An integrated understanding of microbial roles in biogeochemical cycling in anoxic lakes

By

Patricia Q. Tran

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Freshwater and Marine Science)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 05/15/2023

The dissertation is approved by the following members of the Final Oral Committee:

Karthik Anantharaman, Assistant Professor, Bacteriology
Katherine D. McMahon, Professor, Bacteriology, Civil and Environmental Engineering
Emily H. Stanley, Professor, Integrative Biology
Hilary A. Dugan, Assistant Professor, Integrative Biology
Sarah P. Preheim, Assistant Professor, Environmental Health & Engineering, Johns Hopkins University (external)

Abstract

An integrated understanding of microbial roles in biogeochemical cycling in anoxic lakes

By Patricia Q. Tran

Microbial communities are made up of microscopic organisms: viruses, bacteria, archaea, eukaryotes, and fungi. Their ability to exist under various and sometimes extreme conditions allows them to thrive in all corners of the environment. Aquatic environments cover 70% of the earth's surface and much of it is undergoing drastic human-impacted change. Rises in atmospheric temperature are leading to an increase in surface water temperatures, which leads to a cascade of ecological impacts, including oxygen depletion through physical and biological means. In this dissertation, I demonstrate how microbial communities, particularly bacteria, archaea, and phages (viruses that can impact prokaryotes) impact biogeochemical cycling in anoxic lakes. Using holistic approaches that cover both cultivation-dependent and independent methods, I aim to expand the way we approach studying microbial communities by leveraging interdisciplinary strengths to demonstrate that microbes and phages are interacting members that are highly dynamic in time and space.

In **Chapter 1** (Introduction), I summarize the role of freshwater lakes in biogeochemical cycling by focusing on the transformations that occur in the water column. For context, I compare and contrast some of the physical features of Lake Tanganyika and Lake Mendota, the two study sites in this work. I then give a broad summary of what is known about microbial and viral function in anoxic lakes. Together, these form the rationale for the next chapters. In **Chapter 2**, I used physiological and genomic evidence to characterize bacteria that were able to produce hydrogen sulfide under oxic conditions, a process typically associated with dissimilatory sulfate reduction

which occurs under anoxic conditions. This work demonstrates and expands the spatiotemporal scope of hydrogen sulfide sources and sinks in the environment. In **Chapter 3**, I used genome-resolved metagenomics to characterize the potential contribution of bacteria and archaea in Lake Tanganyika, a permanently anoxic lake that happens to be one of the world's deepest and oldest lakes. This revealed that the anoxic hypolimnion of Lake Tanganyika had an extremely high proportion of Archaea, and endemic microorganisms, compared to other freshwater lakes worldwide, giving insight into the interplay between long-term anoxia and evolution. In **Chapter 4**, I provide a commentary that argues for a holistic way to study biogeochemical cycling: from an organismal and methodological perspective. In **Chapter 5**, I apply this holistic framework to assess the impact and interactions between phage and prokaryotes on biogeochemical cycling in a seasonally anoxic freshwater lake. We found that the bacterial community was sensitive to deoxygenation but the viral community was not. A broad range of bacterial taxa were infected by phages, but phages were highly specific. Phage-impacted bacteria were active in methane, sulfur, and nitrogen metabolism. Finally, time-series phage activity data showed the dynamic impact of phage-host interactions for nutrient cycling.

Overall, the knowledge and framework generated in these studies improve our understanding of complex multi-kingdom species interactions and their associations with microbiology, ecology, and biogeochemistry in anoxic environments.

Acknowledgments

There are many people I would like to thank, for having helped me along the way to accomplish this dissertation.

I would like to thank my co-advisors Karthik Anantharaman and Trina McMahon. Thank you both for your unwavering support in my research and professional development. I appreciate that I could make my PhD school experience what I wanted it to be. In fact, it is thanks to you offering me to join your lab that I was able to expand my horizons scientifically and professionally to this degree. Never would I have imagined myself doing a PhD, let alone moving to another country to do it. I learned so much in the past 4.8 years. Thanks for providing a space where I could grow as a person and scientist, and for supporting me (intellectually, financially, etc.) in research, celebrating wins, being patient with rejections, and encouraging my extracurriculars!

I would like to thank my dissertation committee members. I never entered or came out of a committee meeting afraid, but rather felt grateful for the opportunity for us all gather for a few hours to have a meaningful discussion about research. Thank you Emily Stanley for including me in LTER and being supportive of my Lake Mendota research, as early as during my internship in 2018. You introduced me to a community of scientists on campus, which made me excited to join UW-Madison. Thank you Hilary Dugan for the meaningful discussions and advice about research, science, and careers over the years. Thank you Sarah Preheim for joining my committee from afar, I am glad I got to meet you in person in Washington and Switzerland last year. During the last months of writing this dissertation, I often thought about your strategies for writing.

The joy of being co-advised is to be part of two amazing labs: to all members past and present, thanks for your encouragement and support, for sharing your time and energy helping me on projects, and for the thoughtful discussions over a variety of topics. My MSB lab desk

neighbors: I loved that we definitely had the chattiest office, full of energy, music, and ideas sharing. Thanks for all the discussions and coffee breaks!

Some of my colleagues became close friends during my time in Madison. Taylor, I am glad to have connected with such as a thoughtful, honest, and creative person. You showed me the importance of community. I really enjoyed spending time with you all outside of the lab. Kathryn you were such as supportive friend and I loved going to pottery with you. Rachel, I loved climbing with you in Madison, Devil's Lake and in Switzerland, I am happy we went through the PhD journey with similar timelines, and congrats on your recent defense! Coty and Diana, I am glad we lived on the same street and shared a beautiful garden, I enjoyed the sunny days walking to the grocery store and having bubble tea. Maggie, I enjoyed sharing great discussions over a good dinner. Elise you are such an impressive gardener, I recall that the flowers you gave me brought me so much joy. Adrianna, Emily, and Linnea, my friends whom I met in the CFL lunchroom – I value our good times and friendships. Ojaswee, for the dancing, going out and dining and going out!

Finally, I would like to thank my parents Thi Tuyet Hong Do and Quy Bình Tran (who are really mẹ và ba) and my sister Sylvie. Words cannot express my gratitude and love throughout my life. I am always excited to come to visit you and I hope that after graduating I will be able to do it more often and spend celebratory life events by your side. Sylvie thanks for being the best sister ever, for your friendship, giving the best coffee suggestions and your fun attitude. Thank you for your supportive words during phone calls, and for sending me cards - I love looking at them when I need encouragement. I am so glad you can come to visit me in Madison for my dissertation defense. And of course, I would like to thank my partner Max, for your support and love during all these years! I am fortunate to able to enjoy life with you and I'm excited for what's next!

Table of Contents

Abstract *i*

Acknowledgments *iii*

Chapter 1 Introduction *1*

Lakes, anoxia, and their significance **1**

Lake origins **3**
Rift Valley lakes 3
Glacial retreat lakes 4
Physical characteristics and anoxia in these lakes 5

Microbial ecology in freshwater lakes **6**

Overall dissertation motivation **7**

Chapter 2 Physiological and genomic evidence of cysteine degradation and aerobic hydrogen sulfide production in freshwater bacteria *8*

Abstract **9**

Importance **10**

Introduction **10**

Results **15**
Isolates capable of H_2S production in oxic conditions. 15
Detailed microbiological, chemical and genomic characterization in selected isolates 15
Presence of cysteine-degrading organisms and genes in a five-year metagenomic environmental time-series 22

Discussion **24**
The fate of hydrogen sulfide and ammonia, two products of cysteine degradation 24
Genomic structure of the H_2S-producing isolates 25
Challenges associated with measuring oxic H_2S production from organosulfur in the environment 27
Implications of oxic H_2S production by microbes in freshwaters 28

Methods **31**
Enrichment cultures of isolates from a temperate freshwater lake 31
Screening for cysteine degradation into H_2S and ammonia 31
Identification of H_2S producing bacteria using 16S rRNA gene sequencing 32
Detailed characterization of 3 hydrogen-sulfide producing isolates 33
Methods to measure cysteine 34
Methods to measure H_2S using a microsensor 35
Generation of metagenome-assembled genomes 36
Searching for cysteine genes and isolates presence in metagenomic time-series and MAGs 36

Data availability **37**

Acknowledgments **37**

Chapter 3 Depth-discrete metagenomics reveal the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. ***39***

Abstract **40**

Introduction **41**

Methods **42**
Sample collection 42
DNA extraction and sequencing 45
Metagenome assembly, genome binning, dereplication, and selection of genomes 45
Relative abundance across water column depths and gene annotations 46
Identification of phylogenetic markers 46
Phylogenetic tree 47
Taxonomic assignment and comparison of manual vs. automated methods 48
Support for Tanganyikabacteria, a monophyletic sister lineage to Sericytochromatia 49
Comparison of MAG taxonomic diversity in LT vs. LB 50
Metabolic potential analysis and comparison of metabolic potential and connection across three distinct depths in LT 50

Results **51**
The microbiome of Lake Tanganyika 52
How similar is the LT microbiome to that of other deep and ancient freshwater lakes? 59
Depth-dependent contrasts in microbial metabolism in LT: biogeochemical cycling of carbon, nitrogen, and sulfur 62

Discussion **67**
Comparison of microbiomes between two deep rift valley lakes, LT and Baikal: endemism vs. shared lineages 68
Metabolic connections across vertical gradients in LT 69

Conclusion **72**

Data availability **73**

Chapter 4 Biogeochemistry Goes Viral: Towards a Multifaceted Approach To Study Viruses and Biogeochemical Cycling. ***75***

Abstract **76**

The importance of viruses in aquatic biogeochemistry **76**

Techniques to study the roles of the environmental viruses **78**

Looking forward: Combining multifaceted approaches is important to obtain a holistic understanding of ecosystem ecology **81**

Chapter 5 Viral impacts on microbial activity and biogeochemical cycling in a seasonally anoxic freshwater lake ***85***

Abstract **86**

Introduction **86**

Results **89**
Microbial community composition and expression change as a function of anoxia, but abundance does not correlate to activity 91
Viral community composition 95
Comparison of viruses in "total metagenomes" vs. "viromes" 96
Phage activity and impact on nutrient cycles 99
Phage-host interaction diversity and specificity 101
Phages impact biogeochemical processes driven by active microorganisms 103
Phage-host abundance and expression dynamics 105

Discussion **107**
Increased microbial diversity over time and space is not directly correlated to activity 108
Viral impacts on biogeochemistry 109
Collecting viromes in conjunction with total metagenomes allows for recovering more diversity of phages 111
Careful interpretation of phage AMG and their actual activity is needed 113

Methods **114**
Field sampling 114
DNA extraction 115
RNA extraction 115
Biogeochemical field data collection and laboratory analysis 116
DNA sequencing 117
RNA sequencing and processing 118
Metagenome assembly, binning, and MAG quality filtering 118
MAG naming convention 120
Assessing genome coverage and transcript abundance per MAG 121
Assessing the metabolic potential of the MAGs 121
Phylogenetic tree 122
Diversity Index Calculations 122
Viral identification and AMG annotation 122
Phage-host interactions 123
Taxonomic assignment of phages 123

Data availability **123**

Acknowledgments **124**

Chapter 6 Conclusions and Perspectives ***125***

Concluding remarks **125**

Implications for global microbial ecology **125**
Global Patterns of Biodiversity in Deep Lakes 125
Lake Mendota as a model ecosystem 126
Additional samples archived and available to be used for future projects. 126
Environmental change and responses across multi-kingdom microbial community 128
Prioritizing depth-discrete hypolimnetic sampling 128

Anoxic and oxic-anoxic interface hypolimnion microbial community 129

Cryptic Biogeochemical Cycling 130

Environmental viruses: detection, ecology, and implications 131
Viral detection 131
Viral ecology 132
Applications in public health 133

