulfate, manganese, iron, ammonium, and methane.

1.2.2 Lake sediments as environmental archives

The constant delivery and accumulation of material into lake sediments and conditions within limiting the breakdown of organic material enable reconstruction of the history of the lake, its watershed and even the atmosphere by analyzing the deposited material (*e.g.*, Tyson, 2012). In addition to inhibiting the breakdown of organic compounds, the absence of oxygen below the topmost sediment horizons limits the burrowing of respiring animals into the deeper layers. These animals might otherwise disrupt the chronology of the sediment by mixing the deposited layers (Krantzberg, 1985). Thus, the deposition history of inorganic compounds, such as metals, can often be preserved in a stable state in the sediments (Percival and Outridge, 2013). Understanding long-term change in the environment and predicting the possible future changes requires knowledge of such historical changes. Furthermore, as data collected by scientific studies only rarely extends more than 100 years in the past and might have methodological limitations or biases, environmental archives such as sediments are instrumental in investigating the Earth's history. Sediments have been used extensively to document environmental changes in both the lakes themselves, their watersheds, and the atmosphere. For example, pollen from lake cores has been used to reconstruct local paleoclimates over millennial time-periods (Liu et al., 2002; Magny et al., 2001), preserved plankton assemblages have been used to observe changes caused by climate warming (Rühland et al., 2003, 2014), and a wide collection of paleo-proxies have been used for reconstructions of the temporal human and climate dynamics of alpine (Perga et al., 2015) and arctic lakes (Lehnherr et al., 2018).

Study of lake sediments is important to understand the past of these ecosystems and the Earth and use this information to predict how they might react to environmental change. Global anthropogenic changes are ongoing, and their effects might irreversibly change lakes and their

ecosystems even in remote environments. Because these environments are also difficult and expensive to study, many of their features that could be addressed only with modern methods, such as high-throughput sequencing, have remained this far uncharted.

1.3 High-throughput sequencing in the study of microbial communities

The development of methods to directly investigate the genetic material of organisms have ushered in the genomic era of biology (*e.g.*, Hugenholtz, 2002). Major breakthroughs that now enable sequencing of genetic material started from understanding the role (Avery et al., 1944) and structure of DNA (Watson and Crick, 1953), its replication (Meselson and Stahl, 1958), and the nature of the genetic code (Crick et al., 1961; Nirenberg and Matthaei, 1961). Once methods to sequence the nucleotide molecules were developed (Jou et al., 1972; Sanger et al., 1977; Wu, 1972), the genetic information and its relationship to phenotypic features could be studied in-depth. Among other major findings, the sequencing of the 16S small subunit ribosomal RNA gene enabled the separation of prokaryotes to Bacteria and Archaea, resulting in the currently known three domains of life (Woese and Fox, 1977). For the first few decades, chain-termination methods based on Sanger's work prevailed. However, 'Sanger sequencing' can only handle one (up to 900 bp) DNA fragment at a time and requires a relatively large amount of template material. Fortunately, the development of molecular cloning (Cohen et al., 1973) and the PCR (Scharf et al., 1986) enabled the amplification of DNA outside of the original host organism. Finally, the major efforts to sequence the human genome (International Human Genome Sequencing Consortium, 2004; Lander et al., 2001; Venter et al., 2001) and genomes other model organisms (*e.g.*, Mouse Genome Sequencing Consortium, 2002) utilized Sanger sequencing, and revealed the urgent need for cheaper and faster sequencing methods (Schloss, 2008). A total of USD 70 million in grants from the National Human Genome Research Institute

awarded in 2004 towards this development goal led to a surge in novel high-throughput sequencing technologies.

1.3.1 Current sequencing technologies

Most of the new platforms utilize clonal amplification of the template, followed by rounds of sequencing short regions in a massively parallel fashion (Morey et al., 2013). The Illumina chemistry based on sequencing by synthesis on a solid surface (Bentley et al., 2008; Guo et al., 2008) was eventually dominant among these new platforms (Shendure et al., 2017). However, due to the unique features such as cost, read length and error rate of each chemistry, also other methods have found success. For example, instruments based on the Roche 454 chemistry (Margulies et al., 2005) are faster and have longer read lengths, but they are also more expensive and have higher error rates than comparable Illumina-based sequencers (Liu et al., 2012). Indeed, the downsides of most high-throughput sequencing instruments have been the short read lengths (50 bp with some sequencers) and high error rates (~ 0.1%), which have made the computational assembly of longer contiguous pieces of DNA difficult (Salzberg et al., 2012). New sequencing methods with novel chemistries have thus emerged to address these issues, such as Single Molecule Real-Time sequencing from Pacific Biosciences (Eid et al., 2009) and nanopore sequencing from Oxford Biosciences (Clarke et al., 2009), which can produce very long reads (> 20 kbp), but also have high error rates (> 13%; Ardui et al., 2018). Despite these shortcomings, sequencers with chemistries similar to these are slowly reaching maturity (Ardui et al., 2018), and the utilization of long-read sequencers will only increase in the future. However, the shortcomings of high-throughput sequencing are still relatively minor and tolerated because of the major benefits of these technologies. The amount of sequence data produced is massive compared to the earlier methods (as the name ‘high-throughput sequencing’ implies),

and thus, the cost of sequencing has been greatly reduced. Sequencing a megabase of raw DNA cost USD 1000 in 2004, and USD 0.01 by 2017, a decrease of 5 orders of magnitude (Wetterstrand, 2018). High-throughput sequencing has completely revolutionized many fields of biology, among them microbiology.

1.3.2 Use of high-throughput sequencing in microbiology

The use of high-throughput sequencing methods has greatly increased our knowledge on the distribution, diversity and metabolism of microbes. This increase has been very rapid. For example, the Integrated Microbial Genomes database (https://img.jgi.doe.gov/), collects data from the U.S. Department of Energy's microbial sequencing projects conducted at the Joint Genome Institute, and also accepts outside contributions. The database contained 865 bacterial and 53 archaeal genomes in 2008 and expanded to more than 70,000 bacterial and 1,800 archaeal genomes by version 5.0 in January 2019 (Chen et al., 2019). The newer and cheaper sequencing methods have expanded the scope of taxonomic marker gene studies where they are amplified by PCR. At the time when I started this thesis, the most common way to study the composition of any microbial community was by amplifying and sequencing ribosomal 16S rRNA gene fragments (Woese and Fox, 1977) directly from DNA extracts (Stahl et al., 1984), and comparing the sequences to vast online databases (Quast et al., 2013). This has enabled the characterization of microbial diversity in most environments on Earth (Thompson et al., 2017), even while a large portion of the microbes have never been studied in the laboratory. High-throughput sequencing has also facilitated the development of new fields of biology, such as metagenomics (see *e.g.*, Quince et al., 2017), where all the genetic material in a sample is extracted and sequenced without attempting to isolate or culture the microbes (**Figure 1.2**). Assembling together the short sequences into genomes has illuminated the identity of previously

unknown microbes and their metabolic potential. Recently, these 'shotgun metagenomic' techniques enabled the discovery of the Candidate Phyla Radiation (CPR) group, comprising over 15% of phyla in the bacterial domain (Brown et al., 2015; Hug et al., 2016). Likely this group had evaded previous detection, because the PCR primers used in amplicon-based studies did not match their 16S rRNA genes.

High-throughput sequencing has revolutionized the study of microbial ecology, but at the same time the amount of data produced by these methods is unprecedented. For example, the latest generation of Illumina instruments (NovaSeq 6000) can produce up to 6 terabase pairs, or 20 billion reads, of raw sequence per run. Even MiSeq, often used for amplicon studies, produces up to 15 gigabase pairs or 25 million reads. The computational capacity needed to process this deluge of data and the complexity of the results have also increased greatly, which has led to an increased demand of bioinformatic skills and knowledge in microbial ecology.

1.4 Bioinformatics in microbial ecology

The processing of high-throughput data, at least for metagenomes, currently necessitates the use of high-performance computing. For example, state-of-the-art shotgun metagenome assemblers can require up to 1 Tb of memory depending on the complexity of the data (van der Walt et al., 2017), and personal computers are usually configured only with a fraction of this amount. The access to high-performance computing can be expensive – with the cheaper sequencing, access to computational resources has often become the limiting factor for the studies. Fortunately also open-access servers running on donated resources are available, such as MG-RAST (Meyer et al., 2008), which performs a basic phylogenetic and functional analysis from raw metagenomic data. However, the queue times to process submitted data can be long and the platform does not allow customization of pipelines.

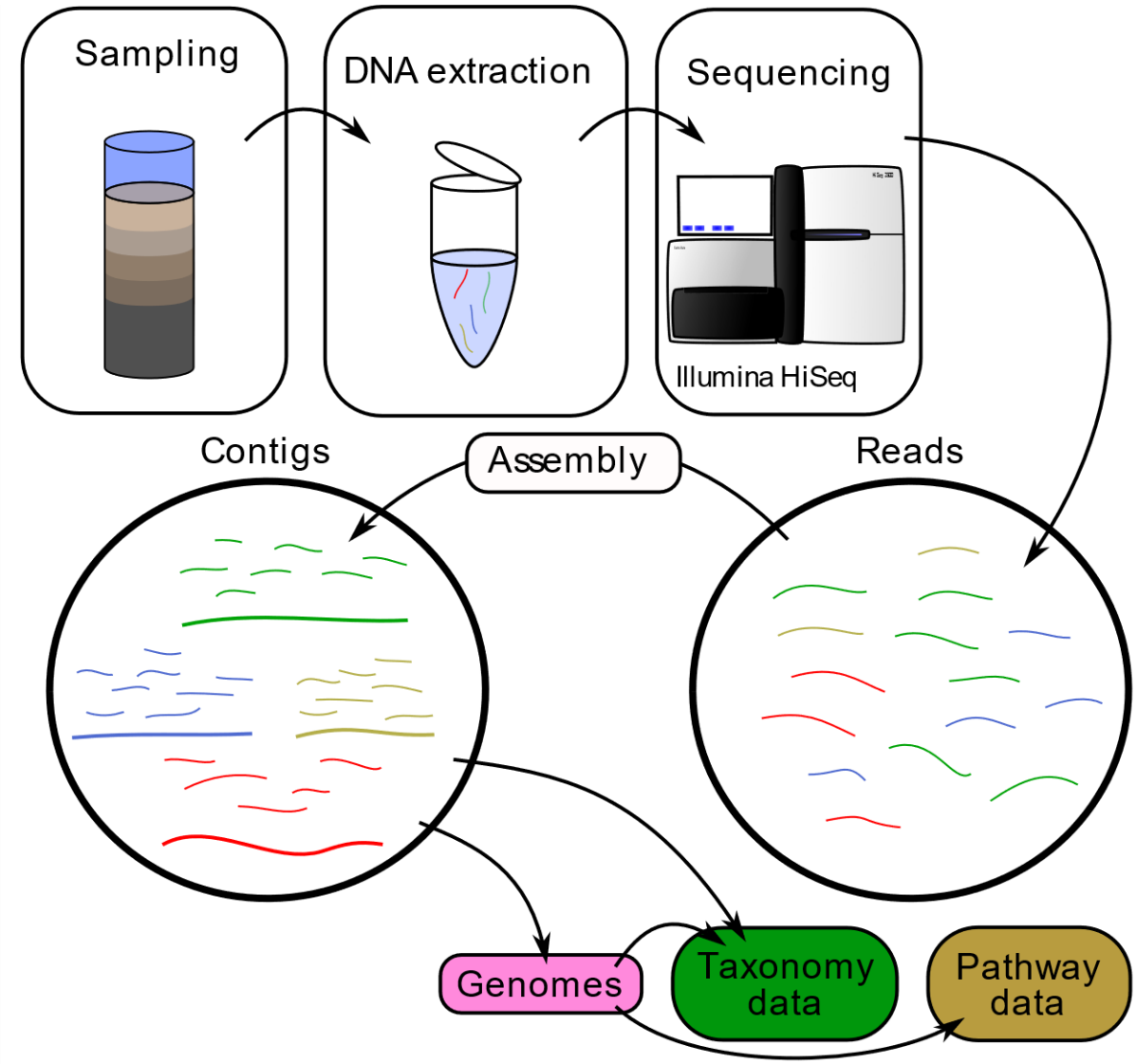

Figure 1.2: Example workflow for a shotgun metagenomic study of sediment microbial communities. Assembly of short reads into contigs enables the reconstruction of representative draft genomes for (closely related) organisms whose DNA is found in the sample in sufficient quantity. Metabolic capabilities of these organisms can be predicted based on the identification of specific functional genes in the reconstructed genomes.

Another problem faced by metagenomic studies is the lack of standardization in the processing of data and reporting of results. Some work has been done to gather best practices in both amplicon (Murray et al., 2015) and metagenomic (ten Hoopen et al., 2017) sequencing. These best practices include, for example, recommendations for sample handling and screening, the use of experimental controls, proper description of important steps in the pipelines, and how to archive and publish data. Guides also exist for the use of appropriate multivariate methods in the analysis of microbial ecology data, depending on the specific research question (Buttigieg and Ramette, 2014). Fortunately, the use of established and popular pipelines such as QIIME (Caporaso et al., 2010) or Mothur (Schloss et al., 2009) for 16S rRNA genes and Anvi'o (Eren et al., 2015) or MG-RAST for metagenomes also facilitates comparisons between different studies. The massive amount of methods available for the processing of amplicon and metagenomic high-throughput sequencing data is both a blessing and a curse for microbial ecology – this enables flexibility in the construction of analysis pipelines, but also makes cross-comparisons among studies often difficult.

Shotgun metagenomics have recently gained popularity over 16S rRNA gene amplicon-based sequencing even in routine description of microbial communities, as more powerful sequencers can now produce a more complete view of the microbial community compared to amplicon studies (Ranjan et al., 2016). However, amplicon-based sequencing still has several benefits over metagenomes: (i) it remains a cheaper alternative for simple analysis of microbial community composition, (ii) processing of data does not require the use of high-performance computing environments, (iii) the coverage of shotgun metagenomes can be lacking, especially in high-diversity environments or for rare functional target genes. Regardless of the sequencing

strategy chosen, the data analysis often proceeds similarly in both cases to investigate patterns of phylogenetic and genetic diversity, and how the prevailing environmental conditions affect them.

1.4.1 Analysis of microbial ecology data

The environmental microbial communities and their interactions with the environment are highly complex. This complexity is reflected in the data obtained both with amplicon and shotgun metagenomic methods. For example, after reducing the variability of the unique amplicon sequences by clustering them into groups representative of a single taxon (*e.g.*, Mahé et al., 2015), a single sample might still contain thousands of these groups. More recently, analyzing exact sequence variants has been shown to have several benefits over variant clusters, among these, an improved comparability across studies (Callahan et al., 2017). However, this approach increases the number of variants to be analyzed even further. Microbial ecology studies also often include metadata of the samples, such as differences in geographical location, pH, redox potential, or other measurements of sample chemistry. This necessitates rigorous design of sampling to avoid bias and accurately address the specific research question (Buttigieg and Ramette, 2014). Multivariate methods, such as reducing the dimensionality of the data by ordination, are necessary beyond 'simple' descriptive analyses (Buttigieg and Ramette, 2014).

Recent developments in computational sciences, namely machine learning algorithms in unsupervised clustering (van der Maaten and Hinton, 2008), ensemble methods (Rokach, 2010) and neural networks (Özesmi et al., 2006), have shown great potential in microbial ecology (*e.g.*, Cordier et al., 2017; Larsen et al., 2012; Pasolli et al., 2016). These methods have many desirable characteristics for handling data from microbial ecology studies. Most of the newer methods are nonparametric, robust, fast and easy to implement, and most importantly they have most often had better performance compared to traditional statistical tools. The major downside of these

methods is that their inner logic might not be available for inspection, and their performance can be evaluated only by analyzing how their output changes in response to the input data. In microbial ecology these methods could greatly facilitate exploratory analyses or studies where the relationships between the response and explanatory variables are unknown. Also, predictive machine learning models could be used to reconstruct phenotypic traits of sequenced microbes (Long et al., 2019), perhaps some day even directly from metagenomic data. Despite their demonstrated performance and outstanding results, machine learning methods are not yet utilized to their full potential in the field of microbial ecology.

1.5 Rationale and objectives for the thesis

This thesis was conceived to shed light on how environmental change during the Anthropocene affects microbial communities in northern freshwater sediments. These communities provide important ecosystem services, such as recycling of carbon, nitrogen, and toxic metals. These elemental cycles are likely strongly affected by the effects of anthropogenic activities on the sediment communities (see *e.g.*, Gerbersdorf et al., 2011; Rudman et al., 2017). To predict how the communities might change in the future, we must understand their current and past states. This thesis focuses on the analysis of northern lakes, because they are both highly vulnerable to the effects of climate change, and because many of them have been strongly impacted by mercury deposition. Analysis of sediment-extracted DNA is used to reconstruct the phylogenetic and metabolic diversity, and a picture of the structural, metabolic, and evolutionary adaptations of the sediment communities to physicochemical constraints. In this thesis, these insights are then utilized to predict effects of environmental change to microbial communities (through analyzing their biogeochemical constraints) in the vulnerable Arctic, and to understand the effects of anthropogenic mercury deposition on the Northern Hemisphere.

The research chapters of this thesis consist of two themes, which are linked through the DNA-based reconstruction of the present and past state of the microbial communities in northern lake sediments. The first research chapter (**Chapters 2)** outlines how the current physicochemical constraints drive the structuring of the sediment microbial communities in Lake Hazen in the Canadian high Arctic. The second research chapter (**Chapter 3**) further describes the diversity of these sediment microbes, and how their genomes are equipped to survive in the cold and oligotrophic conditions. The third research chapter (**Chapter 4**) details how the historical deposition of anthropogenic mercury has affected the microbial communities in lake sediments in the Northern Hemisphere. The final chapter (**Chapter 5**) synthesizes the research contributions of my thesis and gives suggestions for future research directions.

This thesis sheds light on the uncharted diversity of microbes in lake sediments from northern latitudes and their evolutionary response to increasing toxic metal deposition during the Anthropocene. The results provide the baseline on microbial community assembly and functional potential in arctic lake sediments and establish the *merA* gene as a highly sensitive biomarker of historical mercury bioavailability in sediment archives. These results outcomes will be valuable for tracking, understanding, and predicting the effects of anthropogenic change in the environment.

The objective of **Chapter 2** was to provide a baseline of microbial community diversity in the sediments of Lake Hazen in the Canadian High Arctic, and how it is constrained by the prevailing biogeochemical conditions. This study was based on amplifying the small subunit 16S rRNA gene from sediment DNA and assigning taxonomy to similar groups of sequences. The region is currently experiencing climate change related warming, which profoundly impacts these biogeochemical constraints (Lehnherr et al., 2018; St. Pierre et al., 2019b), and we

hypothesized that the ecologically important sediment microbial communities might be severely affected. Furthermore, quite little is still known of the diversity of arctic lake sediments, and the study is among the first few addressing the microbial community structure in these environments (Mohit et al., 2017a; Stoeva et al., 2014a; Thaler et al., 2017; Wang et al., 2016b). This chapter also presented evidence, based on functional mapping of the taxonomy, that the sediment communities in different parts of Lake Hazen might have similar function despite being phylogenetically dissimilar. However, these functional predictions were based solely on the 16S rRNA gene -derived taxonomy of the organisms, and this prompted us to further investigate their functional diversity by shotgun metagenomic sequencing.

The sediment communities in Lake Hazen were also studied in **Chapter 3**, where the scope was extended to analyzing their functional potential with shotgun metagenomics to characterize metabolic adaptations to the prevailing conditions, and to identify the organisms likely catalyzing important nutrient cycles in the lake sediments. Because metagenomic data on arctic freshwater sediments is currently highly scarce (Vigneron et al., 2018), we also aimed to characterize the diversity of recently discovered branches of microbial life in this environment, such as genomes from the Candidate Phyla Radiation. The results of this chapter showed that many of the obtained genomes were from understudied phyla and had a higher prevalence of pathways likely related to energy conservation and membrane chemistry, which might be important adaptations to the prevalent conditions in Lake Hazen sediments. This result also indicates that changes in the sediment microbial community and the microbially catalyzed nutrient cycles are likely in the light of recent climate warming-related changes in the Lake Hazen watershed (Lehnherr et al., 2018; St. Pierre et al., 2019b).

In **Chapter 4** the objective was to expand both geographically and methodologically on the preliminary results of the effect of anthropogenic mercury deposition on the evolution of the microbial mercury reductase *merA* (Poulain et al., 2015). This seminal study by Poulain et al., showed an increase in the effective population size (demographic history) of the *merA* gene (amplified and sequenced from sediment-extracted DNA) coinciding with increased mercury deposition since the Industrial Revolution at the end of the 18th century. The current study was designed to replicate the previous results in other dated freshwater sediment archives on a broad geographic scale and to observe differences between pristine and highly impacted sites. Furthermore, the new study employed high-throughput sequencing, which was likely to capture a higher diversity of the *merA* gene than the previous approach. The results of this study showed that sites located up to 5,500 km apart on the Northern Hemisphere, on average, displayed an increase in the effective population size of *merA* regardless of the level of mercury deposition. All results of the previous study were also replicated with the updated methodology. The signal observed in the sequences of the *merA* gene is likely also sensitive only to the bioavailable fraction of deposited mercury. In this case, it could be used in further studies to track the historical bioavailability of mercury in environmental archives.

Chapter 2: Physicochemical drivers of microbial community structure in sediments of Lake Hazen, Nunavut, Canada

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

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

[2]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada.

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

This article was originally published as: **Ruuskanen, M. O.**, St. Pierre, K. A., St. Louis, V. L., Aris-Brosou, S., and Poulain, A. J. (2018). Physicochemical Drivers of Microbial Community Structure in Sediments of Lake Hazen, Nunavut, Canada. *Front. Microbiol.* 9:1138. doi:10.3389/fmicb.2018.01138.

Author contributions: MOR designed the experiments together with AJP and VStL. MOR, KStP and VstL performed the sampling. KStP and VstL measured the physicochemical data. MOR performed DNA extraction and quality control, analyzed the data and drafted the manuscript. AJP and SAB supervised the project. All authors contributed to the writing and accepted the final version of the manuscript.

2.1 Abstract

The Arctic is undergoing rapid environmental change, potentially affecting the physicochemical constraints of microbial communities that play a large role in both carbon and nutrient cycling in lacustrine environments. However, the microbial communities in such Arctic environments have seldom been studied, and the drivers of their composition are poorly characterized. To address these gaps, we surveyed the biologically active surface sediments in Lake Hazen, the largest lake by volume north of the Arctic Circle, and a small lake and shoreline pond in its watershed. High-throughput amplicon sequencing of the 16S rRNA gene uncovered a community dominated by Proteobacteria, Bacteroidetes, and Chloroflexi, similar to those found in other cold and oligotrophic lake sediments. We also show that the microbial community structure in this Arctic polar desert is shaped by pH and redox gradients. This study lays the groundwork for predicting how sediment microbial communities in the Arctic could respond as climate change proceeds to alter their physicochemical constraints.

2.2 Introduction

While human-induced climate change is a global reality, its effects are amplified in the Arctic, severely impacting freshwater ecosystems there. Indeed, increases in air temperature and precipitation lead to enhanced glacial melt and runoff (Bliss et al., 2014), permafrost thaw (Mueller et al., 2009), and a reduction in ice-cover duration (Vincent and Laybourn-Parry, 2008). In response to these changes, High Arctic lakes can undergo shifts in their temperature, light and nutrient availability, pH, and salinity (Lehnherr et al., 2018; Mueller et al., 2009). Changes in these abiotic factors can be expected to influence the structure of microbial communities which, in turn, can then affect their physicochemical environment, for example through nitrogen fixation, organic carbon mineralization, or sulfate reduction. However, the microbial communities inhabiting polar lake sediments are still poorly characterized, and what drives community composition is relatively unknown. Although over the past few years several studies taking place in the polar regions have used next-generation sequencing to characterize microbial communities (Emerson et al., 2015; Hauptmann et al., 2016; Mohit et al., 2017b; Schütte et al., 2016; Stoeva et al., 2014b; Thaler et al., 2017; Wang et al., 2016b), data on sediment microbial communities in these environments is still sparse. The available data are also biased towards small lakes and thaw ponds, thus underrepresenting large arctic lakes.

To predict how environmental changes might impact future freshwater quality and productivity in the Arctic, we first need to understand the structure of the microbial communities that are mediating the biogeochemical cycles in these environments. This is usually achieved by PCR amplicon sequencing of the 16S rRNA gene, which is commonly used as a phylogenetic marker gene for bacteria and archaea. To move beyond the structural description of a microbial community, we need to understand (i) how the environment is shaping a community, and (ii)

how a community, in turn, shapes its environment. Metrics describing microbial community structure can be correlated with physicochemical variables using multivariate methods, such as Non-metric MultiDimensional Scaling (NMDS), (un-)constrained correspondence analysis, or cluster analysis (Buttigieg and Ramette, 2014). However, most of these approaches remain descriptive, and assume that the relationships between community composition and abiotic factors are linear. To address these limitations, machine learning methods have been used, for instance to predict disease progression from human gut microbiomes (Pasolli et al., 2016), to determine the factors affecting microbial diversity in soil (Ge et al., 2008), or to show that pH controls microbial diversity in acid mine drainage (Kuang et al., 2013). More recently, Beall et al., (2016) identified Operational Taxonomic Units (OTUs) with different abundances between high and low ice conditions in lakes, while Sun et al., (2017) predicted that very low levels of antimony [Sb(V)] and arsenic [As(V)] increase microbial diversity in soils. However, such machine learning approaches have not yet been used to characterize the drivers of microbial diversity in Arctic freshwater environments. Without a full metagenomics or metatranscriptomics dataset, it is difficult to properly describe a functional link between community structure and function. When such data are unavailable, studies have suggested that amplicon-based sequencing data be used to make limited functional predictions of environmental microbial communities (Louca et al., 2016b). This type of functional prediction relies on the presence of taxa known to participate in well characterized biological processes or functions (*e.g.*, oxygenic photoautotrophy, sulfate reduction, and methanogenesis; Aßhauer et al., 2015; Langille et al., 2013; Louca et al., 2016b) but has yet to be applied to undersampled and / or extreme environments such as high arctic lake sediments.

Here, we characterized over a period of two years the microbial community structure in sediments collected from freshwater systems in the Lake Hazen watershed, located in Quttinirpaaq National Park on northern Ellesmere Island, in Nunavut, Canada (82°N, 71°W; **Figure A1**). Bacterial and archaeal 16S rRNA gene amplicon sequencing from environmental DNA samples allowed us to characterize the microbial communities across space and time. Taking advantage of recent developments in machine learning, we determined the physicochemical drivers of the community structures, and use functional mapping of the community structure (Louca et al., 2016b) to make predictions about the sediment microbial communities.

2.3 Material and methods

2.3.1 Collection of sediment cores and associated chemistry

The Lake Hazen watershed is a polar oasis with temperatures higher than usually found at similar latitudes (Keatley et al., 2007) due to the influence of the Grant Land mountains in the northwest. Sediment cores were collected from three water bodies within the watershed: Lake Hazen itself, Pond1, and Skeleton Lake (**Figure A1**). Lake Hazen (74 km long, up to 12 km wide, area 54,200 ha, max. depth of 267 m; **Figure A2a**) is the world's largest lake by volume north of the Arctic Circle. It is primarily fed by runoff from the outlet glaciers of the Grant Land Ice Cap and drained by the Ruggles River to the northeastern coast of Ellesmere Island. Lake Hazen has a relatively stable year-round water temperature of ~3°C (Reist et al., 1995), is fully ice covered in the winter (Latifovic and Pouliot, 2007), and is ultra-oligotrophic (Keatley et al., 2007). Lake Hazen is monomictic, mixing fully in the summer partially influenced by turbidity currents originating from the glacial inflows (Lehnherr et al., 2018). A slight reverse temperature stratification (*i.e.*, lower temperatures right below the ice) develops during the winter. The

surface sediments of Lake Hazen are soft silts, with a total organic carbon content between 3.1 and 8.3%. The bathymetry and geochemistry of Lake Hazen have been thoroughly characterized in Köck et al., (2012). While large lakes like Lake Hazen are rare in the Arctic, small lakes and shallow ponds are a characteristic feature of the Arctic landscape. Skeleton Lake (1.9 ha, max. depth 4.7 m; **Figure A2b**) is fed by permafrost thaw waters, and subsequently drains through two ponds, a wetland, and a small creek before flowing into Lake Hazen (Emmerton et al., 2016). Pond1 (0.1-0.7 ha, max. depth 0.5-1.3 m; **Figures A2c**, **A2d**) is located along the northwestern shore of Lake Hazen. In high glacial runoff years, Pond1 may become hydrologically connected to Lake Hazen as water levels rise (Emmerton et al., 2016; **Figure A2d**). The organic carbon content of the sediments in Pond1 ranges from 7.0 to 10.4% and in Skeleton Lake from 13.0 to 35.1%. Skeleton Lake and Pond1 are fairly productive in the summer with photosynthesis by macrophytes, mosses, and algal mats that cover the sediments (**Figure A2c**), despite their low chlorophyll *a* concentration (Keatley et al., 2007; Lehnherr et al., 2012). Some of the productivity in Skeleton Lake and Pond1 might also be driven by carbon and nutrients originating from fecal matter of birds as both sites are important nesting habitats. In the summer, their water temperature can rise to 19°C, but in the winter, ice cover reaches to the bottom in Pond1 and shallower (< 2 m) parts of Skeleton Lake. The water columns of both Skeleton Lake and Pond1 are depleted of O_2 during the winter because of heterotrophic activity.

Short sediment cores were collected over three field expeditions: (i) in spring 2014 from two sites in Lake Hazen itself (Snowgoose Bay [depth: 44 m] and Deep Hole [258 m]; **Figures A2a**, **A3a**), (ii) in spring 2015 from three sites in Lake Hazen (off John's Island [141 m], Snowgoose Bay [50 m] and Deep Hole [261 m]) plus one site at the centre of Skeleton Lake [4 m] (**Figures A1**, **A3b**), and (iii) in summer 2015 from Pond1 [1.5 m] plus a shallow shoreline

site [0.3 m] in Skeleton Lake (**Figures A1**, **A3c**, **A3d**). In spring, all sites were covered with just less than 2 m of snow-covered ice; in summer, samples were collected during open water (ice-free) conditions. At each site, three intact replicate cores were collected for DNA extraction, and determination of physicochemical profiles and of porewater chemistry. All sediments were collected either with an UWITEC (Mondsee, Austria) gravity corer (deep sites), or manually (shallow sites in Pond1 and Skeleton Lake) into 86 mm inner diameter polyvinyl chloride core tubes. Due to logistical constraints, only a single core was available for DNA extraction from each time and site. Cores for DNA extraction were sectioned in 0.25 cm (spring 2014) or 0.50 cm (summer 2015) intervals immediately after sampling, preserved in Invitrogen™ RNAlater™ (Thermo Fisher Scientific Inc., Waltham, MA, USA), and stored at -18°C before DNA extraction. Contamination of samples was minimized by cleaning the sectioning equipment between each section and wearing non-powdered latex gloves during sample handling. In spring 2015, whole cores were frozen directly after sampling at -18°C, transported back to the University of Ottawa, and sectioned at 1 cm intervals while frozen. Surfaces of the sections in contact with the non-sterile sectioning equipment were scraped clean with bleach-sterilized tools in a laminar flow hood (HEPA 100) before subsampling from the middle of the sections. Redox potential, pH, [H_2S], and dissolved [O_2] 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 (**Tables A1**, **A2**). Redox and [H_2S] data were unavailable for summer 2015 cores because of broken microsensors. For the summer 2015 cores [NO_3^-], [Cl^-], and [SO_4^{2-}] were also measured in sediment porewaters by ion chromatography (**Table A2**). Cores used for analyses of porewater chemistry were sectioned at 1 cm intervals into 50 ml falcon tubes in the field, followed by flushing of any headspace with UHP N_2 before capping. Tubes were then

centrifuged at 4000 rpm, after which the supernatant was filtered through 0.45 µm cellulose nitrate filters into 15 ml tubes, which were then frozen until analysis at the Elemental Analysis and Stable Isotope Ratio Mass Spectrometry Laboratory (Department of Renewable Resources, University of Alberta). Concentrations for H_2S were set to 0 where it was not detected with the microsensors. For the three lowest horizons in the Skeleton Lake 2015 core, [H_2S] was input as the value measured at the deepest sediment depth before the microsensor broke (169.8 mgL^{-1}), as a conservative estimate since its oxidation in the completely anaerobic sediments was likely minimal. When several measurements were made over the sectioning depth used for DNA extraction, concentration readings were averaged. Hereafter, "sediment depth" refers to the lower sediment depth of each sample, measured down from the sediment-water interface. Principal Components Analysis (PCA) was employed to visualize physicochemical differences and relatedness between the different coring sites. For this, the autoplot function from the R package ggfortify 0.4.1 (Horikoshi and Tang, 2017) was used.

2.3.2 Sequencing and data preprocessing

Upon returning to the University of Ottawa, samples for DNA extraction were homogenized, divided into duplicate 250 mg (ww) subsamples, and washed with a buffer (10 mM EDTA, 50 mM Tris-HCl, 50 mM $Na_2HPO_4 \cdot 7H_2O$ at pH 8.0) to remove PCR inhibitors (Poulain et al., 2015; Zhou et al., 1996). DNA was extracted from the duplicate subsamples with PowerSoil® DNA Isolation Kit (MO BIO Laboratories Inc, Carlsbad, CA, USA), and then the duplicate extracts were combined. The 16S rRNA gene fragment was amplified with universal primers in the spring 2014/2015 samples, and primer sets specific to either Bacteria or Archaea in the summer 2015 samples (for details, see **Appendix A**). The extraction kit elution buffer was used as a negative control in screening experiments. Sequencing was completed with Illumina MiSeq

using paired-end 300 bp reads at Molecular Research LP (Shallowater, TX, USA; for details, see **Appendix A**). Sequencing of a single sample per sediment depth per core was deemed sufficient, since no pairwise comparisons of individual samples were conducted in the data analysis. All handling of the samples was conducted in a laminar flow hood (HEPA 100) stainless steel sterile cabinet that was treated with UVC radiation and bleach before each use.

Forward and reverse reads were paired with PEAR 0.9.10 (Zhang et al., 2014a), and libraries were split with QIIME 1.9.1 (Kuczynski et al., 2011). Chimeric sequences were removed with vsearch 2.0.0 (Rognes et al., 2016) utilizing the UCHIME (Edgar et al., 2011), against the SILVA 128 SSU Ref NR99 database (Quast et al., 2013). The reads were clustered into OTUs with Swarm 2.1.9 (Mahé et al., 2015) and singleton OTUs were removed. Counts were normalized using cumulative sum scaling with the Bioconductor package metagenomeSeq 1.18.0 (Paulson et al., 2013). Representative sequences of the OTUs were aligned to the SILVA 128 database (Quast et al., 2013), with SINA Incremental Aligner 1.3.0 (Pruesse et al., 2012). Taxonomy was assigned to the OTUs in SINA, with the Least Common Ancestor method. For phylogeny-based analyses, the alignments were trimmed with trimAl 1.2 using the heuristic "automated1" option (Capella-Gutiérrez et al., 2009) followed by visual inspection in Unipro UGENE 1.26.3 (Okonechnikov et al., 2012). Maximum likelihood phylogenetic trees were built with FastTree 2.1.9 (Price et al., 2010), using the GTR + Γ model of sequence evolution (Aris-Brosou and Rodrigue, 2012).

2.3.3 Data analyses

The number of sequences was tracked throughout each step of the pipeline for quality control (**Table A3**). The taxonomy of OTUs with > 99% sequence identity to the SILVA 128 database was refined to the closest matching entry to facilitate functional mapping. OTUs with

ambiguous, mitochondrial, or plastid assignments were removed with phyloseq 1.20.0 (McMurdie and Holmes, 2013). Negative controls were not sequenced for this study and, as such, we were not able to directly remove possible contamination brought by the DNA extraction kit. Although studies with low microbial biomass (*e.g.*, blood, lungs, dry surfaces) are expected to be more sensitive to contaminants (Glassing et al., 2016; Salter et al., 2014), we tested the impact of possible contaminants by identifying and removing putative contaminating genera from our samples (see **Appendix A**; **Figure A4**). We compared the unmodified data to analyses where we removed 100% of known contaminants from MOBIO PowerSoil DNA extraction kits (Glassing et al., 2016), the kit used in our study. The result of our comparative analyses showed (i) no changes in alpha diversity analyses, (ii) few changes in the clustering analyses (**Figure A5**) and (iii) no changes in the ordination analyses (**Figure A6**), leaving our conclusions unaffected in all cases. Note that *e.g.*, 5 of the 10 most abundant contaminants (Veillonella, Methylobacterium, Prevotella, Tumebacillus, and Oxalobacter) were not found in our samples. In addition to known contaminants from the MOBIO kit used here and described in the main text, we also tested for known contaminants from four additional DNA extraction kits (Salter et al., 2014) that were not used in our study (see **Appendix A**; **Figures A4-A6**). However, as the putative contaminant genera could plausibly be part of the sediment community and the identity of true contaminants are not known, they were not removed from the data. To visually estimate the sequencing depth in our samples, rarefaction curves were constructed from non-normalized data with singleton OTUs included (**Figure A7**). To assess the functional potential of the communities based only on 16S rRNA gene amplicon data, the normalized and curated OTU abundances were mapped to phylogenetically conserved functional groups in a customized database using FAPROTAX 1.0 (Louca et al., 2016b). Briefly, the predictions made by

FAPROTAX are based on references from the literature, and work by mapping OTUs (at any given taxonomic level) to functional groups. The associations are based solely on cultured strains, so that an association between a taxonomic level and a functional group is only made if all representatives at that taxonomic level display the particular function. The total DNA extracted from sediments does not solely represent the metabolically active part of the community, as DNA from both dormant and dead organisms is usually co-extracted (Carini et al., 2017; Klein, 2007; Lennon et al., 2017). Thus, without transcriptomic or proteomic data our functional predictions should be considered hypothetical.

To assess the biological significance of phylogenetic characterizations, samples were analyzed based on two levels of diversity: within and among samples. First, we investigated trends in alpha-diversity (within-sample diversity). Because the contribution of individual taxa to ecosystem processes is likely dependent on their abundance (Nemergut et al., 2013), we chose Simpson's dominance (Morris et al., 2014) as the metric for alpha-diversity. Simpson's dominance is robust to both spurious OTUs and variations in sampling depth between sequencing runs (Pinto and Raskin, 2012). The sequencing depth in our samples (Good's coverage: 76.0% to 97.2%; **Figure A10**) suffices to accurately estimate alpha-diversity (Lundin et al., 2012). To enable comparisons of alpha-diversity and sequencing coverage to other studies, we also calculated Chao1 and Shannon indices, and Good's coverage (**Figure A10**). All this was done based on ten randomized rarefactions of the raw OTU counts with the R package phyloseq 1.20.0 (McMurdie and Holmes, 2013). The relationships between alpha-diversity and its predictors (sample categories or physicochemical variables) were determined with random forests (Breiman, 2001; Liaw and Wiener, 2002). The forests were grown to 5000 trees, using the R package ranger 0.7.0 (Wright and Ziegler, 2017). Selection of the most important

predictors was based on the Gini index by adding predictors one by one in order of decreasing importance (Menze et al., 2009). The best and most parsimonious model was then selected by minimizing the Model Standard Prediction Error (MSPE) for regression random forests (in the case of continuous predictors), or by maximizing Cohen's Kappa for classification random forests (in the case of categorical predictors). The relationships between the most important predictors and Simpson's dominance were estimated with partial dependence plots of the best models with the R package edarf 1.1.1 (Jones and Linder, 2016), which display how model prediction changes as a function of each predictor, while other predictors are fixed to their average value. Thus, each variable's effect on the model prediction is considered independently, and each predictor's relative effect size can be estimated from the variability displayed by the model prediction.

Second, in terms of beta-diversity (between-samples diversity), phylogenetic distances between pairs of samples were calculated with a Double Principle Coordinate Analysis (DPCoA; Pavoine et al., 2004), using OTU abundances and patristic distances estimated from the maximum likelihood tree. The phylogenetic data were limited to OTUs with > 0.01% overall abundance, because of the quadratic increase in runtime per added OTU in DPCoA (Fukuyama et al., 2012). The Bray-Curtis distances between samples were calculated from group abundances in the functional predictions. A Mantel test was used to test for differences between sample physicochemistry, and either their phylogenetic or functionally predicted group distances. Phylogenetic data and functional predictions from spring 2014/2015 were then clustered using the *t*-[d]istributed Stochastic Neighbor Embedding (*t*-SNE) algorithm (van der Maaten and Hinton, 2008), implemented in the R package rtsne 0.13 (Krijthe and van der Maaten, 2017), with "perplexity" set to 5. Clusters were identified with the HDBSCAN algorithm (Campello et

al., 2013), in the package dbscan 1.1.1 (Hahsler et al., 2018), with "minPts" set to 3. Regression and classification random forests were used together with partial dependence plots to identify the most important physicochemical and categorical variables for the clustering patterns, as described above. Two different subsets of the phylogenetic data were analyzed: the most abundant OTUs (> 0.01% abundance) and the dataset limited to OTUs that matched at least one group in the FAPROTAX database. Summer 2015 data were not included in the *t*-SNE analyses because only a single core per lake was available.

The correlations between categorical and continuous variables to beta-diversity were assessed by unconstrained correspondence analysis with "envfit" from the R package vegan 2.4.3 (Oksanen et al., 2016). The variables were fit on NMDS ordinated distance matrices (described above) for both phylogenetic data and functional predictions, and the statistical significance was assessed with 10,000 permutations. For the continuous physicochemical variables, non-linear relationships were analyzed with "ordisurf", which is based on surface fits (contra vector fits in envfit). All *P*-values were Bonferroni-corrected per data set. Random forests (described above) were further used to corroborate these analyses. In the phylogenetic data, OTU abundances were grouped at phylum, class, and order levels for the random forest models, which were all screened for the best and most parsimonious model. Lower (*i.e.*, more exclusive) taxonomic levels were disregarded to increase the ecological meaningfulness of the results (Xu et al., 2014). Partial dependence plots were again generated with the R package edarf to examine the relationships between the most important OTUs and their functionally predicted group abundances to each sample category, spatial, and environmental variable. All these data analyses were done with R 3.4.0 (R Core Team, 2017); the corresponding scripts can be accessed through GitHub (https://github.com/Begia/Hazen16S), the sequencing data can be retrieved from the NCBI

Sequence Read Archive (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA430127), and the geochemical data from the National Centers for Environmental Information online repository (http://accession.nodc.noaa.gov/0171496).

2.4 Results and discussion

2.4.1 Sediment microbial communities are similar to other arctic lakes

Microbial community structure of Lake Hazen and Skeleton Lake sediments in spring 2014/2015 exhibited similarities to other lake sediments in polar (Mohit et al., 2017b; Tang et al., 2013; Wang et al., 2016b) and high-altitude regions (Zhang et al., 2014b) that have comparable ranges of temperature, nutrient and light availability. The most abundant bacterial phyla at our sampling sites were Proteobacteria, Bacteroidetes, Chloroflexi, Actinobacteria, Acidobacteria, Planctomycetes, and Verrucomicrobia (see **Appendix A**; **Figure 2.1**). The dominant archaeal phylum at all sites was Woesearchaeota, similarly to water columns of oligotrophic high-altitude lakes (Ortiz-Alvarez and Casamayor, 2016).

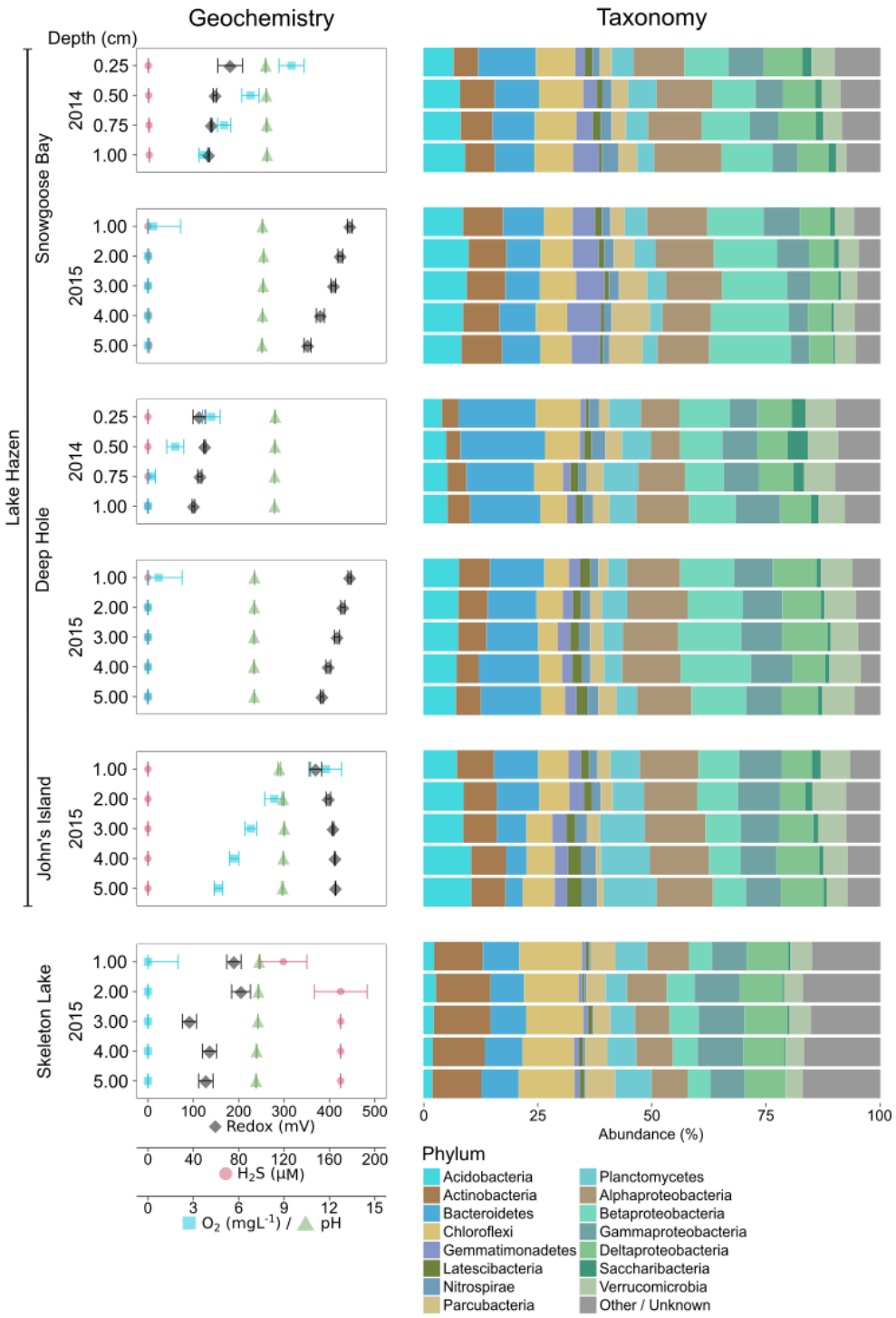

Figure 2.1: Geochemical variability, and microbial community composition of the spring 2014/2015 samples using universal primers. Abundances of taxa have been merged at the phylum level (Proteobacteria at class level). Phyla with less than 1% overall abundance in the data set are merged.

Sediments in Skeleton Lake had higher abundances of Chloroflexi, Actinobacteria, Cyanobacteria, and archaeal phyla, while Acidobacteria were more abundant in Lake Hazen sediments. Differences between the lakes might be driven by better light availability at the sediment-water interface, and higher production of sulfide in Skeleton Lake sediments compared to Lake Hazen. Indeed, all coring sites from Lake Hazen had overlying water columns of more than 40 m, measurable dissolved [O_2] in the top 1 cm (John's Island samples had > 4.7 mgL^{-1} O_2 down to 5 cm), and low [H_2S] (< 1.2 µM). Furthermore, although toxic, low levels of H_2S can enhance cyanobacterial photosynthesis when light intensity is low (Klatt et al., 2015). Hence, Cyanobacteria in Skeleton Lake might be able to photosynthesize below the ice cover in the spring.

In sediments sampled in summer 2015, Chloroflexi was the most abundant bacterial phylum in both Skeleton Lake and Pond1 (**Figure 2.2**). Their high abundance has been previously observed in hypersaline methane-rich springs in the High Arctic (Lamarche-Gagnon et al., 2015). In the current study, the salinity was low ([Cl^-] < 4.4 mgL^{-1}), but Skeleton Lake, and the ponds bordering Lake Hazen are methanogenic (Emmerton et al., 2016). The archaeal communities in Skeleton Lake sediments were dissimilar to those in Pond1. Woesearchaeota were more common in Skeleton Lake, and Euryarchaeota in Pond1 sediments.

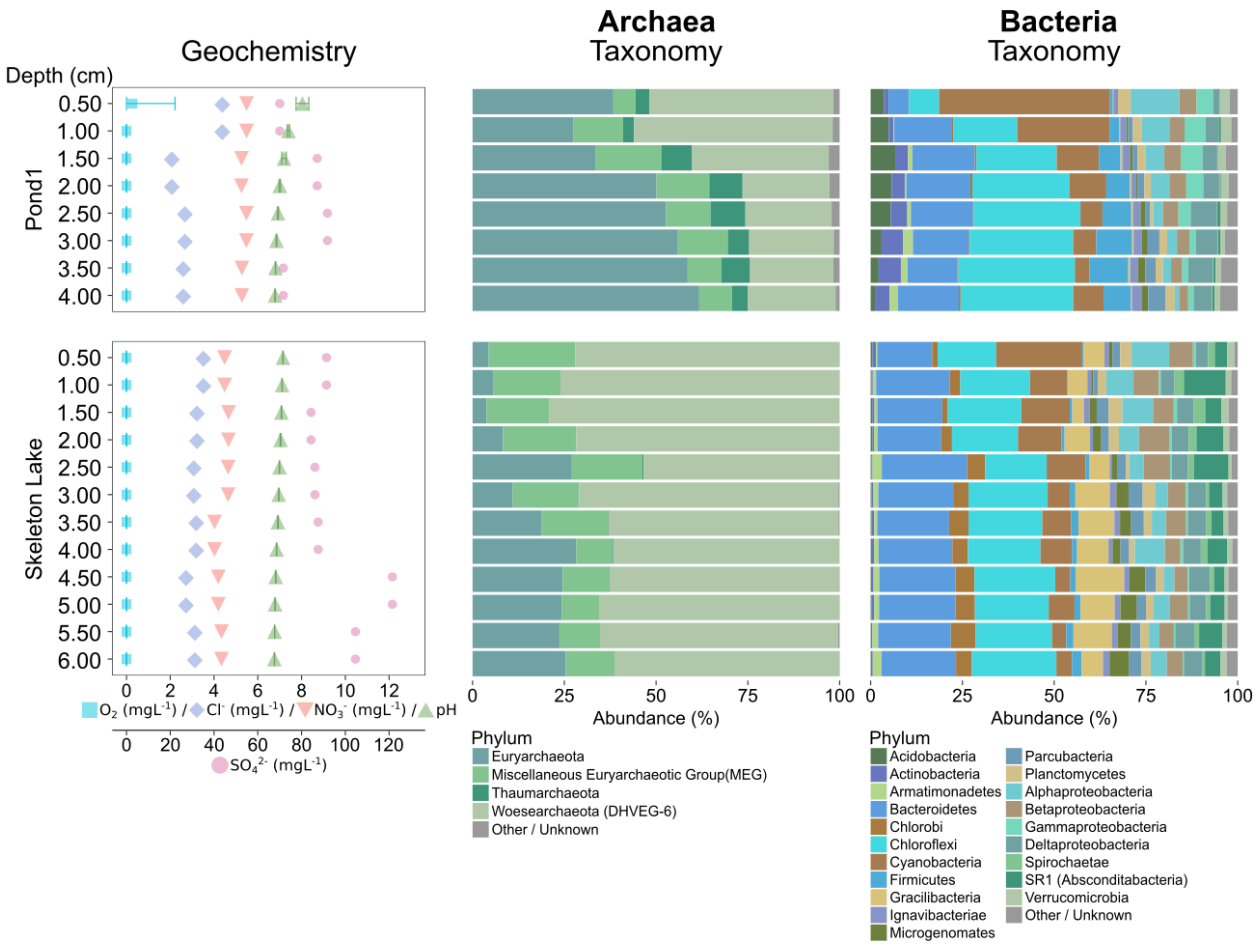

Figure 2.2: Geochemical variability, and microbial community composition of the summer 2015 samples using archaeal and bacterial primers. Abundances of taxa have been merged at the phylum level (Proteobacteria at class level). Phyla with less than 1% overall abundance in each data set are merged.

Mercury methylation had previously been quantified in both Skeleton Lake and Pond1 (Lehnherr et al., 2012), but its microbial actors were unknown. Fourteen OTUs in our data mapped to mercury methylation in our custom functional mapping database, and their 16S sequences, all matched closely to *Methanosphaerula palustris* (Cadillo-Quiroz et al., 2009). The genome of the type strain of this species has been shown to possess the *hgcAB* genes that

strongly predict mercury methylation capability (Gilmour et al., 2013). Other taxa most likely also take part in mercury methylation in these sediments. However, our amplicon-based study might have missed their presence because of primer bias and low 16S database coverage of organisms in these environments.

Intra-lake / pond compositional variability could also be high. For instance, the communities at the two sites sampled in Skeleton Lake in spring (**Figure 2.1**) and summer 2015 (**Figure 2.2**), were strikingly different. Sediments from the deeper site (**Figure A3b**), sampled in spring 2015 under ice cover, had a mostly heterotrophic community dominated by Proteobacteria. Meanwhile, sediments from the shallower site (**Figure A3d**), sampled in summer 2015, were dominated by phototrophs such as Chloroflexi and Cyanobacteria, and anaerobic fermenters such as Bacteroidetes and Gracilibacteria (Thomas et al., 2011; Wrighton et al., 2012). This indicates high spatial heterogeneity of sediment communities in Skeleton Lake sediments. Primer bias and seasonality of the microbial communities in Skeleton Lake might play a part in this, but the question would require further study. Our qualitative observations of the microbial communities are consistent with (i) measurements of high CH_4 emissions from ponds bordering Lake Hazen (Emmerton et al., 2016), (ii) increased [MeHg] in Skeleton Lake (unpublished data), (iii) high autochthonous carbon, and (iv) nitrogen limitation at the sites (St. Louis, unpublished data).

2.4.2 Both redox chemistry and pH drive community diversity and structure

2.4.2.1 Physicochemical data partially explains phylogenetic variability

The phylogenetic variability may be driven by the unique physicochemical properties of each site, which can vary substantially both in time and in space. A PCA of the physical and geochemical variables in spring 2014/2015 shows that samples group by individual

core (**Figure 2.3a**). More specifically, the PCA revealed two major independent (orthogonal) axes of variability: (i) [H_2S]/redox/water depth, and (ii) pH/[O_2] (**Figure 2.3a**). Samples with measurable [H_2S] had lower redox potential and were from shallower sites (mostly from Skeleton Lake; **Table A1**). Samples closer to the sediment/water interface had higher pH and [O_2]. Our sampling likely captured some of the most relevant physicochemical variables constraining the microbial community structures. Indeed, the Euclidean distances between the samples calculated from physicochemical variables were correlated with their phylogenetic DPCoA distances (Mantel-R^2 = 0.57, Bonferroni-corrected $P < 0.01$). The physicochemical variability also correlated with the functionally predicted group abundances (Bray-Curtis distances; Mantel-R^2 = 0.40, Bonferroni-corrected $P < 0.01$). However, only 25% of the OTUs could be mapped to any function and were thus covered by this analysis.

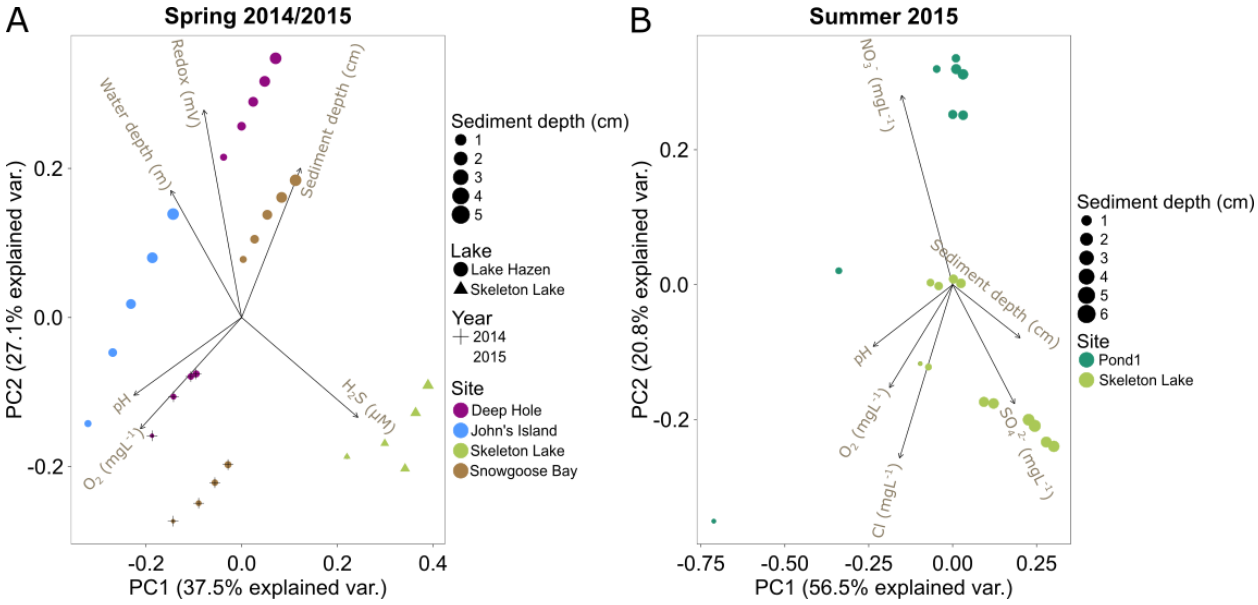

Figure 2.3: PCA biplots of the physicochemical variables. (**a**) Spring 2014/2015 samples. (**b**) Summer 2015 samples.

The PCA on the summer 2015 physicochemical data revealed that these sediments also clustered separately, with Pond1 on one side of PC2 and Skeleton Lake on the other side (**Figure 2.3b**). Most of the differences between these two sites were driven by higher [NO_3^-] in sediments from Pond1 and higher [SO_4^{2-}] in sediments from Skeleton Lake, while pH, [O_2], and [Cl^-] covaried. However, the top 1 cm surface sediments from Pond1 were highly influential in the PCA because of their higher pH and [O_2] than in other samples (**Figure 2.3b**, **Table A2**). This higher pH and [O_2] could reflect the influence of the incoming Lake Hazen waters into Pond1, which tend to be higher in pH and O_2, especially in the summer under the direct influence of the glacial inflows. It is possible that this difference in the scaling of the sites along the PCA reflects different water sources between Pond 1 and Skeleton Lake.

Unlike with the spring 2014/2015 data, the summer 2015 data showed no significant correlations between the physicochemical distances of the samples, and either phylogenetic data or functional predictions (bacterial and archaeal; all Bonferroni-corrected $P > 0.05$). This indicates that the measured physicochemical variability does not explain differences in community structures among samples. Unknown variables, such as redox potential, might be influencing the community assembly at these sites. Furthermore, the two sites in this data set have similar physicochemistry throughout each sediment profile, which probably reduces discriminatory power for this analysis.

2.4.2.2 Higher redox potential and lower sulfide concentration drive alpha-diversity

To identify the drivers of alpha-diversity at Lake Hazen, we fitted random forest models to our data (Touw et al., 2013). Based on Simpson's dominance, diversity in the spring 2014/2015 sediment samples was best predicted by a model including all physicochemical variables (in order of importance: [H_2S], overlying water depth, redox potential, site, lake, pH, sediment

depth, [O_2], and year; pseudo-R^2 = 0.72). These results are consistent with our expectations, as H_2S can be highly toxic to microbial communities (Brouwer and Murphy, 1995; Hoppe et al., 1990). Water depth was the second most important variable explaining alpha-diversity (**Figure 2.4a**). The shallow Lake Hazen sediments at Snowgoose Bay had the highest diversity (**Figure A1**; **Table A1**), which might be driven by high heterogeneity of the sediments, including steep [H_2S] and [O_2] gradients. The Snowgoose Bay site is also under the direct influence of two glacial river outlets, which might contribute to the heterogeneity through increased delivery of nutrients and inorganic matter. Our observations are consistent with previous findings of the positive relationship between sediment heterogeneity and alpha-diversity (Lozupone and Knight, 2007). Redox potential, the third most important continuous variable, also had positive relationship with predicted diversity. The effect was similar in magnitude to [H_2S] and was expected since sulfate reducers are active at low redox potentials. This is consistent with previous studies showing that microbial communities can react quickly to changes in redox potential, changing from aerobic chemoheterotrophy to anaerobic respiration and fermentation (DeAngelis et al., 2010). Finally, we identified pH as a driver of alpha-diversity: since non-extremophilic bacteria need to maintain an optimal intracellular pH of around 7.5 (Booth, 1985), the subsistence of a more diverse community at this pH might be facilitated.

