THE TAXONOMIC AND FUNCTIONAL MICROBIAL DIVERSITY IN LAKE BAIKAL AND OTHER NORTH TEMPERATE LAKES

By

Paul Wilburn

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Integrative Biology—Doctor of Philosophy
Ecology, Evolutionary Biology, and Behavior—Dual Major

2018

ABSTRACT

THE TAXONOMIC AND FUNCTIONAL MICROBIAL DIVERSITY IN LAKE BAIKAL AND OTHER NORTH TEMPERATE LAKES

By

Paul Wilburn

Microorganisms cycle nutrients in every environment on Earth, and their importance in aquatic environments has been recognized for at least 75 years. However, many systems key to better understanding the role specific microbes play in natural environments remain poorly characterized. Lake Baikal is a UNESCO world heritage site. It is the planet's deepest (1642 m), most voluminous (23615 km^3), and oldest (25 to 30 my) lake, containing about 20% of world's unfrozen freshwater. Baikal's size and millions of years of evolutionary development have turned this ancient system into a biodiversity hotspot; however, little is known about its microbial communities. I describe what is the first -omic based survey the microbial communities of Lake Baikal, covering all three basins, multiple depths, and including measured environmental covariates.

In Chapter One, I show that temperature, stratification, nutrients, and dissolved oxygen define major microbial habitats and influenced patterns of community diversity in summer Lake Baikal. The environment, not geographical distance, structured microbial communities in Lake Baikal. The overall main driver of community dissimilarity was temperature. Increases in community diversity are driven by richness in the upper mixed layer and evenness in the deep waters, and those aspects of diversity were associated with different environmental drivers. Next, we used a co-occurrence network to identify lake habitats consistently preferred by groups of co-occurring microorganisms,

discovering two sets of candidate resident and two sets of candidate transient habitat-cohort pairs. Taxonomic makeup reflected the abiotic conditions of those clusters, suggesting key microbial players in each one.

In Chapter Two, I expand microbial community and functional surveys to thirteen additional lakes across Michigan, Minnesota, and Wisconsin, sampled in summer and winter seasons. Lake Baikal indeed harbored microbial communities that were distinct from other north temperate lakes in both seasons, with the next closest communities supported by oligotrophic epilimnia of lakes Superior, Portsmouth, and La Salle. In summer epilimnion of Lake Baikal, which was N-P co-limited at the time of the survey, the enzymes responsible for assimilatory reduction of N species to ammonium and assimilation of ammonium into glutamate were present in ferredoxin-dependent at the low end of N availability gradient, in a trade-off with NADH-dependent, isoforms.

Chapter Three presents 369 high quality draft genomes of microorganisms from Lake Baikal, assembled using computational tools that are currently at the cutting edge of bioinformatics. The metagenome assembled genomes (MAGs) were culture-independent and included the archaea domain, as well as 15 bacterial phyla, four of which have no previously sequenced lineages from Lake Baikal. Most MAGs were small but with large variation. At the same time, genomes assembled from the most stable, aseasonal, and resource environment in the Lake Baikal hypolimnion harbored the smallest genomes with remarkably little size variation, reflecting the oligotrophic environment.

ACKNOWLEDGEMENTS

It is difficult to overstate the role of my advisor, Elena Litchman, in enabling this work. Her efforts provided the funding, resources, and working climate, leaving me with exclusively upbeat “PhD advisor” stories. I am mostly unable to relate to the soul-searching PhD experience folklore told by many others and dramatized in niche comic books. That’s a good thing. Elena’s patience, constructive criticism, and encouragement were relentless from the project’s inception, its development, and ultimately refinement of the final product that is this dissertation. I would also like to thank my committee members: Christopher Klausmeier, Gary Mittelbach, Ashley Shade, James Tiedje, Andrew Allen, and Sarah Evans for their support, guidance, and feedback. They’ve been indispensable with helping to ask the right questions, develop methods, and pointing out ways to more robust interpretations of the results. I look forward to current and future collaborations with all.

Collaborators who have been with me for the past six and a half years were equally indispensable. I thank Ted Ozersky and Kirill Shchapov for two summers and a winter at Lake Baikal, a summer and a winter around lakes in Minnesota and Wisconsin, and countless scientific discussions. Lev Yampolsky and Ed Theriot were also critical in Lake Baikal sampling plan design and primers on bioinformatics and phylogenetics. Pam Woodruff and Allyson Hutchens assisted in the laboratory with nutrient concentration measurements. I am also grateful to the crew of the R.V. Treskov of the Limnological Museum in Listvyanka and field technicians Elena Pislegina and Alexander

Pislegin at the Institute of Experimental Biology, Irkutsk State University for facilitating fieldwork.

I cannot imagine making it through graduate school without the continuous support from my parents, Olivia Wilburn and Alan Mullis. Regularly visiting the parental safe haven in the San Francisco Bay Area has always been an example of how vision and hard work can create and maintain success even in a rapidly changing environment.

Last but not least, I would like to thank Shannon Carvey, who came into my life and halfway through this program. Her optimism, dedication to her work, and love inspired getting through the more arduous second half of writing what follows.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Habitat change and biodiversity loss are some of the most pressing issues in modern times. Addressing these issues echoes fundamental questions in ecology, such as characterizing the mechanisms that organize ecological communities, identifying key community members, and explaining how member traits mediate community function. At the base of arguably every ecosystem are microorganisms. Microbial communities are the ones most responsible for assimilation of inorganic nutrients and remineralization of organic compounds, forming the base of natural food webs. This dissertation focuses on microbial communities in Lake Baikal, Siberia - the world's largest and most ancient lake with many ocean-like properties, such as oligotrophic waters, hydrothermal vents, and the astounding biodiversity at multiple trophic levels. This makes Baikal a candidate connecting piece between what we understand of freshwater and marine ecology. And that is especially important in microbial ecology, which, as a discipline, is presently at an important milestone, attempting to organize a deluge of -omic data from natural systems into a unifying framework.

State of aquatic microbial ecology

Microbial ecology was conceived in the late 19th century when the scientific community realized that microorganisms were pivotal in the "cycle of life". That was remarkable, given the contemporary methods, which were limited to relatively simple models and experiments, and confined to microorganisms that could be observed in culture. Pasteur was first to show a fundamental maxim - that a chemical process

(fermentation of wine) was executed by living bacteria, thus establishing that microbes, in principle, can perform chemical transformations. That was in 1857. The next two decades included multiple discoveries that identified specific microbial taxa capable of key biogeochemical processes, notably sulfur reduction by *Desulfovibrio desulfuricans* and nitrogen fixation by rhizobia – both by M.W. Beijerinck. However, Sergei Winogradsky is now considered by many as the first microbial ecologist for postulating nutrient cycles. By 1880, having discovered microbial sulfur oxidation by *Beggiatoa* and nitrification by *Nitrosomonas*, *Nitrosococcus*, *Nitrobacter* in his early work, he heard of Beijerinck's success with nitrogen-fixing rhizobia. Winogradsky realized that the various microbially-driven transformations of nitrogen and sulfur species were all part of closed nutrient cycles, with cycle components executed by specialized bacterial taxa. During later work in Zurich, he discovered the microbial basis of nitrification, thus completing the understanding of the N cycle that was the paradigm for over 100 years, until the discovery of anammox in 1999 (Strous *et al.* 1999). Winogradsky was the first to isolate nitrifying bacteria and show that the steps necessary for oxidation of ammonia to nitrite and of nitrite to nitrate were separate and could be performed by different bacterial species. Thus, his contribution to microbial ecology was the concept of microbial-mediated nutrient cycling. However, the role of microbes in real natural systems, such as lakes or oceans, remained difficult to quantify.

Early aquatic microbial ecologists acknowledged that, while a key component of natural ecosystems, microorganisms were largely *terra incognita*. Although Lindeman placed microbial "ooze" at the center of Fig. 1 in his landmark publication on the trophic

food web in Cedar Bog Lake, he had little understanding of what controlled microbial dynamics in his study system (Lindeman, 1942). At the nearly the same time, Riley recognized the importance of microorganisms in marine biogeochemistry, and was more outspoken in his dismay that methods to advance understanding of microbial communities did not yet exist (Riley, 1951).

The leap forward came about a decade later with the development of dyes and radioactive tracers. Stains, when combined with isotopes, can be used to measure processes, such as bacterial production rates, independently from other food web constituents. Among the main conclusions was that, across multiple freshwater and marine environments, the large surface-to-area values of bacteria make them the best competitors for scarce dissolved nutrients, compared to phytoplankton, making them particularly important for biogeochemistry in oligotrophic systems (Cotner and Biddanda, 2002). For example, Fuhrman et al. (1989) calculated that heterotrophic bacteria accounted for 70% of carbon and over 80% of particulate nitrogen in the photic zone of the oligotrophic Sargasso Sea. However, the limiting nutrients for primary and bacterial production seemed to differ between freshwater and marine systems.

Nitrogen (N) was established as limiting in most of the oceans, while phosphorus (P) was determined limiting in north temperate lakes (Sterner, 2008). In the realm of freshwater systems, Schindler (1977) used whole-lake manipulations (well-known lake #226) to demonstrate the primacy of P limitation. In his seminal paper, he reasoned that transient or conditional N limitation in lakes is possible, but, given its abundance in the atmospheric reservoir, plankton will ultimately overcome the shortage (Schindler, 1977).

Meanwhile, oceanographers in the mid 1960s already recognized the ammonium deficits in marine anoxic zones (Richards, 1965). Numerous subsequent studies measuring rates of denitrification and the more recently discovered anammox (Strous *et al.* 1999), confirmed widespread N-limitation in the oceans (Francis *et al.*, 2007).

However, as is the case with many simple explanations, the paradigms of P and N limitation in freshwater and marine systems came with multiple exceptions. Since Schindler's lake manipulations, a number of experimental studies pointed to co-limitation by nitrogen and phosphorus, during periods of peak productivity (Elser *et al.*, 1990; Guildford and Hecky, 2000; Sterner, 2008; Harpole *et al.*, 2011; Elser *et al.*, 2007). O'Donnell *et al.* (2017) demonstrated such case for Lake Baikal. Thus, even if the role of N in freshwater systems is transient or conditional, the consistency of exceptions to P limitation clearly makes N an important factor in determining year-round ecosystem productivity and functioning.

The next major advance in aquatic microbial ecology was enabled by environmental DNA sequencing. Surveys of taxonomic diversity came first in the wake of amplicon sequencing of the 16s rRNA gene (Schmidt *et al.*, 1991). By the early 2000s, introduction of next-generation sequencing technologies allowed high throughput characterization of genetic material directly from environmental sources without the use of primers. This ushered an onslaught of the -omic surveys of numerous marine and freshwater environments, where various high-throughput approaches revealed the astounding taxonomic and functional diversity of aquatic microorganisms (Rappé and Giovannoni, 2003; Riesenfeld *et al.*, 2004; DeLong, 2009; Newton *et al.*, 2011). Several

key bacterial and archaeal taxa were associated with specific functions in particular environments. For example, the deep water marine anammox has been associated with Crenarchaeota and about a dozen bacterial genera, which were shared between marine and permanently anoxic freshwater environments (Humbert *et al.*, 2010). At the same time, some clades, like the marine SAR11 and freshwater Actinobacteria *acI* are, as far as we can tell, globally widespread (Rappé *et al.*, 2002; Giovannoni *et al.*, 2005; Šimek *et al.*, 2010). A comprehensive review of the taxonomic and functional discoveries is outside the scope of this introduction. Suffice it to point out that the function (genes) following the form (taxonomy) happens only sometimes, with the connection seemingly mediated by the many abiotic and biotic interactions, opening the door to staggering complexities (Shade, 2017).

In the largest meta-analysis of aquatic organisms to date, Louca *et al.* (2016) found that, while the distribution of functional genes followed abiotic conditions, microbial taxonomic composition had no such trends. However, the authors suggested that both function and form should be considered when characterizing processes in the environment because some genes could catalyze different – and sometimes reverse – processes. For example, variants of the sulfite reductase gene could be involved in either respiratory sulfur reduction or lithotrophic sulfur oxidation. However, sulfite oxidizers were found to be generally more abundant than sulfate respirers in the mesopelagic zone, indicating that dsrAB genes detected there mainly carried out sulfur oxidation. Winogradsky would be excited about that story. In 2005, DeLong expressed hope that "In the near future, ocean microbial genomics will continue to mine complex

community datasets to better understand how community gene content maps onto taxonomic composition, metabolic repertoire and phenotypic expression." After over a decade, that statement is more true than ever before.

This dissertation

Today, many systems with potential treasure troves of biodiversity and evolutionary insights still remain poorly characterized. Even more sorely needed are attempts to formulate the mechanistic explanations for the observed trends in bacterioplankton taxonomy and function. In this dissertation I explore planktonic microorganisms of Lake Baikal and place their diversity and functional repertoire in the context of other north temperate lakes.

Lake Baikal is a UNESCO world heritage site. It is the planet's deepest (1642 m), most voluminous (23615 km^3), and oldest (25 to 30 my) lake, holding as much water as all Laurentian Great Lakes combined or about 20% of world's unfrozen freshwater reserves (Moore *et al.*, 2009). Baikal's size and millions of years of evolutionary development have turned this ancient system into a biodiversity hotspot. Of the approximately 2600 animal species, two-thirds are endemic to the lake and are not found anywhere else (Sherbakov, 1999). This high degree of endemism makes Lake Baikal especially vulnerable to biodiversity loss. Indeed, Baikal has been undergoing warming since long-term monitoring began in 1941, leading to decline in zoo- and phytoplankton species with a concurrent increase of cosmopolitan competitors (Hampton *et al.*, 2008). In addition, as recently as within the last five years, Baikal has

drawn international attention for emerging human-caused (tourism, farming practices, and railroad industry) environmental issues, hallmarked by explosive *Spirogyra* blooms in previously pristine parts of the lake (Volkova *et al.*, 2018). Now, more than at any other time, it is important to advance our knowledge and ability to predict how ecosystems will respond to interacting anthropogenic stressors.

In Chapter One, I show that temperature, stratification, nutrients, and dissolved oxygen define major microbial habitats and influence patterns of community diversity. Co-authors and I show, first of all, that the environment, not geographical distance, structures microbial communities in Lake Baikal. The overall main driver of community dissimilarity is temperature. However, a closer look at the two stratified layers revealed multiple layers of complexity. We used exhaustive model averaging of multiple linear models, to show that increases in community diversity are driven by richness in the upper mixed layer and evenness in the deep waters, and that those aspects of diversity are associated with different environmental drivers. Next, we use data-guided approaches, i.e., a co-occurrence network analysis, to show which lake habitats are consistently preferred by groups of co-occurring microorganisms, rather than assume that habitats are assigned based on its phyto- and zooplankton composition carry over. We contrast environmental preferences of those co-occurring microbial clusters, demonstrate that the taxonomic makeup reflects the abiotic conditions of those clusters, and highlight the potential key microbial players in each one.

In Chapter Two, I expand microbial community and functional surveys to thirteen other lakes across Michigan, Minnesota, and Wisconsin, sampled in summer and winter

seasons. My co-authors and I show that Lake Baikal indeed harbors microbial communities that are distinct from other north temperate lakes, with the next closest communities inhabiting the oligotrophic epilimnia of lakes Superior, Portsmouth, and La Salle. We also show how seasonal "collapse" in the variability of abiotic factors is reflected in microbial communities of most lakes, except those that maintain strong stratification into the winter season. Perhaps the most interesting part of this chapter is the association between N limitation and the preferred mechanisms for N assimilation. In Lake Baikal, the enzymes responsible for assimilatory reduction of N species to ammonium and assimilation of ammonium into glutamate were present in ferredoxin-dependent, as opposed to NADH-dependent, isoforms. Lake Baikal was experimentally shown to be N and P co-limited by O'Donnell *et al.* (2017) at the exact same time the samples for this work were collected. Total nitrogen to phosphorus ratios suggested that in other surveyed lakes N-limitation was unlikely. We argue that N scarcity in Baikal was reflected in the abundance of N-poor ferredoxin-dependent N assimilation enzymes, in a trade-off with N-rich NADH-dependent isoforms. Additionally, we speculate that reducing power available in the particularly oligotrophic summer epilimnia could select for Fd-utilizing microbial taxa, which are mostly photosynthetic, in the summer season. For detailed discussion, see Chapter 3.

Chapter Three presents 369 high quality draft genomes of microorganisms from Lake Baikal, assembled using computational tools that are currently at the cutting edge of bioinformatics. The metagenome assembled genomes (MAGs) are culture-independent and cover the Archaea domain, as well as 15 bacterial phyla, four of which

have no previously sequenced lineages from Lake Baikal. Most MAGs are small but there is a significant variation in genome size. At the same time, genomes assembled from the most stable, aseasonal, and resource (labile carbon)-poor environment in Lake Baikal hypolimnion harbored the smallest genomes with remarkably little variation in genome size. Small genome size and genomic streamlining are common in prokaryotes that thrive in oligotrophic environments, such as *Prochlorococcus* in the oceans (Fernandez-Garcia *et al.*, 2004) and Actinobacteria *acI* in freshwater systems (Ghylin *et al.*, 2014; Kang *et al.*, 2017). Streamlining is thought to conserve scarce nutrients by minimizing DNA synthesis and expression requirements, while at the same time shrinking the overall cell size, which maximized the surface area to volume ratio. Our Baikal MAGs could, therefore, reflect the lake's overall oligotrophic environment, where millions of years allowed microorganisms to optimize the occupancy of available resource niches.

Overall, this dissertation aims to place results from contemporary molecular sequencing approaches into ecological context. Results of Chapters One and Two present biological insights into microbial plankton through a lens of abiotic covariates and biotic co-occurrences between numerous taxa. Chapter Three sets the stage for future model-based work on relationships between phylogenetic diversity and metabolic function in Lake Baikal and other natural systems. My hope is that the dissertation's overall scope will contribute ecological thinking to analyses of microbial communities, as well as help narrow the gap between microbial limnology and oceanography.

REFERENCES

REFERENCES

Behrenfeld MJ, Bale AJ, Kolber ZS, Aiken J, Falkowski PG. (1996). Confirmation of iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean. *Nature* **383**: 508–511.

Boyd PW, Watson AJ, Law CS, Abraham ER, Trull T, Murdoch R, *et al.* (2000). A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature* **407**: 695–702.

Cotner JB, Biddanda BA. (2002). Small players, large role: Microbial influence on biogeochemical processes in pelagic aquatic ecosystems. *Ecosystems* **5**: 105–121.
DeLong EF. (2009). The microbial ocean from genomes to biomes. *Nature* **459**: 200–206.

Elser JJ, Bracken MES, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, *et al.* (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol Lett* **10**: 1135–1142.

Elser JJ, Marzolf ER, Goldman CR. (1990). Phosphorus and Nitrogen Limitation of Phytoplankton Growth in the Freshwaters of North America: A Review and Critique of Experimental Enrichments. *Can J Fish Aquat Sci* **47**: 1468–1477.

Fernandez-Garcia JM, Tandeau de Marsac N, Diez J. (2004). Streamlined Regulation and Gene Loss as Adaptive Mechanisms in Prochlorococcus for Optimized Nitrogen Utilization in Oligotrophic Environments. *Microbiol Mol Biol Rev* **68**: 630–638.

Francis CA, Beman JM, Kuypers MMM. (2007). New processes and players in the nitrogen cycle: The microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J* **1**: 19–27.

Fuhrman JA, Sleeter TD, Carlson CA, Proctor LM. (1989). Dominance of bacterial biomass in the Sargasso Sea and its ecological implications. *Mar Ecol Prog Ser* **57**: 207–217.

Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, *et al.* (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *ISME J* 1–14.

Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL, *et al.* (2005). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82–85.

Guildford SJ, Hecky RE. (2000). Total nitrogen, total phosphorus, and nutrient limitation

in lakes and oceans: Is there a common relationship? *Limnol Oceanogr* **45**: 1213–1223.

Hampton SE, Izmest'Eva LR, Moore M V., Katz SL, Dennis B, Silow E a. (2008). Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia. *Glob Chang Biol* **14**: 1947–1958.

Harpole WS, Ngai JT, Cleland EE, Seabloom EW, Borer ET, Bracken MES, *et al.* (2011). Nutrient co-limitation of primary producer communities. *Ecol Lett* **14**: 852–862.

Humbert S, Tarnawski S, Fromin N, Mallet M-P, Aragno M, Zopfi J. (2010). Molecular detection of anammox bacteria in terrestrial ecosystems: distribution and diversity. *ISME J* **4**: 450–454.

Hutchins DA, Bruland KW. (1998). Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime. *Nature* **393**: 561–564.

Kang I, Kim S, Islam MR, Cho JC. (2017). The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. *Sci Rep* **7**: 1–14.

Lindeman RL. (1942). The Tophic-Dynamic Aspect of Ecology. *Ecology* **23**: 399–417.

Louca S, Parfrey LW, Doebeli M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science (80-)* **353**: 1272–1277.

Moore MVM, Hampton SES, Izmest'eva LLR, Silow E a., Peshkova E V., Pavlov BK. (2009). Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia. *Bioscience* **59**: 405–417.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S, R. J. Newton, *et al.* (2011). A guide to the natural history of freshwater lake bacteria.

O'Donnell DR, Wilburn P, Silow EA, Yampolsky LY, Litchman E. (2017). Nitrogen and phosphorus colimitation of phytoplankton in Lake Baikal: Insights from a spatial survey and nutrient enrichment experiments. *Limnol Oceanogr* 1383–1392.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630–633.

Rappé MS, Giovannoni SJ. (2003). The Uncultured Microbial Majority. *Annu Rev Microbiol* **57**: 369–394.

Richards FA. (1965). Anoxic basins and fjords. In: Riley JP, Skirrow G (eds). *Chemical oceanography Vol. 1*. Academic Press: New York, USA, pp 611–645.

Riesenfeld CS, Schloss PD, Handelsman J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**: 525–52.

Riley G. (1951). Oxygen, phosphate, and nitrate in the Atlantic Ocean. *Bull Bingham Ocean Coll* **13**: 1–124.

Schindler DW. (1977). Evolution of phosphorus limitation in lakes. *Science* **195**: 260–262.

Schmidt TM, DeLong EF, Pace NR. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* **173**: 4371–4378.

Shade A. (2017). Diversity is the question, not the answer. *ISME J* **11**: 1–6.

Sherbakov DY. (1999). Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal. *Trends Ecol Evol* **14**: 92–95.

Šimek K, Kasalický V, Jezbera J, Jezberová J, Hejzlar J, Hahn MW. (2010). Broad habitat range of the phylogenetically narrow R-BT065 cluster, representing a core group of the betaproteobacterial genus limnohabitans. *Appl Environ Microbiol* **76**: 631–639.

Sterner RW. (2008). On the phosphorus limitation paradigm for lakes. *Int Rev Hydrobiol* **93**: 433–445.

Strous M, Gijs Kuenen J, Jetten MSM. (1999). Key Physiology of Anaerobic Ammonium Oxidation. *Appl Environ Microbiol* **65**: 3248–3250.

Volkova EA, Bondarenko NA, Timoshkin OA. (2018). Morphotaxonomy, distribution and abundance of Spirogyra (Zygnematophyceae, Charophyta) in Lake Baikal, East Siberia. *Phycologia* **57**: 298–308.

CHAPTER ONE

ENVIRONMENTAL DRIVERS DEFINE CONTRASTING MICROBIAL HABITATS, DIVERSITY AND FUNCTIONAL REDUNDANCY IN LAKE BAIKAL

Abstract

Understanding how microbial communities respond to environmental change requires the knowledge of the main drivers of their community structure, diversity and potential resilience. For many rapidly changing ecosystems this information is still not available. Lake Baikal in Siberia is the most ancient, deep, voluminous, and biologically diverse lake in the world, with 20 percent of global unfrozen fresh water, that is undergoing rapid warming. Little is known about its bacterioplankton communities and their drivers. In the first extensive survey of Baikal's microbial communities, we show that temperature, stratification, nutrients, and dissolved oxygen, and not the geographic distance, define major microbial habitats and microbial community similarity. Communities in the mixed layer and deep waters exhibited contrasting patterns of richness, diversity and evenness and comprised different cohesive modules in the whole Baikal OTU co-occurrence network. The network exhibited small-world properties that may make it resistant to perturbations but sensitive to changes in the abundances of central, most connected OTUs. Functional redundancy, often associated with higher resilience, was low in cold, open water communities and increased with temperature in the upper mixed layer and with depth in the deep-water samples. Our results suggest that bacterial communities of open waters in Lake Baikal may be more sensitive to warming and other anthropogenic stressors than the littoral communities and may

reorganize significantly in the changing climate. A better understanding of factors structuring bacterial communities in ecosystems that undergo rapid changes, including Lake Baikal and other northern lakes, will allow us to better predict the overall ecosystem responses to anthropogenic stressors.

Significance

This study is the first to identify distinct bacterial assemblages in Lake Baikal during summer stratification and link them to specific environments. Multiple linear regression and model averaging show that community diversity is driven by richness in the mixed layer (ML) and evenness in the deep waters (DW), and we identify environmental covariates that explain these contrasting trends. Network analyses reveal assemblages specific to ML and DW, where phylogeny reflects preferred environments. We then use PICRUSt to predict metagenomes and reveal that redundancy is lowest in the coolest areas of ML, placing them at the greatest risk of microbial functional diversity loss. This is a significant step towards understanding and addressing ongoing challenges faced by biological communities in a changing world.

Introduction

The ecological importance of microorganisms in aquatic systems has been recognized at least since the appearance of “ooze” in Lindeman’s trophic energy transfer diagram (Lindeman, 1942). Their central place in material and energy fluxes is now recognized for nearly all nutrient cycles, with greater relative importance of

prokaryotic organisms in more oligotrophic systems (Cotner and Biddanda, 2002). The advent of next-generation sequencing, combined with environmental monitoring, enabled new discoveries of microbial diversity and function in various aquatic habitats. However, the environmental drivers of microbial community diversity, community structure, function, and stability remain poorly characterized in many aquatic ecosystems, including the world's most ancient (25 My) Lake Baikal – a UNESCO heritage site and known hotspot for endemism of its biota. Baikal is the world's deepest (1643 m) and most voluminous lake, holding about 20% of world's surface unfrozen freshwater (Moore *et al.*, 2009). Of the approximately 2600 plant and animal species in the lake, two-thirds are endemic, including the dominant primary producers, grazers, benthic and pelagic fish and the top predator – world's only freshwater seal (Moore *et al.*, 2009; Hampton *et al.*, 2008).

Hampton *et al.* (Hampton *et al.*, 2008) showed that water temperatures have risen by 1.2 °C over 60 years of high-resolution time series, contributing to an increase in numbers and kinds of non-endemic zooplankton and algal species, with potential consequences for nutrient cycling, food web structure (Moore *et al.*, 2009) and microbial communities. Moreover, the Lake Baikal region is predicted to warm by 3-4°C in the next century (Team *et al.*, 2014), with ongoing changes likely to continue and even accelerate. Because the biota of Lake Baikal, including microbial communities, is adapted to cold temperatures, it may be especially vulnerable to warming. Additionally, other changing environmental factors may also alter the lake's microbial communities.

The vulnerability of communities to environmental change depends in part on their functional redundancy (FR), linked to the number of species that perform a similar function (Holtzendorff *et al.*, 2008). High number of functionally similar species, i.e., high FR, is usually thought of as a mechanism that preserves community functions during disturbance. High bacterial diversity, rapid generation times, and the mobility of bacterial genes through genome rearrangement and horizontal gene transfer enables relatively quick functional loss or gain within a phylogenetic lineage, and, conversely, sharing among phylogenetic lineages. Indeed, prokaryotic functional genes are promiscuous when it comes to species boundaries and are often found in multiple taxa adapted to similar environmental conditions (Martiny *et al.*, 2006), resulting in functional redundancy. The degree to which Lake Baikal microbial communities exhibit such redundancy, and how it may vary over the multiple environmental and geographic gradients is unknown.

Here we present the first comprehensive survey and analysis of microbial plankton in Baikal, spanning all three basins, from open waters to the shallow bays, multiple depths, including the surface layer and depths below 300 m. We reveal major community composition trends that correlate with continuous environmental gradients in a spatial context. We then identify strongest environmental covariates to community composition in a multivariate framework. We use co-occurrence network analyses to reveal clusters of OTUs with contrasting environmental associations and identify clades and taxa most responsible for maintaining the structure of each network cluster. Finally, we use functional prediction tool PICRUSt to estimate functional redundancy and show

that it correlates with temperature and is lowest in open Baikal and in communities resident in its hypolimnion.

Materials and Methods

Our survey was guided by the recorded natural history of Baikal (Kozhov, 1963; Kozhova and Izmest'eva, 1998; Moore *et al.*, 2009) that divides the lake into eight distinct regions (Fig. 1.1a, S1.13). Among them, are Chivirkuy Bay, Proval Bay, and Selenga river plume. Chivirkuy Bay was sampled extensively to capture transition from the shallow innermost bay (9 m depth) to open waters. Proval Bay and Selenga plume stations represented the two most eutrophic areas in Baikal. In total, we collected samples from 24 stations, of which 10 were sampled at various depths for a total of 46 samples.

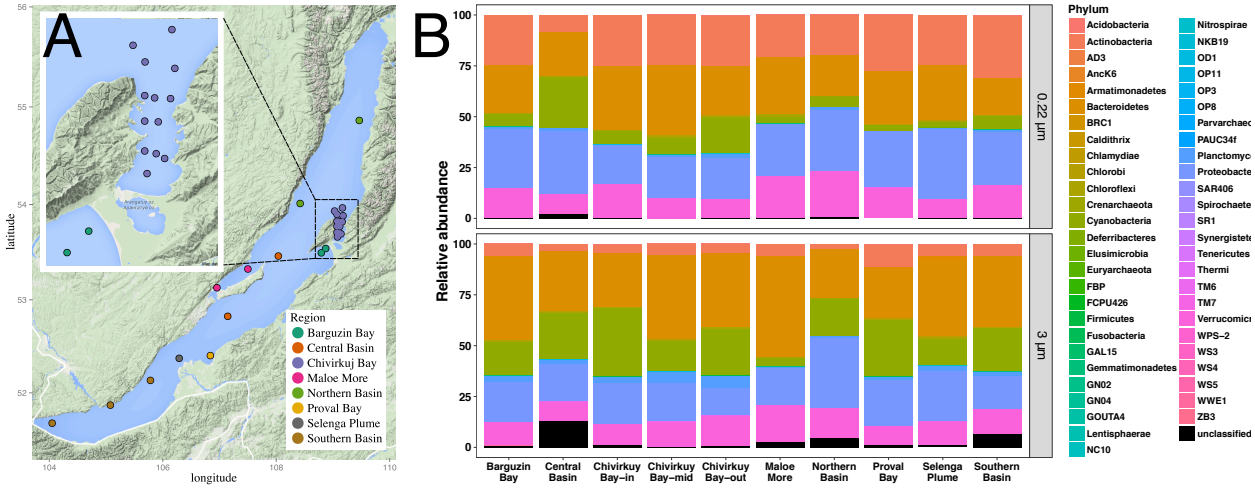


Figure 1.1: (A) Baikal sampling locations. Stations in outside of Chivyrkuy Bay, Proval Bay, and Selenga Plume were sampled at multiple depths. (B) Taxonomic composition of the free-living (0.22 µm, top) and particle-attached (3 µm, bottom) size fractions across s surface samples. While both fractions were mostly dominated by common freshwater phyla, Actinobacteria were enriched in the 3 µm fraction.

Temperature and dissolved oxygen profiles were measured with a YSI Instruments sonde (YSI, Inc., Yellow Springs, OH, USA). Whole water was taken using a Van Dorn sampler (Wildco, Inc., Yulee, Florida, USA), and 5 L were filtered onto 3 μm and 0.22 μm mixed nitrocellulose acetate membranes (EMD Millipore, Billerica, MA, USA) and stored at –20°C in RNAlater (Life Technologies, Grand Island, NY, USA) to capture particle-attached (3 μm filter) and free-living (0.22 μm filter) fractions of microorganisms. Genomic DNA was extracted using the Mo-Bio PowerSoil Kit (Mo-Bio Laboratories, Carlsbad, CA), following manufacturer's protocol. Then, the V4 region of the 16S rRNA gene was sequenced on a MiSeq platform (250PE), as described previously (Kozich *et al.*, 2013). Raw sequences were processed generally following the closed-reference mothur pipeline. Statistical analyses were performed in the R (3.2.2) environment, unless noted otherwise. Multiple regression search and model selection were done with the glmulti package (Calcagno and Mazancourt, 2010). Distance matrices, ordinations, and correlations with environmental variables we calculated with vegan (Oksanen *et al.*, 2018) and phyloseq (McMurdie and Holmes, 2013) packages. We constructed the co-occurrence network using sparCC (Friedman and Alm, 2012), following recommended best practices by Berry and Widder (Berry and Widder, 2014). Next, we identified network modules with the optimum modularity algorithm in the iGraph package (Csárdi and Nepusz, 2006). We followed the WGCNA package (Langfelder and Horvath, 2008) to generate eigenvectors for each module and correlate the first eigenvector (first principal component, PC1) with environmental variables to generate plots and a heatmap (Fig. 1.5). For the functional redundancy measure, the per sample number of

OTUs was divided by the gene content, estimated using PICRUSt software, with the nearest sequenced taxon index (NSTI) cutoff at 0.15, as recommended by the package authors, to select accurate gene copy number predictions (Langille *et al.*, 2013). Next, we validated the predictions using an in-house shotgun dataset available for a subset of samples. For genes that had greater than 10X mean shotgun coverage (4012 genes), PICRUSt predictions had a significant correlation ($r^2 = 0.35$, $p=0.017$) with the metagenomic community characterization, which was consistent with values reported in the original publication (Langille *et al.*, 2013). Detailed explanation of bioinformatics and statistical analyses is in the Supplement.

Results and Discussion

Taxonomic diversity

At a depth of 28059 sequences per sample, we detected 38457 OTUs in the combined 3.0 μm and the 0.22 μm fractions. Non-singletons (6099 in 3.0 μm and 3346 in 0.22 μm) were classified into 61 phyla, dominated by typical freshwater Actinobacteria, Bacteroidetes, Cyanobacteria, Proteobacteria and Verrucomicrobia (Fig. 1.1).

Microbial community composition notably differed from previous studies at Lake Baikal, which explored pelagic communities on Sanger (Bel'kova *et al.*, 2003; Denisova *et al.*, 1999) and Roche 454 (Parfenova *et al.*, 2013; Kurilkina *et al.*, 2016) platforms. In the most recent effort, Kurilkina and colleagues sampled one station depth profile in September and June (Kurilkina *et al.*, 2016) to reveal the main taxonomic groups as

Proteobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Chloroflexi, Acidobacteria, and Cyanobacteria in both seasons, and Caldiserica in September only. Our study did not detect Caldiserica and revealed only a very minor presence of Chloroflexi, while showing a substantial Verrucomicrobia presence (absent in Kurilkina *et al.*) in every sampled region (Fig. 1.1).