References (whole thesis) 135

Appendix 164

Table of Figures

Figure 2.1 Overview of physiological and genomic methods to characterize hydrogen sulfide production from isolates from freshwater lakes. 14
Figure 2.2 Quantitative characterization of cysteine utilization and hydrogen sulfide production in the three freshwater bacterial isolates using high-performance liquid chromatography. 16
Figure 2.3 Genes involved in sulfur and organosulfur key pathways in the three isolates. 19
Figure 2.4 Common and distinguishing features of the three cysteine-degrading isolates. 21
Figure 2.5 Distribution of cysteine desulfurization genes across time and taxa. 24
Figure 2.6 Distribution of cysteine desulfurization genes across time and taxa. 30
Figure 3.1 Map of sampling locations in Lake Tanganyika. 44
Figure 3.2 Phylogeny of Archaeal and Bacterial metagenome-assembled genomes (MAG) recovered from LT. 54
Figure 3.3 Concatenated gene phylogeny of Cyanobacteria and their non-photosynthetic sister-lineages, and comparison of average nucleotide identities. 56
Figure 3.4 Bar plot showing the relative read abundance (RAR) per sample in the 24 metagenomes from Lake Tanganyika. 58
Figure 3.5 Comparison of metagenome-assembled genomes in Lake Tanganyika and Lake Baikal. 60
Figure 3.6 Microbial contributions to carbon, nitrogen and sulfur cycling in Lake Tanganyika 63
Figure 3.7 Summary figure showing the role of different microbial taxa in carbon, nitrogen, and sulfur cycling in Lake Tanganyika. 67
Figure 4.1 Interactions between microbial members and scales of interactions. 77
Figure 4.2 Conceptual framework for maximizing information about viral ecology and biogeochemistry in nature. 82
Figure 5.1 Samples and environmental data. 90
Figure 5.2 Diversity and the top 10 MAGs ranked by abundance, expression, and overexpression. 96
Figure 5.3 Non-metric dimension scale plot of (A) the prokaryotic community (B) and the viral community. 98
Figure 5.4 Comparison of viruses based on dataset origin. 100
Figure 5.5 Phage and AMG expression, and examples of phage-host relationships. 102
Figure 5.6 Methane, nitrogen, and sulfur metabolism and gene expression of 55 MAGs with at least one identified phage. 104
Figure 5.7 Phage-host abundance and activity patterns of Cyanobacteria and Thiobacillaceae hosts and their phages. 106
Figure 6.1 Filters collected during this dissertation. Some have been extracted for this dissertation, and many are stored as archived for future projects. 127

Chapter 1 Introduction

Lakes, anoxia, and their significance

Freshwater lakes are a limited resource for all life forms. Lakes account for 0.009% of the total water on Earth[1]. But despite lakes having a small volume compared to the total water on Earth, freshwater has an outsized influence for humans and the environment. Lakes provide drinkable water, form the habitat for significant wildlife and inland fisheries, and provide places for recreation. Lakes are broadly defined as a body of water surrounded by land. They receive, transform, and circulate input from the terrestrial landscape, such as organic and inorganic matter entering the lake. This matter may be transformed within the lake, released into the atmosphere, percolate into deep groundwater, or flow out into the environment making their way to streams, rivers, and ultimately the oceans. Lakes have often been referred to as "biogeochemical hotspots", because their contribution to biogeochemistry is disproportionally high, and they are not passive systems from which water flows, but rather a dynamic system[2].

Additionally, lakes are sometimes referred to as "sentinels" for climate change, as physical, chemical, and biological indicators that respond to environmental changes[3]. The effects of environmental changes have been observed in historical data, and scientists are predicting its effects on a range of lake types. In a recent review, lake surface water temperatures have increased by around 0.5ºC per decade, whereas deep hypolimnetic temperatures have been stable. However, dissolved oxygen concentrations have decreased in both surface and deep waters[4]. While various

degrees of oxygen depletion are natural in freshwater ecosystems, anoxic zones are becoming increasingly prominent globally[4]. For example, about a quarter of 100 lakes currently classified as monomictic (mix once a year) are predicted to become permanently stratified, and a sixth of dimictic lakes will become monomictic[5]. Practically speaking, a decrease in mixing frequently means that less oxygen will be replenished into the water column. The general mechanism behind this phenomenon is that increased water surface temperatures globally result in greater density differences between warm top waters and cold hypolimnetic waters, making whole water column mixing more difficult. As a result, oxygen from wind mixing in top waters does not diffuse down. Additionally, increased nutrient inputs and warmer surface temperatures are prone to algal and bacterial blooms, which augment oxygen consumption in the waters. However, there is substantial different in individual lake responses.

In many lakes, mostly those with dimitic regions, these oxic-anoxic regime shifts are important for biogeochemical processes[6]. In these anoxic environments or low-oxygen environments, microbes are central to recycling nutrients. For example, while other animals require oxygen, and plants require light, microbes possess unique abilities to harness inorganic carbon and alternative electron acceptors for respiration beyond oxygen enabling them to live and thrive in dark, anoxic, and low nutrient environments. Chemical compounds such as carbon, nitrogen, sulfur and phosphorus compounds are transformed and distributed throughout the water column. In anoxic environments, microbes turn sulfate into toxic hydrogen sulfide, carbon dioxide and nitrate into methane and nitrous oxide, respectively, which are greenhouse gases. Upon water column mixing, these chemical compounds are either further transformed (uptaken) by other organisms or released into

the atmosphere. In the atmosphere, these compounds can travel hundreds to thousands of miles, thereby having a global impact.

Lake origins

This dissertation contains research from two study sites: Lake Tanganyika and Lake Mendota. The two lakes discussed are formed through two major geological events, i.e. created through tectonic activity (e.g., rift valley lakes) or erosion (e.g. glacial erosion).

Rift Valley lakes

Rift valleys form in a lowland region where the Earth's Crust drifts apart. Their marine counterparts are known as Mid Ocean Ridges, which include the East Pacific Rise in the Pacific Ocean for example and are sites of active hydrothermal activity. On land, these rifts are called "Continental Rifts" and form continental rift valleys. Rift valleys can be categorized as active or inactive. The East African Rift is a geological formation in a geologically active area. known as the Afar Triple Junction involving 3 tectonic plates, the Nubian, Somali and Arabian plates. As freshwater flooded the system, a series of important lakes named the East African Great Lakes, which includes Lake Tanganyika were formed. Along the East African Rift are a series of volcanoes, for example Erta Ale (Ethiopia), Mount Kenya (Kenya) and Mount Kilimanjaro (Tanzania).

A much younger lake, Lake Superior, was formed on an inactive rift (the Mid Continental Rift) that stopped diverging about 80 million years ago[7]. Lake Baikal, a lake mentioned in this dissertation as a point of comparison to Lake Tanganyika (**Chapter 3**), is also a lake formed

through rift valley formation (Baikal Rift Zone) and sits atop of a 25-million-year-old continental rift. Lake Baikal is the world's deepest lake and followed by Lake Tanganyika. Lake Tanganyika is an ancient Rift Valley Lake, its shoreline is shared between Tanzania (East), Burundi (North), the Democratic Republic of Congo (West), and Zambia (South). Geologically, it originates from the diversion of the tectonic plates, is the world's longest lake at 419.2 miles, and a depth of 1,436m. Lake Tanganyika is an important source of inland fisheries. It is home to several endemic species of fishes[8].

Glacial retreat lakes

About 30,000 years ago, before the last glaciation, the region now called Madison used to be a river valley. As the glaciers advanced, they deposited material. Following this, about 18,000 years ago, the glaciers retreated and left depressions in the landscapes, which then became lakes. As the glaciers melted, meltwater filled these depressions with water.

Table 1.1 Comparison of Lake Tanganyika and Lake Mendota

	Lake Tanganyika	Lake Mendota
Geological history	Rift valley lake	Glacial retreat
Estimated age	9 to 12 million years	18,000 years
Average depth (m)	570	12.80
Maximum depth (m)	1, 470	25.29
Length (km)	673	9.04
Width (km)	72	6.61

Surface area (km^2)	32 900	39.58
Catchment area (m^2)	231,000	562
Hydrologic Residence time (years)	5500 Hypolimnion: ~2000[9] Epilimnion:~7[9]	~4[10]
Water mixing regime	Meromictic (does not mix)	Dimitic (mixes twice per year)

Physical characteristics and anoxia in these lakes

Lake mixing regimes are yearly patterns of water column mixing. Factors that influence lake mixing include climate and lake bathymetry. Events like wind mixing can also cause partial or complete water column mixing. Biological causes, such as changes in the food web are associated with increased anoxia in lakes. For example, Lake Mendota is a seasonally anoxic lake, but the length of the anoxic period has lengthened by about 10 days after 2008, following the introduction of an invasive species of zooplankton[11].

Lake Tanganyika, is a permanently anoxic lake[12]. In such anoxic lakes that are typically stratified, greenhouse gases can accumulate in the hypolimnion. Rare events, called limnic eruptions, can happen when CO_2 under high pressure (as is the case in deep lakes) is dissolved, but can quickly be released under when the water is unstable. This was the case in Lake Nyos for example, in 1986, where thick layers of CO_2 gas suddenly released into the atmosphere and resulted in major health impacts[13]. In addition to CO_2, other gases can accumulate in the anoxic hypolimnion. For example, Lake Kivu also houses greenhouse gases in the hypolimnion, such as methane, and while there has

been economic pressure for “deep-lake” gas mining, there remain concerns about potential environmental disasters[14,15].

As the length of anoxic periods increases globally, much of our ability to predict its effects rely on a good understanding of current dynamics: what compounds exist in the current anoxic hypolimnion and how are they being processed?

Microbial ecology in freshwater lakes

Microbes (or microorganisms) are small life forms invisible to the naked eye. Ranging from sub-micron sizes to millimeters, microbes span all forms of life including bacteria, archaea, eukaryotes, and viruses. Microbes play key roles in the environment, whether that is for human health, or transforming compounds. Bacterial metabolism is versatile: while animals require oxygen to live, bacteria can live in the complete or partial absence of oxygen. Instead, they are able to use alternative electron acceptors to respire (e.g. nitrate or sulfate). In addition, many compounds are hard to degrade, or “recalcitrant”[16]. Special enzymes, encoded by genes on the DNA of microbes enable them to use these compounds, and further cycle them through the food web. This forms the basis of a concept called the microbial loop, which means that microbes transform compounds and make them available to the food web[17]. Long-term, large-scale spatial studies of lake microbiomes have revealed the core taxa of freshwater lakes, which include organisms from the groups Actinobacteria and Cyanobacteria, for example[18–25]. As a whole however, however relatively few studies have focused on anoxic lakes despite their strong redox gradients and importance to the environment and biogeochemistry.[6,26–29].

Overall dissertation motivation

This dissertation aims to advance empirical and conceptual ways of investigating microbial communities in natural environments:

Theme 1. Biogeochemical cycling: Traditionally, carbon (C), phosphorus (P), and nitrogen (N) are studied in isolation. However, at the ecosystem level, these have contributed to global biogeochemical transformations. At the microbial level, organisms are often not only carbon users or generators but also dependent on sulfur (S), P, and N to form essential proteins and DNA. Classical functional definition such as decomposers, phototrophs, and denitrifiers do not encompass the microbial complexities and interactions that happen in real-time in natural ecosystems. As such, we ask: what would it look like if microbial members were classified outside the boundaries of singular nutrients, and what is the role of complex microbial communities?

Theme 2. Microbial communities across domains of life: The roles of bacteria in biogeochemical functions have been extensively studied in the environment, especially with the rise of genomics and metagenome-assembled genomes, along with the increase of methodologies to solve these problems. Just as bacterial species form communities that interact amongst each other, they also interact across the domains of life. Viruses, which predate on bacteria, can lead to cell lysis, making nutrients available to organisms in the food web. This dissertation addresses how viral and bacterial interactions impact microbially mediated biogeochemical cycling.

Chapter 2 Physiological and genomic evidence of cysteine degradation and aerobic hydrogen sulfide production in freshwater bacteria

Patricia Q. Tran[1,2]*, Samantha C. Bachand[1]*, Jacob C. Hotvedt[1]*, Kristopher Kieft[1,3], Elizabeth A. McDaniel[1,3], Katherine D. McMahon[1,4], Karthik Anantharaman[1s]

1. Department of Bacteriology, University of Wisconsin-Madison, WI, USA
2. Freshwater and Marine Sciences Doctoral Program, University of Wisconsin-Madison, WI, USA
3. Microbiology Doctoral Training Program, University of Wisconsin-Madison, WI, USA
4. Department of Civil and Environmental Engineering, University of Wisconsin-Madison, WI, USA

* Patricia Q. Tran, Samantha C. Bachand and Jacob C. Hotvedt are co-first authors.