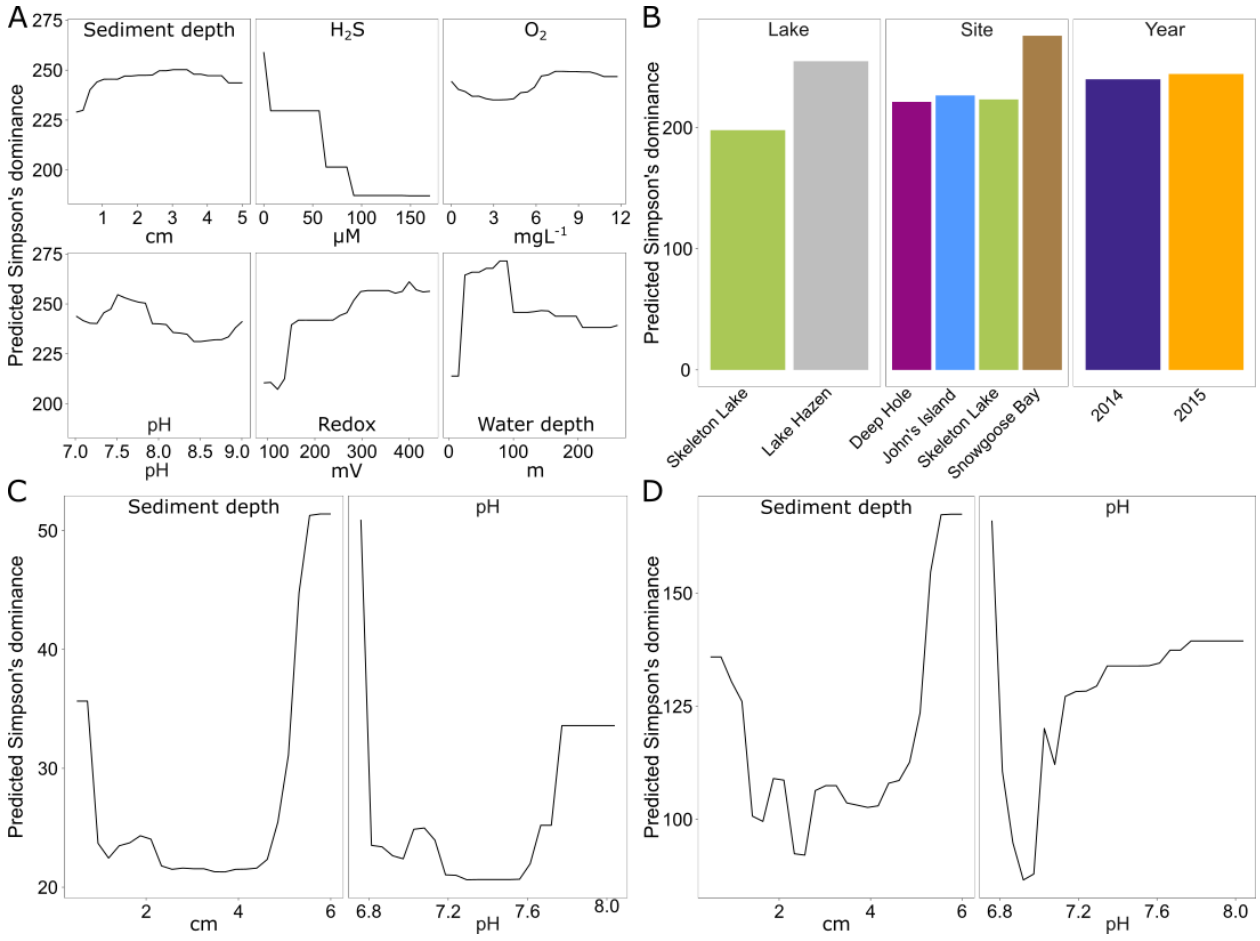

Figure 2.4: Partial dependence of predicted Simpson's dominance on continuous and categorical variables of the random forest model with the smallest prediction error, for each of the data sets. Spring 2014/2015 with universal primers (**a**) 6 continuous variables; (**b**) three categorical variables. (**c**) Summer 2015 with archaeal primers (two continuous variables). (**d**) Summer 2015 with bacterial primers (two continuous variables).

Again, the summer 2015 data set differed from spring 2014/2015 data set, as the best model only included sediment depth and pH as the most important variables for both archaeal (pseudo-R^2 = 0.53) and bacterial data (pseudo-R^2 = 0.32). However, direct comparisons between the spring 2014/2015 data and summer 2015 data sets are difficult, since different sites were

sampled, and different geochemical variables were measured. Archaeal and bacterial alpha-diversity in the summer 2015 data set were highest in the deepest sediments, with a discrete increase at the sediment surface (**Figures 2.4c**, **2.4d**). The increase in diversity might be caused by higher diversity of organisms with obligate aerobic (at the surface sediments) or anaerobic metabolisms (at deeper sediments). Unfortunately, no reliable data could be obtained for [H_2S] or redox potential in these samples because of broken microsensors. Here, pH also seemed to be a factor explaining diversity both for archaea and bacteria, but diversity predictions might be driven only by a few outliers at the extremes of sediment depth.

2.4.2.3 Communities cluster phylogenetically by pH and display similar functional predictions

To independently support these predictions based on random forests, we performed a *t*-SNE cluster analysis on the spring 2014/2015 samples. These samples clustered mostly by individual sediment core for both full phylogenetic data (**Figure 2.5a**) and for data including only the 25% of OTUs that could be functionally predicted (**Figure 2.5b**). The *t*-SNE analysis of functionally predicted data identified only two clusters, one per lake (**Figure 2.5c**). This shows that phylogenetically distinct sediment communities in Lake Hazen have similar functional predictions. Furthermore, Lake Hazen sediments clustered invariably separate from Skeleton Lake sediments in all of these analyses. Random forest classification of the clustering patterns identified pH as the most important predictor for clustering both in the full phylogenetic data set (one predictor; OOB Error = 0%; **Figure A11a**), and for the functionally mapped OTUs (seven predictors; OOB Error = 12%; **Figure A11b**). In the full phylogenetic data set, the sediment communities also appear to be more similar to each other over ranges of pH (**Figure 2.4a**, lower panel). [H_2S] was the most important predictor to explain differences between the clusters in the

functionally predicted data (one predictor; OOB Error = 0%; **Figure A11c**). However, because we sampled only a single site in Skeleton Lake in spring 2015, it remains uncertain if [H_2S] is the only factor affecting the observed difference in the functionally predicted groups in the sediments of the two lakes. Furthermore, heterogeneity of the communities within Skeleton Lake itself could not be addressed with the clustering analysis due to only a single core being analyzed. Regardless, all the phylogenetically distinct communities in Lake Hazen sediments clustered together after functional prediction.

Our results suggest that pH strongly affects phylogenetic community composition in our samples. Indeed, pH has previously been shown to be a major determinant of community composition in similar lake sediments (*e.g.*, Xiong et al., 2012). Sediment microbial communities might be altered in the future because climate change related effects can increase pH in arctic lakes (Kokelj et al., 2005; Mesquita et al., 2010). We observed that the microbial communities in Lake Hazen sediments cored at different sites at different times are phylogenetically distinct from each other and Skeleton Lake sediments. All the samples from Lake Hazen displayed similar functional predictions, while remaining distinct from Skeleton Lake samples. Decoupling between phylogeny and function of microbial communities has previously been observed, *e.g.*, in the global ocean microbial communities (Louca et al., 2016b), and plant-associated environments (Louca et al., 2016a). However, our results rely solely on the analysis of 16S rRNA genes, and therefore lack direct evidence about the actual microbial functioning and activity in the lake sediments. Critical insights could be gained here by employing metagenomics, metatranscriptomics, and (ideally) metaproteomics (Louca et al., 2016a).

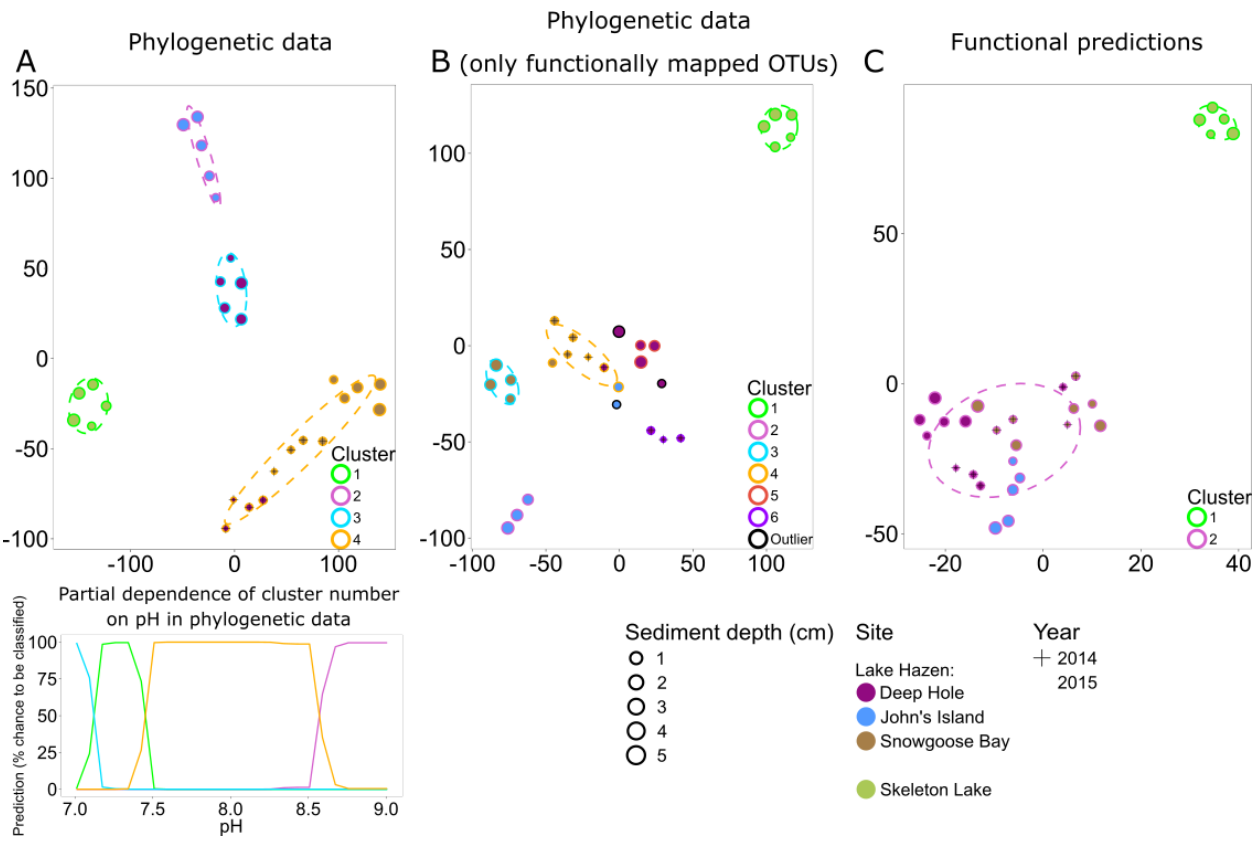

Figure 2.5: *t*-SNE analysis of spring 2014/2015 samples. Phylogenetic dissimilarities were measured with DPCoA, and differences in functional predictions with Bray-Curtis dissimilarity. (**a**) Phylogenetic data, including only OTUs with > 0.01% overall abundance. Partial dependence of cluster number on pH is included for this data below the *t*-SNE plot. (**b**) Phylogenetic data, including only OTUs that were matched to a function (roughly 25% of the full phylogenetic data). (**c**): Functional predictions.

2.4.2.4 Beta-diversity is also driven by redox chemistry and pH

For the spring 2014/2015 bacterial communities, the centroids of the clusters found by NMDS ordinations for both lakes and individual sites were different from each other (Bonferroni-corrected $P < 0.01$), but no year effect could be found (Bonferroni-corrected $P > 0.05$;

Figure 2.6a). As both spring 2014 and spring 2015 samples were also sequenced using the same primer set, we analyzed samples from different years together. The communities in Skeleton Lake sediments were phylogenetically distinct from Lake Hazen sediments, and the communities in individual Lake Hazen cores were also phylogenetically dissimilar to each other. While these patterns are consistent with the *t*-SNE analysis, they are not as clear because NMDS preserves pairwise distances instead of emphasizing them (like *t*-SNE). [H_2S], redox potential, and water depth correlated linearly with phylogenetic distances of the communities (Bonferroni-corrected $P < 0.05$; **Figure 2.6a**). Sediment depth was not linearly correlated with the phylogenetic distances, but the communities at the sediment surface might be more similar to each other than communities deeper in the sediment. This can be observed in the grouping of surface samples together in the middle of the ordination (**Figure 2.6a**). The deepest sediments at John's Island appeared quite unique, which might be due to the presence of O_2 all the way down to 5 cm below sediment surface, whereas O_2 is not found at any other sites below 1 cm (**Figure 2.1**, **Table A1**).

Archaeal communities in sediments from Pond1 and Skeleton Lake (summer 2015) also differed from each other phylogenetically (Bonferroni-corrected $P < 0.01$; **Figure 2.6b**). However, [NO_3^-] was the only physicochemical variable linearly correlated with phylogenetic differences of the communities in the samples (Bonferroni-corrected $P < 0.05$). Similar to archaeal communities, bacterial communities in sediments from Pond1 and Skeleton Lake (summer 2015) were phylogenetically significantly different from each other (Bonferroni-corrected $P < 0.01$; **Figure 2.6c**). [NO_3^-], pH, sediment depth and [Cl^-] correlated linearly with phylogenetic differences between the samples (Bonferroni-corrected $P < 0.05$). The communities in surface sediments of both Pond1 and Skeleton Lake seemed most dissimilar to the other samples from the same core.

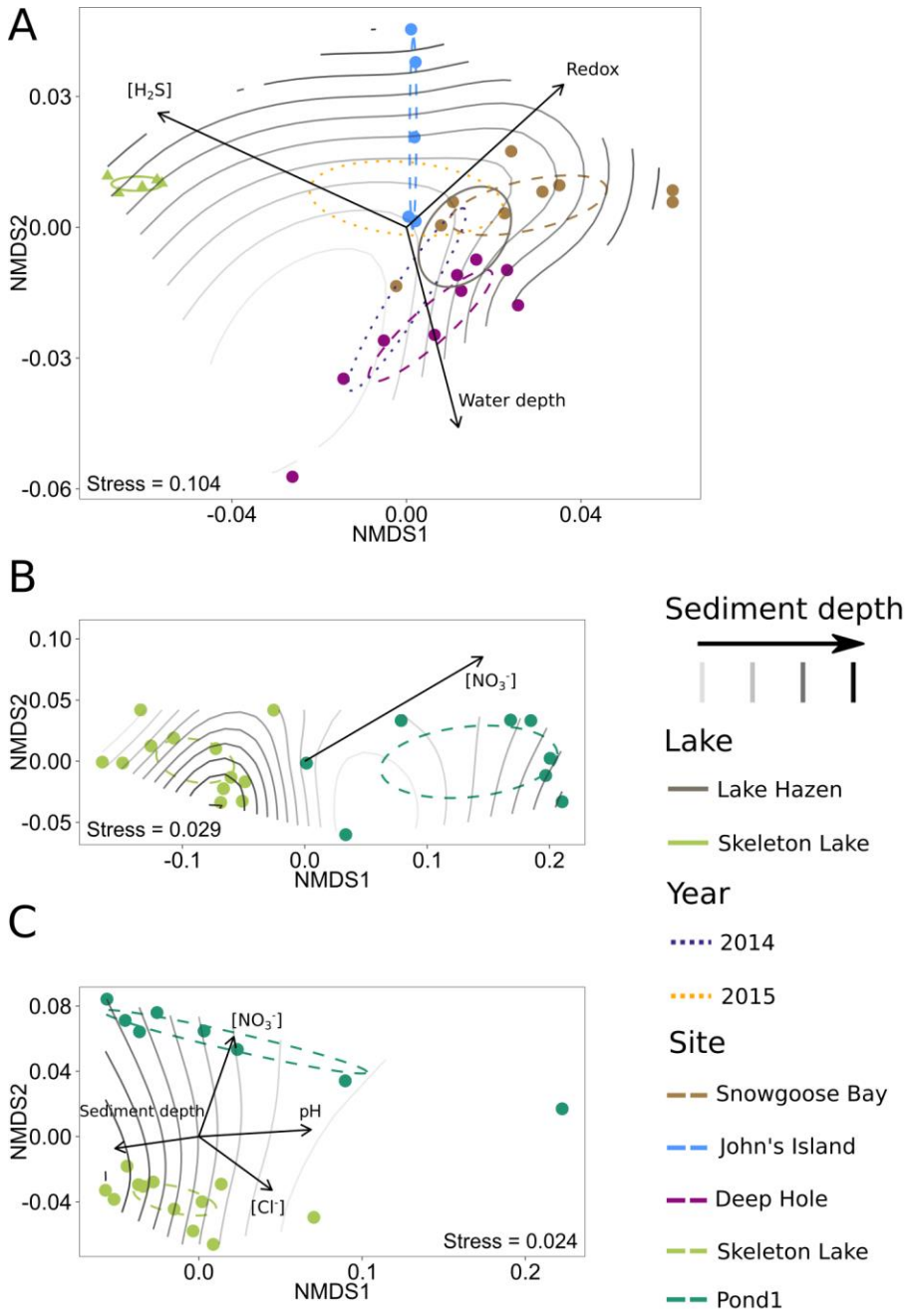

Figure 2.6: NMDS ordinations of phylogenetic DPCoA distances of the samples, with significantly correlated physicochemical variables as vectors. 95%-confidence interval for centroids of sample categories (lake, year, and site, where applicable) is shown with ellipses and sediment depth is overlaid on the plots as a surface fit. (**a**) Spring 2014/2015 with universal primers. (**b**) Summer 2015 with archaeal primers. (**c**) Summer 2015 with bacterial primers.

Altogether, beta-diversity seems to be affected mostly by [H_2S], redox potential and pH. These are the variables that have surfaced in either the ordination or *t*-SNE analysis for both spring 2014/2015 and summer 2015 data sets. In addition, we also observed trends with water depth in spring 2014/2015 data and [NO_3^-] in the summer 2015 data set. The effects of [H_2S] and redox potential are probably linked to toxicity of H_2S and different availability of electron acceptors in the changing redox potential, which together alter the community composition. Water depth in the spring 2014/2015 data set can be seen as a proxy for several factors influencing community structure; the depth of the overlying water column influences both light availability and sediment dynamics, such as differences in sedimentation rate and nutrient inputs, resuspension, and sediment focusing. However, the trends in summer 2015 data with [NO_3^-] are questionable, as (i) [NO_3^-] covaries with sulfate and sediment depth (deeper sediment horizons have lower nitrate and higher sulfate; **Figure 2.3b**; **Table A2**), and (ii) [NO_3^-] in Pond1 is much higher than in Skeleton Lake (**Table A2**).

2.4.3 Taxonomic group abundances vary along physicochemical gradients

We conducted random forest analyses to discover relationships between physicochemical gradients and abundances of taxonomic groups, and our functional predictions (see **Appendix A**; **Tables 2.1**, **2.2**; **Figures A15-A37**). We found an association between increasing levels of [H_2S] and (i) decreasing abundances of aerobic taxa and functionally predicted aerobic groups (putative aerobic ammonia oxidizers, aerobic chemoheterotrophs, aerobic nitrite oxidizers, and predatory/exoparasitic microbes), and (ii) increasing abundances of functionally predicted sulfate respirers, methanogens, and cyanobacteria (cyanobacteria are all photosynthetic and thus mapped to a single group; **Figure A15**). Skeleton Lake sediments had much higher [H_2S] than Lake Hazen sediments, but the community differences linked to [H_2S] are not completely explained by differences between the lakes (**Figure A34**). [H_2S] seems to affect both phylogenetic and functionally predicted community composition, and climate change has previously been thought to result in increased accumulation of sulfur in high arctic lake sediments (Drevnick et al., 2010). Chemical weathering of sulfate containing minerals (*e.g.*, gypsum-$CaSO_4$) following glacial melt and/or permafrost thaw could also increase delivery of SO_4 to waterbodies in the Lake Hazen watershed. Enhanced rates of sulfur cycling in sediments might change the community structure, which might affect other geochemical cycles mediated by the sediment communities.

Table 2.1: Summary of the regression random forest models for continuous variables. Number of predictors (*n*) is shown before and after stepwise model selection, while Mean Standard Prediction Error (MSPE) and pseudo-R2 are shown after selection.

			Before model selection	After model selection		
Data set	Variable	Taxonomic level	*n* (predictors)	*n* (predictors)	MSPE (95% CI)	pseudo-R^2
Spring 2014/2015	H_2S	Class	164	16	132.67 (0.00−267.65)	0.966
		Functional mapping	48	7	97.48 (15.62−179.35)	0.961
	pH	Order	280	8	0.06 (0.03−0.08)	0.875
		Functional mapping	48	3	0.07 (0.03−0.10)	0.851
	O_2	Order	280	3	5.42 (1.37−9.46)	0.560
		Functional mapping	48	7	5.99 (2.45−9.53)	0.514
	Redox potential	Order	280	8	3978.13 (1794.36−6161.90)	0.793
		Functional mapping	48	2	5036.48 (2100.30−7972.66)	0.738
	Sediment depth	Class	164	2	1.10 (0.58−1.63)	0.591
		Functional mapping	48	9	1.62 (0.87−2.38)	0.398
	Water depth	Class	164	2	738.28 (0.00−1484.27)	0.932
		Functional mapping	48	3	3121.19 (1611.48−4630.90)	0.711
Summer 2015/Archaea	pH	Class	13	3	0.04 (0.00−0.11)	0.538
		Functional mapping	10	7	0.04 (0.00−0.12)	0.482
	O_2	Phylum	8	1	0.00 (0.00−0.01)	0.000
		Functional mapping	10	2	0.00 (0.00−0.018)	-0.145
	SO_4^{2-}	Phylum	8	2	77.56 (19.82−135.3)	0.633
		Functional mapping	10	2	105.62 (41.37−169.86)	0.500
	Sediment depth	Order	12	9	0.69 (0.38−1.00)	0.744
		Functional mapping	10	2	0.45 (0.17−0.72)	0.833
	Cl^-	Order	12	3	0.01 (0.04−0.16)	0.729
		Functional mapping	10	7	0.27 (0.00−0.54)	0.270
	NO_3^-	Order	12	9	0.04 (0.01−0.06)	0.869
		Functional mapping	10	2	0.04 (0.01−0.06)	0.867
Summer 2015/Bacteria	pH	Class	85	8	0.04 (0.00−0.09)	0.585
		Functional mapping	26	4	0.04 (0.00−0.10)	0.518
	O_2	Phylum	37	1	0.00 (0.00−0.01)	0.000
		Functional mapping	26	4	0.00 (0.00−0.01)	-0.052
	SO_4^{2-}	Class	85	2	79.64 (43.64−115.64)	0.623
		Functional mapping	26	3	154.05 (74.19−233.9)	0.271
	Sediment depth	Class	85	1	0.38 (0.20−0.56)	0.858
		Functional mapping	26	5	0.99 (0.59−1.39)	0.631
	Cl^-	Order	113	10	0.11 (0.03−0.19)	0.695
		Functional mapping	26	10	0.19 (0.03−0.35)	0.465
	NO_3^-	Order	113	5	0.03 (0.01−0.04)	0.898
		Functional mapping	26	2	0.03 (0.01−0.04)	0.901

Table 2.2: Summary of the classification random forest models for categorical variables. Number of predictors (*n*) is shown before and after stepwise model selection, while Cohen's Kappa values and Out Of Bag (OOB) error rates are shown after selection.

			Before model selection	After model selection		
Data set	Variable	Taxonomic level	*n* (predictors)	*n* (predictors)	Cohen's Kappa	OOB Error (%)
Spring 2014/2015	Site	Class	164	2	1	0
		Functional mapping	48	11	0.9	7.14
	Lake	Phylum	54	1	1	0
		Functional mapping	48	1	1	0
	Sampling year	Class	164	2	0.91	3.57
		Functional mapping	48	3	0.81	7.14
Summer 2015/Archaea	Site	Phylum	8	3	1	0
		Functional mapping	10	1	1	0
Summer 2015/Bacteria	Site	Phylum	37	1	1	0
		Functional mapping	26	3	1	0

Taxonomic groups that increased in abundance with increasing redox potential were aerobic chemoheterotrophs, such as Acidobacteria (Ward et al., 2009), and obligate aerobic methylotrophic Betaproteobacteria (Chistoserdova and Lidstrom, 2013; **Figure A18a**). In addition, the functionally predicted group of methanol oxidizers increased in abundance with increasing redox, which suggests that these organisms are aerobic (Jenkins et al., 1987). However, putative sulfur reducers also showed a positive relationship with redox, which was a surprising result. Most of the taxa mapped with FAPROTAX to this functional group belong to the uncultured genus *Desulfurellaceae* H16, which has been previously detected in anaerobic bioreactors (Wei et al., 2017). Bacteria from the family Desulfurellaceae are typically strict anaerobic sulfur-reducers (Florentino et al., 2017; Greene, 2014), but here seem to be abundant at sites with high redox potential (> 400 mV) and in the presence of oxygen (> 4 mgL-1). To the best of our knowledge, this has not been observed in previous studies.

In the current study, we identified pH as an important driver of the sediment microbial community structure and diversity, similarly to previous studies (see **Appendix A**; Xiong et al., 2012). Random forest analysis showed that the relationships of taxonomic groups to variation in pH were mostly supportive of previous observations in lake sediments (see **Appendix A**; **Figures A16**, **A27a**; Xiong et al., 2012). We also detected an increased abundance of Cyanobacteria at higher pH (**Figures A16**, **A27**), which is in accordance to the generation of alkaline conditions via autotrophic pathways. Similar relationships between pH and Cyanobacteria in the High Arctic have been previously observed in lake microbial mats (Lionard et al., 2012). We also observed a higher abundance of functionally predicted sulfate respirers and methanogens at lower pH. This is in accordance with lower pH optimums of these processes (Ferry, 1993; Hao et al., 1996), than the average pH of 7-8 in our samples.

Finally, results from the random forest analysis showed that abundances of predicted fermenters and intracellular parasites (most of these are known as Amoebae-Resistant Microbes; Greub and Raoult, 2004) increase with water depth (**Figure A20b**). The OTUs identified in our analysis included representatives of, *e.g.*, phylum Chlamydiae (Lory, 2014), and orders Legionnellales (Garrity et al., 2015) and Rickettsiales (Renvoisé et al., 2011). The presence of obligate intracellular parasites indicates a higher abundance of grazing protists, and in the case of Rickettsiales, of arthropods (Renvoisé et al., 2011) at the deeper sites. These organisms might together with fermenting microbes contribute to increased cycling of organic matter and transfer of energy to higher trophic levels (Lei et al., 2014). The increased abundance of microbes involved in organic matter cycling suggests increased delivery of material to the deep basin (*i.e.*, sediment focusing) in Lake Hazen. Furthermore, the longer duration of ice-free periods (Latifovic and Pouliot, 2007; Surdu et al., 2016) and increased runoff (Bliss et al., 2014) seem to have already increased the sediment, carbon, and nutrient inputs to Lake Hazen (Lehnherr et al., 2018).

2.5 Conclusions

Despite extreme conditions in the High Arctic, our results show that lake sediments from this area harbor highly diverse microbial communities that vary both in time and space, but that are mainly shaped by redox and pH. Although the microbial communities in cores sampled at the three sites in Lake Hazen were phylogenetically distinct, they were functionally predicted to exhibit similarities. However, such functional predictions need now to be validated with metagenomics or metatranscriptomics studies, especially when performed on undersampled and extreme environments such as Lake Hazen.

The way such extreme environments will behave in the context of climate change is unclear. On the one hand, the predicted functional similarity of the communities in the backdrop of spatiotemporal microbial heterogeneity could be interpreted as a sign of resilience. However, as rising temperatures have both direct and indirect influences on redox chemistry and pH, the main drivers of microbial communities identified herein, it is very plausible that the current community structure could be disrupted under the climate regime predicted for the Arctic. Future work on Arctic lake sediments should focus on elucidating the functioning of the communities, and long-term studies performed throughout the seasonal regime shifts. As these seasonal shifts drive the redox chemistry, light and nutrient availability in the lakes, they might also affect the structure of microbial communities within.

2.6 Funding

This work was funded by the Finnish Academy of Science and Letters / the Vilho, Yrjö, and Kalle Väisälä Fund (MR), Natural Resources Canada / Polar Continental Shelf Project (VStL), the ArcticNet Centre for Excellence (VStL, AP), the Natural Sciences and Engineering Research Council of Canada (KS, VStL, SAB, AP), and the Canadian Foundation for Innovation (SAB, AP).

2.7 Acknowledgements

We would like to thank Pieter Aukes, Igor Lehnherr, Catherine Wong, Lisa Szostek, and Charles Talbot for their help in the fieldwork; Daniel Gregoire, Mija Ažđajić, Linda Kimpe and Ian Clark for their help and use of tools and facilities for the sediment core handling. We would like to thank the reviewers for their insightful comments that improved the quality of our manuscript. The high-performance computing environment for the data analyses was provided by Ontario's Centre for Advanced Computing.

2.8 Data availability

All analysis scripts generated for this study can be found in the GitHub repository (https://github.com/Begia/Hazen16S) and in the online supplementary file (**Appendix D**), sequencing data is deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA430127), and the geochemical data is deposited in the the NOAA National Centers for Environmental Information database (http://accession.nodc.noaa.gov/0171496).

Chapter 3: Microbial genomes retrieved from High Arctic lake sediments encode for adaptation to cold and oligotrophic environments

Matti O. Ruuskanen[1], Graham Colby[2], Kyra A. St.Pierre[3], Vincent L. St.Louis[4], Stéphane Aris-Brosou[5] and Alexandre J. Poulain[6]

[1,2,5,6]Department of Biology, University of Ottawa, Ottawa, ON, Canada.

[3,4]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada.

[5]Department of Mathematics & Statistics, University of Ottawa, Ottawa, ON, Canada.

Manuscript in press at Limnology and Oceanography (as of September 23, 2019).

doi:10.1002/lno.11334

Author contributions: MOR designed the experiments together with GC, SAB and AJP. GC, KStP and VStL performed the sampling. GC performed the DNA extractions and quality control in the laboratory. KStP and VStL measured the physicochemical data. MOR drafted the manuscript and analyzed the data. AJP and SAB supervised the project. All authors contributed to the writing and accepted the final version of the manuscript.

3.1 Abstract

The Arctic is currently warming at an unprecedented rate, which may affect environmental constraints on the freshwater microbial communities found there. Yet, our knowledge of the community structure and functional potential of High Arctic freshwater microbes remains poor, even though they play key roles in nutrient cycling and other ecosystem services. Here, using high-throughput metagenomic sequencing and genome assembly, we show that sediment microbial communities in the High Arctic's largest lake by volume, Lake Hazen, are phylogenetically diverse, ranging from Proteobacteria, Verrucomicrobia, Planctomycetes, to members of the newly discovered Candidate Phyla Radiation (CPR) groups. These genomes displayed a high prevalence of pathways involved in lipid chemistry, and a low prevalence of nutrient uptake pathways, which might represent adaptations to the specific, cold (~3.5°C) and extremely oligotrophic conditions in Lake Hazen. Despite these potential adaptations, it is unclear how ongoing environmental changes will affect microbial communities, the makeup of their genomic idiosyncrasies, as well as the possible implications at higher trophic levels.

3.2 Introduction

Climate change is transforming Arctic ecosystems: elevated temperatures and increasing precipitation have facilitated permafrost thaw (Romanovsky et al., 2010) and glacial melt (Lehnherr et al., 2018; Milner et al., 2017), leading to impacts on the functioning of aquatic and terrestrial ecosystems, and the natural services they provide. Microbes are major players in the biogeochemical cycling of organic matter and inorganic nutrients; as such studying contemporary Arctic microbial communities is critical for both documenting and predicting how these cycles might respond to ongoing and future environmental change. However, baseline microbial community data from the Arctic are still lacking because these environments are remote and challenging to study. Furthermore, in-depth study of microbial communities has only recently become possible with culture-independent methods like high-throughput sequencing (*e.g.*, Shokralla et al., 2012). As Arctic environments are already responding to climate change at the watershed scale (Lehnherr et al., 2018), and these warming-related changes are projected to continue (IPCC, 2018), there is an urgent need to gather data on the current state of Arctic microbial communities and their function from which to compare in the future.

To date, culture-independent studies of microbial communities in Arctic lakes have been most often performed by amplifying and sequencing taxonomic markers like the 16S rRNA gene (*e.g.*, Crump et al., 2012; Mohit et al., 2017; Ruuskanen et al., 2018; Stoeva et al., 2014; Wang et al., 2016, 2019). However, amplicon-based methods are subject to PCR amplification bias, which might alter the estimates of microbial community composition and diversity. Furthermore, while taxonomy based on a marker gene can be used to predict the functional potential of microbes (Louca et al., 2016b), these predictions are purely hypothetical in the absence of functional data. Metagenomic sequencing enables the reconstruction of nearly complete

Metagenome Assembled Genomes (MAGs) solely from environmental DNA (Zhou et al., 2015). Gene sequences coding for proteins derived from contiguous sequences in the metagenomes can also be used to reconstruct the functional potential of microbial communities in the sampled environment. For example, metagenomic sequencing enabled the discovery of the Candidate Phyla Radiation (CPR; Brown et al., 2015; Hug et al., 2016), consisting of uncultured, deeply branching lineages in bacteria, which had previously evaded detection in purely amplicon-based studies. After their initial discovery, CPR members have been found in a variety of environments, including the deep subsurface (Danczak et al., 2017), marine sediments (León-Zayas et al., 2017), and hypersaline soda lakes (Vavourakis et al., 2018). Presence of CPR bacteria has also been reported in Arctic freshwater environments (Vigneron et al., 2019; Wurzbacher et al., 2017). While they appeared to be absent in 16S rRNA amplicon data, a metagenomic analysis of the same samples revealed them to be highly abundant (Vigneron et al., 2019). However, shotgun metagenomic data from lake sediment microbial communities in Arctic environments are still scarce (Wang et al., 2019b), and both larger lakes and High Arctic lakes remain thus far uncharted by these methods. Such knowledge would definitely expand our understanding of Arctic and global microbial diversity.

To investigate this diversity of microbes and their metabolisms in understudied Arctic lakes, we analyzed shotgun metagenomes from sediment extracted DNA from Lake Hazen, the world's largest High Arctic lake by volume. In a previous study using 16S rRNA gene amplicons, we hypothesized that taxonomically dissimilar sediment communities at different locations in the lake might be functionally similar (Ruuskanen et al., 2018). In the current study, we revisited this question by investigating the taxonomic diversity and functional potential of the sediment microbial community using metagenomics. We identified rarely studied organisms within the

sediment, characterized the metabolic pathways that are over- and underrepresented in these reconstructed genomes (compared to reference data from online repositories as of June 2018), described the ecologically important nutrient cycles that are potentially present in the sediments, and identified which taxa might play key roles in them.

3.3 Material and methods

3.3.1 Sampling and chemical analyses

Lake Hazen (located within Quttinirpaaq National Park on northern Ellesmere Island, Nunavut, Canada; 81.8°N, 71.4°W) is 544 km^2 in surface area, with a maximum depth of 267 m, making it the largest lake by volume north of the Arctic Circle. The area immediately surrounding the lake is a polar oasis with higher than average temperatures for similar latitudes. Temperatures over 0ºC have been observed in this region on more than 80 days per year (Soper and Powell, 1985), which is likely due to thermal shielding by the Grant Land mountains in the northwestern portion of Lake Hazen's watershed (France, 1993). The lake is primarily fed hydrologically by meltwaters of the outlet glaciers of the Grant Land Ice Cap, and has a single major outflow, the Ruggles River, which flows southeastwardly to the eastern coast of Ellesmere Island. Sediment cores were collected from two sites: Deep Hole [261 m] on the 4th of August and Snowgoose Bay [49 m] on the 8th of August 2016 (**Figure B1**). Sampling was conducted from a boat using an UWITEC (Mondsee, Austria) gravity corer with 86 mm inner diameter polyvinyl chloride core tubes. At both sites, triplicate cores were collected: one each for DNA extraction, porewater chemistry and microsensor measurements. While the extracted cores were up to 40 cm in length, the subsectioning was restricted to the topmost 6 cm in all cores, since microprobes cannot be pushed any deeper than this into the sediment cores. The core from which DNA were extracted was sectioned at 0.5 cm intervals in the field, preserved with LifeGuard™ (MO BIO, Carlsbad,

CA), and stored at -18°C until DNA extraction. Because of logistical difficulties involved with sampling in the High Arctic, the sectioning equipment could only be cleaned with non-sterile lake water and bleach between each section before putting the complete sections into sterile 50 mL centrifuge tubes. Non-powdered nitrile gloves were worn while handling the samples. For porewater analyses (NH_4^+, NO_2^-/NO_3^-, SO_4^{2-}, TDP, Cl^-), the core was similarly sectioned at 1 cm intervals and placed in sterile 50 mL centrifuge tubes. The sediment sections were centrifuged at 4500 rpm for 15 minutes to separate the sediments from the porewater. The supernatant was then filtered through 0.45-µm cellulose acetate syringe filters that were first rinsed with a bit of sample water. The remaining filtrate was stored in sterile 15 ml Corning polystyrene centrifuge tubes, and then immediately frozen at -18°C until analyses for NH_4^+, NO_2^-/NO_3^-, SO_4^{2-}, TDP and Cl^-. Porewater chemical concentrations were determined using CALA-certified protocols at the Biogeochemical Analytical Service Laboratory (University of Alberta, Edmonton, AB, Canada). On the final core, 100-µm resolution microprofiles of O_2, redox potential, and pH were measured using Unisense (Aarhus, Denmark) glass microsensors interfaced with the Unisense Field Multimeter immediately upon return to camp. Cores were maintained at ambient temperatures (~4°C) throughout profiling. The microprofiles of these cores have also been described in an earlier study (St. Pierre et al., 2019b).

3.3.2 DNA extraction and sequencing

For the extraction of environmental DNA, triplicate ~0.5 g (wet weight) subsamples were taken from three intervals per sediment core (Deep Hole: 0.0 – 0.5 cm, 1.0 – 1.5 cm, 2.0 – 2.5 cm; Snowgoose Bay: 0.0 – 0.5 cm, 1.5 – 2 cm, 3.0 – 3.5 cm). The samples were first washed once with a saline buffer (10 mM EDTA, 50 mM Tris-HCl, 50 mM $Na_2HPO_4 \cdot 7H2O$ at pH 8.0) to remove inhibitors (Poulain et al., 2015; Zhou et al., 1996), and then DNA was extracted from the

samples (and a negative reagent control) with the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Sample manipulations and extractions were conducted with sterilized equipment in a laminar flow cabinet (HEPA 100). The quality of the DNA was checked with a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, DE, USA) and through confirming the amplification of the *glnA* gene from the extracts (with *E. coli* DNA as the positive and sterile H_2O as the negative control) by PCR and gel electrophoresis (see SI text). The *glnA* gene was chosen as the control for DNA quality because its genomic copy number should be lower than that of the 16S rRNA gene (Stoeva et al., 2014a) and a positive result would better confirm the quality of the DNA. The triplicate DNA extracts were then combined for each core horizon. Library preparation and sequencing was completed by Genome Quebec (Montreal, QC, Canada) with Illumina HiSeq 2500 PE125 in triplicate lanes for each sample.

3.3.3 Preprocessing high-throughput sequencing data

The forward and reverse reads were trimmed and filtered for size and quality with Trimmomatic 0.36 (Bolger et al., 2014) and the data from all six samples were co-assembled with Megahit v1.1.2 (Li et al., 2015). Anvi'o v4 (Eren et al., 2015) was used for database management, following their standard metagenomic workflow. Briefly, reads were mapped to the contigs with Bowtie 2 (Langmead and Salzberg, 2012) and contigs longer than 1 kbp were binned with CONCOCT 0.4.1 (Alneberg et al., 2014). Open reading frames were identified with Prodigal (Hyatt et al., 2010), and functional annotations were inferred based on six reference systems: NCBI's Cluster of Orthologous Genes (COG; Galperin et al., 2015; Tatusov et al., 2003) was run within Anvi'o using DIAMOND (Buchfink et al., 2015) for the protein alignments using the default *E*-value cutoff of 0.001. Following this, Pfam (Finn et al., 2014), TIGRFAM (Selengut et

al., 2007) and Gene Ontology (GO; Ashburner et al., 2000; The Gene Ontology Consortium, 2017) annotations for proteins were added with InterProScan 5.29-68.0 (Jones et al., 2014). Briefly, for both the Pfam and TIGRFAM reference systems, the query protein sequences were searched with HMMER3 (Eddy, 2009) against the respective hidden Markov model databases. Hits for individual query proteins were filtered based on curated model-specific cut-offs and, in the case of Pfam, lower-scoring hits in the same Pfam clan were removed (Jones et al., 2014). Finally, GO terms were associated with the proteins within InterProScan through cross-referencing the Pfam and TIGRFAM annotations with the InterPro database (Hunter et al., 2012).

These bins that were assembled automatically from the contigs were then manually refined in Anvi'o to < 10% contamination (also named "redundancy" in the Anvi'o documentation), based on single copy genes following Campbell et al. (2013) for bacteria, and Rinke et al. (2013) for archaea. These contamination estimates were cross-compared against the lineage-specific marker genes from CheckM v1.0.11 (Parks et al., 2015), and bins that still had CheckM contamination > 10% were further refined manually by splitting. Final completion values for the refined genome bins were also estimated with CheckM. Read coverages of the MAGs in each sample were calculated with Anvi'o, normalized to total number of reads in each sample, and finally scaled to binned reads in each sample. Genome bins were analyzed for the presence of KEGG (Kanehisa et al., 2017) and MetaCyc (Caspi et al., 2018) pathways by mapping GO terms of each bin to Enzyme Classification (EC) categories and reconstructing parsimonious pathways with MinPath (Ye and Doak, 2009). The reconstructed genomes were also analyzed for the presence of marker genes (**Table D1**) and MetaCyc pathways (**Table D2**) for core elemental cycles (C, N, P, and S), together with metal regulation and homeostasis genes.

Abundances of individual MetaCyc pathways were calculated by summing the sample-wise normalized abundances of each MAG indicated to contain the pathway. The abundance of the individual marker genes in the separate samples was calculated from gene-level read coverages in Anvi'o, followed by normalization to total reads in a sample.

To estimate the robustness of the MAG-based community composition in the samples, reads identified as 16S rRNA genes were extracted from the dataset, and assembled with MATAM (Pericard et al., 2018). The 16S rRNA gene contigs were then classified against the SILVA 128 NR95 database (Quast et al., 2013) with the RDP Classifier (Wang et al., 2007). The abundances of the operational taxonomic units (individual 16S rRNA contigs) in MATAM were calculated as the proportion of 16S rRNA reads mapping to each contig per sample.

For phylogenetic and taxonomic comparisons of MAGs to reference data, all non-redundant (one per species) complete bacterial and archaeal genomes, and all the available Candidate Phyla Radiation (CPR; Brown et al., 2015; Hug et al., 2016) genomes, were downloaded from the NCBI GenBank database (Benson et al., 2012). As of June 2018, this comprised 3,362 bacterial, 240 archaeal and 3,561 genomes for CPR. A further 71 CPR genomes were added from Danczak et al. (2017). Open reading frames were used if available, or they were identified de novo with Prodigal. Functional annotations for genes and pathways in these genomes were performed de novo as described above for MAGs. For assessing the phylogeny, sequences of 16 ribosomal proteins were extracted from both the NCBI genomes and our MAGs that were > 70% complete (following Hug et al., 2016; **Table D3**). The sequences were aligned per ribosomal protein with MAFFT 7.402 (Katoh and Standley, 2013) using translated protein sequences, and back-translated to the original nucleotide sequences with TranslatorX (Abascal et al., 2010). Badly aligned sequences were removed, and the alignments were trimmed with trimAl

1.2rev59 using the '-gappyout' mode (Capella-Gutiérrez et al., 2009). Phylogenetic trees were constructed from each ribosomal protein with FastTree 2.1.9, compiled with double precision to estimate accurately short branch lengths (as recommended in the manual), and using the GTR + Γ model of sequence evolution (*e.g.*, Aris-Brosou and Rodrigue, 2012). Sequences with unexpectedly long branches in the individual ribosomal protein trees were then removed with treeshrink (Mai and Mirarab, 2018) with tolerance of false positives '--quantiles' set to 0.01. The trimmed alignments were concatenated for each genome with gap characters added for missing ribosomal proteins. NCBI genome entries with more than 25% gaps and MAGs with more than 50% gaps over the full alignment were then removed.

The higher cut-off on gap proportion for MAGs (50%) compared to the reference genomes (25%) was used to enable inclusion of a higher number of MAGs in the downstream analyses. However, together with the low cut-off used in the binning step (all >1 kbp contigs included), the phylogenetic uncertainty of our taxonomic assignments could have been increased. Thus, we calculated the pairwise phylogenetic distance of each MAG against the reference genomes for each of the ribosomal protein trees with the 'cophenetic' function from ape (Paradis et al., 2004), and the number of incidences of each reference genome in the ten closest genomes for each MAG. The correlation between the proportional incidence of the most commonly found reference genome in each MAG was then compared against the number of different contigs containing ribosomal proteins in them with a least-square linear regression, where the number of contigs was $\log_{10}$-transformed. Finally, the phylogenetic tree containing both the MAGs and the reference genomes was constructed with FastTree 2.1.9 under the GTR + Γ model of sequence evolution, as above.

3.3.4 Taxonomy analysis and functional potential of the microbial community

For further data analysis, only MAGs with < 50% gaps over the complete alignment were preserved (*n* = 55; **Table D4**). The phylogenetic tree containing the MAGs and reference genomes was visualized with ggtree (Yu et al., 2017), and annotated based on NCBI taxonomy. The taxonomy of the MAGs was manually annotated based on the lowest level in a monophyletic clade starting from a node with a support value of at least 0.5 (full tree is included in the SI both as a PDF and a Newick tree file). For MAGs with uncertain taxonomy assignments at the phylum level, 16S rRNA sequences were extracted from their genomes (> 200 bp; *n* = 2) with ssu-align 0.0.1 (Nawrocki, 2009). Also, the 16S rRNA sequences of their closest relative reference genomes based on the ribosomal protein tree were extracted. The MAG 16S rRNA sequences were matched against the NCBI's nr database with BLASTN 2.8.1+ (Zhang et al., 2000) and the top 50 hits for each query were downloaded. These collections consisting of 16S sequences from the MAG itself, *Thermanaerovibrio acidaminovorans* DSM 6589, Candidatus *Caldatribacterium saccharofermentans* OP9-77CS, *Acetomicrobium hydrogeniformans* (NCBI Reference Sequence: NR_116842.1), and the 50 top matches from the BLASTn searches were then aligned and trees built with the Silva Alignment and Tree service (Pruesse et al., 2012), using FastTree under the GTR + Γ model of sequence evolution. The trees were then rooted using the *Acetomicrobium hydrogeniformans* 16S rRNA gene as the outgroup.

Phyloseq 1.24.0 (McMurdie and Holmes, 2013) was used to manage abundance and phylogenetic data. Phylum-level microbial community composition was qualitatively compared between the 16S rRNA assembly and the genome assembly of 55 MAGs, where taxonomy assignments were based on the tree of ribosomal proteins. Differences in the phylum-level 16S rRNA based taxonomy between the samples were also assessed by fitting a generalized linear

model with a quasi-binomial link function (using 'glm'; R Core Team, 2018) to the data. To examine patterns in community structure and function of the MAGs, between-sample distances were first calculated for genome phylogeny using patristic distances with Double Principal Coordinate Analysis (DPCoA; Pavoine et al., 2004). For 16S-based taxonomy, and functional pathways, Bray-Curtis dissimilarities were calculated in vegan 2.5.2 (Oksanen et al., 2018). The distance matrices were ordinated with NMDS, and envfit (Oksanen et al., 2018) was used to correlate the ordinations with sample physicochemistry. To further compare clustering patterns, the distance matrices were dimensionally reduced with *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE; van der Maaten and Hinton, 2008) with the R package rtsne 0.13 (Krijthe and van der Maaten, 2017) and 'perplexity' set to 1.6. Clusters were then detected with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al., 2013) in dbscan 1.1.2 (Hahsler et al., 2018) with 'minPts' set to 2.

The marker genes and MetaCyc pathways were classified to environmentally relevant cellular processes (Carbon (C), Nitrogen (N), Sulfur (S), Phosphorus (P), and toxic metal cycling) and these were divided into categories (**Table D5**). Some pathways were deemed to be misclassified based on their usual taxonomic range specified in the MetaCyc pathway descriptions and subsequently removed from the data. The proportion of genomes that had at least one single marker gene or pathway for a process were quantitatively compared between the MAGs (n = 55) and reference genomes subset to phyla shared with the MAGs (n = 2,486). The homogeneity of pathway abundances between MAGs to the reference genomes was assessed with a Pearson's χ^2 test where *P*-values were estimated based on 10,000 permutations. Differentially abundant processes for each comparison, which contributed more than their equal share (out of n = 46 processes) to the total X^2 score, were visualized with heatmaps of their

Pearson residuals. Finally, the gene-level read coverages (normalized to total number of reads per sample) of N, S, P, and toxic metal processing marker genes in the Lake Hazen sediments were compared between the two sites and between oxic (> 0.0 mg L^{-1}) and anoxic (0.0 mg L^{-1}) samples. The significances of the differences were assessed with pairwise *t*-tests, using 'mt' (False Discovery Rate -based correction for multiple testing) from Phyloseq 1.24.0 (McMurdie and Holmes, 2013). To identify the most abundant MAGs and phyla for each process across the samples, the read coverage of each of the 55 MAGs (relative to total number of reads per sample) was averaged over all six samples (**Table D6**). The relative abundances of the MAGs were also summed together at the phylum level (or class for Proteobacteria) by processes identified through MetaCyc pathways or individual marker genes (**Table D7**). The most abundant phylum in each process was then identified for inclusion in the flow-charts of the N, S, and Hg cycles (**Figure 4.5**).

Primary sequence data produced in this study is available in the NCBI Sequence Read Archive and the 55 medium-quality MAGs are available in GenBank (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA525692). The geochemical data are available in the NSF Arctic Data Center online repository (https://arcticdata.io/catalog/#view/doi:10.18739/A2SJ19Q9P). Shell scripts, and code used in R 3.5.2 (R Core Team, 2018), are available in the supplementary material (**Appendix D**) and through GitHub (https://github.com/Begia/Hazen-metagenome).

3.4 Results and discussion

3.4.1 Reconstruction of Metagenome Assembled Genomes (MAGs)

A total of 115.6 Gbp of sequence in 685 million reads were obtained from the six samples (**Figure B2**). The reads were co-assembled into 5.3 million contigs consisting of 3.7 Gbp of non-redundant sequences. The longest contig was 408.1 kbp and the N50 was 748 bp. The performance of our assembly was slightly better than when the same assembler was used for a complex soil metagenome (Howe et al., 2014; Li et al., 2015), but comparable to other studies of sediments with the same assembler (*e.g.*, Carr et al., 2018; Lau et al., 2018). We sorted the contigs into 1488 bins with < 10% contamination, of which 146 were at least 'medium-quality draft' assemblies with > 50% completion (after Bowers et al., 2017). The remaining bins were 'low-quality draft' assemblies at < 50% completion (n = 1342). On average, 58% of all reads per sample were included in the 1488 bins.

Of these 146 medium-quality draft genomes, 68 were found to be > 70% complete, of which only a subset of 55 MAGs had > 50% of the nucleotides in the concatenated alignment of ribosomal proteins, deemed to be the minimum amount of information for placing them in the genome tree. On average, these 55 MAGs had a completeness of 88.2% based on single copy genes, a G/C ratio of 52.9, an N50 of 22 kbp, a genome size of 3.1 Mb, and contained 3,117 genes with a coding density of 90% (**Table D4**). Note that we used a 1 kbp lower bound on contig size during the binning process, essentially to increase the number of contigs in this step, which is somewhat lower than has recently been used (*e.g.*, 2.5 kbp and 5 kbp in Delmont et al., 2018). However, only eight MAGs (14.5% of the total) appeared to have low N50 values (2.3 - 5 kbp), and the phylogenetic uncertainty in the taxonomic assignments of the MAGs was not correlated to the number of contigs containing ribosomal proteins in each MAG (P = 0.9). Thus,

while this lower 1 kbp cut-off that we used in binning and taxonomy assignment of the MAGs might have increased their fragmentation, this liberal cut-off did not increase the uncertainty of their phylogenetic placement. These are key considerations for the reliability of the downstream analyses, because both the reconstructions of the metabolic pathways of the 55 MAGs (with MinPath) and their phylogenetic placement (with ribosomal proteins) are based on the collection of annotated genes contained in the individual MAGs.

3.4.2 Differential recovery of MAGs compared to 16S rRNA gene data

The most common phyla (or class for proteobacteria) in the MAGs were Verrucomicrobia (14%), Betaproteobacteria (13%), Alphaproteobacteria (12%), Planctomycetes (10%) and Actinobacteria (9%; **Figure 3.1a**). To quantify the reliability of this taxonomic profile, we also binned the raw reads separately with the MATAM pipeline specifically designed to identify 16S rRNA genes (Pericard et al., 2018). This analysis, performed with the default settings of the pipeline, produced 166 OTUs which covered at least 500 bp of the complete 16S rRNA gene, about a third of its length. Among these OTUs, the most common bacterial and archaeal phyla were Betaproteobacteria (20%), Woesearchaeota (13%), Alphaproteobacteria (10%), Actinobacteria (8%) and Deltaproteobacteria (5%; **Figure 3.1b**). As a result, the 55 medium-quality MAGs represented a differential recovery of the total (16S rRNA reads-derived) microbial community, apparently missing at least two major taxonomic groups, Archaea and Firmicutes. This may be due to challenges inherent to identifying taxa based on 16S data and discrepancies between the NCBI and the SILVA databases (Parks et al., 2018), as well as at least two additional issues.

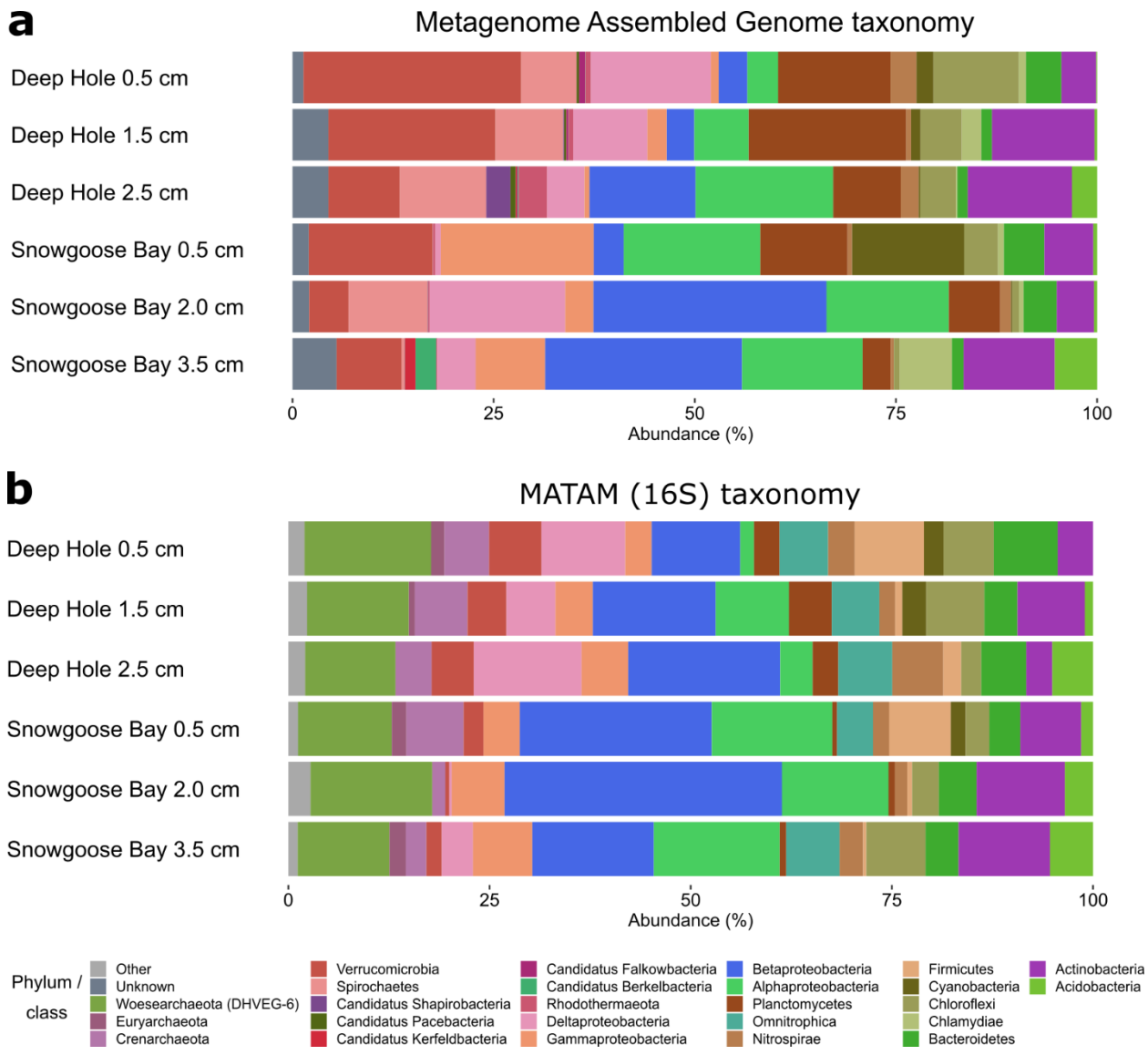

Figure 3.1: Composition of the microbial communities in the samples from Lake Hazen sediments. (**a**) Medium-quality draft MAGs (*n* = 55) with manual taxonomy assignments based on reference genomes in a phylogenetic tree constructed from ribosomal proteins. The phylum Proteobacteria is here subdivided into classes. (**b**) Raw reads binned into 16S rRNA contigs (*n* = 166) with MATAM and automatic taxonomy assignments from the SILVA 128 NR95 database.

First, while we identified 25 archaeal bins in the assembly, only six of them were over 10% complete and none were more than 70% complete. It should be noted that the six sequenced samples had similar community compositions based on the 16S rRNA reads-derived data (Figure 1b; all $P = 1.0$ at the phylum-level), which might have increased the binning difficulty because it utilizes differences in read coverages of contigs between samples (*e.g.*, Alneberg et al., 2014). Second, Firmicutes appeared to be underrepresented among the assembled bins. Only nine low-quality Firmicutes genomes were assembled, of which seven were between 10% and 50% complete, and the rest below 10% complete. This underrepresentation of Firmicutes in the assembled genomes might have been caused by their endosporulation affecting DNA extraction, making them detectable only by marker genes (Filippidou et al., 2015). Despite the difficulties related to the metagenomic assembly, the taxonomic composition of the microbial community in Lake Hazen sediments was very similar to our previous study of the same sites in spring 2015 (Ruuskanen et al., 2018). However, several groups that were found here to be highly prevalent (Archaea, Omnitrophica) were likely missed in the earlier study because of PCR selection bias (Suzuki and Giovannoni, 1996). The community was also similar to previous metagenomic studies of oligotrophic lake sediments (*e.g.*, Wang et al., 2016). One notable exception to previous studies (*e.g.*, Rautio et al., 2011) was that Cyanobacteria were much less common in Lake Hazen. This is likely due to most other arctic studies having focused on shallow thaw ponds, whereas Lake Hazen is ~260 m deep, with light penetration restricted to the upper ~25 m of the water column. Furthermore, other physicochemical characteristics of Lake Hazen (such as ultra-oligotrophy) may not be favorable to Cyanobacteria (St. Pierre et al., 2019b), which might affect their abundance in sediments as well. It should also be noted, that at least the phylum-level microbial community composition in Lake Hazen sediments appeared to be stable from 2015

spring to 2016 summer, despite high sedimentation rates in summer of 2015 resulting from enhanced glacial melt (St. Pierre et al., 2019b).

To understand how the physicochemistry of the lake sediments (**Figure B3**) can drive community structure and function, we used ordination and clustering methods to compare the samples (for an in-depth analysis of Lake Hazen contemporary limnology, see St. Pierre et al., 2019), first based on the 16S-derived data. Here, the sites from Lake Hazen did not have significantly different microbial communities (based on 10,000 permutations; $P = 0.10$), and Cl^- concentration was the only chemical variable with a significant linear correlation with the differences in microbial community structure ($P = 0.03$; **Figure B4**). However, the Cl^- concentration was very low, varying within a narrow range (0.11 – 0.27 mgL^{-1}), and also covarying with both pH and SO_4^{2-} concentration, which are both known to influence the community structure in Lake Hazen (Ruuskanen et al., 2018). Furthermore, we observed no significant differences in community structure along the redox gradient ($P = 0.77$) or O_2 concentration ($P = 0.23$) – which is partly consistent with a previous study of Lake Hazen sediments that showed that the microbial community was not constrained by oxygen concentration, but that the redox gradient was associated with community structure (Ruuskanen et al., 2018). This discrepancy is however not unexpected, as the range of redox potentials of the samples in the current study (**Figure B3**) was much narrower than in the previous study. Furthermore, the O_2 concentration likely plays some role in the community structure and function in Lake Hazen sediments, but the gradient can be extremely steep (**Figure B3**; Ruuskanen et al., 2018), and differences could be only seen with comparisons of much thinner sediment horizons.