Our study revealed compositional differences between the 0.22 μm (free-living) and 3 μm (mostly particle-attached) fractions. Actinobacteria were significantly enriched on the 0.22 μm fraction (Table S1.1), similar to results from lakes in Michigan, USA (Schmidt *et al.*, 2016). Proteobacteria were only marginally more prevalent in the 0.22 μm fraction. We also found enrichment of Bacteroidetes and Cyanobacteria on the 3 μm fraction, the latter owing to filamentous taxa. Indeed, the top three most differentially abundant taxa identified with permutation-based analyses (Cáceres and Legendre, 2009) were all classified as the Nostocales genus *Dolichospermum* and contributed up to 90% of Cyanobacteria in the 3 μm fraction.

Free-living (0.22 μm) surface samples in the open Baikal were dominated by betaproteobacterium *Limnohabitans* sp. Littoral zones were more variable, represented by multiple OTUs classified as *Limnohabitans*, *Synechococcus* and *Actinobacteria* acI. Surface samples of the shallow Proval Bay and the Selenga river plume represented the extreme end of the eutrophic gradient and were dominated by Actinobacteria acI and Verrucomicrobia *Chthoniobacter*. The latter is reportedly incapable of growth on amino acids or organic acids other than pyruvate, suggesting the taxon is likely involved in the breakdown of partially oxidized organic carbon. Together with Bacteroidetes

Sediminibacterium and *Cytophagia*, and Betaproteobacteria *Limonhabitans* and *Polynucleobacter*, the six OTUs make up ~46% of each Proval Bay and Selenga River plume's relative abundance. The deepest samples in this study, collected at 500 m and 300 m, were dominated by Actinobacteria acI and acIV clades as well as ammonia oxidizing Group 1a Crenarchaeota *Nitrosopumilus*, commonly found in oligotrophic marine environments.

Community Richness and Diversity

Free-living community diversity showed opposing trends in the upper mixed layer (ML) and deeper waters (DW). Overall, richness and evenness strongly correlated with temperature in the ML and with depth in DW. For each sample, we calculated the effective number of species (ENS), a measure of diversity (Hill, 1973; Jost, 2006; Leinster and Cobbold, 2012). Hill (Hill, 1973) identified ENS as the number of equally-common species that yields a given value of a diversity index, such as the Shannon H'. Jost *et al.* (Jost, 2006) have argued that because ENS scales linearly with richness of equally-common species, it is the preferred metric for quantitative analyses (see Supplementary Methods). We show that the ENS increased with depth in DW (Fig. 1.2A) and was driven by an increase in evenness (Fig. 1.2C) with no accompanying trend in OTU richness (Fig. 1.2B). In the ML, ENS showed a modest increase at the surface in samples collected at 0 m (Fig. 1.2B). Unlike in DW, higher ML diversity was generated by higher OTU richness (Fig. 1.2B), while evenness remained unaffected. Mixed layer diversity was positively correlated with temperature (Fig. 1.2D), along with

OTU richness (Fig. 1.2E). Deep water diversity and richness were marked by high variability with respect to temperature, and richness showed a marginally significant increase in samples from cooler water. Evenness was not directly correlated with temperature in either layer (Fig. 1.2F). Furthermore, in mixed layer, richness had a greater response (slope) to temperature than diversity did, suggesting that samples at higher temperatures and shallower depths, while supporting the greatest richness, were dominated by a few successful OTUs. This could be due to variable and high resource conditions leading to coexistence of more taxa, including the persistence of rare taxa.

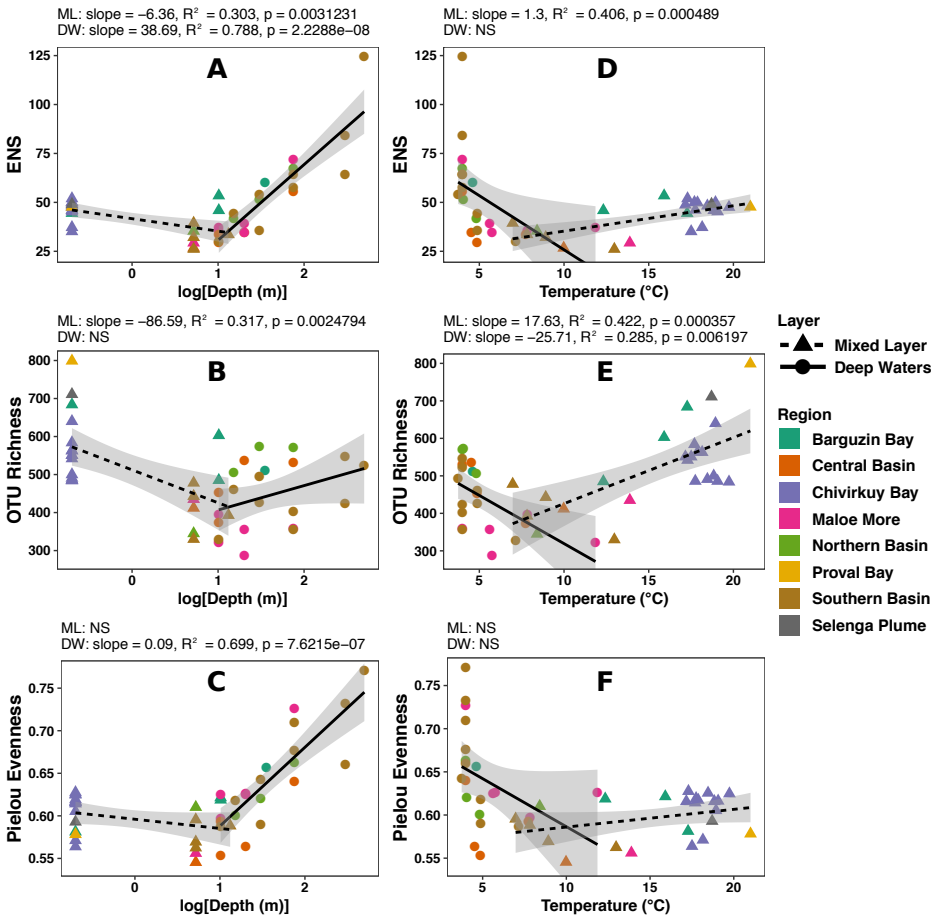


Figure 1.2: Diversity and evenness trends across depth and temperature in the upper mixed layer and deep waters. Effective Number of Species (ENS) was driven by richness in mixed layer and by evenness in deep waters.

Deep waters offer the opposite story. While richness had a significant but weak relationship with depth, ENS and Pielou evenness were very strongly positively correlated with depth (R^2=0.79, p=2.2x10^{-8}; R^2=0.70, p=7.6x10^{-7}), indicating that a substantial increase in diversity in the deep hypolimnion was not driven by richness but by an even community structure.

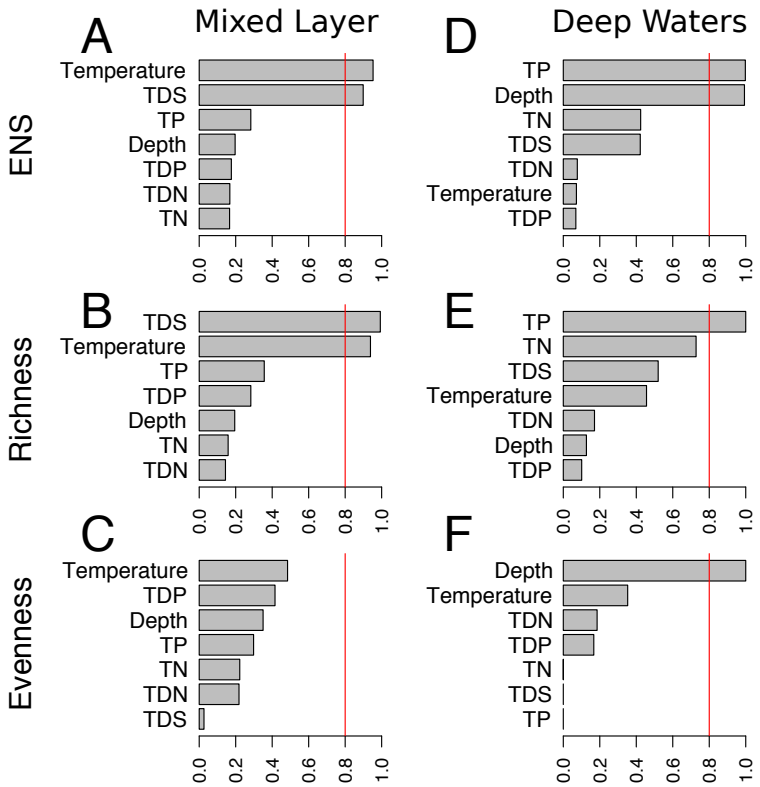


Figure 1.3: Model-averaged importance of environmental predictors for ENS diversity (top), OTU richness (middle) and Pielou Evenness (bottom) in the upper mixed layer (left) and deep waters (right).

These trends reveal an intriguing picture: warmer shallower part of ML and deeper parts of DW both support more diverse communities. But ML diversity, while rich in OTUs, is highly uneven, whereas the deep water communities achieve diversity by increasing

evenness. The trends were similar for the 3 μm size fraction (Fig. S1.12, S1.13). To better resolve underlying drivers that underpin these contrasting community features, we modeled ENS diversity, richness, and evenness with measured environmental variables using iterative multiple regression and model averaging (Calcagno and Mazancourt, 2010).

Diversity trends in the ML were different, compared with deep waters. Diversity in the mixed layer correlated the most with temperature and total dissolved silica (TDS, Fig. 1.3A). These covariates were also responsible for driving the OTU richness, while no measured environmental variable was a reliable predictor of ML evenness (Fig. 1.3B). Depth was not an important predictor for any ML community feature (Fig. 1.3C). These results highlighted the importance of temperature and silica in driving the ML diversity, by elevating OTU richness. In the deep waters, total phosphorus (TP) and depth correlated with ENS (Fig. 1.3D). Furthermore, only TP was a significant predictor for the deep water OTU richness, while only depth was for evenness (Fig. 1.3E, F). Temperature, when controlling for other covariates, was not an important community feature predictor in deep water. The importance of temperature in ML and depth in DW was not unexpected, considering that the surveyed ML spanned hundreds of kilometers across all three Baikal's basins and multiple bays, where temperature ranged from a median 9.5°C in open waters to 21°C in Proval Bay (Fig. S1.7), while DW was characterized by a large range of depths from 10 to 500 m and a relatively constant temperature environment. Additionally, nutrients appeared to associate specifically with OTU richness. Dissolved silica, combined with light availability in ML, is known to

promote diatom growth. Diatom exudates, acting as resources, could create additional niches, enabling coexistence of a greater number of OTUs. In deep waters, TP could promote OTU richness by supporting higher phosphorus demands often attributed to fast-growing copiotrophic organisms (Klausmeier *et al.*, 2004). Logue and colleagues (Logue *et al.*, 2012) found TP alone to be significantly correlated with OTU richness in a survey of Swedish lakes. Our results suggest that eutrophication leads primarily to an increase in the overall number of OTUs, but also to dominance of those with opportunistic lifestyles.

Multivariate trends

Ordination of the Bray-Curtis dissimilarity matrix of the 0.22 µm samples revealed clear grouping, which we designated into four significant clusters (Fig. 1.4A, permANOVA $p<0.001$). The first cluster (eutrophic) comprised samples from shallow and warm areas, specifically inner Chivyrkuy and Proval bays. The second cluster (transition) included samples from the Chivyrkuy Bay to open Lake Baikal gradient, the main tributary Selenga River plume, and the two samples collected in Barguzin Bay – the deepest and most open of the three examined bays. The third cluster (open) contained the bulk of our samples collected in open waters that largely separated in ordination space along a depth gradient. The last cluster (deep) comprised the three deep water samples from 300 and 500 m. *Post hoc* pairwise comparisons of cluster centroids revealed significant differences between each cluster pair (FDR corrected $p<0.05$). Thus, the clusters broadly captured the different habitats of Lake Baikal,

separated by depth and the transitions between open lake waters and bays. Given the confounding effects of abiotic forcing and spatial autocorrelation, we used reciprocal causal modeling (Cushman and Landguth, 2010; Cushman *et al.*, 2013) to test whether selection by abiotic environment or spatial dispersal/distance better explained community composition trends in mixed layer.

Environmental conditions, and not the distance, played a dominant role in structuring free-living mixed layer communities. Reciprocal causal effects models test the hypotheses for significant correlation between the community dissimilarity matrix and each of the geographic distance and environmental distance matrices, while controlling for the other. We found that, when controlling for environment, geographic distance had no correlation with community dissimilarity. However, when controlling for geographic distance, community structure did show a significant correlation with the environment (R^2=0.33, $p<1x10^{-6}$). Our results are consistent with the majority of other studies in freshwater (Logue and Lindström, 2010; Lindström *et al.*, 2006) and marine systems (Sjöstedt *et al.*, 2014), which usually note stronger effects of environment on species composition, compared with dispersal.

Temperature, depth, and total dissolved nitrogen (TDN) were the strongest environmental covariates with multivariate dissimilarity trends among free-living communities (Fig. 1.4). The temperature vector (Fig 4A, R^2=0.71, $p=1x10^{-4}$) indicated the direction of greatest temperature variability along the open waters to bays gradient. We also confirmed presence of the depth covariate among samples collected in the open waters (R^2=0.52, $p=1x10^{-4}$). Interestingly, open water community dissimilarities

also revealed correlation with TDN in approximately the same direction as depth. Because depth can confound the effects of various environmental factors, we further considered an ordination of samples just from the ML (Fig. 1.4B). Among the ML samples, temperature was still the strongest predictor of community dissimilarity (R^2=0.56, p=2x10^{-4}), and TDS was also significant (R^2=0.36, p=0.012). Notably, these were also the only two model-averaged important predictors of OTU richness and ENS in ML (Fig. 1.2A, B). However, TN and TDP were also significant predictors of community dissimilarity in ML (Fig. 1.4B), although they did not show an association with alpha diversity metrics (Fig. 1.2A,B,C). As expected, depth was a not a significant predictor of community dissimilarity in the ML. Lastly, we investigated whether the same environmental variables that predicted differences in microbial community structure also accounted for major abiotic differences between sampled sites. For this, we constructed a PCA of ML samples using all measured environmental factors (Fig. 1.4C). The first two principal components captured 89.8% of the variation. TDS and light had the highest scores of all environmental variables. Furthermore, they were almost parallel to PC1 and PC2, respectively, suggesting that the two were responsible for explaining most of the measured abiotic differences between sampled ML sites. While TDS was a significant environmental covariate in ordination of biotic community structure, light was not (Fig. 1.4B). Indeed, light was also not a significant predictor of biotic alpha diversity metrics (Fig 2). In an opposing example, TDP was the least important predictor of abiotic differences in ML (Fig. 1.4C); however, it was the second most significant predictor of biotic dissimilarity in ML (Fig. 1.4C). These features highlight that changes

in microbial community structure were not correlated with simply the most variable environmental covariates, and that significant associations of particular environmental factors warrant attention in further studies.

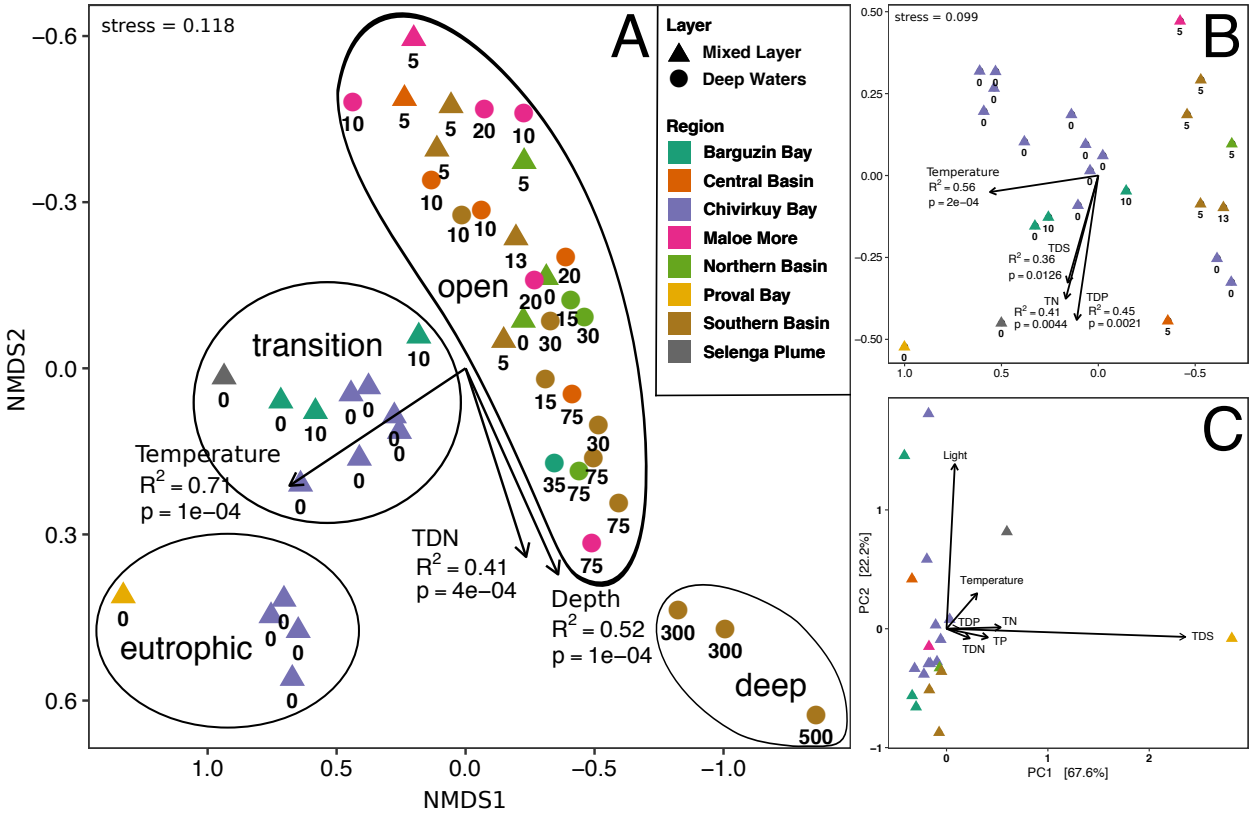


Figure 1.4: Ordination of the Bray-Curtis dissimilarity matrix for all 0.22 µm fraction samples (A) and the mixed layer only samples (B). (C) Principal component analysis of the mixed layer samples in environmental space using all measured covariates. In all panels, colors reflect major recognized regions of Lake Baikal (15, 22). In (A) and (B), numbers below bubbles indicate sample depth (m), and arrows show correlation of environmental covariates with the layout of sample points in ordination space. Each arrow length = R^2; p-values were obtained using permutations. In (A), ellipses were drawn to aid visualization.

OTU co-occurrence networks

To gain more insight into bacterial community structure in Lake Baikal and its dependence on the abiotic drivers and to identify OTUs that tend to co-occur, we

constructed an OTU co-occurrence network. Co-occurrence networks enable data-guided identification of OTU assemblages and correlation of their abundance with continuous environmental variables. In contrast to ANOVA-based permutation procedures, networks do not carry a user bias of arbitrarily defining sample categories, which, even if chosen wisely, result in information loss.

Our network captured the bulk of Baikal's OTUs. We constructed a co-occurrence network for OTUs present in at least 80% of samples (105 nodes and 819 edges, Fig. 1.5; see Supplementary Methods). Importantly, although 105 OTUs appeared to be a large reduction from the total >38,000 detected in Baikal, the network OTUs amounted to over >81% of cumulative relative abundance of every sample in >85% of samples. For the remainder of the samples, coverage averaged 63% ± 17%. Environments that were better observed had more OTUs in the networks. Lower coverage environments included the notable outliers collected at 500 m, 300 m in Southern Basin and at the surface of the eutrophic Proval Bay and Selenga River plume. Thus, with network analyses we used non-categorized continuous data to directly capture the vast majority of microbial community and environmental covariate data, revealing the dominant groups of co-occurring OTUs and their association with major environmental drivers. The resulting network exhibited the "small world" properties. Small world networks are characterized by high connectivity between neighboring nodes and low connectivity between distant nodes (Watts and Strogatz, 1998), creating clusters or modules of consistently co-occurring OTUs. Within each module, OTUs with many connections (central nodes) are thought to reflect the processes that bring together module

members. Small world networks may be more robust to perturbations but changes to the abundances of the central, well-connected OTUs may disproportionately affect the whole modules (Comte *et al.*, 2015).

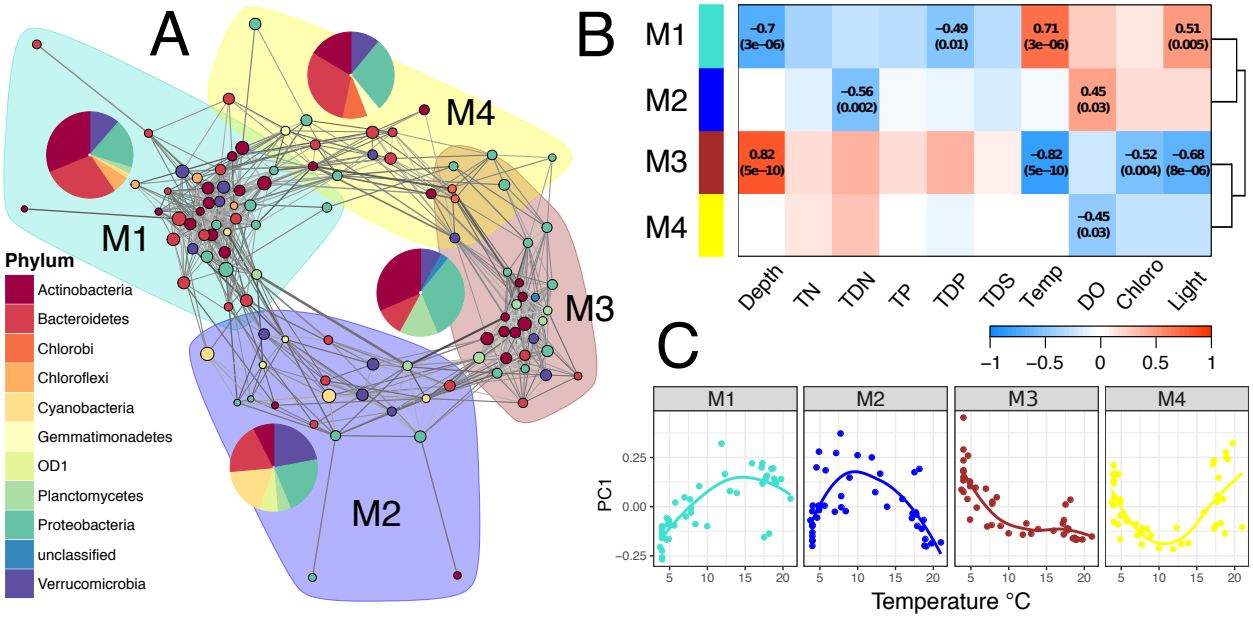


Figure 1.5: (A) Co-occurrence network of OTUs across all samples. Edge grayscale hue reflects pairwise correlation strength (darker edges show stronger correlations), and thickness indicates edge betweenness score. Modules were defined using maximum modularity optimization. For each module, the first principal component (eigenvector, PC1) of just that module's constituent OTUs abundance matrix was used to summarize the dominant abundance trends across sampled sites. (B) PC1 trends are summarized in a heatmap, where numbers are Spearman correlation coefficients with p-values in parentheses below. Empty cells indicate non-significant results. (C) Example PC1 trends are shown with respect to temperature.

Using the optimal modularity approach (Brandes *et al.*, 2008; Csárdi and Nepusz, 2006), we identified four modules of OTUs that tend to co-occur across sampled sites (Fig. 1.5A). Strong cohesion (see Supplementary Methods) within Module 1 (M1; clustering coefficient, CC_{M1}=0.68; Table S1.8) and Module 3 (M3; CC_{M3}=0.82) suggested common drivers for the consistently co-occurring OTUs. In contrast, low

cohesion within Module 2 (M2; CC_{M2}=0.54) and Module 4 (M4; CC_{M4}=0.58) pointed to a looser internal structure.

Module eigenvectors (Langfelder and Horvath, 2008) or M1 and M3 revealed opposing monotonic relationships with respect to environmental covariates (detailed introduction in Supplementary Methods). M1 showed a negative monotonic relationship with depth (Spearman ρ=-0.70, $p=3x10^{-6}$) and a positive relationship with temperature (ρ=0.71, $p=3x10^{-6}$). In contrast, M3 had a strong positive relationship with depth (ρ=0.82, $p=5x10^{-10}$) and a negative relationship with temperature (ρ=-0.82, $p=5x10^{-10}$). As expected, the direction of the opposing trends for the two clusters was reversed for light, owing to its inverse relationship with depth (Fig. 1.5B). Thus, we found that a large fraction of Baikal's planktonic prokaryotes can be classified into one of the two cohesive clusters: mixed layer "warm" M1 and deeper waters "cold" M3 cluster.

Modules M2 and M4 (Fig. 1.5a) showed inverse opposing unimodal (non-monotonic) trends with depth, temperature and light (Fig. 1.5B, C). Weaker clustering within these modules suggested less cohesion, possibly due to the inclusion of taxa with weaker habitat preferences or terrestrial or riverine dispersal. M2 showed highest cumulative relative abundance at an intermediate temperature of about 10°C, with a strongly positive relationship with oxygen. In contrast, M4 showed a negative relationship with oxygen, abundance peaks at low and high temperature extremes and minimum abundance at approximately 11°C. Specifically, M4 was high in abundance in the immediate surface (0 m) and deep water, but with a sharp drop in abundance at 5 m depth.

Phylogenetic signal in modules

Membership in the modules had a weakly significant phylogenetic signal (Fig. S1.17). This, combined with different habitat preferences among the modules, suggested different taxonomic groups displayed preferential environmental associations, thus supporting the importance of niche differences and phylogenetic niche conservatism.

M1 and M4 had a high percentage of *Bacteroidetes*, known as opportunistic degraders of high molecular weight organic matter, such as proteins and carbohydrates, with genomes containing numerous carbohydrate-active enzymes covering a large spectrum of substrates from plant, algal and animal origin (Thomas *et al.*, 2011). In aquatic systems, *Bacteroidetes* have been noted to follow pulses of organic matter inputs and cyanobacterial blooms (Newton *et al.*, 2011). This further casts M1 as a warm water ML module and M4 as the loose product of sediment and terrestrial input.

M1 is the only module to contain members (three OTUs) of the Chloroflexi phylum. Although Chloroflexi have been found in diverse environments, including oxygenated (its type genus) and anoxic (Overmann, 2008) hot springs, the CL500-11 clade – a subclass of the deep-ocean SAR202 clade – has been suggested as characteristic of oxygenated hypolimnia in deep lakes, including Crater Lake (Urbach *et al.*, 2007), Lake Biwa (Okazaki *et al.*, 2013) and the Laurentian Great Lakes (Denef *et al.*, 2015). However, the three Chloroflexi OTUs detected in our Lake Baikal network were from the Chloroflexi class (two OTUs) and Roseiflexales order (one OTU), and, as part of M1, appear to prefer warmer and shallower environments in the lake.

The cold deep-water cluster M3 had the highest relative abundance of OTUs in unclassified phyla. While the module also contained characteristically terrestrial Gammaproteobacteria *Pseudomonas* and *Acinetobacter*, their low connectivity suggested they are unlikely to play a central role in M3 processes.

M2 is the only module with Cyanobacteria; the module's three most abundant OTUs – indeed, second and fourth most abundant in the entire >38,000 OTU dataset – classified as *Synechococcus*. Cumulative relative abundance of *Synechococcus* OTUs peaked at ~15 m, reflecting the approximate location of deep chlorophyll maximum at stratified stations. To further understand internal structure of the four modules, we focused on central OTUs that played important roles in creating the module structure.

Central otus in each module reflect module ecology

The meaning of a node (in our case OTU) position in the network is the subject of much discussion. A central OTU has a significant positive correlation with a large number of that cluster's OTUs. We offer one abiotic and one biotic interpretation, resulting from two explanations for the presence of network clusters in the first place. First, clusters of OTUs that correlate with important environmental covariates, like temperature, could reflect abiotic habitat filtering. In this case, OTUs in modules with higher clustering coefficients would more consistently reflect environmental conditions associated with that module. For example, every OTU in M3, which has the highest clustering coefficient (Table S1.8), shows clear negative relationship with temperature (Fig. 1.5C). A second explanation requires an assumption that co-occurrences reflect

biotic species interactions. Then, a central OTU could be hypothesized to directly increase the abundance of other OTUs, e.g., by synthesizing or otherwise making available a limiting resource. It is important to point out that the two explanations are not mutually exclusive and could both be parts of ecological mechanisms that give rise to observed clusters of co-occurring microbial species.

The most globally central (most connected) OTUs are likely to belong to most populous module with the greatest number of connections (M1). The most central OTU in the network is OTU137, classified as an autotrophic methylotroph LD19, which was previously reported as a summer-fall bloomer in Lakes Michigan and Muskegon, MI (Fujimoto *et al.*, 2016). Although not very abundant in our dataset, its high centrality suggests it is indicative of an ecological process that is at least partially responsible for supporting other OTUs in the ML.

OTU001 (*Limnohabitans*) – also in M1 – has a high overall abundance and the network position that straddles the balance between centrality in M1 and connectedness to M2. Its high relative abundance and ubiquity across the lake suggest important roles in lake ecology. Connectedness to two modules could reflect this ubiquity and further indicate an influence of OTU001 on OTUs that occupy both niches. More broadly, *Limnohabitans* genus has many ecotypes. Its fine-scale phylogeny and functioning, shown experimentally with isolates (Salcher, 2014), revealed diverse lifestyles. In fact, three OTUs out of 105 in our network were classified as *Limnohabitans*, and one of the other OTUs (OTU104) was in the hypolimnetic M3. Separation of different *Limnohabitans* OTUs into M1 and M3 with opposing spatial occurrences and

environmental preferences suggests they are indeed ecotypes (Fuhrman, 2009) with distinct functions. Future studies can experimentally assess *Limnohabitans* ecotype responses to temperature and other environmental drivers.

The most connected OTU in the hypolimnetic M3 was Planctomycetes OTU022, classified into the Phycisphaerales order. Little is known about the ecological role of Planctomycetes in aquatic environments (Newton *et al.*, 2011). The order is known for its distinct visual appearance, and a recent report of a 3D reconstruction of Phycisphaerales cellular membrane revealed characteristic deep invaginations in the cellular envelope (Santarella-Mellwig *et al.*, 2013). It is possible the increased surface area of Phycisphaerales becomes useful in the oligotrophic hypolimnion of Baikal. In contrast to the most central but not abundant OTU in M1, OTU022 is also the most abundant in M3, which is the most tightly clustered module in the network (Table S1.8). These data suggest Phycisphaerales may substantially contribute to nutrient cycling below the thermocline with direct impacts for the rest of M3.

Functional redundancy

FR was calculated as the ratio of the number of OTUs in the sample that could be characterized using PICRUSt to the number of unique KEGG orthologs (KOs) in each sample, increased with temperature in the mixed layer (Fig. 1.6) and was highest in bays. This may be because littoral (nearshore) zones are expected to exhibit more variable environments, supporting co-existence of functionally redundant species.

In deep waters, functional redundancy increased with depth (Fig. 1.6). Our results warrant further in-depth analyses of microbial communities using techniques like isotope probing of phylogenetic microarrays, that track resource utilization, to reveal potential trade-offs between resource partitioning in unstable and streamlining in constant conditions. Functional redundancy, where more than one species perform a similar function, enables greater functional stability of ecosystems in the face of perturbations, including environmental change (Walker, 1992).

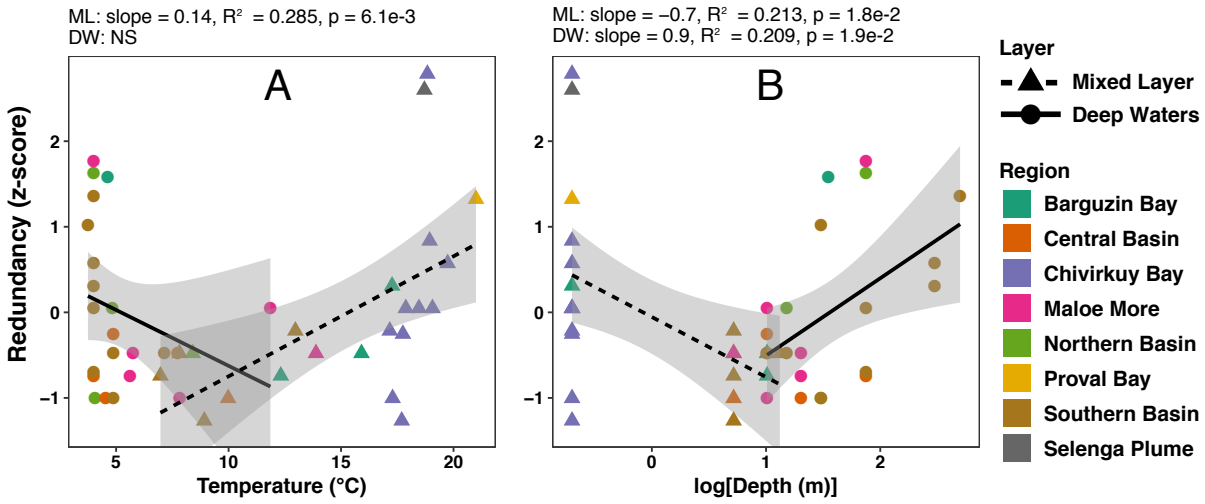


Figure 1.6: Functional redundancy (number of OTUs/number of KOs) across sampled temperature (A) and depth (B) gradients. The number of OTUs are the number of OTUs that had the PICRUSt-inferred genomes and KOs were determined using the PICRUSt tool.

Our first extensive survey of Lake Baikal's bacterioplankton revealed that temperature and nutrients are the major drivers of the microbial community composition, diversity and functional redundancy, so that the anthropogenic changes in these factors would likely significantly alter planktonic microbial communities. The OTU co-occurrence network analysis identified two major clusters, associated with the upper

mixed layer and deep waters, with a detectable phylogenetic signal in each cluster composition. The “small world” properties of the network suggest that the communities may be resilient to perturbations but the changes in abundances of the central OTUs may result in the network rearrangement. The lower functional redundancy in cold, low nutrient open water communities compared to warmer, higher nutrient waters may result in their higher vulnerability to changing environment.

APPENDIX

APPENDIX

SUPPLEMENTARY INFORMATION FOR CHAPTER ONE

Supplementary Materials and Methods

Sampling and environmental contextual data

The survey was guided by the recorded natural history of Baikal (Kozhova and Izmest'eva, 1998; Moore *et al.*, 2009) that divides the lake into eight distinct regions. We collected samples from 24 spatial locations, where 10 of those locations were sampled at various depths for a grand total of 46 samples across the lake (Fig. S1.1). The cruise took place on board the R.V. Treskov on August 3-17, 2013.

Temperature and dissolved oxygen profiles were measured with a YSI Instruments sonde (YSI, Inc., Mod4 Springs, OH, USA). Light profiles were measured using Walz model US- SQS/L model Li-185B light probe (Heinz Walz GmbH, Germany), connected to a Li-Cor Quantum photometer (LI-COR Biosciences, Lincoln, NE, USA).