This chapter has been accepted for publication in mSystems on March 29, 2023.

Abstract

The sulfur-containing amino acid cysteine is found in the environment, including in freshwater lakes. Biological cysteine degradation can result in hydrogen sulfide (H_2S), a toxic and ecologically relevant compound that is a central player in biogeochemical cycling in aquatic environments. Here, we investigated the ecological significance of cysteine in oxic freshwater, using isolated cultures, controlled experiments, and multi-omics. We screened bacterial isolates enriched from natural lake water for their ability to produce H_2S when provided cysteine. We identified 29 isolates (*Bacteroidota, Proteobacteria,* and Actinobacteria) that produced H_2S. To understand the genomic and genetic basis for cysteine degradation and H_2S production, we further characterized 3 isolates using whole-genome sequencing (using a combination of short-read and long-read sequencing) and tracked cysteine and H_2S levels over their growth ranges: *Stenotrophomonas maltophilia* (Gammaproteobacteria), *Stenotrophomonas bentonitica* (Gammaproteobacteria) and *Chryseobacterium piscium* (Bacteroidota). Cysteine decreased and an H_2S increased, and all 3 genomes had genes involved in cysteine degradation. Finally, to assess the presence of these organisms and genes in the environment, we surveyed a five-year time series of metagenomic data from the same isolation source (Lake Mendota, WI, USA) and identified their presence throughout the time series. Overall, our study shows that diverse isolated bacterial strains can use cysteine and produce hydrogen sulfide under oxic conditions, and we show evidence using metagenomic data that this process may occur more broadly in natural freshwater lakes. Future considerations of sulfur cycling and biogeochemistry in oxic environments should account for H_2S accumulation from the degradation of organosulfur compounds.

Importance

Hydrogen sulfide (H_2S), a naturally occurring gas with both biological and abiotic origins, can be toxic to living organisms. In aquatic environments, H_2S production typically originates from anoxic (lacking oxygen) environments such as sediments, or the bottom layers of thermally stratified lakes. However, the degradation of sulfur-containing amino acids such as cysteine, which all cells and life forms rely on, can be a source of ammonia and H_2S in the environment. Unlike other approaches for biological H_2S production such as dissimilatory sulfate reduction, cysteine degradation can occur in the presence of oxygen. Yet, little is known about how cysteine degradation influences sulfur availability and cycling in freshwater lakes. In our study, we identified diverse bacteria from a freshwater lake that can produce H_2S in the presence of O_2. Our study highlights the ecological importance of oxic H_2S production in natural ecosystems and necessitates a change in our outlook on sulfur biogeochemistry.

Introduction

In most natural environments, hydrogen sulfide gas (H_2S) production is usually attributed to defined groups of bacteria and archaea[30,31], and occurs primarily in anoxic environments. During the process of dissimilatory sulfate reduction, sulfate acts as a terminal electron acceptor, and is converted to hydrogen sulfide. However, other pathways for H_2S production exist, namely assimilatory sulfate reduction, in which H_2S contributes to cell growth and increased biomass, and the desulfurylation (desulfurization) of sulfur-containing amino acids such as cysteine which can lead to production of pyruvate, ammonia, and H_2S[32]. It is believed that assimilatory sulfate reduction contributes to growth but does not release H_2S from the cell, while dissimilatory sulfate

reduction and cysteine degradation can both contribute to growth and release ecologically relevant nitrogen and sulfur compounds into the ecosystem.

Microbes are responsible for several steps of the assimilatory and dissimilatory pathways, which enable sulfur intermediates (sulfate, thiosulfate, sulfate, hydrogen sulfide) to flow in the environment. Sulfur cycling in freshwater ecosystems can be ecologically significant, especially in places where strong redox gradients exist[33]. For example, in high arctic lakes, intermediate sulfur-compounds (other than sulfate and hydrogen sulfide) are suggested to serve as biogeochemical hubs, because more organisms had genes to transform the intermediates, than for sulfate-reduction and hydrogen sulfide oxidation[34]. Cysteine, a sulfur containing amino acid, is proposed to be an overlooked source of both carbon[35] and sulfur. Additionally, seston (particles in water comprised of both living and non-living organisms) contains organosulfur-containing lipids which settle into the sediments, and contributes to the sulfur pool, even in highly oligotrophic lakes such as Lake Superior[36]. We also note that some of these transformations can occur abiotically as well. For example, measuring sulfur isotopes can reveal if a compound has been formed through biological or abiotic processes, and abiotic transformations of sulfur compounds has been observed in geothermal springs environments[37]. Therefore, the sulfur cycle can be complex to parse out, because it is driven by both biotic and abiotic factors, and has multiple intermediates, for which the details of multiple specific pathways remain to be studied.

In seasonally stratified lakes consisting of oxygenated warm water (epilimnion) floating atop colder anoxic waters (hypolimnion), H_2S is often abundant in the hypolimnion[38,39], due to oxygen demand driving terminal electron acceptor depletion. However, an overlooked player in the pool

of available H_2S are the use of organosulfur compounds, such as cysteine, by microbes. Cysteine is required to produce many proteins and is also important for protein structure. It is one of the two amino acids (methionine being the other) that contains a sulfur group; however, the sulfhydryl group on cysteine is more reactive and can lead to H_2S formation. Methionine can also lead to H_2S formation, for example in the human body[40,41].

Like all amino acids, cysteine also contains an amine group that will form ammonia once the molecule is degraded. In the case of cysteine degradation (desulfurylation) by microbes, this also leads to H_2S production. H_2S is ecologically relevant because it can be toxic to plants and animals. During periods of anoxia, H_2S can accumulate to levels beyond the threshold for living organisms, and can cause massive fish kills[42]. Unlike other H_2S sources, cysteine desulfurylation can occur under oxic conditions (in *E.coli*)[43], thereby expanding the environmental scope of this sulfur pool. Indeed, cysteine can be desulfurylated under oxic conditions in the laboratory, but the natural prevalence of this process in freshwater lakes and other oxic environments remains unknown. We expect that H_2S production in oxic environments (during the mixed water column periods of the year, and throughout the stratified period in the mixed epilimnion) could result from cysteine breakdown by microbes.

In this study, we investigated the prevalence of organosulfur degradation (desulfurylation) in a freshwater lake, using both laboratory and genomic evidence, to advance our understanding of oxic sulfur cycling in aquatic ecosystems (Figure 2.1). First, we grew bacterial isolates enriched from Lake Mendota's oxic epilimnion to quantify H_2S and ammonia production, which informs the potential for organosulfur degradation in an oxygenated aquatic environment. We found 18

isolates producing H_2S under oxic conditions. We selected three H_2S-producing isolates for detailed characterization using full-genome sequencing and chemical analyses to track cysteine concentrations and H_2S accumulation during their growth: *Stenotrophomonas maltophilia* (Gammaproteobacteria), *Stenotrophomonas S. bentonitica* (Gammaproteobacteria) and *Chryseobacterium piscium* (Bacteroidota). In all three isolates, cysteine decreased and H_2S increased over their exponential growth curve under oxic conditions. Finally, we contextualized our laboratory results using a time-series of metagenomic data from the same isolation source (Lake Mendota, WI, USA), to study the temporal importance of organosulfur degradation. We found that genes for cysteine desulfurylation were present and abundant throughout the time-series suggesting that the ability to degrade cysteine is well represented in Lake Mendota.

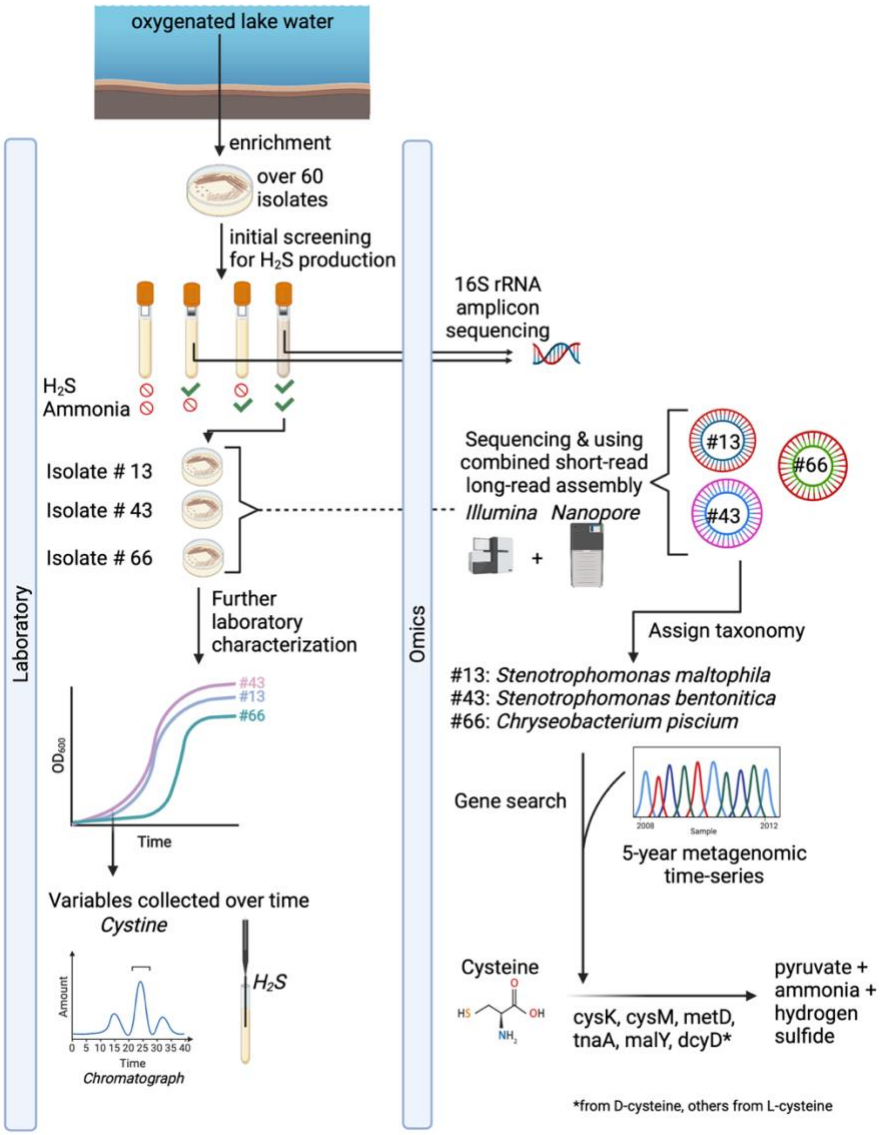

Figure 2.1 Overview of physiological and genomic methods to characterize hydrogen sulfide production from isolates from freshwater lakes. Isolates were enriched from oxygenated lake water, resulting in about over 60 isolates. Isolates were screened for H_2S and ammonia production using qualitative H_2S production assay and qualitative ammonia assay. Isolates that produced H_2S were selected for 16S rRNA gene sequencing. Then based on the taxonomic results from 16S rRNA gene sequencing, three distinct isolates were selected for whole-genome sequencing using a combination of short- and long-read sequencing. Genome characterization of functional potential

and taxonomic classification was conducted on the isolates. Screening of genes involved in cysteine degradation (Table S4) was conducted in the isolates and a five-year metagenomic time series of Lake Mendota (2008-2012).

Results

Isolates capable of H_2S production in oxic conditions.

To answer the question of whether bacteria could produce H_2S in the presence of oxygen, we grew freshwater isolates in pure culture originally recovered from the water column of temperate eutrophic Lake Mendota. We grew the isolates in moderately rich media (R2A, see methods) under control and treatment conditions (cysteine addition) and tracked H_2S production after 24 hours (Figure S1, Table S1 at https://doi.org/10.6084/m9.figshare.21711488). Using qualitative H_2S measurements, we found that 18 isolates produced H_2S and ammonia when grown in the presence of amended cysteine. We performed 16S rRNA gene sequencing on the 29 isolates that produced H_2S, regardless of whether they produced ammonia or not when amended with cysteine. Isolates that produced both H_2S and ammonia were identified as *Stenotrophomonas rhizophila* (Betaproteobacteria), *Stenotrophomonas maltophilia* (Betaproteobacteria), *Citrobacteria gillenii* (Gammaproteobacteria), and *Chryseobacterium sp.* (Bacteroidota), whereas those producing H_2S but not ammonia were identified as *Pseudomonas arsenicoxydans, Pseudomonas mandelii, Pseudomonas migulae, Pseudomonas thivervalensis,* and *Microbacterium flavescens*.