To assess the validity of these 16S-derived results, we turned to the 55 MAGs, which also showed no significant differences in terms of community structure among samples, or in terms of correlations between community structure and physicochemistry (all $P > 0.10$). Similarly, a clustering analysis was unable to fully separate the sites based on these 55 MAGs (**Figure B5a**), but not on the 16S rRNA contig data that exhibited more differences among sites (with identical settings as for the MAGs; **Figure B5b**); note that these clustering analyses are by nature however qualitative, as no statistical tests were performed. Finally, we saw no separation of the samples in clustering when using functional pathway data from the 55 MAGs (derived with MetaCyc; **Figure B6**). This was also likely due to the small differences in the community composition of the two sites. As the communities appeared to be similar at both sampling sites, we pooled the data from each site together for all downstream analyses, in order to assess the extent of phylogenetic and potential metabolic microbial diversity in those lake sediments.

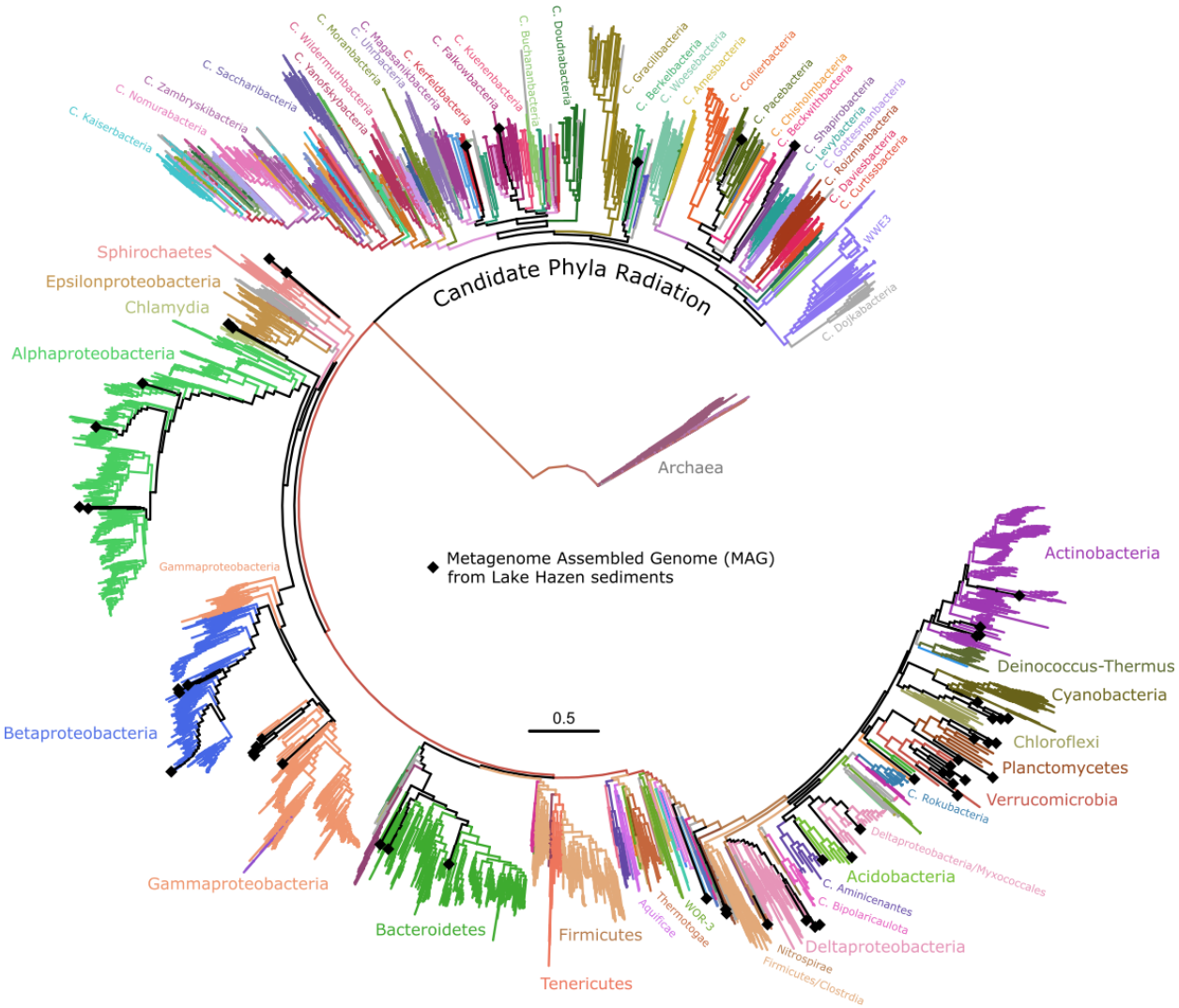

Figure 3.2: Phylogenetic tree of MAGs aligned against reference genomes annotated at phylum-level based on the NCBI taxonomy database. Genomes identified in NCBI as belonging to each phylum are indicated with different colors, which are matching those in **Figure 3.1**. Black diamonds indicate the phylogenetic placement of the MAG assembled in the current study. Scale bar are in unit of number of substitutions per site. Full tree with SH-like node support values is included as supplementary files (**Appendix D:** Full_tree_with_supports.pdf and Full_tree_with_supports.nwk).

3.4.3 Lake Hazen sediments harbor phylogenetically diverse bacteria

To place the 55 MAGs in a phylogenetic and taxonomic context, we reconstructed a tree adding 5,942 reference genomes to our MAG collection (**Figure 3.2**). The analysis showed that five of the MAGs (9%) likely represented CPR bacteria from previously known, NCBI-classified, candidate phyla (namely: Kerfeldbacteria, Falkowbacteria, Berkelbacteria, Pacebacteria and Shapirobacteria). This result highlights the importance of sequencing samples from rarely studied environments, such as High Arctic lakes. Two of these MAGs displayed similar characteristics to previously studied CPR (Hug et al., 2016), such as small genome size (< 1.3 Mbp) and a low number of genes (1309 and 1255 open reading frames; **Table D4**). In addition, three MAGs could not be classified to any known phylum by either their ribosomal proteins, or their partial 16S rRNA gene sequences (SH-like aLRT support < 50%). We note however that among these three unclassified MAGs, LH_MA_65_9 was likely related to uncultured bacteria close to candidate division OP11 (**Figure B7**), and LH_MA_57_9 was likely related to bacteria close to candidate division OP10 (**Figure B8**), a division mostly sampled from lake sediments. For example, the sediments of Upper Mystic Lake (Massachusetts, USA, www.ncbi.nlm.nih.gov/nuccore/DQ166697.1) contained the closest match to LH_MA_65_9 (99% BLAST identity).

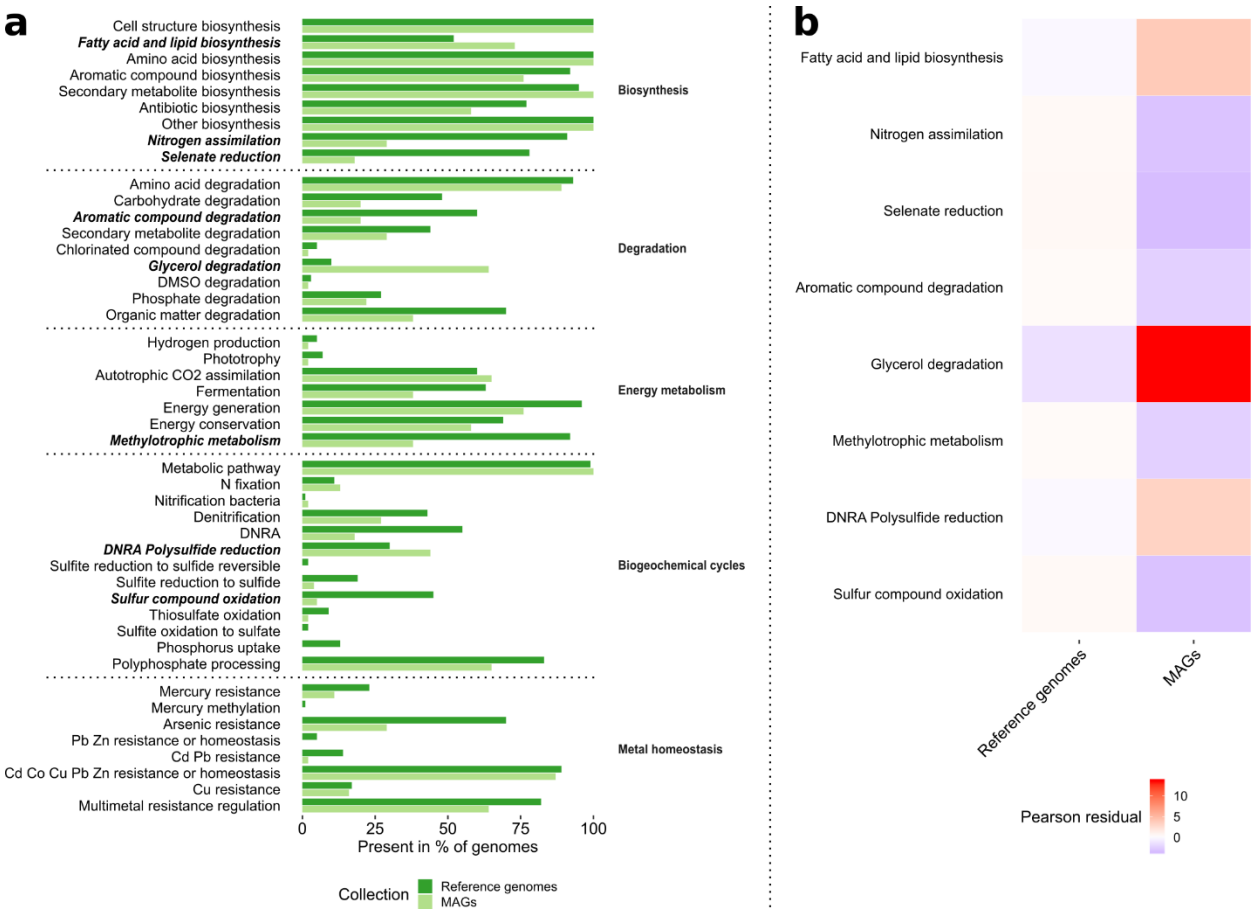

Figure 3.3: Cellular processes and metabolism in reference genomes and MAGs. (**a**) Percent representation of genomes in each data set (reference genomes and MAGs) with at least one marker pathway or gene for each process or metabolism. Names of pathways or processes with the most differentially abundant presence in MAGs compared to the reference genomes subset to common phyla are indicated in bold italics. (**b**) The most differentially abundant processes according to the Pearson's χ^2 test ($P = 0.001$) as a heatmap of their Pearson residual scores.

3.4.4 Lake Hazen MAGs are enriched in genes to strive at low temperatures and oligotrophy

To test if these reconstructed genomes from Lake Hazen possess unique metabolic features, we quantified the presence of select pathways in both the MAGs and a set of reference genomes at the same taxonomic rank (phylum level; n = 2,486 genomes), and compared their prevalence (**Figure 3.3**). In particular, we found that the marker genes and pathways for cellular metabolism and nutrient cycling (**Table D5**) were significantly different (Pearson's χ^2; X^2 = 361.25 ; P = 0.0001 based on 10,000 permutations).

Among the most different pathways, three were overrepresented in the MAGs (**Figure 3.3b**). Among those three, glycerol (glycerophosphodiester) degradation and fatty acid / lipid biosynthesis were more prevalent in the MAGs (**Figure 3.3b**). This is likely linked with temperature tolerance and energy conservation strategies. Indeed, both cold stress (Chintalapati et al., 2004) and starvation (Lever et al., 2015) can induce changes in the lipid composition of microbial cell membranes. A high prevalence of fatty acid desaturases was also recently seen in several Antarctic lake metagenomes (Koo et al., 2018), suggesting that these pathways are important for psychrotrophic microbes – given that water temperatures are around 3.5°C below a depth of 50 m (St. Pierre et al., 2019b). Alternatively, triacylglycerols could be utilized for energy storage in the form of lipid droplets (Alvarez and Steinbüchel, 2002). In addition to energy storage, lipids, in particular when present as droplets, might also play a role in regulating the stress response in bacteria (Zhang et al., 2017). Both of these roles could be beneficial to the bacteria harboring them in the Lake Hazen sediments, as most nutrients and oxygen are delivered to the lake in glacial meltwater during summer (St. Pierre et al., 2019b). This short period of higher nutrient and oxygen availability correlates with a temporary jump in microbial activity

(St. Pierre et al., 2019b), whose energy stores might be triacylglycerols that are then gradually released to maintain metabolism during the long winter.

The third overrepresented pathway, Dissimilatory Nitrite Reduction to Ammonia (DNRA) / Polysulfide reduction, shows that the MAGs are also enriched in membrane-linked molybdopterin oxidoreductases genes of the NrfD/PsrC family that includes genes coding for tetrathionate-, dimethyl sulfoxide-, polysulfide-, and nitrite reductases (Jormakka et al., 2008). It is more likely that these genes in MAGs (matching the annotation Pfam 03916) are related to sulfur reduction than DNRA, because the more specific markers for DNRA, such as nrfH, nirB and nirD, were rarer in the MAGs than in the reference genomes (**Figure 3.3a**). Furthermore, the same marker gene (nrfD) has been found for instance in metagenomes from sediments of saline (Ferrer et al., 2011) and freshwater lakes (Lin et al., 2011), and associated there with anaerobic respiration with oxidized sulfur compounds. The nrfD gene is thus likely important also for anaerobic respiration in the mostly anoxic sediments of Lake Hazen.

Underrepresented pathways in the Lake Hazen sediment MAGs (**Figure 3.3b**) might be underutilized and thus subject to genome streamlining, which is common in oligotrophic organisms (Giovannoni et al., 2014). These pathways include, for example, nitrogen assimilation as NO_3^- and NH_4^+. Likely, because of the low concentration of inorganic nitrogen in Lake Hazen (St. Pierre et al., 2019b) it might be directly assimilated only by a small number of organisms. The majority of microbes might instead resort to recycling of existing organic nitrogen compounds or nitrogen fixation to fulfill their needs for this nutrient (**Figure 3.3**). Similarly, aromatic compound degradation, methylotrophic metabolism and sulfur compound oxidation (largely desulfonation) were rarer in the MAGs than reference genomes. The lower prevalence of these pathways might also be a consequence of low nutrient and substrate availability in Lake

Hazen and the low number of methanogens (this study, Emmerton et al., 2016; Ruuskanen et al., 2018; St. Pierre et al., 2019b). In addition, selenate reduction was underrepresented in the MAGs. This pathway is usually important for anaerobic metabolism (Staicu and Barton, 2017), but in Lake Hazen its lower prevalence could similarly reflect the low availability of selenium because the site is distant from anthropogenic sources (*e.g.*, Chapman et al., 2010). It should be noted, that marker genes used to identify the aforementioned pathways may also have been missed due to technological biases, on our focusing the analysis on only the 55 most complete MAGs, and / or their incompleteness. Despite these potential biases however, the underrepresentation of these specific pathways in the MAGs makes sense in the light of the conditions in Lake Hazen sediments. Indeed, Lake Hazen is rapidly changing with the increased delivery of sediment, nutrients, organic carbon and contaminants, perpetuated by enhanced glacial melt throughout the watershed (Lehnherr et al., 2018). The dense and turbid glacial waters entering Lake Hazen flow directly to the bottom of the lake (St. Pierre et al., 2019b), and this might effectively lead to the disappearance of oligotrophic ecological niches in the sediments with increased productivity.

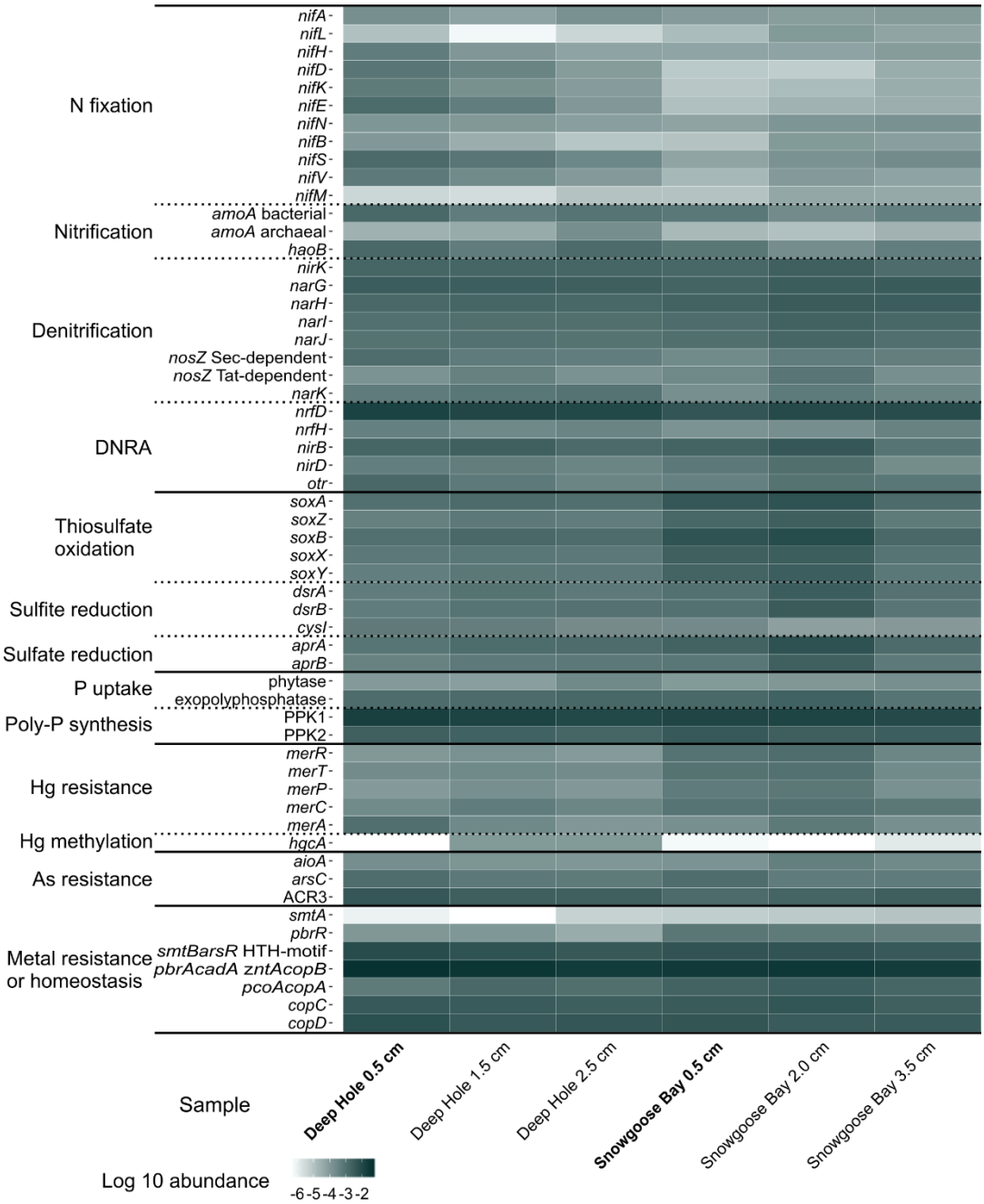

Figure 3.4: Normalized read coverages of individual marker genes from the whole metagenomic data across the samples. Samples from oxic environments (O_2 concentration > 0.0 mg L^{-1}) are indicated in bold. Differences in abundances of individual marker genes between the two sites (Deep Hole and Snowgoose Bay) and oxygen levels (oxic and anoxic) were not significant (pairwise t-test; all $P > 0.8$ after FDR correction). Gene family/domain annotations used to identify the individual marker genes are shown in **Table D5**.

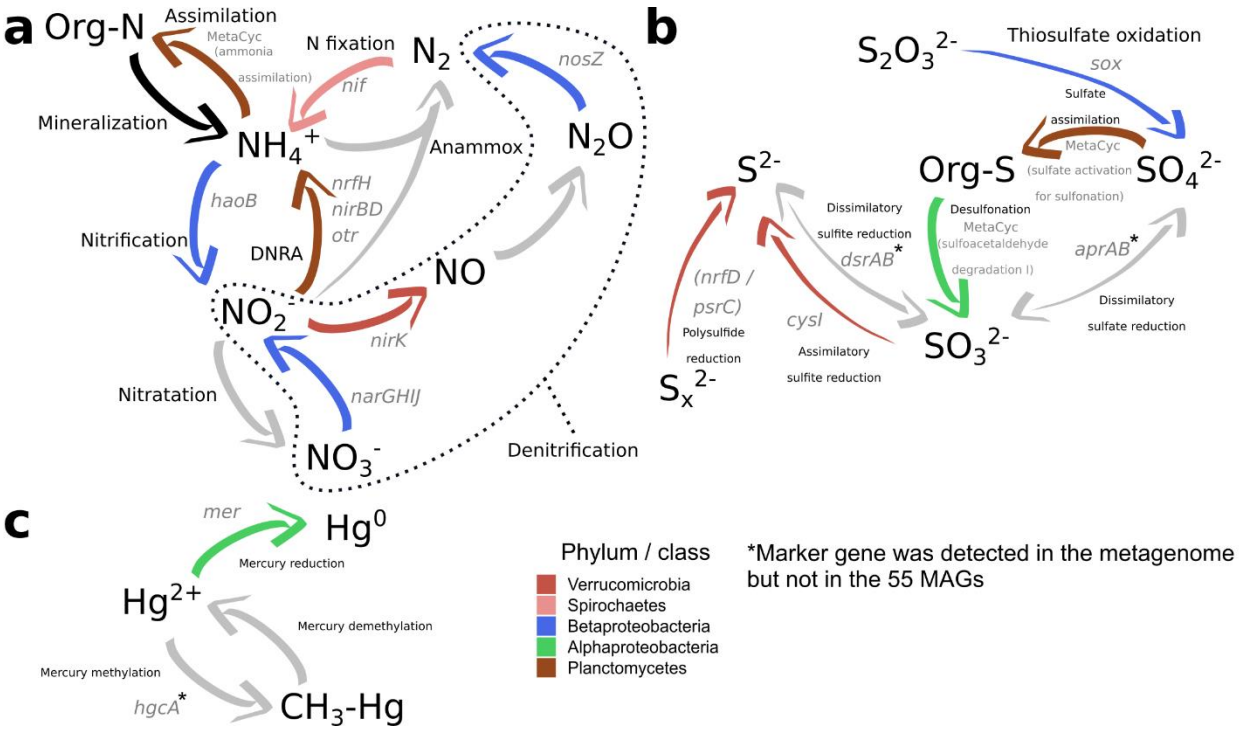

Figure 3.5: Biogeochemical cycling in the Lake Hazen sediment. Most common phyla participating in each process are indicated. The phylum Proteobacteria is here subdivided into classes. Marker gene symbols or 'MetaCyc' for marker pathways are indicated in gray text next to each process. Gray arrows denote processes that were not found in the 55 MAGs. If the gene is not indicated, a specific marker for the process could not be identified in the annotation. (**a**) Nitrogen cycle. DNRA: Dissimilatory Reduction of Nitrite to Ammonia. (**b**) Sulfur cycle. (**c**) Mercury cycle.

3.4.5 Spatial homogeneity of nutrient / toxic metal cycling in Lake Hazen sediments

Finally, to investigate more closely nutrient and toxic metal cycling in the sediments and evaluate the possible contribution of different microbes to these cycles, we identified marker genes or pathways for these relevant processes. However, because the most parsimonious pathways are calculated using marker genes present in a single genome bin, it is possible that

low quality bins could have strongly biased pathway inference (Ye and Doak, 2009). To alleviate this issue, we analyzed differences in abundances of all individual marker genes found in all assembled contigs in the metagenome longer than 1 kbp – and not only the 55 MAGs – using gene-level normalized read coverages summarized per each marker gene (**Figure 3.4**).

Comparing within and between sites, as well as oxic (surface samples; < 0.5 cm) vs. anoxic samples (deeper samples are completely anoxic at both sites (**Figure B3**), we found that none of the individual marker genes for the nutrient cycles were differentially abundant between the two sites (pairwise *t*-tests, all $P > 0.9$ after FDR correction) or between the oxic and anoxic sediment horizons (pairwise *t*-tests, all $P > 0.8$ after FDR correction). Thus, for the next analysis we averaged the normalized abundances of the 55 MAGs across all six samples (the two sites and three depths) to identify the likely key organisms in the nitrogen, sulfur, and mercury cycles – which improved our ability to investigate these more fine-grain patterns of nutrient cycling.

Based on these average read coverages, we identified the most abundant MAGs that included markers for nitrogen, sulfur and mercury cycling (**Table D6**), as summarized in **Table D7** for each ecologically important process in the nutrient cycles and shown in **Figure 3.5**. Proteobacteria (mostly beta- and alpha-) were overall the most abundant phylum that had marker genes or pathways for nitrogen and sulfur cycling, but other phyla were often the most abundant when looking at individual processes (**Appendix B**; **Table D7**). The metagenome also contained markers for the full sulfur cycle, although the genes for dissimilatory sulfate reduction to sulfite (*aprAB*) and dissimilatory sulfite reduction to sulfide (*dsrAB*) were not detected specifically in the 55 MAGs (**Appendix B**; **Figure 3.5b**). While no marker genes for mercury methylation (*hgcA*; Pfam 3599) were found in the MAGs (but were present in 32 reference genomes and 27 low quality bins from the metagenome), six reconstructed genomes harbored *mer*-operon

genes involved in mercury resistance, with Alphaproteobacteria as the most abundant of these (**Figure 3.5c**). The nitrogen and sulfur cycles in the Lake Hazen sediments appeared to be closely intertwined, catalyzed by taxonomically diverse lineages. Certain MAGs, such as LH_MA_37_3, likely from *Geobacter*, and LH_MA_65_9, might represent in the lake ecosystem highly important organisms that can fix dissolved nitrogen gas, and are thus capable of returning it back into the biological cycles. Furthermore, other MAGs, such as LH_MA_55_1, might represent microbes that play roles simultaneously in both sulfur and nitrogen cycles (**Appendix B**). These results are consistent with our previous work that, based on 16S rRNA gene amplicon data, predicted the presence of key functional groups such as aerobic ammonia oxidizers (LH_MA_61_7, likely from Nitrosomonadales), nitrate reducers (*e.g.*, LH_MA_28_10, likely from *Rhodoferax*) and sulfate reducers (presence of *aprAB* genes in the metagenomic dataset) in the Lake Hazen sediments (Ruuskanen et al., 2018).

3.5 Conclusions

We show that in addition to being unique in its location (81°N), dimensions, and volume, Lake Hazen hosts a phylogenetically diverse set of microbes whose reconstructed genomes contain a high prevalence of pathways that make them fit for thriving in a cold (~3.5°C) and oligotrophic environment. This diversity includes organisms from recently discovered groups, such as the Candidate Phyla Radiation, and from uncultured branches of the tree of life. Because metagenomic surveys from arctic lake sediments are currently scarce, our results provide the scientific community with a compositional baseline and functional potential of these ecosystems. Indeed, as the oligotrophic niches that extant microbes inhabit are likely to be affected by ongoing environmental changes in Lake Hazen, changes in the microbial community at the base of the lake ecosystem might have unforeseen consequences, with possible repercussions leading even to higher trophic levels.

3.6 Acknowledgements

We would like to thank the Poulain and Aris-Brosou lab members at University of Ottawa, and members of the Virta and Hultman labs at University of Helsinki for their helpful comments during data analysis and writing of the manuscript. We would also thank the Editors and two anonymous reviewers for their comments, which greatly improved the quality of this study. This work was funded by the Natural Resources Canada / Polar Continental Shelf Project (VStL, AJP), the Natural Sciences and Engineering Research Council of Canada (VStL, SAB, AJP), the ArcticNet Centre for Excellence (VStL, AJP), the Canadian Foundation for Innovation (SAB, AJP) and the Finnish Academy of Science and Letters / the Vilho, Yrjö, and Kalle Väisälä Fund (MR).

Chapter 4: Demographics of the microbial mercury reductase coincide with increasing anthropogenic mercury in the Northern Hemisphere

Matti O. Ruuskanen[1], Stéphane Aris-Brosou[1,2] and Alexandre J. Poulain[1]

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

[2]Department of Mathematics & Statistics, University of Ottawa, Ottawa, ON, Canada

Manuscript is currently under review at the ISME Journal: Multidisciplinary Journal of Microbial Ecology, revisions requested by October 2019 (as of September 23, 2019).

Author contributions: MOR designed the experiments together with AJP and SAB. MOR performed most of the sampling (some samples were contributed), laboratory work, data analysis, and drafted the manuscript. AJP and SAB supervised the project. All authors contributed to the writing and accepted the final version of the manuscript.

4.1 Abstract

Anthropogenic mercury remobilization has considerably increased since the Industrial Revolution in the late 1700s. The Minamata Convention on Mercury is a United Nations treaty (2017) aiming at curbing mercury emissions. Unfortunately, evaluating the effectiveness of such a global treaty is hampered by our inability to determine the lag in aquatic ecosystem responses to a change in atmospheric mercury deposition. Whereas past metal concentrations are obtained from core samples, there are currently no means of tracking historical metal bioavailability or toxicity. Here, we recovered DNA from dated sediment cores collected in Canada and Finland and reconstructed the past demographics of the microbial mercuric reductase (MerA) - an enzyme involved in Hg detoxification - using Bayesian relaxed molecular clocks. We found that the evolutionary dynamics of *merA* exhibited a dramatic increase in effective population size starting from 1783.8 ± 3.9 CE, which coincides with both the Industrial Revolution, and with independent measurements of atmospheric Hg concentrations. We show that even low levels of anthropogenic mercury affected the evolutionary trajectory of microbes in the northern hemisphere, and that microbial DNA encoding for detoxification determinants stored in environmental archives can be used to track historical pollutant toxicity.

4.2 Introduction

Mercury (Hg) is a naturally occurring toxic metal that is globally distributed because of the volatility of its reduced form (Driscoll et al., 2013). Under anoxic conditions, microbes can transform inorganic Hg into methylmercury, which is bioaccumulated in organisms and biomagnified throughout food webs. Hg is naturally remobilized from geological sources but anthropogenic emissions have dramatically increased since the Industrial Revolution, which took place around the end of the 18th century (Selin, 2009). As a consequence of anthropogenic activity, the concentration of Hg in the atmosphere is estimated to be almost three times higher than in pre-industrial times, and about eight times that of 2000 BCE, when Hg started being used by human civilizations (Amos et al., 2013). These estimates are based on biogeochemical models, backed by historical sources (Nriagu, 1994), archeological evidence (Brooks, 2012), and measures of total Hg in sediment cores (Cooke et al., 2009; Elbaz-Poulichet et al., 2011). However, historical response of ecosystem components to Hg has mostly relied on preserved museum specimens (Vo et al., 2011), because direct tracking of Hg toxicity and bioavailability in environmental archives, such as ice and sediment cores or permafrost, is currently impossible.

Even though human activities have contributed to increase the amount of Hg in the environment, we do not know how historical Hg deposition has affected key microbial players which often control the amount of Hg available to food webs (Barkay and Wagner-Dobler, 2005). This is a major knowledge gap that potentially hinders policy development, because effective risk reduction strategies depend on a comprehensive understanding of the present and past effects of toxic pollutants on ecosystems (Hsu-Kim et al., 2018). One possibility to track bioavailable Hg through time is to monitor how microbial systems responded to historical toxic Hg levels. The best-known microbial Hg detoxification mechanism, the *mer*-operon, encodes

proteins that efficiently detect, transport and reduce organic and inorganic forms of Hg to its volatile form, Hg^0, which then diffuses out of the cell (Barkay and Wagner-Dobler, 2005; Mathema et al., 2011b). It is thought that the *mer*-operon evolved billions of years ago, in marine hydrothermal environments under strong geogenic Hg pressure, with subsequent constraints by light, salinity and redox conditions shaping the evolution of the mercuric reductase MerA (encoded by the *merA* gene; Boyd and Barkay 2012). *MerA* is (i) specific to Hg detoxification (Barkay et al., 2003), (ii) omnipresent in the environment (Barkay et al., 2010; Osborn et al., 1997), and (iii) easily exchanged between coexisting microbial populations via horizontal gene transfers (Barkay et al., 2003). Therefore we can expect that *merA* variants be maintained in a given environment based on the selective advantage these variants provide to the community, rather than segregating based on their taxonomy (vertical inheritance). As such, *merA* is an ideal candidate to provide molecular insights into historical exposure to Hg. Previous results suggested an association between *merA* phylogeny and Hg deposition in a remote region (Poulain et al., 2015), but this work was limited in its spatial scope and by the lack of variation in historical Hg deposition. To address these shortcomings, we tested if the evolutionary response of *merA* can be used as a proxy for historical changes in the toxicity, and hence bioavailability, of Hg in sedimentary archives collected over a much broader spatial scale.

Here, we show that regardless of site or deposition rate, the local effective population sizes of *merA* increased at an unprecedented rate starting at the beginning of the Industrial Revolution in the Northern Hemisphere. No such response was observed for our control housekeeping gene, *rpoB*, encoding for the β subunit of the bacterial RNA polymerase. Our results imply that even small changes in Hg loadings can have long-lasting impacts on the evolutionary dynamics of microbes detoxifying this pollutant, and that the *merA* gene is a

sensitive maker, that can be used to track historical Hg bioavailability over broad, continental scales.

4.3 Results and discussion

4.3.1 Total mercury concentrations affect *merA* diversity but not its abundance

To understand how microbes responded to different levels of anthropogenic Hg deposition, we collected sediment cores from eight lakes in Canada and Finland, as well as a river in Finland (**Figure 4.1**). We first tested if the abundance of *merA* (quantified with droplet digital PCR) correlated with total mercury concentration ([THg]), sediment deposition date (using radiometric and other dating techniques; see Methods), or sampling depth (from sediment surface) in these sediment cores (**Figure C1**). We assessed these results against a single-copy housekeeping gene, *glnA*, as our first control gene. We found that the copy numbers of *merA* did not change with sediment depth, nor the estimated deposition date, or even with [THg] (**Figure C2a**). However, the number of *glnA* copies, our control gene, decreased with sediment depth (Figure S2B-C). These results contrast with previous studies reporting an increase in *merA* copy numbers over long-term (up to 90 years) industrial Hg contamination of various soils (Frossard et al., 2018) or with bioavailable Hg in snowpacks (Larose et al. 2013; **Figure C3**). However, such studies focused on spatial variation, and did not address changes over time or geochemical gradients – at each location.

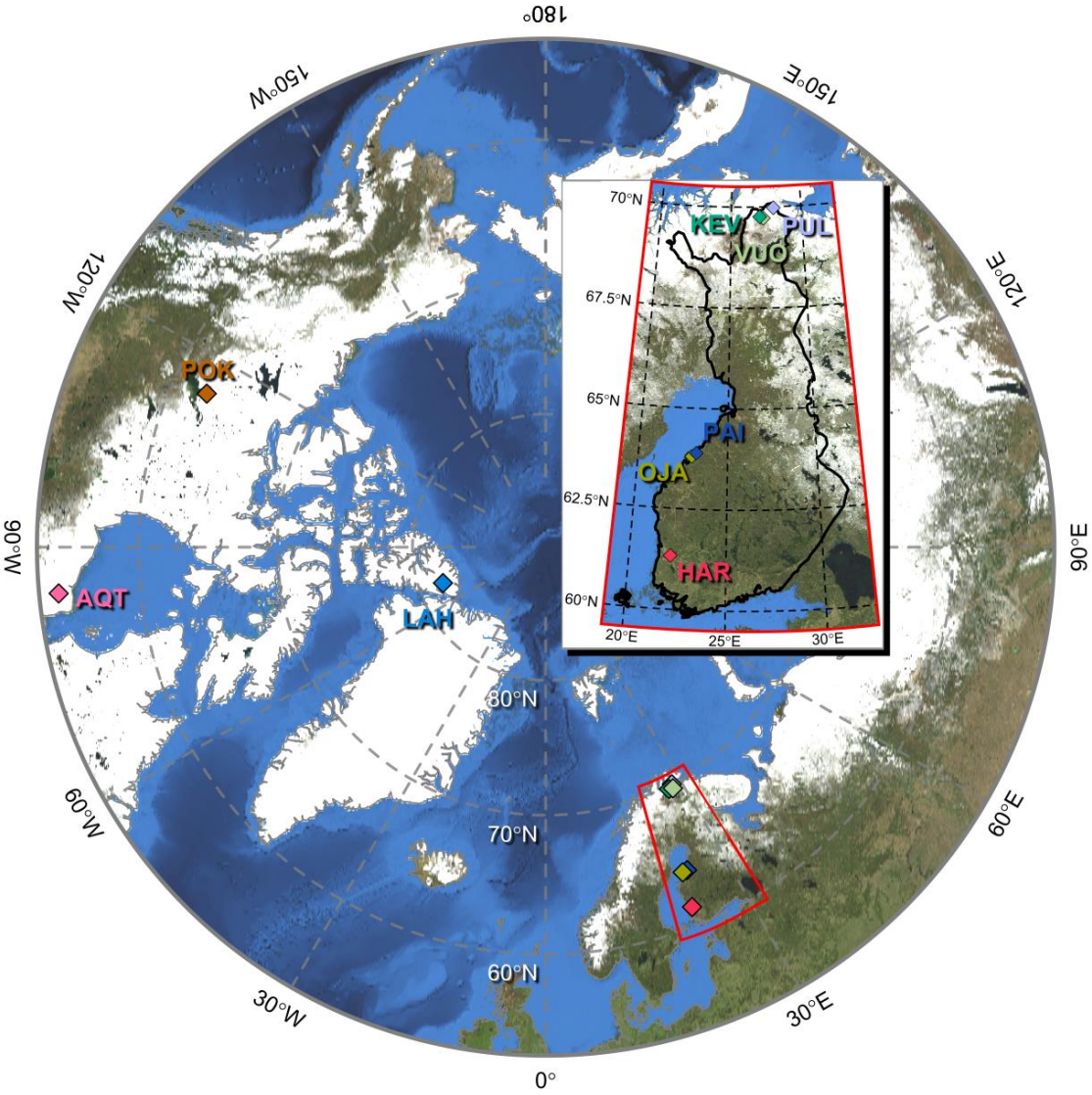

Figure 4.1: Locations of sampling sites in Finland (blown up in inset) and northern Canada. By alphabetical order: AQT: Lake Aquatuk; HAR: River Kokemäenjoki; KEV: Lake Kevojärvi; LAH: Lake Hazen; OJA: Lake Öjanjärvi; PAI: Lake Päiväjärvi; POK: Pocket Lake; PUL: Lake Pulmankijärvi; VUO: Lake Vuolimuš Cieskuljávri.

Given the lack of variation in the abundance of *merA* along [THg] gradients, there could instead be changes in its genetic diversity, suggesting adaptation to increasing Hg selective pressure. To address this question, we compared the diversity of *merA* sequences among samples (beta-diversity), both over varying [THg] and temporal gradients in the sediment cores. In this case, we used an additional control gene, the single-copy housekeeping gene *rpoB*, encoding for the RNA polymerase β subunit, to provide more accurate evolutionary information; indeed, *glnA* amplicons proved too short (< 156 bp) for this subsequent step of the analysis. These beta-diversity analyses showed that *rpoB* variants were similar among sites ($P = 0.09$) and countries ($P = 0.88$). On the other hand, *merA* variants differed significantly among sites ($P = 0.01$), while countries showed similar variational responses ($P = 0.30$; **Figure 4.2**). This pattern suggests that *merA* diversity is affected by [THg] gradients, contrary to *rpoB*. To confirm this observation, we fitted General Additive Models (GAM) of [THg] and dating on the ordinations of *merA* and *rpoB* diversity. All these fits were significant ($P < 1 \times 10^{-17}$), with *merA*'s beta-diversity correlating linearly with both [THg] and temporal (dating) gradients. The diversity of *rpoB* failed to show any significant response to any of these gradients, only exhibiting phylogenetic similarity at the sediment surfaces (**Figure 4.2**). These contrasted results suggest that the two genes, *merA* and *rpoB*, have responded differently to increase Hg deposition over time. As *merA* encodes for a protein (MerA), whose sole known function is to detoxify Hg, it can be posited that *merA*'s evolutionary trajectory could have been affected by historical changes in Hg deposition, at least in the Northern Hemisphere.

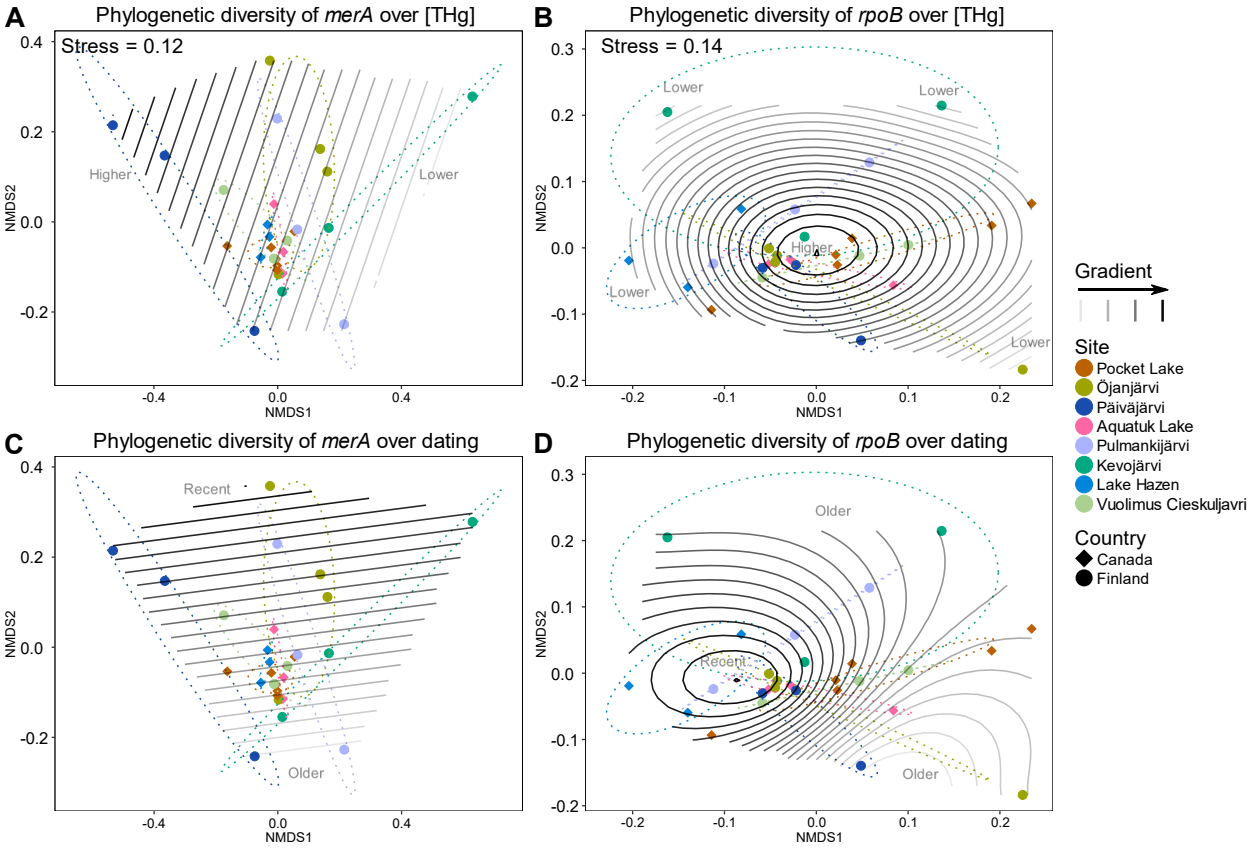

Figure 4.2: Beta-diversity of *merA* and *rpoB*. Correlations of the NMDS ordination with [THg] (**a**, **b**) and dating (**c**, **d**). Samples from the same site are color-coded and grouped with ellipses. Stress values for the NMDS ordinations are indicated in panels **a** and **b**, and are identical for **c** and **d**, respectively. All fits of the general additive models for [THg] and dating (shown as gradients) were significant ($P < 0.01$).

4.3.2 Swift evolutionary response of *merA* from the onset of the Industrial Revolution

To test this hypothesis, we reconstructed both the phylogeny and past demographics of microbes carrying *merA* genes with Bayesian relaxed molecular clocks (Drummond et al. 2006). First, the trees constructed for *merA* showed a significant association ($P < 0.01$) with the sediment horizon (*i.e.*, depth/date, as a categorical variable) in all lakes except for Vuolimus Cieskuljavri, which may be due to its low [THg] (max [THg] = 50.4 ng g^{-1} dw; **Figure 4.3**). This clustering of phylogenetic clades by dates suggests a rapid evolutionary turnover, where *merA* gene variants are replaced from one horizon to the next. Interestingly, such a high turnover was however not observed for our control gene, *rpoB*, where this association between sediment horizon and phylogeny was significant in only three out of eight (38%) sites.

Even if the significance of this gene-specific association of clades with dates cannot be rejected (Binomial test, $P = 0.73$), *rpoB* is considered to be a reliable phylogenetic marker gene in bacteria (Case et al., 2007) and such associations might reflect the structuring of the microbial communities in sediments in response to physical and geochemical gradients (Ruuskanen et al., 2018). Furthermore, we noted that community structure beta-diversity patterns (inferred with *rpoB*) differed from that of *merA*'s. This lack of a clear signal prompted us to go beyond a mere phylogenetic assessment, and to further evaluate how changes in [THg] have affected the evolutionary dynamics of the two genes.

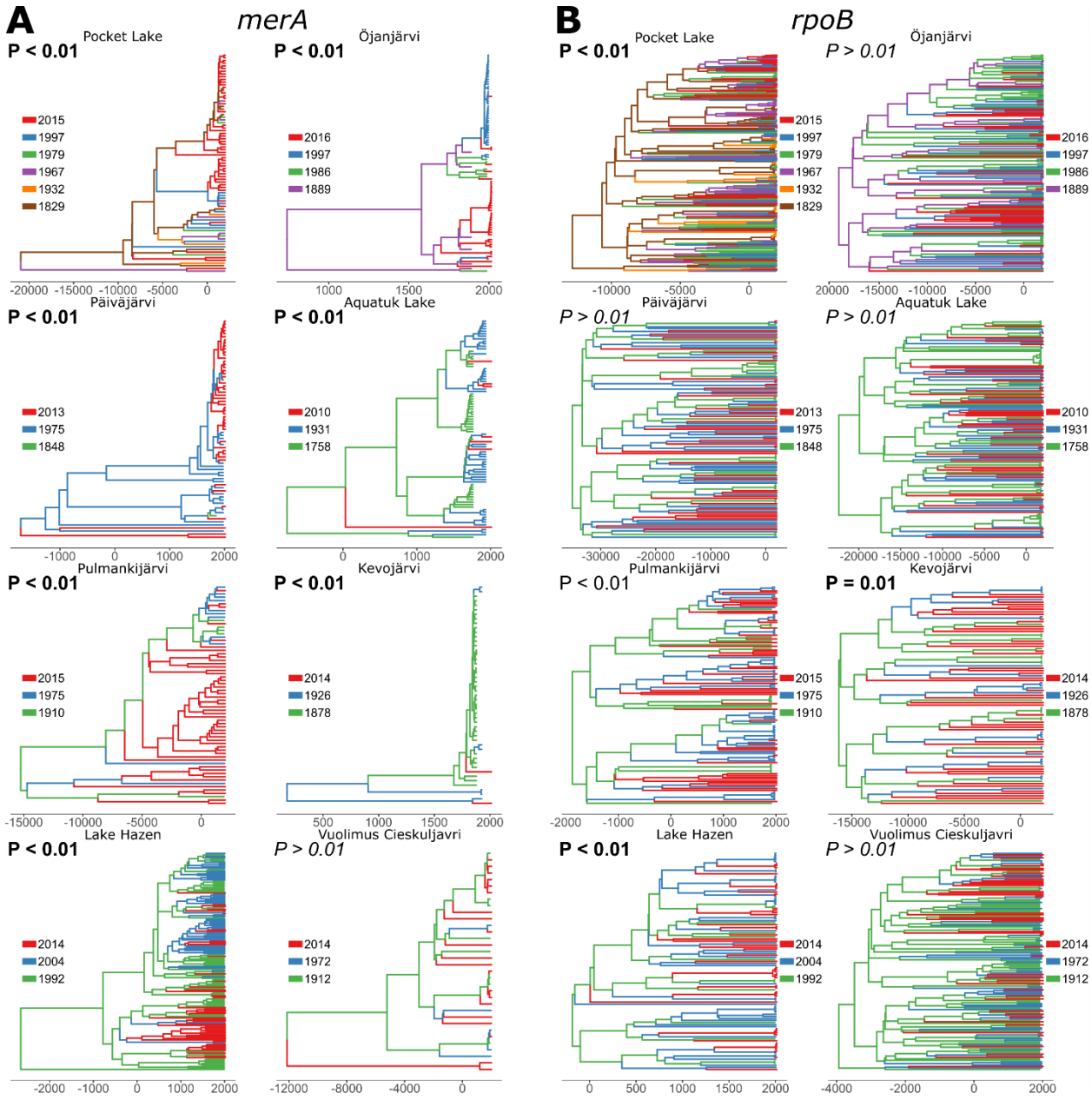

Figure 4.3: Maximum *a posteriori* trees from the relaxed molecular clock analyses. Trees are shown for *merA* (**a**) and *rpoB* (**b**) at each site. Sampling horizons were color-coded as shown next to each tree and are identical across trees for the topmost three horizons. The approximate dates affixed to the sampling horizons were estimated externally. The *x*-axes are in unit of CE years, and resultant of the BEAST analysis calibrated with the external dating.

Our reconstructions of *merA* demographics based on Bayesian skylines showed sigmoidal increases in the scaled effective population sizes (N_e) of this gene over time at most sites, matching the trends in [THg] (**Figure C4a**). In contrast to *merA*, the N_e of *rpoB* varied widely (**Figure C4b**) and, critically, did not seem to coincide with any trends in [THg] at any of the sites, which is consistent with the results of our GAM analyses (**Figure 4.2**). The magnitude of the response of *merA* to increased Hg deposition was likely related to changes in bioavailable Hg causing a positive selective pressure on *merA* (as hypothesized by Poulain et al., 2015), rather than solely to changes in [THg]. Indeed, the total concentration of a metal is rarely a good predictor of its toxicity, and this is true for Hg (Barkay et al., 2003). Rather, metal speciation, *i.e.*, the complexes that a metal forms with inorganic and organic ligands in a solution, affects how metals interact with living cells, causing toxicity (Hughes and Poole, 1991). Environmental archives only record [THg] but, as environmental conditions change over time, we can expect that microbes present in the water column or in surface sediments be exposed to different levels of Hg exhibiting varying toxicity. Furthermore, Hg bioavailability might differ among sites due to differences in local geochemistry affecting Hg speciation (*e.g.*, [dissolved organic matter]; Chiasson-Gould, Blais, and Poulain 2014; Mangal et al. 2019). For instance, brief decreases in median *merA* N_e at Öjanjärvi and Päiväjärvi coincided with peaks in [THg], whereas no changes in *merA* N_e were observed with a higher [THg] peak in Pocket Lake (**Figure C4a**).

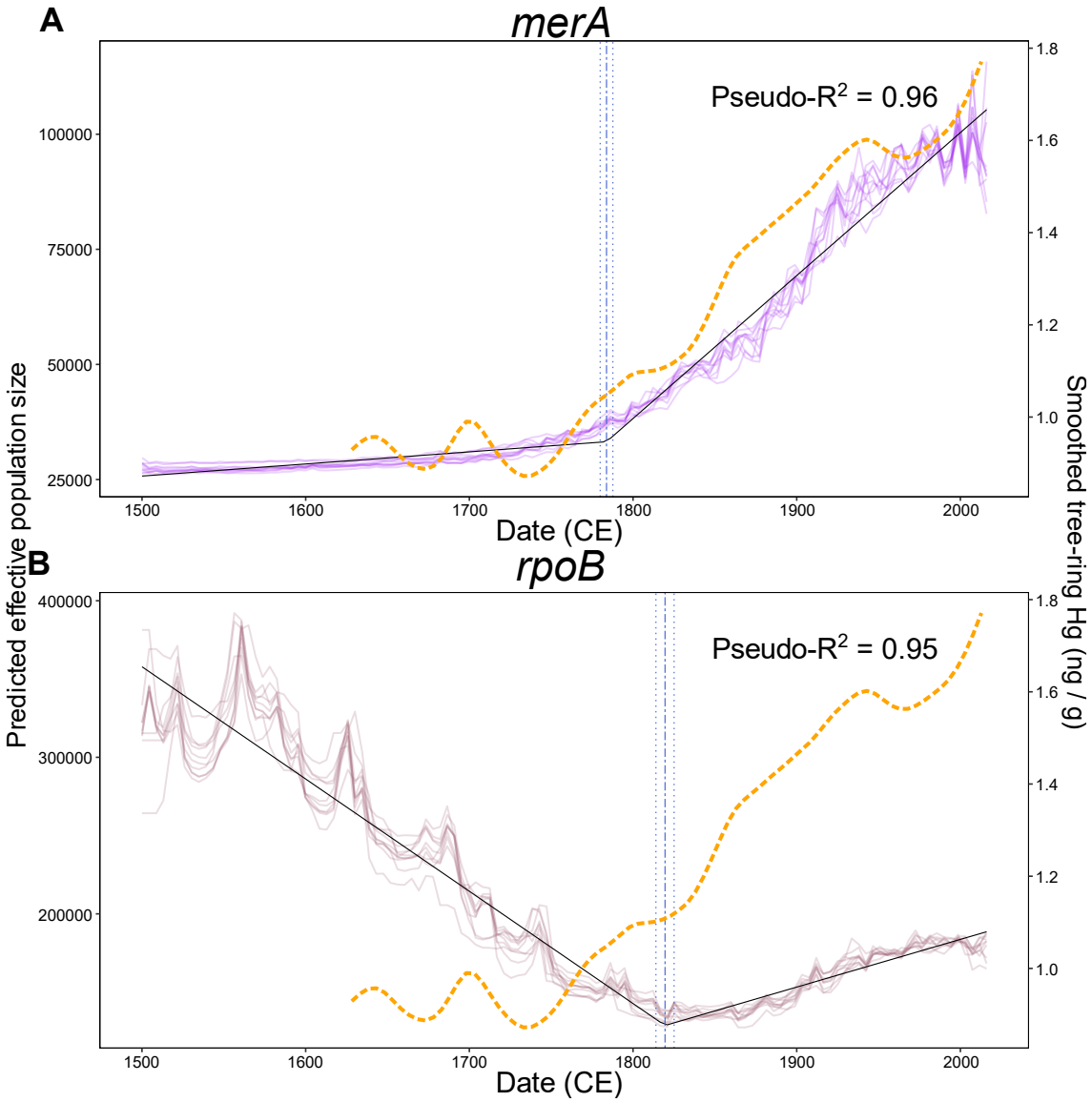

Figure 4.4: Partial dependence plots of predicted effective population size as a function of calendar date for *merA* (**a**) and *rpoB* (**b**). Results are shown from fitting ten randomized models for each gene. The pseudo-R^2 scores quantify average prediction accuracy of the models, which also included the sampling site as a variable. The black lines show the segmented linear model fits and the blue lines indicate their breakpoints (dot-dash lines) surrounded by their 99% CI (dotted lines). Dashed orange lines show approximate atmospheric Hg concentration and were redrawn using data by Clackett, Porter, and Lehnherr (2018), from their high-resolution analysis of tree-rings in subarctic Canada.

Finally, to average over these site-specific responses, we fitted random forest models to the reconstructed demographics for each gene, over all the sampled sites. We found that *merA*'s demographics initially exhibited a slow increase of N_e, followed by a sharp increase through time (pseudo-R^2 = 0.96; **Figure 4.4a**). Again, these trends were not observed for *rpoB* (pseudo-R^2 = 0.95; **Figure 4.4b**). To determine the dates at which these demographic dynamics changes occurred, we fitted segmented regressions of the partial dependence of N_e *vs.* date.

The highly variable and decreasing N_e of our control gene, *rpoB*, changed around 1819.6 ± 5.6 CE (99% CI; adjusted R^2 = 0.93) towards less variation and a slightly increasing trend. It is possible that the changes in the N_e of *rpoB* were caused by structuring of the sediment communities along geochemical gradients, as hypothesized earlier in our beta-diversity analysis. These gradients, such as redox potential, are known to affect microbial community structure in the sediments (Ruuskanen et al., 2018), and could impact the reconstructed N_e of *rpoB* because of its strong association with the phylogeny of the organisms (Case et al., 2007). However, the temporal trends we observed might also be connected to the increasing atmospheric pCO_2 and its delivery into freshwater ecosystems (Weiss et al., 2018). Increased pCO_2 concentration in the lakes and their watersheds affects microbes both directly and indirectly (Weiss et al., 2018; Yu and Chen, 2019), and could cause geographically widespread changes in the N_e of *rpoB*. Furthermore, the variability of the *rpoB* N_e also appears to have decreased since the early 19th century (**Figure 4.4b**), possibly mirroring the concurrent decrease in geographical heterogeneity of global climate conditions (Neukom et al., 2019).

Contrary to our control gene, the dynamics of N_e for *merA* changed around 1783.8 ± 3.9 CE (99% CI; adjusted R^2 = 0.98), date after which the rate of increase rose by an order of magnitude. The latest global increases in Hg deposition originated from the increased wood

burning in North America (Pérez-Rodríguez et al., 2018) at the end of the 18th century, and coal combustion fueling the Industrial Revolution in Europe, which began around 1760 (Allan et al., 2013; Stearns, 2018). This increase in anthropogenic Hg in the atmosphere was recently tracked using high-resolution tree-ring data in subarctic Canada (Clackett et al., 2018). We overlaid the smoothed data of atmospheric Hg concentrations from this study on our reconstructed demographics of *merA* (**Figure 4.4a**). The temporal coherence of *merA* N_e, derived solely using sediment DNA, and that of atmospheric Hg^0 levels obtained from tree rings, is striking.

Our results show that across our sites, located up to 5,500 km apart in the Northern Hemisphere, historical anthropogenic Hg emissions likely affected the evolutionary dynamics of *merA* – allowing us to track variations in Hg toxicity, and hence bioavailability, through time. The likely effects of bioavailable Hg on the demographic history of *merA* were rapid and long lasting, and highly sensitive as they were observed even at relatively low levels of Hg deposition. With implementation of the Minamata convention underway (Minamata Convention on Mercury), gaining insights into historical responses of biota to changes in Hg deposition is critical. Indeed, a recent study showed divergence between mercury levels in aquatic wildlife and atmospheric values, especially in the last two decades (Wang et al. 2019), that might be caused by climate change. The overriding and opposite effect that climate change may have on a reduction of Hg emissions requires that governments and environmental agencies worldwide be equipped with the means to accurately assess the efficiency of a reduction of Hg emissions on ecosystem health. We posit that our finding can be applied to any globally distributed contaminants for which microbes have evolved specific detoxification determinants.