Whole water was taken using a Van Dorn closing bottle (Wildco, Inc., Yulee, Florida, USA). From each sample, (a) 1L was filtered onto a GF/F glass fiber filter (GE Healthcare Bio-Sciences, Pittsburgh, PA, USA) for Chl a measurements and frozen at –20°C; (b) 5 L were sequentially filtered onto 3 µm and 0.22 µm mixed nitrocellulose acetate membranes (EMD Millipore, Billerica, MA, USA) and stored at –20°C in RNAlater (Life Technologies, Grand Island, NY, USA) to capture particle-attached and free-living fractions of microorganisms (Lincoln *et al.*, 2014); (c) 50 mL were frozen at –20°C for total nutrient analysis; and (e) 15 mL were filtered through a 0.22 µm mixed nitrocellulose acetate membrane (EMD Millipore) and frozen at –20°C for dissolved nutrient analysis.

Samples were transported to the USA in insulated containers cooled with liquid nitrogen. In the USA, samples for molecular analyses were stored at –80°C and samples for nutrient analyses were stored at –20°C.

Temperature and light profile modeling

Temperature values at collection sites were estimated from YSI instruments sonde profiles. YSI sonde profiled temperature at each station down to between 40 m to 50 m. Five temperature values per second were recorded on the downcast (lowering approximately 1 m per second) and used for modeling. For each station, high-order polynomial functions were fit to the data, and point estimates were used to infer temperature values at exact depths used for water sample collection (Fig. S1.8). For sites samples collected at 75 m, 300 m and 500 m, temperature was assigned a value of 4°C, based on literature values (Kozhov, 1963, 41).

Light values at collection sites were estimated from profiles collected using the Li-Cor Quantum photometer (LI-COR Biosciences, Lincoln, NE, USA). Profiles were measured down to site bottom or maximum 20 m depth. The light extinction function $I_D = I_0 e^{-kD}$, where I_D = light intensity at depth, I_0 = light intensity at the surface, k = extinction coefficient (0.035 for pure water) and D = depth, was fit to the to the data to estimate light values at collection sites (Fig. S1.9).

Nutrient Measurements

Samples for nutrient analysis were thawed and digested with potassium persulfate at 120°C for 30 min. Total and dissolved nitrogen were measured colorimetrically on a Shimadzu UV-240-PC spectrophotometer (Shimadzu Scientific Instruments, Columbia, Maryland, USA) at 224 nm using the 2nd derivative method (Crumpton *et al.*, 1992). Total and dissolved phosphorus was measured using the orthophosphate method on a Lachat Instruments Quick Chem 8500 autoanalyzer (Lachat Instruments, Loveland, Colorado, USA). Chl *a* filters were extracted with ethanol, and Chl *a* was determined fluorimetrically (Welschmeyer, 1994).

DNA extraction and amplicon sequencing

Genomic DNA was extracted using the Mo-Bio PowerSoil Kit (Mo-Bio Laboratories, Carlsbad, CA), following manufacturer's protocol. Then, the V4 region of the 16S rRNA gene was sequenced using dual-index primers as described previously (Kozich *et al.*, 2013). After PCR amplification, the products were normalized and pooled. The pool was loaded on an Illumina MiSeq v2 flow cell and sequenced with a standard 500 cycle reagent kit for paired-end 250 bp reads (PE250). Base calls were done with Real Time Analysis software v1.18.54. Output of RTA was demultiplexed and converted to FastQ with Illumina Bcl2Fastq v1.8.4.

Amplicon sequence processing

The FastQ output files were processed using mothur, following general MiSeq protocol and the options below (Kozich *et al.*, 2013; Schloss and Westcott). Sequences were aligned to a full SILVA v.119 database. Chimeras were removed with UCHIME in mothur environment. The remaining sequences were classified with a naïve Bayesian RDP classifier (Wang *et al.*, 2007) and the Greengenes (August 2013 release) database. Sequences classified as Mitochondria, Chloroplasts and Eukaryota were removed, resulting in 197,738 unique sequences that were clustered into OTUs at 97% similarity. Consensus taxonomy for each OTU was determined following mothur protocol. Coverage for sequenced samples varied between 24348 and 76838 (Fig. S1.11, left), and was rarefied to the lowest coverage sample i.e. 24348 reads (Fig. S1.11, right).

Multiple regression

We used the brute force approach to multiple linear regression (exhaustive search, followed by model selection) to determine the relationship of community richness, diversity, and evenness with measured environmental variables. Richness was taken as the number of OTUs in each sample. This comparison was possible because our sampling effort was rarefied in mothur (see above). The effective number of species (ENS), represented diversity and was calculated by raising the natural number *e* to the power equal to the Shannon diversity index H' (Jost, 2006; Leinster and Cobbold, 2012). It was important to use ENS because, unlike diversity indices, such as Shannon and Simpson, ENS has a linear response to change in community diversity – a necessary

property of a response variable in quantitative modeling (Jost, 2006; Leinster and Cobbold, 2012). Statistical modeling and model selection were performed with the glmulti package in R (Calcagno and Mazancourt, 2010).

Multivariate statistical analyses

Distance matrix. Ordination of the OTU abundance matrix was performed in the R environment (version 3.2.2) using vegan (Oksanen *et al.*, 2016) and phyloseq (McMurdie and Holmes, 2013) packages. First, we compared different techniques for calculating community dissimilarities. We used the Mantel test to reveal correlations between the Bray-Curtis, Jaccard, unweighted Unifrac, and weighted Unifrac distance matrices. Results showed greatest differences between presence-absence distance metrics and abundance-weighted distance metrics. However, within each of the two categories, the metrics were approximately interchangeable. Thus, we decided to use the Bray-Curtis metric for better compatibility with other studies. The high correlation of weighted Unifrac matrix indicated that phylogenetic information did not largely affect distance matrix results.

Test for dispersion (geographic distance) was done with reciprocal causal modeling using partial Mantel tests. The use of reciprocal partial mantel tests for reciprocal causal modeling is advocated in the literature by Cushman (Cushman and Landguth, 2010; Cushman *et al.*, 2013), where he criticized simple Mantel tests for their Type I error rates, and proposed reciprocal partial mantel as a solution (Cushman and Landguth, 2010), later expanding to a more sophisticated "relative support" technique - also based on

partial Mantel tests (Cushman *et al.*, 2013). Rousset criticized all Mantel-based options in favor of mixed effects regression methods (Guillot and Rousset, 2013) because they are not robust when there is autocorrelation in the environment matrix. In a summary review, Cushman conceded that LME is indeed the best method for decoupling the effect of spatial dispersal and selection; however, that doesn't mean reciprocal partial mantel tests are inappropriate, especially if autocorrelation in data is weak (Shirk *et al.*, 2017). Cushman pointed out that reciprocal partial mantel tests performed almost as well as mixed linear models if conclusions were based on R^2 effect sizes, and not just p-values. Fortunately, *our data does not have significant spatial autocorrelation with respect to environment* (Moran's I > 0.05). Thus, we proceeded using Cushman's method for its simplicity.

Geographic distance between sampled sites was calculated with the geosphere R package using the Vincenty Ellipsoid model. Mantel test was performed using the vegan R package. Because the geographic distance matrix was not normally distributed (Fig. S1.20), we used the rank-based Kendall test statistic option in the Mantel tests.

Ordination was done in the phyloseq package (McMurdie and Holmes, 2013) and modified for visualization with a custom R script. NMDS plots are freely rotatable and scalable, and Fig. 1.4 panels A and B were thus adjusted for greater visual clarity.

Correlation with environmental variables was done with the envfit function in the vegan package, which correlates continuous environmental values with separation of points in ordination space. Significance was calculated by bootstrapping using 9999 permutations.

Network construction and analyses

Co-occurrence matrix and network construction. The OTU co-occurrence matrix was calculated for OTUs that were present in 80% of all samples (109 OTUs) with sparCC software (Friedman and Alm, 2012), as recommended in best practices for co-occurrence network construction by Berry and Widder (Berry and Widder, 2014). Significance values for correlations were bootstrapped, as described by sparCC authors, and then corrected for multiple testing using FDR manually in R with a custom script. Positive co-occurrences with adjusted $p < 10^{-5}$ were used for downstream analyses. The network was constructed, displayed and analyzed using the iGraph package (Csárdi and Nepusz, 2006) in the R environment. Two pairs of OTUs that were only connected to each other but not the rest of the network were removed, resulting in 105 OTUs in all subsequent analyses.

Comparison with simulated networks. We compared basic network statistics with null distributions of two types of simulated networks. First, we created 10,000 random Erdös-Rényi networks, which were undirected, without loops and used the gnm model with the same number of nodes (105) and edges (814) as our Baikal network. We also created 10,000 small-world Watts-Strogatz networks with 105 nodes and replacement probability p=0.05. Average path length and the clustering coefficient (transitivity) of the Lake Baikal network were compared to distributions of simulated networks (Fig. S1.16). Two-tailed *p*-values for Baikal network statistics were calculated as the number of simulated observations greater than the absolute values of the Baikal network divided by the total number of simulations.

The network exhibited small world properties. Both overall CS (0.618) and the average path length (AP=2.62) were much higher than random Erdös-Rényi simulations (CC mean=0.150, p=0, n=10,000; AP mean=1.93, p=0, n=10,000; Fig. S1.16a) and higher than small-world Watz-Strogatz simulations (CC mean=0.514; p=0, n=10,000; AP mean=2.27, p=0, n=10,000; Fig. S1.16b). The basic network statistics are summarized in Table S1.2.

Small-world properties place greater topological importance on central nodes (nodes with many connections) and to a lesser extent bottlenecks in maintaining the network structure. Hubs can be defined as nodes with high eigenvector centrality, and bottlenecks as high betweenness nodes. Hubs are well-connected, in many cases to other well-connected nodes and are therefore considered topologically more central. In an ecological co-occurrence network, they have been hypothesized to mediate processes important to their neighbors (Fuhrman, 2009). For example, a central OTU may produce a common good limiting resource, like a vitamin.

Bottleneck nodes are positioned along a high number of shortest paths between pairs of other nodes. This is called high betweenness. Fig. 1.4 displays betweenness values for each edge as proportional to its width. In a small world network with few bridges between modules, bottlenecks are the only way one module can interact with another. Therefore, bottlenecks have been hypothesized to mediate feedbacks between ecologically meaningful OTU assemblages (Fuhrman, 2009). These features emphasize the strength of network methods to detect candidate keystone taxa not necessarily based on their large abundance but on how they affect other network players.

Community (module) detection. Network modules were identified using the optimal modularity algorithm (Brandes *et al.*, 2008) in the iGraph package. Optimal modularity calculates the arrangement and membership of clusters that gives the highest modularity score, as defined by Newman *et al* (Newman and Girvan, 2004) and adopted by the iGraph authors.

Environmental trends of modules. To summarize OTU abundance trends across sampled sites, we used the first principal components of OTU abundance matrices (eigenOTUs (Langfelder and Horvath, 2008)) for OTU members of each of the four modules. The resulting four eigenOTUs were used to assess correlation of modules with environmental factors, such as temperature, nutrients, oxygen, and chlorophyll levels. For example, module M1 has 39 member OTUs. They are together because, by definition, they have similar occurrence patterns across samples (see exact abundances in Fig. S1.18, S1.19). But how can we summarize all OTU abundances in one vector? One option is to sum their relative abundances in each sample and get per sample cumulative values. However, this option is heavily biased towards the more abundant OTUs. Indeed, the difference between the most and least abundant OTUs in M1 is approximately two orders of magnitude. A better option is to use the widely practiced principal component analysis (PCA), which weighs OTUs by their power to explain variation in other OTUs. In PCA, the first principal component (PC1) is the best summary of OTU abundances in each sample. Thus, we ran an independent PCA for OTU members of each of the four modules and used PC1 for each module as a summary of its abundance across samples. The PC1s were correlated with measured environmental variables (Fig 5b, c).

Phylogenetic signal calculation. We first constructed a phylogenetic tree of OTUs by computing pairwise DNA distances using the K80 model and constructing a tree with the bionj algorithm (Gascuel, 1997). Phylogenetic signal was detected using a modified phylo.signal.disc R script developed by Enrico Rezende (Universidad Autònoma de Barcelona). This function was used in several studies to reveal a phylogenetic signal in discrete traits (Nichols *et al.*, 2013; Montesinos-Navarro *et al.*, 2012; Bauer *et al.*, 2012; Moro *et al.*, 2015; Valiente-Banuet and Verdú, 2007; Verdú and Pausas, 2007). It works with a tree of network OTUs and treats module assignment of each OTU as a character state with random transitions. The strongest possible phylogenetic signal in the OTU tree would be indicated by *num* - 1 transitions, where *num* is the number of modules. To quantitatively assess the phylogenetic signal, the script first uses maximum parsimony to generate a null model by calculating the minimum number of character transitions needed to create permuted character states of the OTUs for 10,000 simulations. Then, the actual number of transitions in the tree of Baikal OTUs with known module membership is compared with a distribution of transitions obtained from the permuted simulations. A p-value for a phylogenetic signal is calculated from comparing the true number of transitions with a null distribution of transitions obtained from the permuted simulations (Fig. S1.16).

Gene content and functional redundancy calculations

Gene content estimation. Gene content per sample was estimated using the PICRUSt (Langille *et al.*, 2013) python package. PICRUSt functional inference works in three steps. It uses a reference tree(McDonald *et al.*, 2012), constructed from the entire Greengenes

database (DeSantis *et al.*, 2006) of 408,315 curated full-length 16s sequences, and a table of 6910 KEGG ortholog (KO) abundances for 2,590 known genomes that have identifiers in the Greengenes tree. First, PICRUSt prunes the full reference tree to known genomes and uses phylogenetic modeling to reconstruct ancestral states. The result is estimated copy numbers for each KO in each ancestor in the pruned tree. Second, PICRUSt infers gene content for the tips of the entire Greengenes reference tree. This prediction is generated by an average of the contents of extant and inferred ancestral genomes, both weighted exponentially by the reciprocal of phylogenetic distance, in line with previous research that reported an inverse exponential relationship between 16s phylogenetic distance and gene content conservation (Zaneveld *et al.*, 2010). In the last step, user-supplied OTUs are matched to identifiers in the reference tree, their abundances are normalized by predicted 16s copy numbers, and a table containing copy numbers of their KOs is created. Prediction of the sample-wide genome content (metagenome) is a simple addition of KO copy numbers for all OTUs in the sample, weighted by their relative abundances. Our preliminary comparison of the PICRUSt-inferred metagenomes with the sequenced metagenomes from a subset of our samples (Langille *et al.*, 2013) reveals a good agreement (Wilburn et al. in prep.), verifying the reliability of the PICRUSt approach.

PICRUSt offers a quality control metric NSTI (nearest sequenced taxon index) for the per-OTU accuracy of gene content prediction. For example, because OTUs are clustered at 97% sequence similarity, an OTU with NSTI < 0.03 is identical to a known sequenced genome. Since phylogenetic distance from known genomes is the best indicator of

predicted genome quality, we worked with PICRUSt results based on OTUs with NSTI < 0.15, as per PICRUSt authors recommendation (Langille *et al.*, 2013).

In Baikal dataset, OTUs compatible with the PICRUSt pipeline were identified within mothur using the make.biom command as described in the MiSeq SOP. Of the PICRUSt-compatible OTUs, 486 non-singleton OTUs with NSTI < 0.15 were used to predict metagenomic content of samples for redundancy analysis.

Functional redundancy was calculated by dividing the number of non-zero OTUs by the number of unique non-zero KOs in each sample. This calculation ensures that each KO present in the sample is counted only once and does not depend on the number of gene copies. With more OTUs contributing the same KOs, the ratio (functional redundancy) goes up.

Supplementary Results

Temperature, light and nutrient profiles

Most stations were thermally stratified, while others exhibited weaker stratification because of the storm at the end of the cruise (Fig. S1.8). At the surface, temperature ranged from 7.0°C in northernmost SB to 21°C in Proval Bay with a median at 17.4°C. In the open waters, the median temperature was 9.5°C, consistent with expected values (Kozhova and Izmest'eva, 1998).

Light profiles with models are shown in Fig. S1.9, top. Light extinction coefficients (Fig. S1.9, bottom) were greatest at the single station in the eutrophic Proval Bay, high in the

Selenga River plume and lowest in Maloe More. Central Baikal stations, sampled along the gradient from shallow bay to open lake conditions, showed variable values.

Nutrient profiles are shown in Fig. S1.10. Total nitrogen (TN) ranged from 3.0 µM in southern Maloe More Straight to 25.2 µM in Proval Bay with both an overall and open waters medians at 10.0 µM. TP ranged from 0.16 µM in northernmost South Baikal to 1.4 µM in Proval Bay with lakewide and open waters medians at 0.35 and 0.38 µM. Molar TN/TP ratios varied from 5.6 in southern Maloe More to 56.9 in southern Central Baikal; median TN/TP was 26.1. Chl a ranged from 0.66 µg L1 in mid-Chivirkuy Bay to 10.5 µg L^{-1} in Proval Bay with the median at 2.0 µg µg L^{-1}. Nutrients increase, but with mild trends at depths above 80 m (not significant) with a visible increase at the deeper stations at 300 and 500 m. While relationships with TN, TP, DN, DP, DS and chlorophyll were all significant using all samples, only TN and chlorophyll remained significant with two high nutrient outlier stations (Proval Bay and Selenga Shallow) removed. The remaining positive relationship illustrated an expected correlation between nutrient load and primary productivity.

Multiple regression

Trends on the particle-attached 3um fraction were similar to the free-living 0.22 um fraction (Fig. S1.12, S1.13)

Multivariate statistical analyses

Dissimilarity matrices based on phylogenetic distance metrics (unweighted and weighted unifrac) correlated very highly with phylogeny-free Jaccard and Bray-Curtis matrices (Fig. S1.14) and were concluded not to add significant additional information to modeling the multivariate community structure.

Ordination of the 3um fraction showed similar results to the 0.22um fraction (Fig. S2.16).

Supplementary Figures

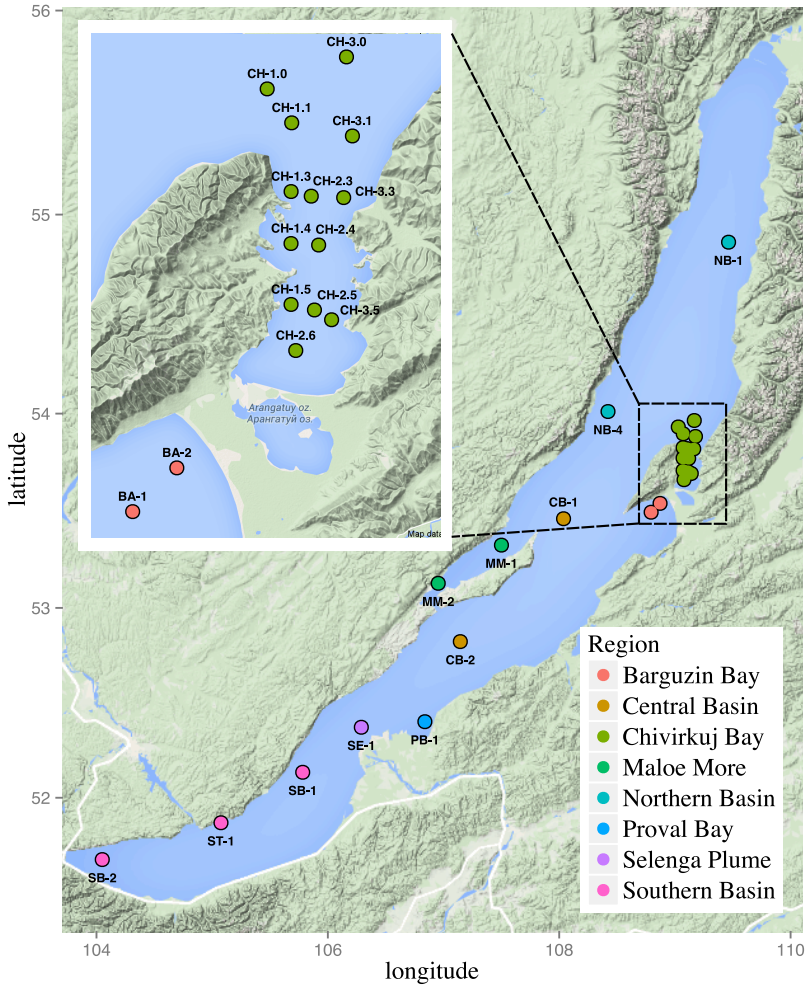


Figure S1.7: Sampling sites at Lake Baikal. Colors reflect major recognized regions of Lake Baikal (Kozhov, 1963; Kozhova and Izmest'eva, 1998).

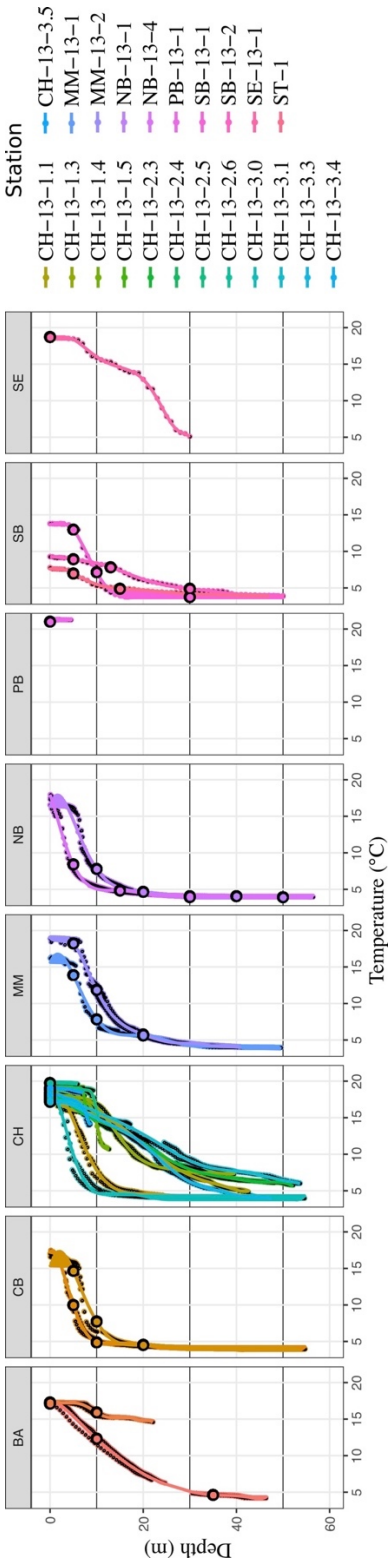


Figure S1.8: Temperature profiles and location of collected samples in the water column.

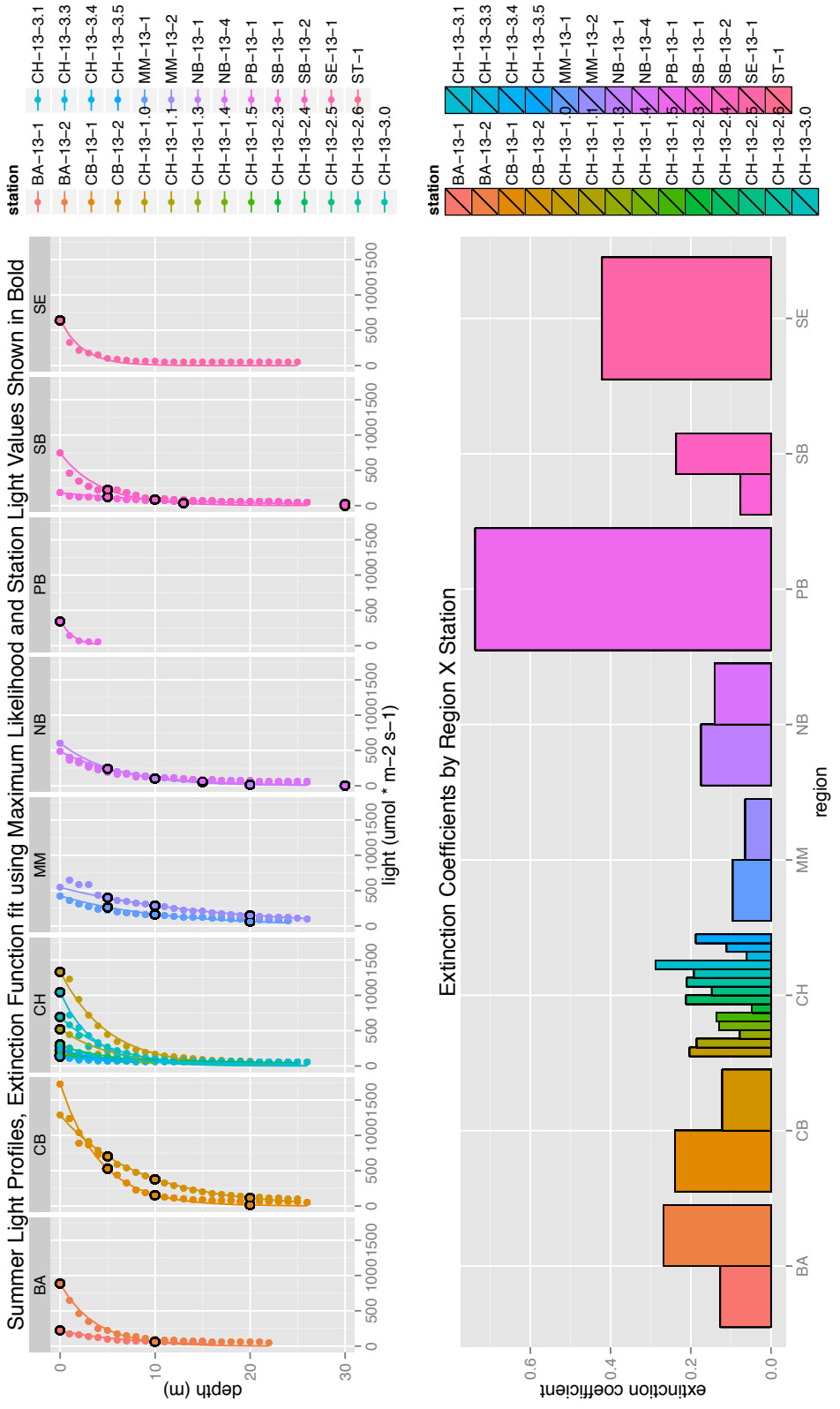


Figure S1.9: Light profiles, fitted models and extinction coefficients.

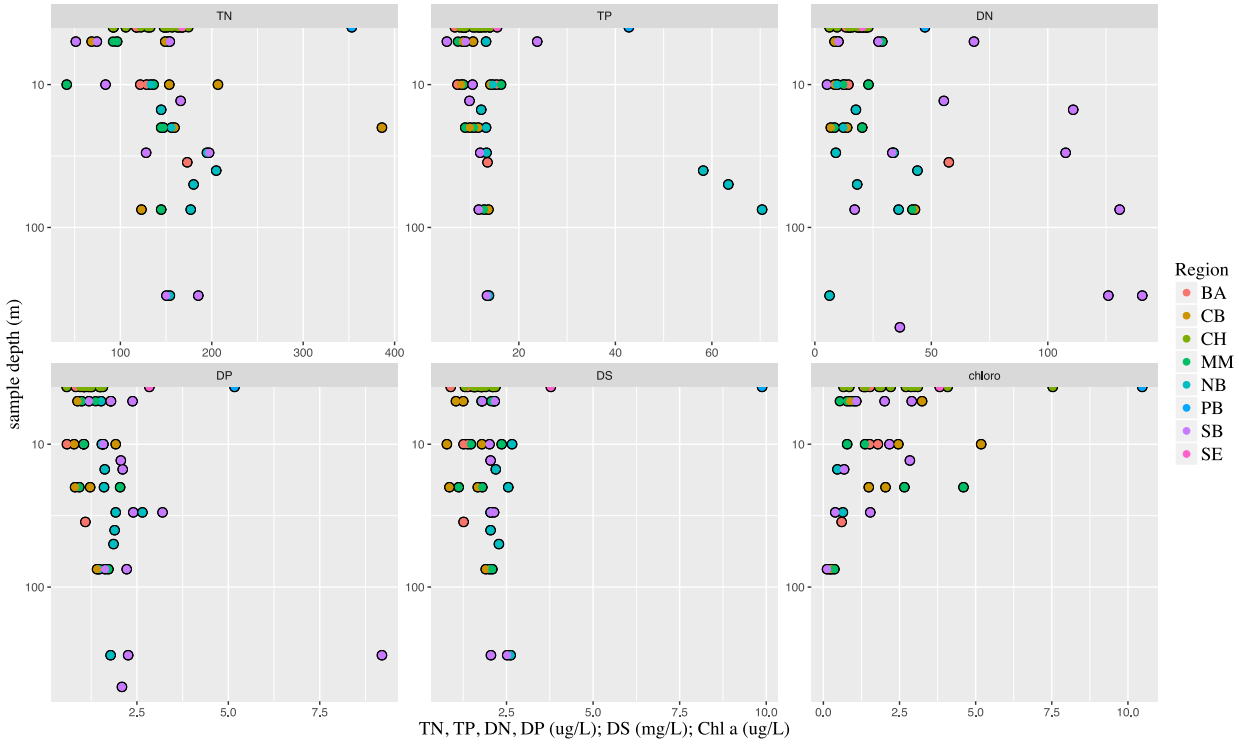


Figure S1.10: Nutrients show non-significant increase with depth.

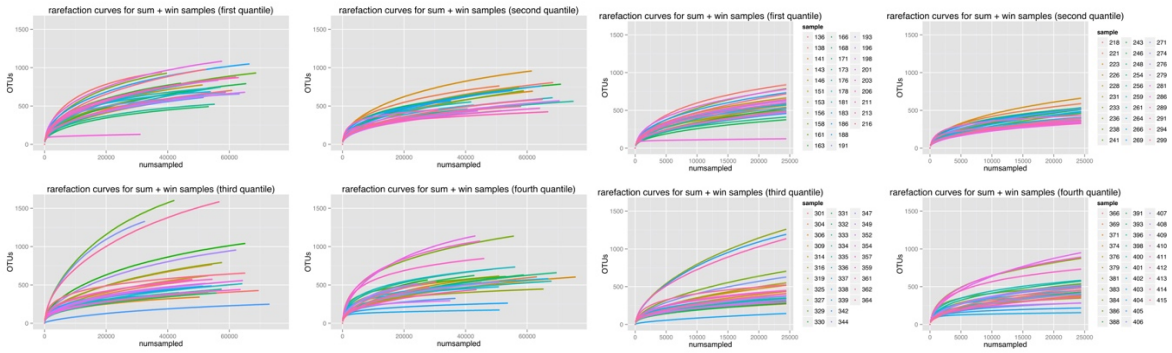


Figure S1.11: Rarefaction curves showing sampling depth per sample (left) and rarefied samples (right)

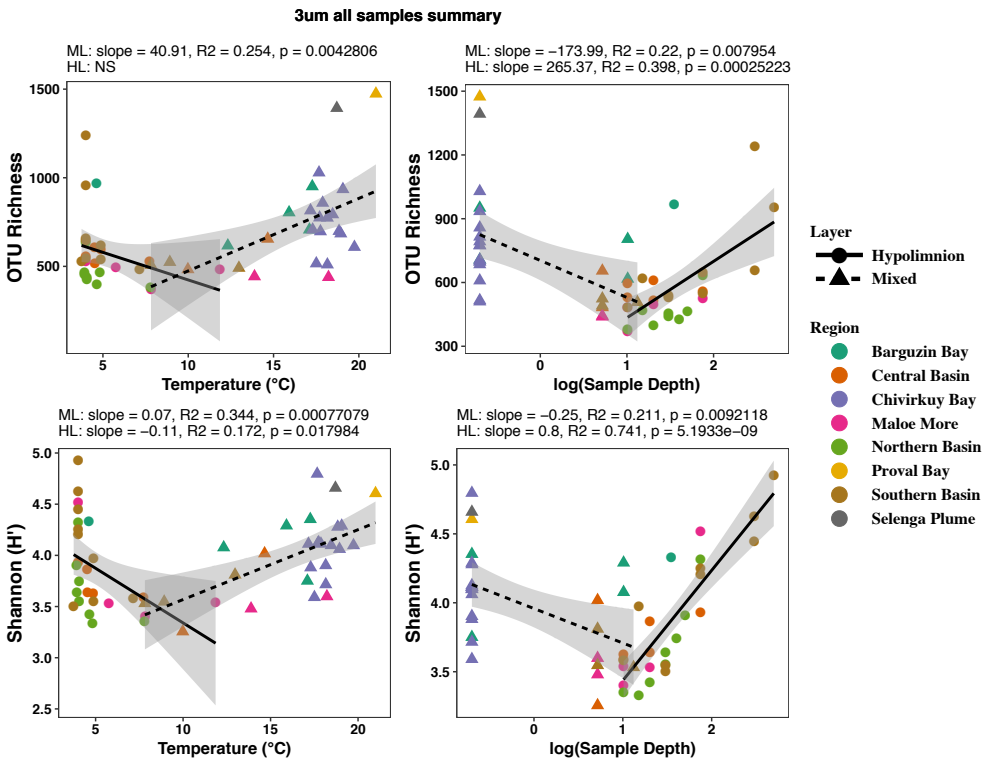


Figure S1.12: OTU richness and Shannon diversity on the 3 um fraction size, showing the same trends as the 0.22 um fraction.

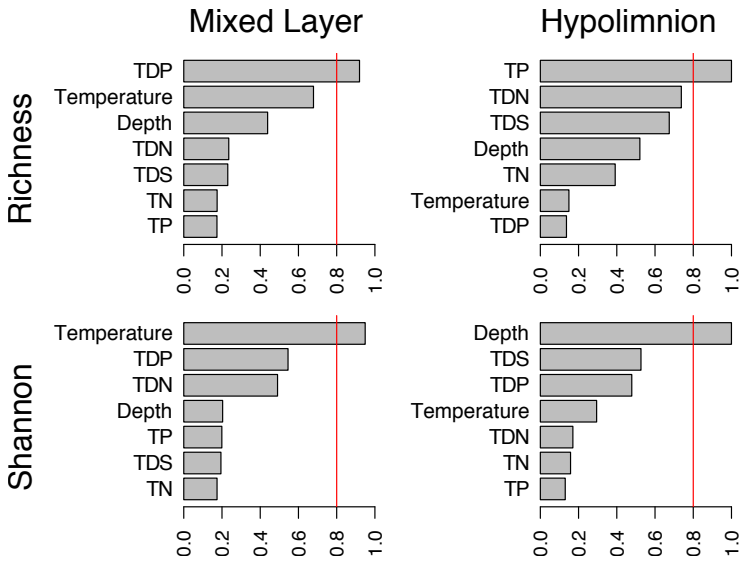


Figure S1.13: Model-averaged importance of environmental predictors for OTU richness (top) and Shannon diversity (bottom) in the mixed layer and the hypolimnion – on the 3 μm size fraction.