Detailed microbiological, chemical and genomic characterization in selected isolates

Next, we selected 3 isolates (#43, #13 and #66) representing distinct species based on 16S rRNA sequence taxonomy (97% identity), that produced H_2S for further characterization. These detailed

characterizations include OD_{600}-based growth rates and paired quantitative measurements of cysteine and H_2S concentrations in the spent medium. Cysteine addition resulted in concommital H_2S production over time (Figure 2.2, Table S2 at https://doi.org/10.6084/m9.figshare.21711509). We also tested which forms of cysteine (L-cysteine or D-cysteine) the isolates used. The Gammaproteobacteria isolates (panels CB) used L- and D- cysteine at similar rates, but *C. piscium* used D-cysteine at a greater rate than L-cysteine. (Figure 2.2, Table S3 at https://doi.org/10.6084/m9.figshare.21711497).

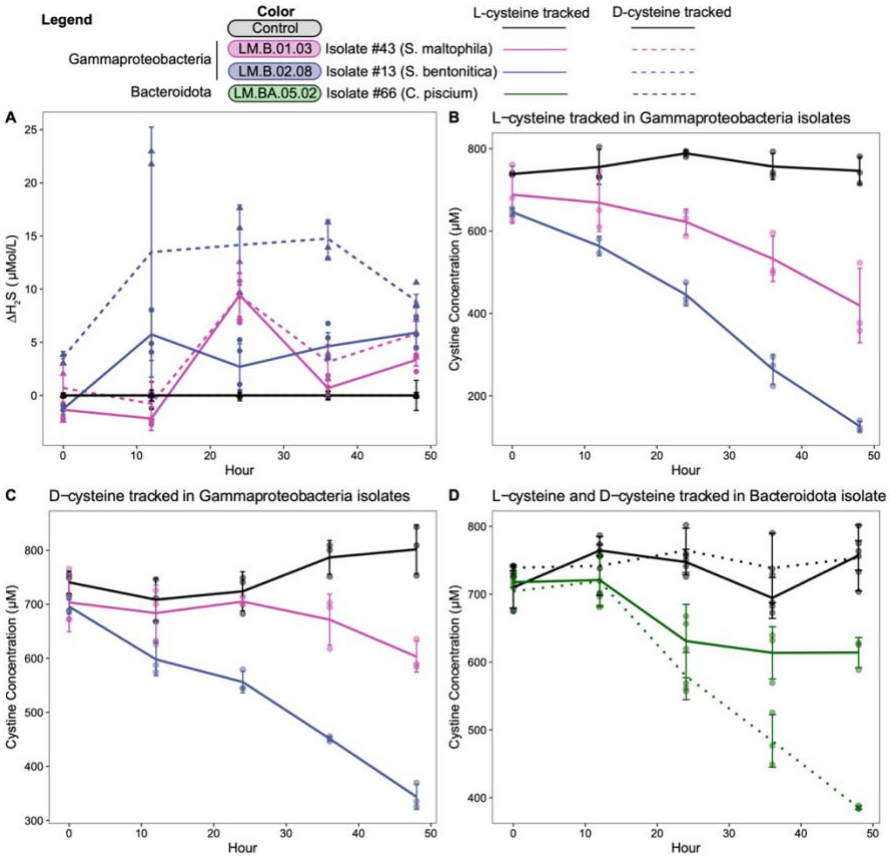

Figure 2.2 Quantitative characterization of cysteine utilization and hydrogen sulfide production in the three freshwater bacterial isolates using high-performance liquid chromatography. A. Higher amounts of H_2S were produced by the isolates compared to negative controls (meaning no bacteria was grown in the culture) over the course of 50 hours. **B, C, D.**

Identification of different forms of cysteine that can be degraded. L-cysteine decreased in all isolates compared to the control (B, D). D-cysteine also decreased over time in all samples except the negative control, however, the net amount decreased was less compared to L-cysteine. Cysteine concentrations were measured as cystine as described in the methods. Because Gammproteobacteria isolates (#43 and #13) were assessed in a different experimental run than the Bacteroidota isolate (#66), but using the exact same instruments and methods, plots B, C and D are separated by HPLC runs. Due to large sample volume, it was not possible to test all isolates and conditions in one HPLC run.

Next, we performed whole-genome sequencing using combined short-read and long-read sequencing on these 3 isolates. Isolate "13-LM-B-02-08" (referred to as #13), had a genome size of 4.18 Mbp, GC content of 66.8%, and was classified to the phylum Proteobacteria, class Gammaproteobacteria, order Xanthomonadales, family Xanthomonadaceae, genus Stenotrophomonas, and species Stenotrophomonas maltophilia_O. Isolate #13 was fully circular and assembled in 1 scaffold. Isolate 43-LM-B-01-03 (referred to as #43) had a genome size of 4.3 Mbp, GC content of 66.5%, and was classified to the phylum Proteobacteria, class Gammaproteobacteria, and species Stenotrophomonas bentonitica. Isolate #43 was assembled in 2 scaffolds. Finally, Isolate LM_BA_5.2 (referred to as #66) had the smallest genome at 1.37 Mbp, GC content of 33.7%, and was classified to the phylum Bacteroidota, class Bacteroidia, order Flavobacteriales, family Weeksellaceae, genus Chryseobacterium, and species Chrysobacterium piscium. Isolate #66 was assembled in 7 contigs. All three genomes were estimated to have 100% genome completeness according to GTDB-tk.

We performed whole-genome sequencing because functional information such as gene content cannot be predicted reliably from 16S rRNA gene sequencing alone. The full genome of Isolate 43 was assembled into a single circular genome, and taxonomically assigned to *Stenotrophomonas maltophilia*. Unlike the 16S rRNA sequence which assigned it to *S. rhizophila*, the full-genome was actually closer to *S. maltophilia*. The full genome of isolate #13, could be assembled into 2 long contigs, and was taxonomically assigned to *S. bentonitica*. The *Chryseobacter* genome was assembled into one circular genome, and assigned to *Chryseobacterium piscium*. All three genomes were estimated to be 100% complete based on CheckM. Overall, the 16S rRNA gene amplicon sequencing performed prior agreed with full-genome sequencing assignment in some cases, and in others, the whole-genome sequencing assignment allowed finer taxonomic resolution (such as in the case of Isolate 13).

Overall, whole-genome sequencing provided more information about the isolates' metabolic potential. The genomic content was then used to inform how or why H_2S might be produced in oxic environments, as shown in the laboratory experiments. We used gene annotations of the 3 isolates to infer the presence of genes involved in cysteine metabolism, namely those involved in cysteine degradation to ammonia, pyruvate, and H_2S: *metC*, *malY*, *tnaA*, *cysM*, *cysK* (which involve the use of L-cysteine as the substrate), and *dcyD* (which involves the use of D-cysteine as the substrate) (Table S4 at https://doi.org/10.6084/m9.figshare.21711491). However, we note that these genes may have other enzymatic activities, such as cysteine biosynthesis instead of degradation (Table S5 at https://doi.org/10.6084/m9.figshare.21711494) (Figure 2.3).

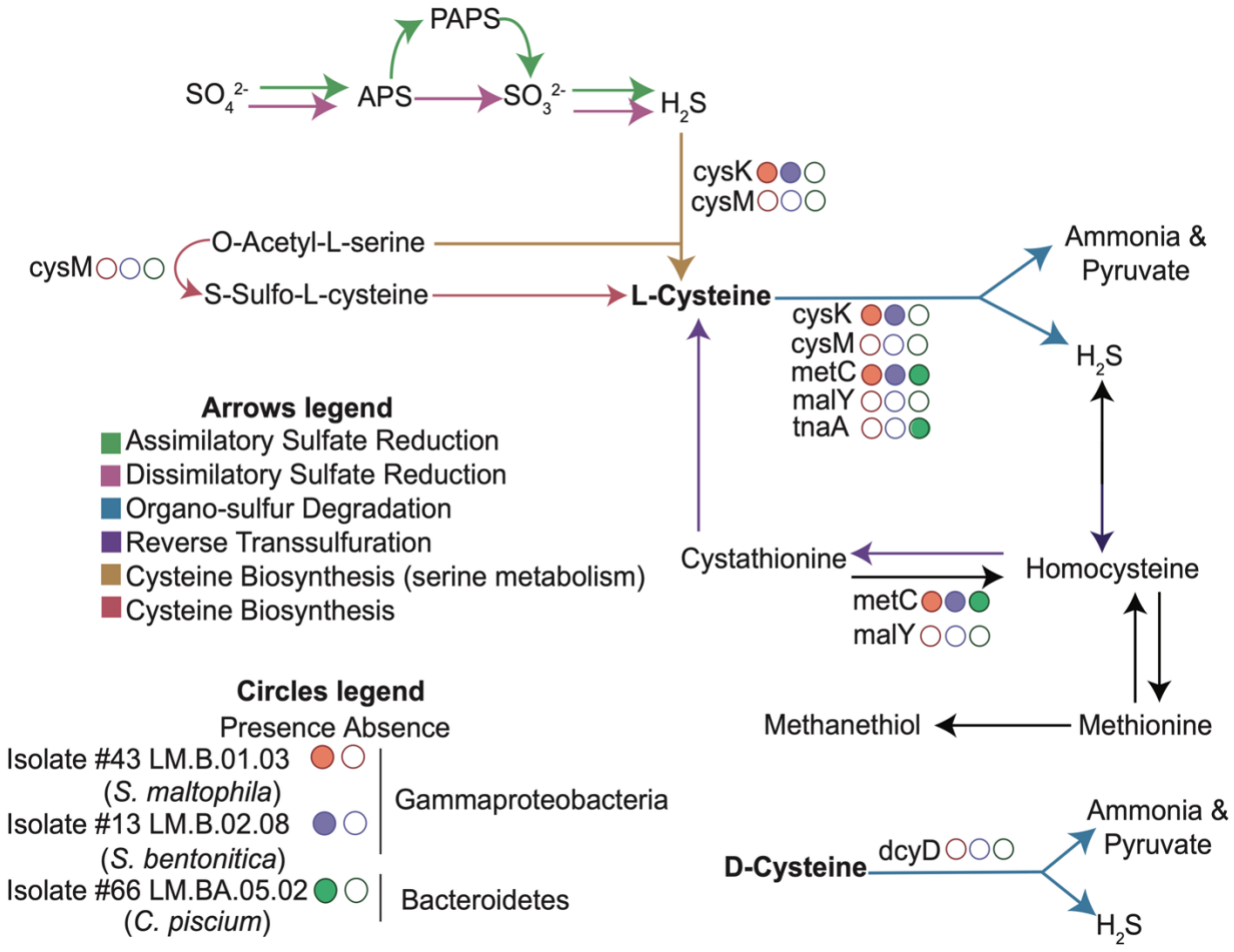

Figure 2.3 Genes involved in sulfur and organosulfur key pathways in the three isolates. Several pathways for hydrogen sulfide and cysteine production exist in microorganisms. The presence/absence of key genes (*cysK*, *cysM*, *malY*, *metC*, *tnaA*, *sseA*, *aspB*, and *dcyD* along the blue arrows) in the three isolate genomes are shown by filled (present) circles, and unfilled (not present) circles.

Leveraging the full-genomic content of the 3 isolates (Table S6 at https://doi.org/10.6084/m9.figshare.21711500), we proposed a joint cellular map based on

identified metabolic functions and pathways in the genomes (Figure 2.4). All three isolates contained pathways associated with central carbon metabolism: including the TCA cycle, glycolysis, gluconeogenesis, the pentose phosphate pathway and the glyoxylate cycle. They could all generate fatty acids using fatty acid biosynthesis, and catabolize fatty acids through the beta-oxidation pathway. As expected, they had genes for cysteine metabolism including *metC*, *malY*, *tnaA*, *cysM*, *cysK*, and *dcyD*, cysteine biosynthesis pathways from homocysteine and serine, as well pathways for degradation of other amino acids including methionine.