4.4 Material and methods

4.4.1 Sampling and sample processing

Intact sediment cores were obtained from eight lakes in Canada and Finland, and a river in Finland, which were selected based on previous data on their Hg deposition history (**Table 4.1**, **Figure 4.1**). Low Hg contamination (max [THg] < 100 ng g^{-1} dw) sites included Lake Hazen (LAH) in NU sampled in May 2015 in the Canadian High Arctic (Köck et al., 2012), and Lake Kevojärvi (KEV), Lake Vuolimus Cieskuljavri (VUO), and Lake Pulmankijärvi (PUL) subarctic sites sampled in April 2016 in Finland (Rekolainen et al., 1986). To sample moderate Hg contamination (max [THg] from 100 to 1,000 ng g^{-1} dw) through long-range transport we obtained sediment-extracted DNA from a previous study of Aquatuk Lake (AQT), ON, Canada (Poulain et al., 2015), sampled in August 2010; this site also allowed us to validate our original findings using different approaches for quantification and sequencing. To represent freshwater sediments with a history of strong anthropogenic Hg contamination (max [THg] > 1,000 ng g^{-1} dw), we sampled Pocket Lake (POK), YT, Canada in July 2016, under the direct influence of toxic metal deposition from the Giant Mine Au roaster (Thienpont et al., 2016). Lake Öjanjärvi (OJA) and Lake Päiväjärvi (PAI) were sampled in May 2016 in Kokkola, Finland, as they are affected by metal deposition from the Ykspihlaja industrial park (Vuori, 2009). River Kokemäenjoki (KOK) was sampled at the dam reservoir in May 2016 in Harjavalta, Finland. The site has a history of strong Hg pollution from a chlor-alkali plant upstream and a nearby Cu/Ni/Au/Ag smelter (Schultz et al., 1995). Sizes of their watersheds (if not previously available) were calculated based on digital elevation models and online tools based on ArcGis (ESRI, Redlands, CA, USA; Ontario Ministry of Natural Resources and Forestry - Provincial Mapping Office 2013; Suomen metsäkeskus 2017). Sites in Finland were sampled with an HTH

sediment corer (Renberg and Hansson, 2008) (Pylonex AB, Umeå, Sweden) except for Kevojärvi, where a wedge type ice finger sampler (1.5 m) filled with a dry ice/coolant mixture was used to preserve the annual laminae in the sediment. At Lake Hazen and Pocket Lake, a UWITEC gravity corer (Mondsee, Austria) was used. Most of the cores were sectioned at 1 cm intervals in the field within hours of sampling, and were subsampled from the middle of the sections (avoiding the edges) with sterilized tools to minimize cross-contamination and edge deformations in the cores. The sections from Finnish cores were frozen on dry ice or in a -80°C freezer immediately after sectioning. The Pocket Lake core was sectioned after sampling with sterilized tools at 0.5 cm intervals, which were frozen at -20°C. The whole core from Lake Hazen was frozen at -18°C in the field and sectioned later at 1 cm intervals while frozen. Both Lake Hazen and Lake Kevojärvi sections were then subsampled with sterilized tools from the middle of the sections. No chemical preservatives were used for any samples, all sample containers were sterile, and bleach-sterilized powder-free nitrile gloves were worn throughout sample handling. After the initial freezing, the samples were shipped frozen, stored at < -18°C, and thawed only directly before DNA extraction.

Table 4.1: Geomorphological characterization of the sampling sites.

Peak [THg]	Site	Abbr.	Sampling coordinates	Lake area	Watershed area	Catchment area to lake area ratio [3]	Mean / max. water depth	References
2,739 ng g^{-1} dw	Kokemäenjoki / Harjavalta	HAR	61.33207°N 22.12895°E	NA (river)	26100 km^2	NA (river)	NA (river)	(Finnish Environment Institute SYKE, 2018b)
1,906 ng g^{-1} dw	Pocket Lake	POK	62.50897°N 114.37377°W	0.048 km^2	0.095 km^2	1.98	2 m /6 m	(Gibson et al., 1998; Reid, 1997)
1,057 ng g^{-1} dw	Öjanjärvi	OJA	63.80657°N 22.99994°E	12 km^2	[1]3970 km^2	331	1.6 m / 9 m	(Bone et al., 2016)
415 ng g^{-1} dw	Päiväjärvi	PAI	63.88923°N 23.24727°E	0.13 km^2	[2]0.95 km^2	7.31	Unk. / 1.8 m	(Finnish Environment Institute SYKE, 2018a)
110 ng g^{-1} dw	Aquatuk Lake	AQT	54.3281°N 84.5686°W	8.28 km^2	[2]231 km^2	27.9	Unk. / 12 m	(Rühland et al., 2014)
126 ng g^{-1} dw	Pulmankijärvi	PUL	69.97459°N 28.00415°E	11.2 km^2	800 km^2	71.4	11 m / 34 m	(Ilmast and Sterligova, 2002; Mansikkaniemi and Syrilä, 1965; Petäjä, 1964)
91 ng g^{-1} dw	Kevojärvi	KEV	69.75804°N 26.99785°E	1.1 km^2	1470 km^2	1340	11.1 m / 35 m	(Kuusisto, 1981)
63 ng g^{-1} dw	Lake Hazen	LAH	81.8245°N 70.71604°W	540 km^2	6860 km^2	12.7	95 m / 267 m	(Emmerton et al., 2016; Köck et al., 2012)
50 ng g^{-1} dw	Vuolimus Cieskuljavri	VUO	69.73197°N 27.09612°E	0.39 km^2	[2]22.5 km^2	57.7	Unk. / 1.9 m	(Pike, 2013)

[1] Watershed area combined with hydrologically connected Luodonjärvi (area 73 km^2, mean depth 2.6 m, max depth 11m).

[2] Estimated in this study.

[3] Calculated based on lake and watershed area.

4.4.2 DNA extraction, chemistry and dating

Samples from the cores were homogenized, and DNA extraction was done in duplicate from 0.25 g (wet weight) subsamples of each section. The subsamples were first washed with a buffer (10 mM EDTA, 50 mM Tris-HCl, 50 mM Na2HPO4·7H2O at pH 8.0) to remove PCR inhibitors (Poulain et al., 2015; Zhou et al., 1996). Environmental DNA, consisting of nucleic acids from alive, dormant and dead organisms in the sediment, was extracted with the MOBIO PowerSoil® DNA Isolation Kit (Carlsbad, CA, USA) and the duplicate extracts were combined. Duplicate negative control extractions were made from the wash buffer to assess kit contamination and combined after extraction.

For chemistry and dating, the core horizons were freeze dried and their water content was measured. THg data for the AQT samples were obtained from a previous study (Poulain et al., 2015), and in other cores it was quantified based on thermal decomposition, gold amalgamation, and atomic absorption with a MA3000 mercury analyzer (Nippon Instruments Corporation, College Station, TX, USA) at the Laboratory for the Analysis of Natural and Synthetic Environmental Toxins (LANSET, University of Ottawa, ON, Canada; see SI archive for primary measurements). Marine sediment certified reference materials MESS-3 and MESS-4 (National Research Council of Canada), and Buffalo River Sediment NIST-2704 (National Institute of Standards and Technology) were used as calibration controls for [THg] measurements.

Annually laminated (*i.e.* varved) sediments deposit in Kevojärvi, which enabled dating these sediments in high resolution by varve counting. The dating for the Kevojärvi core in the current study (KEV) was calibrated to a parallel core (KEVO-1) dated earlier by a highly resolved ^{137}Cs profile, where the peaks from the radioactive fallout from Chernobyl accident in 1986 and the 1960's nuclear testing were used to adjust the varve chronology (E. Haltia, University of Turku, personal communication, February 2019). The ^{210}Pb and ^{137}Cs dating of the other cores for the current study (VUO, PUL, PAI, OJA, KOK and HAR) was completed with an Ortec High Purity Germanium Gamma Spectrometer (Oak Ridge, TN, USA) at the LANSET facility at University of Ottawa. Certified Reference Materials obtained from International Atomic Energy Association (Vienna, Austria) were used for efficiency corrections, and results were analyzed using ScienTissiME (Barry's Bay, ON, Canada). Constant Rate of Supply (CRS) models were used for all primarily dated cores, and cross-calibrated against the peak in ^{137}Cs (see SI archive). The peak was estimated to occur at 1986 in the Finnish cores (VUO, PUL, PAI, OJA, KOK, HAR) as deposition from the Chernobyl disaster (Kansanen et al., 1991).

Geochronology data from previous studies was used directly for the AQT core (Poulain et al., 2015), and the LAH and POK cores, where the profiles were adjusted for different sampling times. The core from Lake Hazen was obtained from the same exact site with the same equipment as a previously ^{210}Pb-dated core (Lehnherr et al., 2018) and its sedimentation-adjusted dating was used for the new core. For Pocket Lake, the ^{210}Pb dating of a previously analyzed core (Thienpont et al., 2016) was used, since we observed no differences in the THg profiles of the cores (P = 0.58; **Figure C7**). For individual horizons in the cores with no direct ^{210}Pb measurements, dates were estimated either by interpolating between measured horizons or extrapolating the models by fitting a second order polynomial to the modeled dates (all R^2 > 0.97; **Figure C8**). Extrapolations were only performed down to 1750 CE, since sediment chronology can be underestimated below supported ^{210}Pb values (Cooke et al., 2010).

4.4.3 Gene quantification with droplet digital PCR (ddPCR)

Copy numbers of *merA* and *glnA* were quantified from all DNA samples (n = 126; **Figure C1**) with the Bio-Rad QX200 ddPCR system, and the primary data was processed with the QuantaSoft suite (Bio-Rad, Hercules, CA, USA). Briefly, droplets were generated from PCR reactions according to manufacturer's guidelines and the reactions were run according to conditions outlined in **Table C1**, using primers for *merA* developed by Gill (2012) and for *glnA* by Hurt et al. (2001). Droplet digital PCR enables quantification of target sequences in samples, similarly to quantitative / real-time PCR, but with simplified workflow, and increased accuracy and precision compared to the other copy number quantification methods (Pinheiro et al., 2012). The accuracy of the assay was controlled with previously quantified plasmid templates containing the target gene for both *merA* and *glnA*, and elution kit buffer extractions as negative controls. The baselines for positive droplets in each sample were adjusted based on the negative

and positive controls. The copy numbers (and 95% CI) for both genes were then normalized to copies per ng DNA in each sample.

4.4.4 Amplicon sequencing

The samples were screened for presence of *merA* and *rpoB* with conventional PCR, using primers for *merA* developed by Wang et al. (2011), and primers for *rpoB* designed for the current study. The primer sets for nested PCR of *merA* were chosen for the current study to enable comparisons to the results of Poulain et al. (2015), where different methods were used to analyze the same extracted DNA samples from Lake Aquatuk. Details of the PCR methods are outlined in **Table C1**. We sequenced the single-copy housekeeping gene *rpoB* as a control instead of *glnA* because it enabled the design of a PCR amplicon similar in size to *merA* (and longer than with *glnA*, which proved too short for diversity analyses). Furthermore, *rpoB* is also a marker gene for bacterial phylogeny (Case et al., 2007), which is useful for discerning between the variation in overall microbial community structure and variation specific to the *merA* Hg resistance gene. The primers for *merA* had been designed to target Proteobacteria and Firmicutes (Wang et al. 2011) and to specifically contain the motif for the terminal cysteine residues in the *merA* gene. The primers for *rpoB* were designed for this study with a broad range of specificity in mind (notable phyla not covered: Cyanobacteria, Gracilibacteria, Microgenomates and SR1). Despite our best efforts (*e.g.*, fresh reagents and enzymes, changing laboratory space, and equipment), negative controls sometimes turned up positive for *merA* (**Figure C9**), but never for *rpoB*. The contamination of the *merA* negative controls was likely due to presence of contaminating DNA in the reagents (Glassing et al., 2016; Salter et al., 2014) and the high combined number of cycles (25 + 35) in the nested PCR approach. Thus, we sequenced 30 samples for *rpoB*, and 30

samples and a kit negative control for *merA* (in total of 9 cores; see **Figure C1**) with Illumina MiSeq (paired-end 250 bp) at Génome Québec (Montreal, QC, Canada).

4.4.5 Sequencing data processing

All processing was completed similarly for both *merA* and *rpoB* unless otherwise stated. The reads were paired with pear v0.9.10 (Zhang et al., 2014a) with > 100 bp overlap and the stringency *P*-value set to 0.0001. The primers and barcodes were removed, sequences were truncated to the first ambiguous base call or where the Phred quality score fell under 28 in a window of 2 bp, and sequences shorter than 150 bp were removed with QIIME 1.9.1 (Caporaso et al., 2010). Chimeric sequences were removed with vsearch v2.0.0 (Rognes et al., 2016) utilizing the uchime algorithm (Edgar et al., 2011) against databases of 614 *merA* or 286 *rpoB* sequences truncated to the amplified region (gene databases are included in **Appendix D**). To filter non-target reads, the sequences were translated to amino acids with EMBOSS 6.5.7 (Rice et al., 2000), searched against custom HMMs (included in **Appendix D**) constructed from the *merA* and *rpoB* databases with HMMER 3.1b2 (Finn et al., 2015), and sequences not matching the profiles (*merA*: $E > 1.0 \times 10^{-25}$; *rpoB*: $E > 1.0 \times 10^{-5}$) were removed. The nucleotide sequences were then dereplicated and clustered with Swarm 2.1.9 (Mahé et al., 2015), and variant abundances were summarized with custom perl and R scripts (see **Appendix D**). The sequences were subset to cluster seeds, aligned with Muscle (Edgar, 2004) through TranslatorX (Abascal et al., 2010) to find a common reading frame, and trimmed to the first aligned codon position. The *merA* sequences that did not encode a Tyr605/Phe605 residue required for activity (Barkay et al., 2003) were removed, and sequences that bridged gaps in the alignments (except for positions after 343 bp for *rpoB*) with < 10% occupancy were removed in both *merA* and *rpoB* alignments. The sequences were realigned with MAFFT (Katoh and Standley, 2013) through

TranslatorX, trimmed with trimAl 1.2rev59 (Capella-Gutiérrez et al., 2009) using the '-gappyout' option, and sequences still containing gaps were removed.

Out of 2,580 quality controlled *merA* variants in the samples, 37 (1.4%) were also present in the sequenced negative control (kit extract). These were first removed (**Figure C10**), and the remaining read coverage was assessed over estimated calendar dates (**Figure C11**). Maximum Likelihood trees were constructed from the sequences that passed all quality control steps (n = 2,551) with FastTree 2.1.9 (Price et al., 2010) using the GTR + Γ model of sequence evolution (*e.g.*, Aris-Brosou and Rodrigue 2012). Long branches, potentially indicating sequencing errors, were then removed with TreeShrink 1.0.0 (Mai and Mirarab, 2017) using a false positive error rate of FDR = 0.01, and the trees were re-reconstructed with FastTree under the same model as before.

4.4.6 ddPCR data analyses

For all ddPCR analyses, the data was limited to samples with estimated dating > 1750 CE, and the HAR samples were removed because of the unreliable (mixed) core profile. Briefly, correlation of gene copy numbers of *merA* and *glnA* with [THg], dating, and sediment depth (distance from surface of the sediment) were assessed with linear mixed-effects models using 'lmer' from lme4 (Bates et al., 2015). The copy numbers were $\log_{10}$-transformed (after assessing the normality of residuals), all variables were scaled and mean-centered, and sampling site was used as a random effect (random intercept). Also, identified outliers (see **Figure C1**; Pulmankijärvi, points around 1900 CE for *glnA*) were removed from the final model. Significance of model variables was assessed from 95% confidence intervals of the transformed variables, and the effect size of significant variables was estimated over their range from model predictions. To compare our ddPCR data against previous studies of *merA* copy numbers across

[THg] (Frossard et al., 2018; Larose et al., 2013), we averaged and $\log_{10}$-transformed (after assessing the normality of residuals) the normalized copy numbers of *merA,* and averaged [THg] over the top 5 cm of sediment samples at each site, and investigated their correlation across the sites with a least square linear regression ('lm' in R).

4.4.7 Diversity analyses

Abundances of *merA* and *rpoB* variants within each samples were normalized with cumulative sum scaling with metagenomeSeq (Paulson et al., 2013). Patterns in beta-diversity (between-sample diversity) of the genes were analyzed with a Double Principle Coordinate Analysis using phylogenetic distances of the variants followed by Non-metric Multidimensional Scaling (NMDS) on the sample distance matrix with phyloseq (McMurdie and Holmes, 2013). Influence of [THg] ($\log_{10}$-transformed ng g^{-1} dw), date (CE), and sampling site on the patterns observed in the ordinations were assessed with the functions 'ordisurf' and 'factorfit' in the vegan package (Oksanen et al., 2016).

4.4.8 Bayesian demographic reconstructions

To reconstruct the phylogeny and past population dynamics at the studied lakes, the *merA* and *rpoB* variants were first subset per sampling site and variants present in more than one sample per site were removed to eliminate cross-contamination. Because of computational limitations, if there were more than 50 variants in a sample, they were reduced to the most phylogenetically diverse variants by calculating patristic distances with ape (Paradis et al., 2004), followed by Ward's hierarchical clustering (Murtagh and Legendre, 2014), and cutting the tree to 50 groups. The timed phylogenies of *merA* and *rpoB* were reconstructed for each lake with BEAST v1.8.0 (Drummond et al., 2012) using the GTR + Γ model of sequence evolution (*e.g.*, Aris-Brosou and Rodrigue 2012), under a skyline prior on the speciation process, and an uncorrelated lognormal

prior on rates (Drummond et al., 2006) (mean = 0.001, stdev = 1.0); an additional lognormal prior was placed on tree heights (mean = 10, stdev = 1.0, offset = 500 years); chains were started from UPGMA trees. Markov chain Monte Carlo samplers were run in duplicate for up to two billion generations (or until convergence), using a thinning of 20,000 to decorrelate samples; converged duplicate runs were combined after removing burn-in periods (conservatively, between 5 to 35 % of each chain length, according to mixing of each chain). The maximum *a posteriori* trees were summarized from the combined collections of trees for each site and gene with TreeAnnotator v1.8.0 (Drummond et al., 2012). Due to computational limitations, the POK combined trees had to be thinned to half the sampling frequency of other sites for calculating the summary tree. The association between deposition date of the sample and phylogenies of the gene variants in each lake was tested with BaTS (Parker et al., 2008) using 10^3 replicates on a subsample of 10^4 trees sampled from the posterior distributions (Poulain et al., 2015); *P*-values were based on the Association Index (Wang et al. 2001). The ancestral demographics estimated with effective population size (N_e) were reconstructed through Bayesian Coalescent Skyline plots (Drummond et al., 2005) using 25 intervals and piecewise-constant spline regressions.

4.4.9 Random forest modeling of N_e and breakpoint analysis

To model overall trends in the demography sizes of both genes, random forests were grown to 5,000 trees with ranger (Wright and Ziegler, 2015). The median of the estimated N_e was used as the dependent variable with $^{210}Pb/^{137}Cs$ dates (CE) and sampling sites as predictors (**Figure C5**). To asses stability of model predictions, the data was randomly split ten times while taking care of class imbalance among sites. In the splits, 80% of the data was used for training of each model and 20% for testing. Partial dependence of model predictions on the predictors was analyzed in each of the ten models with edarf (Jones and Linder, 2016). The effect of date was important for

the accuracy of model predictions for both genes: a sensitivity analysis showed a decrease in the pseudo-R^2, when date was omitted from the models, on average from 0.96 to 0.41 for *merA* and from 0.95 to 0.80 for *rpoB*. The alternative models where [THg] was added as a predictor had similar performance and results to the main models (**Figure C6**), but with a slightly delayed onset of the increase in *merA* N_e and earlier onset for the increase in *rpoB* N_e. However, the prediction accuracies of the alternative models, or their results, did not change with the addition of [THg]. Also, its values were imputed (by the overall mean) for missing dates, *i.e.* before the lowest extent of the [THg] measurements in each core. Thus, we chose the parsimonious model without [THg]. Finally, segmented regression analyses were performed on both the *merA* and *rpoB* partial dependence plots of the relationship between date and estimated N_e with segmented (Muggeo, 2017), based on a single breakpoint and a starting value of 1800.

4.5 Data and software availability

All code and primary data in this study are available through GitHub (https://github.com/Begia/merA-evolution/). In addition, shell scripts and code written for this study is available in **Appendix D**. The raw sequence data been submitted in the NCBI SRA read archive (https://www.ncbi.nlm.nih.gov/sra/PRJNA539962, release pending) and the quality controlled *merA* and *rpoB* variant sequences are available in GenBank (release pending).

4.6 Acknowledgements

We would like to thank Marko Virta and his research group at the University of Helsinki for providing resources and facilities while in Finland, Timo Saarinen (University of Turku) for also assisting with field work with Sami Jokinen, Jukka Mattila (Water Protection Association of the River Kokemäenjoki), Eeva-Kaarina Aaltonen (Pohjanmaan Vesi ja Ympäristö ry), and Juhani Hannila (City of Kokkola). Sofia Perin and Julian Evans assisted in analysis of the ddPCR data.

The high-performance computing environments were provided by Ontario's Centre for Advanced Computing and CSC – IT Center for Science Ltd. We thank Linda Kimpe at the University of Ottawa for her help with sediment dating. This work was funded by the Natural Sciences and Engineering Research Council of Canada (A.J.P. & S.A-B.), the Canadian Foundation for Innovation (A.J.P., S.A-B.), an Invitational Fellowship from the Japanese Society for the Promotion of Science (S.A-B.), and the Finnish Academy of Science and Letters/the Vilho, Yrjö, and Kalle Väisälä Fund (M.O.R.).

Chapter 5: Research synthesis

In summary, my thesis expands our knowledge on the diversity and evolution of microbial communities in the sediments of high-latitude lakes (> 54ºN) and how they are affected by past and present anthropogenic disturbances. High-throughput sequencing of environmental DNA and reconstructions of phylogenetic and functional gene diversity based on this improves on the baseline knowledge of these communities and enable the tracking of historical anthropogenic mercury deposition and its bioavailability.

5.1 Summary of research contributions

In **Chapter 2** I focused on identifying the main drivers of community composition in Lake Hazen and two small freshwater ponds in its watershed. This study was the first microbial taxonomic marker gene analysis conducted on the sediments of large arctic lakes instead of small ponds. Indeed, Lake Hazen is the largest lake by water volume north of the Arctic Circle. While machine learning has been increasingly applied within microbial ecology (Qu et al., 2019), for example to identify microbes which presence depended on the ice cover in lakes (Beall et al., 2016) and to predict soil productivity based on the microbiome composition (Chang et al., 2017), here, machine learning was directly employed to reach many of the conclusions. I applied both unsupervised clustering and partial dependence of Random Forest-based predictions on the most important predictor variables. As a result, the ability of the models to accurately predict physicochemical gradients based on the varying abundance of taxonomic marker genes in the samples was quite remarkable (prediction accuracy by pseudo-R^2 > 0.7 for 4 out of 6 variables). This result shows how the structure of the microbial community is intimately linked with the physicochemical gradients, and supports the use of microbial communities as quantitative

biosensors (Smith et al., 2015) and for environmental biomonitoring (Cordier et al., 2017, 2019), where predictions could be constructed with machine learning. The analysis of the taxonomy-based functional predictions with machine learning also indicated that the phylogenetically dissimilar communities in Lake Hazen might have similar functional potential. Similar results have been observed in, for example, microbial communities in oceans (Louca et al., 2016b) and in connection to bromeliads (Louca et al., 2016a). Finally, the description of the microbial community structure in Lake Hazen is an important addition to the baseline information of global microbial diversity. This lake is also an important model system to follow and predict the effects of climate change on vulnerable Arctic environments (Lehnherr et al., 2018; St. Pierre et al., 2019b), and the establishment of baseline data for the microbial communities enables further tracking of these changes in the future.

In **Chapter 3** I expanded on the marker-gene based taxonomic assessment of the Lake Hazen sediment communities by metagenomic sequencing and genome assembly, using new samples from two of the sites studied in **Chapter 2**. This study provides, to date, the broadest shotgun metagenomic data set from arctic lake sediments (in comparison to Wang et al., 2019). As warming-related changes are ongoing in the Lake Hazen watershed (Lehnherr et al., 2018), it is unclear if and how the communities can adapt to the increases in, for example, input of nutrients, turbidity, and sedimentation rate. However, it is highly likely that together with the ecological state of the lake, the structure of the community will be significantly altered by these changes. My results also expand our knowledge on the biology of the CPR groups, as they had this far evaded detection in the Arctic but appear to be quite common at least in the Lake Hazen sediments. These organisms appear to thrive in environments with highly limited energy availability, such as subsurface environments (Brown et al., 2015; Danczak et al., 2017),

hypersaline soda lakes (Vavourakis et al., 2018) and now also in oligotrophic arctic lake sediments. Furthermore, the likely genomic adaptations to the cold and oligotrophic conditions in the Lake Hazen sediments are further supported by similar discoveries from Antarctic lakes (Koo et al., 2018), with similarly low temperatures and poor energy and nutrient availability. The data gathered in **Chapters 2** and **3** establishes an excellent baseline both for community structure and functional potential in Lake Hazen sediments, enabling the tracking of changes in the microbial communities of this well-studied model system (France, 1993; Köck et al., 2012; Lehnherr et al., 2018; St. Pierre et al., 2019a, 2019b) in the future.

My final research chapter (**Chapter 4**) presents supporting evidence to the recent research hypotheses (Poulain et al., 2015) that (i) sequences of microbial DNA can record the effects of a selection pressure, (ii) the signal is preserved after the removal of such pressure (in the anoxic part of the sediment), and (iii) can then be used in reconstruction of the past trends in it. My results from the effect of past mercury toxicity on microbial resistance replicated those of the preliminary study on a single lake (Poulain et al., 2015). However, this was completed on an expanded geographical scale at continental level and methodology and in addition the results indicated that the signal is indeed sensitive to only the bioavailable fraction of mercury. The bioavailability of mercury in the water and sediment phases is influenced by the inorganic and organic ligands in solution, and is dependent, for example, on the type of dissolved organic matter in different lakes (Chiasson-Gould et al., 2014; Mangal et al., 2019). Furthermore, the evolutionary trajectory of the microbes was swiftly impacted after the onset of the Industrial Revolution and even in lakes with low total-mercury concentrations.

5.2 Ecological implications

The effects of anthropogenic change have a global reach, and despite the role of microbial communities as, for example providers of important ecosystem services, such as recycling of carbon and other nutrients, the responses of the communities to these effects have remained largely unexplored (Pointing et al., 2016). My thesis helps to assess the effects of the two most important anthropogenic changes impacting microbes in aquatic ecosystems: climate change and pollution (Labbate et al., 2016). My research also aids in addressing several open questions in microbial ecology, such as: 'How resilient are microbial communities and different functional groups to ecosystem disturbance?' and 'How do fundamental shifts in environmental conditions impact the trajectory of microbial evolution?' (Antwis et al., 2017).

Firstly, the establishment of a baseline for the structure and functional potential of microbial communities in arctic lake sediments enables tracking their changes in the future. Here, Lake Hazen could serve as an excellent model system to follow how the warming climate changes the arctic ecosystems. Sediment and nutrient delivery into Lake Hazen has already increased due to only a ~1ºC rise in the summer temperatures, and the majority of the nutrients are deposited directly into the sediments in the dense turbidity currents (Lehnherr et al., 2018; St. Pierre et al., 2019b). The ongoing and increasing delivery of organic carbon and nutrients will likely lead to a higher rate of microbial respiration with oxygen and nitrate, which subsequently lower the redox potential at least close to the glacial inputs of the lake (St. Pierre et al., 2019b). The sediment communities are now known to be structured along the redox and pH gradients (**Figures 2.4**, **2.5**, **2.6**), and their genomes contain features which might be beneficial in the cold and oligotrophic conditions (**Figure 3.3**). With even higher rates of microbial respiration and limited oxygen delivery into the sediments (often only the first cm at the top is oxic; **Figure 2.1**),

it is possible that some sediments might turn completely anoxic. This could increase the prevalence of anoxic bacteria such as Chloroflexi, which genomes were shown to contain several denitrification marker genes (**Table D6**). Furthermore, the (mostly) aerobic Acidobacteria and Actinobacteria appeared to be sensitive to changes in redox potential (**Figure A18**), and their abundances likely strongly decrease in more anoxic sediments. These bacteria participate, for example., in the breakdown of complex organic matter and production of bioactive secondary metabolites (see Barka et al., 2015; Kielak et al., 2016). The resilience of these communities to the anticipated environmental changes remains to be seen, but the research conducted in my thesis can serve as an important reference point for future studies.

Secondly, my thesis elucidated how microbial communities in high latitude lakes have reacted to long-term anthropogenic pollution. The signal discovered in the ancestral *merA* sequences is likely sensitive only to bioavailable mercury, because the detoxification of mercury through this pathway is only catalyzed inside the bacterial cells (Boyd and Barkay, 2012). Despite large differences in the total-mercury concentrations in the lakes and the mercury deposition trends over time (**Figure C1**), the signal was observed at all sites examined in the study. Thus, even small increases in the bioavailable fraction of mercury could induce an evolutionary response in the microbes. Specifically, the response was an increase in the estimated N_e of *merA*, which mirrored Hg deposition in the environment (**Figure 4.4**). In the seminal study, selection was found to act on the amino acid residues in the terminal region of the *merA* gene, which might confer advantages to the microbes in adjusting to increased Hg stress (Poulain et al., 2015). Furthermore, because of the global reach of gaseous mercury (Amos et al., 2013; Gworek et al., 2017), it is likely that an increase in the N_e of *merA* could be recovered globally from the (chronologically preserved) sediment records of lakes in contact with the

atmosphere. Genetic signals connected with anthropogenic toxic metal deposition, such as increased N_e, could likely be detected in other microbial resistance genes in environmental archives. For example, the airborne deposition of toxic metal(loid)s (with specific microbial resistance mechanisms) has increased over 10-fold since preindustrial time for lead and cadmium (McConnell and Edwards, 2008), and up to 17-fold for arsenic at industrial sites (Thienpont et al., 2016).

5.3 Limitations and future directions

The results of this thesis are based mostly on environmental DNA extracted from the samples. Thus, the functional and phylogenetic inferences made in these studies rely on comparisons with online databases, which might contain inaccuracies, such as incorrect annotations or sequencing errors (Bengtsson-Palme et al., 2016). Furthermore, extracted DNA is partly from dead or dormant micro-organisms present in the samples, and is known to obscure diversity estimates (Carini et al., 2017). Another important consideration when working with environmental DNA is the proper sampling design and use of techniques to control for contamination and degradation of the molecules. Unfortunately, a large number of current DNA-based microbial biodiversity studies are lacking in these aspects (Dickie et al., 2018), because the field is still quite young and the methods are still underdeveloped.

In **Chapters 2** and **3**, the sampling design and handling could have been improved to match latest recommendations for sequencing-based microbial ecology studies, such as by Dickie et al. (2018). For example, we could have benefited from analyzing several biological replicates from each site, sampling the same exact sites in subsequent years, and analyzing the same sediment depths in all cores. In **Chapter 2**, several primer pairs were used for different sample sets which impeded cross-comparisons between sites and seasons. The different sample

sets were originally planned to be analyzed separately and were sequenced at different time points (and thus with different primer sets). The use of negative DNA extraction controls and their sequencing could have reduced the effort expended in, for instance, assessing the effect of putative contaminating sequences in **Chapter 2** (**Figures A4-A6**). In **Chapter 4**, special care was taken to reduce DNA contamination, but we were not able to get rid of contaminants in our laboratory experiments in **Chapter 4** despite our best efforts (**Figure C9**). This contamination likely originated from the DNA extraction or PCR reagents (*e.g.*, Glassing et al., 2016), and was removed after sequencing by comparing the variants found in the samples to those found in a negative control (**Figure C10**). In the future, care should be taken to assess the presence of problematic contaminants in kits and reagents already prior to DNA extraction from the samples. This is an important consideration, as suboptimal choices made in data analysis can be reverted but obtaining new or more samples from the environment is usually impossible, or at least prohibitively expensive and time consuming.

Anthropogenic effects such as climate change and toxic metal pollution profoundly impact the microbial communities providing crucial ecosystem services, and we are only beginning to understand some of the consequences of these effects. This thesis allows the use of Lake Hazen as a well-established model system to track the warming-related changes in the microbial communities in future studies. The sampling should be extended to other parts of the lake to further gauge the spatial variability of the communities, while also tracking the changes in previously sampled sites. It would be beneficial to use metagenomics and transcriptomics instead of amplicon sequencing in these studies, as sequencing costs will likely continue to decline in the future. While the use of proteomics and metabolomics in microbial ecology is still challenging (Wang et al., 2016a), their use in a systems biology-approach would provide with

valuable insights into the ecological function of the microbes and their interactions with the environment (Aguiar-Pulido et al., 2016).

To understand how microbial communities might adapt in response to the changing conditions, research should be conducted on the evolutionary trajectory of microbial metal resistance under anthropogenic pressures. The geographic extent of the increase in the N_e of *merA* should be examined to find out if the signal is limited to the Northern Hemisphere and if it differs at sites with high geogenic or ancient anthropogenic mercury deposition. As this signal (increasing N_e of *merA*) is likely sensitive only to bioavailable mercury, it could be further developed as a biomarker to track the historically bioavailable mercury in environmental archives. Also, to understand the mechanistic details on how the *merA* gene has evolved under the anthropogenic pressures, ancient-like mercury reductase genes (*i.e.*, the deepest sampled variants) could be generated in the laboratory by site-directed mutagenesis. The expression of these genes in heterologous hosts could be used to test the mercury reduction activity of the enzymes *in vitro*. It should also be investigated if similar evolutionary signals could be found for other microbial toxic metal resistance genes, which could also serve as sensitive biomarkers for their deposition. Genes such as *pbr* for lead (Jaroslawiecka and Piotrowska-Seget, 2014), *czc* for cadmium/zinc/cobolt (Jain and Bhatt, 2014), *cadAB* for cadmium (Abbas et al., 2018), and *ars*-operon genes for arsenic (Ben Fekih et al., 2018; Zhu et al., 2017) would be prime candidates for additional studies.

References

Abascal, F., Zardoya, R., and Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucl. Acids Res.*, gkq291. doi:10.1093/nar/gkq291.

Abbas, S. Z., Rafatullah, M., Hossain, K., Ismail, N., Tajarudin, H. A., and Abdul Khalil, H. P. S. (2018). A review on mechanism and future perspectives of cadmium-resistant bacteria. *Int. J. Environ. Sci. Technol.* 15, 243–262. doi:10.1007/s13762-017-1400-5.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data. *Evol Bioinform Online* 12s1, EBO.S36436. doi:10.4137/EBO.S36436.

Allan, M., Le Roux, G., Sonke, J. E., Piotrowska, N., Streel, M., and Fagel, N. (2013). Reconstructing historical atmospheric mercury deposition in Western Europe using: Misten peat bog cores, Belgium. *Science of The Total Environment* 442, 290–301. doi:10.1016/j.scitotenv.2012.10.044.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods* 11, 1144–1146. doi:10.1038/nmeth.3103.

Al-Rousan, S., Pätzold, J., Al-Moghrabi, S., and Wefer, G. (2004). Invasion of anthropogenic CO2 recorded in planktonic foraminifera from the northern Gulf of Aqaba. *Int J Earth Sci (Geol Rundsch)* 93, 1066–1076. doi:10.1007/s00531-004-0433-4.

Alvarez, H. M., and Steinbüchel, A. (2002). Triacylglycerols in prokaryotic microorganisms. *Appl. Microbiol. Biotechnol.* 60, 367–376. doi:10.1007/s00253-002-1135-0.

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Mol. Biol. Rev.* 59, 143–169.

Amos, H. M., Jacob, D. J., Streets, D. G., and Sunderland, E. M. (2013). Legacy impacts of all-time anthropogenic emissions on the global mercury cycle. *Global Biogeochemical Cycles* 27, 410–421. doi:10.1002/gbc.20040.

Andreotti, R., Pérez de León, A. A., Dowd, S. E., Guerrero, F. D., Bendele, K. G., and Scoles, G. A. (2011). Assessment of bacterial diversity in the cattle tick *Rhipicephalus* (*Boophilus*) microplus through tag-encoded pyrosequencing. *BMC Microbiology* 11, 6. doi:10.1186/1471-2180-11-6.

Antwis, R. E., Griffiths, S. M., Harrison, X. A., Aranega-Bou, P., Arce, A., Bettridge, A. S., et al. (2017). Fifty important research questions in microbial ecology. *FEMS Microbiol Ecol*

93. doi:10.1093/femsec/fix044.

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 46, 2159–2168. doi:10.1093/nar/gky066.

Aris-Brosou, S., and Rodrigue, N. (2012). "The essentials of computational molecular evolution," in *Evolutionary Genomics* Methods in Molecular Biology. (Humana Press, Totowa, NJ), 111–152. doi:10.1007/978-1-61779-582-4_4.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556.

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi:10.1093/bioinformatics/btv287.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J. Exp. Med.* 79, 137–158.

Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H.-P., et al. (2015). Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev* 80, 1–43. doi:10.1128/MMBR.00019-15.

Barkay, T., Kritee, K., Boyd, E., and Geesey, G. (2010). A thermophilic bacterial origin and subsequent constraints by redox, light and salinity on the evolution of the microbial mercuric reductase. *Environmental Microbiology* 12, 2904–2917. doi:10.1111/j.1462-2920.2010.02260.x.

Barkay, T., Miller, S. M., and Summers, A. O. (2003). Bacterial mercury resistance from atoms to ecosystems. *FEMS Microbiology Reviews* 27, 355–384. doi:10.1016/S0168-6445(03)00046-9.

Barkay, T., and Wagner-Dobler, I. (2005). Microbial transformations of mercury: Potentials, challenges, and achievements in controlling mercury toxicity in the environment. *Advances in applied microbiology* 57, 1–52.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* 67. doi:10.18637/jss.v067.i01.

Beall, B. F. N., Twiss, M. R., Smith, D. E., Oyserman, B. O., Rozmarynowycz, M. J., Binding, C. E., et al. (2016). Ice cover extent drives phytoplankton and bacterial community structure in a large north-temperate lake: implications for a warming climate: Effect of ice cover on microbial community structure. *Environmental Microbiology* 18, 1704–1719. doi:10.1111/1462-2920.12819.

Ben Fekih, I., Zhang, C., Li, Y. P., Zhao, Y., Alwathnani, H. A., Saquib, Q., et al. (2018). Distribution of Arsenic Resistance Genes in Prokaryotes. *Front Microbiol* 9. doi:10.3389/fmicb.2018.02473.

Bengtsson-Palme, J., Boulund, F., Edström, R., Feizi, A., Johnning, A., Jonsson, V. A., et al. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *PROTEOMICS* 16, 2454–2460. doi:10.1002/pmic.201600034.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Research* 41, D36–D42. doi:10.1093/nar/gks1195.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* 456, 53–59. doi:10.1038/nature07517.

Bliss, A., Hock, R., and Radić, V. (2014). Global response of glacier runoff to twenty-first century climate change. *Journal of Geophysical Research: Earth Surface* 119, 717–730. doi:10.1002/2013JF002931.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bölter, M., and Müller, F. (2016). Resilience in polar ecosystems: From drivers to impacts and changes. *Polar Science* 10, 52–59. doi:10.1016/j.polar.2015.09.002.

Bone, A., Haldin, L., Koivisto, A.-M., Mäenpää, E., Mäensivu, M., Pakkala, J., et al. (2016). *Luodon-Öjanjärveen laskevien vesistöjen vesienhoidon toimenpideohjelma 2016-2021*. Etelä-Pohjanmaan ELY-keskus Available at: http://www.doria.fi/handle/10024/124451 [Accessed November 15, 2018].

Booth, I. R. (1985). Regulation of cytoplasmic pH in bacteria. *Microbiol Rev* 49, 359–378.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35, 725–731. doi:10.1038/nbt.3893.

Boyd, E., and Barkay, T. (2012a). The mercury resistance operon: from an origin in a geothermal environment to an efficient detoxification machine. *Front. Microbio.* 3, 349. doi:10.3389/fmicb.2012.00349.

Boyd, E. S., and Barkay, T. (2012b). The mercury resistance operon: From an origin in a geothermal environment to an efficient detoxification machine. *Frontiers in Microbiology* 3, 349. doi:10.3389/fmicb.2012.00349.

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32. doi:10.1023/A:1010933404324.

Brooks, W. E. (2012). “Industrial Use of Mercury in the Ancient World,” in *Mercury in the Environment*, ed. M. Bank (University of California Press), 19–24. doi:10.1525/california/9780520271630.003.0002.

Brouwer, H., and Murphy, T. (1995). Volatile sulfides and their toxicity in freshwater sediments. *Environmental Toxicology and Chemistry* 14, 203–208. doi:10.1002/etc.5620140204.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., et al. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523, 208.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol* 90, 543–550. doi:10.1111/1574-6941.12437.

Cadillo-Quiroz, H., Yavitt, J. B., and Zinder, S. H. (2009). Methanosphaerula palustris gen. nov., sp. nov., a hydrogenotrophic methanogen isolated from a minerotrophic fen peatland. *Int. J. Syst. Evol. Microbiol.* 59, 928–935. doi:10.1099/ijs.0.006890-0.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11, 2639–2643. doi:10.1038/ismej.2017.119.

Campbell, J. H., O’Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., et al. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *PNAS* 110, 5540–5545. doi:10.1073/pnas.1303090110.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. in *Advances in Knowledge Discovery and Data Mining* Lecture Notes in Computer Science., eds. J. Pei, V. Tseng, L. Cao, H. Motoda, and G. Xu (Springer, Berlin, Heidelberg), 160–172. doi:10.1007/978-3-642-37456-2_14.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi:10.1093/bioinformatics/btp348.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7, 335–336. doi:10.1038/nmeth.f.303.

Carini, P. (2019). A “Cultural” Renaissance: Genomics Breathes New Life into an Old Craft. *mSystems* 4, e00092-19. doi:10.1128/mSystems.00092-19.

Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., and Fierer, N. (2017). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature*

Microbiology 2, 16242. doi:10.1038/nmicrobiol.2016.242.

Carr, S. A., Schubotz, F., Dunbar, R. B., Mills, C. T., Dias, R., Summons, R. E., et al. (2018). Acetoclastic *Methanosaeta* are dominant methanogens in organic-rich Antarctic marine sediments. *The ISME Journal* 12, 330–342. doi:10.1038/ismej.2017.150.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Appl. Environ. Microbiol.* 73, 278–288. doi:10.1128/AEM.01177-06.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 46, D633–D639. doi:10.1093/nar/gkx935.

Chang, H.-X., Haudenshield, J. S., Bowen, C. R., and Hartman, G. L. (2017). Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. *Front Microbiol* 8. doi:10.3389/fmicb.2017.00519.

Chapman, P. M., Adams, W. J., Brooks, M., Delos, C. G., Luoma, S. N., Maher, W. A., et al. (2010). *Ecological Assessment of Selenium in the Aquatic Environment*. CRC Press.

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., et al. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* 47, D666–D677. doi:10.1093/nar/gky901.

Chiasson-Gould, S. A., Blais, J. M., and Poulain, A. J. (2014). Dissolved organic matter kinetically controls mercury bioavailability to bacteria. *Environ. Sci. Technol.* 48, 3153–3161. doi:10.1021/es4038484.

Chintalapati, S., Kiran, M. D., and Shivaji, S. (2004). Role of membrane lipid fatty acids in cold adaptation. *Cell. Mol. Biol. (Noisy-le-grand)* 50, 631–642.

Chistoserdova, L., and Lidstrom, P. M. E. (2013). "Aerobic methylotrophic prokaryotes," in *The Prokaryotes*, eds. E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson (Springer Berlin Heidelberg), 267–285. doi:10.1007/978-3-642-30141-4_68.

Chu, H., Fierer, N., Lauber, C. L., Caporaso, J. G., Knight, R., and Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environmental Microbiology* 12, 2998–3006. doi:10.1111/j.1462-2920.2010.02277.x.

Clackett, S. P., Porter, T. J., and Lehnherr, I. (2018). 400-Year Record of Atmospheric Mercury from Tree-Rings in Northwestern Canada. *Environ. Sci. Technol.* 52, 9625–9633. doi:10.1021/acs.est.8b01824.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 4, 265–270. doi:10.1038/nnano.2009.12.

Cohen, A. S. (2003). *Paleolimnology: The History and Evolution of Lake Systems*. Oxford University Press.

Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3240–3244.

Cook, A. M., Laue, H., and Junker, F. (1998). Microbial desulfonation. *FEMS Microbiol. Rev.* 22, 399–419. doi:10.1111/j.1574-6976.1998.tb00378.x.

Cooke, C. A., Balcom, P. H., Biester, H., and Wolfe, A. P. (2009). Over three millennia of mercury pollution in the Peruvian Andes. *Proceedings of the National Academy of Sciences* 106, 8830–8834. doi:10.1073/pnas.0900517106.

Cooke, C. A., Hobbs, W. O., Michelutti, N., and Wolfe, A. P. (2010). Reliance on 210Pb Chronology Can Compromise the Inference of Preindustrial Hg Flux to Lake Sediments. *Environ. Sci. Technol.* 44, 1998–2003. doi:10.1021/es9027925.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., et al. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environ. Sci. Technol.* 51, 9118–9126. doi:10.1021/acs.est.7b01518.

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology* 27, 387–397. doi:10.1016/j.tim.2018.10.012.

Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General Nature of the Genetic Code for Proteins. *Nature* 192, 1227. doi:10.1038/1921227a0.

Crump, B. C., Amaral-Zettler, L. A., and Kling, G. W. (2012). Microbial diversity in arctic freshwaters is structured by inoculation of microbes from soils. *ISME J* 6, 1629–1639. doi:10.1038/ismej.2012.9.

Danczak, R. E., Johnston, M. D., Kenah, C., Slattery, M., Wrighton, K. C., and Wilkins, M. J. (2017). Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome* 5. doi:10.1186/s40168-017-0331-1.

DeAngelis, K. M., Silver, W. L., Thompson, A. W., and Firestone, M. K. (2010). Microbial communities acclimate to recurring changes in soil redox potential status. *Environmental Microbiology* 12, 3137–3149. doi:10.1111/j.1462-2920.2010.02286.x.

Dearing, J. A., Battarbee, R. W., Dikau, R., Larocque, I., and Oldfield, F. (2006). Human–environment interactions: learning from the past. *Reg Environ Change* 6, 1–16. doi:10.1007/s10113-005-0011-8.

Dearing, J. A., Jones, R. T., Shen, J., Yang, X., Boyle, J. F., Foster, G. C., et al. (2008). Using

multiple archives to understand past and present climate–human–environment interactions: the lake Erhai catchment, Yunnan Province, China. *J Paleolimnol* 40, 3–31. doi:10.1007/s10933-007-9182-2.

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., Rappé, M. S., et al. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* 3, 804–813. doi:10.1038/s41564-018-0176-9.

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., et al. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources* 18, 940–952. doi:10.1111/1755-0998.12907.

Drevnick, P. E., Muir, D. C. G., Lamborg, C. H., Horgan, M. J., Canfield, D. E., Boyle, J. F., et al. (2010). Increased accumulation of sulfur in lake sediments of the High Arctic. *Environ. Sci. Technol.* 44, 8415–8421. doi:10.1021/es101991p.

Driscoll, C. T., Mason, R. P., Chan, H. M., Jacob, D. J., and Pirrone, N. (2013). Mercury as a global pollutant: Sources, pathways, and effects. *Environmental Science & Technology* 47, 4967–4983. doi:10.1021/es305071v.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4, e88. doi:10.1371/journal.pbio.0040088.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22, 1185–1192. doi:10.1093/molbev/msi103.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29, 1969–1973. doi:10.1093/molbev/mss075.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23, 205–211.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi:10.1093/bioinformatics/btr381.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 323, 133–138. doi:10.1126/science.1162986.

Elbaz-Poulichet, F., Dezileau, L., Freydier, R., Cossa, D., and Sabatier, P. (2011). A 3500-year record of Hg and Pb contamination in a mediterranean sedimentary archive (the Pierre Blanche Lagoon, France). *Environ. Sci. Technol.* 45, 8642–8647. doi:10.1021/es2004599.

Emerson, J. B., Varner, R. K., Johnson, J. E., Owusu-Dommey, A., Binder, M., Woodcroft, B. J., et al. (2015). Linking sediment microbial communities to carbon cycling in high-latitude lakes. *AGU Fall Meeting Abstracts* 21.

Emmerton, C. A., St. Louis, V. L., Lehnherr, I., Graydon, J. A., Kirk, J. L., and Rondeau, K. J. (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, 5849–5863. doi:10.5194/bg-13-5849-2016.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, e1319. doi:10.7717/peerj.1319.

Ferrer, M., Guazzaroni, M.-E., Richter, M., García-Salamanca, A., Yarza, P., Suárez-Suárez, A., et al. (2011). Taxonomic and Functional Metagenomic Profiling of the Microbial Community in the Anoxic Sediment of a Sub-saline Shallow Lake (Laguna de Carrizo, Central Spain). *Microbial Ecology* 62, 824–837.

Ferry, J. G. (1993). *Methanogenesis: Ecology, physiology, biochemistry & genetics*. Springer Science & Business Media.

Filippidou, S., Junier, T., Wunderlin, T., Lo, C.-C., Li, P.-E., Chain, P. S., et al. (2015). Under-detection of endospore-forming Firmicutes in metagenomic data. *Computational and Structural Biotechnology Journal* 13, 299–306. doi:10.1016/j.csbj.2015.04.002.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res* 42, D222–D230. doi:10.1093/nar/gkt1223.

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., et al. (2015). HMMER web server: 2015 update. *Nucl. Acids Res.* 43, W30–W38. doi:10.1093/nar/gkv397.

Finnish Environment Institute SYKE (2018a). Päiväjärvi (84.048.1.002). *Järviwiki*. Available at: http://www.jarviwiki.fi/wiki/P%C3%A4iv%C3%A4j%C3%A4rvi_(84.048.1.002) [Accessed November 15, 2018].

Finnish Environment Institute SYKE (2018b). Vesistöennusteet: Kokemäenjoen vesistöalue - Harjavalta. Available at: http://wwwi2.ymparisto.fi/i2/35/q3510450y/wqfi.html [Accessed November 15, 2018].

Florentino, A. P., Stams, A. J. M., and Sánchez-Andrea, I. (2017). Genome sequence of *Desulfurella amilsii* strain TR1 and comparative genomics of Desulfurellaceae family. *Frontiers in Microbiology* 8. doi:10.3389/fmicb.2017.00222.

France, R. L. (1993). The Lake Hazen Trough: A late winter oasis in a polar desert. *Biological Conservation* 63, 149–151. doi:10.1016/0006-3207(93)90503-S.

Frossard, A., Donhauser, J., Mestrot, A., Gygax, S., Bååth, E., and Frey, B. (2018). Long- and short-term effects of mercury pollution on the soil microbiome. *Soil Biology and Biochemistry* 120, 191–199. doi:10.1016/j.soilbio.2018.01.028.

Fukuyama, J., MCMURDIE, P. J., DETHLEFSEN, L., RELMAN, D. A., and HOLMES, S. (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pac Symp Biocomput*, 213–224.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 43, D261–D269. doi:10.1093/nar/gku1223.

Garrity, G. M., Bell, J. A., and Lilburn, T. (2015). "*Legionellales ord. nov*," in *Bergey's Manual of Systematics of Archaea and Bacteria*, eds. W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun, P. DeVos, et al. (Chichester, UK: John Wiley & Sons, Ltd), 1–1. doi:10.1002/9781118960608.obm00098.

Ge, Y., He, J., Zhu, Y., Zhang, J., Xu, Z., Zhang, L., et al. (2008). Differences in soil bacterial diversity: driven by contemporary disturbances or historical contingencies? *ISME J* 2, 254–264. doi:10.1038/ismej.2008.2.

Gerbersdorf, S. U., Hollert, H., Brinkmann, M., Wieprecht, S., Schüttrumpf, H., and Manz, W. (2011). Anthropogenic pollutants affect ecosystem services of freshwater sediments: the need for a "triad plus x" approach. *J Soils Sediments* 11, 1099–1114. doi:10.1007/s11368-011-0373-0.

Gibson, J. J., Reid, R., and Spence, C. (1998). A six-year isotopic record of lake evaporation at a mine site in the Canadian subarctic: results and validation. *Hydrological Processes* 12, 1779–1792. doi:10.1002/(SICI)1099-1085(199808/09)12:10/11<1779::AID-HYP694>3.0.CO;2-7.

Gill, H. (2012). The Effect of Aluminium Industry Effluents on Sediment Bacterial Communities. PhD Thesis. University of Ottawa, Ottawa, ON, Canada. Available at: http://www.ruor.uottawa.ca/handle/10393/23423 [Accessed September 20, 2019].

Gilmour, C. C., Podar, M., Bullock, A. L., Graham, A. M., Brown, S. D., Somenahally, A. C., et al. (2013). Mercury methylation by novel microorganisms from new environments. *Environmental Science & Technology* 47, 11810–11820. doi:10.1021/es403075t.

Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* 8, 1553–1565. doi:10.1038/ismej.2014.60.

Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B., and Chiodini, R. J. (2016). Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens* 8, 24. doi:10.1186/s13099-016-0103-7.

Greene, A. C. (2014). "The family Desulfurellaceae," in *The Prokaryotes*, eds. E. Rosenberg, E.

F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson (Berlin, Heidelberg: Springer Berlin Heidelberg), 135–142. doi:10.1007/978-3-642-39044-9_312.

Greub, G., and Raoult, D. (2004). Microorganisms resistant to free-living amoebae. *Clin. Microbiol. Rev.* 17, 413–433.

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., et al. (2008). Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* 105, 9145–9150. doi:10.1073/pnas.0804023105.

Gworek, B., Dmuchowski, W., Baczewska, A. H., Brągoszewska, P., Bemowska-Kałabun, O., and Wrzosek-Jakubowska, J. (2017). Air Contamination by Mercury, Emissions and Transformations—a Review. *Water Air Soil Pollut* 228. doi:10.1007/s11270-017-3311-y.

Hahsler, M., Piekenbrock, M., Arya, S., and Mount, D. (2018). *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. Available at: https://CRAN.R-project.org/package=dbscan [Accessed June 4, 2018].

Hao, O. J., Chen, J. M., Huang, L., and Buglass, R. L. (1996). Sulfate-reducing bacteria. *Critical Reviews in Environmental Science and Technology* 26, 155–187. doi:10.1080/10643389609388489.

Hauptmann, A. L., Markussen, T. N., Stibal, M., Olsen, N. S., Elberling, B., Bælum, J., et al. (2016). Upstream freshwater and terrestrial sources are differentially reflected in the bacterial community structure along a small Arctic river and its estuary. *Frontiers in Microbiology* 7. doi:10.3389/fmicb.2016.01474.

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Marengo Orsini, J. A., et al. (2007). Understanding and attributing climate change.

Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* 5, 1571–1579. doi:10.1038/ismej.2011.41.

Hoppe, H., Gocke, K., and Kuparinen, J. (1990). Effect of H2S on heterotrophic substrate uptake, extracellular enzyme activity and growth of brackish water bacteria. *Marine Ecology Progress Series* 64, 157–167. doi:10.3354/meps064157.

Horikoshi, M., and Tang, Y. (2017). *ggfortify: Data visualization tools for statistical analysis results*. Available at: https://cran.r-project.org/web/packages/ggfortify/index.html.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *PNAS*, 201402564. doi:10.1073/pnas.1402564111.

Hsu-Kim, H., Kucharzyk, K. H., Zhang, T., and Deshusses, M. A. (2013). Mechanisms regulating mercury bioavailability for methylating microorganisms in the aquatic

environment: a critical review. *Environ. Sci. Technol.* 47, 2441–2456. doi:10.1021/es304370g.

Hsu-Kim, H., S. Eckley, C., and E. Selin, N. (2018). Modern science of a legacy problem: mercury biogeochemical research after the Minamata Convention. *Environmental Science: Processes & Impacts* 20, 582–583. doi:10.1039/C8EM90016G.

Huang, J., Zhang, X., Zhang, Q., Lin, Y., Hao, M., Luo, Y., et al. (2017). Recently amplified arctic warming has contributed to a continual global warming trend. *Nature Climate Change* 7, 875. doi:10.1038/s41558-017-0009-5.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nature Microbiology* 1, 16048. doi:10.1038/nmicrobiol.2016.48.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology* 3, reviews0003.1. doi:10.1186/gb-2002-3-2-reviews0003.

Hughes, M. N., and Poole, R. K. (1991). Metal speciation and microbial growth—the hard (and soft) facts. *Microbiology* 137, 725–734. doi:10.1099/00221287-137-4-725.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 40, D306–D312. doi:10.1093/nar/gkr948.

Hunting, E. R., and Kampfraath, A. A. (2013). Contribution of bacteria to redox potential (Eh) measurements in sediments. *Int. J. Environ. Sci. Technol.* 10, 55–62. doi:10.1007/s13762-012-0080-4.

Hurt, R. A., Qiu, X., Wu, L., Roh, Y., Palumbo, A. V., Tiedje, J. M., et al. (2001). Simultaneous Recovery of RNA and DNA from Soils and Sediments. *Appl. Environ. Microbiol.* 67, 4495–4503. doi:10.1128/AEM.67.10.4495-4503.2001.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.

Ilmast, N. V., and Sterligova, O. P. (2002). Biological characteristics of European whitefish in Lake Pulmankijarvi, northern Finland. *Ergebnisse der Limnologie* 57, 359–366.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945. doi:10.1038/nature03001.

IPCC (2018). *Global warming of 1.5 C An IPCC Special Report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. , eds. V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, et al. In Press.

Iwata, H., Tanabe, S., Sakai, N., and Tatsukawa, R. (1993). Distribution of persistent organochlorines in the oceanic air and surface seawater and the role of ocean on their global transport and fate. *Environ. Sci. Technol.* 27, 1080–1098. doi:10.1021/es00043a007.

Jain, S., and Bhatt, A. (2014). Molecular and in situ characterization of cadmium-resistant diversified extremophilic strains of Pseudomonas for their bioremediation potential. *3 Biotech* 4, 297–304. doi:10.1007/s13205-013-0155-z.

Jannasch, H. W., and Jones, G. E. (1959). Bacterial Populations in Sea Water as Determined by Different Methods of Enumeration1. *Limnology and Oceanography* 4, 128–139. doi:10.4319/lo.1959.4.2.0128.

Jaroslawiecka, A., and Piotrowska-Seget, Z. (2014). Lead resistance in micro-organisms. *Microbiology* 160, 12–25. doi:10.1099/mic.0.070284-0.

Jenkins, O., Byrom, D., and Jones, D. (1987). Methylophilus: a new genus of methanol-utilizing bacteria. *International Journal of Systematic and Evolutionary Microbiology* 37, 446–448.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Jones, Z., and Linder, F. (2016). edarf: Exploratory data analysis using random forests. *The Journal of Open Source Software* 1. doi:10.21105/joss.00092.

Jormakka, M., Yokoyama, K., Yano, T., Tamakoshi, M., Akimoto, S., Shimamura, T., et al. (2008). Molecular mechanism of energy conservation in polysulfide respiration. *Nat Struct Mol Biol* 15, 730–737. doi:10.1038/nsmb.1434.

Jou, W. M., Haegeman, G., Ysebaert, M., and Fiers, W. (1972). Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* 237, 82–88. doi:10.1038/237082a0.

Kalnina, L., Stivrins, N., Kuske, E., Ozola, I., Pujate, A., Zeimule, S., et al. (2015). Peat stratigraphy and changes in peat formation during the Holocene in Latvia. *Quaternary International* 383, 186–195. doi:10.1016/j.quaint.2014.10.020.

Kämpfer, P. (2015). "*Sphingobacteriia class. nov.*," in *Bergey's Manual of Systematics of Archaea and Bacteria* (John Wiley & Sons, Ltd). doi:10.1002/9781118960608.cbm00013.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45, D353–D361. doi:10.1093/nar/gkw1092.

Kansanen, P. H., Jaakkola, T., Kulmala, S., and Suutarinen, R. (1991). Sedimentation and

distribution of gamma-emitting radionuclides in bottom sediments of southern Lake Päijänne, Finland, after the Chernobyl accident. *Hydrobiologia* 222, 121–140. doi:10.1007/BF00006100.

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30, 772–780. doi:10.1093/molbev/mst010.

Katsev, S., and Crowe, S. A. (2015). Organic carbon burial efficiencies in sediments: The power law of mineralization revisited. *GEOLOGY* 43, 607–610. doi:10.1130/G36626.1.

Keatley, B. E., Douglas, M. S. V., and Smol, J. P. (2007). Limnological characteristics of a High Arctic oasis and comparisons across northern Ellesmere Island. *Arctic* 60, 294–308.

Kibblewhite, M., Tóth, G., and Hermann, T. (2015). Predicting the preservation of cultural artefacts and buried materials in soil. *Science of The Total Environment* 529, 249–263. doi:10.1016/j.scitotenv.2015.04.036.

Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A., and Kuramae, E. E. (2016). The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front Microbiol* 7. doi:10.3389/fmicb.2016.00744.

Klatt, J. M., Haas, S., Yilmaz, P., de Beer, D., and Polerecky, L. (2015). Hydrogen sulfide can inhibit and enhance oxygenic photosynthesis in a cyanobacterium from sulfidic springs. *Environ Microbiol* 17, 3301–3313. doi:10.1111/1462-2920.12791.

Klein, D. A. (2007). Microbial Communities in Nature: a Postgenomic Perspective. *Microbe Magazine* 2, 591–595. doi:10.1128/microbe.2.591.1.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 41, e1–e1. doi:10.1093/nar/gks808.

Köck, G., Muir, D., Yang, F., Wang, X., Talbot, C., Gantner, N., et al. (2012). Bathymetry and Sediment Geochemistry of Lake Hazen (Quttinirpaaq National Park, Ellesmere Island, Nunavut). *Arctic*, 56–66.

Kokelj, S. V., Jenkins, R. E., Milburn, D., Burn, C. R., and Snow, N. (2005). The influence of thermokarst disturbance on the water quality of small upland lakes, Mackenzie Delta region, Northwest Territories, Canada. *Permafrost and Periglacial Processes* 16, 343–353. doi:10.1002/ppp.536.

Koo, H., Hakim, J. A., Morrow, C. D., Crowley, M. R., Andersen, D. T., and Bej, A. K. (2018). Metagenomic Analysis of Microbial Community Compositions and Cold-Responsive Stress Genes in Selected Antarctic Lacustrine and Soil Ecosystems. *Life (Basel)* 8, 29. doi:10.3390/life8030029.