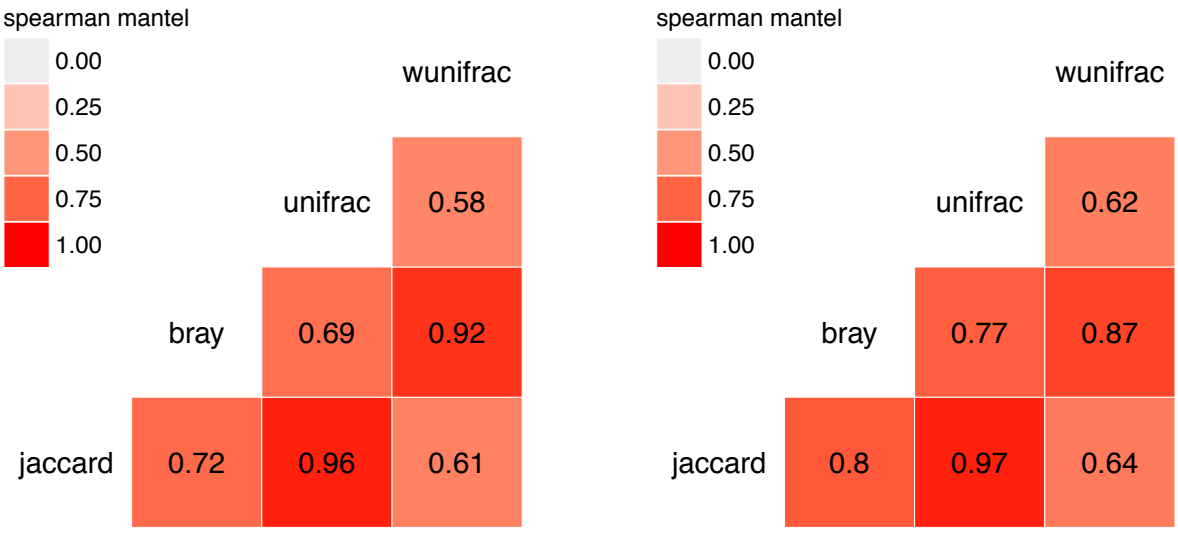


Figure S1.14: Pairwise Mantel test correlations between distance matrices based on phylogeny-free and phylogenetically-informed metrics for the 0.22 um (left) and 3um (right) size fractions. Unweighted distance calculators treat all OTUs equally, while weighted versions emphasize differences among the more abundant OTUs.

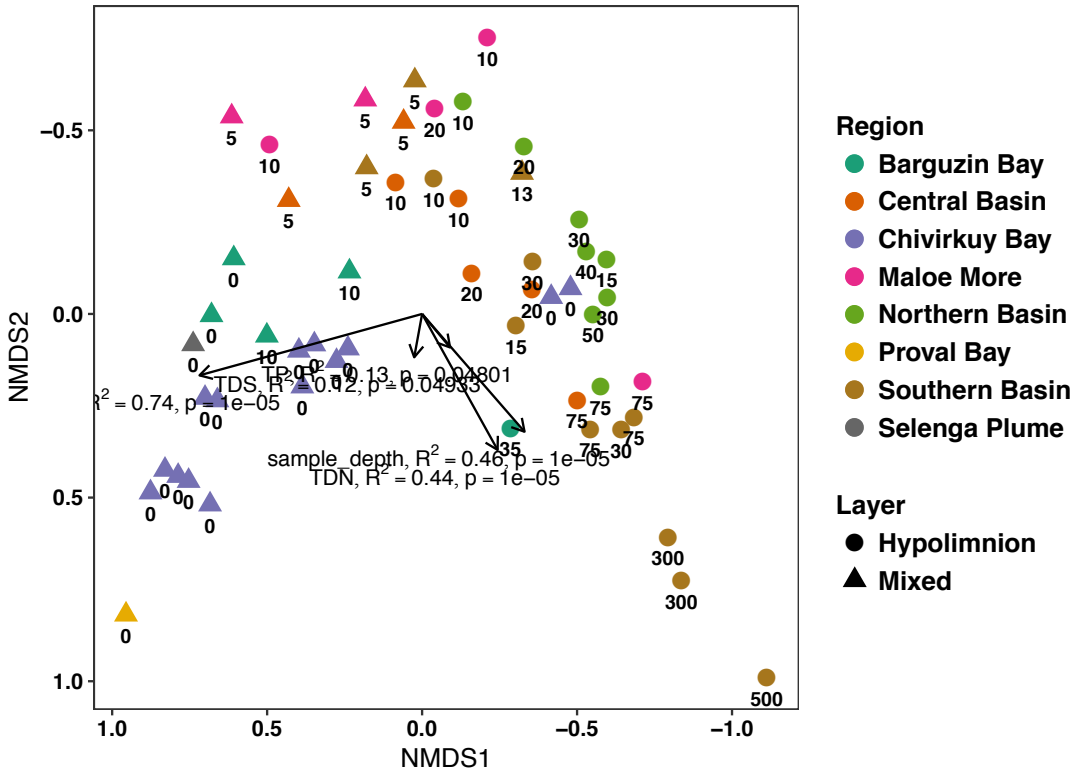
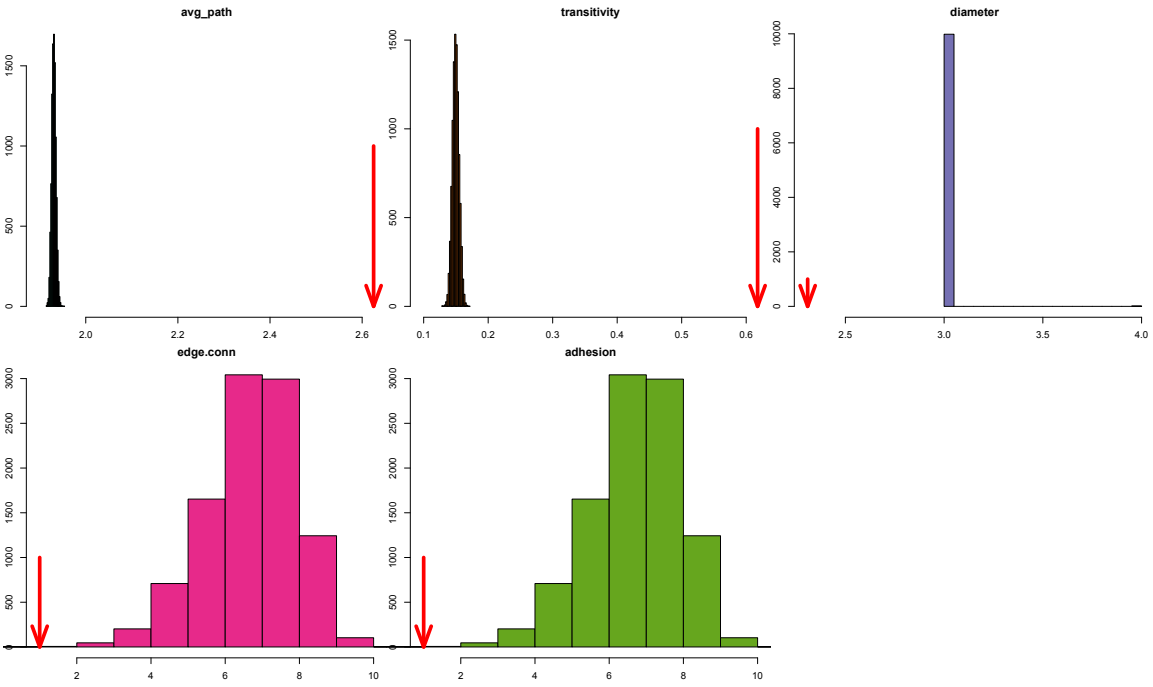


Figure S1.15: NMDS ordination of the particle-attached 3um fraction



B

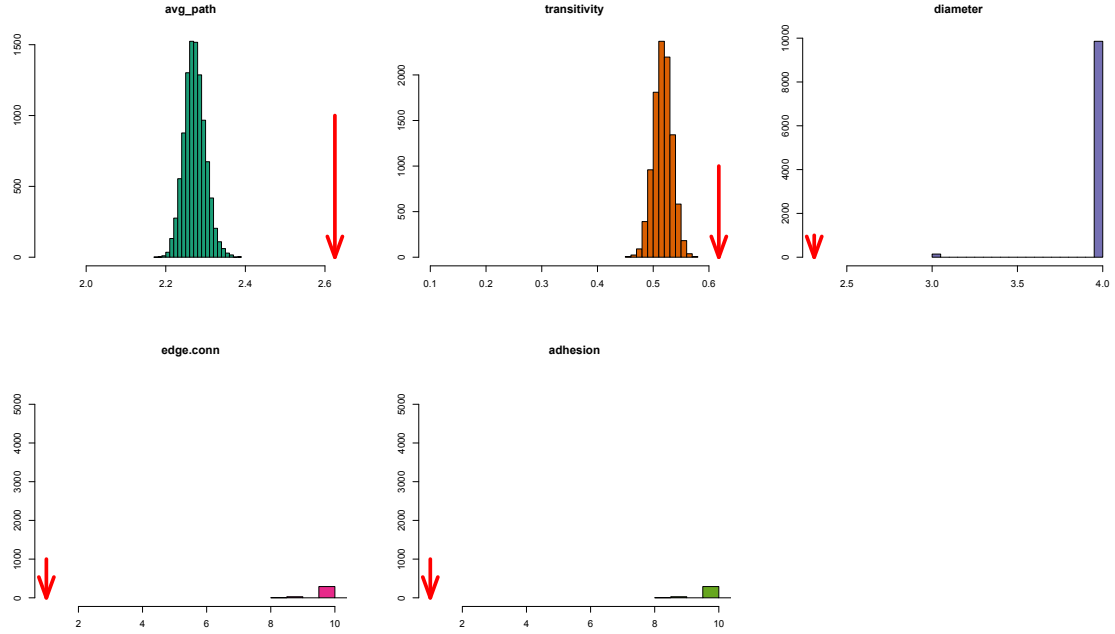


Figure S1.16: Network statistics comparison with simulated Erdos-Renyi (A) and Watts-Strogatz (B), run for 10,000 simulations each. Red arrows indicate statistics for Baikal networks.

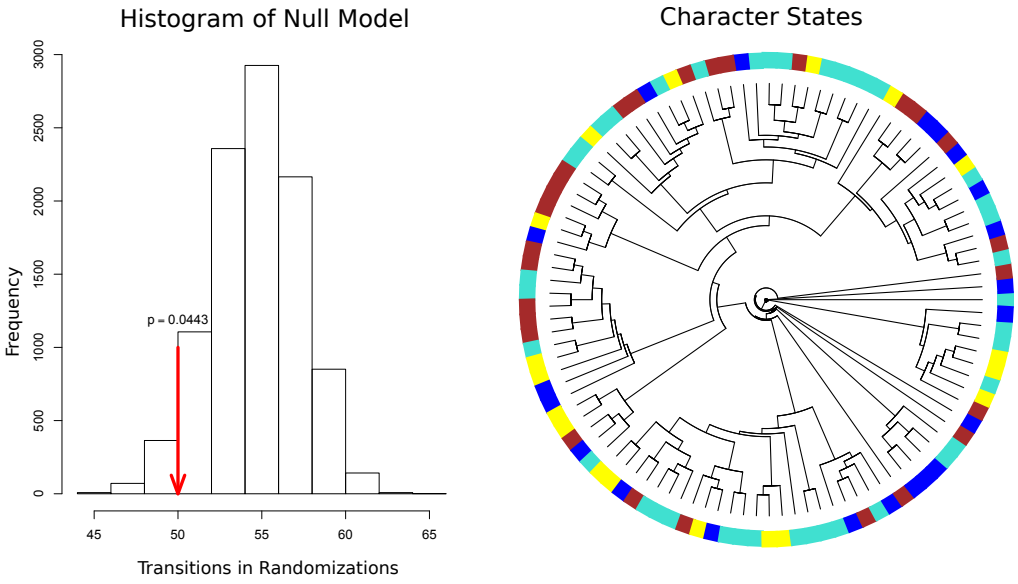


Figure S1.17: Phylogenetic signal for co-occurrence network modules as discrete character states for individual OTUs.

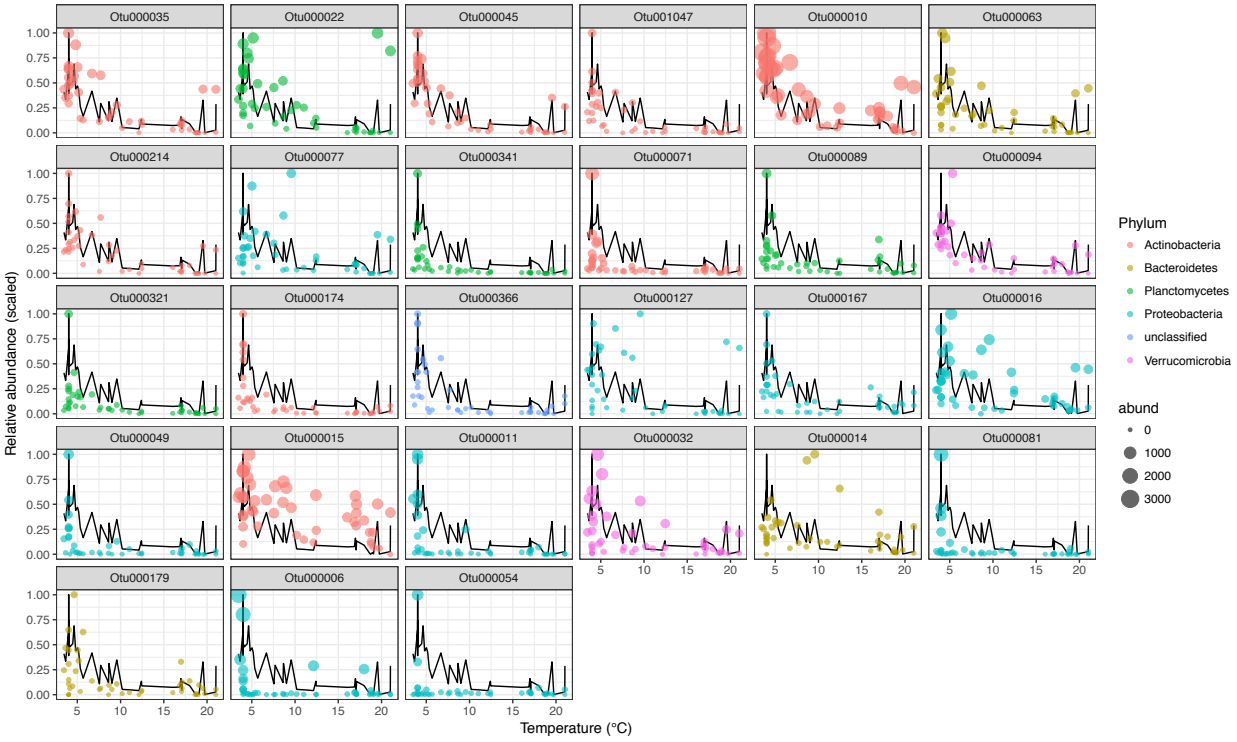


Figure S1.18: OTU abundance (scaled to 0-1 range) in the **M1** (ML module) are shown in bubbles. The PC1 trend for M1 (scaled to 0-1 range) is shown as a black line in each panel. OTU panels are arranged in order of decreasing centrality (connectedness). Bubble sizes indicate actual relative abundance values of the OTUs to communicate which OTUs were generally more or less abundant in Baikal.

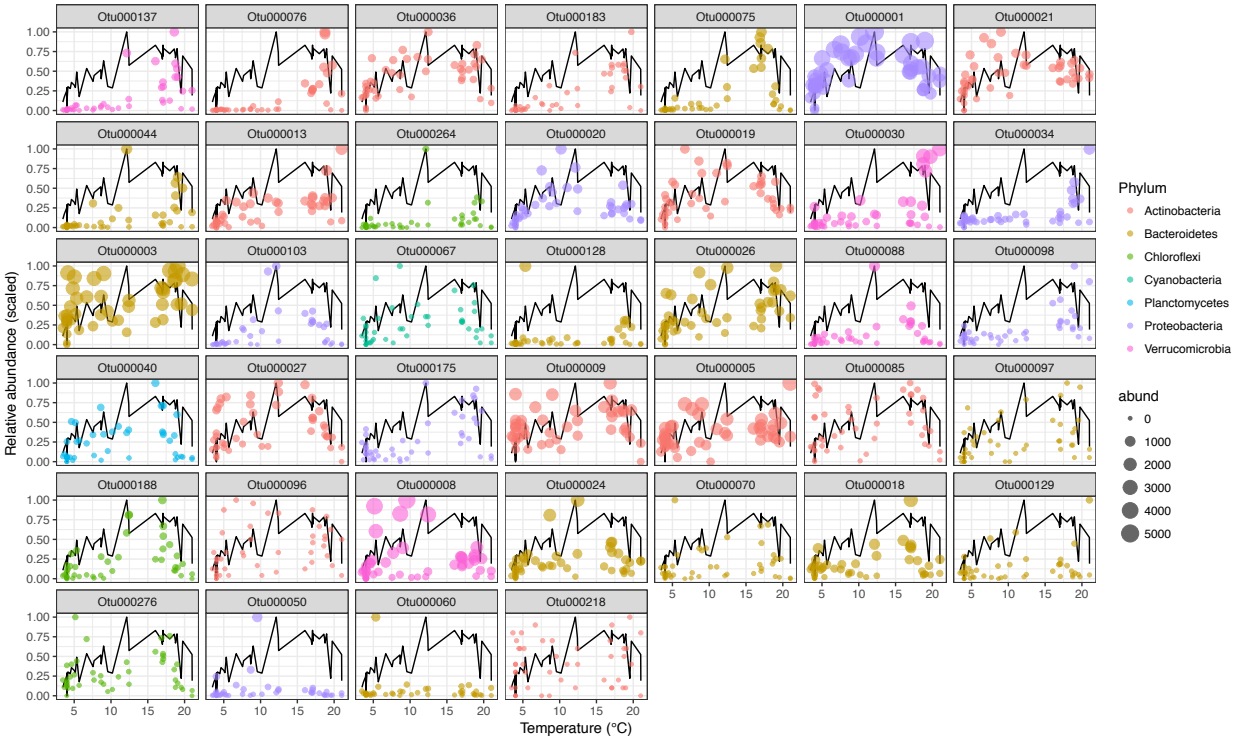


Figure S1.19: OTU abundance (scaled to 0-1 range) in the **M3** (DW module) are shown in bubbles. The PC1 trend for M1 (scaled to 0-1 range) is shown as a black line in each panel. OTU panels are arranged in order of decreasing centrality (connectedness). Bubble sizes indicate actual relative abundance values of the OTUs to communicate which OTUs were generally more or less abundant in Baikal.

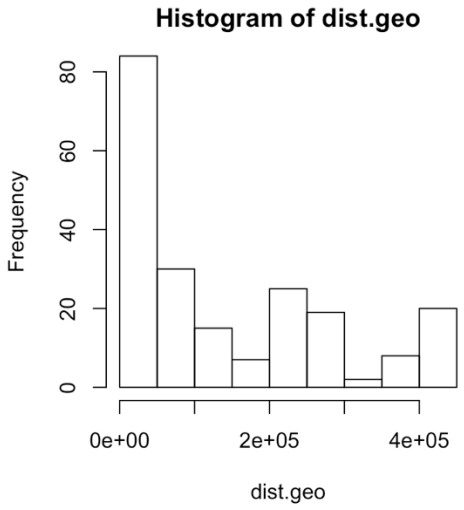


Figure S1.20: The non-normal frequency distribution of the geographic distance matrix.

Table S1.1: Differential relative abundances of the top 5 phyla on the 0.22 µm and 3 µm size fractions. Student's t-test identified significant differences between mean abundances of four phyla.

	0.22 µm mean(sd) %	**3 µm mean(sd) %**	**df**	**t-stat**	**p**
Actinobacteria	22.828(6.04)	5.361(2.38)	11.731	8.514	2.31E-06
Bacteroidetes	25.656(5.11)	34.2(8.21)	15.064	-2.793	0.01360339
Cyanobacteria	8.238(7.27)	19.4(8.22)	17.733	-3.217	0.00485304
Proteobacteria	26.204(5.09)	20.77(5.76)	17.731	2.235	0.03854459
Verrucomicrobia	14.442(4.8)	12.457(2.55)	13.709	1.155	0.26766726

Table S1.2: Statistics for the four detected modules in the Baikal co-occurrence network. M1 and M3 have the highest transitivity (clustering coefficient) values.

Module	**Average Path**	**Clustering Coefficient**
M1	1.626	0.684
M2	1.797	0.540
M3	1.410	0.823
M4	1.795	0.583

REFERENCES

REFERENCES

Bauer U, Clemente CJ, Renner T, Federle W. (2012). Form follows function: morphological diversification and alternative trapping strategies in carnivorous Nepenthes pitcher plants. *J Evol Biol* **25**: 90–102.

Berry D, Widder S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol* **5**: 219.

Brandes U, Delling D, Gaetler M. (2008). On Modularity Clustering. *IEEE Trans Knowl Data Eng* **20**: 172–188.

Calcagno V, Mazancourt C De. (2010). glmulti : An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *J Stat Softw* **34**: 1–29.

Crumpton W, Isenhart T, Mitchell P. (1992). Nitrate and organic N analyses with second-derivative spectroscopy. *Limnol Ocean* **37**: 907–9.

Csárdi G, Nepusz T. (2006). The igraph software package for complex network research. *InterJournal Complex Syst* **1695**: 1695.

Cushman SA, Landguth EL. (2010). Spurious correlations and inference in landscape genetics. *Mol Ecol* **19**: 3592–3602.

Cushman SA, Wasserman TN, Landguth EL, Shirk AJ. (2013). Re-evaluating causal modeling with mantel tests in landscape genetics. *Diversity* **5**: 51–72.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Friedman J, Alm EJ. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**: e1002687.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Gascuel O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14**: 685.

Guillot G, Rousset F. (2013). Dismantling the Mantel tests. *Methods Ecol Evol* **4**: 336–344.

Jost L. (2006). Entropy and diversity. *Oikos* **113**: 363–375.

Kozhov M. (1963). Lake Baikal and Its Life. Springer Netherlands: Dordrecht.

Kozhova OM, Izmest'eva LR. (1998). Lake Baikal : evolution and biodiversity. Backhuys Publishers.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–5120.

Langfelder P, Horvath S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Langille MGIM, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J a J, *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–21.

Leinster T, Cobbold CA. (2012). Measuring diveristy: the importance of species similarity. *Ecology* **93**: 477–489.

Lincoln S a., Wai B, Eppley JM, Church MJ, Summons RE, DeLong EF. (2014). Reply to Schouten: Planktonic Euryarchaeota are a significant source of archaeal tetraether lipids in the ocean. *Proc Natl Acad Sci* **111**: 1409439111-.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–8.

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

Montesinos-Navarro A, Segarra-Moragues JG, Valiente-Banuet A, Verdú M. (2012). Plant facilitation occurs between species differing in their associated arbuscular mycorrhizal fungi. *New Phytol* **196**: 835–44.

Moore MVM, Hampton SES, Izmest'eva LLR, Silow E a., Peshkova E V., Pavlov BK. (2009). Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia. *Bioscience* **59**: 405–417.

Moro MF, Silva IA, de Araújo FS, Nic Lughadha E, Meagher TR, Martins FR. (2015). The role of edaphic environment and climate in structuring phylogenetic pattern in seasonally dry tropical plant communities. *PLoS One* **10**: e0119166.

Newman M, Girvan M. (2004). Finding and evaluating community structure in networks.

Phys Rev E **69**: 1–16.

Nichols E, Uriarte M, Bunker DE, Favila ME, Slade EM, Vulinec K, *et al.* (2013). Trait-dependent response of dung beetle populations to tropical forest conversion at local and regional scales. *Ecology* **94**: 180–189.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, *et al.* (2016). vegan: Community Ecology Package. http://cran.r-project.org/package=vegan.

Schloss P, Westcott S. MiSeq SOP. https://www.mothur.org/wiki/MiSeq_SOP (Accessed November 30, 2015).

Shirk AJ, Landguth EL, Cushman SA. (2017). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Mol Ecol Resour* 55–67.

Valiente-Banuet A, Verdú M. (2007). Facilitation can increase the phylogenetic diversity of plant communities. *Ecol Lett* **10**: 1029–36.

Verdú M, Pausas JG. (2007). Fire drives phylogenetic clustering in Mediterranean Basin woody plant communities. *J Ecol* **95**: 1316–1323.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–7.

Welschmeyer N. (1994). Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. *Limnol Oceanogr* **39**: 1985–1992.

Zaneveld JR, Lozupone C, Gordon JI, Knight R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res* **38**: 3869–3879.

REFERENCES

REFERENCES

Bel'kova NL, Parfenova V V., Kostornova TY, Denisova LY, Zaichikov EF. (2003). Microbial biodiversity in the water of Lake Baikal. *Microbiology* **72**: 203–212.

Berry D, Widder S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol* **5**: 219.

Brandes U, Delling D, Gaetler M. (2008). On Modularity Clustering. *IEEE Trans Knowl Data Eng* **20**: 172–188.

Cáceres M De, Legendre P. (2009). Associations between species and groups of sites:\nindices and statistical inference. *Ecology* **90**: 3566–3574.

Calcagno V, Mazancourt C De. (2010). glmulti : An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *J Stat Softw* **34**: 1–29.

Comte J, Lovejoy C, Crevecoeur S, Vincent WF. (2015). Co-occurrence patterns in aquatic bacterial communities. *BGD Biogeosciences Discuss* **12**: 10233–10269.

Cotner JB, Biddanda BA. (2002). Small players, large role: Microbial influence on biogeochemical processes in pelagic aquatic ecosystems. *Ecosystems* **5**: 105–121.

Csárdi G, Nepusz T. (2006). The igraph software package for complex network research. *InterJournal Complex Syst* **1695**: 1695.

Cushman SA, Landguth EL. (2010). Spurious correlations and inference in landscape genetics. *Mol Ecol* **19**: 3592–3602.

Cushman SA, Wasserman TN, Landguth EL, Shirk AJ. (2013). Re-evaluating causal modeling with mantel tests in landscape genetics. *Diversity* **5**: 51–72.

Denef VJ, Mueller RS, Chiang E, Liebig JR, Vanderploeg HA. (2015). Chloroflexi CL500-11 populations that predominate keep lake hypolimnion bacterioplankton rely on nitrogen-rich DOM metabolism and C1 compound oxidation. *Appl Environ Microbiol*.

Denisova LY, Bel'kova NL, Tulokhonov II, Zaĭchikov EF. (1999). Diversity of bacteria at various depths in the southern part of Lake Baikal as detected by 16S rRNA sequencing. *Microbiology* **68**: 475–483.

Friedman J, Alm EJ. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**: e1002687.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Fujimoto M, Cavaletto J, Liebig JR, McCarthy A, Vanderploeg HA, Denef VJ. (2016). Spatiotemporal distribution of bacterioplankton functional groups along a freshwater estuary to pelagic gradient in Lake Michigan. *J Great Lakes Res* **42**: 1036–1048.

Hampton SE, Izmest'Eva LR, Moore M V., Katz SL, Dennis B, Silow E a. (2008). Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia. *Glob Chang Biol* **14**: 1947–1958.

Hill M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427–432.

Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, Garczarek L. (2008). Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium Prochlorococcus marinus PCC 9511. *J Biol Rhythms* **23**: 187–199.

Jost L. (2006). Entropy and diversity. *Oikos* **113**: 363–375.
Klausmeier C, Litchman E, Daufresne T, Levin S. (2004). Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. *Nature* **429**: 171–174.

Kozhov M. (1963). Lake Baikal and Its Life. Springer Netherlands: Dordrecht.

Kozhova OM, Izmest'eva LR. (1998). Lake Baikal : evolution and biodiversity. Backhuys Publishers.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–5120.

Kurilkina MI, Zakharova YR, Galachyants YP, Petrova DP, Bukin YS, Domysheva VM, *et al.* (2016). Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. *FEMS Microbiol Ecol* **92**: 1–13.

Langfelder P, Horvath S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Langille MGIM, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J a J, *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–21.

Leinster T, Cobbold CA. (2012). Measuring diveristy: the importance of species similarity. *Ecology* **93**: 477–489.

Lindeman RL. (1942). The Tophic-Dynamic Aspect of Ecology. *Ecology* **23**: 399–417.

Lindström ES, Forslund M, Algesten G, Bergström A-K. (2006). External control of bacterial community structure in lakes. *Limnol Oceanogr* **51**: 339–342.

Logue JB, Langenheder S, Andersson AF, Bertilsson S, Drakare S, Lanzén A, *et al.* (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J* **6**: 1127–36.

Logue JB, Lindström ES. (2010). Species sorting affects bacterioplankton community composition as determined by 16S rDNA and 16S rRNA fingerprints. *ISME J* **4**: 729–738.

Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci* **103**: 12552–12557.

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

Moore MVM, Hampton SES, Izmest'eva LLR, Silow E a., Peshkova E V., Pavlov BK. (2009). Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia. *Bioscience* **59**: 405–417.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S, R. J. Newton, *et al.* (2011). A guide to the natural history of freshwater lake bacteria.

Okazaki Y, Hodoki Y, Nakano SI. (2013). Seasonal dominance of CL500-11 bacterioplankton (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol Ecol* **83**: 82–92.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, *et al.* (2018). vegan: Community Ecology Package.

Overmann J. (2008). Green Nonsulfur Bacteria. *Life Sci*.

Parfenova VV, Gladkikh AS, Belykh OI. (2013). Comparative analysis of biodiversity in the planktonic and biofilm bacterial communities in Lake Baikal. *Microbiology* **82**: 91–101.

Salcher MM. (2014). Same same but different: Ecological niche partitioning of planktonic freshwater prokaryotes. *J Limnol* **73**: 74–87.

Santarella-Mellwig R, Pruggnaller S, Roos N, Mattaj IW, Devos DP. (2013). Three-Dimensional Reconstruction of Bacteria with a Complex Endomembrane System. *PLoS Biol* **11**

Schmidt ML, White JD, Denef VJ. (2016). Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton phyla. *Environ Microbiol* **18**: 1212–1226.

Sjöstedt J, Martiny JBH, Munk P, Riemann L. (2014). Abundance of broad bacterial taxa in the sargasso sea explained by environmental conditions but not water mass. *Appl Environ Microbiol* **80**: 2786–2795.

Team CW, Pachauri RK, Meyer LA. (2014). IPCC, 2014: climate change 2014: synthesis report. Contribution of Working Groups I. *II III to Fifth Assess Rep Intergov panel Clim Chang IPCC, Geneva, Switz* **151**.

Thomas F, Hehemann JH, Rebuffet E, Czjzek M, Michel G. (2011). Environmental and gut Bacteroidetes: The food connection. *Front Microbiol* **2**: 1–16.

Urbach E, Vergin KL, Larson GL, Giovannoni SJ. (2007). Bacterioplankton communities of Crater Lake, OR: Dynamic changes with euphotic zone food web structure and stable deep water populations. *Hydrobiologia* **574**: 161–177.

Walker BH. (1992). Biodiversity and Ecological Redundancy. *Conserv Biol* **6**: 18–23.

Watts DJ, Strogatz SH. (1998). Collectivedynamics of ‘small-world’ networks. *Nature* **393**: 440–442.

CHAPTER TWO

ESTIMATED NITROGEN ASSIMILATION STRATEGIES REFLECT SUMMER RESOURCE LIMITATION IN NORTH TEMPERATE LAKES

Abstract

Microorganisms are critical facilitators of biogeochemical and trophic processes in aquatic systems. Understanding drivers of their community composition in different systems and seasons is an active area of research. We present a 16s sequencing survey (139 samples) covering the world's oldest, most voluminous and deepest Lake Baikal, ten diverse lakes in Minnesota (including Lake Superior, prairie and Canadian shield lakes and a flooded meromictic iron mine), and three lakes in southwest Michigan during the summer and ice-covered periods. This survey places Baikal microbial communities at an end of a gradient of microbial composition, with the most similar microbial samples in the dataset collected from other oligotrophic areas, such as Lake Superior and the epilimnia of highly stratified lakes in the summer season. Total nitrogen and phosphorus, and oxygen were strong environmental covariates with community trends in winter. In contrast, temperature was the strongest environmental covariate in the summer with no significance in winter. Oxygen was also important only in the summer hypolimnion, but not in the mixed layer. Predicted functional repertoire at surveyed sites revealed that nitrogen availability strongly influenced the metabolic strategy of inorganic N uptake and assimilation via the glutamate pathway. Low N availability was associated with higher per-sample abundance of nitrogenase and the nitrate/nitrite transporters. N limitation also favored organisms that source electrons for

reduction and assimilation of N species from photosynthetic activity with Ferredoxin-dependent isoforms, in a direct trade-off with those that use NADH-dependent enzymes, coupled to electrons from catabolic processes. We propose that N conservation and the reducing power sources available in summer epilimnia are important factors that select microbial taxa in the summer season.

Introduction

The ecological impact of microorganism-driven nutrient cycling has long been recognized in aquatic systems, and various -omics approaches recently revealed the incredible taxonomic and functional diversity of aquatic microorganisms. In freshwater systems, microbial research efforts often focus on time-series collection at a single location (e.g., lake Mendota) or examining particular systems, like bog lakes (Linz *et al.*, 2017), alpine lakes (Urbach *et al.*, 2001), or lakes in the high Arctic. However, fewer studies attempt to contextualize findings across diverse habitats and large spatial scales (but see recent (He *et al.*, 2017; Mehrshad *et al.*, 2018).

Furthermore, our understanding of the functioning of aquatic ecosystems in general, and their microbial processes in particular, is largely restricted to the summer ice-free season. Of the world's 117 million lakes, the majority lie between 45° and 75° latitude and experience seasonal ice cover (Verpoorter *et al.*, 2014). However, academic schedules, assumptions about winter dormancy, and the logistics of winter field work have all but prevented systematic studies of under-ice aquatic habitats (Hampton *et al.*, 2015, 2017). Ongoing worldwide reduction in the duration of lake ice cover (Benson *et*

al., 2012) has been shown to influence phytoplankton ecology, including phenology and trophic role (Weyhenmeyer *et al.*, 1999), but many questions about the consequences of ice loss remain. Particularly little is known about the under-ice microbiome, restricting our understanding of the year-round functioning of aquatic systems and precluding accurate prediction of their response to ongoing climate change.

Resource limitation is one of the main forces that structure communities. The dominant paradigm for freshwater lakes is that phosphorus limits productivity. The canonical explanation for the primacy of P, as opposed to N, is the abundance of N in the atmospheric reservoir and its availability to aquatic systems through N-fixation. In the seminal work on this subject, Schindler speculated that transient N limitation in lakes is indeed possible, but that plankton ultimately overcome the shortage (Schindler, 1977). However, since then, a number of experimental studies suggested numerous exceptions, most commonly pointing to co-limitation by nitrogen and phosphorus, during periods of peak productivity (Elser *et al.*, 1990; Guildford & Hecky, 2000; Sterner, 2008; Harpole *et al.*, 2011; Elser *et al.*, 2007). O'Donnell *et al.* demonstrated such case for Lake Baikal in Siberia (O'Donnell *et al.*, 2017) – the world's deepest and most ancient freshwater lake, which holds about as much water as the Laurentian Great Lakes combined (Moore *et al.*, 2009). Even if the role of N is transient or conditional, the consistency of exceptions to P limitation clearly makes N an important factor in determining year-round ecosystem productivity and functioning. Given these important roles of N, it is vital to improve the mechanistic understanding of the way microbial communities affect the dynamics of N in freshwater systems.