Despite these similarities, the three isolates also have distinguishing characteristics amongst them (Table S6 at https://doi.org/10.6084/m9.figshare.21711500, Figure 2.3). For example, while all isolates encoded genes for sulfur oxidation (sulfur dioxygenase), genes for thiosulfate oxidation were present in the two *Stenotrophomonas* isolates but not *Chryseobacterium*. The *Chryseobacterium* isolate encoded for a urease suggesting the use of organic nitrogen in the form of urea but this was absent in the two *Stenotrophomonas* isolates. Finally, genes for sugar utilization were identified in the two *Stenotrophomonas* isolates but not in *Chryseobacterium*.

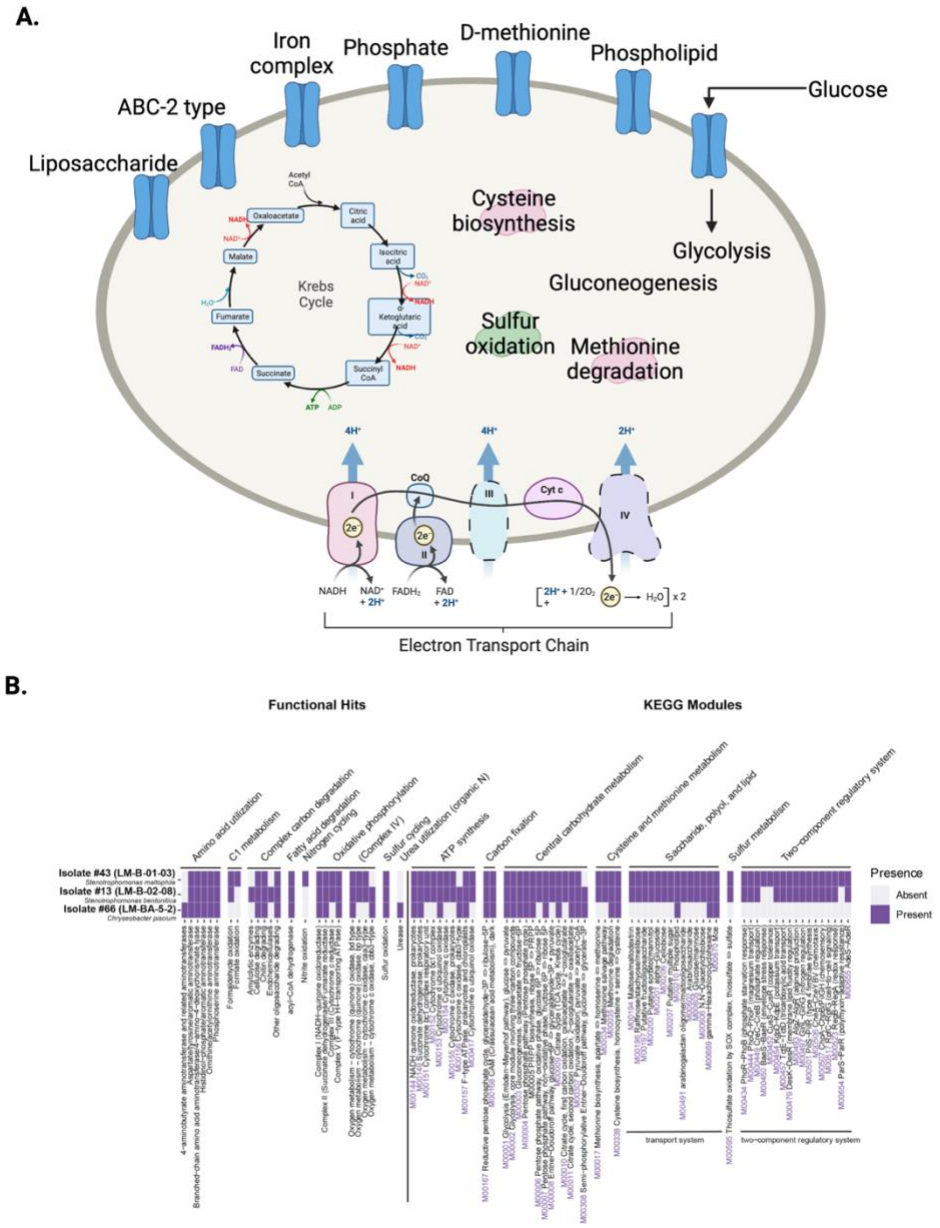

Figure 2.4 Common and distinguishing features of the three cysteine-degrading isolates. A. Cellular map showing important metabolic pathways and transporters which were common to all three isolates. A complete list is found in Table S6. The KEGG module identifiers are listed in purple whenever relevant. **B.** Heat map showing selected metabolic functions and pathways in the three isolates. A complete list is found in Table S6.

Presence of cysteine-degrading organisms and genes in a five-year metagenomic environmental time-series

To put these laboratory results and lab-grown organisms into a natural environment context, we leveraged a previously published metagenomic time-series collected from the oxygenated upper mixed layer of Lake Mendota spanning 2008-2012 (97 time points) [24], and reassembled and rebinned the data (See Methods) to search for the presence of cysteine degradation genes in the microbial communities.

First, we searched the time-series to see if organisms in our study were also present in the time series. To do this, we linked the 16S rRNA gene sequences of the isolated organisms to the assembled metagenomes (i.e. contigs) from the time series. We found that while the 16S rRNA sequences were also present in the time series (Table S7 at https://doi.org/10.6084/m9.figshare.21711530, S8 at https://doi.org/10.6084/m9.figshare.21711536, S9 at https://doi.org/10.6084/m9.figshare.21711548), and broadly distributed over time, these scaffolds were not part of binned genomes. Therefore, little information about these isolates would be gathered from metagenomic data only. As such, the full-genome sequencing we performed was particularly helpful in understanding the full genomic structure of the H_2S-producing organisms.

Second, we searched for the 6 genes associated with cysteine degradation and H_2S production (Table S5 at https://doi.org/10.6084/m9.figshare.21711494) (in binned and unbinned contigs)

using KEGG HMMs with hmmsearch. In total, we searched over 22 million amino acid sequences and identified 1882 hits to the 5 marker genes found in the isolate metagenomes, *dcyD* was not found (Figure 2.5, Table S10 at https://doi.org/10.6084/m9.figshare.21711527). *cysK* and *malY* were the genes with the most corresponding matches at any time point, followed by *metC* and *cysM*. Only 2 scaffolds contained *tnaA*. Overall, after normalizing for metagenome read size per sample, there was no obvious temporal trend of the genes, although genes were found throughout the 5-year time series.

Among these cysteine-degrading gene sequences, several were identified in binned metagenome-assembled genomes (MAGs) (Figure 2.5, Table S11 at https://doi.org/10.6084/m9.figshare.21711545), which allowed for the assignment of taxonomy. Overall, we identified 139 genes to be distributed in genomes of organisms from Actinobacteria, Bacteroidota, Chloroflexota, Cyanobacteria, Planctomycetes, Proteobacteria, and Verrucomicrobia, representing common freshwater lineages [25]. The *tnaA* gene was only present in Bacteroidota, but other genes were more broadly distributed among taxonomic groups.

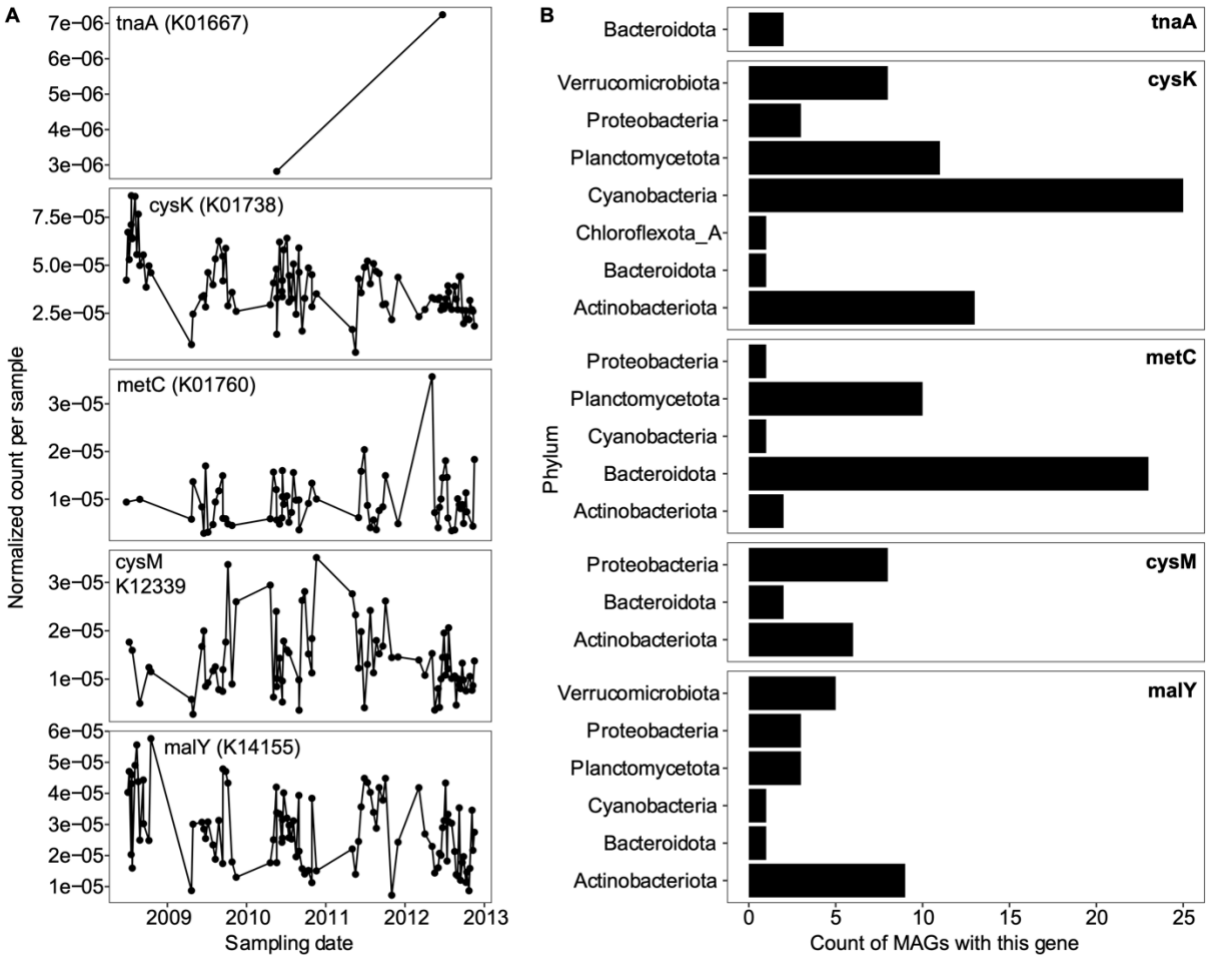

Figure 2.5 Distribution of cysteine desulfurization genes across time and taxa. A. Counts of cysteine desulfurization genes (5 genes) in a 5-year time-series of Lake Mendota, normalized by total number of annotated genes per metagenome sample. We also searched for dcyD but did not identify it in any sample. B. Taxonomy of the metagenome-assembled genomes in which those genes were found.

Discussion

The fate of hydrogen sulfide and ammonia, two products of cysteine degradation

H_2S blocks the binding of oxygen during aerobic respiration making it toxic to the cell. Therefore, we hypothesize that most microbes aerobically degrading cysteine will excrete H_2S from the cell as was seen with the 29 isolates from this study producing H_2S when grown in media containing a cysteine source. Additionally, the 3 isolates that underwent further characterization were quantitatively shown to deplete cysteine levels when producing H_2S (Figure 2). While it is possible that organisms may use H_2S internally as a sulfur source, we did not identify any genes encoding sulfide quinone oxidoreductases, flavocytochrome c dehydrogenases, or other genes for the oxidation or transformation of H_2S. H_2S can also be used as a protective compound against antibiotics in aerobic environments by some bacteria; however, due to the stress H_2S accumulation has on the cell, this defense mechanism is only used in extreme situations [44].

The H_2S-producing isolates fell into two groups when grown in media with cysteine: ammonia-producing and ammonia-consuming. Because ammonia is a common metabolic precursor, the production or consumption of ammonia alone does not indicate whether the cell is undergoing the cysteine degradation pathway. This may be why we see variation in whether the 29 isolates produced or consumed ammonia despite all producing H_2S (Figure S1).