Krantzberg, G. (1985). The influence of bioturbation on physical, chemical and biological parameters in aquatic environments: A review. *Environmental Pollution Series A, Ecological and Biological* 39, 99–122. doi:10.1016/0143-1471(85)90009-1.

Krijthe, J., and van der Maaten, L. (2017). *rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation*. Available at: https://cran.r-project.org/web/packages/Rtsne/index.html [Accessed November 29, 2017].

Kuang, J.-L., Huang, L.-N., Chen, L.-X., Hua, Z.-S., Li, S.-J., Hu, M., et al. (2013). Contemporary environmental variation determines microbial diversity patterns in acid mine drainage. *ISME J* 7, 1038–1050. doi:10.1038/ismej.2012.139.

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. (2011). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Bioinformatics* CHAPTER, Unit10.7. doi:10.1002/0471250953.bi1007s36.

Kuusisto, E. (1981). *Suomen vesistöjen lämpötilat kaudella 1961-1975*. Vesihallitus. National Board of Waters Available at: https://helda.helsinki.fi/handle/10138/31402 [Accessed November 15, 2018].

Labbate, M., Seymour, J. R., Lauro, F., and Brown, M. V. (2016). Editorial: Anthropogenic Impacts on the Microbial Ecology and Function of Aquatic Environments. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01044.

Lamarche-Gagnon, G., Comery, R., Greer, C. W., and Whyte, L. G. (2015). Evidence of in situ microbial activity and sulphidogenesis in perennially sub-0 °C and hypersaline sediments of a high Arctic permafrost spring. *Extremophiles* 19, 1–15. doi:10.1007/s00792-014-0703-4.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotech* 31, 814–821. doi:10.1038/nbt.2676.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Larose, C., Prestat, E., Cecillon, S., Berger, S., Malandain, C., Lyon, D., et al. (2013). Interactions between Snow Chemistry, Mercury Inputs and Microbial Population Dynamics in an Arctic Snowpack. *PLOS ONE* 8, e79972. doi:10.1371/journal.pone.0079972.

Larsen, P. E., Field, D., and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods* 9, 621–625. doi:10.1038/nmeth.1975.

Latifovic, R., and Pouliot, D. (2007). Analysis of climate change impacts on lake ice phenology in Canada using the historical satellite data record. *Remote Sensing of Environment* 106, 492–507. doi:10.1016/j.rse.2006.09.015.

Lau, N.-S., Zarkasi, K. Z., Md Sah, A. S. R., and Shu-Chien, A. C. (2018). Diversity and Coding Potential of the Microbiota in the Photic and Aphotic Zones of Tropical Man-Made Lake with Intensive Aquaculture Activities: a Case Study on Temengor Lake, Malaysia. *Microb Ecol*. doi:10.1007/s00248-018-1283-0.

Lehnherr, I., Louis, V. L. S., Sharp, M., Gardner, A. S., Smol, J. P., Schiff, S. L., et al. (2018). The world's largest High Arctic lake responds rapidly to climate warming. *Nature Communications* 9, 1290. doi:10.1038/s41467-018-03685-z.

Lehnherr, I., St. Louis, V. L., and Kirk, J. L. (2012). Methylmercury cycling in high arctic wetland ponds: Controls on sedimentary production. *Environ. Sci. Technol.* 46, 10523–10531. doi:10.1021/es300577e.

Lei, Y.-L., Stumm, K., Wickham, S. A., and Berninger, U.-G. (2014). Distributions and biomass of benthic ciliates, foraminifera and amoeboid protists in marine, brackish, and freshwater sediments. *J. Eukaryot. Microbiol.* 61, 493–508. doi:10.1111/jeu.12129.

Lennon, J. T., Muscarella, M. E., Placella, S. A., and Lehmkuhl, B. K. (2017). How, when, and where relic DNA biases estimates of microbial diversity. *bioRxiv*, 131284. doi:10.1101/131284.

León-Zayas, R., Peoples, L., Biddle, J. F., Podell, S., Novotny, M., Cameron, J., et al. (2017). The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the candidate superphylum Parcubacteria (OD1). *Environmental Microbiology* 19, 2769–2784. doi:10.1111/1462-2920.13789.

Lever, M. A., Rogers, K. L., Lloyd, K. G., Overmann, J., Schink, B., Thauer, R. K., et al. (2015). Life under extreme energy limitation: a synthesis of laboratory- and field-based investigations. *FEMS Microbiol Rev* 39, 688–728. doi:10.1093/femsre/fuv020.

Lewis, S. L., and Maslin, M. A. (2015). Defining the Anthropocene. *Nature* 519, 171–180. doi:10.1038/nature14258.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi:10.1093/bioinformatics/btv033.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22.

Lin, W., Jogler, C., Schüler, D., and Pan, Y. (2011). Metagenomic Analysis Reveals Unexpected Subgenomic Diversity of Magnetotactic Bacteria within the Phylum Nitrospirae. *Appl. Environ. Microbiol.* 77, 323–326. doi:10.1128/AEM.01476-10.

Lionard, M., Péquin, B., Lovejoy, C., and Vincent, W. F. (2012). Benthic cyanobacterial mats in the High Arctic: Multi-layer structure and fluorescence responses to osmotic stress. *Front. Microbiol.* 3. doi:10.3389/fmicb.2012.00140.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of Next-Generation Sequencing Systems. *BioMed Research International*. doi:10.1155/2012/251364.

Liu, X., Shen, J., Wang, S., Yang, X., Tong, G., and Zhang, E. (2002). A 16000-year pollen record of Qinghai Lake and its paleo-climate and paleoenvironment. *Chin. Sci. Bull.* 47, 1931. doi:10.1360/02tb9421.

Long, G. S., Hussen, M., Dench, J., and Aris-Brosou, S. (2019). Identifying genetic determinants of complex phenotypes from whole genome sequence data. *BMC Genomics* 20, 470. doi:10.1186/s12864-019-5820-0.

Lory, S. (2014). "The Phylum Chlamydiae," in *The Prokaryotes* (Springer, Berlin, Heidelberg), 497–499. doi:10.1007/978-3-642-38954-2_151.

Louca, S., Jacques, S. M. S., Pires, A. P. F., Leal, J. S., Srivastava, D. S., Parfrey, L. W., et al. (2016a). High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution* 1, 0015. doi:10.1038/s41559-016-0015.

Louca, S., Parfrey, L. W., and Doebeli, M. (2016b). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi:10.1126/science.aaf4507.

Lozupone, C. A., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 104, 11436–11440. doi:10.1073/pnas.0611525104.

Lundin, D., Severin, I., Logue, J. B., Östman, Ö., Andersson, A. F., and Lindström, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? *Environmental Microbiology Reports* 4, 367–372. doi:10.1111/j.1758-2229.2012.00345.x.

Magny, M., Guiot, J., and Schoellammer, P. (2001). Quantitative Reconstruction of Younger Dryas to Mid-Holocene Paleoclimates at Le Locle, Swiss Jura, Using Pollen and Lake-Level Data. *Quaternary Research* 56, 170–180. doi:10.1006/qres.2001.2257.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ* 3, e1420. doi:10.7717/peerj.1420.

Mai, U., and Mirarab, S. (2017). TreeShrink: Efficient Detection of Outlier Tree Leaves. in *Comparative Genomics* Lecture Notes in Computer Science., eds. J. Meidanis and L. Nakhleh (Springer International Publishing), 116–140.

Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19, 272. doi:10.1186/s12864-018-4620-2.

Mangal, V., Stenzler, B. R., Poulain, A. J., and Guéguen, C. (2019). Aerobic and Anaerobic Bacterial Mercury Uptake is Driven by Algal Organic Matter Composition and Molecular Weight. *Environ. Sci. Technol.* 53, 157–165. doi:10.1021/acs.est.8b04909.

Mansikkaniemi, H., and Syrilä, S. (1965). *Main features of the glacial and postglacial development of Pulmanki valley in northernmost Finland*. Instituti Geographici Universitatis Turkuensis.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature* 437, 376–380. doi:10.1038/nature03959.

Mathema, V. B., Thakuri, B. C., and Sillanpää, M. (2011a). Bacterial mer operon-mediated detoxification of mercurial compounds: a short review. *Arch Microbiol* 193, 837–844. doi:10.1007/s00203-011-0751-4.

Mathema, V. B., Thakuri, B. C., and Sillanpää, M. (2011b). Bacterial mer operon-mediated detoxification of mercurial compounds: a short review. *Arch Microbiol* 193, 837–844. doi:10.1007/s00203-011-0751-4.

McConnell, J. R., and Edwards, R. (2008). Coal burning leaves toxic heavy metal legacy in the Arctic. *PNAS* 105, 12140–12144. doi:10.1073/pnas.0803564105.

McMurdie, P. J., and Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8, e61217. doi:10.1371/journal.pone.0061217.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., et al. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10, 213. doi:10.1186/1471-2105-10-213.

Meselson, M., and Stahl, F. W. (1958). The replication of DNA in Escherichia coli. *Proceedings of the national academy of sciences* 44, 671–682.

Mesquita, P. S., Wrona, F. J., and Prowse, T. D. (2010). Effects of retrogressive permafrost thaw slumping on sediment chemistry and submerged macrophytes in Arctic tundra lakes. *Freshwater Biology* 55, 2347–2358. doi:10.1111/j.1365-2427.2010.02450.x.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386. doi:10.1186/1471-2105-9-386.

Milner, A. M., Khamis, K., Battin, T. J., Brittain, J. E., Barrand, N. E., Füreder, L., et al. (2017). Glacier shrinkage driving global changes in downstream systems. *PNAS* 114, 9770–9778. doi:10.1073/pnas.1619807114.

Minamata Convention on Mercury Available at: http://www.mercuryconvention.org/ [Accessed April 25, 2019].

Mohit, V., Culley, A., Lovejoy, C., Bouchard, F., and Vincent, W. F. (2017a). Hidden biofilms in a far northern lake and implications for the changing Arctic. *NPJ Biofilms Microbiomes* 3. doi:10.1038/s41522-017-0024-3.

Mohit, V., Culley, A., Lovejoy, C., Bouchard, F., and Vincent, W. F. (2017b). Hidden biofilms in a far northern lake and implications for the changing Arctic. *npj Biofilms and Microbiomes* 3. doi:10.1038/s41522-017-0024-3.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., and Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* 110, 3–24. doi:10.1016/j.ymgme.2013.04.024.

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., et al. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecol Evol* 4, 3514–3524. doi:10.1002/ece3.1155.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi:10.1038/nature01262.

Mueller, D. R., Van Hove, Patrick., Antoniades, D., Jeffries, M. O., and Vincent, W. F. (2009). High Arctic lakes as sentinel ecosystems: Cascading regime shifts in climate, ice cover, and mixing. *Limnol. Oceanogr.* 54, 2371–2385. doi:10.4319/lo.2009.54.6_part_2.2371.

Muggeo, V. M. R. (2017). *segmented: Regression Models with Break-Points / Change-Points Estimation*. Available at: https://CRAN.R-project.org/package=segmented [Accessed November 30, 2018].

Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M., and Loy, A. (2015). Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *The ISME Journal* 9, 1152–1165. doi:10.1038/ismej.2014.208.

Murray, D. C., Coghlan, M. L., and Bunce, M. (2015). From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows. *PLoS One* 10. doi:10.1371/journal.pone.0124671.

Murtagh, F., and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J Classif* 31, 274–295. doi:10.1007/s00357-014-9161-z.

Nawrocki, E. (2009). Structural RNA homology search and alignment using covariance models. Available at: http://eddylab.org/software/ssu-align/ [Accessed December 13, 2018].

Nealson, K. H. (1997). Sediment bacteria: who's there, what are they doing, and what's new? *Annu Rev Earth Planet Sci* 25, 403–434. doi:10.1146/annurev.earth.25.1.403.

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., et al. (2013). Patterns and processes of microbial community assembly. *Microbiol Mol Biol Rev* 77, 342–356. doi:10.1128/MMBR.00051-12.

Neukom, R., Steiger, N., Gómez-Navarro, J. J., Wang, J., and Werner, J. P. (2019). No evidence for globally coherent warm and cold periods over the preindustrial Common Era. *Nature* 571, 550–554. doi:10.1038/s41586-019-1401-2.

Nirenberg, M. W., and Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences* 47, 1588–1602. doi:10.1073/pnas.47.10.1588.

Nriagu, J. O. (1994). Mercury pollution from the past mining of gold and silver in the Americas. *Science of The Total Environment* 149, 167–181. doi:10.1016/0048-9697(94)90177-5.

Okonechnikov, K., Golosova, O., Fursov, M., and The UGENE Team (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi:10.1093/bioinformatics/bts091.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2016). *vegan: Community ecology package*. Available at: https://CRAN.R-project.org/package=vegan.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018). *vegan: Community Ecology Package*. Available at: https://CRAN.R-project.org/package=vegan [Accessed June 4, 2018].

Ontario Ministry of Natural Resources and Forestry - Provincial Mapping Office (2013). Ontario Flow Assessment Tool. Available at: https://www.javacoeapp.lrc.gov.on.ca/geonetwork/srv/en/main.home?uuid=b5802601-471c-443c-8b67-83260d09c3e2 [Accessed November 15, 2018].

Ortiz-Alvarez, R., and Casamayor, E. O. (2016). High occurrence of Pacearchaeota and Woesearchaeota (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes. *Environmental Microbiology Reports* 8, 210–217. doi:10.1111/1758-2229.12370.

Osborn, A. M., Bruce, K. D., Strike, P., and Ritchie, D. A. (1997). Distribution, diversity and evolution of the bacterial mercury resistance (*mer*) operon. *FEMS Microbiol. Rev.* 19, 239–262.

Özesmi, S. L., Tan, C. O., and Özesmi, U. (2006). Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecological Modelling* 195, 83–93. doi:10.1016/j.ecolmodel.2005.11.012.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412.

Parducci, L., Väliranta, M., Salonen, J. S., Ronkainen, T., Matetovici, I., Fontana, S. L., et al. (2015). Proxy comparison in ancient peat sediments: pollen, macrofossil and plant DNA. *Philos Trans R Soc Lond B Biol Sci* 370. doi:10.1098/rstb.2013.0382.

Parker, J., Rambaut, A., and Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* 8, 239–246. doi:10.1016/j.meegid.2007.08.001.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114.

Parks, J. M., Johs, A., Podar, M., Bridou, R., Hurt, R. A., Smith, S. D., et al. (2013). The Genetic Basis for Bacterial Mercury Methylation. *Science* 339, 1332–1335. doi:10.1126/science.1230667.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology* 12, e1004977. doi:10.1371/journal.pcbi.1004977.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Robust methods for differential abundance analysis in marker gene surveys. *Nat Methods* 10, 1200–1202. doi:10.1038/nmeth.2658.

Pavoine, S., Dufour, A.-B. A.-B., and Chessel, D. (2004). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J. Theor. Biol.* 228, 523–537. doi:10.1016/j.jtbi.2004.02.014.

Percival, J. B., and Outridge, P. M. (2013). A test of the stability of Cd, Cu, Hg, Pb and Zn profiles over two decades in lake sediments near the Flin Flon Smelter, Manitoba, Canada. *Science of The Total Environment* 454–455, 307–318. doi:10.1016/j.scitotenv.2013.03.011.

Pérez-Rodríguez, M., Silva-Sánchez, N., Kylander, M. E., Bindler, R., Mighall, T. M., Schofield, J. E., et al. (2018). Industrial-era lead and mercury contamination in southern Greenland implicates North American sources. *Sci.Total Env.* 613–614, 919–930. doi:10.1016/j.scitotenv.2017.09.041.

Perga, M.-E., Frossard, V., Jenny, J.-P., Alric, B., Arnaud, F., Berthon, V., et al. (2015). High-resolution paleolimnology opens new management perspectives for lakes adaptation to climate warming. *Front. Ecol. Evol.* 3. doi:10.3389/fevo.2015.00072.

Pericard, P., Dufresne, Y., Couderc, L., Blanquart, S., Touzet, H., and Birol, I. (2018). MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in

metagenomes. *Bioinformatics* 34, 585–591. doi:10.1093/bioinformatics/btx644.

Petäjä, A. (1964). Depth charts of some lakes in Utsjoki, Finnish Lapland. *Ann. Univ. Turku. A, II* 32, 346–349.

Pike, G. (2013). Understanding temporal and spatial temperature variation at the local scale in a high latitude environment.

Pinheiro, L. B., Coleman, V. A., Hindson, C. M., Herrmann, J., Hindson, B. J., Bhat, S., et al. (2012). Evaluation of a Droplet Digital Polymerase Chain Reaction Format for DNA Copy Number Quantification. *Anal Chem* 84, 1003–1011. doi:10.1021/ac202578x.

Pinto, A. J., and Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLOS ONE* 7, e43093. doi:10.1371/journal.pone.0043093.

Pointing, S. B., Fierer, N., Smith, G. J. D., Steinberg, P. D., and Wiedmann, M. (2016). Quantifying human impact on Earth's microbiome. *Nature Microbiology* 1, 16145. doi:10.1038/nmicrobiol.2016.145.

Post, E., Forchhammer, M. C., Bret-Harte, M. S., Callaghan, T. V., Christensen, T. R., Elberling, B., et al. (2009). Ecological Dynamics Across the Arctic Associated with Recent Climate Change. *Science* 325, 1355–1358. doi:10.1126/science.1173113.

Poulain, A. J., Aris-Brosou, S., Blais, J. M., Brazeau, M., Keller, W. (Bill), and Paterson, A. M. (2015). Microbial DNA records historical delivery of anthropogenic mercury. *ISME J* 9, 2541–2550. doi:10.1038/ismej.2015.86.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE* 5, e9490. doi:10.1371/journal.pone.0009490.

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi:10.1093/bioinformatics/bts252.

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of Machine Learning in Microbiology. *Front Microbiol* 10. doi:10.3389/fmicb.2019.00827.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucl. Acids Res.* 41, D590–D596. doi:10.1093/nar/gks1219.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35, 833–844. doi:10.1038/nbt.3935.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria:

R Foundation for Statistical Computing Available at: https://www.R-project.org/.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-project.org/ [Accessed March 4, 2019].

Rabus, R., Hansen, T. A., and Widdel, F. (2006). Dissimilatory Sulfate- and Sulfur-Reducing Prokaryotes. *The Prokaryotes*, 659–768. doi:10.1007/0-387-30742-7_22.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications* 469, 967–977. doi:10.1016/j.bbrc.2015.12.083.

Rapp, J. Z., Bienhold, C., Offre, P., and Boetius, A. (2016). Polysaccharide degradation potential of bacterial communities in Arctic deep-sea sediments (1200-5500 m water depth). in Available at: http://epic.awi.de/41571/.

Rautio, M., Dufresne, F., Laurion, I., Bonilla, S., Vincent, W. F., and Christoffersen, K. S. (2011). Shallow freshwater ecosystems of the circumpolar Arctic1. *Écoscience; Sainte-Foy* 18, 204–222.

Razumov, A. S. (1932). *Mikrobiologija* 1, 131–146.

Reid, B. (1997). *Evaporation studies at mine tailings ponds in the Northwest Territories, Canada*.

Reist, J. D., Gyselman, E., Babaluk, J. A., Johnson, J. D., and Wissink, R. (1995). Evidence for two morphotypes of Arctic char (*Salvelinus alpinus* (L.)) from Lake Hazen, Ellesmere Island, Northwest Territories, Canada. *Nordic Journal of Freshwater Research (Sweden)*.

Rekolainen, S., Verta, M., and Liehu, A. (1986). *The effect of airborne mercury and peatland drainage on sediment mercury contents in some Finnish forest lakes*. Vesihallitus. National Board of Waters.

Renberg, I., and Hansson, H. (2008). The HTH sediment corer. *Journal of Paleolimnology* 40, 655–659.

Renvoisé, A., Merhej, V., Georgiades, K., and Raoult, D. (2011). Intracellular Rickettsiales: Insights into manipulators of eukaryotic cells. *Trends in Molecular Medicine* 17, 573–583. doi:10.1016/j.molmed.2011.05.009.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16, 276–277. doi:10.1016/S0168-9525(00)02024-2.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi:10.1038/nature12352.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi:10.7717/peerj.2584.

Rokach, L. (2010). Ensemble-based classifiers. *Artif Intell Rev* 33, 1–39. doi:10.1007/s10462-009-9124-7.

Romanovsky, V. E., Smith, S. L., and Christiansen, H. H. (2010). Permafrost thermal state in the polar Northern Hemisphere during the international polar year 2007–2009: a synthesis. *Permafrost and Periglacial Processes* 21, 106–116. doi:10.1002/ppp.689.

Rudman, S. M., Kreitzman, M., Chan, K. M. A., and Schluter, D. (2017). Evosystem Services: Rapid Evolution and the Provision of Ecosystem Services. *Trends in Ecology & Evolution* 32, 403–415. doi:10.1016/j.tree.2017.02.019.

Rühland, K. M., Hargan, K. E., Jeziorski, A., Paterson, A. M., Keller, W. (Bill), and Smol, J. P. (2014). A Multi-Trophic Exploratory Survey of Recent Environmental Changes using Lake Sediments in the Hudson Bay Lowlands, Ontario, Canada. *Arctic, Antarctic, and Alpine Research* 46, 139–158. doi:10.1657/1938-4246-46.1.139.

Rühland, K., Priesnitz, A., and Smol, J. P. (2003). Paleolimnological Evidence from Diatoms for Recent Environmental Changes in 50 Lakes across Canadian Arctic Treeline. *Arctic, Antarctic, and Alpine Research* 35, 110–123. doi:10.1657/1523-0430(2003)035[0110:PEFDFR]2.0.CO;2.

Ruuskanen, M. O., St. Pierre, K. A., St. Louis, V. L., Aris-Brosou, S., and Poulain, A. J. (2018). Physicochemical Drivers of Microbial Community Structure in Sediments of Lake Hazen, Nunavut, Canada. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.01138.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12, 87. doi:10.1186/s12915-014-0087-z.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 22, 557–567. doi:10.1101/gr.131383.111.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74, 5463–5467. doi:10.1073/pnas.74.12.5463.

Scharf, S. J., Horn, G. T., and Erlich, H. A. (1986). Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science* 233, 1076–1078.

Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nature Biotechnology* 26, 1113–1115. doi:10.1038/nbt1008-1113.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported

Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi:10.1128/AEM.01541-09.

Schroeder, W. H., and Munthe, J. (1998). Atmospheric mercury—An overview. *Atmospheric Environment* 32, 809–822. doi:10.1016/S1352-2310(97)00293-8.

Schultz, T., Korhonen, P., and Virtanen, M. (1995). A mercury model used for assessment of dredging impacts. *Water, Air, and Soil Pollution* 80, 1171–1180. doi:10.1007/BF01189779.

Schütte, U. M. E., Cadieux, S. B., Hemmerich, C., Pratt, L. M., and White, J. R. (2016). Unanticipated geochemical and microbial community structure under seasonal ice cover in a dilute, dimictic Arctic lake. *Frontiers in Microbiology* 7. doi:10.3389/fmicb.2016.01035.

Schuur, E. a. G., McGuire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., et al. (2015). Climate change and the permafrost carbon feedback. *Nature* 520, 171–179. doi:10.1038/nature14338.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., et al. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.* 35, D260-264. doi:10.1093/nar/gkl1043.

Selin, N. E. (2009). Global biogeochemical cycling of mercury: A review. *Annual Review of Environment and Resources* 34, 43–63. doi:10.1146/annurev.environ.051308.084314.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345–353. doi:10.1038/nature24286.

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* 21, 1794–1805. doi:10.1111/j.1365-294X.2012.05538.x.

Simon, J. (2002). Enzymology and bioenergetics of respiratory nitrite ammonification. *FEMS Microbiology Reviews* 26, 285–309. doi:10.1111/j.1574-6976.2002.tb00616.x.

Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., et al. (2015). Natural bacterial communities serve as quantitative geochemical biosensors. *mBio* 6, e00326-15. doi:10.1128/mBio.00326-15.

Solomon, S., Qin, D., Manning, M., Averyt, K., and Marquis, M. (2007). *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. Cambridge university press.

Soper, J. H., and Powell, J. M. (1985). *Botanical studies in the Lake Hazen Region, northern Ellesmere Island, Northwest Territories, Canada*.

St. Pierre, K. A., St. Louis, V. L., Lehnherr, I., Gardner, A. S., Serbu, J. A., Mortimer, C. A., et al. (2019a). Drivers of Mercury Cycling in the Rapidly Changing Glacierized Watershed of the High Arctic's Largest Lake by Volume (Lake Hazen, Nunavut, Canada). *Environ. Sci. Technol.* 53, 1175–1185. doi:10.1021/acs.est.8b05926.

St. Pierre, K. A., St. Louis, V. L., Lehnherr, I., Schiff, S. L., Muir, D. C. G., Poulain, A. J., et al. (2019b). Contemporary limnology of the rapidly changing glacierized watershed of the world's largest High Arctic lake. *Scientific Reports*. doi:10.1038/s41598-019-39918-4.

Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* 224, 409–411. doi:10.1126/science.224.4647.409.

Staicu, L. C., and Barton, L. L. (2017). "Bacterial Metabolism of Selenium—For Survival or Profit," in *Bioremediation of Selenium Contaminated Wastewater*, ed. E. D. van Hullebusch (Cham: Springer International Publishing), 1–31. doi:10.1007/978-3-319-57831-6_1.

Staley, J., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology* 39, 321–346. doi:10.1146/annurev.mi.39.100185.001541.

Stearns, P. N. (2018). *The Industrial Revolution in World History*. 4th ed. Routledge doi:10.4324/9780429494475.

Stoeva, M. K., Aris-Brosou, S., Chételat, J., Hintelmann, H., Pelletier, P., and Poulain, A. J. (2014a). Microbial community structure in lake and wetland sediments from a High Arctic polar desert revealed by targeted transcriptomics. *PLOS ONE* 9, e89531. doi:10.1371/journal.pone.0089531.

Stoeva, M. K., Aris-Brosou, S., Chételat, J., Hintelmann, H., Pelletier, P., and Poulain, A. J. (2014b). Microbial community structure in lake and wetland sediments from a High Arctic polar desert revealed by targeted transcriptomics. *PLoS ONE* 9, e89531. doi:10.1371/journal.pone.0089531.

Sun, W., Xiao, E., Xiao, T., Krumins, V., Wang, Q., Häggblom, M., et al. (2017). Response of soil microbial communities to elevated antimony and arsenic contamination indicates the relationship between the innate microbiota and contaminant fractions. *Environ. Sci. Technol.* doi:10.1021/acs.est.7b00294.

Suomen metsäkeskus (2017). Valuma-alueen määritystyökalu. Available at: https://metsakeskus.maps.arcgis.com/apps/webappviewer/index.html?id=4ab572bdb631439d82f8aa8e0284f663 [Accessed November 15, 2018].

Surdu, C. M., Duguay, C. R., and 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, 941–960. doi:10.5194/tc-10-941-2016.

Suzuki, M. T., and Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62, 625–630.

Takai, K., and Horikoshi, K. (2000). Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl Environ Microbiol* 66, 5066–5072.

Tang, C., Madigan, M. T., and Lanoil, B. (2013). Bacterial and Archaeal diversity in sediments of West Lake Bonney, McMurdo Dry Valleys, Antarctica. *Appl. Environ. Microbiol.* 79, 1034–1038. doi:10.1128/AEM.02336-12.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41. doi:10.1186/1471-2105-4-41.

ten Hoopen, P., Finn, R. D., Bongo, L. A., Corre, E., Fosso, B., Meyer, F., et al. (2017). The metagenomic data life-cycle: standards and best practices. *Gigascience* 6, 1–11. doi:10.1093/gigascience/gix047.

Thaler, M., Vincent, W. F., Lionard, M., Hamilton, A. K., and 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. doi:10.3389/fmars.2016.00275.

Thamdrup, B., and Canfield, D. E. (2000). "Benthic Respiration in Aquatic Sediments," in *Methods in Ecosystem Science*, eds. O. E. Sala, R. B. Jackson, H. A. Mooney, and R. W. Howarth (New York, NY: Springer New York), 86–103. doi:10.1007/978-1-4612-1224-9_7.

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi:10.1093/nar/gkw1108.

Thienpont, J. R., Korosi, J. B., Hargan, K. E., Williams, T., Eickmeyer, D. C., Kimpe, L. E., et al. (2016). Multi-trophic level response to extreme metal contamination from gold mining in a subarctic lake. *Proc Biol Sci* 283. doi:10.1098/rspb.2016.1125.

Thomas, F., Hehemann, J.-H., Rebuffet, E., Czjzek, M., and Michel, G. (2011). Environmental and gut Bacteroidetes: The food connection. *Front Microbiol* 2. doi:10.3389/fmicb.2011.00093.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi:10.1038/nature24621.

Thomson Reuters (2019). Trash found littering ocean floor in deepest-ever submarine dive | CBC News. *CBC*. Available at: https://www.cbc.ca/news/world/trash-littering-ocean-floor-deepest-submarine-dive-1.5134717 [Accessed June 5, 2019].

Thrash, J. C. (2019). Culturing the Uncultured: Risk versus Reward. *mSystems* 4, e00130-19. doi:10.1128/mSystems.00130-19.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., et al. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform* 14, 315–326. doi:10.1093/bib/bbs034.

Tyson, R. (2012). *Sedimentary Organic Matter: Organic facies and palynofacies*. Springer Science & Business Media.

UNEP (2014). *Minamata convention on mercury*.

van den Berg, E. M., Elisário, M. P., Kuenen, J. G., Kleerebezem, R., and van Loosdrecht, M. C. M. (2017a). Fermentative Bacteria Influence the Competition between Denitrifiers and DNRA Bacteria. *Front Microbiol* 8. doi:10.3389/fmicb.2017.01684.

van den Berg, E. M., Rombouts, J. L., Kuenen, J. G., Kleerebezem, R., and van Loosdrecht, M. C. M. (2017b). Role of nitrite in the competition between denitrification and DNRA in a chemostat enrichment culture. *AMB Express* 7. doi:10.1186/s13568-017-0398-x.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605.

van der Walt, A. J., van Goethem, M. W., Ramond, J.-B., Makhalanyane, T. P., Reva, O., and Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18. doi:10.1186/s12864-017-3918-9.

Vavourakis, C. D., Andrei, A.-S., Mehrshad, M., Ghai, R., Sorokin, D. Y., and Muyzer, G. (2018). A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome* 6, 168. doi:10.1186/s40168-018-0548-7.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi:10.1126/science.1058040.

Vergin, K. L., Urbach, E., Stein, J. L., DeLong, E. F., Lanoil, B. D., and Giovannoni, S. J. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Appl Environ Microbiol* 64, 3075–3078.

Vigneron, A., Cruaud, P., Mohit, V., Martineau, M.-J., Culley, A. I., Lovejoy, C., et al. (2018). Multiple Strategies for Light-Harvesting, Photoprotection, and Carbon Flow in High Latitude Microbial Mats. *Front Microbiol* 9. doi:10.3389/fmicb.2018.02881.

Vigneron, A., Lovejoy, C., Cruaud, P., Kalenitchenko, D., Culley, A., and Vincent, W. F. (2019). Contrasting Winter Versus Summer Microbial Communities and Metabolic Functions in a Permafrost Thaw Lake. *Front. Microbiol.* 10, 1656. doi:10.3389/fmicb.2019.01656.

Vincent, W. F., and Laybourn-Parry, J. (2008). *Polar lakes and rivers: Limnology of Arctic and Antarctic aquatic ecosystems*. Oxford University Press.

Vo, A.-T. E., Bank, M. S., Shine, J. P., and Edwards, S. V. (2011). Temporal increase in organic mercury in an endangered pelagic seabird assessed by century-old museum specimens. *Proc Natl Acad Sci U S A* 108, 7466–7471. doi:10.1073/pnas.1013865108.

Vuori, K.-M. (2009). *Kokkolan edustan merialueen sedimenttien toksisuus ja ekologinen riskinarviointi*. Helsinki: Suomen ympäristökeskus: Edita Publishing.

Wang, D.-Z., Kong, L.-F., Li, Y.-Y., and Xie, Z.-X. (2016a). Environmental Microbial Community Proteomics: Status, Challenges and Perspectives. *Int J Mol Sci* 17. doi:10.3390/ijms17081275.

Wang, F., Outridge, P. M., Feng, X., Meng, B., Heimbürger-Boavida, L.-E., and Mason, R. P. (2019a). How closely do mercury trends in fish and other aquatic wildlife track those in the atmosphere? – Implications for evaluating the effectiveness of the Minamata Convention. *Science of The Total Environment* 674, 58–70. doi:10.1016/j.scitotenv.2019.04.101.

Wang, N. F., Zhang, T., Yang, X., Wang, S., Yu, Y., Dong, L. L., et al. (2016b). Diversity and composition of bacterial community in soils and lake sediments from an arctic lake area. *Front Microbiol* 7, 1170. doi:10.3389/fmicb.2016.01170.

Wang, N., Guo, Y., Li, G., Xia, Y., Ma, M., Zang, J., et al. (2019b). Geochemical-Compositional-Functional Changes in Arctic Soil Microbiomes Post Land Submergence Revealed by Metagenomics. *Microbes Environ.*, ME18091. doi:10.1264/jsme2.ME18091.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi:10.1128/AEM.00062-07.

Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E., and Simmonds, P. (2001). Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* 75, 11686–11699. doi:10.1128/JVI.75.23.11686-11699.2001.

Wang, Y., Boyd, E., Crane, S., Lu-Irving, P., Krabbenhoft, D., King, S., et al. (2011). Environmental Conditions Constrain the Distribution and Diversity of Archaeal merA in Yellowstone National Park, Wyoming, U.S.A. *Microb Ecol* 62, 739–752. doi:10.1007/s00248-011-9890-z.

Ward, N. L., Challacombe, J. F., Janssen, P. H., Henrissat, B., Coutinho, P. M., Wu, M., et al. (2009). Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. *Appl. Environ. Microbiol.* 75, 2046–2056. doi:10.1128/AEM.02294-08.

Watson, J. D., and Crick, F. (1953). A structure for deoxyribose nucleic acid.

Wei, H., Wang, J., Hassan, M., Han, L., and Xie, B. (2017). Anaerobic ammonium oxidation-denitrification synergistic interaction of mature landfill leachate in aged refuse bioreactor: Variations and effects of microbial community structures. *Bioresource Technology* 243, 1149–1158. doi:10.1016/j.biortech.2017.07.077.

Weiss, L. C., Pötter, L., Steiger, A., Kruppert, S., Frost, U., and Tollrian, R. (2018). Rising pCO2 in Freshwater Ecosystems Has the Potential to Negatively Affect Predator-Induced Defenses in Daphnia. *Current Biology* 28, 327-332.e3. doi:10.1016/j.cub.2017.12.022.

Wetterstrand, K. (2018). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute (NHGRI)*. Available at: www.genome.gov/sequencingcostsdata [Accessed April 21, 2019].

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088.

Wright, M. N., and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv:1508.04409 [stat]*.

Wright, M. N., and Ziegler, A. (2017). **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software* 77. doi:10.18637/jss.v077.i01.

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337, 1661–1665. doi:10.1126/science.1224041.

Wu, R. (1972). Nucleotide Sequence Analysis of DNA. *Nature New Biology* 236, 198–200. doi:10.1038/newbio236198a0.

Wurzbacher, C., Nilsson, R. H., Rautio, M., and Peura, S. (2017). Poorly known microbial taxa dominate the microbiome of permafrost thaw ponds. *ISME J* 11, 1938–1941. doi:10.1038/ismej.2017.54.

Xiong, J., Liu, Y., Lin, X., Zhang, H., Zeng, J., Hou, J., et al. (2012). Geographic distance and pH drive bacterial distribution in alkaline lake sediments across Tibetan Plateau. *Environ Microbiol* 14, 2457–2466. doi:10.1111/j.1462-2920.2012.02799.x.

Xu, Z., Malmer, D., Langille, M. G. I., Way, S. F., and Knight, R. (2014). Which is more important for classifying microbial communities: who's there or what they can do? *ISME J* 8, 2357–2359. doi:10.1038/ismej.2014.157.

Ye, Y., and Doak, T. G. (2009). A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLOS Computational Biology* 5, e1000465. doi:10.1371/journal.pcbi.1000465.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (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, 28–36. doi:10.1111/2041-210X.12628.

Yu, R.-Q., Reinfelder, J. R., Hines, M. E., and Barkay, T. (2018). Syntrophic pathways for microbial mercury methylation. *The ISME Journal* 12, 1826–1835. doi:10.1038/s41396-018-0106-0.

Yu, T., and Chen, Y. (2019). Effects of elevated carbon dioxide on environmental microbes and its mechanisms: A review. *Sci. Total Environ.* 655, 865–879. doi:10.1016/j.scitotenv.2018.11.301.

Zalasiewicz, J., Waters, C. N., Ivar do Sul, J. A., Corcoran, P. L., Barnosky, A. D., Cearreta, A., et al. (2016). The geological cycle of plastics and their use as a stratigraphic indicator of the Anthropocene. *Anthropocene* 13, 4–17. doi:10.1016/j.ancene.2016.01.002.

Zhang, C., Yang, L., Ding, Y., Wang, Y., Lan, L., Ma, Q., et al. (2017). Bacterial lipid droplets bind to DNA via an intermediary protein that enhances survival under stress. *Nature Communications* 8, 15979. doi:10.1038/ncomms15979.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014a). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi:10.1093/bioinformatics/btt593.

Zhang, J., Yang, Y., Zhao, L., Li, Y., Xie, S., and Liu, Y. (2014b). Distribution of sediment Bacterial and Archaeal communities in plateau freshwater lakes. *Appl Microbiol Biotechnol* 99, 3291–3302. doi:10.1007/s00253-014-6262-x.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, 203–214. doi:10.1089/10665270050081478.

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* 62, 316–322.

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S. G., and Alvarez-Cohen, L. (2015). High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. *mBio* 6, e02288-14. doi:10.1128/mBio.02288-14.

Zhu, Y.-G., Xue, X.-M., Kappler, A., Rosen, B. P., and Meharg, A. A. (2017). Linking Genes to Microbial Biogeochemical Cycling: Lessons from Arsenic. *Environ. Sci. Technol.* 51, 7326–7339. doi:10.1021/acs.est.7b00689.

Zolitschka, B., Behre, K.-E., and Schneider, J. (2003). Human and climatic impact on the environment as derived from colluvial, fluvial and lacustrine archives—examples from the Bronze Age to the Migration period, Germany. *Quaternary Science Reviews* 22, 81–100. doi:10.1016/S0277-3791(02)00182-8.

Appendix A: Supporting information for Chapter 2

A.1 Sequencing details

For spring 2014/2015 samples, the V3-V4 region of the bacterial 16S rRNA gene was sequenced in two separate runs using the primers 341F (5'-CCT ACG GGN GG CWG CAG-3') and 805R (5'-GAC TAC HVG GGT ATC TAA TCC-3'; Herlemann et al., 2011). The use of these primers is recommended for environmental studies (Klindworth et al., 2013). Summer 2015 samples were sequenced for the V3-V5 region of the archaeal 16S with the primers Arch349F (5'-GYG CAS CAG KCG MGA AW-3') and Arch806R (5'-GGA CTA CVS GGG TAT CTA AT-3'; Takai and Horikoshi, 2000). The same samples were also sequenced for the V1-V3 region of the bacterial 16S with primers 27Fmod (5'-AGR GTT TGA TCM TGG CTC AG-3'; Vergin et al., 1998) and 519Rmodbio (5'-GTN TTA CNG CGG CKG CTG-3'; Andreotti et al., 2011). The primers for each sequencing were used in a 28 cycle PCR with barcodes on the forward primer, using the HotStarTaq Plus Master Mix Kit (Qiagen, Germantown, MD, USA). Temperature cycling in PCR consisted of: 94°C, 3 min; 28 cycles of {94°C, 30 s; 53°C, 40 s; 72°C, 1 min}, with a final elongation at 72°C for 5 min. The PCR products were then checked in a 2% agarose gel and pooled in equal proportions based on the molecular weights and DNA concentrations of the products, followed by purification with calibrated Ampure XP beads (Beckman Coulter, CA, USA). All sequencing was performed on the Illumina MiSeq platform producing paired-end 300 bp reads according to the manufacturer's guidelines, by Molecular Research LP (Shallowater, TX, USA).

A.2 Assessing contamination from DNA extraction kits

Briefly, putative contaminating genera from MOBIO PowerSoil kit (Glassing et al., 2016) and four other extraction kits (Salter et al., 2014) not used in this study were identified from the assigned taxonomy of the OTUs in all our data (**Figure A4**). The complete data analysis script was then run with either 100% of the putative MOBIO PowerSoil contaminants or 10%, 20%, 50% and 100% of abundances of all the putative contaminants removed from the OTU tables. None of our primary conclusions (from analysis of alpha- and beta-diversity, *t*-SNE clustering, and differential analyses of taxa and functionally predicted groups along physicochemical gradients) were affected by removal of 100% of the putative MOBIO PowerSoil kit contaminants (**Figures A5**, **A6**). Furthermore, no effects were observed when 10% of abundances of all the putative contaminant genera were removed, but at 20% reduced abundance, the clustering patterns and physicochemical gradient analyses were visibly affected (data not shown). Beta-diversity analyses were more robust to reducing the abundances up to 20%, but at 50% reduced abundance water depth did not significantly correlate with the phylogenetic differences in the spring 2014/2015 data set. Results of our alpha-diversity analyses were not affected by even 100% removal of all the putative contaminant genera (data not shown).

A.3 Overview of the microbial community structure

Lake Hazen sediments in spring 2014/2015 were dominated by Proteobacteria (38%; **Figure 2.1**). The largest Proteobacterial classes in Lake Hazen were Alphaproteobacteria (12%) and Betaproteobacteria (11%). Other major phyla were Bacteroidetes (10%), Acidobacteria (8%), Chloroflexi (7%), and Actinobacteria (7%). The most abundant archaeal phyla in Lake Hazen

sediments was Woesearchaeota (formerly "DHVEG-6"; 0.2%, including bacteria in the total), and Thaumarchaeota (0.02%). Skeleton Lake sediments were also dominated by Proteobacteria (32%), but the classes differed from those in Lake Hazen: Deltaproteobacteria (9%), Gammaproteobacteria (9%) and Alphaproteobacteria (8%). Other major phyla in Skeleton Lake sediments were Chloroflexi (12%), Actinobacteria (11%), Bacteroidetes (8%), Planctomycetes (6%), and Cyanobacteria (6%). The most abundant archaeal phyla in Skeleton Lake sediments were Woesearchaeota (1%, including bacteria), followed by Euryarchaeota (0.9%).

In the summer 2015 sediment samples from Skeleton Lake, the archaeal community primarily consisted of Woesearchaeota (67%), Euryarchaeota (17%) and the Miscellaneous Euryarchaeotic Group (MEG; 16%; **Figure 2.2**). The Pond1 archaeal community mostly consisted of Euryarchaeota (47%), Woesearchaeota (32%), MEG (12%), and Thaumarchaeota (6%). The most abundant bacterial phyla in the 2015 summer Skeleton Lake sediments were Chloroflexi (20%), Bacteroidetes (20%), Proteobacteria (16%, in order of abundance: Beta-, Alpha-, Delta-, and Gamma-), Cyanobacteria (9%), Gracilibacteria (8%), and SR1 (Absconditabacteria; 5%; **Figure 2.2**). The most abundant bacterial phyla in Pond1 sediments were Chloroflexi (24%), Proteobacteria (18%, in order of abundance: Alpha-, Delta-, Beta- and Gamma-), Bacteroidetes (15%), Cyanobacteria (15%), and Firmicutes (6%).

The most common functionally mapped groups in Lake Hazen in spring 2014/2015 were aerobic chemoheterotrophs (while ranking second in Skeleton Lake), and cyanobacteria (as they are grouped together at phylum level in FAPROTAX) in Skeleton Lake in spring 2014/2015 (**Figure A8**). Besides cyanobacteria, sulfate reducers were the second most common functionally mapped group in Skeleton Lake sediments, while almost absent from sediments in Lake Hazen. Finally, aerobic ammonia and nitrite oxidizers, as well as intracellular parasites were more

prevalent in Lake Hazen than in Skeleton Lake sediments. Overall, functionally mapped groups associated with aerobic metabolism seemed to be more prevalent in Lake Hazen than in Skeleton Lake sediments.

Skeleton Lake and Pond1 had somewhat similar functional predictions, as methanogenesis dominated in both sediments (**Figure A9**). About 2/3 of the archaeal functional mapping groups were methanogenic; mercury methylators, and nitrogen fixing archaea were present at both sites. In the bacterial data from 2015, cyanobacteria, and aerobic chemoheterotrophs were the most abundant functional predictions.

A.4 Taxonomic and functionally predicted group abundances along physicochemical gradients

The best random forest models had pseudo-R^2 values between -0.15 (*i.e.*, worse than fitting a straight line) and 0.97 with 1 to 16 predictors for continuous physicochemical variables (**Figures A15**−**A32**). For the models of categorical variables, the OOB error rates varied from 0% to 7.14% with 1 to 11 predictors (**Figures A33**−**A37**). Most of the taxonomic models had the best prediction accuracy on the order level (8 out of 20 models), which was the lowest taxonomic rank included. The predictions of taxonomic data were generally better than functionally mapped data for the same variable. The average improvement in MSPE from functionally mapped to taxonomic data was 66% for spring 2014/2015, 434% for summer 2015 archaeal data (improvement in the Cl^- model from 0.27 to 0.01 MSPE was highly influential) and 54% for summer 2015 bacterial data. The improvement was probably caused by the lower coverage of the functional mapping of the OTUs in the data since random forests usually perform better with more data. The high improvement in the archaeal data set was mostly

caused by the difference between the two [Cl^-] models. Highest pseudo-R^2 values from the models were obtained for [H_2S] (0.96−0.97, only spring 2014/2015), water depth (0.71−0.93, only spring 2014/2015), [NO_3^-] (0.87−0.90, only summer 2015), pH (0.48−0.88), sediment depth (0.40−0.86) and redox potential (0.74−0.79, only spring 2014/2015). All variables except [O_2] in summer 2015 (pseudo-$R^2 \leq 0$), could be linked to changes in taxonomical and/or functional mapping group abundances along their gradients. Variability in the predictors explained by our regression random forest models was high for both taxonomic and functional mapping groups. The few most important groups in the models were selected among up to 280 taxonomic or up to 48 functional mapping groups. They also represented groups that we expected to have the strongest response to the variable in question.

Trends of observed relationships between phylogenetic and functional mapping groups to the most relevant physicochemical variables ([H_2S], redox potential, pH, water depth, and [NO_3^-]) are described below, with further detail than in the main manuscript.

A.4.1 Continuous variables (regression rf)

Higher [H_2S] was linked to lower abundance of 16 bacterial classes, *e.g.*, within Acidobacteria and Bacteroidetes in the spring 2014/2015 data set (**Figure A15a**). These taxa are primarily aerobic heterotrophs (for example, Kämpfer, 2015; Rapp et al., 2016; Ward et al., 2009). The functional mapping showed a lower abundance of aerobic ammonia oxidizers, chemoheterotrophs and nitrite oxidizers, together with predatory or exoparasitic microbes in the spring 2014/2015 data set (**Figure A15b**). In the spring 2014/2015 functionally mapped data, higher [H_2S] correlated with increasing abundances of cyanobacteria, methanogens, and sulfate respirers (**Figure A15b**).

In the spring 2014/2015 phylogenetic data, the abundance of 4 orders within the phylum Acidobacteria, the order Gemmatimonadales, and the orders Methylophilales and TRA3-20 within Betaproteobacteria increased with increasing redox potential (**Figure A18a**). However, the abundance of order Anaerolineales decreased with increasing redox (**Figure A18a**). In the spring 2014/2015 functionally mapped data, methanol oxidizers and sulfur respirers had positive trends with higher redox (**Figure A18b**). The only group decreased in abundance with increasing redox potential was the strictly anaerobic order Anaerolineales from the phylum Chloroflexi.

In the spring 2014/2015 phylogenetic data, increasing pH was linked with increasing abundance of the order Ignavibacteriales, and uncultured/unknown orders within phyla Omnitrophica and Verrucomicrobia (**Figure A16a**). Orders Frankiales, Coriobacteriales, Bacteriodales, Rhodobacterales, and Desulfuromonadales decreased in abundance with increasing pH in the spring 2014/2015 data set (**Figure A16a**). In the summer 2015 archaeal communities, the abundance of an uncultured class in phylum Woesearchaeota (DHVEG-6) increased with increasing pH (**Figure A21a**). An uncultured class within the Miscellaneous Euryarchaeotic Group (MEG) decreased in abundance with increasing pH (**Figure A21a**). In the bacterial communities, the abundances of phyla Cyanobacteria and Proteobacteria increased with increasing pH (**Figure A27a**). 17 phyla decreased in abundance with increasing pH, for example, Actinobacteria, Bacteroidetes, Chlorobi, Chloroflexi, Firmicutes, Gracilibacteria, Microgenomates, and Absconditabacteria (**Figure A27a**). Iron respirers, photoheterotrophs, and sulfate respirers decreased in abundance with increasing pH (**Figure A16b**). In summer 2015 archaeal data, abundances of mercury methylators were positively related to pH (**Figure A21b**). Also, several methanogenic groups had a negative relationship with pH (**Figure A21b**). In summer 2015 bacterial data, increasing pH was linked with increasing abundances of aerobic

ammonia oxidizers, anoxygenic phototrophs and chemoheterotrophs, cyanobacteria, and ureolytic microbes (**Figure A27b**). However, sulfur oxidizers, fermenters, iron respiring and sulfate respiring microbes decreased in abundance with increasing pH (**Figure A27b**).

Our results correspond to previous results on the role of pH as a factor controlling lake sediment microbial community structure and diversity (Xiong et al., 2012). However, in the current study the abundance of an Alphaproteobacterial order declined towards higher pH (**Figure A16a**), contrary to previous observations (Xiong et al., 2012). Our results are also somewhat contrary to positive relationships of Actinobacteria and Bacteroidetes with pH in arctic soil (Chu et al., 2010). However, the discrepancies might be explained by the different pH ranges and taxonomic levels examined in these studies.

In the spring 2014/2015 data two classes, Subgroup 2 within phylum Acidobacteria, and Flavobacteria, had higher abundance at deeper sites (**Figure A20a**). From functional mapping groups, fermenters and intracellular parasites, also had positive relationship with water depth, while iron respirers displayed a U-shaped trend (either low or high abundance at deeper sites; **Figure A20b**).

Higher [NO_3^-] in the summer 2015 archaeal data was affiliated with increased abundance of 5 uncultured or unidentified orders within the Soil Crenarchaeotic Group, and decreased abundance of orders Methanomicrobiales, Methanosarcinales, and Thermoplasmatales (**Figure A26a**). In the summer 2015 archaeal data the abundances of an uncultured Bacteroidetes order, and the orders Chlorobiales, Fibrobacterales, Burkholderiales, and Desufovibrionales were lower at higher [NO_3^-] (**Figure A32a**). In the summer 2015 archaeal functionally mapped data, aerobic ammonia oxidizers and methanogens (disproportionation of methyl groups) had higher

abundances at higher [NO_3^-] (**Figure A26b**). In the summer 2015 bacterial functionally mapped data, fermenters and photoheterotrophs had lower abundances at higher [NO_3^-] (**Figure A32b**).

A.4.2 Categorical variables (classification rf)

The most important taxa for classifying samples between the four sites in the spring 2014/2015 data were classes Subgroup 2 within phylum Acidobacteria, and Flavobacteria (**Figure A33a**). The most important functional mapping groups for classifying samples by sites in the spring 2014/2015 data were aerobic chemoheterotrophs, cyanobacteria, dark oxidizers of sulfur compounds (including sulfide and thiosulfate oxidation), fermenters, intracellular parasites, nitrate denitrifiers, photoheterotrophs, exoparasites, and ureolytic microbes (**Figure A33b**). In summer 2015 data, the archaeal phyla Bathyarchaeota, Thaumarchaeota, and Woesearchaeota (DHVEG-6; **Figure A36a**), together with the bacterial phylum Gracilibacteria were most important for classifying the sites (**Figure A37a**). In summer 2015 data sets, the functionally predicted methanogens (by disproportionation of methyl groups; **Figure A36b**), anoxygenic photoautotrophs, fermenters, and methanotrophs were most important in differentiating between the two sites (**Figure A37b**).

In the spring 2014/2015 data, the abundance changes in a single group of microbes, methanogenic Euryarchaeota, could be used to divide the samples accurately (OOB Error = 0%) by origin to Lake Hazen and Skeleton Lake; the group was absent in Lake Hazen and present in Skeleton Lake (**Figure A34**).

A.5 Sediment samples

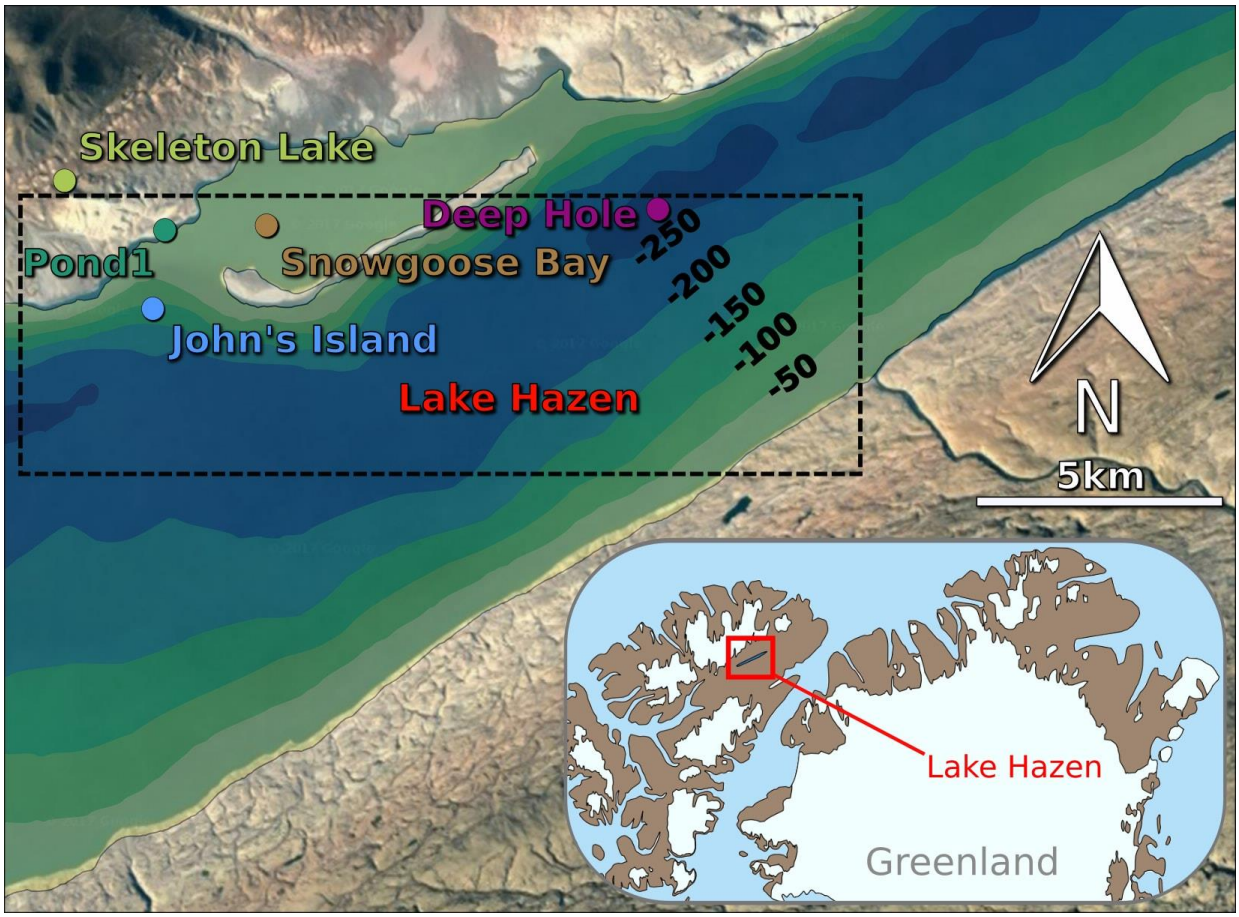

Figure A1: Map of the sampling sites and location of the sampling area. Enclosed depth map: (Köck et al., 2012) map data: Google Earth / Terrametrics (2017).

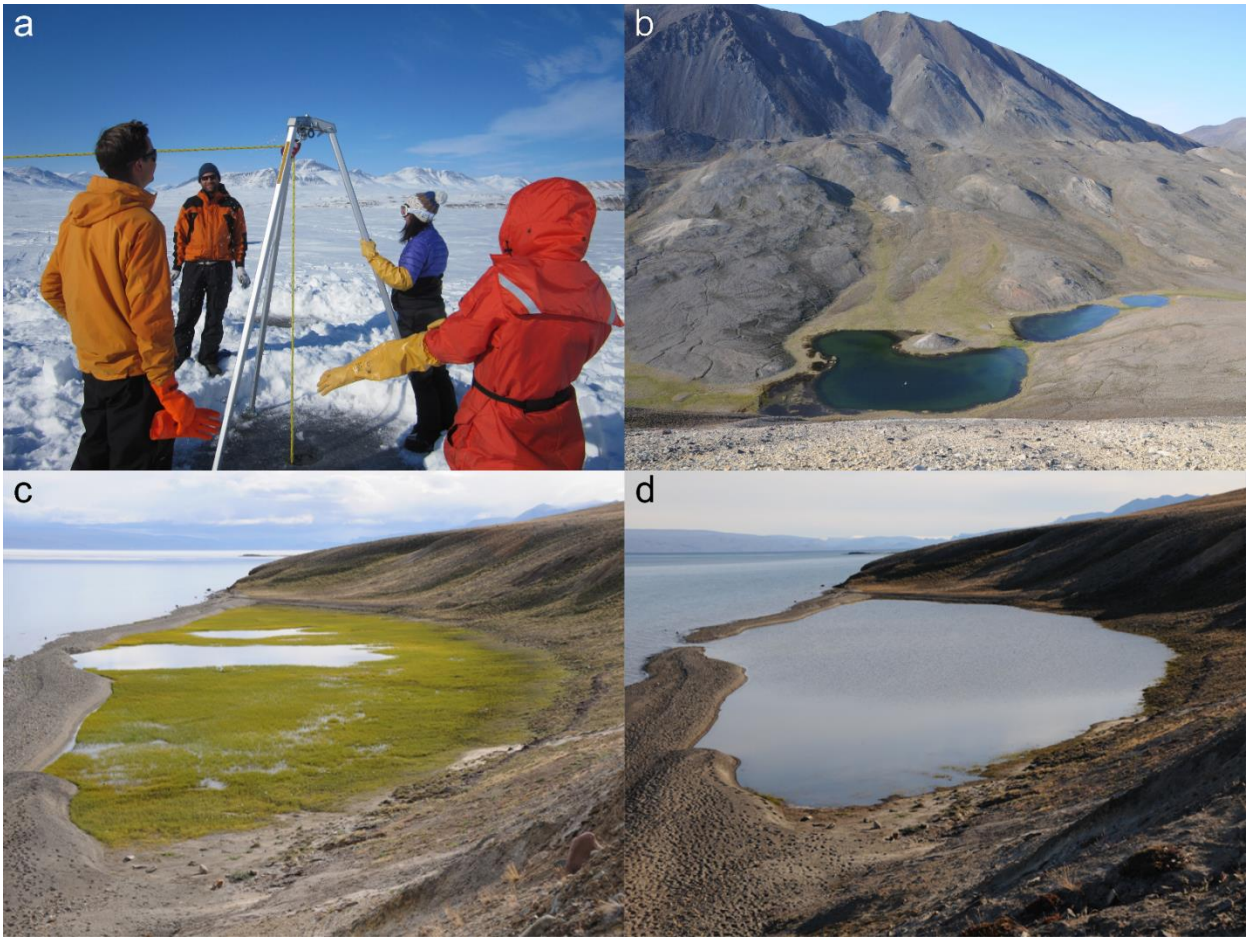

Figure A2: Photographs of the sampling area. (**a**) Sampling at the Deep Hole site on Lake Hazen in spring 2015, looking NW. (**b**) View over Skeleton Lake (leftmost) in summer 2015. (**c**) Pond1 on July 15th, 2010, looking SW towards Lake Hazen (on the left), showing the development of vegetation. (**d**) Pond1 on July 19th, 2010, showing a change in water level after formation of the hydrological connection to Lake Hazen.