Bacterial assimilation of inorganic N is a process of N species reduction, where reducing power is sourced from photosynthesis or catabolic pathways. Its first three enzymatically committed steps are catalyzed by nitrate reductase, nitrite reductase, and glutamate synthase (GS). There are two known isoforms of each one of these enzymes that differ in their electron cofactors. Multiple lines of evidence, coming from diverse phylogenies, including sequenced cyanobacteria and microeukaryotes, and the location and expression of the different enzyme isoforms in higher plants, have associated the ferredoxin-dependent (Fd) forms with nitrogen assimilation coupled to photosynthesis, and the NADH-dependent enzymes with obtaining reducing power from respiration (García-Gutiérrez *et al.*, 2018; Esteves-Ferreira *et al.*, 2018; Bernard & Habash, 2009). Community-wide prevalence of one enzyme isoform over another would reflect community shift in the direction of the most competitive N assimilation strategy in N-limited systems.

Here we present an amplicon survey (139 samples) of microbial communities covering a wide range of temperate lakes, the world's oldest, most voluminous and deepest Lake Baikal and thirteen diverse lakes in Minnesota, Wisconsin and Michigan (including Lake Superior, prairie and the Canadian Shield lakes, a flooded meromictic former iron mine, and the oligotrophic, mesotrophic and eutrophic lakes in MI) during the summer and ice-covered periods. We extend our previous findings from Lake Baikal to place its community composition in the context of other lakes, which were chosen to span a gradient of several environmental factors. We also identify and contrast main environmental drivers of community composition between the summer and winter

seasons in the study lakes. Finally, we use a functional prediction tool PICRUSt (Langille *et al.*, 2013) to estimate the functional repertoire in each sample and report two results. First, we confirm that nitrogen limitation, as inferred by TN:TP ratios, was associated with community shifts towards enrichment in nitrogenase genes – a trend consistent across all surveyed lakes. To our knowledge, we are first to report a trade-off between Fd-based and NADH-based nitrogen assimilation in N-limiting conditions, possibly brought on by N conservation and bioenergetics of acquiring reducing power.

Materials and Methods

Study sites

Summer sampling of Lake Baikal was guided by its recorded natural history (Kozhov, 1963) that divides the lake into eight distinct regions. We collected samples from 24 spatial locations, covering all eight regions, where 10 locations were sampled at various depths for a total of 46 samples across the lake (Fig. 2.1). The cruise took place on board the R.V. *Treskov* on August 3-17, 2013. Winter sampling at Baikal was done on two days (March 18 and 26, 2013) at the site of long-term monitoring “Station 1”, which was also part of the Summer survey, in the Southern Basin approximately 2.2 km offshore from Irkutsk State University Baikal Biological Station in Bolshie Koty.

Thirteen additional lakes (Table 2.1, Fig. 2.1) were sampled during the period of winter ice cover (March 16-20, 2015) and during the open water summer period (July 8-14, 2015). Lakes were chosen to represent a range of physical, chemical and biological conditions in order to assess how winter and summer microbial communities differ in a

diverse set of north temperate lakes. Two Lake Superior stations were selected in part due to logistical reasons, specifically ease of access over stable ice cover in winter. The Madeline Ice Road site (near Bayfield, WI; station 5 on Fig. 2.1) is near a maintained ice road connecting Madeline Island to the mainland and allowed us easy access to an ice-covered portion of Lake Superior. The Vandecraig Bay station (near Washburn, WI; station 4 on Fig. 2.1), situated in Chequamegon Bay, also has regular ice-cover with easy access. These two stations differ in physical, chemical and biological parameters (deep, oligotrophic site vs. shallow, oligo-mesotrophic nearshore site). Lakes Barrs, Briar, and Pike were chosen for their proximity to urban areas but different water clarity and depth. Lakes Burntside and Nels are both on the Canadian Shield, and, while Burntside is well-developed with vacation houses on its shores, Nels Lake is remote, accessible only by all wheel drive vehicles and has no shore development. Mille Lacs is the largest inland lake by surface area in Minnesota, but is relatively shallow. In contrast, lake La Salle is one of the deepest in MN with high water clarity. Lastly, lake Portsmouth is the deepest inland lake in Minnesota (137 m), and is a former iron pit mine. It was flooded in 1964 and since turned into a recreational lake, stocked with fish. Its depth, combined with small surface area, led us to hypothesize that it is meromictic with year-round stratification. Sampling was conducted from the surface of the ice in March and from a small boat during the ice-free period in July. In both seasons we collected water column CTD sonde profiles, and water samples from different depths for chemical and biological analyses.

Table 2.1: Sampled lakes in Michigan, Minnesota, and Wisconsin

Number on Fig. 1	Lake, Location	Lake Size, km^2	Site depth, m	Water sampling depths, m	Summer sampling date	Winter sampling date
1	Barrs, MN	0.52	6	0, 5	7/12/15	3/16/15
2	Briar, MN	0.3	5.5	0, 4.5	7/12/15	3/16/15
3	Pike, MN	1.97	13	0, 6, 12	7/8/15	3/16/15
4	Superior (Vandecraig Bay), WI	82102.6	8	0, 7.5	7/9/15	3/17/15
5	Superior (Madeline Ice Road), WI	82102.6	47	0, 15, 30, 45	7/9/15	3/17/15
6	Portsmouth, MN	0.5	93	0, 15, 25, 35, 80	7/10/15	3/18/15
7	Mille Lacs, MN	536.1	8.5	0, 7	7/10/15	3/18/15
8	Burntside, MN	28.9	26	0, 5, 14, 25	7/13/15	3/19/15
9	Nels, MN	0.74	8	0, 7	7/13/15	3/19/15
10	La Salle, MN	0.9	60.5	0, 7, 20, 35, 57	7/14/15	3/20/15
11	Gull, MI	8.8	31	0, 10, 13, 17, 25, 32	7/21/15	2/09/15
12	Wintergreen, MI	0.15	7	0, 4, 5.5, 6	7/29/15	2/12/15
13	Lawrence, MI	0.049	13	0, 3.5, 10.5, 12	7/29/15	2/11/15

Sample collection

We are reporting results based on data from several sampling expeditions. Methods for Lake Baikal are described in Chapter 1, and lakes in Michigan were sampled

following those protocols. Minnesota and Wisconsin lakes were sampled using similar techniques, as described in detail below.

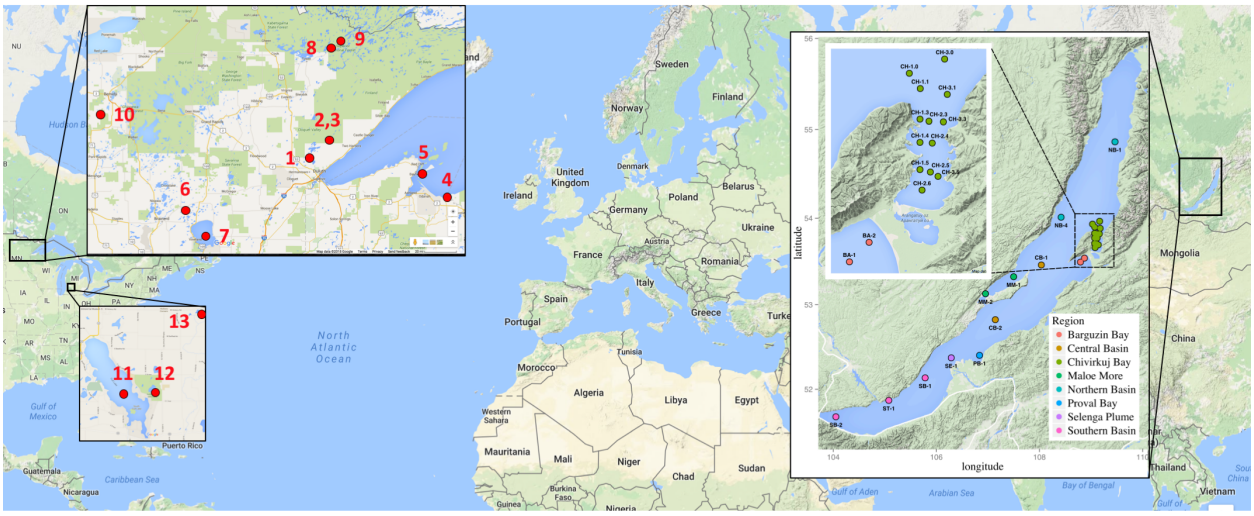


Figure 2.1: Map of sampled locations

Several physical characteristics were measured during winter and summer periods. In the ice-covered period, we visually estimated a percentage of snow cover on the ice. Average snow thickness was determined from measurements at 5 locations near the sampling site. The thickness and clarity of ice were recorded as well. Water and ice+snow light attenuation was measured with LI-COR probe equipped with a quantum LI-192 cosine sensor (LI-COR Biosciences., Lincoln, USA). Light attenuation by ice+snow was estimated by measuring light levels in air and directly under the ice+snow surface, pressing the surface of the light sensor to the underside of the ice. The sampling hole in the ice was covered with a black, light-proof plastic during the light measurements to avoid light contamination. Water column profiles of water temperature, dissolved oxygen, chlorophyll fluorescence, pH, fluorescent DOM and

conductivity were measured by EXO2 multiparameter sonde (YSI Inc., Yellow Springs, USA), recording approximately five values per second on the downcast (lowering approximately 1 m per second) and used for modeling (see below).

Water samples were collected for chlorophyll *a*, total phosphorus (TP), total nitrogen (TN), soluble reactive phosphorus (SRP), and nitrate and ammonia (NO_3^- and NH_4^+) with a 3.7L Van Dorn Water Sampler at discrete depths (Table 1). At minimum, water was collected at lake surface (or immediately under the ice in winter) and 0.5 m above lake bottom. For nutrient measurement, water was collected into 2 L acid-washed bottles and stored in dark coolers until return to the lab and processing.

Chlorophyll *a* analysis was performed in duplicate; samples (volume 100-400 mL) were filtered at a low vacuum onto cellulose nitrate membrane filters (0.2-*μ*m pore size; 45-mm ø) under low light conditions and stored frozen in the dark until analysis. Chlorophyll a was extracted into 90% acetone at 4°C in the dark for 18 to 24 hours before analysis and measured fluorometrically on a Turner Designs model 10-AU fluorometer (Welschmeyer, 1994).

For DNA, 1 L were sequentially filtered onto 3 µm and 0.22 µm mixed nitrocellulose acetate membranes (EMD Millipore, Billerica, MA, USA) and stored at –20°C in RNAlater (Life Technologies, Grand Island, NY, USA) to capture particle-attached and free-living fractions of microorganisms.

Nutrient measurements

Methods for Lake Baikal are described in Chapter One, and samples from Michigan lakes were analysed following those protocols. Minnesota and Wisconsin lakes were processed using similar techniques, as described below.

TP and TN analyses were conducted on duplicate whole water samples from each sampling depth. After collection, water for TP and TN analyses were transferred to 60-ml acid-washed plastic bottles and frozen until analysis. TP was determined spectrophotometrically by the molybdate blue method on Shimadzu UV-1800 spectrophotometer following a potassium persulfate oxidation at high temperature (Murphy & Riley, 1962). TN samples were acidified prior to analysis with 40 μL of 6N HCl. Samples were run on the Shimadzu TOC-VSH total carbon/nitrogen analyzer, where both total carbon and total nitrogen were quantified.

Duplicate samples for Soluble Reactive Phosphorus (SRP), NO_3^- and NH_4^+ were filtered through 0.22 μm cellulose nitrate membrane filters into separate 60 ml acid-washed plastic bottles and frozen until analysis. SRP concentrations were determined after adding color reagent and read on a Shimadzu UV Spectrophotometer UV-1800 at 880 nm. NO_3^- analysis was carried out on a QuickChem 8000 Lachat FIA Automated Ion Analyzer where the nitrate was reduced to nitrite by a copperized cadmium column and determined spectrophotometrically using the Low Flow method (Wendt, 2001). NH_4^+ was analyzed fluorometrically (Holmes *et al.*, 1999) on a Turner Designs 10-AU fluorometer at low light-conditions. Final NH_4^+ concentrations were calculated based on Taylor *et al* (Taylor *et al.*, 2007).

DNA extraction and amplicon sequencing

Genomic DNA was extracted using the Mo-Bio PowerSoil Kit (Mo-Bio Laboratories, Carlsbad, CA), following manufacturer's protocol. Then, the V4 region of the 16S rRNA gene was sequenced using dual-index primers as described previously (Kozich *et al.*, 2013). After PCR amplification, the products were normalized and pooled. The pool was loaded on an Illumina MiSeq v2 flow cell and sequenced with a standard 500 cycle reagent kit for paired-end 250 bp reads (PE250). Base calls were done with Real Time Analysis software v1.18.54. Output of RTA was demultiplexed and converted to FastQ with Illumina Bcl2Fastq v1.8.4.

Amplicon sequence processing

The FastQ output files were processed using mothur, following general MiSeq protocol and the options below (Kozich *et al.*, 2013; Schloss *et al.*, 2009). Sequences were aligned to a full SILVA v.123 database. Chimeras were removed with UCHIME in mothur (version 1.36) environment. The remaining sequences were classified with a naïve Bayesian RDP classifier (Wang *et al.*, 2007) and the Greengenes (August 2013 release) database. Sequences classified as Mitochondria, Chloroplasts and Eukaryota were removed, resulting in 197,738 unique sequences that were clustered into OTUs at 97% similarity. Consensus taxonomy for each OTU was determined following mothur protocol. Coverage for sequenced samples was rarefied to the lowest coverage sample at 27005 reads.

Statistical analyses

All analyses were performed in the R (3.2.2) environment. Temperature values at collection sites were estimated from YSI instruments sonde profiles. The YSI sonde profiled temperature at each station down to between 40 m to 50 m. For each station, high-order polynomial functions were fit to the data, and point estimates were used to infer temperature values at exact depths used for water sample collection (Fig. 2.1). For Lake Baikal sites samples collected at 75 m, 300 m and 500 m, temperature was assigned a value of 4°C, based on literature values (Kozhov, 1963).

Distance matrices, ordinations, and correlations with environmental variables we calculated with vegan (Oksanen *et al.*, 2016) and phyloseq (McMurdie & Holmes, 2013) packages. For consistency of comparing the effect of environmental covariates, we used only stations sampled in both seasons. The per sample gene content was calculated using PICRUSt (Langille *et al.*, 2013) with a quality cutoff NSTI<0.15. PICRUSt coverage varied between 54-87% of non-singleton OTUs and was well-represented across phylogenetic clades (Fig. S2.13). In other words, PICRUSt estimation did not introduce phylogenetic gaps in metagenomic estimates.

Estimated gene (KO) abundances were normalized to abundance of *recA* (K03553) in each sample, and the ratios were scaled linearly to a 0 to 1 range across all samples. Because *recA* is a single copy universal marker, functional KO to *recA* ratios represented abundances of genes per average cell in a community. And while different genes have highly variable baseline abundances, scaling of each gene across all samples allowed comparing different genes in the same normalized abundance range.

Results

Abiotic parameters

The abiotic parameters differed significantly between the seasons in all lakes (Fig. 2.2a, Fig. S2.10a, $p<0.01$). As expected, winter samples were at the low temperature position in multivariate space. Additionally, winter samples were also at the low end of the TN and TP gradients. The oxygen gradient was defined primarily by the lower saturation levels in hypoxic hypolimnia.

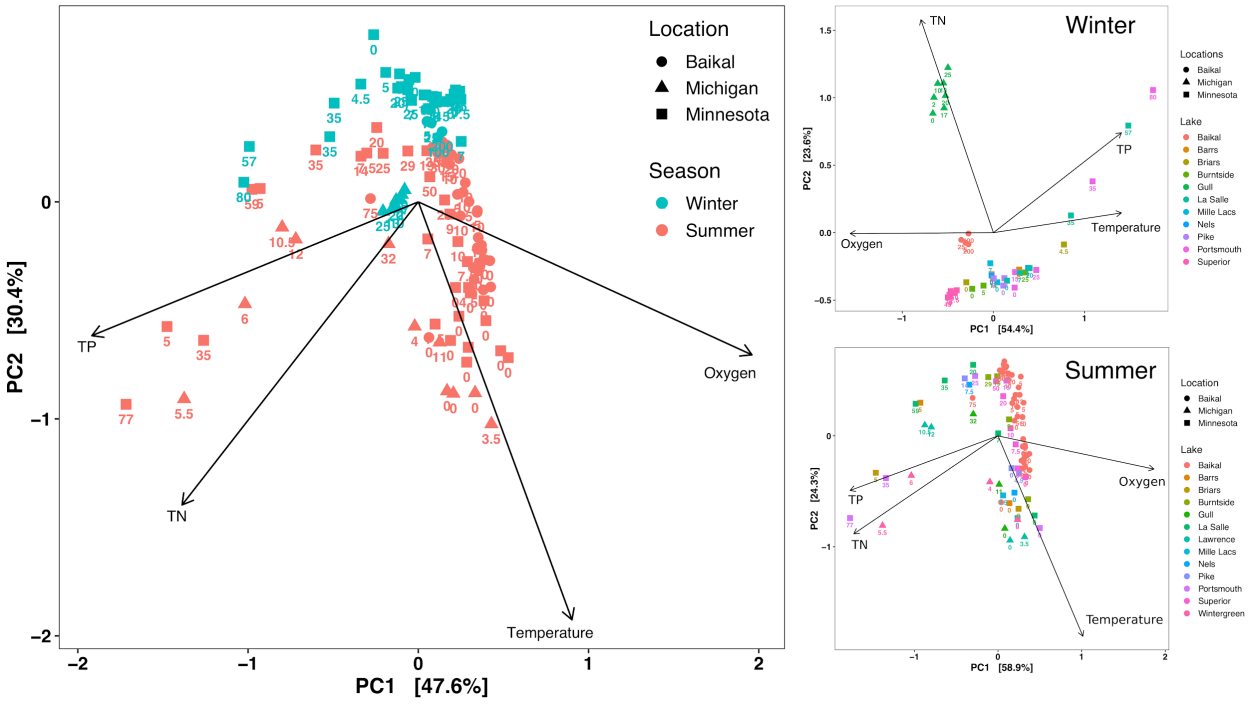


Figure 2.2: Abiotic differences between lakes in winter and summer seasons. Sample depths (m) are indicated below each point.

Beta dispersion of winter samples was significantly lower than that of summer samples (Fig. S2.10a, $p<0.03$), indicating lower variability of abiotic conditions in the winter season, with exceptions of lakes Portsmouth and La Salle, which had poor

seasonal mixing. Winter samples collected in the hypolimnia of those lakes were warmer, less oxygenated, and contained higher P, compared to all other samples, suggesting stratification in the winter season at those sites. Conversely, the summer epilimnia of those lakes were oligotrophic, with N and P levels approaching the lowest P concentrations in Lake Baikal samples (Fig. 2.2C). Interestingly, the winter Gull Lake, MI samples, while clustered together, differed markedly from all other locations, mostly due to high N content (Fig 2.2).

Microbial community composition and environmental covariates

At a depth of 27005 sequences per sample, we detected 133185 OTUs in the combined 3.0 µm (particle-attached) and the 0.22 µm (free-living) fractions. Of those, 42688 were non-singletons on 3.0 µm, and 49786 on the 0.22 µm. The non-singletons across the two fractions were classified into 63 phyla. Free-living samples were dominated by Actinobacteria, Bacteroidetes, Cyanobacteria, Proteobacteria, and Verrucomicrobia (surface sample composition in Fig. 2.3, 2.4). *Sediminibacterium, Limnohabitans* and *Synechococcus* were the most abundant taxa in the dataset. However, our study design was explicitly aimed to cover diverse environments (Fig. 2.1), and the relative abundances of individual taxa, as well as each phylum, varied widely between locations and seasons. For example, the hypolimnion of Portsmouth lake in both seasons was dominated by an obligate anaerobe in the Verruco-5 class, sulfur-reducing taxa *Desulfococcus*, members of the Desulfobulbaceae family and

sulfur-reducing Parvarchaea. To compare taxonomic community composition across lakes and seasons, we first considered ordinations of Bray-Curtis dissimilarities.

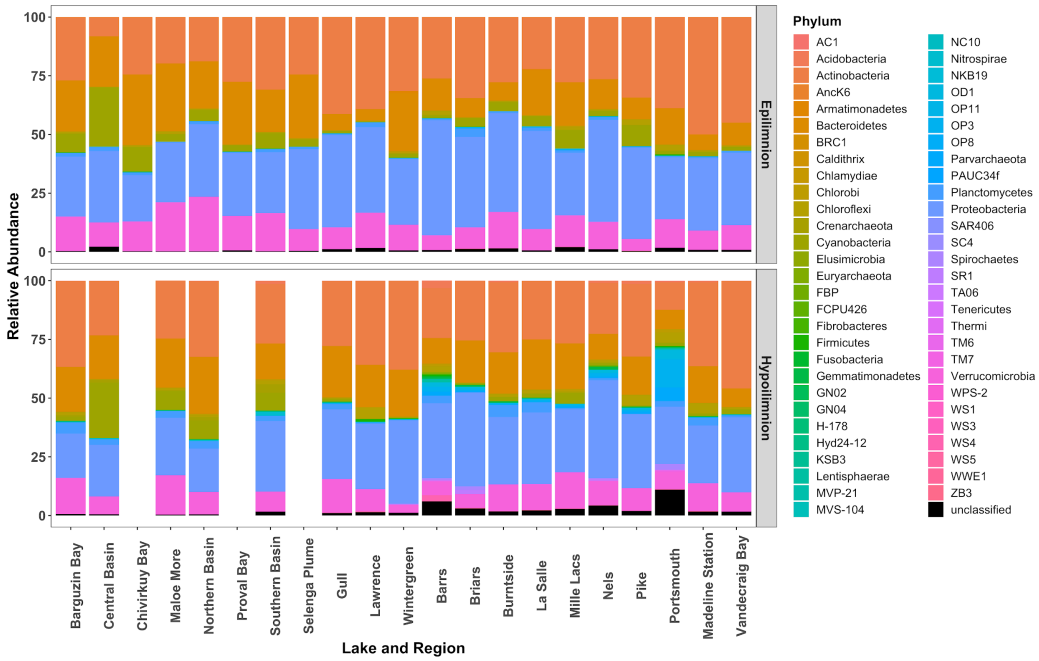


Figure 2.3: Taxonomic composition of free-living microorganisms in summer.

At the broadest scale, the communities showed significant differences in the centroids ($p<1x10^{-4}$) between summer and winter (Fig. 2.5, S2.10B) in ordination space. A substantial overlap occurred between winter samples and those from summer lake hypolimnia, while epilimnetic samples separated almost completely. In both winter and summer, Lake Baikal occupied one extreme of the community dissimilarity space (Fig. 2.5B,C). At the other extreme were samples collected from below the thermo- and chemocline of the deep meromictic Portsmouth lake in Minnesota. Communities most similar to Lake Baikal were collected in Lake Superior in both seasons. In the summer, additional samples similar to Baikal included the epilimnia of lakes LaSalle and

Portsmouth in Minnesota, and Gull Lake in Michigan, which were all oligotrophic in that season.

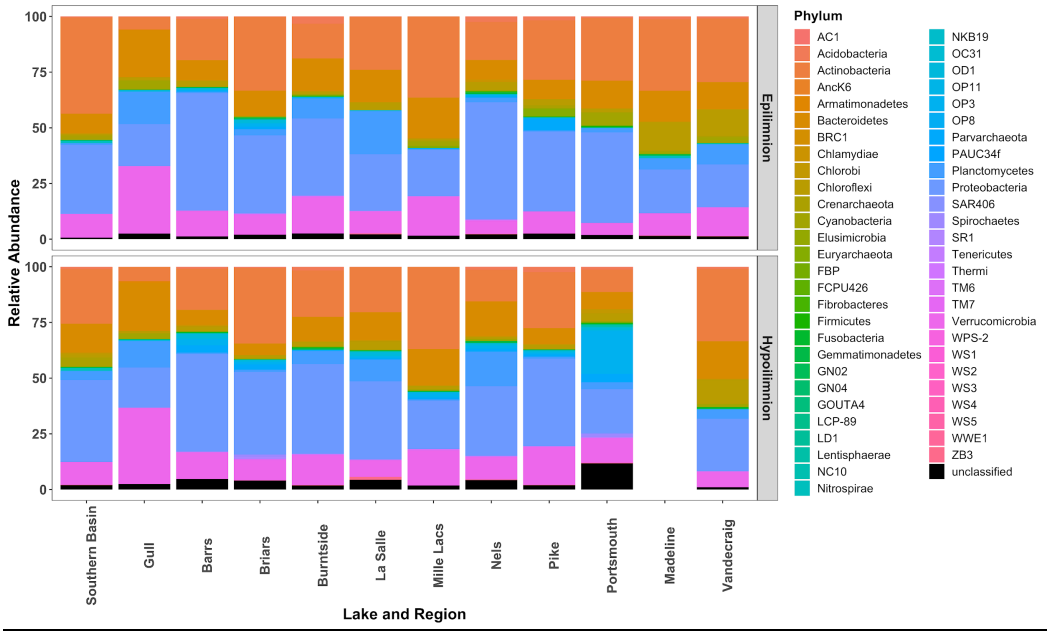


Figure 2.4: Taxonomic composition of free-living microorganisms in winter.

Considering two seasons separately, total nitrogen was the strongest environmental covariate separating community composition in the winter (R^2=0.81) and a significant one in the summer (R^2=0.58), with dissolved and total P also significant in both seasons (Fig. 2.4). Temperature was only significant in summer, where it was a strong covariate (R^2=0.55).

A closer look at summer epilimnia revealed that oxygen did not correlate with microbial community structure in those systems (Fig. S2.12). TN was still the most significant covariate (R^2=0.63), while total P was not, and dissolved nitrogen had greater explanatory power than dissolved phosphorus (R^2=0.50, R^2=0.36). Our next step was to

examine the functional repertoire of samples the most productive season and region i.e., summer epilimnia, through the lens of N availability.

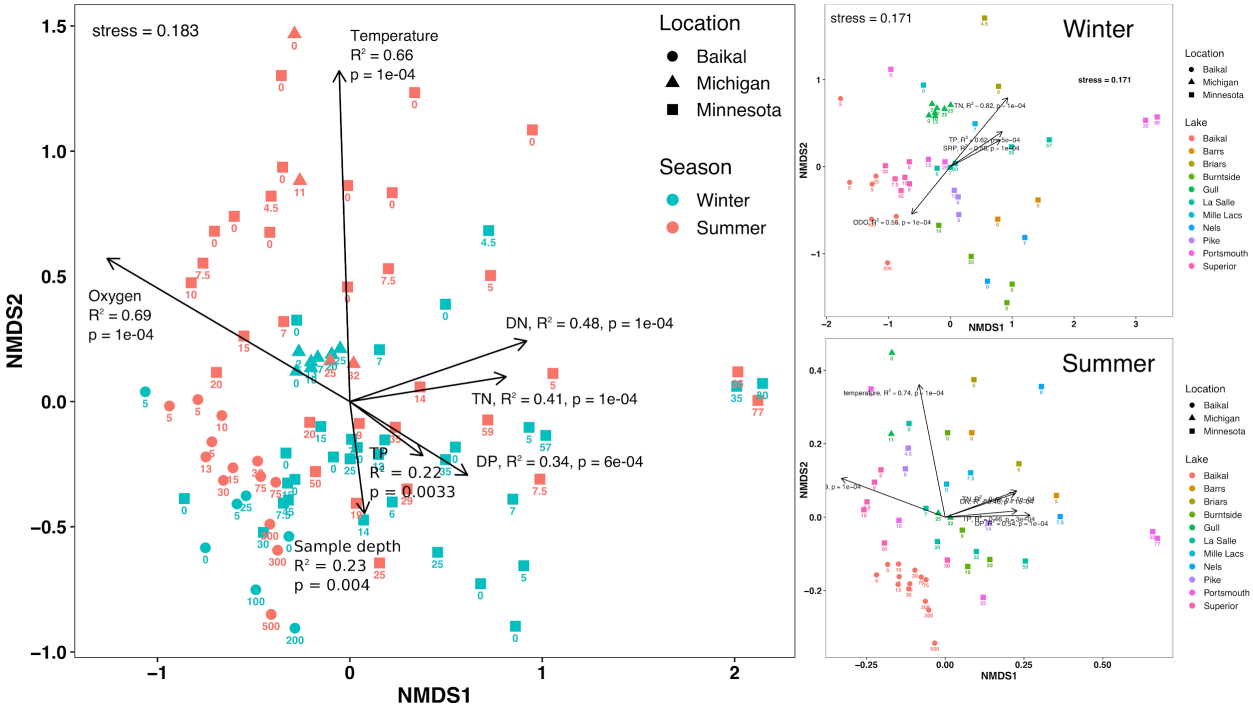


Figure 2.5: Biotic dissimilarities in winter and summer. Only stations sampled in both seasons were included for consistency. Numbers below points are sample depths (m).

Nitrogen assimilation strategies

The PICRUSt-derived estimates of nitrogenase gene abundance were higher under lower N availability in the summer mixed layer (ML) samples. This trend was consistent across all lakes and was significant against both TDN (Fig. 2.6, R^2=0.30, $p<1.1x10^{-5}$) and TN/TP ratios (Fig. S2.14, R^2=0.23, $p<3x10^{-4}$). Because PICRUSt metagenome estimates are reflections of the functional community structure, these results indicate community shifts towards taxa capable of N fixation. The nitrogenase enrichment in the summer epilimnia also correlated with TN but not with any other measured

environmental variable. This signal also became non-significant when winter samples (Fig. S2.14) and the summer hypolimnion were included, suggesting a decreased selective pressure for nitrogen fixation or presence of limiting resources other than nitrogen in those environments.

Inorganic nitrogen transporter genes (Nrt) followed the nitrogenase abundance trends, showing a significant decrease in the per sample abundance with increasing N availability, as estimated by TDN (Fig. 2.6, R^2=0.38, $p<3.7x10^{-7}$), as well as TN/TP ratios (Fig. S2.15, R^2=0.32, $p<3.7x10^{-6}$), in summer ML across all sites. The trend was noticeably stronger at the low end of N availability, represented by the Lake Baikal samples. Indeed, at higher N availability, lakes Superior in Minnesota and Gull and Lawrence lakes in Michigan showed a large variability in community-wide Nrt prevalence at similar TDN values (Fig. 2.6), with Nrt abundance in Lake Superior almost as high as in many Lake Baikal samples and the Lawrence Lake samples had Nrt content near zero. Importantly, just as in case of nitrogenase, the trends were absent in the winter, and the addition of winter samples to the analyses eliminated significance (Fig. S2.15). For Baikal summer epilimnion, N co-limitation was experimentally shown at the same time with the same water samples as those used in this report (O'Donnell *et al.*, 2017). Altogether, these observations revealed an association between N availability and N acquisition strategy, but only in systems that are expected to experience N limitation.

In the metabolic machinery for inorganic nitrogen reduction to ammonia and assimilation of ammonia into glutamate, ferredoxin (Fd)-dependent N assimilation

enzymes showed a clear trade-off with NADH-dependent isoforms, with respect to available nitrogen, but, again, only in the summer epilimnion samples from Lake Baikal (Fig. 2.7), which exclusively comprised the low end of the TDN gradient (Fig. 2.6). NADH-dependent forms of nitrate reductase, nitrite reductase, and glutamate synthase all reached a plateau with increasing dissolved nitrogen in the Lake Baikal epilimnion samples. This was matched by a decrease in the FD-dependent forms in the same samples in similar but not exactly opposite amounts. These results reflected shifts community composition away from the organisms that rely on photosynthesis-dependent N assimilation in the presence of greater available nitrogen, as measured by TDN. At the high end of the TDN gradient, occupied by summer mixed layers of lakes other than Bakal, the trade-off remained at high (for NADH) and low (for Fd) plateaus.

Discussion

We present a survey of microbial communities in diverse northern lakes, including the planet's most voluminous, deepest and most ancient Lake Baikal in Siberia, the world's largest (by surface area) Lake Superior, a meromictic former iron pit mine i.e., Portsmouth Lake, and the classic "Wetzel lakes" Gull, Lawrence, and Wintergreen in southwest Michigan, familiar to most aquatic ecologists. We reveal seasonal differences in environmental drivers of microbial community structure, generalized across wide spatial scales. We are also, to our knowledge, first to report shifts in microbial community composition that result in a nitrogen-dependent functional trade-off of

nitrogen assimilation strategies and hypothesize that the trade-off is based on availability of electron sources and is manifested only during N limitation.