Genomic structure of the H_2S-producing isolates

Overall, the three isolates selected for whole genome sequencing revealed genes for cysteine degradation into H_2S. Based on laboratory studies, they were able to produce H_2S in the presence of oxygen. The isolates were obligate aerobes, presenting interesting questions about these organisms' life history.

Stenotrophomonas maltophilia, Stenotrophomonas bentonitica, and *Chryseobacterium piscium* are have been shown to be present in natural environments. *S. maltophilia* is a cosmopolitan bacterium in nature, and found in a range of natural environments, particularly in association with plants[45]. *S. bentonitica* was originally characterized in bentonite formations, was predicted to have high tolerance to heavy metals [46], and has been observed in arctic seawater [47]. *C. piscium* was isolated from a fish in the arctic ocean [48], but its ecological significance in the oceans remains unknown. This previously described *C. piscium* strain LMG 23089 was not reported to produce H_2S yet our genetic and physiological analyses suggest that it has the enzymatic machinery to degrade cysteine.

One possible explanation for this discrepancy is that LMG 23089 was previously grown on SIM medium to test H_2S. The SIM medium is used for physiological study of sulfur reduction, indole production and motility (abbreviated SIM). The SIM medium offers visual representation of H_2S production via the reduction of thiosulfate, it reacts with iron salts in the media, and media changes to darker black color. This method of assessing H_2S is lower resolution than the modern H_2S probes which measure µM concentrations. As a side test on isolate #66, H_2S was not produced when thiosulfate was provided, but H_2S was produced when cysteine was provided.

One finding of this study was that none of the 6 genes searched for cysteine degradation into H_2S and ammonia was common to *all* three isolates, despite all three isolates showing the same cysteine-decrease, ammonia-increase, and H_2S-increase over time. This could be explained by alternative, perhaps less straightforward pathways for H_2S production. One pathway is led by a gene named cystathionine gamma-lyase (“CTH” or “CSE”). In some bacteria and mammals, this

enzyme is involved in H_2S production [49]. An HMM search for this enzyme showed that it was present in Isolate #13, 43 and #66. While it was not initially included in the initial methods and study, this could hint to another commonality among oxic H_2S producing organisms.

Challenges associated with measuring oxic H_2S production from organosulfur in the environment

Extrapolating these laboratory results to widespread distribution of organosulfur degradation in the natural environment necessitates several steps, namely because of the major knowledge gaps that exist concerning the sulfur cycle in freshwater lakes, and because bridging the gap between cultivation-based, omics-based [50], and field-based experiments is needed. Foremost, the identity, distribution, and availability of organosulfur compounds broadly across lakes globally is currently mostly unknown. Cysteine is notoriously difficult to measure, and many previous studies characterizing the amino acid composition of the water column only measure the sulfur-containing organosulfur compound taurine [51,52]. One of the difficulties in studying the fate of cysteine in oxic environments is that it can be abiotically and spontaneously oxidized into cystine [53], which *Escherichia coli* has been shown to uptake [54]. In a study of *E.coli* K-12 that lacked a cysteine transporter, cysteine could enter the cell through transporters dedicated to other amino acids, when no amino acid alternatives other than cysteine were present in the medium [43].

Organic sulfur in the form of cysteine is an important organosulfur amino acid, and is important in protein folding and function [55]. As such, there is a difference in the fates of cysteine when it exists bound in cell walls, versus when cysteine is free in the water column and available for degradation by bacteria. While cysteine has been shown to contribute to the carbon pool and carbon flow in lakes [35], more quantitative field measurements are necessary to support whether cysteine

also serves as a sulfur pool. Yet, other forms of organosulfur have important significance in aquatic environments. In marine environments, for example, DMSP (dimethylsulfoxonium propionate) is a critical component of the marine organosulfur cycle [56].

Additionally, current differences between computational gene similarity searches versus *in vivo* enzymatic functions are challenging to assess for the genes responsible for the cysteine degradation into pyruvate, ammonia, and H_2S. One reason is that the enzymatic activity of the gene has mostly been described in model organisms such as *E. coli*, and it has been shown that gene expression can be induced by genetic factors, or environmental factors such as metals [57]. At least 6 genes have been proposed to have this enzymatic activity, yet each gene may serve different functions *in situ*, and it is difficult to assert directionality of enzymatic function based on metagenomic or genomic analyses only. To this end, the isolated bacterial strains from this study, which are non-model organisms, and originate from the natural freshwater lake environment, may be used for further detailed biochemical, physiological, and microbiological studies. Further characterization of these bacterial isolates using gene knockout, gene induction, or heterologous gene expression studies may inform the functional activity of these genes in nature.

Implications of oxic H_2S production by microbes in freshwaters

This study demonstrates the potential for H_2S production by microbes in lake ecosystems to occur in the presence of oxygen, using genomic and physiological evidence from pure culture bacterial isolates, and screening of long-term metagenomic time-series. By combining lake-to-laboratory experiments, we show that multiple bacterial strains spanning Gammaproteobacteria,

Betaproteobacteria, Actinobacteria, and Bacteroidota are all able to produce H_2S under oxic conditions, and at temperatures that would be ecologically relevant for surface lake water during the summer. Surface water temperatures in Lake Mendota can reach up to 27°C, and the top few meters of water surface are saturated in oxygen. Worldwide, maximum lake surface temperature can range between 23 to 31°C [58].

Unlike dissimilatory sulfate reduction, bacteria using cysteine to generate ammonia, pyruvate, and H_2S, is not dependent on sulfate as an initial reaction substrate. Increased sulfate concentrations are shown to lead to higher sulfate reduction rates in shallow eutrophic freshwater, the sulfur originating from algal decay for example [59]. While Lake Mendota is a low-iron and high-sulfate lake [60], not all lakes have elevated sulfate levels, and therefore, H_2S production might previously not have been thought of as relevant to study. However, sulfur-containing amino acids can have many origins. In lakes, concentrations of amino acids (free dissolved and combined) often reflect the input and outputs of the lake [61,62]. For example, amino acids contributed a detectable amount to the nitrogen cycle, and bacterial utilization of amino acids contributes to nitrogen pool and cycling in natural ecosystems, [62] although cysteine amino acids were not measured in that study.

Freshwater lakes that are dimictic can stratify in temperature and oxygen during the summer, and oxygen concentrations vary throughout the year. In the fall and spring, oxygen is abundant, and cysteine degradation into H_2S could be a relevant process for the sulfur pool, and H_2S fluxes to the atmosphere could be significant since wind is prevalent. Under ice during the winter, where oxygen is plentiful, H_2S could be produced but also rapidly consumed or oxidized. On the other hand, gases would be trapped under ice. During summer, the anoxic hypolimnion and sediments

are known H_2S sources due to dissimilatory sulfate reduction, but density gradients would prevent H_2S from reaching the atmosphere. However, the oxygenated mixed epilimnion could be an H_2S source through organosulfur degradation. If we consider the importance of oxic hydrogen sulfide production, which could occur year-round, the H_2S pool and the scope of sulfur transformations may be greater than anticipated, if we focus solely on the anoxic hypolimnion (Figure 2.6). Future work aiming to understand the broader distribution of sulfur-containing amino acids and other organosulfur compounds in freshwaters, their fates and transformations, as well as their contribution to H_2S production, will inform global sulfur biogeochemical cycling.

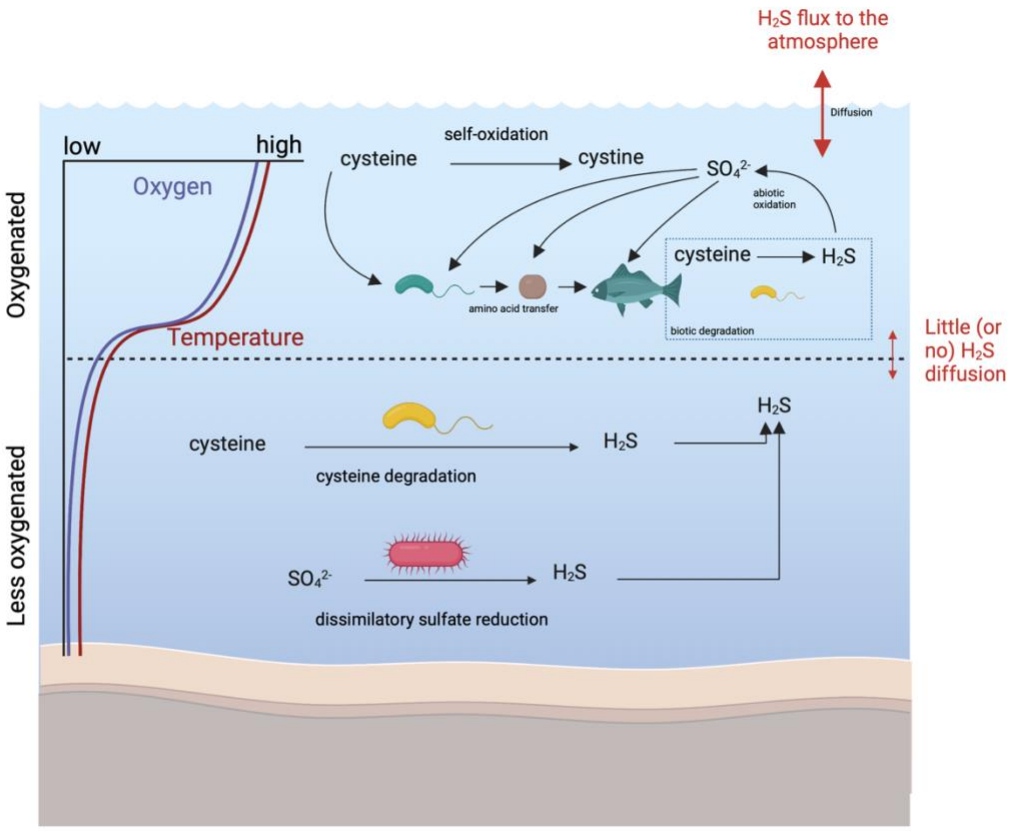

Figure 2.6 Distribution of cysteine desulfurization genes across time and taxa. Oxygenated seasons and part of the lake water columns are shown with an asterisk. Significant research gaps

include cysteine concentrations in the natural environment over time, hypothesized H_2S fluxes across different layers in the lake water column, and contribution of different H_2S sources in the hypolimnion. In all seasons, portions of the water column can be oxygenated.

Methods

Enrichment cultures of isolates from a temperate freshwater lake

Lake Mendota (43°06′24″ N 89°25′29″ W) is a temperate eutrophic lake in South-Central Wisconsin, in Madison, WI, USA. Lake Mendota is part of the Long-Term Ecological Research Network North Temperate Lakes (NTL-LTER, https://lter.limnology.wisc.edu/about/lakes). Lake Mendota encounters annual stratification and annual seasonal anoxia in the hypolimnion. Lake water was collected on September 14, 2018 from an integrated water sample (0m to 12m) from Lake Mendota at the "Deep Hole" station (43°05'54", 89°24'28"), where the maximum depth is 23.5m. The water samples were collected during stratification from the oxygenated epilimnion. The lake water was collected in pre-acid washed 2 L sampling bottles using a flexible PVC tube, and brought back on shore within hours for immediate processing. Serial dilution was performed, and lake bacteria were grown on PCB (plate count media broth) agar media, at room temperature (~21^{o}C), in the lab in the light. The PCB media was made of: 1 L water, 5g/L of yeast extract, 10g/L of tryptone and 2g/L of dextrose/D-glucose. If grown on solid media, 10g of agar per 1L media was added. Enrichment resulted in about 60 isolates.

Screening for cysteine degradation into H_2S and ammonia

Isolates were able to grow on PCB and R2A media. R2A media is a culture medium for bacteria that typically grow in freshwater. It is less "nutrient-rich" than PCB media, and therefore slightly

closer to natural lake water than PCB. For the screening of the isolates for H_2S production, we grew them on R2A media. Each isolate had two treatments: grown in R2A media without cysteine for the control, and grown in R2A media with amended cysteine as the treatment.