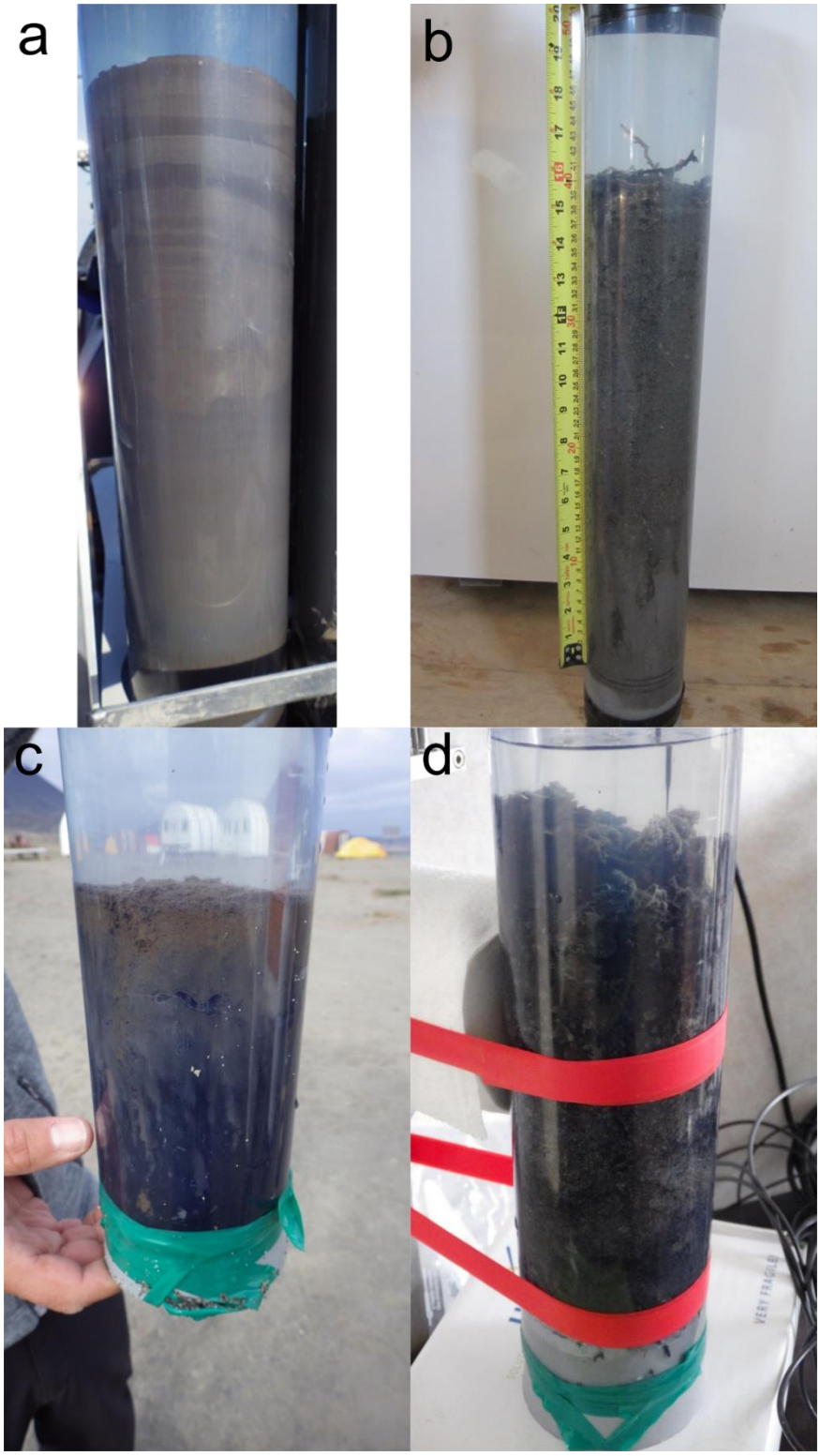

Figure A3: Photos of sediment cores from (**a**) Lake Hazen Deep Hole in spring 2014, (**b**) Skeleton Lake (deeper site) in spring 2015, (**c**) Pond1 in summer 2015, and (**d**) Skeleton Lake (shallower site) in summer 2015.

Table A1: Physicochemical data for spring 2014/2015 samples. Values with higher resolution than sampling for DNA extraction were averaged over the range in question.

Lake	Site	Year	Water depth (m)	Sediment depth (cm)	H_2S (µM)	pH	Redox (mV)	O_2 (mgL^{-1})
Lake Hazen	Snowgoose Bay	2014	44	0.25	0.37	7.78	181.02	9.49
				0.5	0.66	7.82	147.86	6.78
				0.75	1.01	7.85	139.69	5.04
				1	1.17	7.87	133.67	3.69
		2015	50	1	0	7.57	444.72	0.34
				2	0.33	7.65	423.43	0
				3	0	7.61	408.65	0
				4	0.23	7.57	379.81	0
				5	0.79	7.54	351.84	0
	Deep Hole	2014	258	0.25	0	8.40	112.85	4.19
				0.5	0	8.39	125.02	1.81
				0.75	0	8.37	114.62	0.20
				1	0	8.37	100.41	0
		2015	261	1	0	7.04	444.35	0.71
				2	0	7.02	429.24	0
				3	0	7.01	416.65	0
				4	0	7.01	397.13	0
				5	0	7.02	383.14	0
	John's Island	2015	141	1	0	8.68	369.22	11.77
				2	0	8.92	397.32	8.35
				3	0	9.01	407.65	6.80
				4	0	8.95	411.88	5.70
				5	0	8.90	413.02	4.67
Skeleton Lake		2015	4	1	119.31	7.35	189.38	0
				2	169.83	7.30	204.96	0
				3	169.83	7.27	91.41	0
				4	169.83	7.19	135.81	0
				5	169.83	7.15	127.47	0

Table A2: Physicochemical data for summer 2015 samples. Values with higher resolution than sampling for DNA extraction were averaged over the range in question.

Site	Water depth (m)	Sediment depth (cm)	pH	O_2 (mgL^{-1})	NO_3^- (mgL^{-1})	Cl^- (mgL^{-1})	SO_4^{2-} (mgL^{-1})
Skeleton Lake	0.3	0.5	7.15	0	4.48	3.51	91.38
		1	7.11	0	4.48	3.51	91.38
		1.5	7.08	0	4.66	3.22	84.34
		2	7.04	0	4.66	3.22	84.34
		2.5	6.99	0	4.65	3.07	86.11
		3	6.96	0	4.65	3.07	86.11
		3.5	6.92	0	4.02	3.19	87.55
		4	6.86	0	4.02	3.19	87.55
		4.5	6.82	0	4.19	2.72	121.51
		5	6.79	0	4.19	2.72	121.51
		5.5	6.77	0	4.34	3.12	104.62
		6	6.76	0	4.34	3.12	104.62
Pond1	1.5	0.5	8.04	0.27	5.49	4.38	70.03
		1	7.4	0	5.49	4.38	70.03
		1.5	7.2	0	5.26	2.07	87.14
		2	7.005	0	5.26	2.07	87.14
		2.5	6.925	0	5.48	2.67	91.83
		3	6.855	0	5.48	2.67	91.83
		3.5	6.81	0	5.27	2.59	71.69
		4	6.79	0	5.27	2.59	71.69

A.6 Data analysis quality control

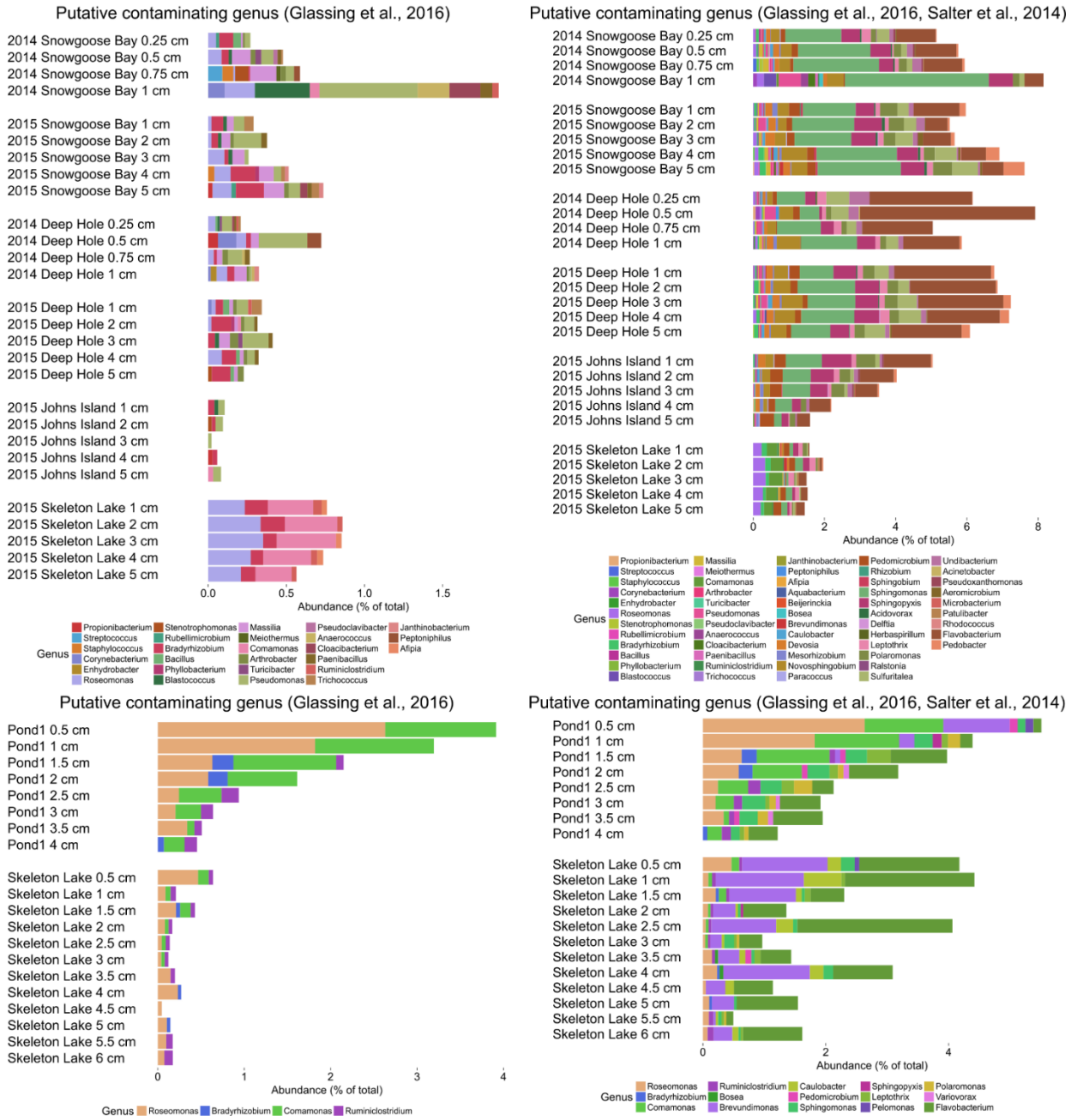

Figure A4: Per sample abundance of putative contaminant genera in spring 2014/2015 samples and summer 2015 samples identified on the left panel in the MOBIO PowerSoil kit (Glassing et al., 2016) and on the right panel in other kits and reagents (Salter et al., 2014). Only bacterial data is shown for summer 2015 since no putative contaminant archaea have been identified.

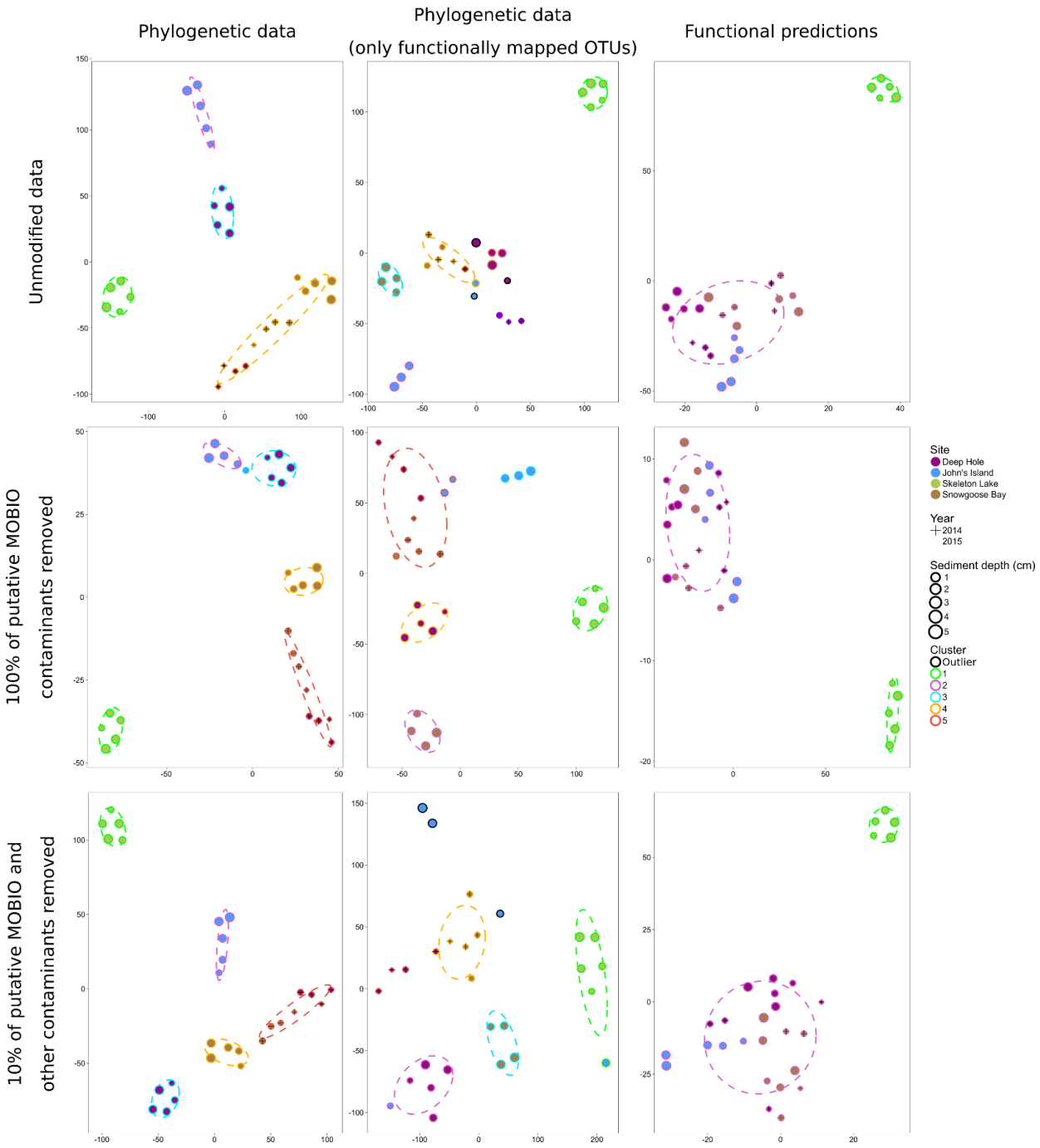

Figure A5: Effects of removing putative contaminant genera on the *t*-SNE clustering patterns in spring 2014/2015 data. Unmodified data compared to analyses where 100% of genera identified in the MOBIO PowerSoil kit (Glassing et al., 2016) and 10% of both MOBIO and other kit contaminants (Salter et al., 2014) were removed.

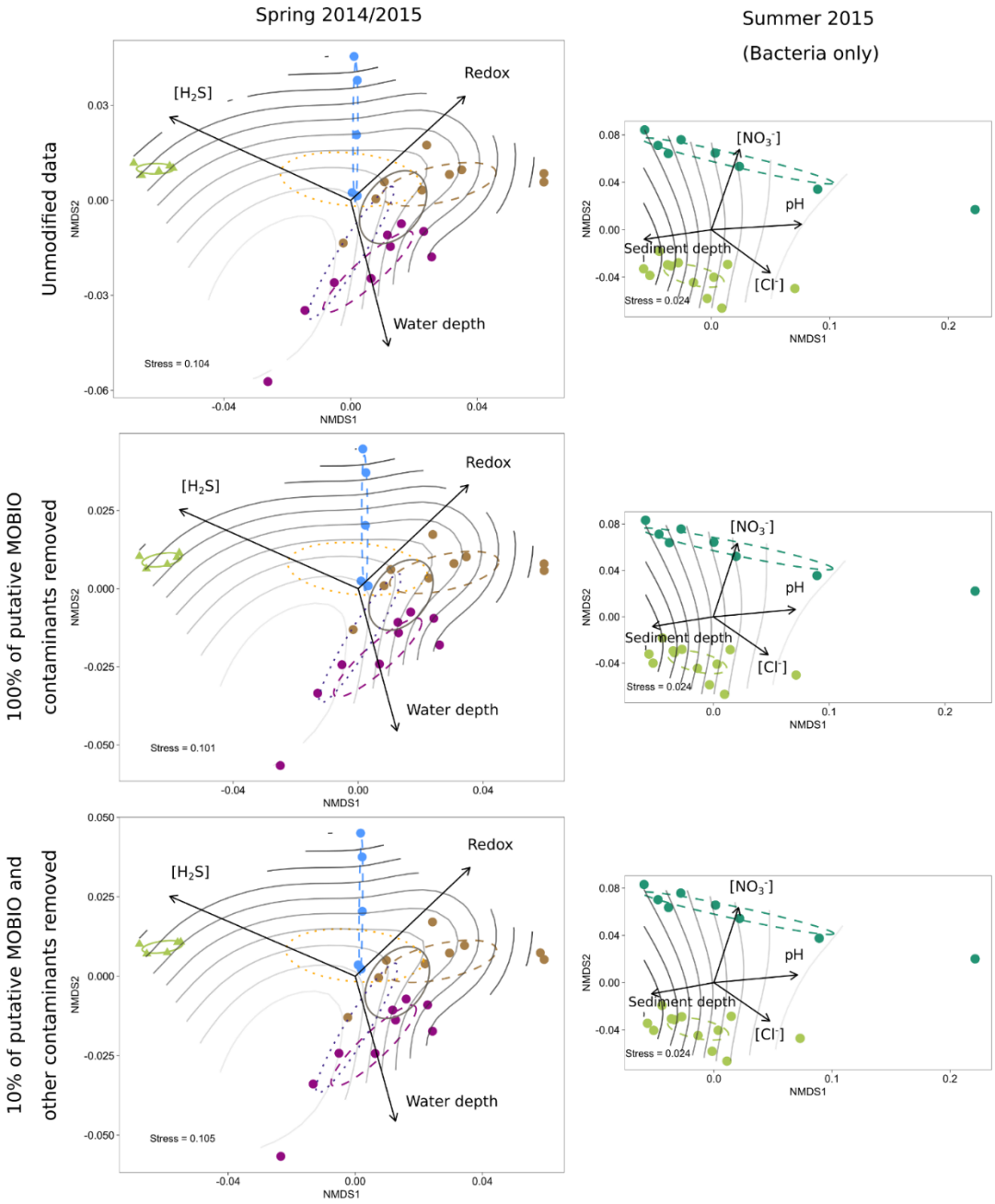

Figure A6: Effects of removing putative contaminant genera on the NMDS ordination patterns in spring 2014/2015 and summer 2015 bacterial data. Unmodified data compared to analyses where 100% of genera identified in the MOBIO PowerSoil kit (Glassing et al., 2016) and 10% of both MOBIO and other kit contaminants (Salter et al., 2014) were removed. Only bacterial data is shown for summer 2015 since no putative contaminant archaea have been identified.

Table A3: Number of reads and OTUs after each handling step.

	Data set:		
	spring 2014/2015	summer 2015 archaeal	summer 2015 bacterial
Raw reads	5,395,074	4,894,213	4,753,402
Reads after pairing	5,357,889	4,723,603	4,333,772
Reads after sample pruning & QC	560,323	390,236	148,030
Reads after chimera picking	482,308	361,960	132,832
Unique (dereplicated) reads	240,833	169,365	87,590
Total clusters (OTUs)	75,600	36,177	39,359
Non-singleton OTUs	16,933	4,865	5,415
OTUs after sample pruning & taxonomic QC	15,176	1,067	5,240
OTUs with > 0.1 ‰ overall abundance	2,655	1,067	2,928
Functionally mapped OTUs	3,772 (25%)	153 (14%)	1,238 (24%)
Unique functional mapping groups	48	10	26

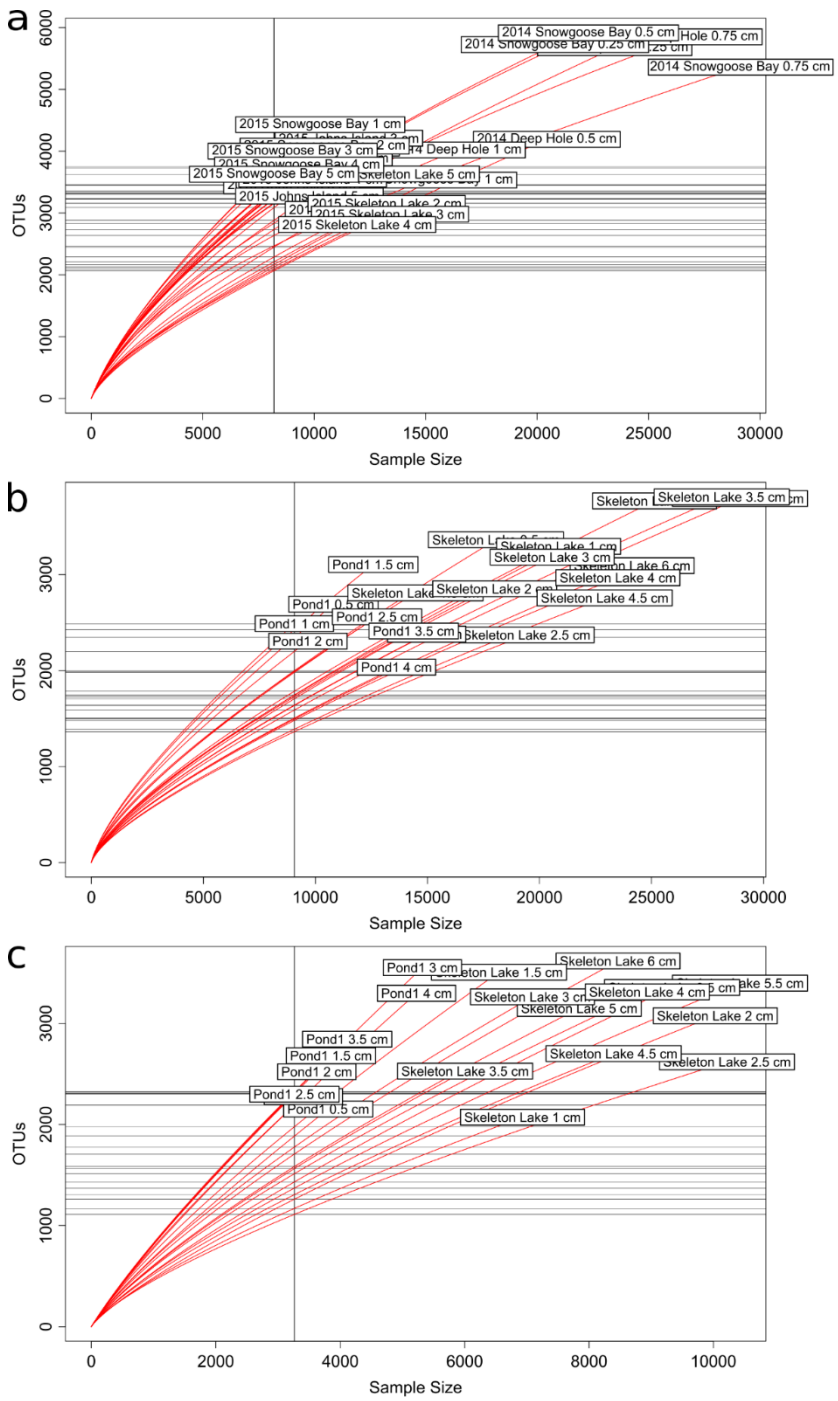

Figure A7: Rarefaction curves of the data sets showing number of OTUs as a function of rarefied read counts on non-normalized data. Vertical line shows the smallest number of reads in a sample in the data set in question, while horizontal lines show number of OTUs retained for samples at this rarefaction depth. (**a**) Spring 2014/2015 with universal primers. (**b**) Summer 2015 with archaeal primers. (**c**) Summer 2015 with bacterial primers.

A.7 Functional mapping

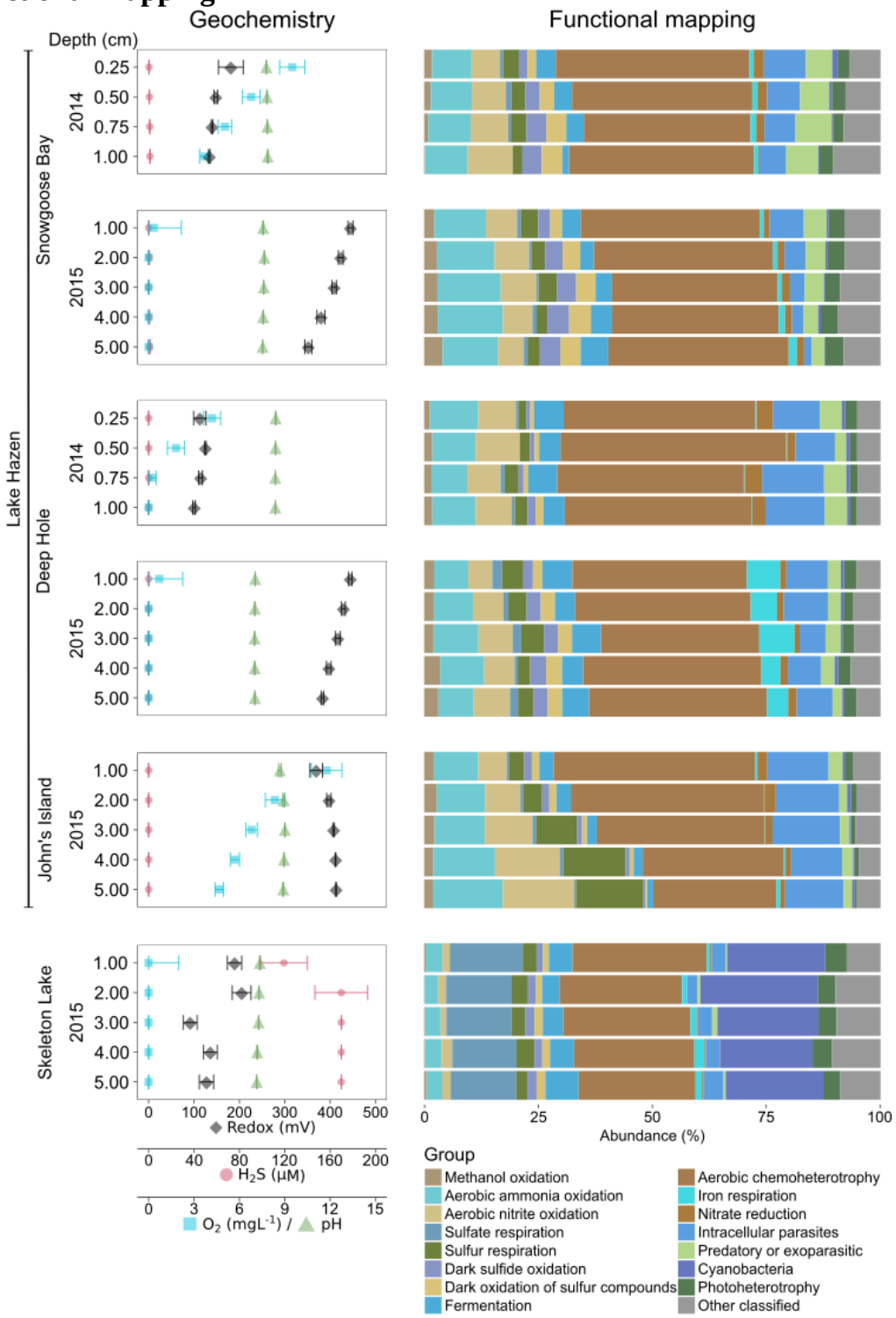

Figure A8: Geochemical variability, and functional mapping group composition of the spring 2014/2015 samples using universal primers. Groups with less than 1% overall abundance in the data set are merged as "Other classified".

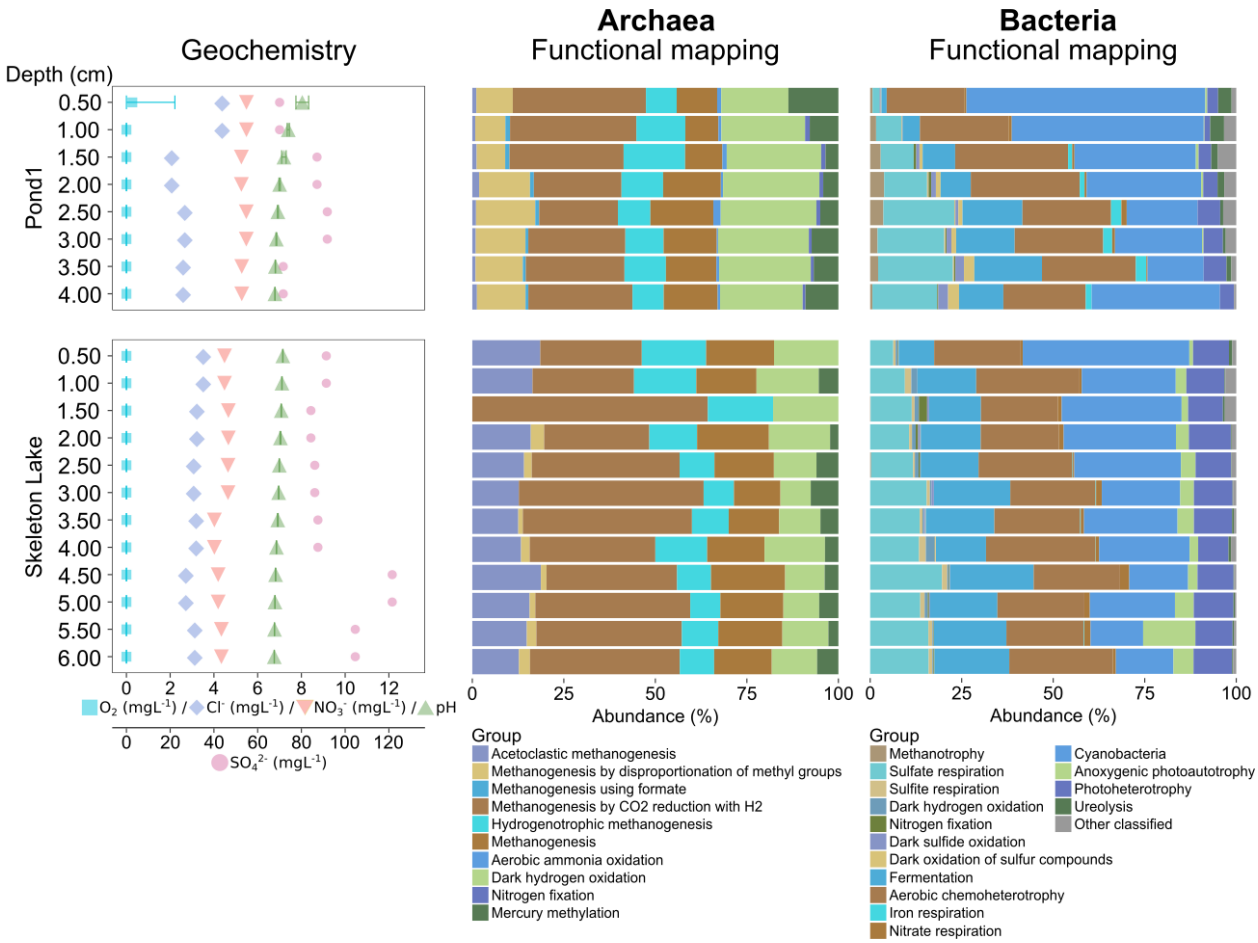

Figure A9: Geochemical variability, and functional mapping group composition of the summer 2015 samples using archaeal and bacterial primers. Groups with less than 1% overall abundance in each data set are merged as “Other classified”.

A.8 Alpha- and beta-diversity

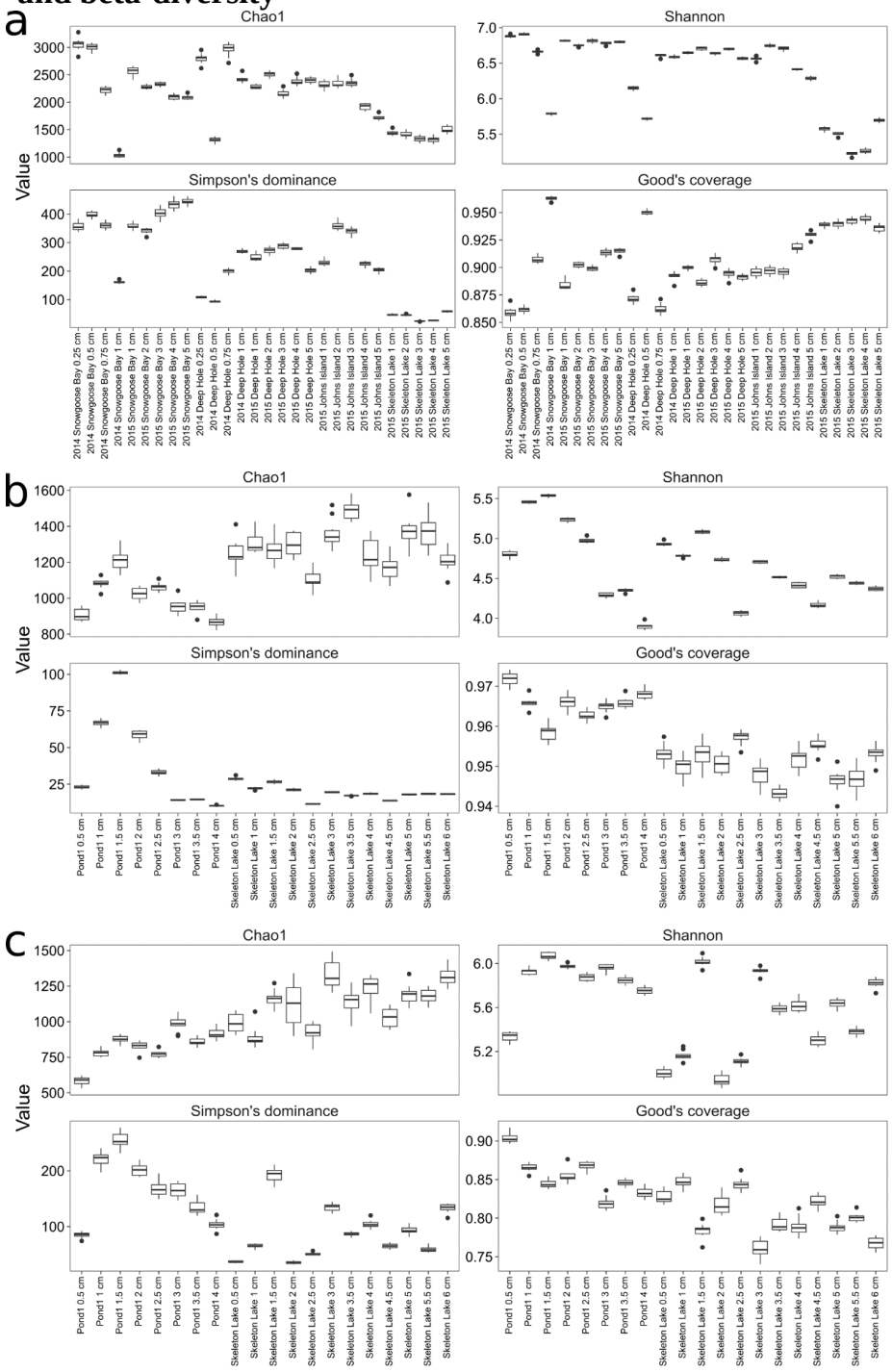

Figure A10: Richness and diversity of the communities summarized by Chao1, Shannon, and Simpson's Dominance (InvSimpson) indices, with Good's Coverage. (**a**) Spring 2014/2015 with universal primers. (**b**) Summer 2015 with archaeal primers. (**c**) Summer 2015 with bacterial primers.

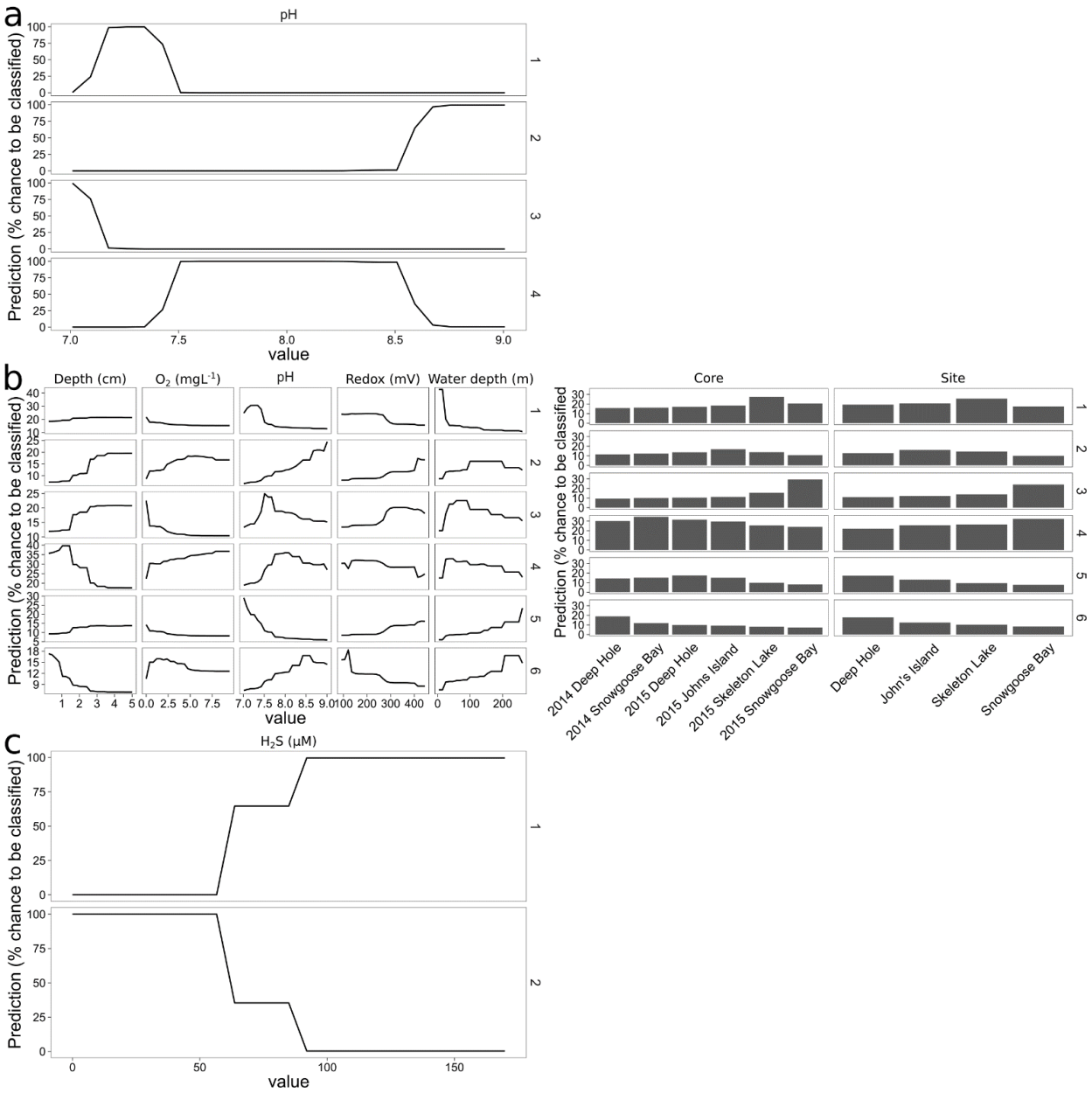

Figure A11: Partial dependence of random forest model prediction of cluster group (after recursive feature elimination) from spring 2014/2015 data set with universal primers. (**a**) Data set only including OTUs with > 0.1 ‰ overall abundance, with DPCoA distance matrix. (**b**) Data set only including OTUs that were matched to a function through FAPROTAX, with DPCoA distance matrix. (**c**) Functionally mapped data, with Bray-Curtis distance matrix.

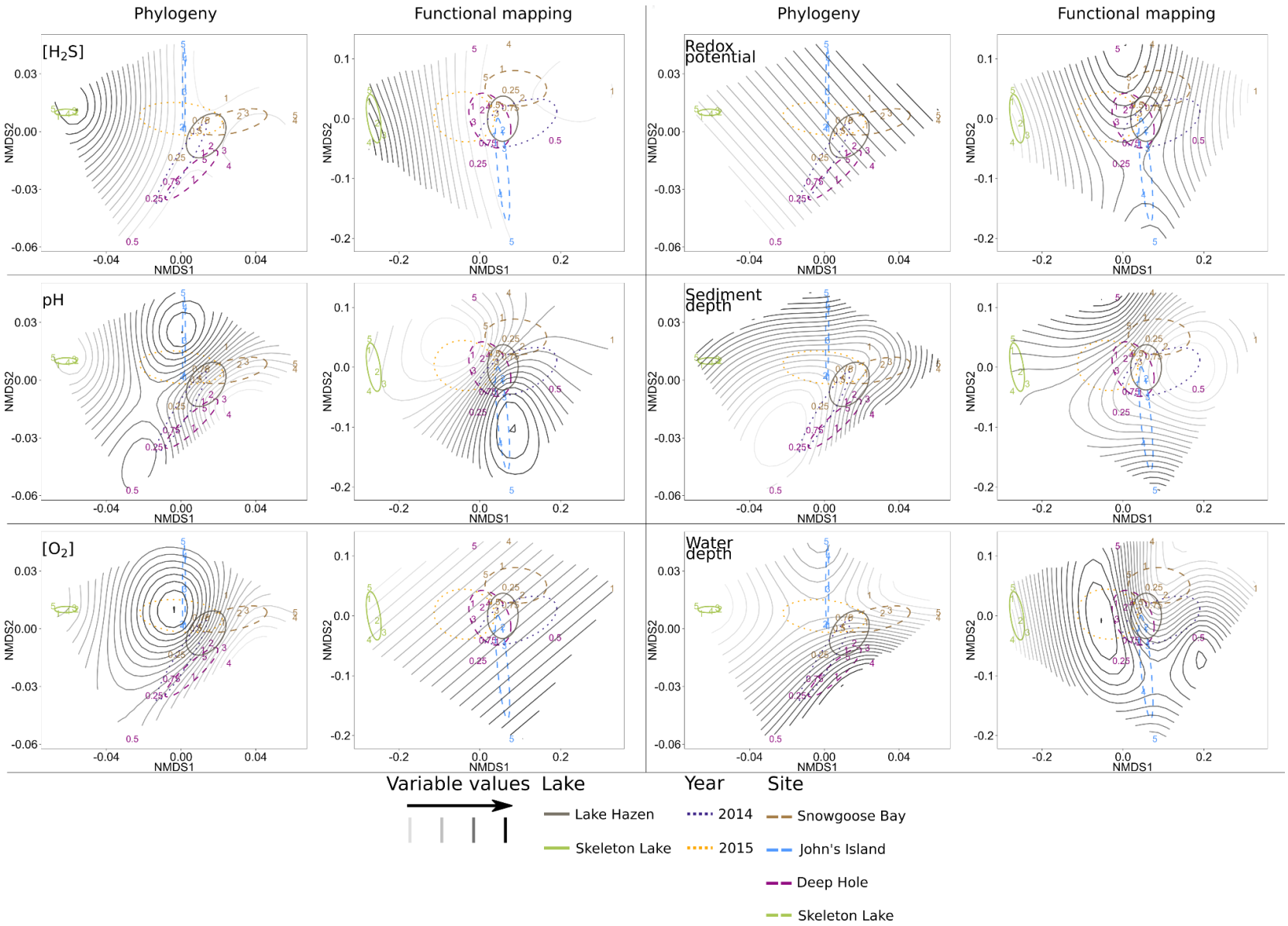

Figure A12: NMDS ordinations of the spring 2014/2015 data. Phylogenetic distances of samples through DPCoA ("Phylogeny" columns) and Bray-Curtis dissimilarity of the functional mappings ("Functional mapping" columns), with surface fits of all measured physicochemical variables, as indicated on the phylogeny column, on rows. 95%-confidence interval for centroids of sample categories (lake, year, and site) are shown with ellipses and lower extent of sampling depth in cm is indicated as the marker of each data point.

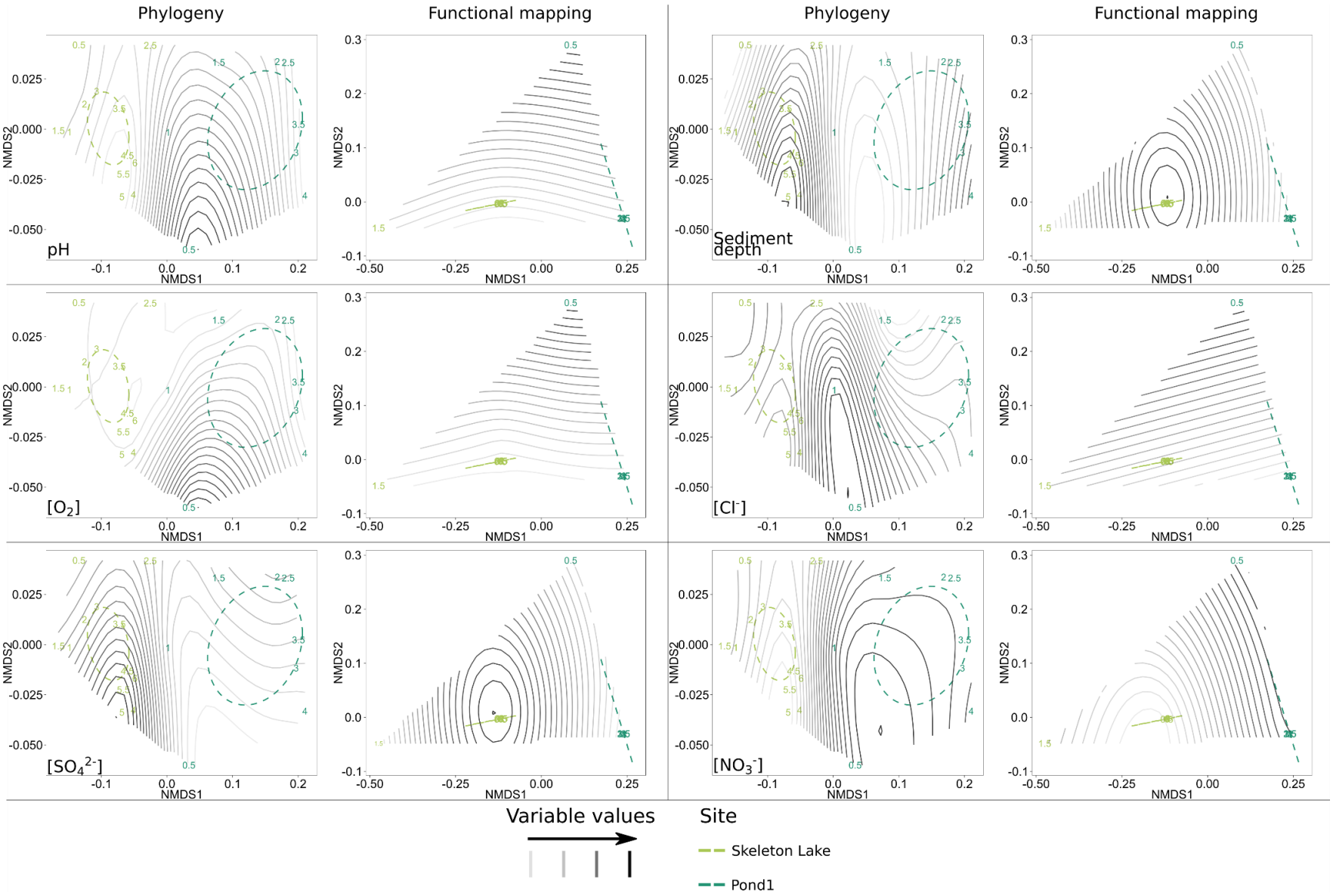

Figure A13: NMDS ordinations of the summer 2015 archaeal data. Phylogenetic distances of samples through DPCoA ("Phylogeny" columns) and Bray-Curtis dissimilarity of the functional mappings ("Functional mapping" columns), with surface fits of all measured physicochemical variables, as indicated on the phylogeny column, on rows. 95%-confidence interval for centroids of sampled sites are shown with ellipses and lower extent of sampling depth in cm is indicated as the marker of each data point.

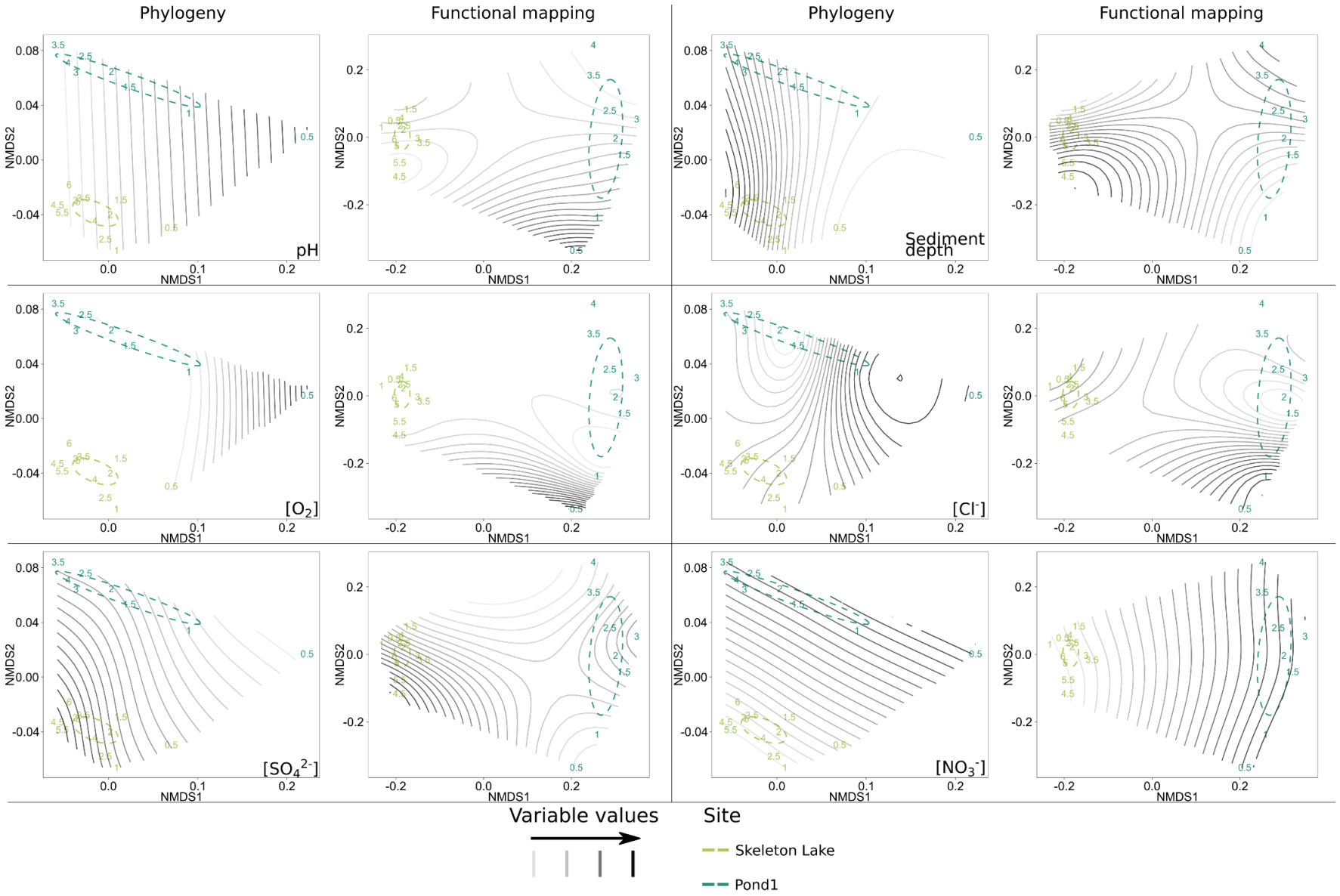

Figure A14: NMDS ordinations of the summer 2015 bacterial data. Phylogenetic distances of samples through DPCoA ("Phylogeny" columns) and Bray-Curtis dissimilarity of the functional mappings ("Functional mapping" columns), with surface fits of all measured physicochemical variables, as indicated on the phylogeny column, on rows. 95%-confidence interval for centroids of sampled sites are shown with ellipses and lower extent of sampling depth in cm is indicated as the marker of each data point.