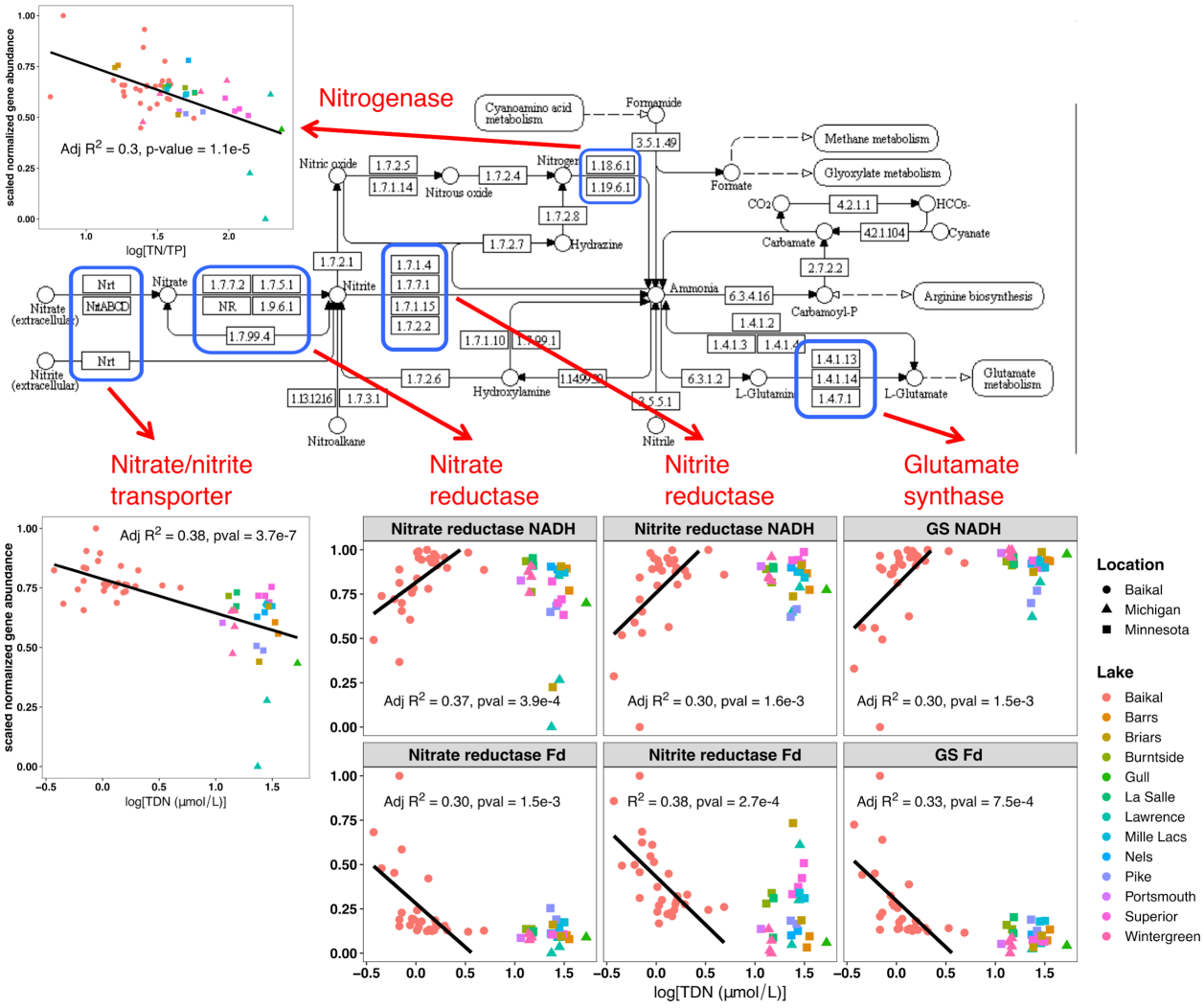


Figure 2.6: Regression of transporters and nitrogen assimilation genes with log dissolved nitrogen

The dominant phyla were generally consistent with findings from other freshwater temperate lakes (Newton *et al.*, 2011; Logue *et al.*, 2012), with notable exceptions, such as the hypolimnion of likely meromictic Portsmouth Lake, MN. While the communities in the Portsmouth Lake's epilimnion appeared similar to other small oligotrophic lakes,

such as LaSalle Lake in Minnesota, and Gull and Lawrence lakes in Michigan, below the chemocline at 25 m the environment became completely anoxic (Fig. S2.9). In that environment, communities were markedly different, dominated by sulfur-reducing bacteria and archaea (see Results). Parvarchea, present in Portsmouth Lake's hypolimnion, have been found in anoxic environments of arctic drained lake basins (Kao-Kniffin *et al.*, 2015) and marine sediment mats (Wong *et al.*, 2017), and members of the Desulfobulbaceae family, including *Desulfococcus*, are known sulfate reducers (Almstrand *et al.*, 2016; Rosenberg, 2014). Similarly, the bottom hypolimnion in lake La Salle were differentially abundant in *Methylotenera mobilis*, which has been methylamine oxidation in lake sediments (Kalyuzhnaya *et al.*, 2006) and members of the obligately anaerobic Phycisphaerales family (Fukunaga *et al.*, 2009). These Portsmouth hypolimnion communities were remarkably similar between summer and winter (Fig. 2.5A).

Summer communities were structured along a strong temperature gradient (Fig. 2.5C), with nutrients, including TN/TP ratios also significant in the epilimnia. Although oxygen was a significant covariate in summer samples, it varied greatly with depth, owing to lake stratification, which was present in most lakes at the time of summer sampling. This prompted us to analyze of the epilimnetic and hypolimnetic samples separately, which revealed, as expected, that oxygen was a significant covariate only in the hypolimnia but had no effect on microbial community structure in the epilimnia (Fig. S2.12).

In contrast to summer, oxygen saturation, TN, and TP were significant environmental covariates in the ice-covered season (Fig. 2.5B). The multivariate trends held even with exclusion of lakes stratified in the winter season (Portsmouth and La Salle, Fig. S2.13) and retaining samples from Lake Baikal, which is oxygenated at all depths year-round. The absence of temperature significance was expected, because it is less variable across lakes and depths in the winter season. Furthermore, TN/TP ratios were not significant covariates in winter either. This is noteworthy because it indicated a possibly diminished role of nitrogen limitation during the ice-covered period. Thus, our findings highlight the importance of heterotrophy under the ice, suggesting that oxygen availability, rather than temperature or nutrient limitation, is a key force in structuring communities in the winter season, as well as the hypoxic hypolimnia of deep lakes in the summer.

Nitrogen assimilation machinery reflected resource limitation

Correlation of the per-sample gene abundances with environmental variables indicated that N availability during possible N limitation had consistent effects on its assimilation mechanism. O'Donnell *et al.* (2017) experimentally established N co-limitation in Lake Baikal in the summer using water from the samples presented in the current report. Extending analyses to the rest of the lakes, we use TN/TP ratios and TDN as a proxy for N availability.

Nitrogenase and inorganic nitrogen transporters (Fig. 2.6) were enriched in low N availability samples, but only in the summer and not in winter (Fig. S2.14, S2.15). We

propose that the absence of nitrogenase and Nrt trends in winter is due to the lack of N limitation in that season, where microbial communities could be limited by other environmental factors, such as light, carbon, or oxygen. The relative importance of oxygen reserves in the ice-covered period, when lakes' gas exchange with the air is made impossible, has long been known (Hampton *et al.*, 2017). Our multivariate analyses, when separated by season and stratification layers, confirm that oxygen is only important in structuring microbial communities in winter and only in the hypolimnia of summer lakes. It is important to point out that in the case of nitrogenase, depletion was only observed with respect to TN/TP ratios and not TDN, even in the summer ML. This suggests a greater association between intracellular nitrogen reserves and the competitive advantage of having nitrogen fixing capabilities. Thus, our results highlight the larger role of nutrient limitation in the mixed layer during the summer season, which is reflected in community structure that favors different ways of assimilating N.

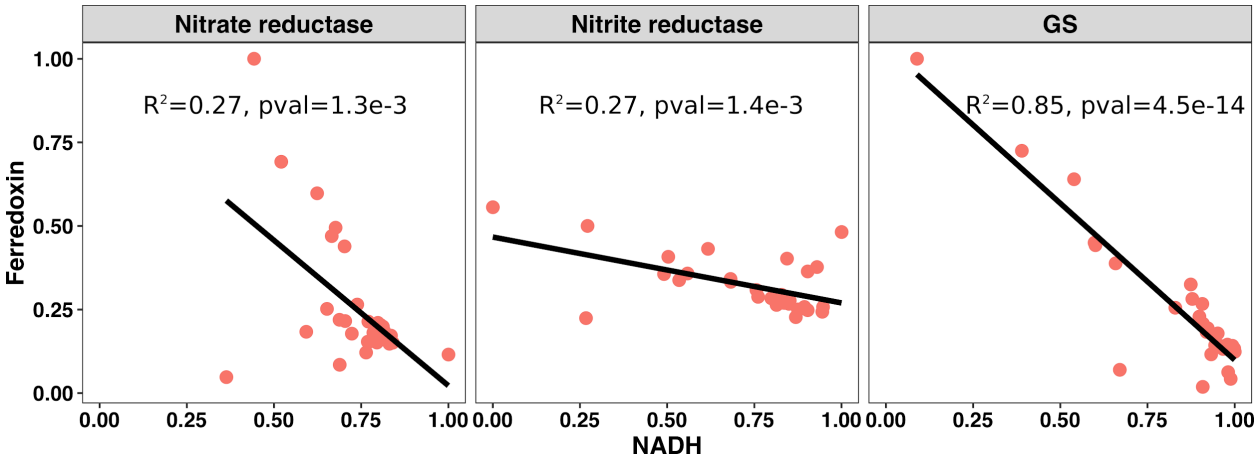


Figure 2.7: Apparent trade-off between Fd- and NADH-dependent N assimilation enzyme isoforms in Lake Baikal

Among our most intriguing findings is the apparent nitrogen-dependent trade-off between Fd- and NADH- cofactor isoforms of enzymes in the N assimilation pathway (Fig. 2.6). In these downstream steps of N assimilation, where N species get reduced to ammonia and then assimilated into glutamate, N limitation appears to favor Fd-dependent enzymatic isoforms. The N-limited Lake Baikal samples got progressively depleted in Fd-dependent enzymes with increasing nitrogen availability, in an apparent trade-off with NADH-dependent enzymes, which got enriched. Analogous trade-offs have long been known in higher plants, where Fd-dependent nitrate reductase and GS are active in the leaves, and are thus associated with photosynthesis, while NADH-dependent isoforms are active in the roots (Lea & Miflin, 2003). Classic protein studies of germinating seeds recorded a switch from NADH- to Fd- dependent isoforms, once cotyledons emerge (Matoh & Takahashi, 1982) and pointed to light as an activator of Fd-GS (Hecht *et al.*, 1988). Indeed, in our dataset, the relative abundance of cyanobacteria increased with decreasing available nitrogen (Fig. S2.16), and all sequenced cyanobacteria, including *Prochlorococcus* have the Fd-dependent GS; however, some also carry the NADH isoforms (Temple *et al.*, 1998; Muro-Pastor *et al.*, 2005). We offer two explanations for the observed trade-off in our study. The first has to do with intracellular nitrogen conservation.

It is possible that in a N-limited system, organisms that use N conserving cofactors to assimilate nitrogen become more competitive. NADH is a nitrogen-rich molecule, containing 7 N atoms (and a 7/2 N/P atomic ratio), while Ferredoxin only contains four N atoms on its four cysteine residues. Various strategies of N conservation pursued by

hallmark marine oligotrophic prokaryotes, have been widely reported, and include reduced G-C content, substitution of N-rich amino acid residues, even at the cost of creating overall bulkier proteins, and losses of entire pathways, such as the absence of DNA repair mechanisms in *Prochlorococcus* (Grzymski & Dussaq, 2012). While nitrogen and phosphorus concentrations in lakes are usually positively correlated (Sterner, 2008), our observed gene trends did not manifest in analyses with TP and TDP, further implying an important role of nitrogen. However, the Fd-NADH trade-off still existed, albeit more weakly, even at the high end of the N availability gradient, where many lakes presumably are no longer N-limited. Our second explanation deals with the energetics of heterotrophy, as opposed to autotrophy, as electron sources.

Enrichment in Fd-dependent N assimilation genes could indicate increased competitiveness of photosynthetic taxa in low N environments. Clearly, that is the strategy followed by nitrogen-fixing cyanobacteria, such as some members of *Anabaena*. In other words, the oligotrophic locations in Lake Baikal, where TDN is at its lowest, may simply not support enough heterotrophy to make the downstream NADH-based N assimilation competitive.

Combining our two proposed explanations, it is possible that Fd-based assimilation of N becomes more advantageous under N limitation and high light in summer epilimnia. In the winter season and in the summer hypolimnia either or both of these conditions are relaxed, and the N dependent Fd-NADH cofactor tradeoff is no longer detectable.

Conclusion

Our study identifies major environmental factors structuring temperate freshwater microbial communities across extensive spatial scale and between seasons. We show that different abiotic factors were associated with the community structure in the winter, compared to the summer season. We highlight how N limitation shapes the functional composition of microbial communities by selecting for species capable of N fixation or with the nitrogen assimilation genes that rely on N-free Fd cofactors (in a trade-off with NADH-dependent N-rich isoforms) that draw reducing power from photosynthesis, presumably from the abundant light available in summer epilimnia. Our results show what abiotic factors structure pelagic microbial communities in a diverse set of temperate lakes in two different seasons and underscore the importance of N availability in determining the functional repertoire of these communities.

APPENDIX

APPENDIX

SUPPLEMENTARY INFORMATION FOR CHAPTER TWO

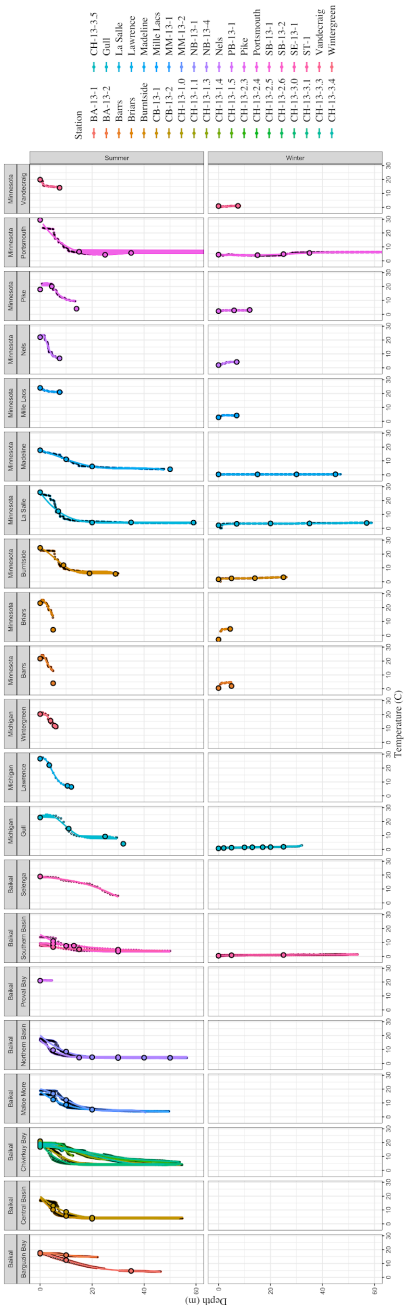


Figure S2.8: Temperature profiles each surveyed station, as recorded by a YSI sonde (solid black circles), fitted polynomial functions (colored lines), and estimated temperatures at depths, from which microbial samples were collected (large black outline circles).

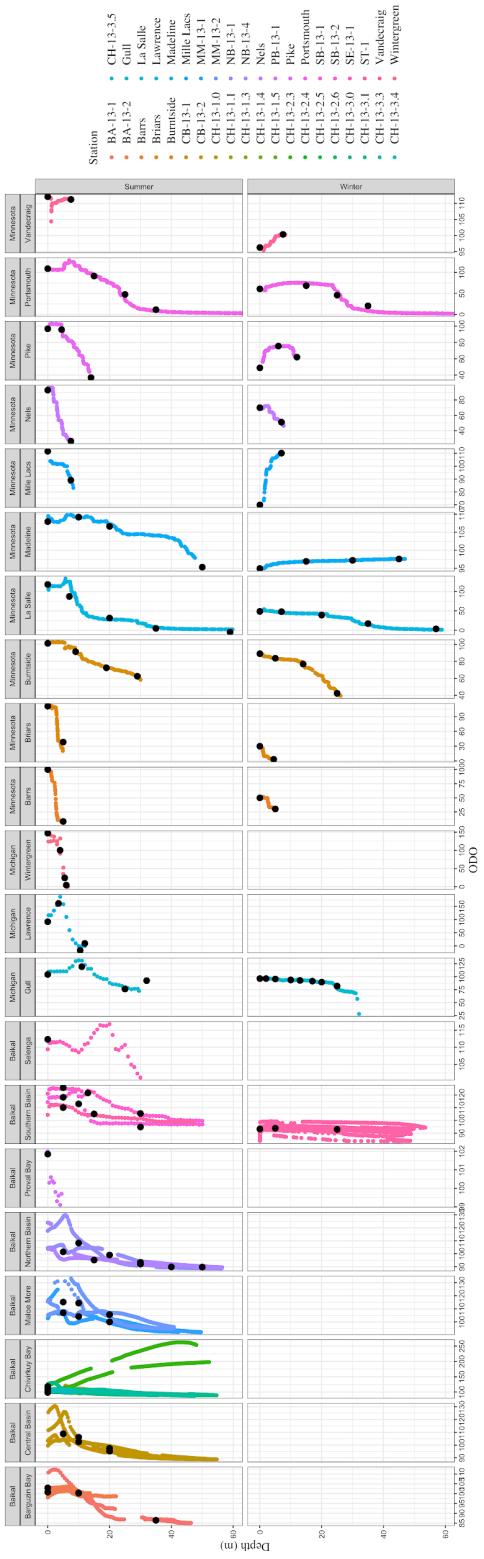


Figure S2.9: Oxygen profiles each surveyed station, as recorded by a YSI sonde (solid black circles), fitted polynomial functions (colored lines), and estimated temperatures at depths, from which microbial samples were collected (large black outline circles).

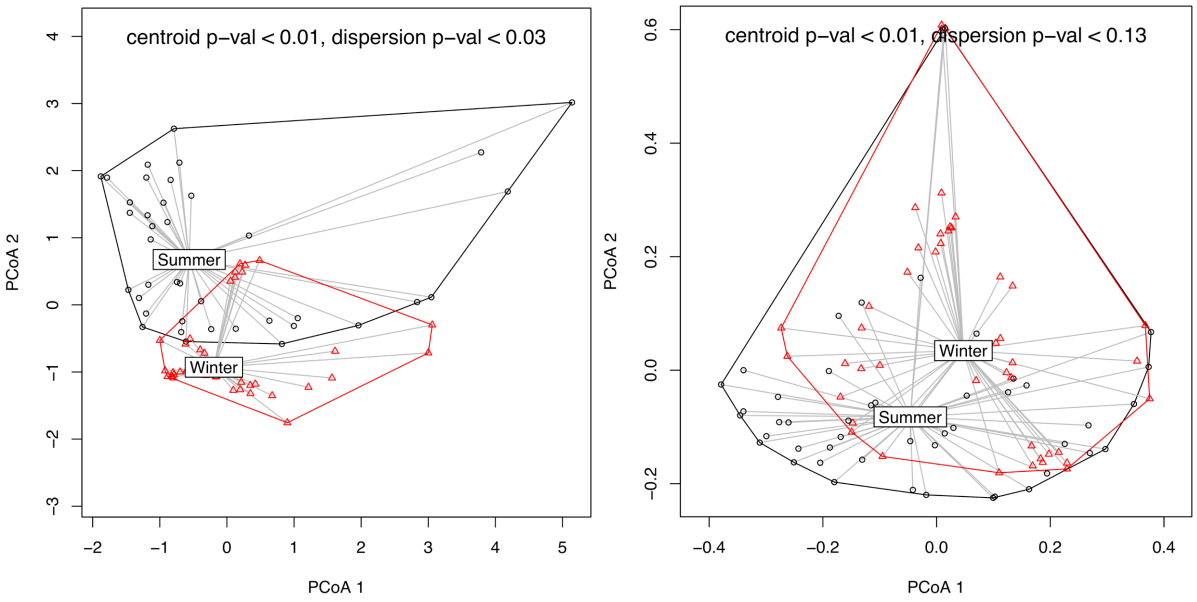


Figure S2.10: Principal Coordinate Analysis of sampled stations, based on abiotic (A) and biotic (B) measurements.

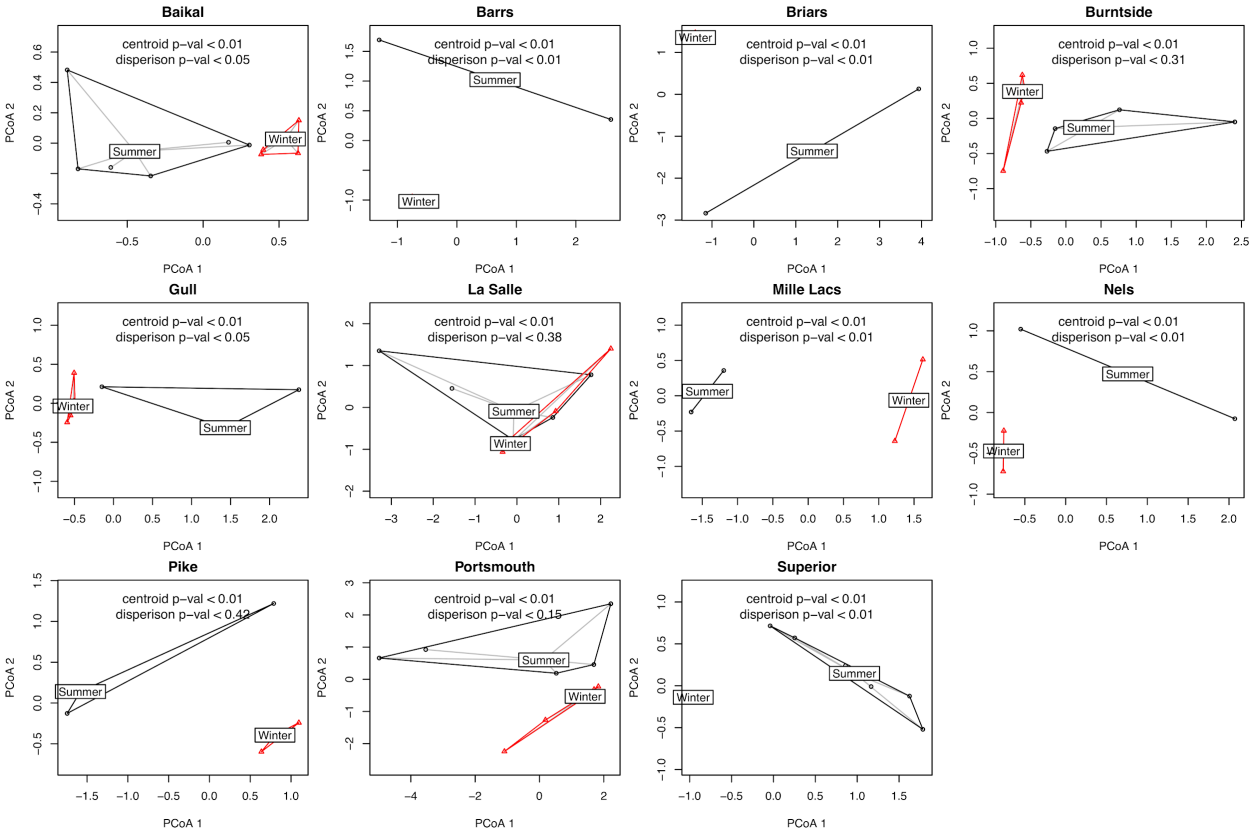


Figure S2.11: Beta dispersion around the centroid of measured abiotic variables for each lake, compared between summer and winter seasons. All lakes showed significant differences in abiotic conditions (centroids) between the two seasons. Also, all lakes, except Pike, Burntside, La Salle, and Portsmouth showed greater abiotic variation (dispersion) in summer, compared to winter.

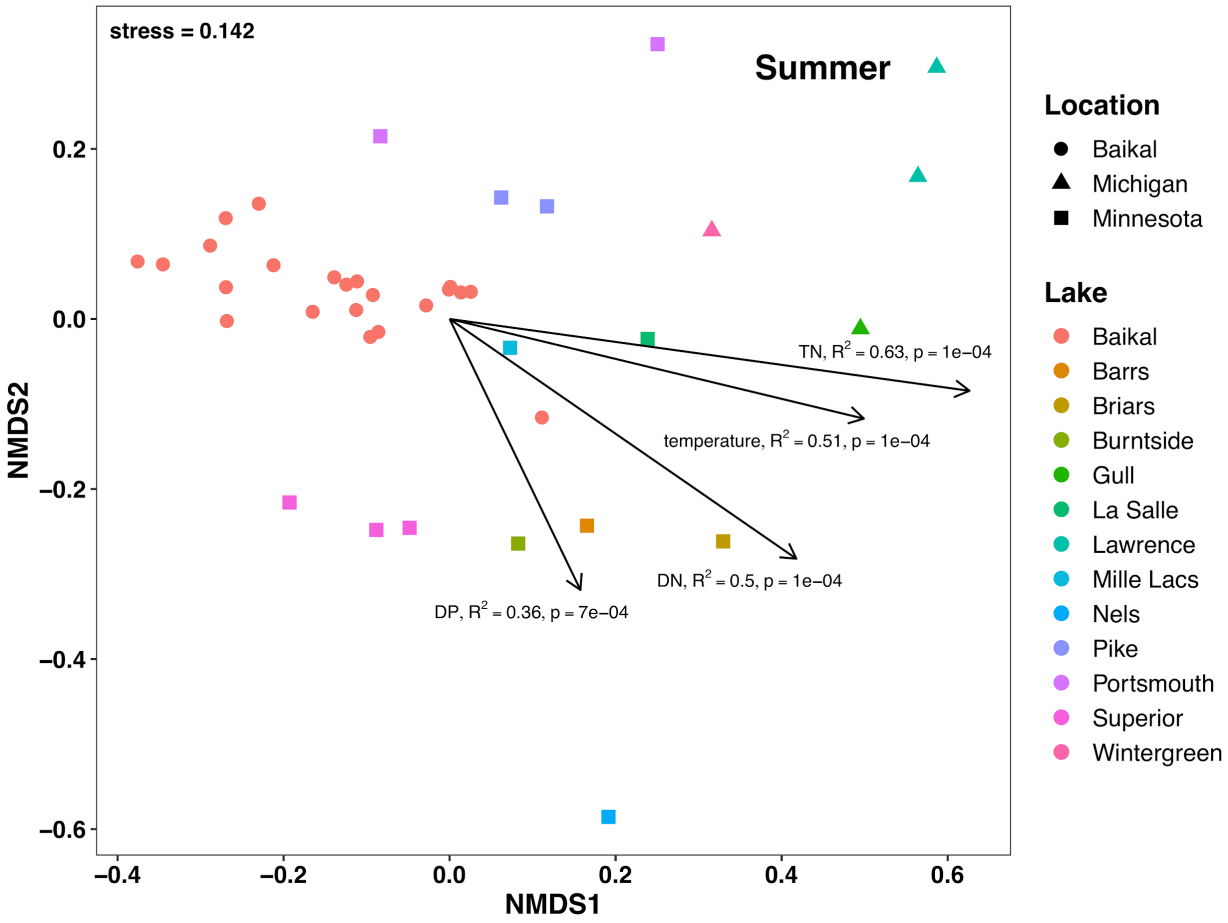


Figure S2.12: Microbial community similarity across lakes for summer epilimnetic samples. Total N was the most significant covariate, while total P was not, and dissolved nitrogen had greater explanatory power than dissolved phosphorus. Oxygen did not predict microbial community structure in these samples.

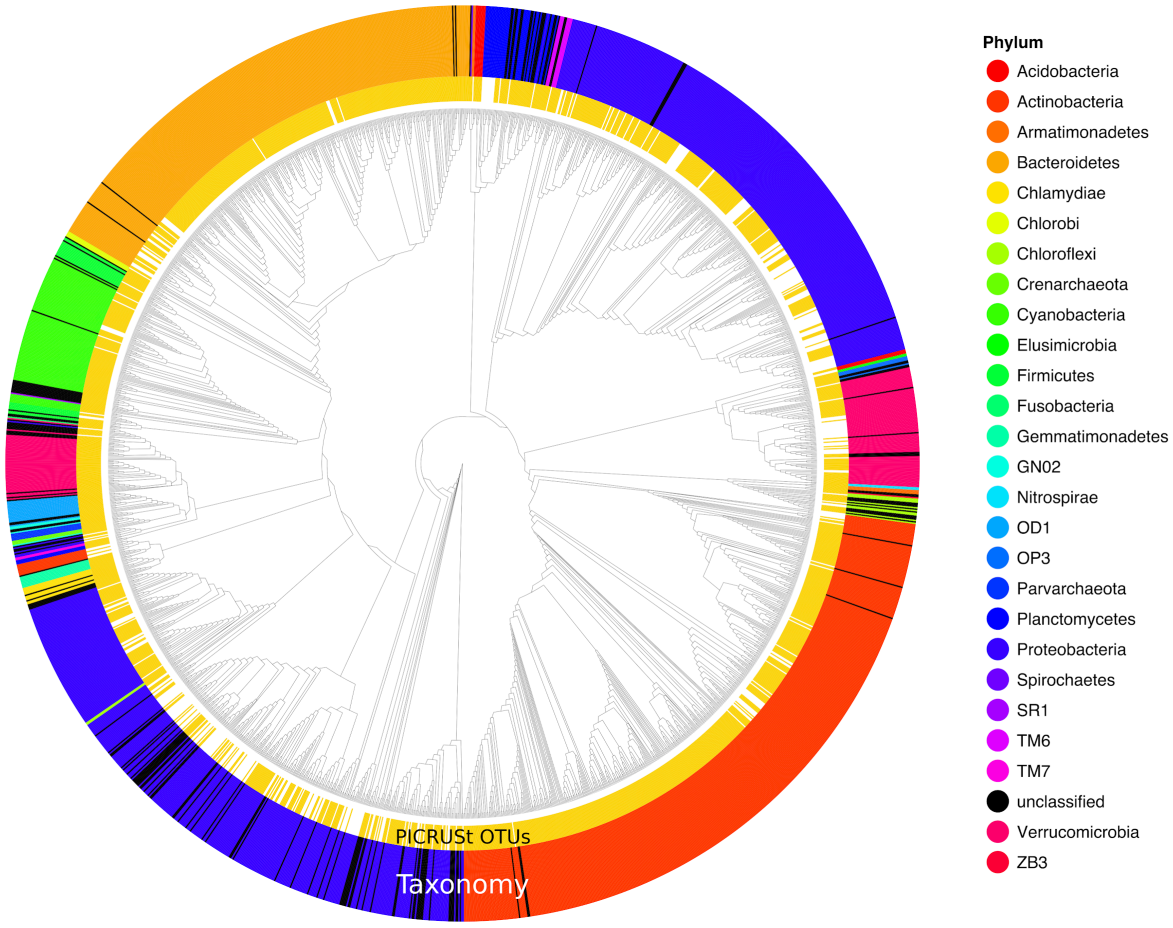


Figure S2.13: PICRUSt coverage expressed as percentage of PICRUSt-compatible OTUs in each sample (54-87%). The differential abundances of the same set of PICRUSt-compatible OTUs across samples and the estimated copy number of each functional gene (KEGG Orthologs, KOs) in the PICRUSt-compatible OTU genomes that was used to produce the bulk per-sample KO abundances.

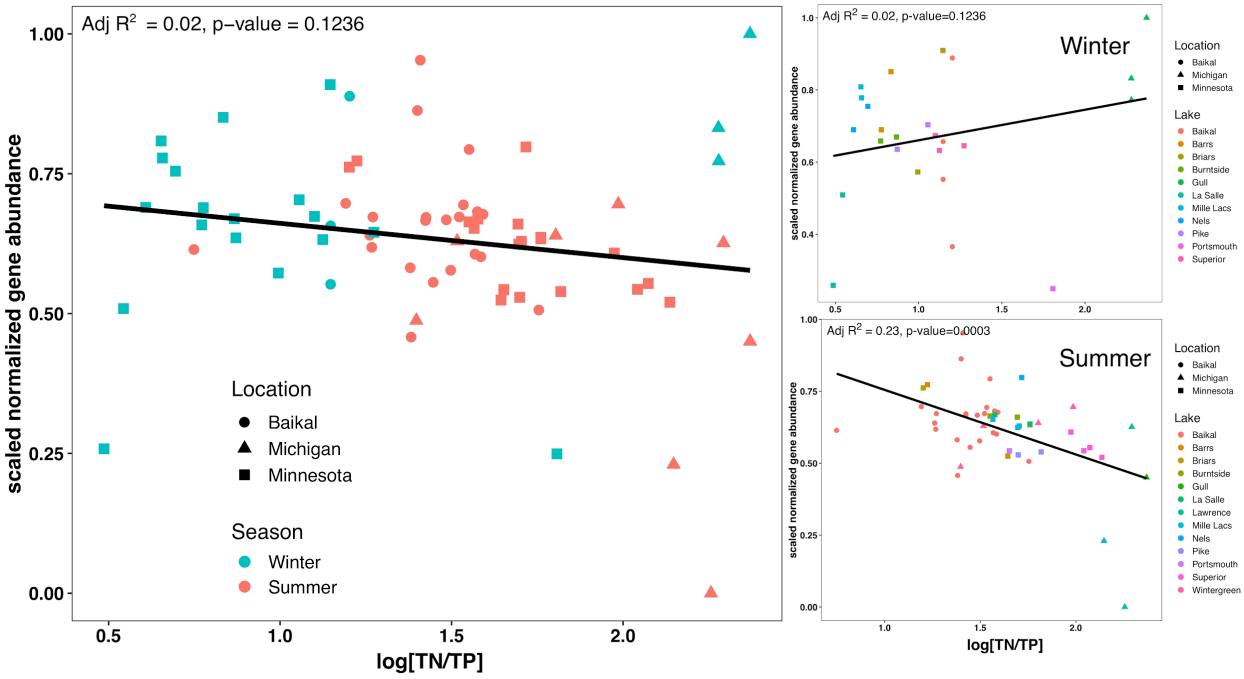


Figure S2.14: Nitrogenase per-sample content in combined winter and summer surface samples (left), and winter and summer seasons. Nitrogenase abundance decreased with log TN/TP only in the summer mixed layer, reflecting community shifts towards greater N_2 fixation under N-limiting conditions.

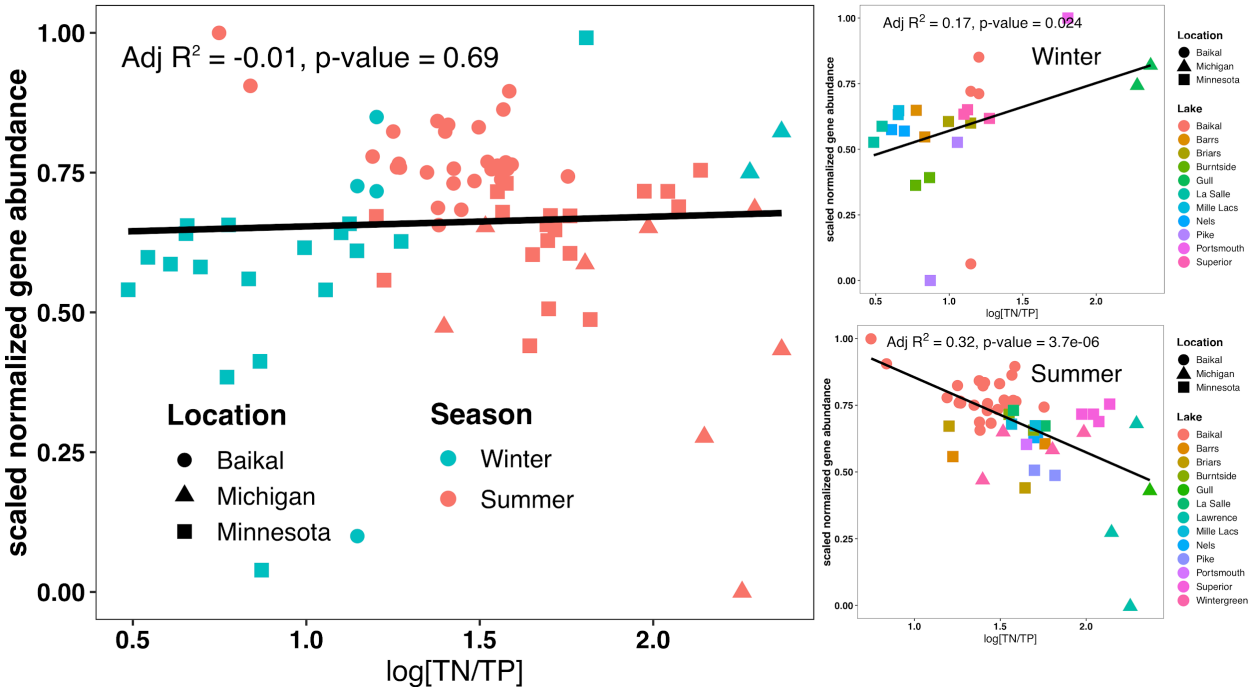


Figure S2.15: Nitrate and nitrite transporter gene (Nrt) per-sample abundance vs. log (TN/TP) in A) Both winter and summer samples. B) Winter samples, C) Summer samples

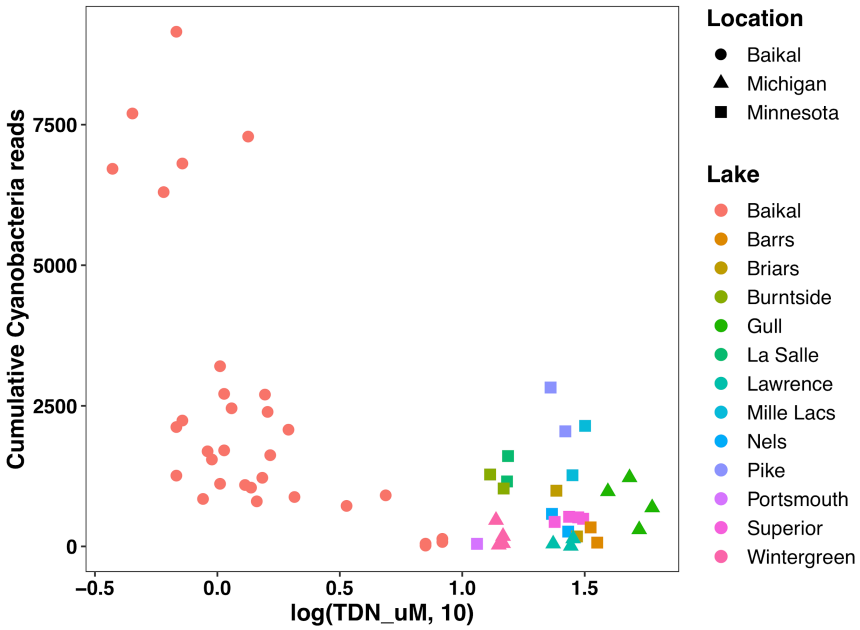


Figure S2.16: Cyanobacteria increased with more limiting N

REFERENCES

REFERENCES

Almstrand R, Pinto AJ, Figueroa LA, Sharp JO. (2016). Draft Genome Sequence of a Novel Desulfobacteraceae Member from a Sulfate-Reducing Bioreactor Metagenome. *Genome Announc* **4**.