R2A media consisted of (1 L of water): 0.5g of casein, 0.5g of dextrose, 0.5g of starch, 0.5g of yeast extract, 0.3g of K_2HPO_4, 0.3g of sodium pyruvate, 0.25g of peptone, 0.25g of beef extract, 0.024g of $MgSO_4$, and autoclave. To make the same media for plates, we added 15g of agar before autoclaving. For controls, isolates were grown in the media without cysteine amendments. For "treatments", 2mM cysteine was added.

To assess the amount of cysteine degradation into H_2S and ammonia, we screened each of the 60 isolates for H_2S and/or ammonia accumulation. To test H_2S production, we grew the strains individually in liquid media, in culture tubes with loose fitting caps, thus letting oxygen in. They were also shaken at room temperature. We and used lead acetate test strips (Fisher Scientific, USA) to qualitatively assess H_2S accumulation in the headspace. A darkening of the strip shows that H_2S was produced. To test ammonia concentrations after 24 hours, we measured samples at time zero and 24 hours using Ammonia Salicylate Reagent Powder Pillows and Ammonia Cyanurate Reagent Powder Pillows (Hatch Reagents) and used spectrophotometry at the 655nm wavelength. The three isolates were tested by thioglycolate broth test and were determined to be obligate aerobes.

Identification of H_2S producing bacteria using 16S rRNA gene sequencing

Colony PCR and DNA extractions were conducted using the EtNa Crude DNA Extraction and ExoSAP-ITTM PCR Product Cleanup protocols on the isolates that tested positive for producing H_2S. Full length 16S rRNA gene products were generated for sequencing using universal 16S rRNA primers (27f, 1492r)[63].DNA concentration yields were measured using the QuBit dsDNA HS assay kit (QuBit). DNA was sequenced at the University of Wisconsin-Madison Biotechnology Center (Madison, WI, USA) The program 4Peaks [64] was used to clean the base pairs by quality-checking followed by homology search using BLASTn against the NCBI Genbank database (accessed December 2019) [65] to identify the sequences.

Detailed characterization of 3 hydrogen-sulfide producing isolates

We selected 3 isolates that could aerobically produce H_2S for further detailed characterization. We selected these isolates because some of the 18 isolates that produced H_2S when grown with cysteine had identical 16S rRNA sequences, therefore we chose isolates that had distinct 16S rRNA sequences for full-genome sequencing. Additionally, using 16S rRNA gene sequencing of the isolates, one was only assigned to *Stenotrophomonas sp.*, and we believed that whole-genome sequencing would enable us to get a higher taxonomic confirmation and more complete information.

We performed DNA extraction using the PowerSoil Powerlyzer kit (Qiagen) without protocol modifications, and sent the genomic DNA for full genome sequencing at the Microbial Genome Sequencing Center (MIGS) (Pittsburg, PA) with combined short read Illumina and long read nanopore sequencing. The data was processed by MIGS to assemble the short-reads (Illumina Next Seq 2000) and long-reads (Oxford Nanopore Technologies) into full-genomes. Quality control and

adapter trimming was performed with bcl2fastq (Illumina) and porechop (https://github.com/rrwick/Porechop) for Illumina and Oxford Nanopore Technologies (ONT) MinION sequencing respectively. Hybrid assembly with Illumina and ONT reads was performed with Unicycler [66]**.** Genome annotation of the 3 isolates was done with Prokka v.1.14.5[67], using the --rfam setting.

Genome completeness and contamination were estimated using CheckM v.1.1.3 [68] *lineage_wf*. Taxonomic classification was conducted using GTDB-tk v.0.3.2[69] with the database release r95. The full-genome taxonomic classification agreed with the prior 16S rRNA gene sequencing results, but we were further able to identify Isolate 43 as *Stenotrophomonas bentonitica*. We ran METABOLIC-G v.4.0[70] to identify genes associated with cysteine degradation and other metabolic pathways.

Growth measurements of the three isolates were measured using OD_{600} with a spectrophotometer, with measurements every 1 hour. The isolates were grown in R2A broth media, shaken in an incubator at 27°C. Aliquots were collected over the growth range for cysteine measurements across the growth curves, as described below. A H_2S microsensor (Unisense) was used to measure H_2S over time.

Methods to measure cysteine

Cysteine concentrations were measured as cystine, as described in [71] (https://osf.io/9k8a6/). One of the reasons for measuring cystine instead of cysteine is that in oxic environments, cysteine is oxidized rapidly into cystine[53,54]. Additionally, unless LC-MS is used, cysteine can be difficult to

measure directly. Samples were diluted 5:4:1 Sample:DI H_2O:DMSO and left at room temperature for at least 24 hours. Chromatographic analysis was performed on an Agilent 1260 Infinity II with an Agilent Zorbax Eclipse Plus C18 RRHT 4.6x50mm, 1.8µm, with Guard column. Column temperature was maintained at 40°C using an Agilent 1260 TCC (G1316A).

Gradient elution was performed using Mobile phase A (MPA) consisting of 10mM Na_2HPO_4, 10mM Sodium tetraborate decahydrate, in DI H_2O, adjusted to pH 8.2 with HCl, filtered to 0.45µm. Mobile phase B (MPB) consisted of 45:45:10 Acetonitrile:Methanol:DI H_2O. Gradient used for elution was as follows: 0 minutes 98% MPA, 2% MPB; 0.2 minutes 98% MPA, 2% MPB; 6.67 minutes 46% MPA, 54% MPB; 6.77 minutes 0% MPA, 100% MPB; 7.3 minutes 0% MPA, 100% MPB; 7.4 minutes 98% MPA, 2% MPB; 8 minutes 98% MPA, 2% MPB. Flow rate was 2.0mL/min. The pump used was an Agilent Infinity Series G1311B quat pump. Pre-column derivatization was performed using an Agilent 1260 ALS (G1329B) with an injector program. Detection was performed using an Agilent 1260 Infinity II MWD (G7165A) at 338nm with 10nm bandwidth. Reference was 390nm with 20nm bandwidth. Recovery was tested during method development. Recoveries of cystine ranged from 87.2-101.5%, with an average of 92.1%.

Methods to measure H_2S using a microsensor

Aliquots of at least 1 mL were taken from cultures at desired times after inoculation. We used the Unisense H2S microsensor probe (https://unisense.com/products/h2s-microsensor/) following exactly the manufacturer's method for making standards and calibrating. H_2S concentrations were measured by suspending the H_2S probe in the aliquot, and leaving it in place until the measurement stabilized over 2 minutes. Because the measurement fluctuates over the course of these minutes,

we excluded data gathered while the probe was stabilizing in the sample and averaged the value for each time point.

Generation of metagenome-assembled genomes

Sequencing of the Lake Mendota time series for 2008-2012 was previously conducted at the Joint Genome Institute[24], containing 97 time points (and therefore 97 metagenomic datasets)[72]. In summary, raw reads were quality filtered using fastp[73], and individually assembled using metaSPAdes[74]. Each metagenome was reciprocally mapped to each individual assembly using BBMap v38.07[75] with 95% sequence identity cutoff. Differential coverage mapping to all samples was used to bin contigs into metagenome-assembled genomes (MAGs) using Metabat2 v.2.12.1[76]. Bins were quality assessed with CheckM v.1.1.2[68], dereplicated with dRep v.2.4.2[77], and classified with GTDB-tk v.0.3.2[78] with default settings. This resulted in a total of 116 MAGs from Lake Mendota (Table S12 at https://doi.org/10.6084/m9.figshare.21711551) which are available for download at the Open Science Framework (https://osf.io/qkt9m/).

Searching for cysteine genes and isolates presence in metagenomic time-series and MAGs

Genes for cysteine degradation were identified using HMMsearch v3.1b2 [79]. HMMs were downloaded from KoFam[80], accessed May 2020). The KO numbers for the six cysteine degradation genes are: *metC* (K01760), *cysK* (K01738), *cysM* (K12339), *malY* (K14155), *tnaA* (K01667), and *dcyD* (K05396) (Table S4 at https://figshare.com/s/c7ea24f2f5c8773ea385). The HMM files are those published by KOfam, with the modification of manual addition of the TC

thresholds, curated by KEGG. HMM-based homology searches were conducted on the 97 Lake Mendota metagenomes assemblies as described above.

Data availability

The 16S rRNA sequences for the 29 H_2S-producing isolates, and the whole-genome sequences (nucleotides and amino acids) for isolates #13, #43 and #66 are available on OSF at the following doi: https://doi.org/10.17605/OSF.IO/G25EQ. The isolates genomes are also deposited on NCBI in the Bioprojects: PRJNA776273, PRJNA776272 and PRJNA839079. The 97 metagenomes were previously published in [72] and are available through JGI's IMG/M and Genome Portal.

Acknowledgments

We are thankful to Anna Schmidt for collecting the original lake water from Lake Mendota in 2018, and to Adam Breister and Elizabeth Zanetakos who enriched the bacterial isolates during summer 2018. We thank Trina McMahon's Lab and the Long-Term Ecological Research Network, and the Center for Limnology for their field support and prior work on Lake Mendota. We are thankful for the University of Wisconsin's Water Science and Engineering Laboratory for the use of their HPLC instrumentation, and James Lazarcik for training and assistance with the instruments.

This work was supported by the USDA National Institute of Food and Agriculture (NIFA) under grant: Hatch project 1025641. Patricia Tran and Kristopher Kieft received the support from the Anna Grant Birge Memorial Award from the Center for Limnology for support for the project in

2019. Patricia Tran is supported by the Natural Science and Engineering Research Council (NSERC) of Canada Doctoral Fellowship and a Wisconsin Distinguished Graduate Award Fellowship from the University of Wisconsin-Madison. Samantha Bachand was supported by the National Science Foundation (NSF) Research Experience Undergraduate (REU) Award, and the University of Wisconsin-Madison' Holstrom Environmental Research Fellowship. Kristopher Kieft was supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison, and a William H. Peterson Fellowship Award from the Department of Bacteriology, University of Wisconsin-Madison. Elizabeth McDaniel was supported by a fellowship through the Department of Bacteriology at the University of Wisconsin - Madison. This research was performed in part using the Wisconsin Energy Institute computing cluster, which is supported by the Great Lakes Bioenergy Research Center as part of the U.S. Department of Energy Office of Science. We thank the U.S. Department of Energy Joint Genome Institute for sequencing and assembly (CSP 394) of the Lake Mendota metagenomes.

Chapter 3 Depth-discrete metagenomics reveal the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika.

This chapter is published in the ISME Journal. Citation:

Tran, Patricia. Q., Samantha C. Bachand, Peter B. McIntyre, Benjamin Kraemer, Yvonne Vadeboncoeur, Ismael A. Kimirei, Rashid Tamatamah, Katherine D. McMahon, Karthik Anantharaman. (2021) "Depth-discrete metagenomics reveal the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika". *ISME J* 15, pg. 1971–1986. https://doi.org/10.1038/s41396-021-00898-x

Supplementary material for this chapter is found on the publisher's website.

Abstract

Lake Tanganyika (LT) is the largest tropical freshwater lake, and the largest body of anoxic freshwater on Earth's surface. LT's mixed oxygenated surface waters float atop a permanently anoxic layer and host rich animal biodiversity. However, little is known about microorganisms inhabiting LT's 1470 meter deep water column and their contributions to nutrient cycling, which affect ecosystem-level function and productivity. Here, we applied genome-resolved metagenomics and environmental analyses to link specific taxa to key biogeochemical processes across a vertical depth gradient in LT. We reconstructed 523 unique metagenome-assembled genomes (MAGs) from 34 bacterial and archaeal phyla, including many rarely observed in freshwater lakes. We identified sharp contrasts in community composition and metabolic potential with an abundance of typical freshwater taxa in oxygenated mixed upper layers, and Archaea and uncultured Candidate Phyla in deep anoxic waters. Genomic capacity for nitrogen and sulfur cycling was abundant in MAGs recovered from anoxic waters, highlighting microbial contributions to the productive surface layers via recycling of upwelled nutrients, and greenhouse gases such as nitrous oxide. Overall, our study provides a blueprint for incorporation of aquatic microbial genomics in the representation of tropical freshwater lakes, especially in the context of ongoing climate change, which is predicted to bring increased stratification and anoxia to freshwater lakes.