A.9 Partial dependence plots of gradient analysis random forests

A.9.1 Continuous variables: spring 2014/2015

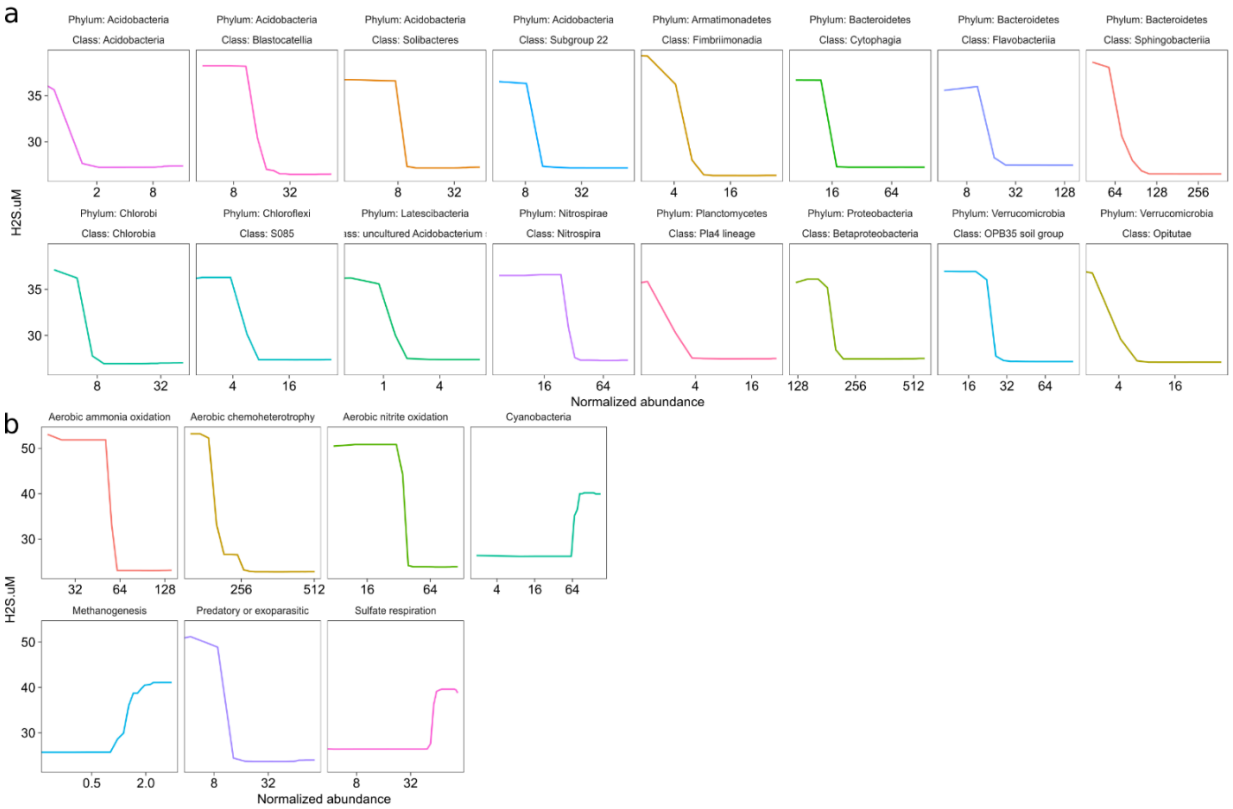

Figure A15: Partial dependence of random forest model prediction of [H_2S] from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

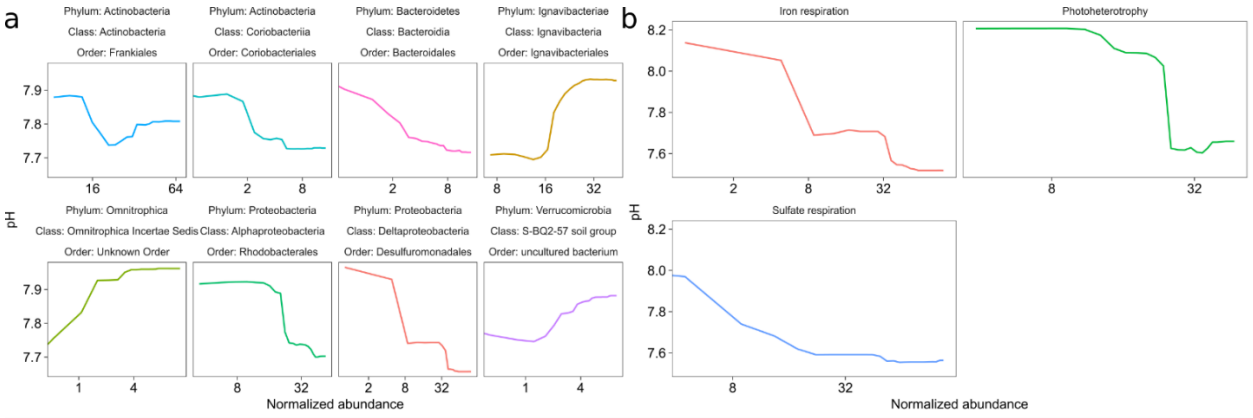

Figure A16: Partial dependence of random forest model prediction of pH from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

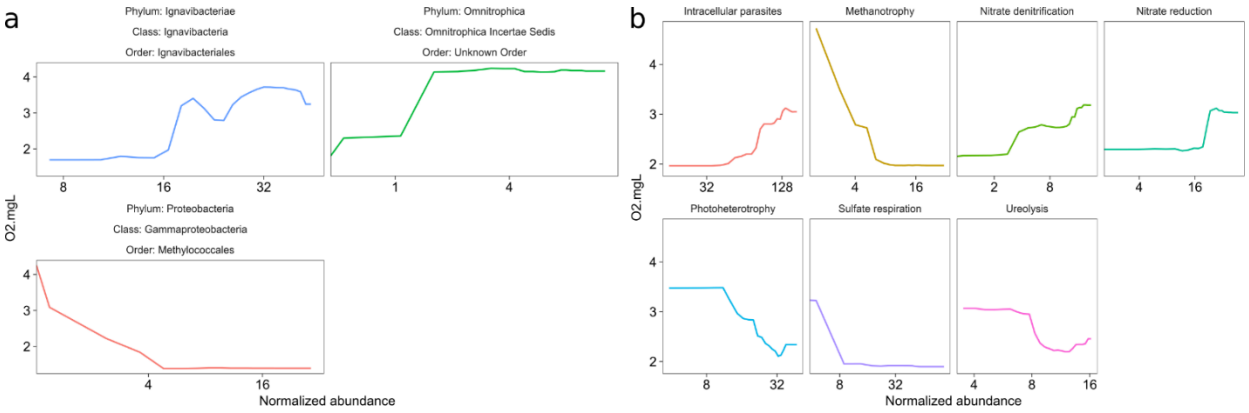

Figure A17: Partial dependence of random forest model prediction of [O_2] from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

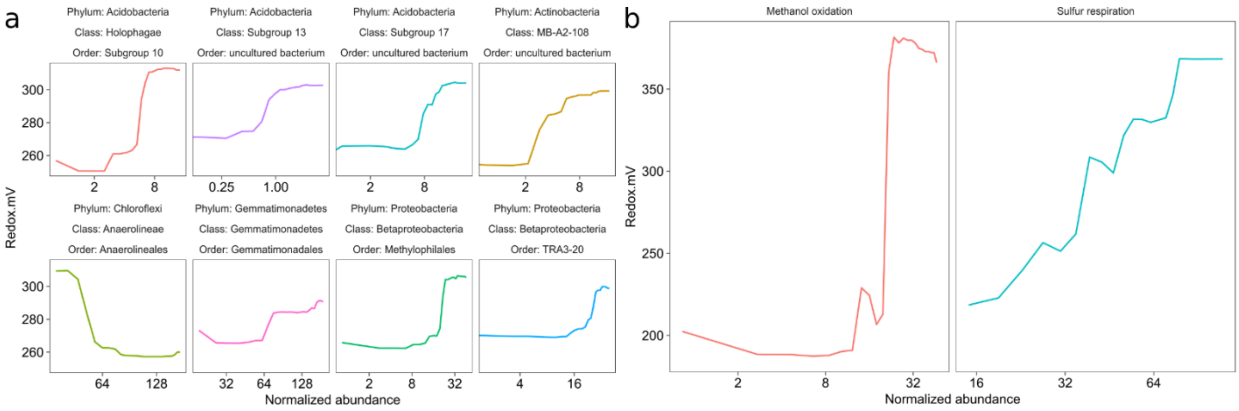

Figure A18: Partial dependence of random forest model prediction of redox potential from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

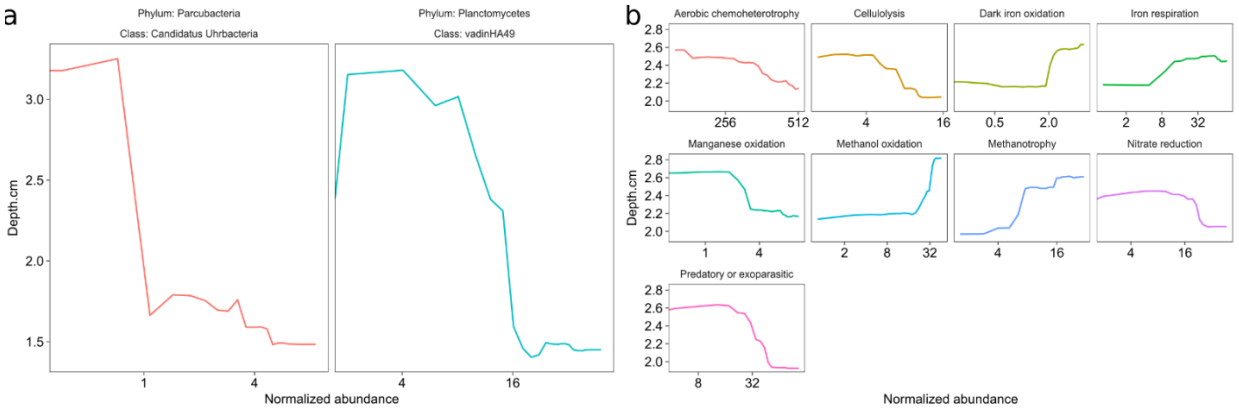

Figure A19: Partial dependence of random forest model prediction of sediment depth from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

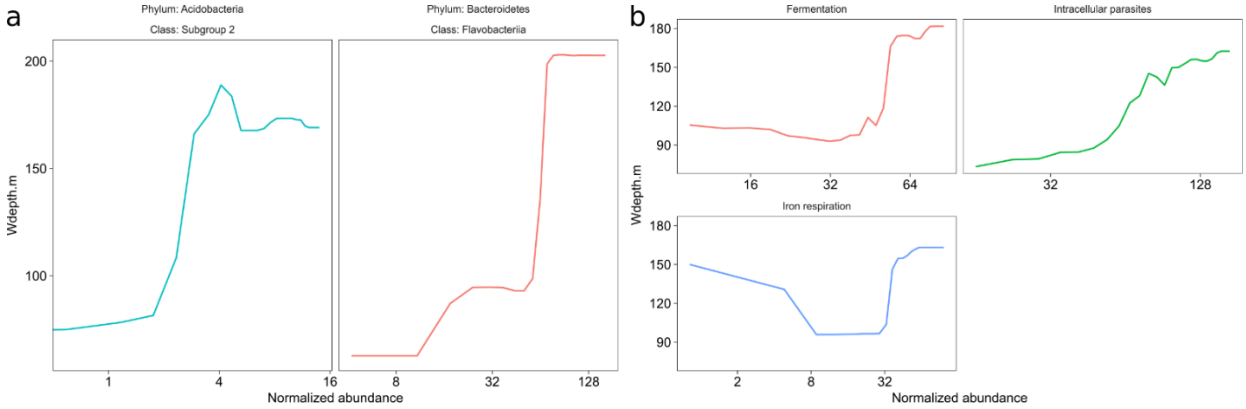

Figure A20: Partial dependence of random forest model prediction of water depth from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

A.9.2 Continuous variables: summer 2015

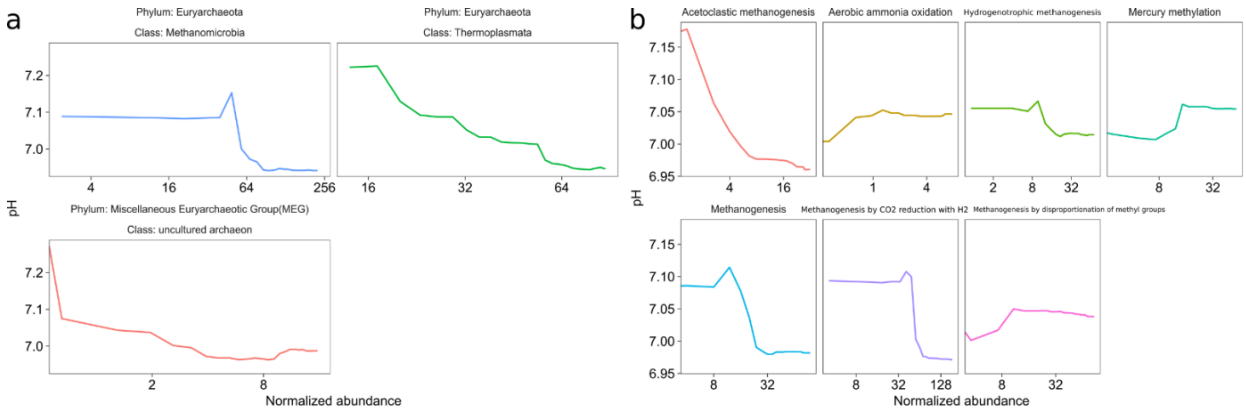

Figure A21: Partial dependence of random forest model prediction of pH from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

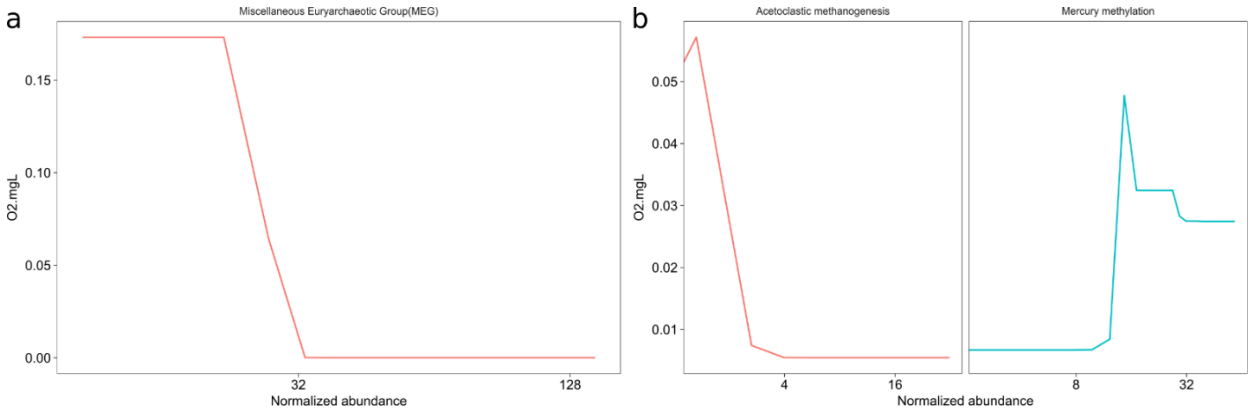

Figure A22: Partial dependence of random forest model prediction of [O_2] from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

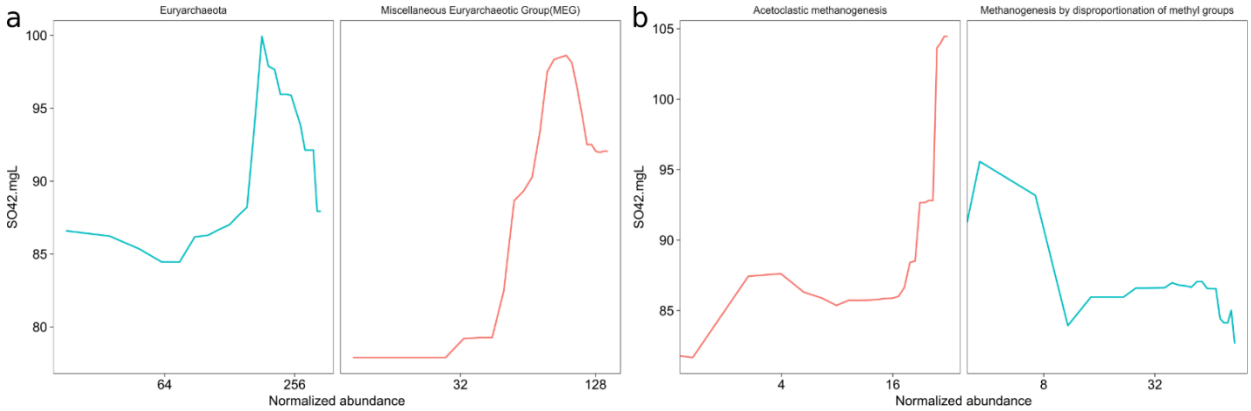

Figure A23: Partial dependence of random forest model prediction of [SO_4^{2-}] from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

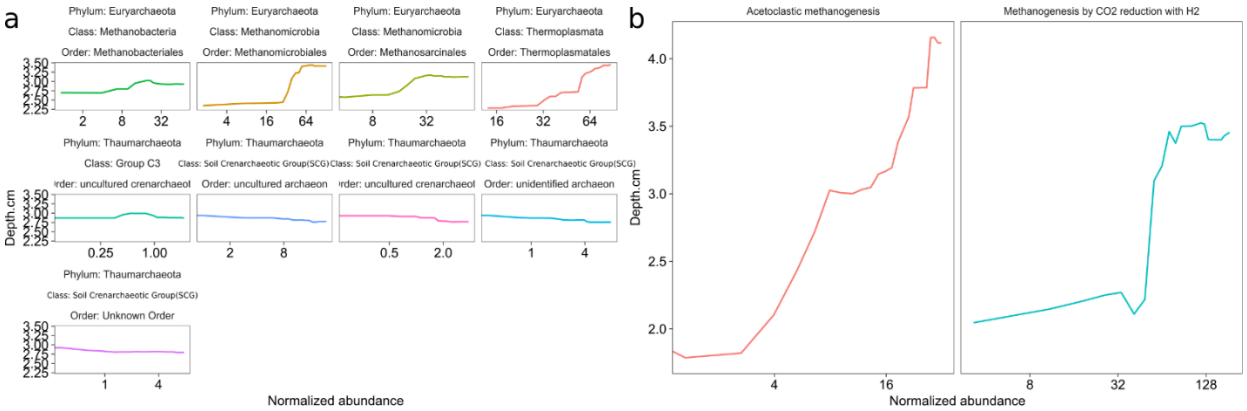

Figure A24: Partial dependence of random forest model prediction of sediment depth from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

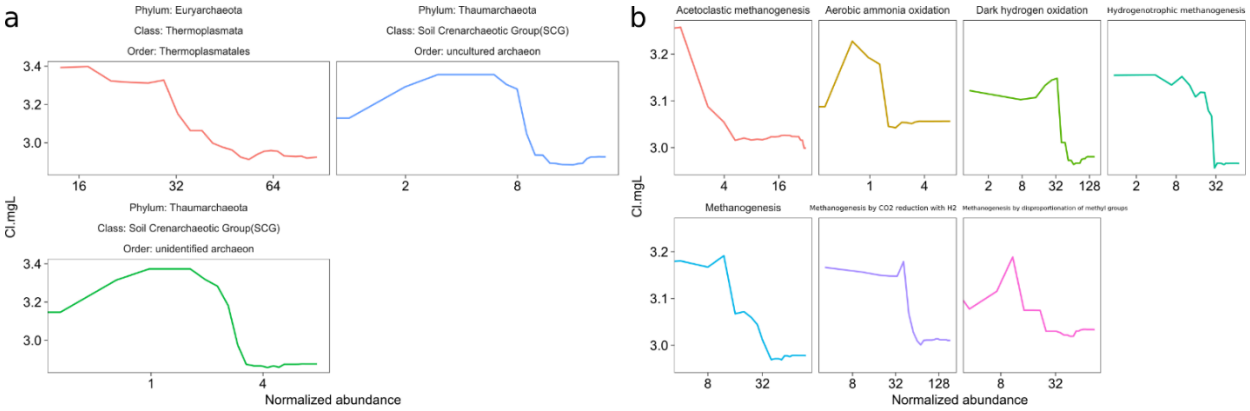

Figure A25: Partial dependence of random forest model prediction of [Cl^-] from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

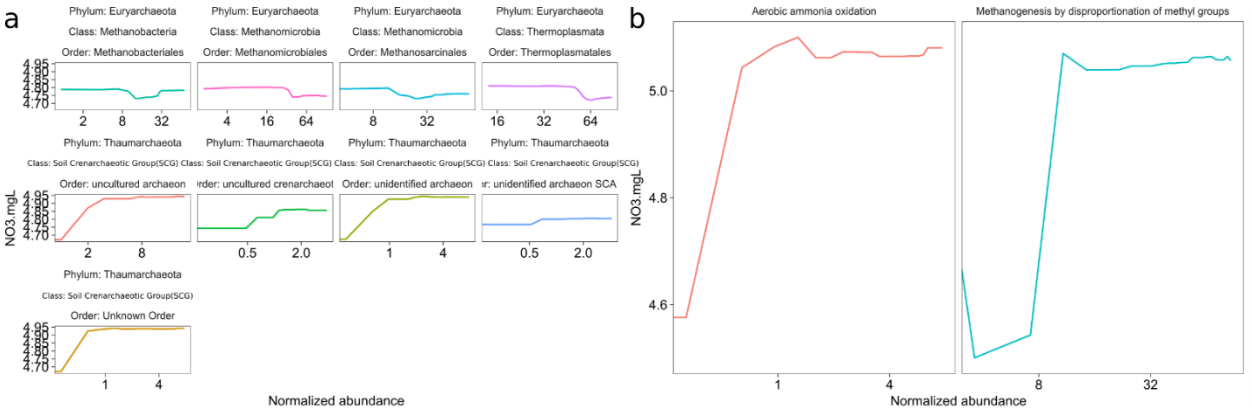

Figure A26: Partial dependence of random forest model prediction of [NO_3^-] from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

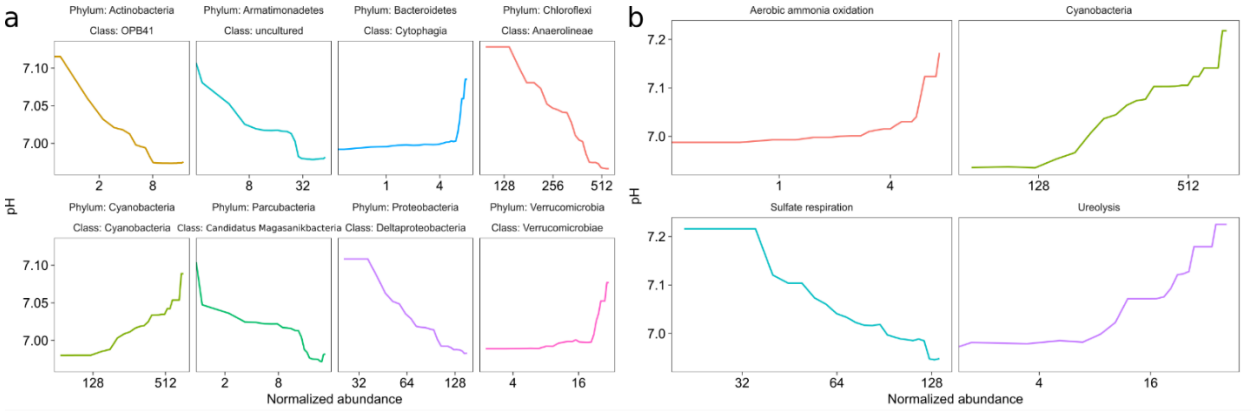

Figure A27: Partial dependence of random forest model prediction of pH from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

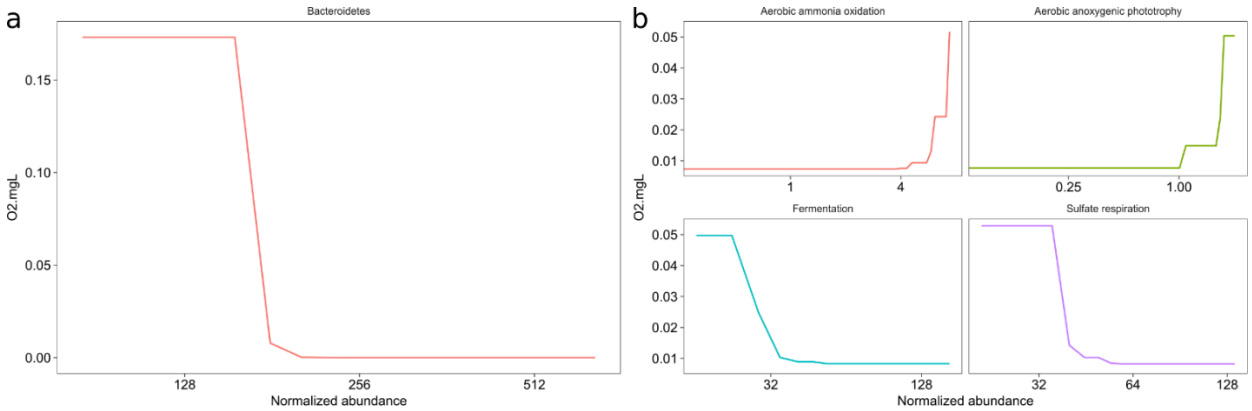

Figure A28: Partial dependence of random forest model prediction of [O_2] from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

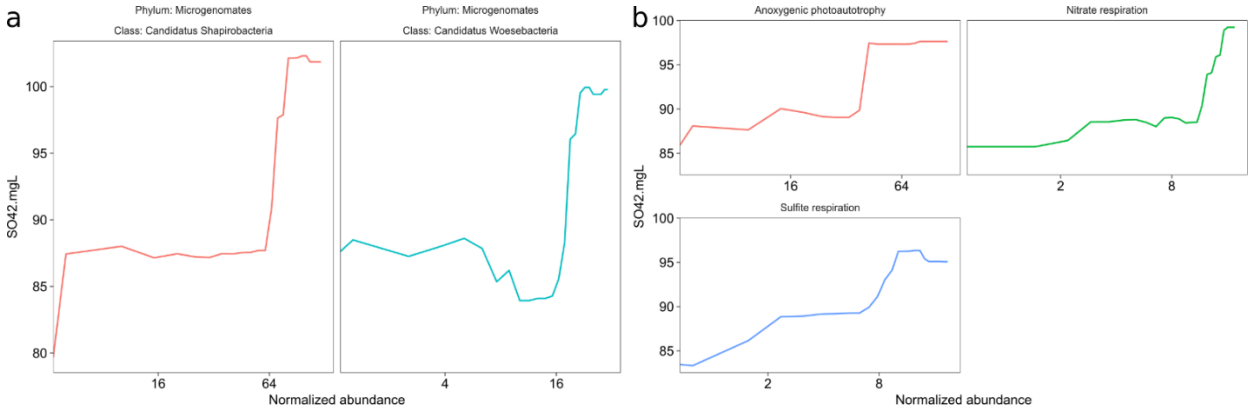

Figure A29: Partial dependence of random forest model prediction of [SO_4^{2-}] from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

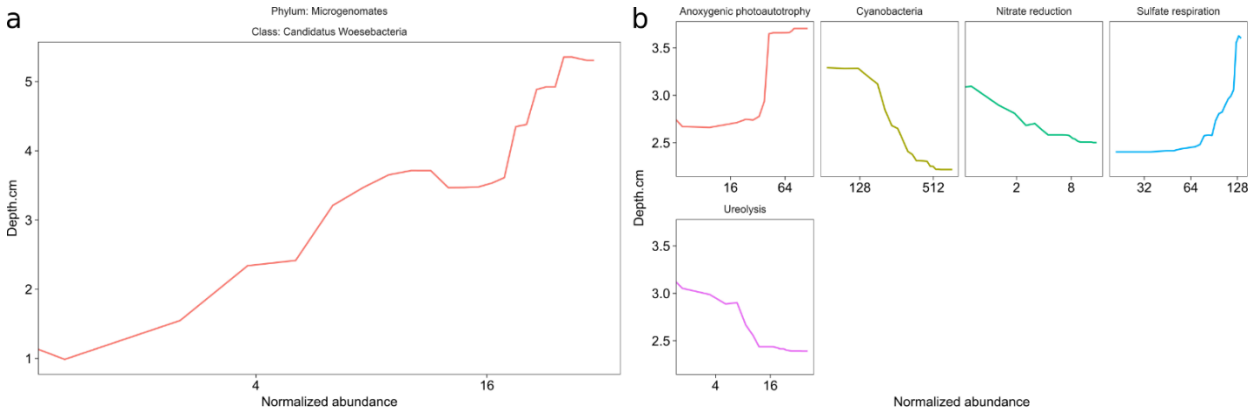

Figure A30: Partial dependence of random forest model prediction of sediment depth from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

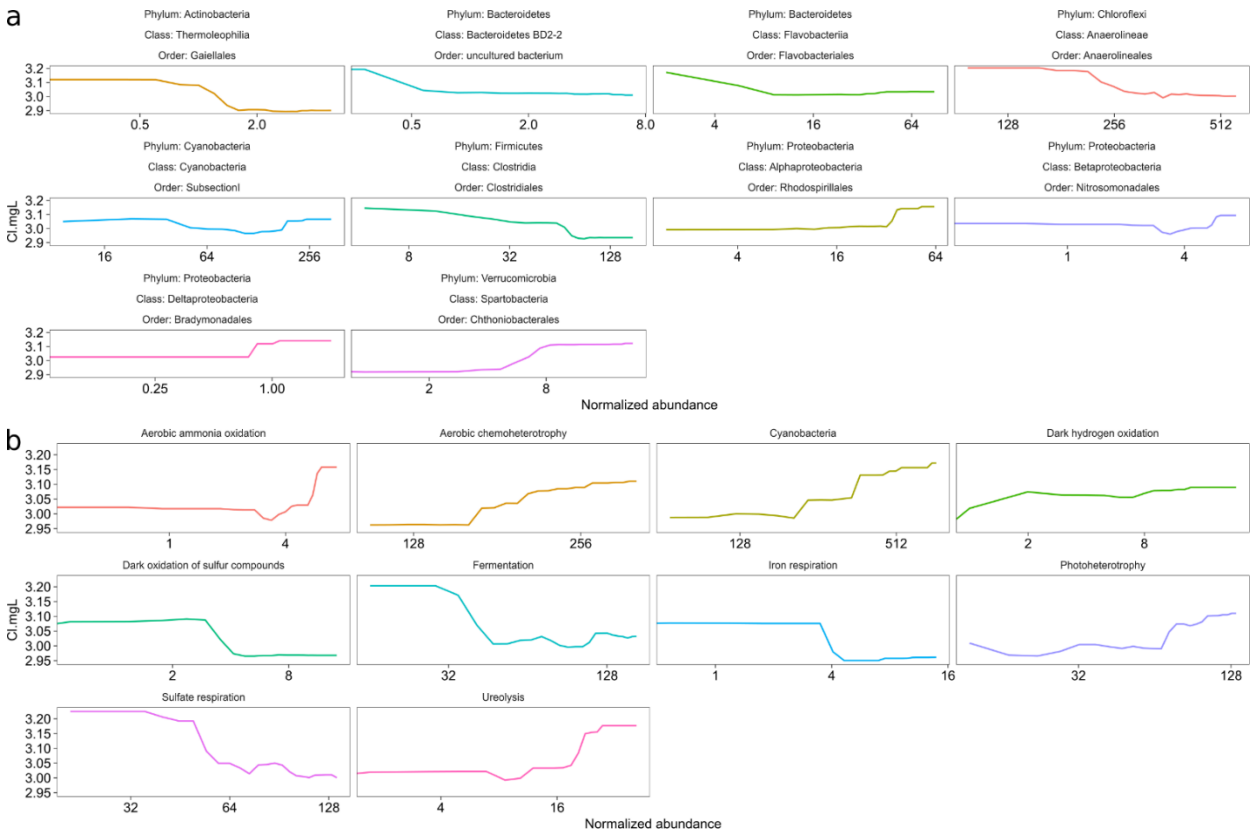

Figure A31: Partial dependence of random forest model prediction of [Cl^-] from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

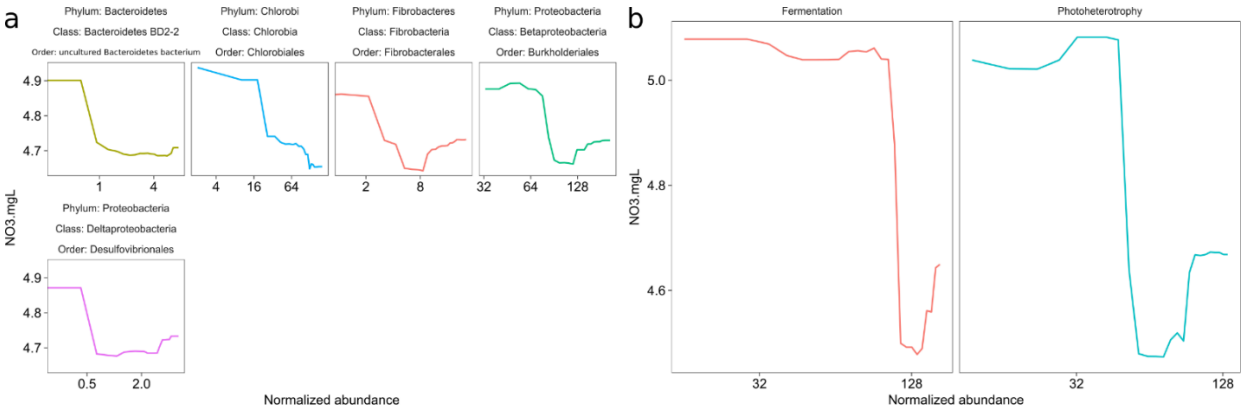

Figure A32: Partial dependence of random forest model prediction of [NO_3^-] from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

A.9.3 Categorical variables: spring 2014/2015

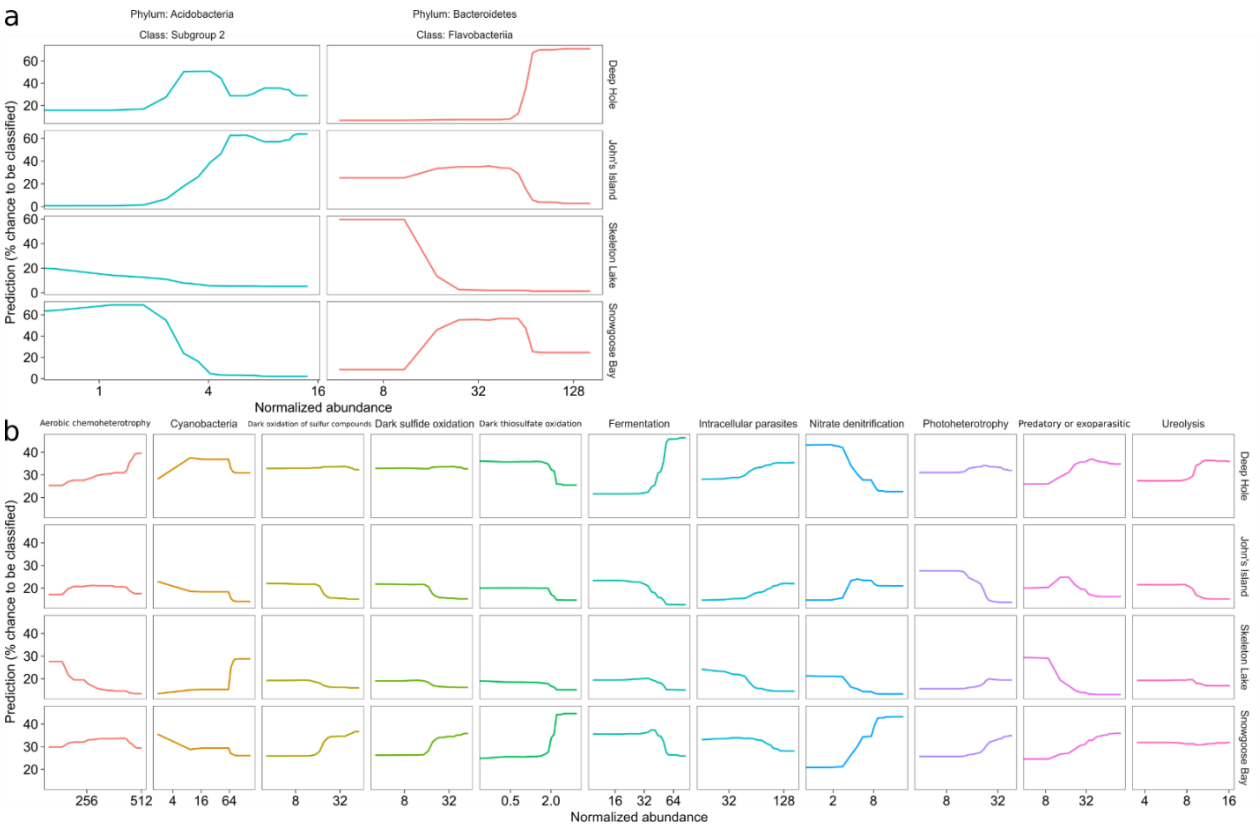

Figure A33: Partial dependence of random forest model prediction of sampling site from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

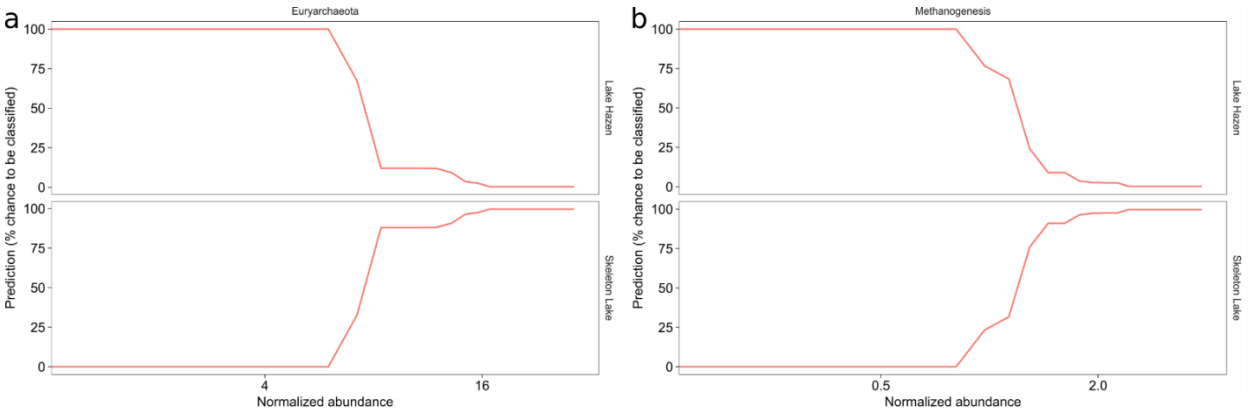

Figure A34: Partial dependence of random forest model prediction of sampled lake from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

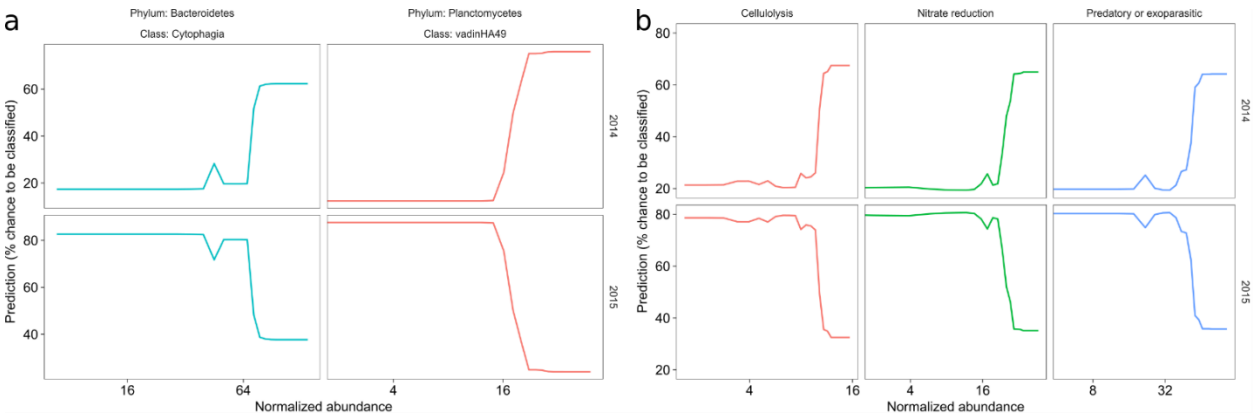

Figure A35: Partial dependence of random forest model prediction of sampling year from spring 2014/2015 data set with universal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

A.9.4 Categorical variables: summer 2015

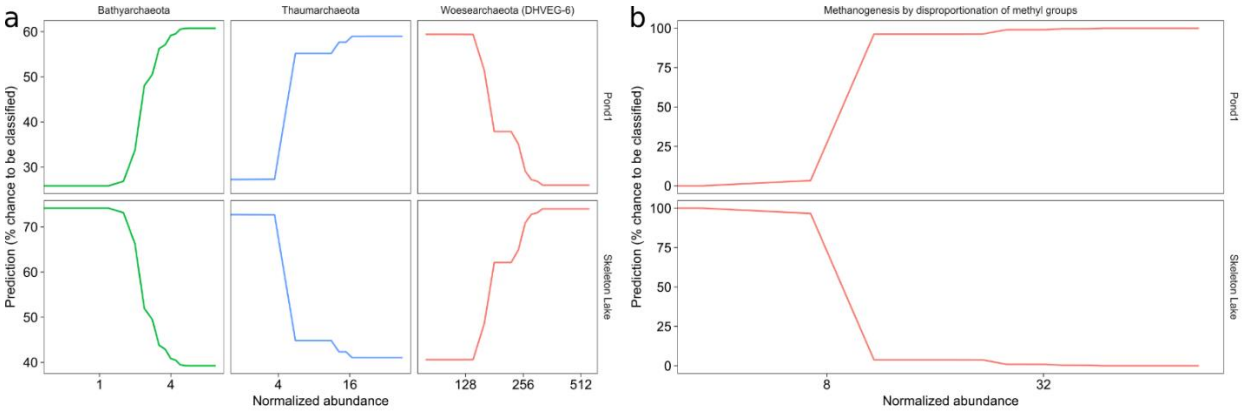

Figure A36: Partial dependence of random forest model prediction of sampling site from summer 2015 data set with archaeal primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

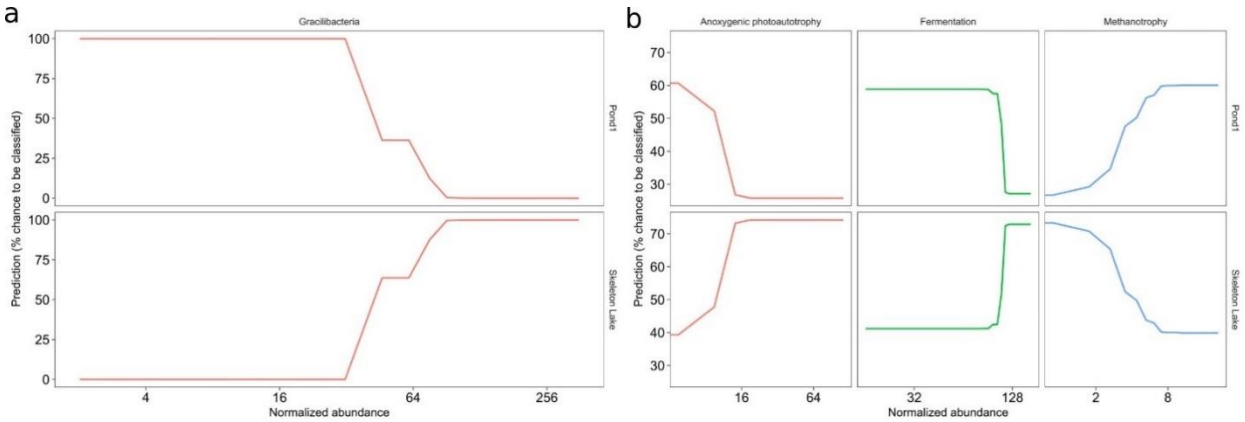

Figure A37: Partial dependence of random forest model prediction of sampling site from summer 2015 data set with bacterial primers. (**a**) Phylogenetic data. (**b**) Functionally mapped data.

Appendix B: Supporting information for Chapter 3

B.1 Sampling

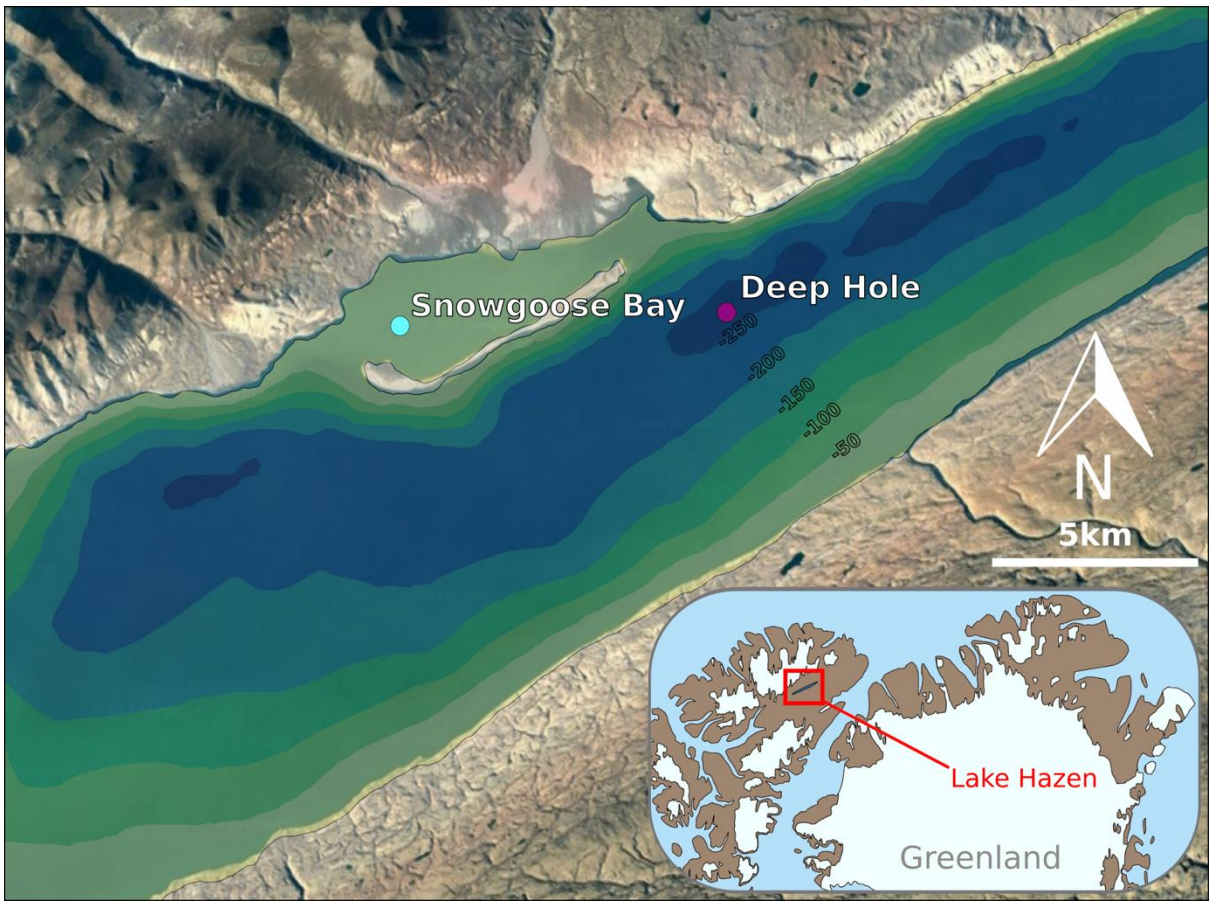

Figure B1: Sampling sites at Lake Hazen and location of the site on northern Ellesmere Island, Nunavut, Canada (inlay). Landsate-7 satellite image courtesy of the U.S. Geological Survey. Bathymetric data adapted from Köck et al. (2012).

B.2 DNA extraction and sequencing

PCR for the *glnA* gene to check for DNA quality was performed with the primers GS1β (5'-GAT GCC GCC GAT GTA GTA-3') and GS2γ (5'-AAG ACC GCG ACC TTY ATG CC-3'), which generate a 153 or 156 bp fragment of the gene (Hurt et al., 2001). Amplifications were performed in 25 µL reaction volumes, each containing 12.5 µL of EconoTaq PLUS GREEN 2X Master Mix (Lucigen Corporation, Middleton, WI, USA), 1 µL of each forward and reverse primer (25 µM), 1 µL of template (DNA extract) and 9.5 µL of H_2O. Reaction cycling consisted of initial disassociation at 94ºC for 2 min, followed by disassociation at 94ºC for 30s, annealing at 55ºC for 30s and elongation at 72ºC for 30s for 30 times, with a final elongation at 72ºC for 5 min. Amplification from all DNA extracts, but not the H_2O PCR controls, was confirmed by electrophoresis.

B.3 Assembly

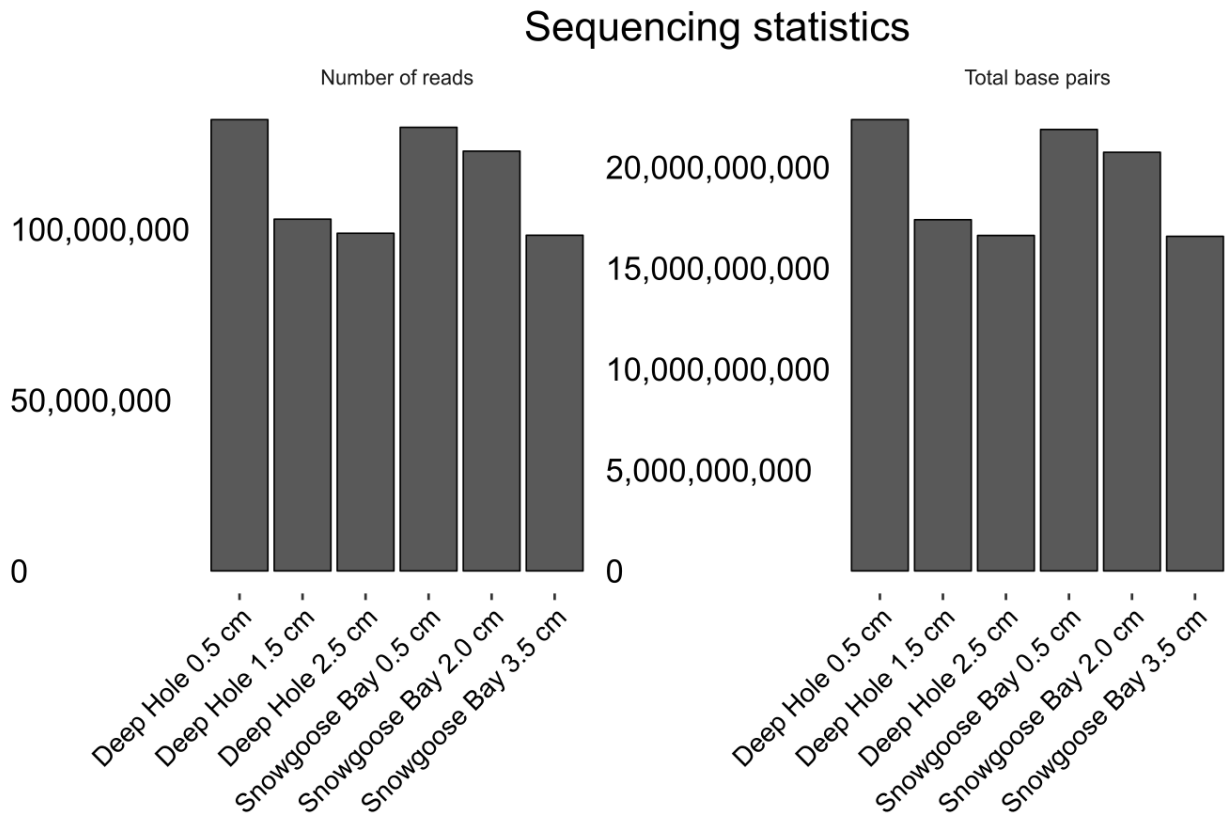

Figure B2: Raw number of reads and base pairs per sample.

B.4 Chemistry

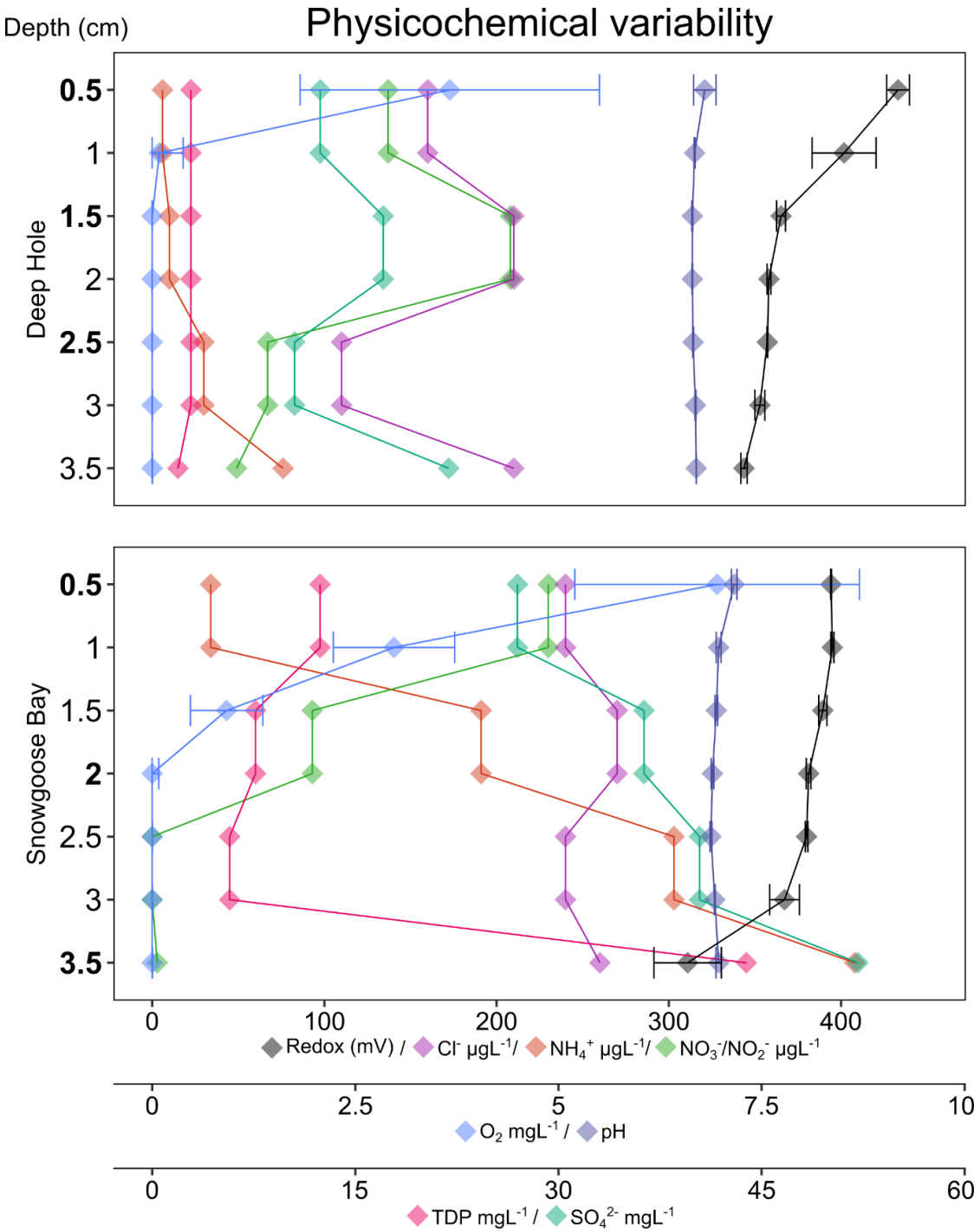

Figure B3: Physicochemical variability in the samples from Lake Hazen sediments. Specific scales for measured variables are indicated as three separate x-axes. Sediment depths where DNA was extracted from in each core are indicated with bold font.

B.5 Community structure

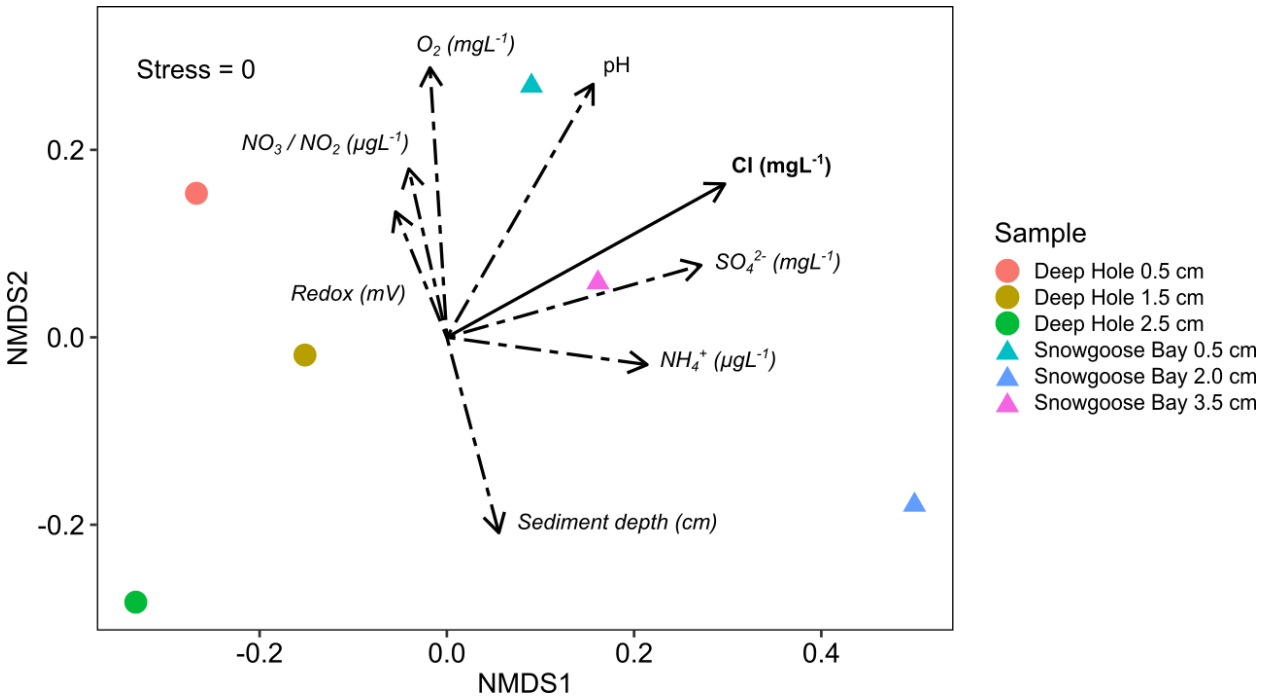

Figure B4: NMDS ordination of Bray-Curtis dissimilarity of the microbial communities (*n* = 166) of the six samples, based on assembly of 16S rRNA fragments. NMDS stress is indicated in the top left corner and vectors for physicochemical variables are overlaid. Continuous variables other than [Cl^-] (solid vector; $P = 0.03$) were not significantly linearly correlated with the ordination (dashed vectors; $P > 0.1$), and the samples could not be differentiated by site ($P = 0.1$).

B.6 Marker genes and pathways

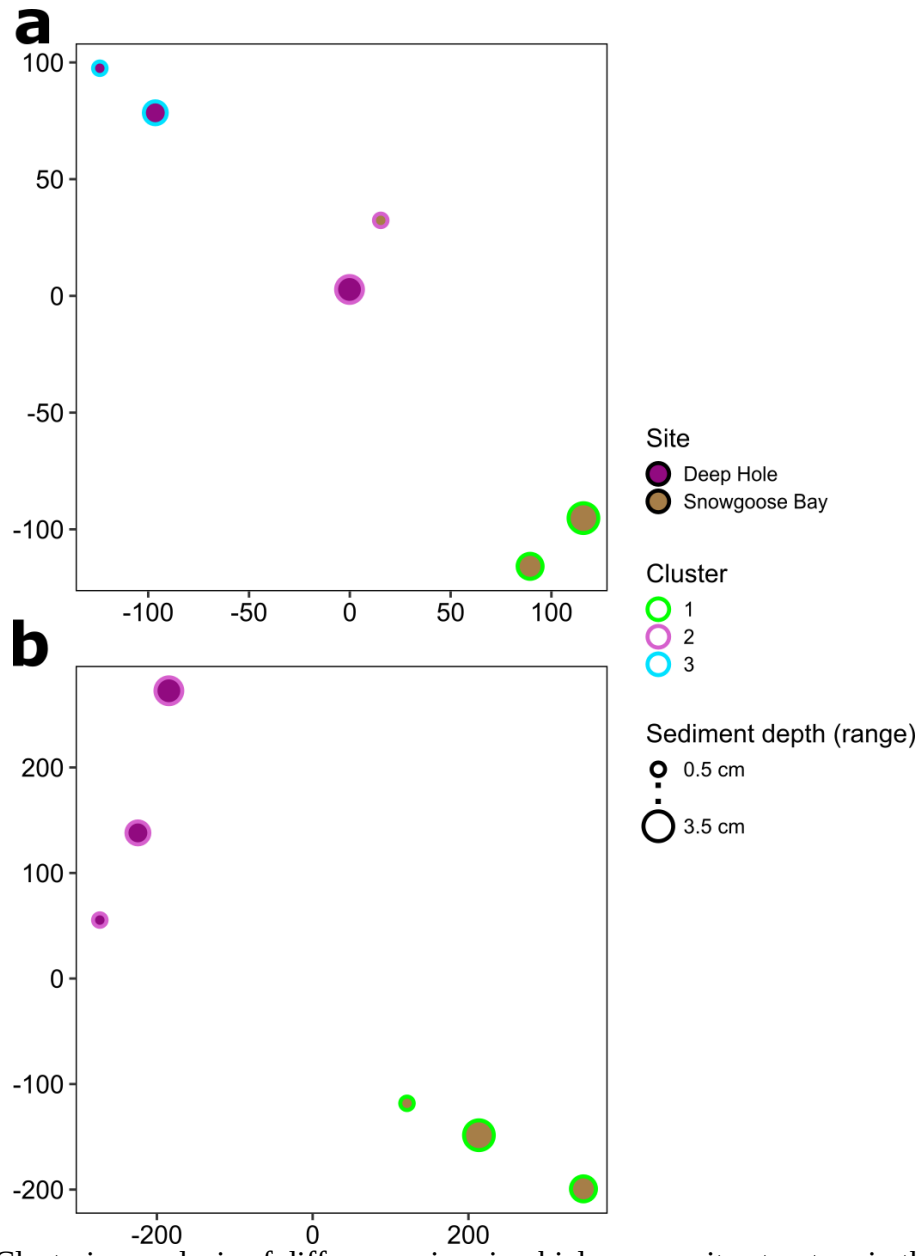

Figure B5: Clustering analysis of differences in microbial community structure in the sediment samples from Lake Hazen. (**a**) Medium quality MAGs (*n* = 55) with manual taxonomy assignments based on reference genomes in a phylogenetic tree constructed from ribosomal proteins. (**b**) Raw reads binned into 16S rRNA contigs (*n* = 166) with automatic taxonomy assignments from the SILVA 128 NR95 database.

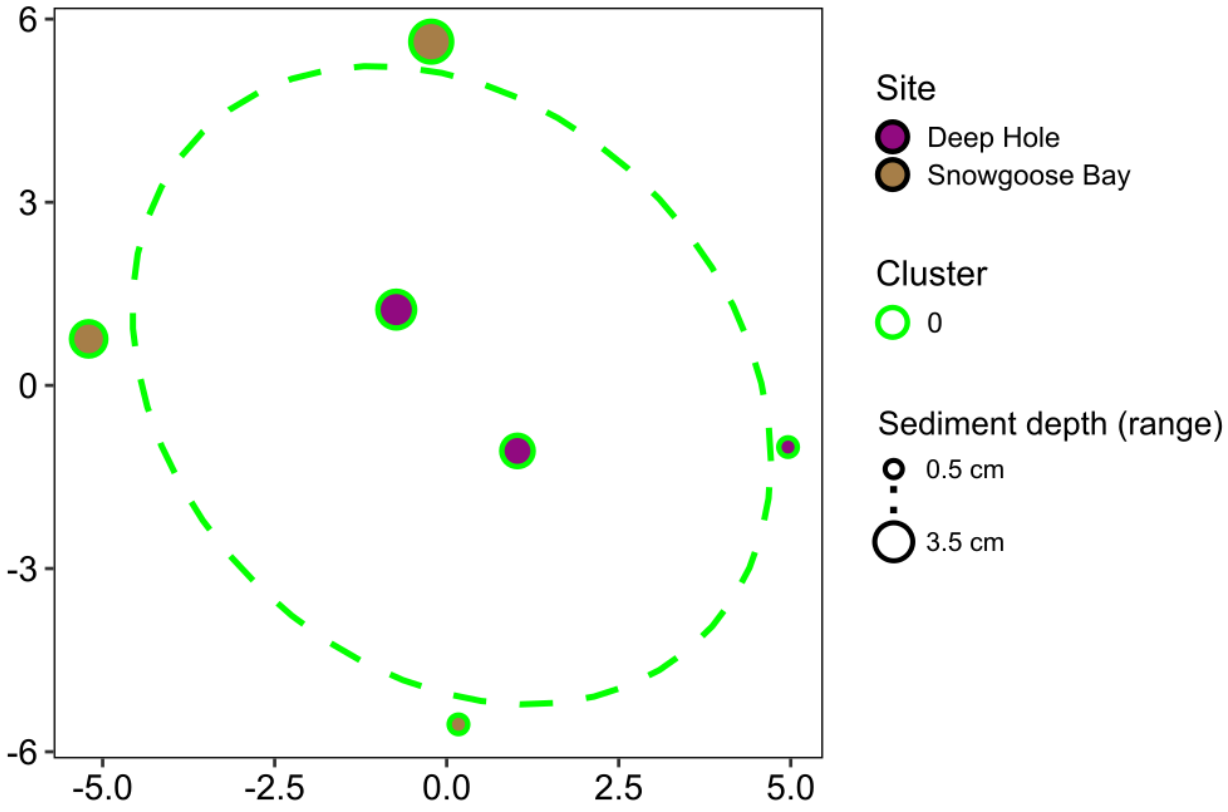

Figure B6: Clustering analysis of differences in MetaCyc functional pathway abundances in the 55 MAGs from sediment samples from Lake Hazen. All samples were assigned to 'Cluster 0' meaning that they are outliers, and no clusters were found in the analysis.

B.7 Recently discovered and unfamiliar bacteria are found in Lake Hazen sediments

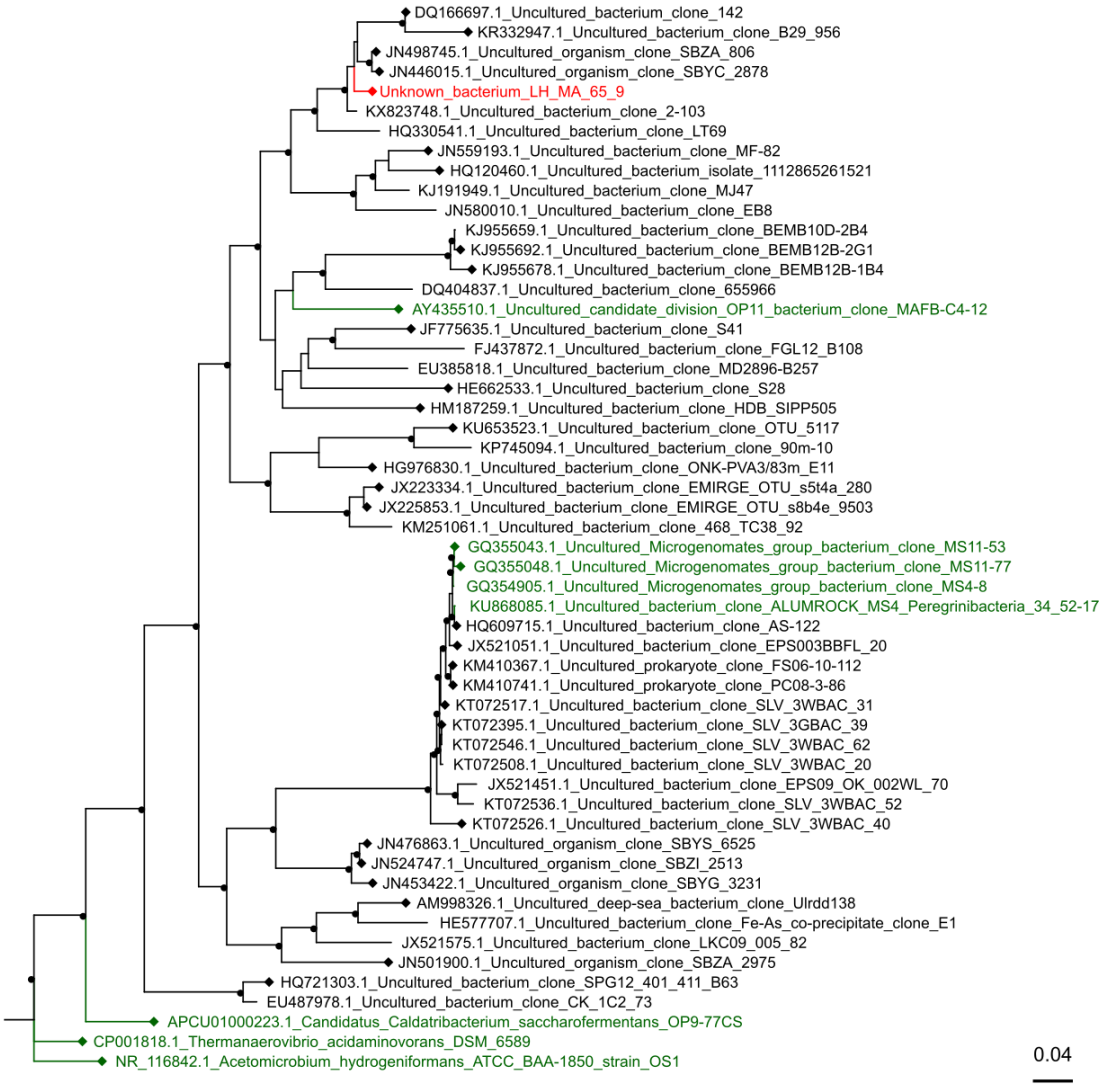

Figure B7: Phylogenetic tree of the taxonomically uncharacterized MAG, LH_MA_65_9, based on the alignment of its 16S rRNA gene. The MAG is shown in red and taxonomically classified 16S sequences in green. Circles on nodes indicate > 0.85 support in FastTree and diamonds on tips indicate higher than 90% nucleotide identity with the LH_MA_65_9 16S rRNA sequence over the aligned region. The tree was rooted with the 16S rRNA gene of *Acetomicrobium hydrogeniformans* (NCBI Reference Sequence: NR_116842.1).