Benson BJ, Magnuson JJ, Jensen OP, Card VM, Hodgkins G, Korhonen J, *et al.* (2012). Extreme events, trends, and variability in Northern Hemisphere lake-ice phenology (1855–2005). *Climatic Change* **112**:299–323.

Bernard SM, Habash DZ. (2009). The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. *New Phytol* **182**:608–620.

Elser JJ, Bracken MES, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, *et al.* (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol Lett* **10**:1135–1142.

Elser JJ, Marzolf ER, Goldman CR. (1990). Phosphorus and nitrogen limitation of phytoplankton growth in the freshwaters of north america: A review and critique of experimental enrichments. *Can J Fish Aquat Sci* **47**:1468–1477.

Esteves-Ferreira AA, Inaba M, Fort A, Araújo WL, Sulpice R. (2018). Nitrogen metabolism in cyanobacteria: metabolic and molecular control, growth consequences and biotechnological applications. *Crit Rev Microbiol* **44**:541–560.

Fukunaga Y, Kurahashi M, Sakiyama Y, Ohuchi M, Yokota A, Harayama S. (2009). Phycisphaera mikurensis gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. *J Gen Appl Microbiol* **55**:267–275.

García-Gutiérrez Á, Cánovas FM, Ávila C. (2018). Glutamate synthases from conifers: gene structure and phylogenetic studies. *BMC Genomics* **19**:65.

Grzymski JJ, Dussaq AM. (2012). The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J* **6**:71–80.

Guildford SJ, Hecky RE. (2000). Total nitrogen, total phosphorus, and nutrient limitation in lakes and oceans: Is there a common relationship? *Limnol Oceanogr* **45**:1213–1223.

Hampton SE, Galloway AWE, Powers SM, Ozersky T, Woo KH, Batt RD, *et al.* (2017). Ecology under lake ice. *Ecol Lett* **20**:98–111.

Hampton SE, Moore MV, Ozersky T, Stanley EH, Polashenski CM, Galloway AWE.

(2015). Heating up a cold subject: prospects for under-ice plankton research in lakes. *Journal of Plankton Research* **37**:277–284.

Harpole WS, Ngai JT, Cleland EE, Seabloom EW, Borer ET, Bracken MES, *et al.* (2011). Nutrient co-limitation of primary producer communities. *Ecol Lett* **14**:852–862.

He S, Stevens SLR, Chan L-K, Bertilsson S, Glavina Del Rio T, Tringe SG, *et al.* (2017). Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled Genomes. *mSphere* **2**.

Hecht U, Oelmüller R, Schmidt S, Mohr H. (1988). Action of light, nitrate and ammonium on the levels of NADH- and ferredoxin-dependent glutamate synthases in the cotyledons of mustard seedlings. *Planta* **175**:130–138.

Holmes RM, Aminot A, Kérouel R, Hooker BA, Peterson BJ. (1999). A simple and precise method for measuring ammonium in marine and freshwater ecosystems. *Can J Fish Aquat Sci* **56**:1801–1808.

Kalyuzhnaya MG, Bowerman S, Lara JC, Lidstrom ME, Chistoserdova L. (2006). Methylotenera mobilis gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family Methylophilaceae. *Int J Syst Evol Microbiol* **56**:2819–2823.

Kao-Kniffin J, Woodcroft BJ, Carver SM, Bockheim JG, Handelsman J, Tyson GW, *et al.* (2015). Archaeal and bacterial communities across a chronosequence of drained lake basins in Arctic Alaska. *Sci Rep* **5**:18165.

Kozhov M. (1963). Lake baikal and its life. Springer Netherlands: Dordrecht

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**:5112–5120.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**:814–821.

Lea PJ, Miflin BJ. (2003). Glutamate synthase and the synthesis of glutamate in plants. *Plant Physiol Biochem* **41**:555–564.

Linz AM, Crary BC, Shade A, Owens S, Gilbert JA, Knight R, *et al.* (2017). Bacterial community composition and dynamics spanning five years in freshwater bog lakes. *mSphere* **2**.

Logue JB, Langenheder S, Andersson AF, Bertilsson S, Drakare S, Lanzén A, *et al.* (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J* **6**:1127–1136.

Matoh T, Takahashi E. (1982). Changes in the activities of ferredoxin- and NADH-glutamate synthase during seedling development of peas. *Planta* **154**:289–294.

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

Mehrshad M, Salcher MM, Okazaki Y, Nakano S-I, Šimek K, Andrei A-S, *et al.* (2018). Hidden in plain sight-highly abundant and diverse planktonic freshwater Chloroflexi. *Microbiome* **6**:176.

Moore MV, Hampton SE, Izmest'eva LR, Silow EA, Peshkova EV, Pavlov BK. (2009). Climate change and the world's "sacred sea"—Lake Baikal, siberia. *Bioscience* **59**:405–417.

Muro-Pastor MI, Reyes JC, Florencio FJ. (2005). Ammonium assimilation in cyanobacteria. *Photosyn Res* **83**:135–150.

Murphy J, Riley JP. (1962). A modified single solution method for the determination of phosphate in natural waters. *Anal Chim Acta* **27**:31–36.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75**:14–49.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, *et al.* (2016). vegan: Community Ecology Package.

O'Donnell DR, Wilburn P, Silow EA, Yampolsky LY, Litchman E. (2017). Nitrogen and phosphorus colimitation of phytoplankton in Lake Baikal: Insights from a spatial survey and nutrient enrichment experiments. *Limnol Oceanogr* **62**:1383–1392.

Rosenberg E. (2014). The Prokaryotes. 4th edition. Springer: New York.

Schindler DW. (1977). Evolution of phosphorus limitation in lakes. *Science* **195**:260–262.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**:7537–7541.

Sterner RW. (2008). On the phosphorus limitation paradigm for lakes. *Int Rev Hydrobiol* **93**:433–445.

Taylor BW, Keep CF, Hall RO, Koch BJ, Tronstad LM, Flecker AS, *et al.* (2007). Improving the fluorometric ammonium method: matrix effects, background fluorescence, and standard additions. *Journal of the North American Benthological Society* **26**:167–177.

Temple SJ, Vance CP, Stephen Gantt J. (1998). Glutamate synthase and nitrogen assimilation. *Trends Plant Sci* **3**:51–56.

Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ. (2001). Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnol Oceanogr* **46**:557–572.

Verpoorter C, Kutser T, Seekell DA, Tranvik LJ. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophys Res Lett* **41**:6396–6402.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**:5261–5267.

Welschmeyer NA. (1994). Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. *Limnol Oceanogr* **39**:1985–1992.

Wendt K. (2001). Determination of Nitrate/Nitrite in surface and wastewaters by flow injection analysis (low flow method). QuikChem® Method 10-107-04-1-J.

Weyhenmeyer GA, Blenckner T, Pettersson K. (1999). Changes of the plankton spring outburst related to the North Atlantic Oscillation. *Limnol Oceanogr* **44**:1788–1792.

Wong HL, Visscher PT, White RA, Smith D-L, Patterson MM, Burns BP. (2017). Dynamics of archaea at fine spatial scales in Shark Bay mat microbiomes. *Sci Rep* **7**:46160.

CHAPTER THREE

METAGENOME ASSEMBLED GENOMES OF NOVEL MICROBIAL LINEAGES FROM LAKE BAIKAL IN SUMMER AND WINTER SEASONS

Abstract

Microorganisms are the essential agents of biogeochemical cycling in aquatic systems. Recent -omic advances uncovered their immense taxonomic diversity and functional repertoire. However, the connection between taxonomy and function remains elusive. This is in large part due to the narrow phylogenetic breadth of sequenced microbial genomes, brought on by difficulties in cultivating microorganisms from natural, in particular oligotrophic, environments. Here we present 369 high quality draft metagenome assembled genomes (MAGs) from Lake Baikal, Siberia. They were binned from assemblies comprising 22 sites with wide spatial and depth coverage of the lake in summer, and two samples collected from shallow and deep waters in the winter season. Baikal is the world's most ancient, deep, and voluminous freshwater body. It is a biodiversity hotspot that we hope will contribute important evolutionary insights to genome collections. Our MAGs are culture-independent and span the archaea domain and 15 bacterial phyla, four of which have no previously sequenced representatives from the lake. Most genomes are small but with large variation. At the same time, the most stable, aseasonal, and resource poor sites in the Lake Baikal hypolimnion harbored the smallest genomes with remarkably little size variation. These results could reflect Baikal's overall oligotrophic environment, where millions of years allowed microorganisms to maximize occupancy of available resource niches. We hope this

report will set the stage for future model-based work on relationships between phylogenetic diversity and metabolic function in this and other natural systems.

Introduction

Microorganisms in aquatic systems play a key role in nutrient cycling and energy transfer. In the last two decades, environmental meta -omic studies revealed incredible taxonomic and functional diversity, inferred distributions of phylogenetic lineages, and associated gene abundances with environmental gradients. However, sequenced genomes of microorganisms commonly found in those environments remain the critical and often missing piece in understanding microbial ecology in a way that unifies information about environmental factors, metabolic strategies, evolutionary relationships, and biotic interactions. This is because biochemical processes take place inside living cells, where genes are organized into controlled metabolic pathways, and knowledge of gene duplication or loss in those pathways is critical in understanding their function. For these and other reasons, complete genomes are also necessary for acting as references in high-throughput environmental transcriptomic and proteomic studies of novel and poorly characterized natural systems. Thus, while bulk gene content in environmental samples provides insights on the metabolic potential of a community, complete genomes are needed for understanding of function, ecology, and evolution of cells.

Current technologies allow three main ways of obtaining complete microbial genomes: from cultured isolates, physically sorted single cell genomes (SAGs), and

using bioinformatic approaches to construct metagenome assembled genomes (MAGs) (Bowers *et al.*, 2017). Genomes obtained from isolates benefit from possible field addition experiments and controlled laboratory assays that can experimentally associate genes and regulatory elements with functional responses. However, lineages characteristic of oligotrophic environments (the "uncultivated majority") are notoriously difficult or impossible to culture (Bowers *et al.*, 2017; Margesin, 2017; Gilbert & Dupont, 2011). SAGs circumvent this limitation by using high throughput flow cytometry or dilution-to-extinction to isolate single cells. Genomes are then obtained after amplifying the single cell genomic content, most commonly with multiple displacement amplification (MDA). This culture-independent method has generated thousands of novel genomes from diverse environments, and has been used to characterize common freshwater lineages, including Actinobacteria (Garcia *et al.*, 2013; Kang *et al.*, 2017) and Verrucomicrobia (Martinez-Garcia *et al.*, 2012). An additional advantage of SAGs is the certainty in the absence of contamination. However, SAGs generally suffer from poor completion rates, and amplification biases introduced by MDA can skew information on gene copy numbers and miss some elements altogether.

MAGs are an effective alternative. They are generated by processing a metagenomic assembly of contigs and grouping some of those contigs into bins, based on matching tetranucleotide composition and coverage-based contig co-occurrence patterns across samples (Kang *et al.*, 2015). The bins are then evaluated for completion and contamination based on presence of an expected set of single copy markers, and those that pass a set threshold are then considered MAGs (Parks *et al.*, 2015). The

quality of MAGs has rapidly improved with recent developments in assembly and binning computational tools, surpassing completion rates of most SAGs (Hugerth *et al.*, 2015; Bowers *et al.*, 2017; Parks *et al.*, 2017). It is important to keep in mind that each MAGs is an average of a population. While this warrants caution in interpreting nucleotide polymorphisms in the averaged features, mapping of short reads from multiple environments onto the MAG reference can reveal which polymorphisms across short reads are candidate local adaptations within a population represented by a MAG. More MAGs have recently been obtained from different environments (Parks *et al.*, 2017), many important ecosystems still lack adequate characterization of the functional diversity of microbial communities, such as Lake Baikal and other ancient freshwater lakes.

Lake Baikal, Siberia is the world's deepest (1642 m), largest (volume 23,000 km^3), and oldest (25-30 mya) lake, holding approximately 20% of the world's unfrozen freshwater (Moore *et al.*, 2009). Likely owing to its size and age, Baikal is an island of biodiversity and endemism; of the roughly 2600 plant and animal species that inhabit the lake, approximately two-thirds are not found anywhere else (Moore *et al.*, 2009; Kozhov, 1963; Soma *et al.*, 2001), earning it a place among UNESCO world heritage sites.

Lake Baikal spans diverse environments. Its water column gets briefly stratified in late July to mid August and has two major phytoplankton bloom periods: the late winter and early spring season, and during summer stratification. In winter, the under-ice planktonic community remains active, and a dense layer of diatoms and other algae

develops at the ice-water interphase (Osipova *et al.*, 2010; Bondarenko *et al.*, 2012; Katz *et al.*, 2015). Although historically cold year-round, the lake waters are experiencing rapid warming, accompanied by a decline in endemic zoo- and phytoplankton species and a rise of cosmopolitan competitors (Moore *et al.*, 2009; Hampton *et al.*, 2008; Katz *et al.*, 2015). However, although Baikal's zoo- and phytoplankton have been monitored for over 70 years, relatively little is known of its bacterioplankton community diversity and function. The microbial plankton has been investigated with 16S amplicon sequencing (Denisova *et al.*, 1999; Kurilkina *et al.*, 2016). However, studies based on the 16S rRNA gene carry a PCR amplification bias and only provide information on the taxonomic makeup without directly capturing functional genes (but see Chapter One and Chapter Two).

Here we present 369 high quality MAGs from Lake Baikal, assembled and binned using shotgun sequencing across 24 samples, collected in summer and winter under the ice. Our sampling plan was designed to maximize coverage of environmental diversity and includes all three Baikal basins, a 0-500 m depth profile of the Southern Basin, the eutrophic Proval Bay and Selenga River plume, mesotrophic Barguzin Bay and Maloe More, and a eutrophic-oligotrophic gradient from inner to outer Chivirkuy Bay. In addition, two (out of 24) samples were collected from Southern Basin in the winter from under the ice at 0 and 250 m. We expand phylogenetic diversity of sequenced genomes from Lake Baikal by approximately ten fold, including 15 MAGs in 4 phyla with no to-date sequenced representatives from the lake. We show that the majority of genomes are small, but the the high size variability suggests possible niches

for generalist and specialist taxa. We also emphasize that the hypolimnion stands out as harboring by far the smallest and least variable in size MAGs, possibly reflecting the most constant conditions of any of the sampled environments.

Materials and Methods

We sequenced 24 samples using two HiSeq lanes. Many of the DNA samples contained low concentration of DNA, and the alternative low input DNA prep kit was used. Libraries were prepared using the Rubicon Genomics ThruPLEX DNA-Seq Kit, following the manufacturer's protocol. After completion of library prep whey were QC'd and quantified using a combination of Qubit dsDNA HS, Caliper LabChip HS DNA and Kapa Biosystems Illumina Quantification qPCR assays. Libraries were pooled in equimolar proportions for multiplexed sequencing. This pool was loaded on two lanes of an Illumina HiSeq 2500 Rapid Run flow cell (v2) and sequencing was carried out using HiSeq SBS reagents (v2) in a 2x250bp format. Base calling was done by Illumina Real Time Analysis (RTA) v1.18.64 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v1.8.4. All computational steps were performed on the Michigan State University High Performance Computing Cluster (HPCC).

Raw read assessment

Each of the two HiSeq 2500 lanes produced a forward R1 and a reverse R2 read for each sample. The the two R1 reads were concatenated, as were the two R2 reads, producing one R1 and one R2 collection of short reads per sample. Raw read

properties, including distribution of read lengths, per base position quality, and presence of adapters, were assessed with FastQC package.

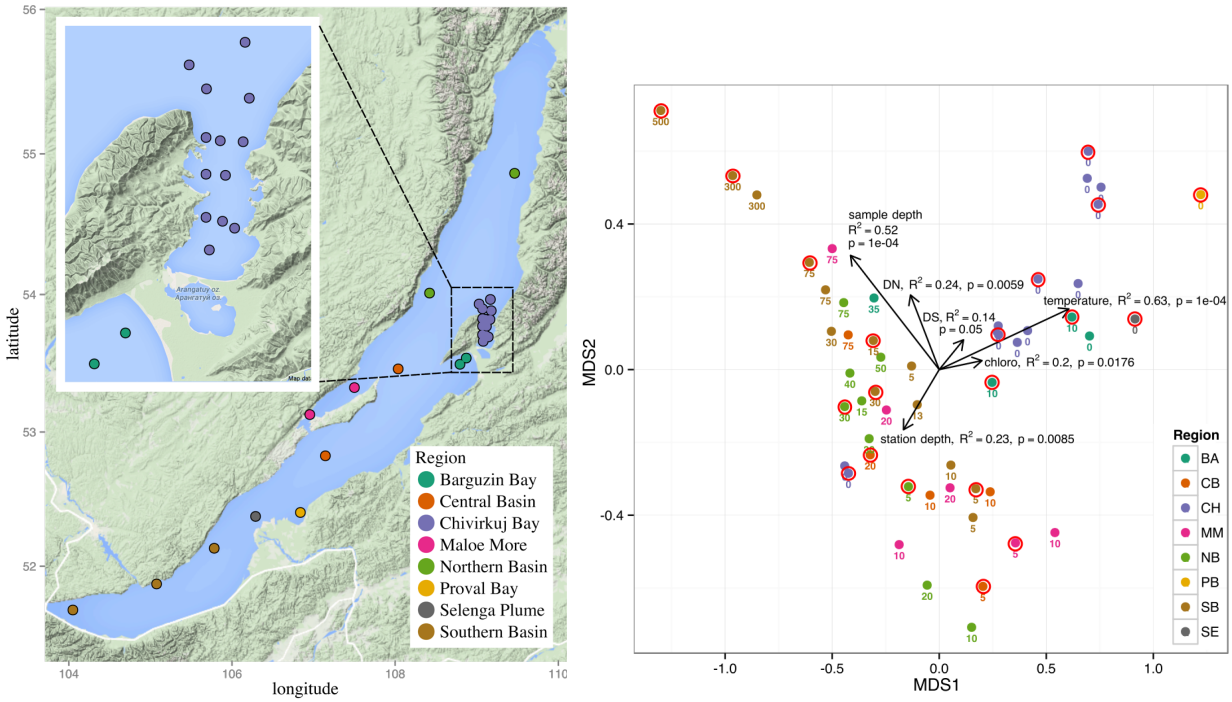


Figure 3.1: Sampling stations (left), with multiple depths per station. The choice of shotgun samples was motivated by the amplicon survey in Chapter 1 (right). Numbers below samples indicate depths (m).

Trimming and QC of Raw Sequences

Trimming of the Illumina adapters is required to avoid confusing the assembler. Additional options aim to improve overall sequence quality. A common approach is to use a sliding window to track specified base quality, trim when necessary, and keep the sequence if its final length exceeds a cutoff value. We used Trimmomatic to remove adapters, short reads, and reads with poor quality sliding windows. Other programs will do this, too; however, Trimmomatic has speed and low memory use to its credit.

Trimming parameter values dictate how many sequences we discard, where lower values of Q are more permissive and higher ones more stringent. We explored a few options and ultimately chose the lowest cutoff value Q10 to make sure we had enough sequences to provide adequate coverage (see Results). FastQC was used again to check post-trimmed sequences.

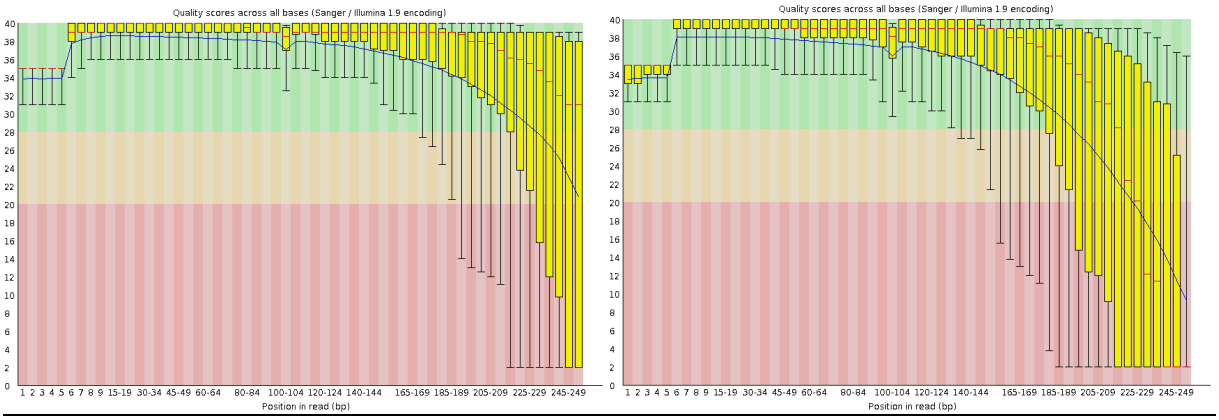


Figure 3.2: Raw sequence length distribution.

Minhash diagnostics for sample co-assembly

Based on a number of trial assemblies involving individual samples, preliminary contigs suffered from low coverage. This was due to the shortage of reverse reads that did not pass the trimming step, even when using a relatively permissive quality cut-off (see above). Pooling multiple samples for co-assembly can increase coverage of the assembled contigs. However, if co-assembled samples are drawn from sources with substantially dissimilar community composition, excessive variation in the to-be-assembled features the can result in few features reaching consensus and therefore failing to assemble. The solution is to identify groups of samples with similar enough composition to warrant pooling. In order to achieve this, we compared sample sequence

composition using sourmash software, developed by Dr. Titus Brown. It first decomposes each dataset into k-mers of a specified size and then performs principal coordinate analysis based on pairwise Jaccard dissimilarities (Ondov *et al.*, 2016). We chose k77 and k99 for sourmash k-mer parameters because three trial assemblies of individual samples using the metaSPAdes assembler (see methods below) had the greatest N50 at k77 but the greatest number of long contigs at k99. Results suggested pooling our 24 samples into twelve groups of sample pairs (Fig. 3.4).

Assembly

A number of research groups are actively developing competing assemblers (e.g. metaVelvet, megahit, metaSPAdes), each one offering advantages in ease of use, speed, low memory usage, or quality of results. For preliminary trials, as shown in later sections, we used megahit for its speed. For final assemblies, we used metaSPAdes because it is known to produce longer contigs, at the expense of speed and memory footprint, with the options listed below. Assembly evaluation was done with metaQUAST.

MetaSPAdes assembler options

- `-12` and `-s`: use both paired and unpaired reads.
- `--meta` option changes scoring and penalty schemes.
- `-k 21,33,55,77,99,127` directs to include more k-mer length to assemble. The manual recommends this approach with low coverage datasets.

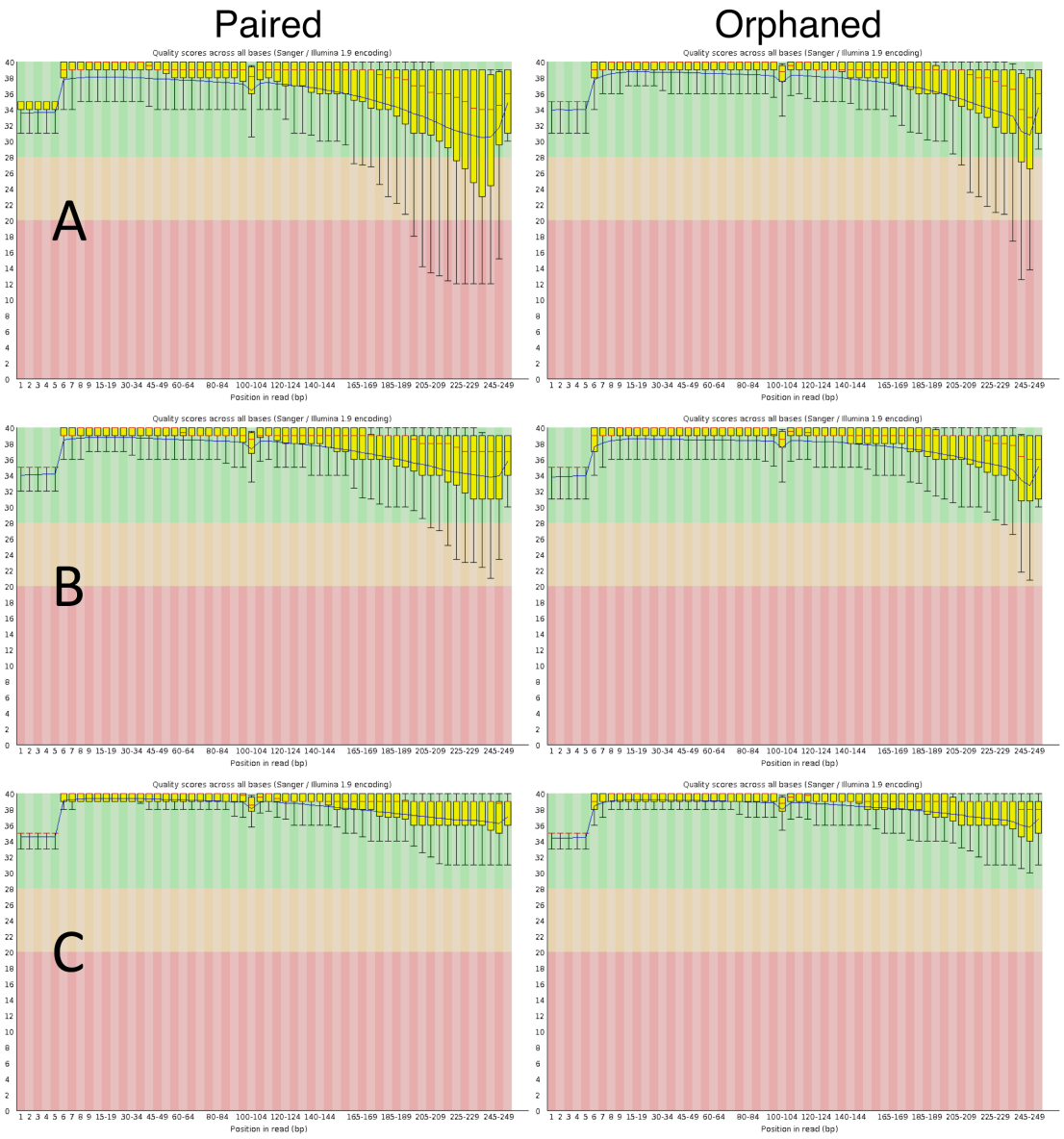


Figure 3.3: Sequence length distribution after filtering using: (A) q10 and 50 minimum length, (B) q20 and 50 minimum length, and (C) Q30

Short read mapping

Alignment of short reads onto each of the twelve assemblies was done with bowtie2 in local mode using the very-sensitive threshold. The default end-to-end alignment was designed to work with single genomes, not considering SNPs. However, when dealing

with metagenomic datasets, local alignment is preferable because it allows multiple feature variants to exist on different contigs.

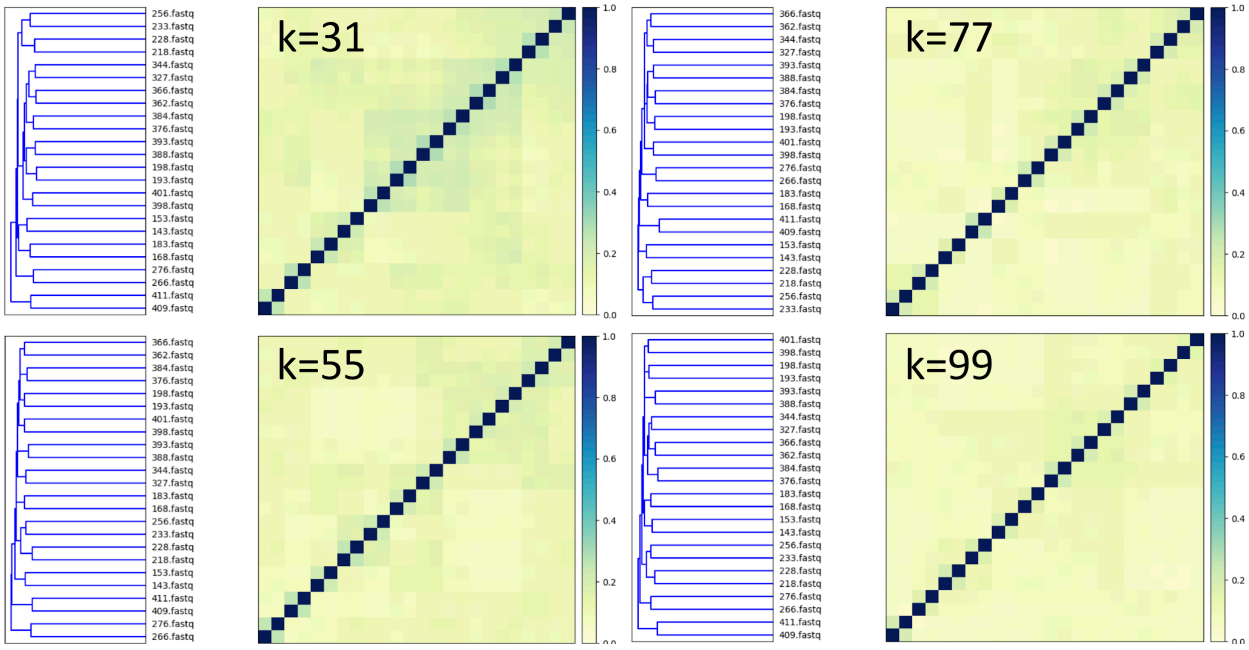


Figure 3.4: Similarity of samples, based on k-mer decomposed short reads. Specific pairs of samples showed consistent similarity at all investigated k-mer values.

Binning

We used the metaBAT2 software suite, maintained by Lawrence Berkeley National Laboratory, which matches tetranucleotide composition and contig co-occurrence patterns to create contig bins (Kang *et al.*, 2015). It has been previously shown that different microbial genomes have distinct tetranucleotide composition biases (Saeed *et al.*, 2012; Mrázek, 2009; Teeling *et al.*, 2004; Pride *et al.*, 2003). MetaBAT has calculated empirical intra- and inter-species frequency distances by analyzing 1,414 unique, completed bacterial genomes from the NCBI database. Contig co-occurrences were determined based on coverage of each contig from the previous step. Information

on tetranucleotide composition and contig co-occurrences is then combined for each contig pair, and putative bins are formed iteratively from the distance matrix.

Checking bins for completeness and contamination

We used CheckM software (Parks *et al.*, 2015) to assess bin completion and contamination. For each putative contig bin, CheckM first uses HMM searches to identify a minimum 10 out of the 43 phylogenetically informative markers, as identified by Brown et al (Brown *et al.*, 2015), to place a bin into a reference phylogenetic tree, which is based on the Lawrence Berkeley database, using pplacer (Matsen *et al.*, 2010). Next, for each bin checkM uses its placement in the reference tree and the gene composition of the 2052 finished IMG reference genomes, which are also part of the reference tree, to determine a set of lineage-specific single copy markers expected to be contained in the putative MAGs. A suitable set of marker genes for assessing a genome includes the distance of the putative MAG relative to the surrounding reference genomes and the amount of variation in these genomes. The degree of completion is simply the number of single copy markers present in a bin divided by the number expected. Contamination results from multiple copies of the same single-copy marker present in the same bin.

MAG tree construction

HMMer 3.1.2b was used to search for 139 bacterial single copy core genes, identified by Campbell *et al.*, within each genome (Campbell *et al.*, 2013). To create this

collection, Campbell et al. analyzed the occurrence of protein families in 1516 complete bacterial genomes distributed around bacterial tree of life, and identified protein families that occurred only once in at least 90% of all genomes. Then, MUSCLE package was used to align each detected example of all core genes (Edgar, 2004). For each genome, the final alignment of all of the detected and aligned single copy core genes of the possible 139 were then concatenated together. Gblocks was then used to trim the alignment to remove universal gaps (Talavera & Castresana, 2007). This trimmed and concatenated alignment served as the final alignment which was used to build the phylogenomic tree. Fasttree was then used to construct an approximate-maximum likelihood tree, using the JTT model of protein evolution and the CAT model of site-varying evolution (Price *et al.*, 2010).

Results and Discussion

We assembled 369 high quality bacterial and archaeal genomes from Lake Baikal, comprising four 100 percent complete genomes with zero contamination, 16 additional genomes at >99% completion and 0-2.5% contamination, with the rest at >80% completion and <10% contamination. Their phylogeny spans 15 phyla, of which four phyla have never been described in Baikal. Owing to the high quality of our assemblies and the abundance of coverage-based contig co-occurrence information available across 24 samples (Fig. 3.1), even our lowest quality MAGs were more than twice as complete as those reported previously by Cabello-Yeves *et al.* (Cabello-Yeves *et al.*,

2018), while maintaining low contamination levels, which were not accounted for in the aforementioned study.

Sequence quality filtering

The HiSeq 2500 produced 749,587,488 raw paired-end reads. Although the output was within specifications, quality assessment (Fig. 3.2, Table 3.1) indicated that the reverse run had substantially lower base quality at the longer end of the reverse reads. This is a common indicator of polymerase degradation on the reverse run. However, by the time the deviation was discovered, the sequencing center no longer retained the DNA samples for possible re-sequencing. The resulting shortage of reverse reads dictated the downstream bioinformatic choices.

We used a permissive quality cutoff at the trimming step to maximize the amount of available data. The FastQC sequence quality control software suggests three threshold levels at Q10, Q20, and Q30, corresponding to low, medium, and high quality (Fig. 3, Table 3.1). However, having high quality inputs is becoming less relevant because modern assemblers, including megahit and metaSPAdes used here, integrate base quality in their scoring and penalty schemes (Li *et al.*, 2015; Bankevich *et al.*, 2012). At the same time, the most stringent cutoff at Q30 resulted in letting through only 13.47% of the original reads (Table 3.1). To further test the effects of available data on assembly quality, we co-assembled all samples with megahit using data from the four different trimming levels shown in Table 3.1. Assembly using the lowest cutoff level (Q10) produced the largest overall assembly with longest maximum length contig and

an N50 that was comparable to assemblies using Q20 datasets (Table 3.2). Assembly using Q30 cutoff data was inferior by all metrics. Lastly, the final deciding factor in using Q10 data instead of Q20 was making more short read data available for mapping to the assembly. A greater range of possible feature coverage provides increased statistical power for detection of enriched or depleted features and associating them with environmental conditions. Likewise, a larger range of contig coverage values is important for calculating accurate contig co-occurrences in contig binning procedures.

Table 3.1: Raw and trimmed sequence statistics. Quality statistics for four Trimmomatic quality cutoff thresholds. The most stringent cutoff at Q30 had the lowest percentage of passing reads.

quality cutoff	minlen	raw R1+R2	passed paired	passed orphaned	% passed paired	% passed total
10	50	749,587,488	432,210,262	315,498,176	57.66	99.75
20	50	749,587,488	392,794,042	340,852,470	52.4	97.87
20	91	749,587,488	377,131,618	167,332,423	50.31	72.64
30	135	749,587,488	68,394,718	32,576,450	9.12	13.47

Minhash k-mer decomposition and diagnostics

To maximize coverage in our assemblies, we identified samples with similar k-mer composition that could be effectively pooled together for co-assembly. The initial step requires decomposition of short reads from each sample into k-mers of determined length, and the best k-mer length is that which will be used for the co-assemblies. To cover the range of possible k-mer lengths, we used sourmash to decompose samples at

k-mer 33, 55, 77, and 99, generate a pairwise Jaccard dissimilarity matrix for each k-mer decomposition, run principal coordinate analysis on each matrix, and construct a hierarchical tree of sample dissimilarities (Fig. 3.4).

Table 3.2: Preliminary co-assemblies using all samples with sequences filtered at various cutoff schemes. Results indicated that sequences filtered at Q10 and Q20 produced similar results. We chose to proceed with Q10 to maximize coverage.

trim	megahit ver	k-mer	#contigs	total bp	max length	N50
Q10	1.1.1	k141	8,637,252	6,034,799,935	268,764	794
Q10	1.3	k141	8,640,384	6,034,582,651	244,377	793
Q20	1.1.1	k141	7,950,994	5,588,730,087	198,422	808
Q20_k91	1.1.1	k91	7,900,803	5,541,374,810	249,687	804
Q30	1.1.1	k99	8,448,240	4,610,404,357	114,443	607

At all k-mer lengths, the lowest level sample pairs clearly showed the greatest pairwise correlation, compared with higher levels, which showed little clustering, as indicated by shallow branch lengths (Fig. 3.4). For smaller k-mer values, samples showed overall greater pairwise correlation scores. The smaller k-mer value heatmaps also revealed additional larger areas of sample similarity correlation. For example, at k=31 there there is a detectable large block of 8 samples between samples 384 and 398 (Fig. 3.4, k=31). However, cautious interpretation is advised for these correlation blocks at low k-mer values because shorter k-mers are more likely to give false positives, as is mentioned in the sourmash manual.

Based on these results, we chose to co-assemble in 12 groups of sample pairs, as identified by sourmash hierarchical clustering (Fig. 3.4, all k-mers), for two reasons.

First, branch lengths indicated a strong pairwise correlation signal. Second, this was a robust approach because the lowest level pairs were consistent between all investigated k-mer values, in contrast with variable higher level hierarchies. In addition, the clustering reflected biological meaning. For first example, although at all k-mer lengths, samples 409 and 411 appear as outgroups, and they were the only two Winter samples in our entire dataset. These two samples comprised assembly Group 12 (Table 3.3, Table 3.4). For second examples, samples 384 and 376 (Group 10, Table 3.3, Table 3.4) were clustered as pairs at all k-mer values, and they were the deep water samples, collected at 300 and 500 meters in the Summer.

Assembly and Coverage

We used metaSPAdes (Bankevich *et al.*, 2012) to assemble contigs from 12 groups of paired samples (Table 3.3) and bowtie2 to map short reads from each sample onto each of the 12 assemblies. In this process, we aimed to maximize two metrics: contig length and short read coverage. While assembly quality in many respects varied between groups, we achieved the goal of N50>1kb for every sample group assembly. This was biologically meaningful because average bacterial genes are roughly 1kb in length, and N50>1kb meant that most contigs in each of our assemblies were longer than an average gene.

We used the -very-sensitive-local option in bowtie2 (Langmead & Salzberg, 2012), which increased accuracy of short read alignment at the cost of computational resources and time. It was important to maximize coverage information for the contigs

Table 3.3: Comparison of final metaSPAdes assemblies for the twelve grouped sample pairs.

Assembly Group	Total length	Total length (>= 50000 bp)	Largest contig	# contigs (>= 50000 bp)	N50	L50
1	761,274,632	363,43,909	363,770	390	1,284	118,234
2	760,276,931	18,014,213	323,679	227	1,298	122,582
3	846,614,893	19,542,259	330,799	242	1,250	146,058
4	756,660,347	21,085,508	301,636	262	1,345	116,081
5	676,115,055	21,625,468	301,770	261	1,398	95,819
6	725,831,599	32,246,503	462,599	373	1,502	94,013
7	824,889,360	19,290,219	365,495	241	1,190	148,647
8	861,341,705	15,669,593	244,004	197	1,101	182,195
9	740,759,481	12,441,279	237,269	162	1,207	135,339
10	590,193,777	6,420,465	274,265	81	1,144	117,228
11	1,199,453,315	13,792,993	679,166	185	1,187	228,282
12	598,244,224	13,117,212	305,660	161	1,066	124,331

because contig binning partly relies on contig co-occurrences, and those co-occurrences are calculated based on their differential coverage coverage across different samples. Short read mapping showed a variable number of reads from each sample recruited by the assemblies. As expected, reads from samples which were part of the co-assembled groups in the first place aligned best with those assemblies at 77-96%. Groups 12 and 10 produced overall smallest assemblies, with the shortest maximum contigs, lowest N50, and they recruited the lowest number of reads across many samples. This is interesting because group 10 comprised deep water samples

collected at 500 and 300 meters, and group 12 was composed of the only two Winter samples, collected at 5 and 250 meters. It is important to note that the amplicon survey in Chapter One established that deep water samples in Baikal had the highest Pielou Evenness. Is it possible that the high evenness effectively diluted the sequencing effort for group 10, spreading coverage across many similarly distributed species and therefore hindering deep sequencing of any of particular organism or any particular gene feature, preventing assembly of long continuous stretches.

Binning and bin refinement into mags

We used metabat2 (Kang *et al.*, 2015), which processes tetranucleotide composition and coverage-based contig co-occurrence information to assign contigs into bins. Each group produced between 590 and 611 bins. Next, we used CheckM software (Parks *et al.*, 2015) to assess bin completion and contamination (see Methods) and discard bins below 80% completion and over 10% contamination. Only about 5% of the original bins passed this cutoff, and those that did were renamed MAGs. Each assembly produced between 26 and 44 high quality MAGs (Table 3.4), with the notable exception of group 10, which yielded 13 MAGs, likely due to the relatively lower quality of the group 10 assembly (Table 3.3). Nonetheless, all MAGs were composed of notably long contigs, indicated by large mean N50 values for each group (Table 3.4), which were more than an order of magnitude larger than those for the bulk contigs (Table 3.3). The number of predicted genes indicated coding density in the 0.9-0.96 range. This was noteworthy

Table 3.4: Summary properties of final MAGs (>80% completion, <10% contamination), evaluated for each group. The deep water Group 10 had the lowest number of MAGs, which were, in turn, on average composed of the shortest contigs, possibly directly contributing to having the least mean number of predicted genes per MAG.

Group	# Bins	Complet. Mean (CV)	Contam. Mean (CV)	Genome Size Mean (CV)	Contig N50 Mean (CV)	Predicted Genes Mean (CV)
1	34	92.8 (0.05)	2 (1)	2,546,612 (0.49)	42,583 (0.9)	2,495 (0.46)
2	39	90.1 (0.07)	1.9 (1.05)	2,261,156 (0.46)	26,064 (0.83)	2,215 (0.38)
3	35	91.9 (0.05)	2.1 (0.95)	2,518,436 (0.39)	27,225 (0.75)	2,466 (0.34)
4	35	88.8 (0.07)	1.6 (0.62)	1,938,059 (0.35)	27,011 (0.73)	1,945 (0.32)
5	34	90.5 (0.08)	1.8 (1.11)	2,013,669 (0.37)	32,005 (0.68)	2,007 (0.34)
6	44	91.5 (0.07)	1.9 (1.05)	2,213,990 (0.4)	34,763 (0.91)	2,162 (0.35)
7	26	88.5 (0.07)	2.3 (0.87)	2,068,900 (0.37)	32,096 (0.81)	2,037 (0.33)
8	27	87.9 (0.07)	1.8 (0.56)	1,946,160 (0.36)	23,692 (0.8)	1,962 (0.31)
9	26	88.9 (0.07)	1.8 (0.56)	1,884,547 (0.35)	21,701 (0.61)	1,888 (0.29)
10	13	85.6 (0.05)	2 (0.5)	1,562,828 (0.21)	17,742 (0.31)	1,576 (0.18)
11	30	90.5 (0.07)	2.8 (0.71)	2,739,712 (0.51)	22,798 (0.97)	2,729 (0.44)
12	26	89.6 (0.07)	1.6 (0.62)	2,353,846 (0.4)	25,432 (0.72)	2,262 (0.36)

because it reflected the expected bacterial coding density (85-95%) (Land *et al.*, 2015), further validating our MAGs as high quality draft genomes.

Abundance of small genomes in an oligotrophic environment

Most of the reconstructed genomes reported here are smaller than average (Giovannoni *et al.*, 2014); however, the substantial variation in genome sizes could reflect large physiological diversity of microbiota in Lake Baikal. While most genomes were indeed small (approximately 1.5-2.5 Mb), MAGs from each assembly group, with the exception of Group 10, showed a notable second mode at around twice the size of their respective group mean (Fig. 3.5). For Groups 1-9 and 12, the larger MAGs averaged approximately 4 Mb, and even larger at about 6 Mb for group 11. Such large differences between the two genome size categories (CV approx. 0.4 across groups) was not explained by variation in completeness (CV approx. 0.06 across groups). The MAG sizes were also not correlated with contamination levels in any group. Because small genomes are usually associated with streamlined metabolisms, often seen in oligotrophic environments (Swan *et al.*, 2013; Lynch, 2006; García-Fernández *et al.*, 2004; Giovannoni *et al.*, 2005; Tripp *et al.*, 2010), we suggest that the majority of Baikal microbiota are specialists, adapted to an environment with low resource diversity. However, the large MAGs could indicate presence of microbial generalists. Of particular interest is the absence of the large MAG mode in Group 10, which comprised samples from Lake Baikal hypolimnion, collected at 500 and 300 m. This group also contains the smallest MAGs, potentially indicating that the constant conditions of the hypolimnion of

the world's most ancient lake have effectively excluded microbial generalists, while selecting for small genomes that specialize on processing the few available resources.

New MAG phylogenetic lineages

In this report we substantially expand knowledge of phylogenetic diversity of sequenced microorganisms from Lake Baikal, based on alignments of 139 phylogenetically informative core genes, identified by Campbell *et al.*, within each MAG (Campbell *et al.*, 2013). Our MAG phylogeny comprises the Archaea domain, which we used as root (outgroup) and 15 bacterial phyla. The most represented bacterial phyla included Actinobacteria, Verrucomicrobia, Bacteroidetes, and Betaproteobacteria, reflecting diversity captured in our 16S V4 amplicon survey (Chapter One). However, MAGs revealed 13 lineages of Archaea that have been overlooked by the amplicon survey, possibly due to PCR primer amplification biases. We also detected numerous deep clades within every phylum, which were missed by Cabello-Yeves *et al.* (2018), including four entire phyla, e.g., Chloroflexi, Gammaproteobacteria, Ignavibacterium, and the superphylum group CPR, which do not have any previously reported representatives. This was likely a consequence of our study assembling approximately ten times the number of MAGs as the previous study. For example, there appear to be 16 distinct taxa in Betaproteobacteria (Fig. 3.6), of which Cabello-Yeves *et al.* (2018) only detected one, and similar ratios of currently reported to known lineages can be made for other phyla. Of the newly characterized lineages, Chloroflexi, Ignavibacteria and CPR are all poorly understood and are largely uncultured.

We did not detect the Chloroflexi CL-500-11 in the current study. Chloroflexi members have been found in diverse environments, including oxygenated (its type genus) and anoxic (Overmann, 2001) hot springs. However, of particular interest is the CL500-11 clade (class Anaerolineae), where numerous studies cast it as a key constituent of oxygenated hypolimnia of diverse lakes on global scale (Mehrshad *et al.*, 2018), including Crater Lake (Urbach *et al.*, 2001, 2007), Lake Biwa (Okazaki *et al.*, 2013) and the Laurentian Great Lakes (Denef *et al.*, 2015; Fujimoto *et al.*, 2016). However, our results indicate CL500-11 levels below detection limits of the current study. This warrants discussion because we are reporting the most comprehensive metagenomic survey of Lake Baikal microbiota to date. These results are also consistent with our amplicon survey in Chapter One, where three Chloroflexi 16S V4 amplicon OTUs were detected across 54 investigated samples in Lake Baikal, but two OTUs were from the Chloroflexi class and the remaining OTU was from the Roseiflexales order, and all appeared to prefer shallower and warmer environments in the lake. Furthermore, we could not find evidence of CL500-11 in Lake Baikal in any other study. Although a meta-study by He *et al.* (2017) included a figure indicating a small (<2%) presence of CL500-11 in Baikal at 0 m and 20 m, it does not include a reference in the main text or the Supplement. The only to-date published Lake Baikal study to include samples from exclusively 0 and 20 m is that of Cabello-Yeves *et al.*, and the authors are missing the entire Chloroflexi phylum altogether (Fig. 3.6). Thus, we must emphasize a conspicuous absence of CL500-11 in Lake Baikal, including summer epilimnion, summer hypolimnion and winter under-ice samples, collected from 0 and

250 meters. This is intriguing, considering the high abundance of the CL500-11 Chloroflexi clade in oxygenated hypolimnia of other lakes, documented by numerous other independent studies. We speculate that sources of nitrogen could differ between Baikal and other lakes. Indeed, genomic reconstruction of CL500-11 by *Denef et al*. (2015) showed that the clade preferentially relies on catabolism on high molecular weight, nitrogen-rich compounds. However, in contrast to the established paradigm, where lakes are generally considered P-limited, Baikal is N and P co-limited, at least in the summer season (O'Donnell *et al.*, 2017). In such system, N-rich compounds may be scarce and thus not favor a specialist, such as CL500-11. It is our hope that future field and laboratory studies will help to further quantify the mechanism for CL500-11 environmental preferences.

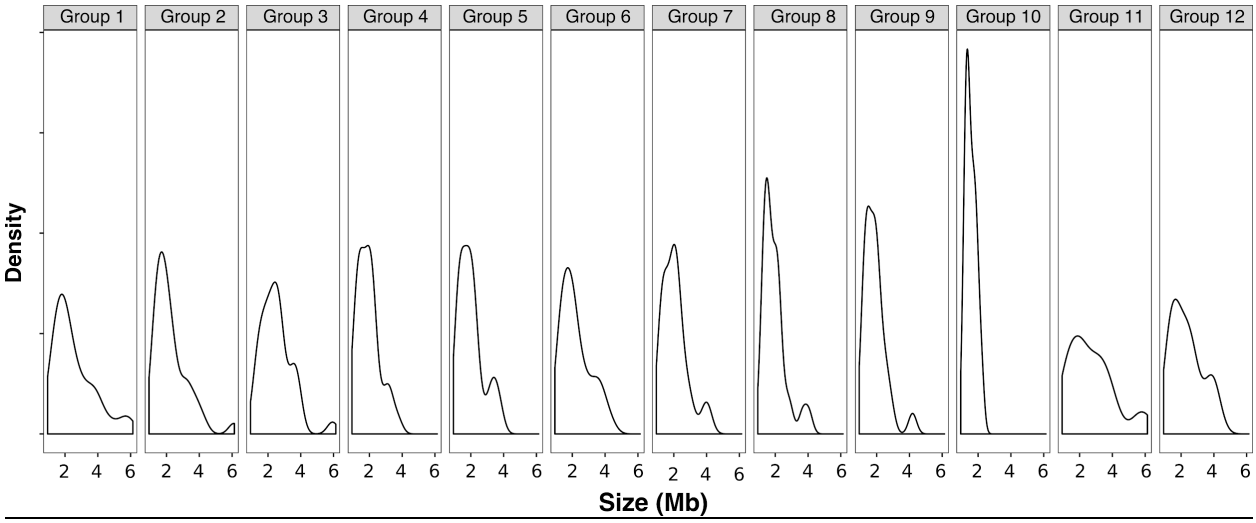


Figure 3.5: MAG genome size distribution was skewed to the left with a smaller second mode at larger values in every group, except Group 10.

We also report diverse lineages of Actinobacteria and Verrucomicrobia MAGs, which agrees with the high abundance of both phyla in Baikal in our amplicon study (Chapter

One). Actinobacteria are some of the most abundant aquatic microbial taxa (Newton *et al.*, 2011). Recent genomic studies reveal that members of the most typical freshwater clade acI have particularly streamlined genomes, often unable to synthesize certain vitamins, amino acids, and reduced sulphur compounds (Neuenschwander *et al.*, 2018).

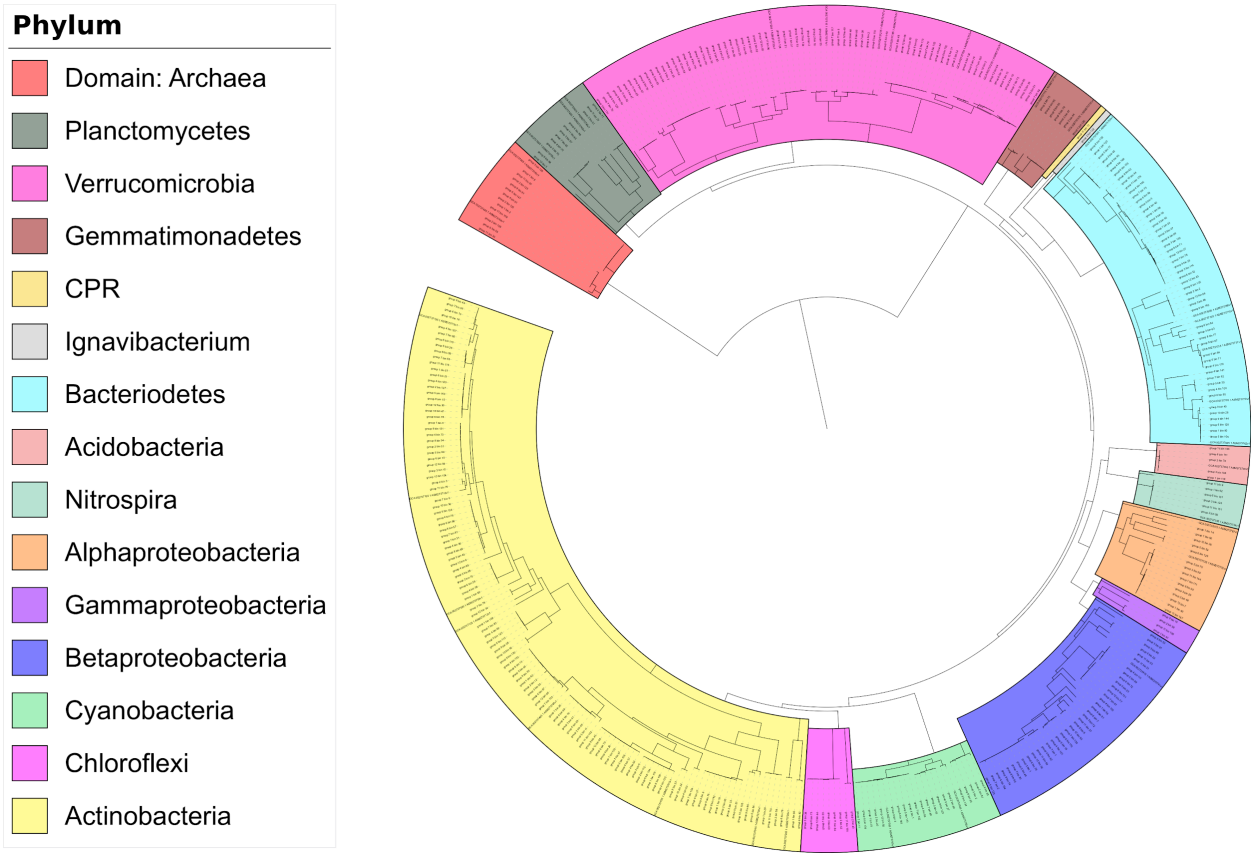


Figure 3.6: Phylogenetic tree of our MAGs, combined with those reported by Cabello-Yeves *et al.* (2018), indicated by the longer labels at tree tips. Phylogeny key starts with the Archaea domain and proceeds clockwise.

These deficiencies are not only enabled but can be evolutionarily favored where certain "leaky" common goods unavoidably become available in the shared environment. What results are increased dependencies on co-occurring organisms. The evolutionary

mechanism for creating these scenarios is known as the Black Queen hypothesis (Morris *et al.*, 2012). Previous genomic analyses of Verrucomicrobia revealed their potential role as polysaccharide degraders in freshwater systems (Martinez-Garcia *et al.*, 2012), but with significant differences between functional profiles, possibly reflecting differences in processing autochthonous (internally derived) and terrestrial carbon (He *et al.*, 2017). The functional potential of novel taxa in Baikal could be investigated in comparative studies with publicly available genomes.

Conclusions and future directions

Whether taxonomically diverse microorganisms pursue different metabolic strategies in response to environmental conditions, such as temperature, light, and nutrient levels, has been a subject of many large scale analyses and discussion (Louca *et al.*, 2016; Martiny *et al.*, 2015). Reconstructed high quality MAGs from Lake Baikal (369 in total) provide a perspective on the phylogeny, metabolism, and distribution of 15 Phyla, including 4 previously uncharacterized phyla in the world's most ancient freshwater system, unraveling remarkable genomic diversity of pelagic freshwater MAGs in the summer and winter seasons. Although the high level phylogenetic makeup resembles many freshwater environments, at least one taxon hailed as hallmark (CL500-11) is missing from the hypolimnion. We further report the small size of most constructed MAGs, including those with high completion and low contamination rates, and highlight the variability and bimodality of MAG size distributions in all environments, except arguably the most aseasonal and ancient - the Lake Baikal hypolimnion. These results

will lead to further investigations in comparative phylogenetics, functional genomics, and ecology.

Future directions in phylogenetics will certainly include curation by collapsing shallow MAG clades, especially if those clades were assembled from different sample groups e.g. the 10 clades shown in Fig. 3.7. The curated MAGs will need to be compared with more publicly available sequenced genomes from multiple environments, such as the Baltic Sea (Hugerth *et al.*, 2015) and other environments (Parks *et al.*, 2017), for more comprehensive inference on their worldwide distribution and possible specificity to Lake Baikal.

Future functional analyses will focus on genomic content of Lake Baikal MAGs. One goal would be detection of various signatures of cold adaptation, including GC content, protein isoelectric points, and enrichment of functional genes and regulatory elements previously associated with success in cold environments (Rodrigues *et al.*, 2009; Bakermans *et al.*, 2009). An additional approach would determine intra-population variability of genomic features within each MAG. Because MAGs are population averages, alignment of short reads will identify polymorphism sites. The resulting SNP analyses from the different sampled sites in Lake Baikal could associate variability in genomic features with differences in environmental conditions.

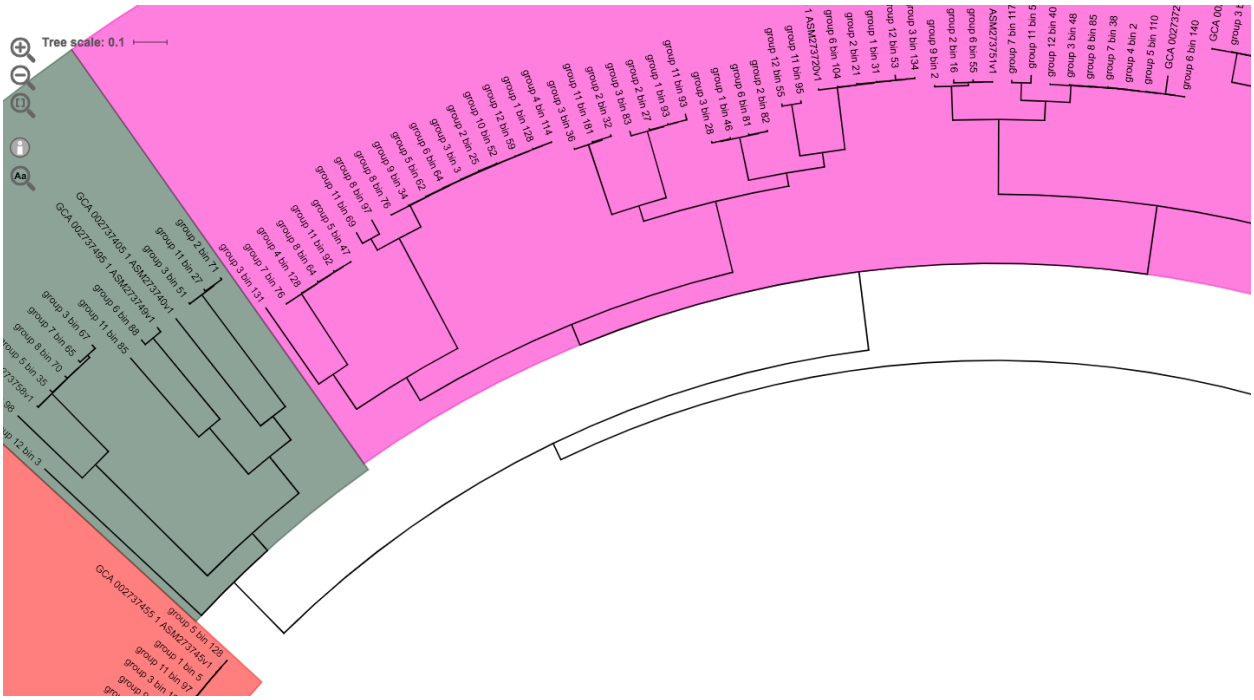


Figure 3.7: MAGs within shallow clades that come from different assembly groups are candidates for merging.

Future ecological modeling will use single-copy markers, such as *recA,* to determine coverage-based abundances of MAGs across sampled sites in Baikal. Then, the knowledge of MAGs genomic content can be combined with measured environmental conditions across sampled sites to model MAG abundances with genes and environment as predictors. Such ecological models have been used as tools for calculating the so-called fourth corner matrix, which identifies genes-by-environment interactions that significantly impact MAG abundance (Brown *et al.*, 2014). We believe this approach promises significant advances, especially in the framework of traits and abiotic factors. By quantifying the relationships between traits and different environments, it answers the question how microbes, in an ancient and diverse place like Lake Baikal, use genes to process the environment to increase their fitness.

etermining a set of “fitness relevant genes” gets us one step closer to understanding the context-dependent basis for distribution of life on earth.

REFERENCES

REFERENCES

Bakermans C, Sloup RE, Zarka DG, Tiedje JM, Thomashow MF. (2009). Development and use of genetic system to identify genes required for efficient low-temperature growth of Psychrobacter arcticus 273-4. *Extremophiles* **13**:21–30.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**:455–477.

Bondarenko NA, Belykh OI, Golobokova LP, Artemyeva OV, Logacheva NF, Tikhonova IV, *et al.* (2012). Stratified distribution of nutrients and extremophile biota within freshwater ice covering the surface of Lake Baikal. *J Microbiol* **50**:8–16.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, *et al.* (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**:725–731.

Brown AM, Warton DI, Andrew NR, Binns M, Cassis G, Gibb H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods Ecol Evol* **5**:344–352.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, *et al.* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**:208–211.

Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV, *et al.* (2018). Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. *Appl Environ Microbiol* **84**.

Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, *et al.* (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci USA* **110**:5540–5545.

Denef VJ, Mueller RS, Chiang E, Liebig JR, Vanderploeg HA. (2015). Chloroflexi CL500-11 Populations That Predominate Deep-Lake Hypolimnion Bacterioplankton Rely on Nitrogen-Rich Dissolved Organic Matter Metabolism and C1 Compound Oxidation. *Appl Environ Microbiol* **82**:1423–1432.

Denisova LI, Bel'kova NL, Tulokhonov II, Zaĭchikov EF. (1999). [Diversity of bacteria at various depths in the southern part of Lake Baikal as detected by 16S rRNA sequencing]. *Mikrobiologiia* **68**:547–556.

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

Fujimoto M, Cavaletto J, Liebig JR, McCarthy A, Vanderploeg HA, Denef VJ. (2016). Spatiotemporal distribution of bacterioplankton functional groups along a freshwater estuary to pelagic gradient in Lake Michigan. *J Great Lakes Res* **42**:1036–1048.

Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R, *et al.* (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J* **7**:137–147.

García-Fernández JM, de Marsac NT, Diez J. (2004). Streamlined regulation and gene loss as adaptive mechanisms in Prochlorococcus for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev* **68**:630–638.

Gilbert JA, Dupont CL. (2011). Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* **3**:347–371.

Giovannoni SJ, Cameron Thrash J, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* **8**:1553–1565.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**:1242–1245.

Hampton SE, Izmest'Eva LR, Moore MV, Katz SL, Dennis B, Silow EA. (2008). Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia. *Glob Chang Biol* **14**:1947–1958.

He S, Stevens SLR, Chan L-K, Bertilsson S, Glavina Del Rio T, Tringe SG, *et al.* (2017). Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled Genomes. *mSphere* **2**.

Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, *et al.* (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol* **16**:279.

Kang DD, Froula J, Egan R, Wang Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**:e1165.

Kang I, Kim S, Islam MR, Cho J-C. (2017). The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. *Sci Rep* **7**:42252.

Katz SL, Izmest'eva LR, Hampton SE, Ozersky T, Shchapov K, Moore MV, *et al.*

(2015). The “Melosira years” of Lake Baikal: Winter environmental conditions at ice onset predict under-ice algal blooms in spring. *Limnol Oceanogr* **60**:1950–1964.

Kozhov M. (1963). Lake baikal and its life. Springer Netherlands: Dordrecht

Kurilkina MI, Zakharova YR, Galachyants YP, Petrova DP, Bukin YS, Domysheva VM, *et al.* (2016). Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. *FEMS Microbiol Ecol* **92**.

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, *et al.* (2015). Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**:141–161.
Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**:357–359.

Li D, Liu C-M, Luo R, Sadakane K, 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.

Louca S, Parfrey LW, Doebeli M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**:1272–1277.

Lynch M. (2006). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* **60**:327–349.

Margesin R (ed). (2017). Psychrophiles: from biodiversity to biotechnology. Springer International Publishing: Cham

Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, *et al.* (2012). Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. *PLoS ONE* **7**:e35314.

Martiny JBH, Jones SE, Lennon JT, Martiny AC. (2015). Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**:aac9323.

Matsen FA, Kodner RB, Armbrust EV. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**:538.

Mehrshad M, Salcher MM, Okazaki Y, Nakano S-I, Šimek K, Andrei A-S, *et al.* (2018). Hidden in plain sight-highly abundant and diverse planktonic freshwater Chloroflexi. *Microbiome* **6**:176.

Moore MV, Hampton SE, Izmest’eva LR, Silow EA, Peshkova EV, Pavlov BK. (2009).

Climate change and the world's "sacred sea"—Lake Baikal, siberia. *Bioscience* **59**:405–417.

Morris JJ, Lenski RE, Zinser ER. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**.

Mrázek J. (2009). Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol* **26**:1163–1169.

Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. (2018). Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J* **12**:185–198.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75**:14–49.

Okazaki Y, Hodoki Y, Nakano S. (2013). Seasonal dominance of CL500-11 bacterioplankton (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol Ecol* **83**:82–92.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, *et al.* (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**:132.

Osipova SV, Bondarenko NA, Obolkina LA, Permyakov AP, Dudareva LV, Timoshkin OA. (2010). Communities in the ultrapure ice of Lake Baikal: distinctive characteristics and adaptive strategies of ice inhabitants. In:*20th IAHR International Symposium on Ice*, Vol. 1.

Overmann J. (2001). Green Nonsulfur Bacteria. In:*Encyclopedia of life sciences*, John Wiley & Sons, Ltd (ed), John Wiley & Sons, Ltd: Chichester, UK.

O'Donnell DR, Wilburn P, Silow EA, Yampolsky LY, Litchman E. (2017). Nitrogen and phosphorus colimitation of phytoplankton in Lake Baikal: Insights from a spatial survey and nutrient enrichment experiments. *Limnol Oceanogr* **62**:1383–1392.

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

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, *et al.* (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**:1533–1542.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2--approximately maximum-likelihood

trees for large alignments. *PLoS ONE* **5**:e9490.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**:145–158.

Rodrigues DF, da C Jesus E, Ayala-Del-Río HL, Pellizari VH, Gilichinsky D, Sepulveda-Torres L, *et al.* (2009). Biogeography of two cold-adapted genera: Psychrobacter and Exiguobacterium. *ISME J* **3**:658–665.

Saeed I, Tang S-L, Halgamuge SK. (2012). Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* **40**:e34.

Soma Y, Tanaka A, Soma M, Kawai T. (2001). 2.8 million years of phytoplankton history in Lake Baikal recorded by the residual photosynthetic pigments in its sediment core. *Geochem J* **35**:377–383.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**:11463–11468.

Talavera G, Castresana J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**:564–577.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**:938–947.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

Urbach E, Vergin KL, Larson GL, Giovannoni SJ. (2007). Bacterioplankton communities of Crater Lake, OR: dynamic changes with euphotic zone food web structure and stable deep water populations. *Hydrobiologia* **574**:161–177.

Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ. (2001). Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnol Oceanogr* **46**:557–572.