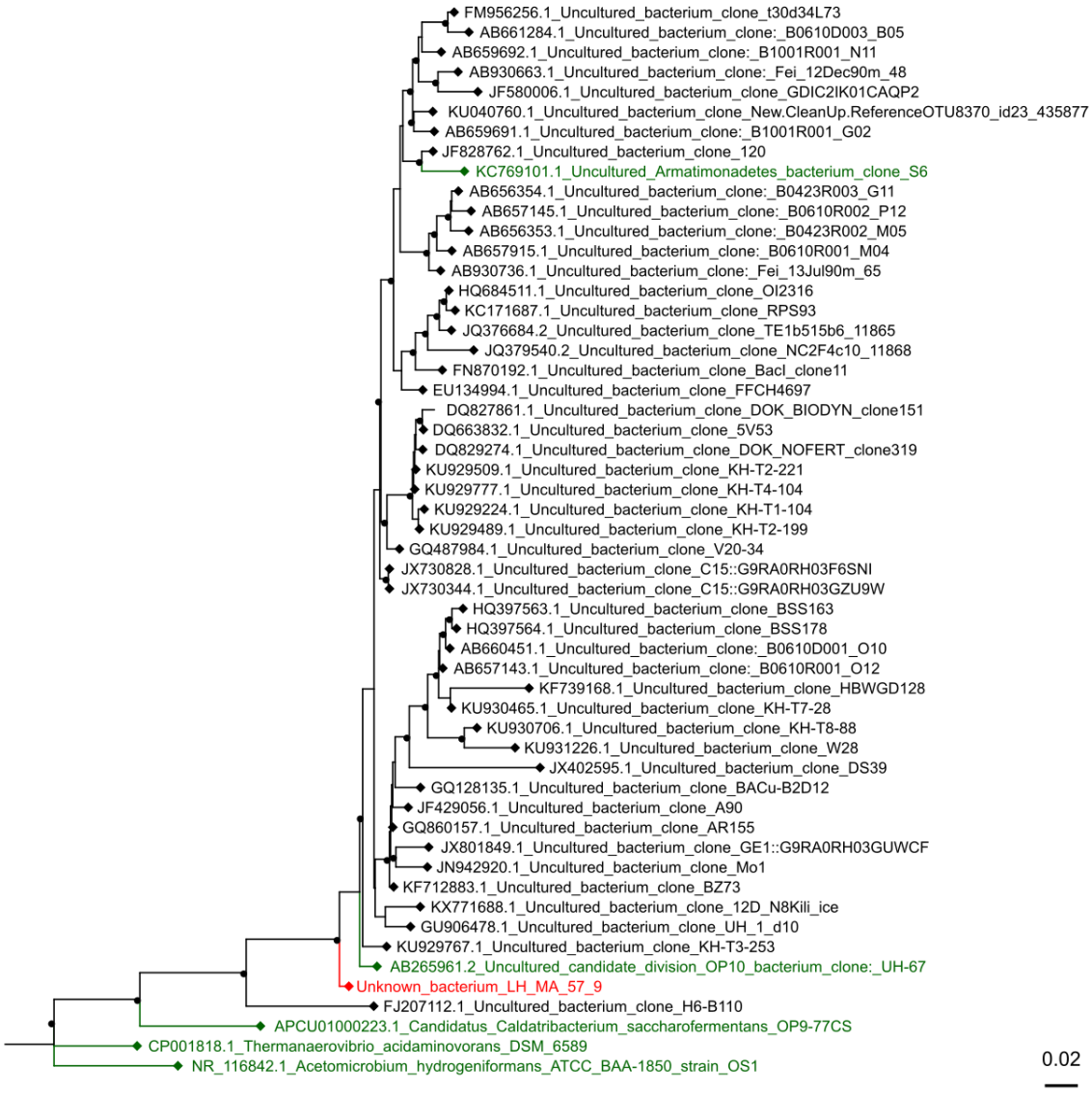

Figure B8: Phylogenetic tree of the taxonomically uncharacterized MAG, LH_MA_57_9, based on the alignment of its 16S rRNA gene. The MAG is shown in red and taxonomically classified 16S sequences in green. Circles on nodes indicate > 0.85 support in FastTree and diamonds on tips indicate higher than 90% nucleotide identity with the LH_MA_57_9 16S rRNA sequence over the aligned region. The tree was rooted with the 16S rRNA gene of *Acetomicrobium hydrogeniformans* (NCBI Reference Sequence: NR_116842.1).

B.8 Nitrogen and sulfur cycles in the sediments

A Spirochaetes MAG was the most common nitrogen (N_2) fixer, together with three less abundant *Geobacter* MAGs (**Figure 3.5a**). Three Planctomycetes MAGs were the most common that harbored *nirAB* genes involved in Dissimilatory Nitrite Reduction to Ammonia (DNRA, NO_2^- to NH_4^+). Planctomycetes MAGs were also the most common that had ammonia (NH_4^+) assimilation genes. Together, these three processes channel inorganic nitrogen into organic matter. Conversely, nitrogen is lost from aquatic systems by nitrification followed by denitrification, which releases it as NO, N_2O or N_2 (**Figure 3.5**). Betaproteobacteria, specifically two Nitrosomonadales MAGs, were the only one which had the *haoB* gene involved in nitrification (NH_4^+ oxidation to NO_2^-), and the most abundant MAG harboring the *nosZ* marker gene for N_2O reduction (to N_2), respectively. A *Rhodoferax* MAG (from Betaproteobacteria) was the most abundant nitrate (NO_3^-) reducer, harboring *nar*-genes. The ratio of DNRA to denitrification is important for the fate of nitrogen in the sediment as the former cycles nitrite back to ammonia and the latter reduces it to a volatile form. In Lake Hazen, 15 MAGs (27%) had genes involved in denitrification and 8 (15%) in DNRA. The mean relative abundances of the MAGs that had denitrification genes were also higher. However, the activity of these organisms in the two processes cannot be deduced purely from the analysis of metagenomic data, and marker genes for both processes were sometimes found in the same MAGs. A Myxococcales MAG (from Deltaproteobacteria) was the most abundant MAG with the *nrfH*-gene (involved in DNRA), but the organism was perhaps also capable of (*nirK*-mediated) nitrite reduction to the volatile nitrous oxide (N_2O), a key denitrification process. The relative contribution of these processes both in the community and in the same organism varies depending on the environment (van den Berg et al., 2017b) and can even be influenced by bacteria not directly involved in them (van den Berg et al., 2017a). It is not certain if denitrification (leading to losses of N) is more common

in the Lake Hazen sediments. The lake appears to be a net sink of NO_3^-, NO_2^- and NH_4^+ (St. Pierre et al., 2019b), but because the top 1 to 2 cm of the sediment are oxidized (**Figure B3**), denitrification is unlikely to take place there. On the other hand, because of the high sediment deposition in Lake Hazen (St. Pierre et al., 2019b), the nitrogen species that are deposited together with the particulate matter might get buried quite rapidly into the anoxic layers, where denitrification by these organisms could take place. However, further study would be required to assess the fate of nitrogen and its seasonal trends in Lake Hazen.

In the sulfur cycle, sulfate (SO_4^{2-}) in the sediment can either be reduced anaerobically by dissimilatory sulfate reducers to sulfite (SO_3^{2-}) or aerobically by assimilation into organic sulfur (**Figure 3.5b**). Marker genes for the dissimilatory pathway (*aprAB*) were detected in the > 1kb contigs, and the assimilatory pathway was detected in 10 of the 55 MAGs, with Planctomycetes as the most abundant phylum. The assimilatory pathway is likely utilized in the aerobic top sediments to synthesize cysteine and methionine, while the strictly anaerobic dissimilatory pathway is likely an electron sink deeper in the sediment. Sulfite can then be produced in the sediment directly by the dissimilatory pathway or desulfonated from organic sulfonates (Cook et al., 1998). Here, Alphaproteobacteria was the most abundant phylum with the desulfonation pathway. The 'NrfD-type' family (Pfam 03916) that we annotated as 'DNRA Polysulfide reductase' matches enzymes associated with both nitrite reduction to ammonia (Simon, 2002) and sulfur reduction (Jormakka et al., 2008). In **Figure 3.5**, we annotated polysulfide reduction with this gene rather than DNRA since we had more specific models for DNRA marker genes (namely *nrfH*). The marker gene for this more generic sulfur reduction (including tetrathionate-, DMSO-, and polysulfide-reductases) was found in several MAGs, of which Verrucomicrobia was the most abundant, followed by Deltaproteobacteria and Betaproteobacteria. The more specific *cysI*-mediated assimilatory sulfite reduction

marker was only found in two Verrucomicrobia bacteria (also harboring the other sulfur reduction marker gene).

Only few MAGs had markers for multiple sulfur reduction pathways, and none could probably catalyze the complete reduction of sulfate to sulfide. Furthermore, we could not find the markers for *dsrAB* or *aprAB* genes from the MAGs while they were found, respectively, from 61 and 54 repository genomes using the same pipeline. They were also found in the low-quality bins and unbinned contig data (**Figure 3.4**). The *dsrAB* genes encode proteins that catalyze the reduction of sulfite to sulfide, or act in reverse in the oxidation of sulfide (Müller et al., 2015). The *aprAB* genes similarly encode for proteins catalyzing the reversible reduction of sulfate to sulfite (Rabus et al., 2006). The absence of these pathways in the 55 MAGs shows that these pathways are likely rare, but definitely not absent in the Lake Hazen sediments. Finally, sulfur is also oxidized in the sediment; a single MAG from Nitrosomonadales contained the *soxAXYZ*-genes.

B.9 Nutrient cycling capabilities of individual MAGs

Certain MAGs appeared to have highly versatile metabolisms. The only MAG with marker genes for sulfur oxidation (*soxAXYZ*) from the order Nitrosomonadales (LH_MA_55_1) likely represents an organism capable of assimilatory sulfite reduction to sulfide (as it has a *cysI* gene) and is also important in nitrogen cycling (**Table D6**): the MAG contained both a respiratory nitrate reductase (*narI*) and a nitrite reductase (*nirBD*), which produce ammonium (NH_4^+) as the product. The MAG also contained a nitrous oxide reductase (*nosZ*), which could be used for respiration. Furthermore, the MAG of this likely autotrophic organism contained a wide array of biosynthetic pathways, *e.g.*, a pentose phosphate pathway, and a reverse TCA cycle likely utilized for carbon fixation. A *Geobacter* (LH_MA_37_3) MAG included marker genes for N fixation (*nifS, nifV*), ammonia assimilation (*gltB*), nitrate reduction to nitrite (*narGHIJ*), DNRA / Polysulfide reduction (*nrfD*), and likely has wide

biosynthetic capabilities. Similarly, two Chloroflexi (LH_MA_58_17 and LH_MA_61_3) MAGs and a Spirochaete (LH_MA_58_12) MAG had several marker genes for both nitrogen and sulfur cycle processes. The MAG LH_MA_65_9, which was unclassified at the phylum-level, can likely both fix nitrogen (*nifS*) and utilize nitrite through DNRA (*nrfH*) and ammonia assimilation (*amtB*, *gltB*). None of the CPR MAGs had any of the nutrient cycling marker genes or MetaCyc pathways.

Appendix C: Supporting information for Chapter 4

C.1 Supplementary figures and tables

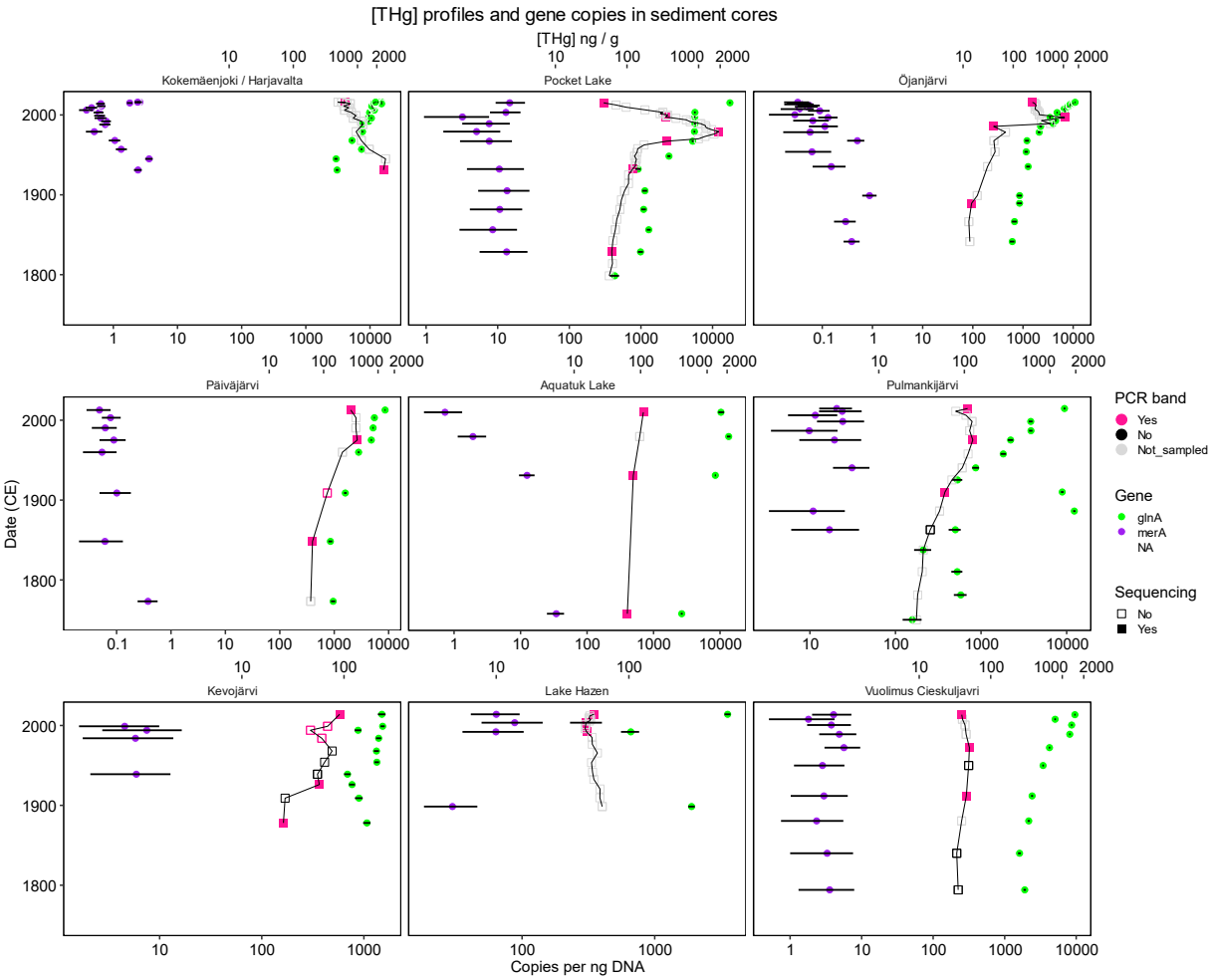

Figure C1: Dated gene quantification with droplet digital PCR and amplicon sequencing. Total Hg profiles and gene copy numbers in of merA and glnA the sediment cores. The samples that were sequenced are shown as solid squares. CE dates were ^{210}Pb-derived, except for Kevojärvi (varve counting) and Pocket Lake and Lake Hazen (cross-comparisons to previously ^{210}Pb-dated cores).

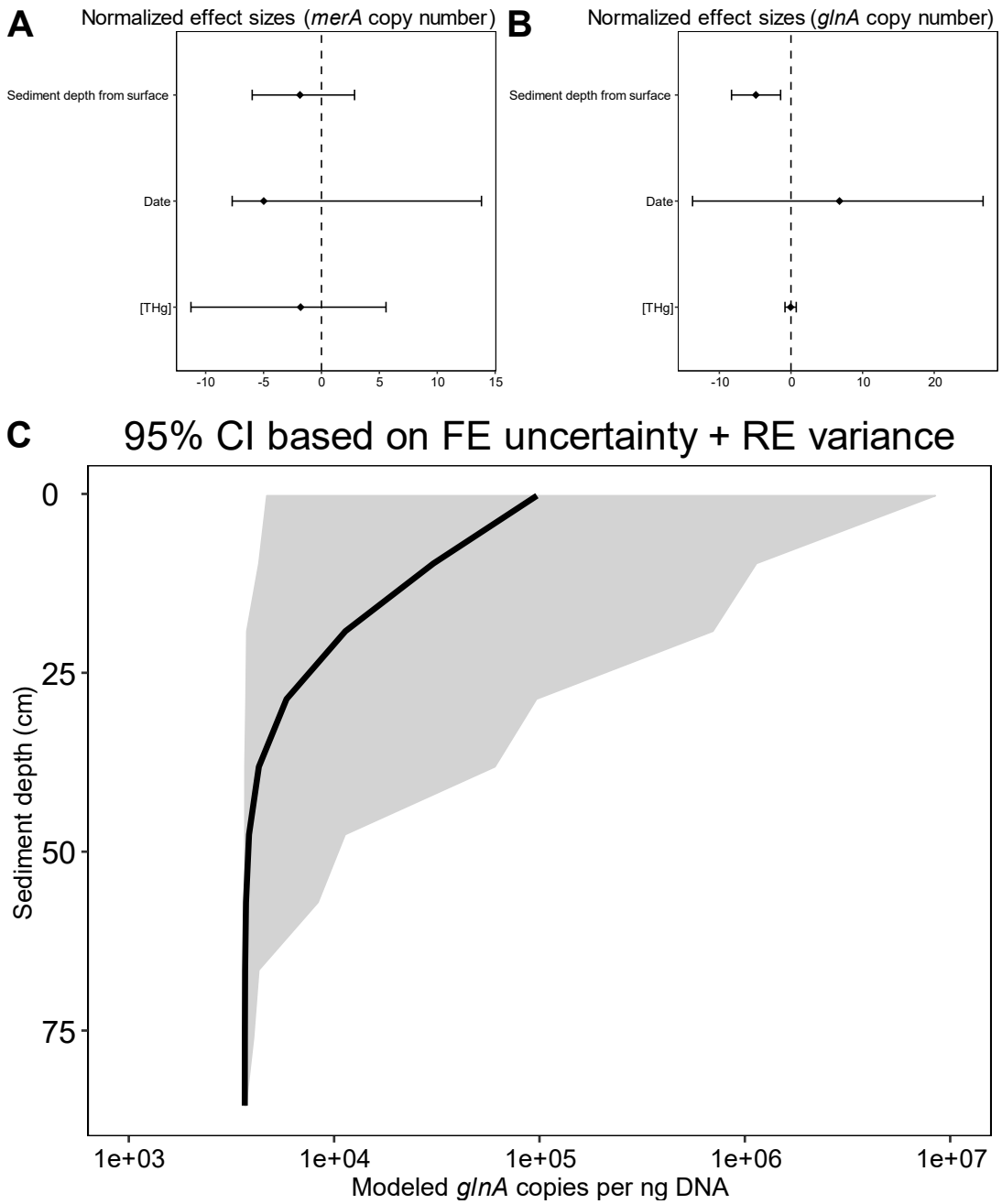

Figure C2: Drivers of gene abundances – droplet digital PCR analyses. Results of the mixed linear models showing the 95% confidence intervals of model terms for (**a**) merA and (**b**) glnA copy numbers on scaled and mean-centered variables, and (**c**) estimated effect size of sediment depth (distance from sediment surface) on the copy numbers of *glnA*. FE = Fixed Effect, RE = Random Effect.

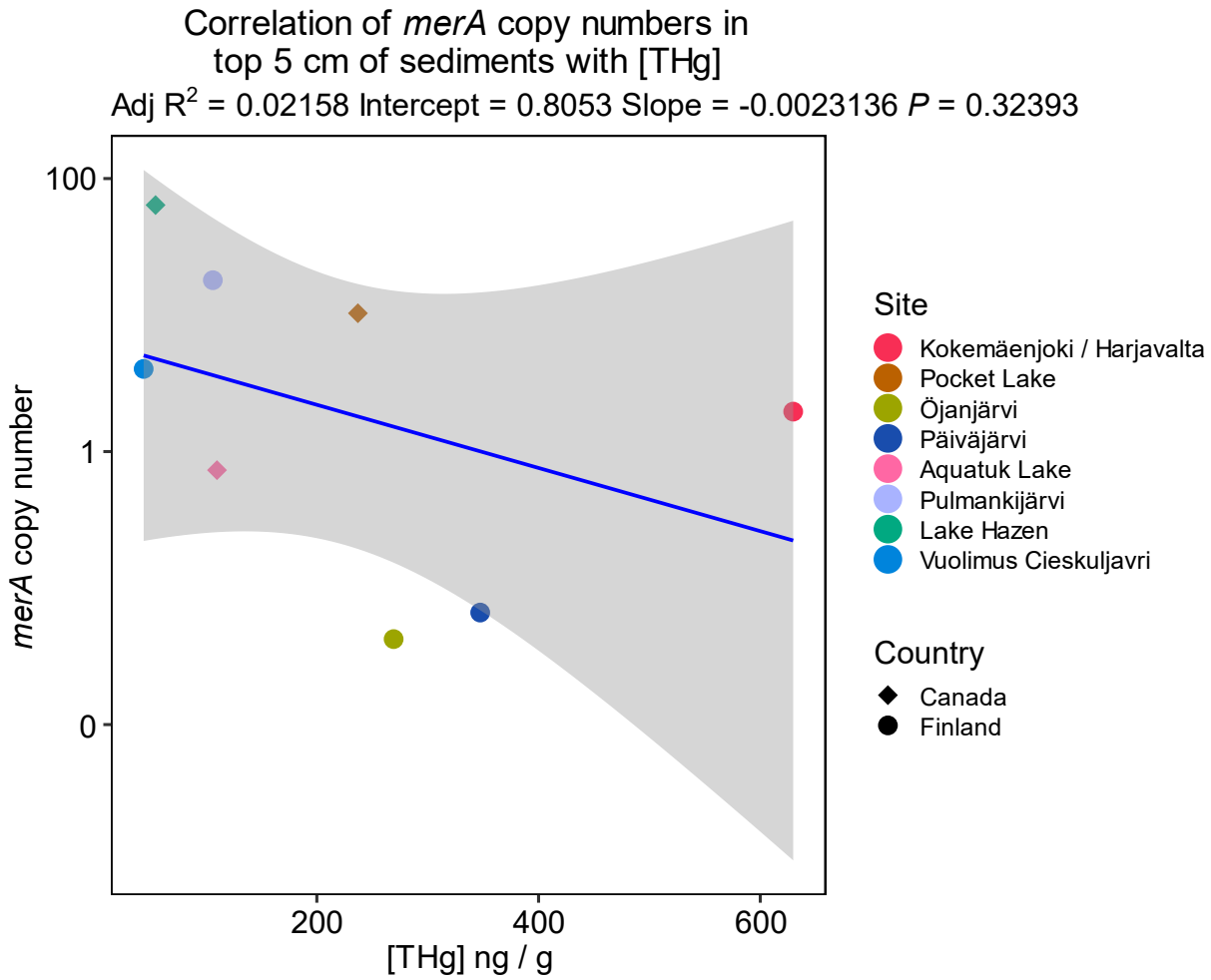

Figure C3: Gene abundance as a function of total mercury concentrations. Correlation between mean *merA* copy number and mean [THg] in the topmost 5 cm of sediment at each site. Least-square model fit is shown in blue, with its 95% confidence interval shown as the shaded region.

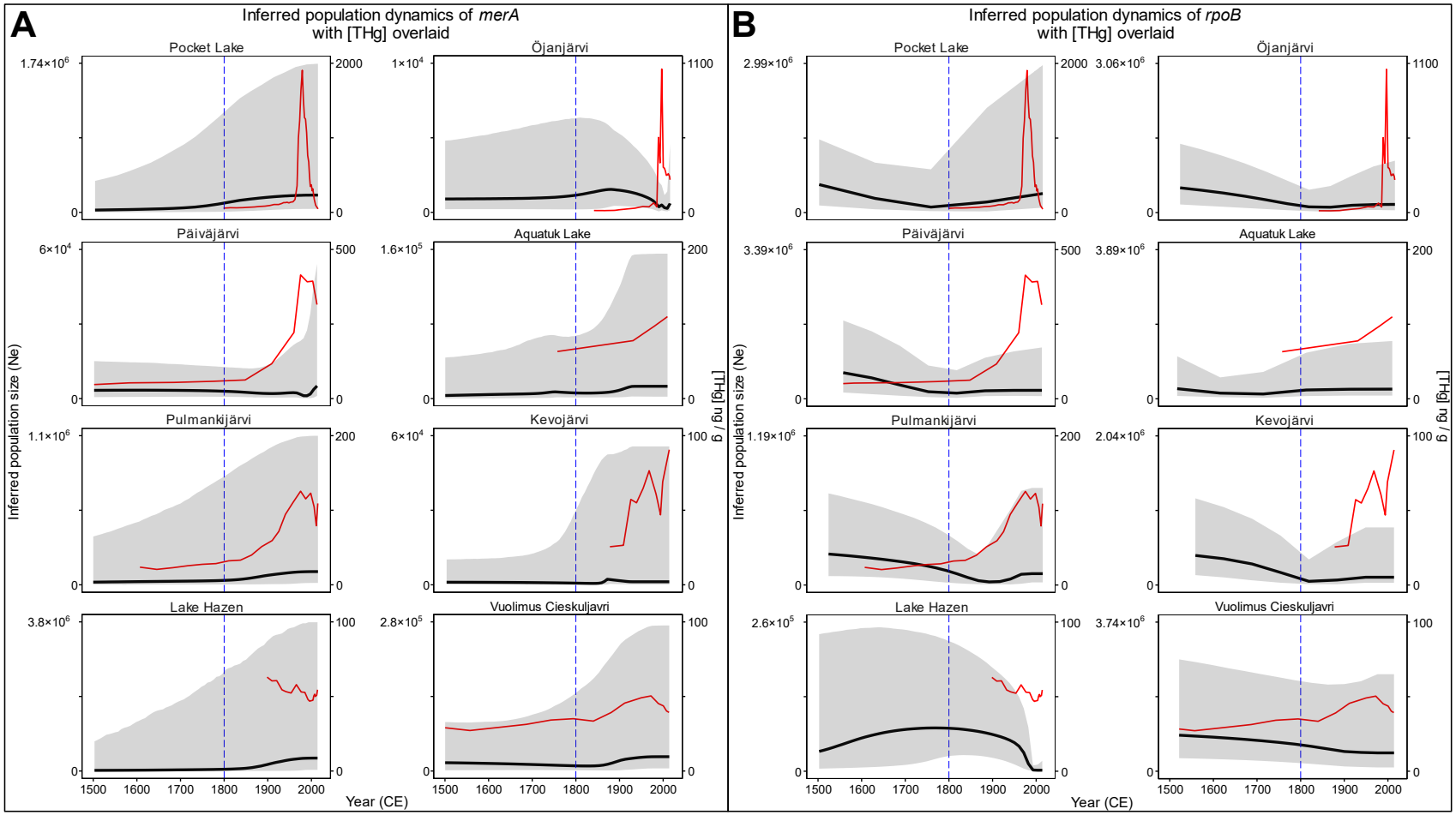

Figure C4: Demographic reconstructions. These are shown for *merA* (**a**) and *rpoB* (**b**) over calendar date with [THg] overlaid. In each panel, the black thick line shows the median of inferred effective population size and the grey area shows the 95% credible interval, the red line is the measured [THg], and the vertical blue dashed line indicates year 1800 CE, the approximate onset of the Industrial Revolution in the Northern Hemisphere.

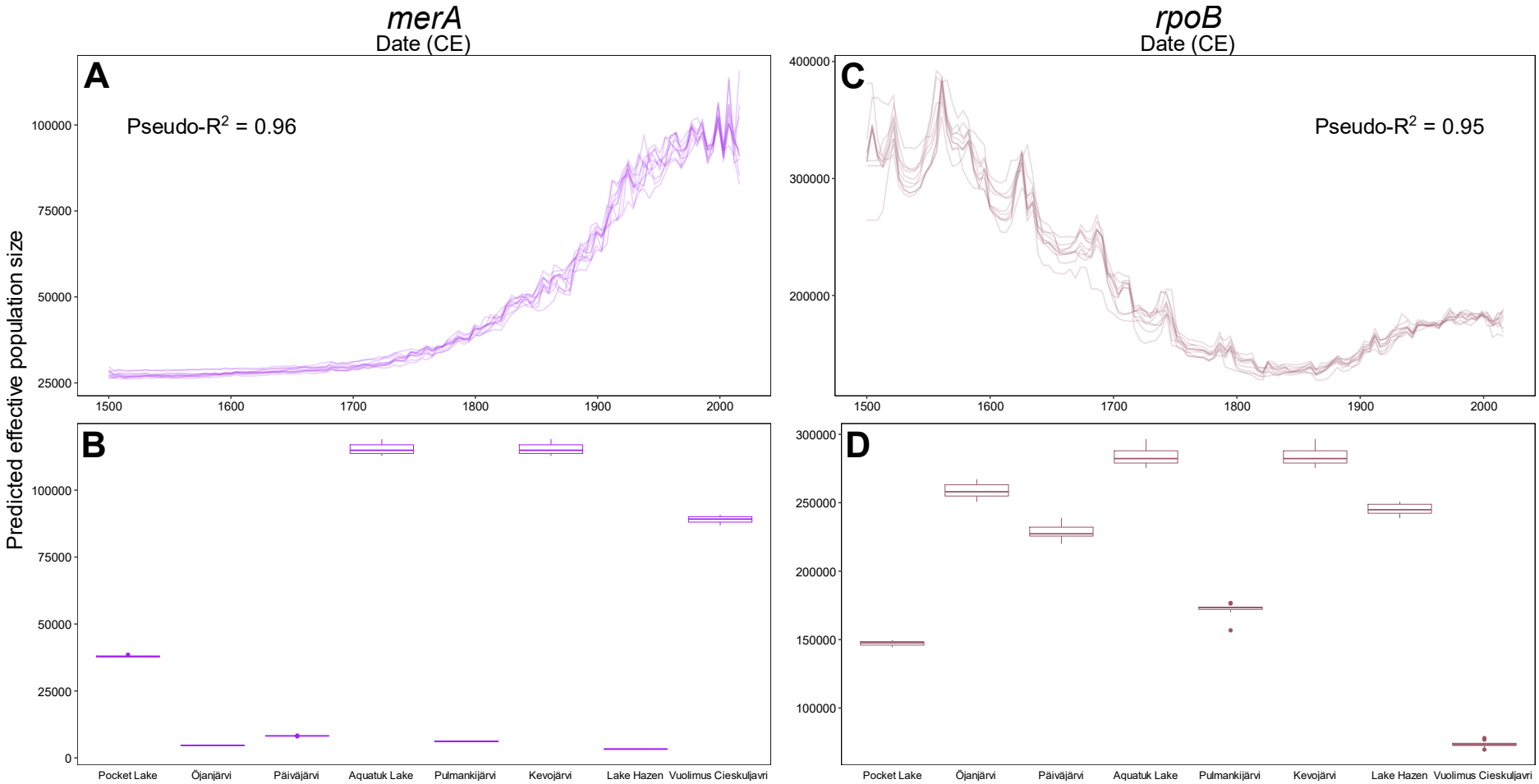

Figure C5. Partial dependence of predicted effective population sizes. Shown for *merA* (**a**, **b**) and *rpoB* (**c**, **d**) on their two predictors in the random forest models: CE date (**a**, **c**), and sampling site (**b**, **d**). Prediction accuracies are indicated in the top panels as pseudo-R^2. Each line in a and b and the variability per site shown in the box plots (**b**, **d**) shows the predictions of each of the 10 models trained and tested on different random splits of the data.

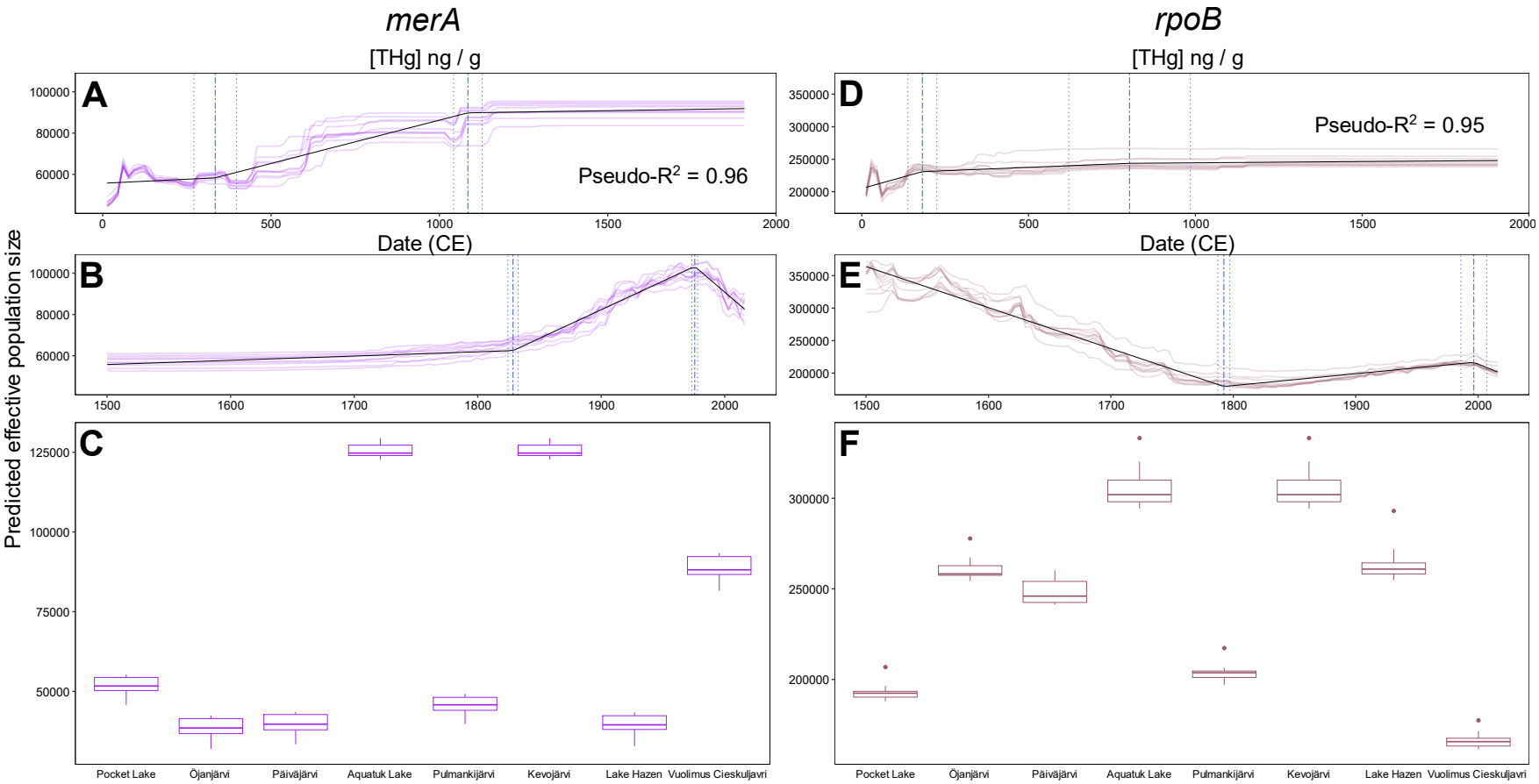

Figure C6: Partial dependence of predicted effective population sizes. Shown for *merA* (**a**, **b**, **c**) and *rpoB* (**d**, **e**, **f**) for the alternate random forest models, based on [THg] (**a**, **d**), CE date (**b**, **e**) and sampling site (**c**, **f**). Prediction accuracies are indicated in the top panels as pseudo-R^2. Each line in **a**, **b**, **d** and **e**, and the variability per site shown in the box plots (**c**, **f**) shows the predictions of each of the 10 models trained and tested on different random splits of the data. Three-part segmented linear models (black lines) were fit to the [THg] and date results for *merA* ([THg] R^2 = 0.88; date R^2 = 0.95) and *rpoB* ([THg] R^2 = 0.46; date R^2 = 0.95). The blue lines show the breakpoint estimates (dot-dash lines) and their 99% CIs (dotted lines).

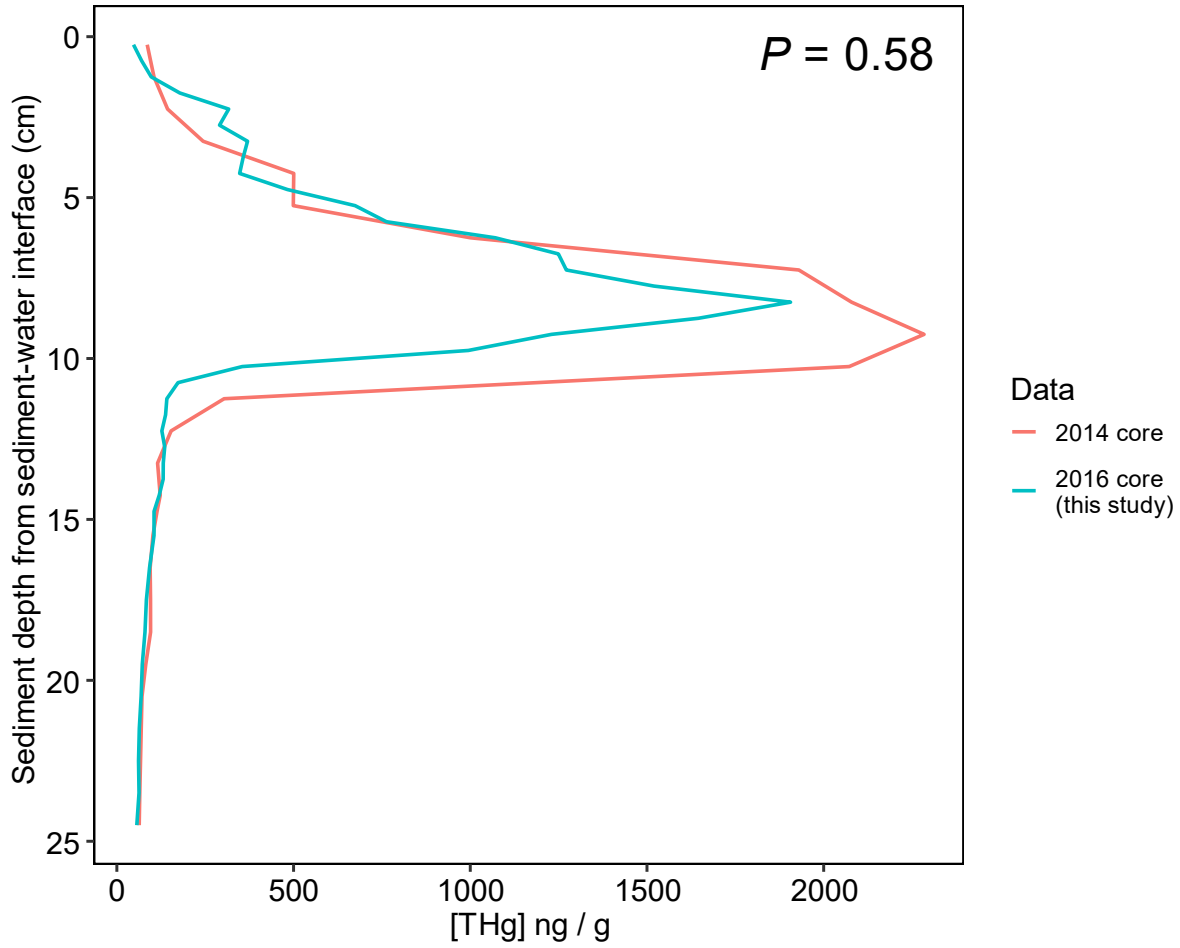

Figure C7: DNA extraction, chemistry and dating. Comparison of previously published data of [THg] in a sediment core from Pocket Lake (red line; Thienpont et al., 2016) to [THg] data from the core analyzed in the current study. The *P*-value is based on the Kolmogorov-Smirnov test.

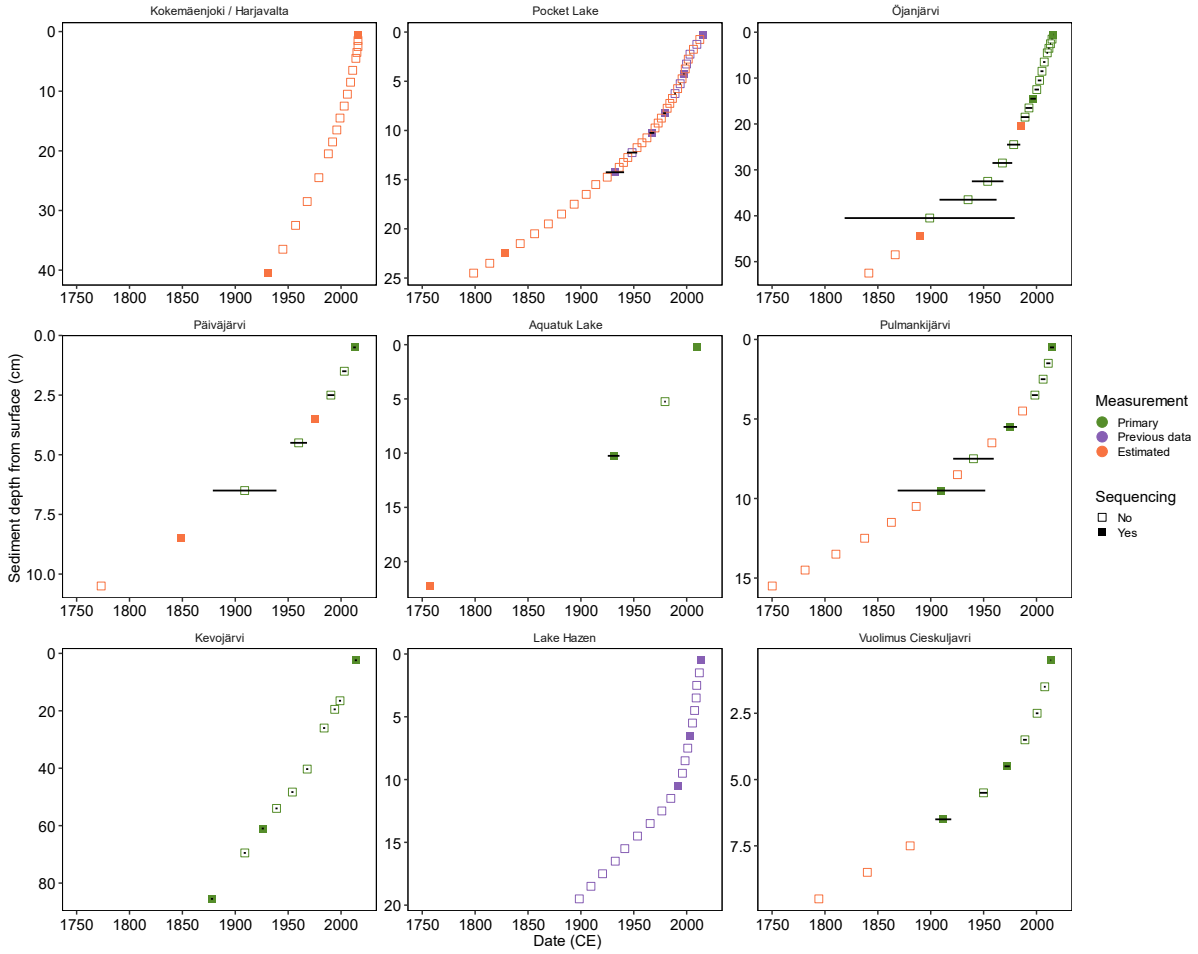

Figure C8: Dating profiles of sediment cores examined in the current study. Samples used for high-throughput sequencing of the *merA* and *rpoB* genes are shown as filled squares. All fits of second order polynomials to the measured dates, used in the extrapolations, had $R^2 > 0.97$. The ^{210}Pb profile of the Kokemäenjoki / Harjavalta core appeared to be mixed; the dating shown here, without error bars, was not used in subsequent analyses.

Table C1: Gene quantification with droplet digital PCR and amplicon sequencing. PCR primer pairs used in the study.

Target gene	Name	Used in	Sequence (5' – 3')	Product length	Master mix (per reaction)		PCR conditions	Reference
Mercuric reductase *(merA)*	qmerA1 F	ddPCR	CAT GAC GGT GCA GGA ACT G	101 bp	QX200 MM (2x) H_2O 10 µM FW/RV primers Template **Total**	11 µL 8.33 µL (ea.) 0.33 µL 2 µL **22 µL**	95°C 5 min 40x {95°C 30 s, 60°C 1 min} 4°C 5 min 90°C 5 min 10°C ∞	(Gill, 2012)
	qmerA1 R		GCT GCT TCA CAT CCT TGT TG					
	NlfF	Nested PCR (long product)	CCA TCG GCG GCA CYT GCG TYA A	1247 bp	EconoTaq PLUS MM (2x) H_2O 10 µM FW/RV primers Template **Total**	25 µL 19 µL (ea.) 2.5 µL 1 µL **50 µL**	94°C 2 min 25x {94°C 30s, 61°C 30s, 72°C 90s} 72°C 5 min 10°C ∞	(Wang et al., 2011)
	NlfR		CGC YGC RAG CTT YAA YCY YTC RRC CAT YGT					
	NsfF	Nested PCR (short product)	ACA CTG ACG ACA TGG TTC TAC A	308 bp	EconoTaq PLUS MM (2x) H_2O 10 µM FW/RV primers Template from NlfF-NlfR PCR **Total**	12.5 µL 9 µL (ea.) 1.25 µL 1 µL **25 µL**	94°C 2 min 35x {94°C 30s, 61°C 30s, 72°C 30s} 72°C 5 min 10°C ∞	(Wang et al., 2011)
	NsfR		CGC YGC RAG CTT YAA YCY YTC RRC CAT YGT (same as NlfR)					
	NsfF_CS1	Illumina MiSeq sequencing	ACA CTG ACG ACA TGG TTC TAC AAT CCG CAA GTN GCV ACB GTN GG	352 bp	(Similar to NsfF – NsfR above)			(Wang et al., 2011) Common sequence 1 or 2 tags added as required by Illumina MiSeq
	NlfR_CS2		TAC GGT AGC AGA GAC TTG GTC TCG CYG CRA GCT TYA AYC YYT CRR CCA TYG T					
Glutamate synthetase *(glnA)*	GS1β	ddPCR	GAT GCC GCC GAT GTA GTA	153-156 bp	QX200 MM (2x) H_2O 10 µM FW/RV primers 1:10 diluted template **Total**	11 µL 8.33 µL (ea.) 0.33 µL 2 µL **22 µL**	95°C 5 min 40x {95°C 30 s, 60°C 1 min} 4°C 5 min 90°C 5 min 10°C ∞	(Hurt et al., 2001)
	GS2γ		AAG ACC GCG ACC TTY ATG CC					
Ribosomal polymerase subunit B *(rpoB)*	rpoB_ssF	PCR	CDG AAG GYC CRA ACA TYG	375 bp	EconoTaq PLUS MM (2x) H_2O 10 µM FW/RV primers Template **Total**	12.5 µL 9 µL (ea.) 1.25 µL 1 µL **25 µL**	94°C 2 min 35x {94°C 30s, 53°C 30s, 72°C 30s} 72°C 5 min 10°C ∞	(This study)
	rpoB_ssR		CYT GRC GYT GCA TGT TRG					
	rpoB_ssF_CS1	Illumina MiSeq sequencing	ACA CTG ACG ACA TGG TTC TAC ACD GAA GGY CCR AAC ATY G	419 bp	(Similar to rpoB_ssF – rpoB_ssR above)			(This study) Common sequence 1 or 2 tags added as required by Illumina MiSeq
	rpoB_ssR_CS2		TAC GGT AGC AGA GAC TTG GTC TCY TGR CGY TGC ATG TTR G					

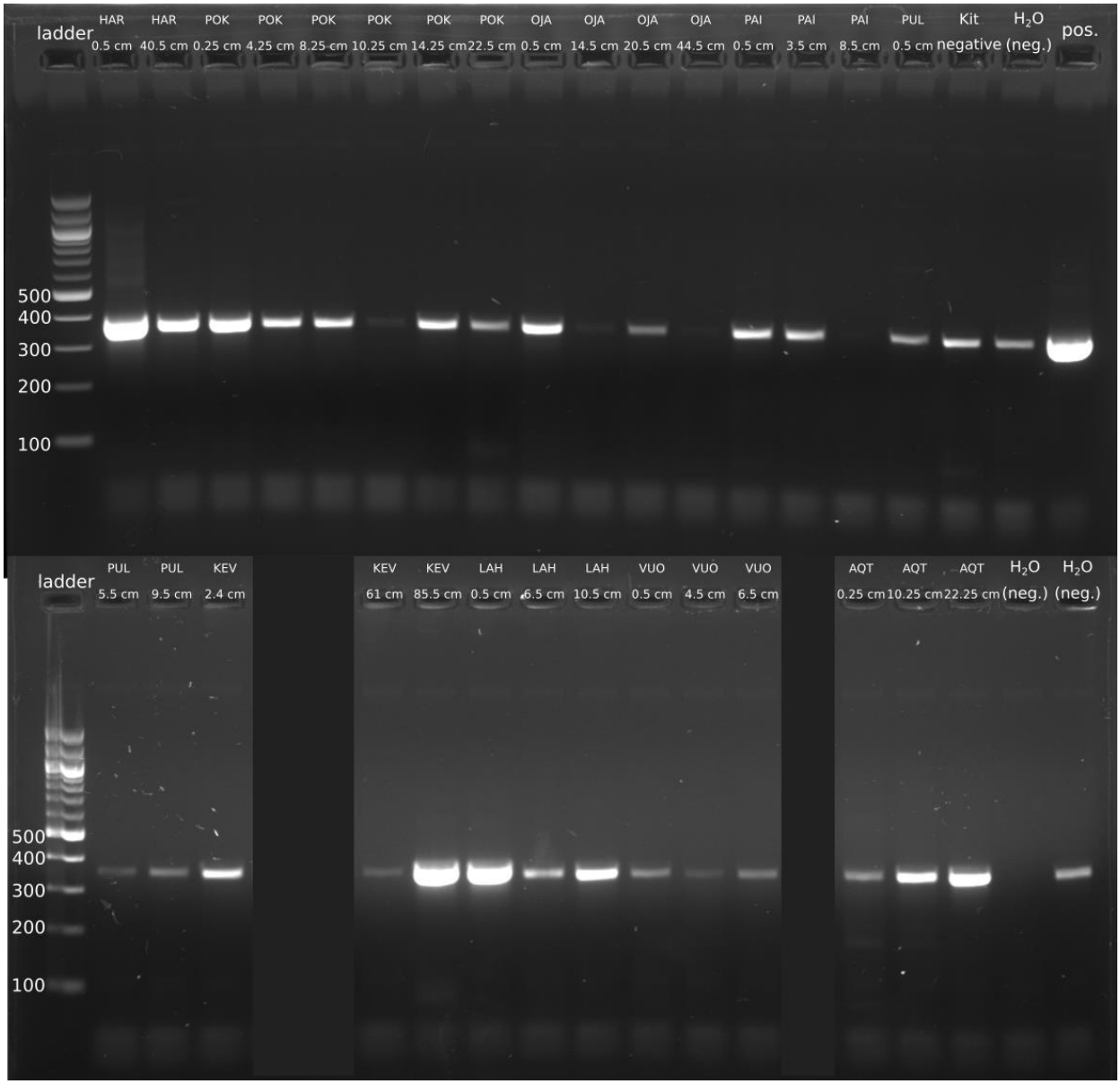

Figure C9: Amplicon sequencing. PCR screening of the samples selected for sequencing of the merA gene using first the NlfF/NlfR primers followed by the NsfF-CS1/NlfR-CS2 primers. The reaction conditions are outlined in **Table C1**. Electrophoresis was performed in a 1.5% agarose gel in a 1× Sodium-Borate buffer with 0.5 µg mL-1 EtBr at 11V / cm for 30 min. NEB QuickLoad Purple 100 bp ladder was used (5 µL) and 10 µL of each sample + 2 µL loading dye was loaded. The positive control was a Tn501 plasmid containing the *mer*-operon.

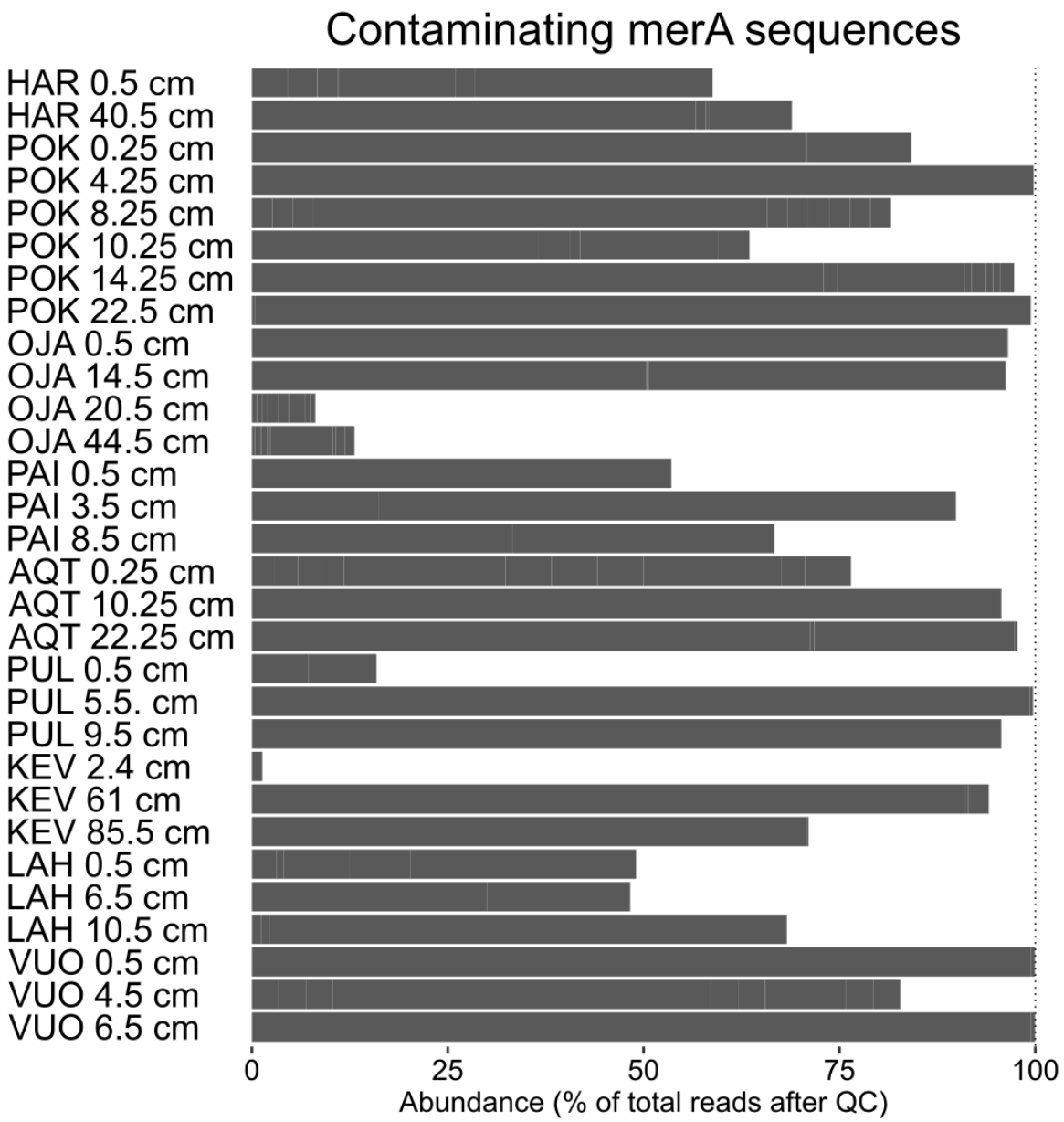

Figure C10: Sequencing data processing. Proportion of total reads in the samples identified as contaminants (any *merA* variants present in the negative control).

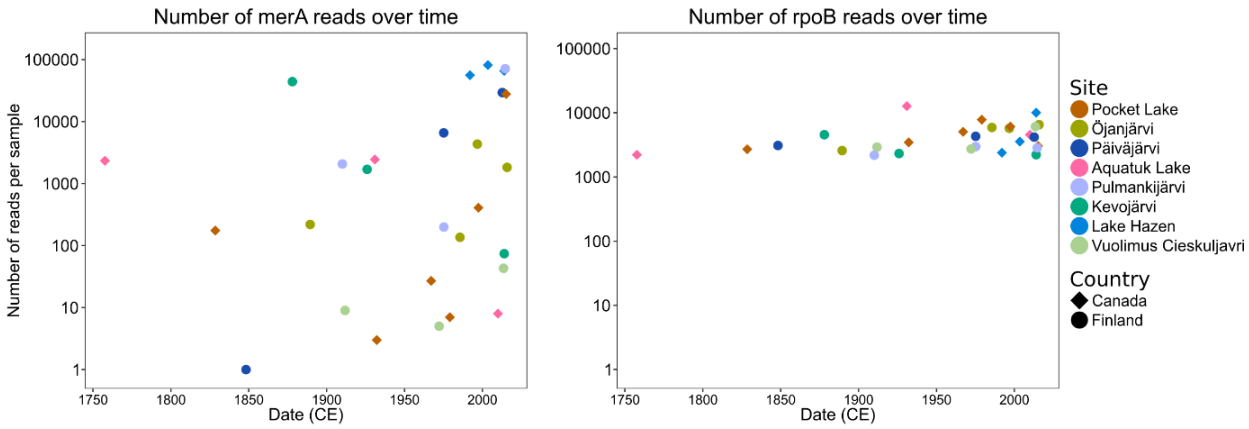

Figure C11: Sequencing data processing. Number of *merA* and *rpoB* reads per sample after quality control, with estimated calendar dates of the samples to show data coverage at different time points.

Appendix D: Supporting online information

Appendix D can be found online at (https://github.com/Begia/PhD-thesis) and contains the following supporting information:

For **Chapter 2:**

I. Sequence handling shell scripts:
    a. One shell script for each of the three data sets (Spring 2014/2015, Summer 2015 Archaea, Summer 2015 Bacteria): *Dataset*_Sequence_handling_shell_script.txt

II. Custom R scripts (used by the sequence handling scripts) to construct OTU tables:
    a. swarm_construct_otu_table.R, and uc_to_OTU_table.R

III. Custom FAPROTAX databases used to map OTU abundances into functions by taxonomy
    a. FAPROTAX_Hazen.txt

IV. Data analysis script which uses the primary data to produce all the figures, tables, and other results in the manuscript:
    a. Data_analysis_scripts.R

For **Chapter 3:**

I. Primary shell scripts and two accessory (parallel) shell scripts used to process the primary data from raw Illumina reads
    a. Primary_shell_scripts.sh, IPRS_array_proka.sh, and Metagenome_array_proka.sh

II. Data analysis script which uses the primary data to produce all the figures, tables, and other results in the manuscript:

    a. Data_analysis.R

III. **Tables D1, D2, D3, D4, D5, D6** & **D7**

IV. Full phylogenetic tree (see **Figure 3.2)** with annotated support values and full NCBI taxonomy of the genomes:

    a. Full_tree_with_supports.pdf, and Full_tree_with_supports.nwk

For **Chapter 4:**

I. Shell scripts used to process the primary data from Illumina read files:

    a. merA_runs.sh, and rpoB_runs.sh

II. Custom R and Python scripts (used by the shell scripts) to construct gene variant tables:

    a. swarm_construct_otu_table.R, and uc_to_OTU_table.py

III. Data analysis script which uses the primary data to produce all the figures, tables, and other results in the manuscript:

    a. analysis_script.R