Introduction

Located in the East African Rift Valley, Lake Tanganyika (LT) holds 16% of the Earth's freshwater and is the second-largest lake by volume. By its sheer size and magnitude, LT exerts a major influence on biogeochemical cycling on regional and global scales[81,82]. For instance, LT stores over 23 Tg of methane below the oxycline [82], and about 14,000,000 Tg of carbon in its sediments[81]. Over the past centuries, LT's rich animal biodiversity has been a model for the study of species radiation and evolution [8]. In contrast, the microbial communities in LT that drive much of the ecosystem-scale productivity remains largely unknown.

LT is over ten million years old and oligotrophic, which provides a unique ecosystem to study microbial diversity and function in freshwater lakes, specifically tropical lakes. The comparatively thin, oxygenated surface layer of this ancient, deep lake harbors some of the most spectacular fish species diversity on Earth[83], but surprisingly ~80% of the 1890 km3 of water is anoxic. Being meromictic, its water column is permanently stratified. This causes a large volume of anoxic and nutrient-rich bottom waters to be thermally isolated from the upper ~70 m of well-lit, nutrient-depleted surface waters. Despite stratification, periodically in response to sustained winds, pulses of phosphorus and nitrogen upwelling from deep waters replenish the oxygenated surface layers and sustain its productivity[84].

The physicochemical environment of LT is quite distinct from other ancient lakes[19,85]. For example, Lake Baikal (LB), located in Siberia, is a seasonally ice-covered, of comparable depth (1642 m for LB, and 1470 m for LT), yet their thermal and oxygen profiles are drastically distinct. While LT is a permanently stratified layer with a large layer of anoxic waters, LB's deepest layers

are oxygenated. Previous work on LT's microbial ecology documented spatial heterogeneity in microbial community composition, especially with depth[86,87]. The observed differences were primarily related to thermal stratification, which leads to strong gradients in oxygen and nutrient concentrations. Early evidence from LT suggests that anerobic microbially driven nitrogen cycling such as anerobic ammonium oxidation (anammox) is an important component of nitrogen cycling[88]. However, the emergent effects of depth-specific variation in microbial communities on biogeochemical and nutrient cycling in LT remain largely unknown.

Here, we investigated microbial community composition, metabolic interactions, and microbial contributions to biogeochemical cycling along ecological gradients from high light, oxygenated surface waters to dark, oxygen-free and nutrient-rich bottom waters of LT. Our comprehensive analyses include genome-resolved metagenomics to reconstruct hundreds of bacterial and archaeal genomes, which were used for metabolic reconstructions at the resolution of individual organisms and the entire microbial community, and across different layers in the water column. In addition, we compared the microbial ecology in two contrasting ancient, and deep rift-formed lakes (LT and LB) to address questions about ecology, evolution, and endemism of microorganisms in freshwater lakes. Our work offers a window into the understudied microbial diversity of LT and serves as a case study for investigating microbial roles and links to biogeochemistry in globally distributed anoxic and deep freshwater lakes.

Methods

Sample collection

Samples were collected in LT, located in Central-East Africa, and based around the Kigoma and Mahale regions of Tanzania. We sampled near Kigoma because there are established field sites there and data that go back more than a decade. We sampled near Mahale National Park because of our focus on conservation in the nearly intact nearshore ecosystems there. LT has a strong latitudinal gradient (~673 km), therefore sampling in these two locations was also a way to capture some of that latitudinal variability.

From 2010 to 2013, samples for chlorophyll a, conductivity, dissolved oxygen (DO), and temperature were collected. Secchi depths were collected over 2 years in 2012 and 2013, from a previous research cruise using a YSI 6600 sonde with optical DO and chlorophyll-a sensors. Those measurements were collected down to ~150 m. Twenty-four water samples were collected in 2015 for metagenome sequencing. A summary of all samples is available in **Table S1**. The samples were collected around two stations termed Kigoma and Mahale based on nearby cities. Water samples were collected with a vertically oriented Van Dorn bottle in 2015 and metadata is listed in **Table S1**. For the metagenomes, three Van Dorn bottle casts (10 L) in Kigoma and Mahale were collected, in which depth-discrete samples were collected across a vertical gradient, down to 1200 m at the maximum depth (Figure 3.1 and Table S1). In addition, five surface samples were collected from the Mahale region, and one surface sample from the Kigoma region (Figure 3.1). Methods to produce the map in Figure 3.1 using the geographic information system ArcMap are described in the **Supplementary Text**.

We filtered as much as we could until the filter clogged, which was usually between 500 and 1000 mL. This usually took about 10–20 min of hand pumping. The water was not prefiltered, but

filtered directly through a 47 mm 0.2-µm pore nitrocellulose filter, which was stored in a 2-mL tube with RNAlater. Filters were frozen immediately and brought back on dry ice to UW-Madison.

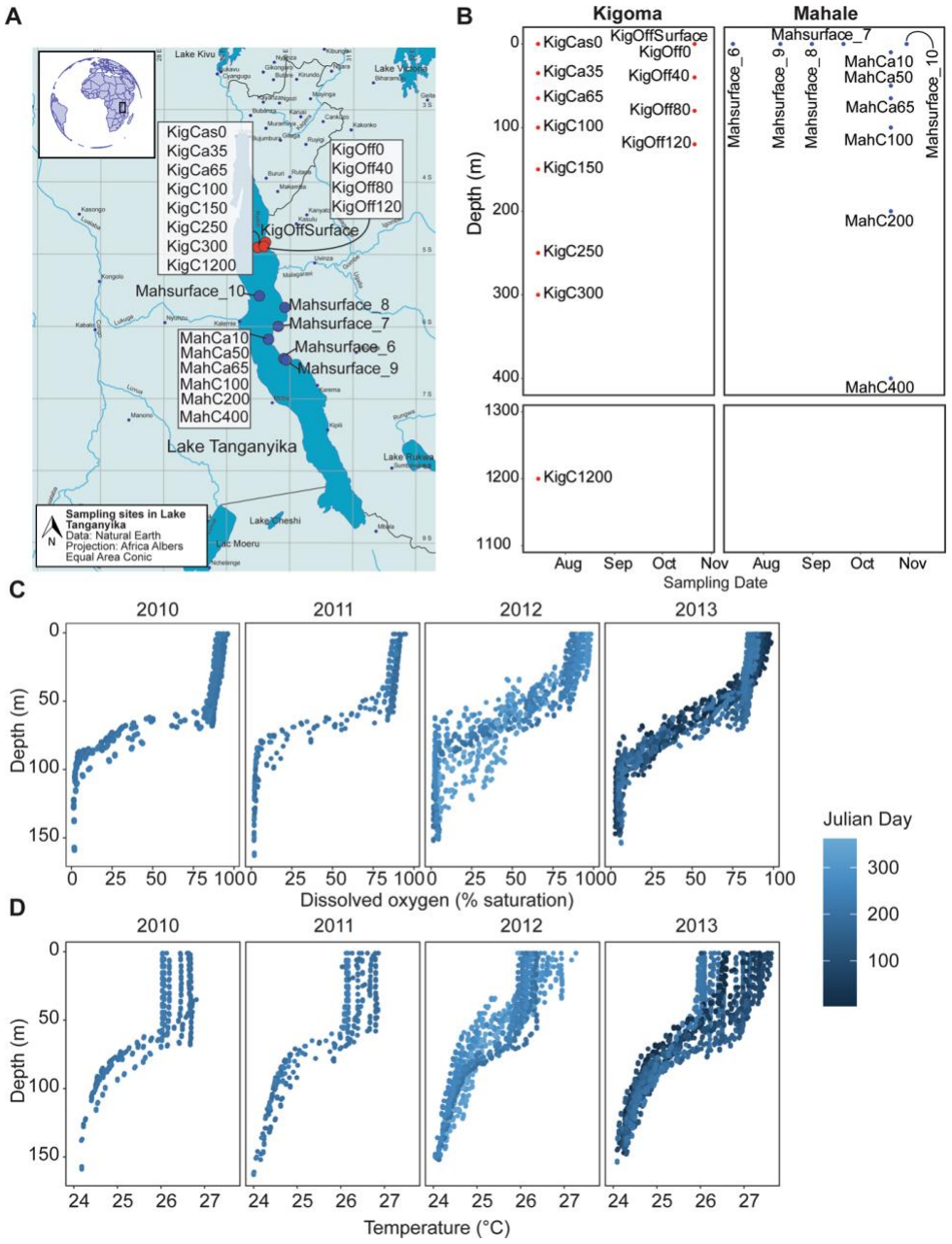

Figure 3.1 Map of sampling locations in Lake Tanganyika. The inset figure shows the approximate location of LT in a spherical global projection. Lakes and rivers are identified in blue;

cities are shown in black. Rivers, lakes, and populous cities are labeled. The locations of the 24 samples collected are labeled, corresponding to Kigoma (red) or Mahale (blue), which are stations in Lake Tanganyika. The white shaded box represents location for which we collected depth-discrete samples (cast). B Sampling locations along the vertical transect of LT. Samples are depicted foremost by location (Kigoma vs. Mahale), followed by sampling date on the x-axis and depth on the y-axis. Short sample names are written next to each dot. C Dissolved oxygen (% saturation) in Lake Tanganyika, in percentage of saturation, from 2010 to 2013. D Temperature profiles, in degree Celsius, from 2010 to 2013 in Lake Tanganyika.

DNA extraction and sequencing

DNA extractions were performed using the MP Biomedicals FastDNA Spin Kit with minor protocol modifications as described previously[89] back in Madison, WI. Metagenomic DNA was sequenced at the Joint Genome Institute (JGI) (Walnut Creek, CA) on the HiSeq 2500 platform (Illumina, San Diego, CA, U.S.A.), which produces 2 × 150 base pairs (bp) reads with a targeted insert size of ~240 bp.

Metagenome assembly, genome binning, dereplication, and selection of genomes

Each of the 24 individual samples were assembled de novo by JGI to obtain 24 metagenomes assemblies. Briefly, raw metagenomic sequence reads were quality filtered, then assembled using MetaSPADEs v3.12 (default parameters)[74]. Each metagenome was binned individually using: MetaBat1[90], MetaBat2[91] v2.1.12, and MaxBin2 v2.2.4[92]. We consolidated the bins generated by these three different tools using DASTool v1.1.0[93] (default parameters except for --score_threshold 0.4), resulting in a total of 3948 MAGs. We dereplicated these MAGs using dRep v2.3.2 with the

default settings[77]. A total of 821 unique clusters were created. Genome completion was estimated using CheckM v1.0.11[68] and DAStool. To select a set of medium and high-quality genomes (≥50% completeness, <10% contamination) for downstream analyses, we used the minimum genome information standards for metagenome-assembled genomes[94], which yielded a total of unique dereplicated 523 MAGs (Table S2). In total, these representative MAGs represent clusters containing 3948 MAGs. The 523 MAGs are used as the dataset for this study.

Relative abundance across water column depths and gene annotations

We mapped each metagenomic paired-read set to each of the 523 MAGs using BBMap [75], with default settings, to obtain a matrix of relative coverage (as a proxy for abundance across the samples) vs. MAG. We used pileup.sh implemented in BBMap, which calculates the coverage values, while also normalizing for the length of the scaffold and genome size [75]. We combined the mapping table from the 24 metagenomes, and summed the total coverage based on the associated MAGs identifier for each scaffold. To normalize the coverage by the metagenome size, we divided the coverage by the number of total reads per metagenome. Open reading frames of the scaffolds were identified using Prodigal v2.6.3 [95].

Identification of phylogenetic markers

A set of curated Hidden Markov Models (HMM) for 16 single-copy ribosomal proteins (rpL2, rpL3, rpL4, rpL5, rpL6, rpL14, rpL14, rpL15, rpL16, rpL18, rpL22, rpL24, rpS3, rpS3, rpS8, rpS10, rpS17, rpS19) [96] was used to identify these genes in each MAG using hmmsearch (HMMER 3.1b2) [79] with the setting --cut_tc (using custom-derived trusted cutoffs). The esl-reformat.sh from hmmsearch was used to extract alignment hits.

Phylogenetic tree

To create the concatenated gene phylogeny, we used reference MAGs, which represented a wide range of environments including marine, soil, hydrothermal environments, coastal and estuarine environments originally built from references [97]. We used hmmsearch to identify the 16 ribosomal proteins for the bacterial tree, and 14 ribosomal proteins for the archaeal tree as described previously[97] (also described in "Identification of phylogenetic markers"). All identified ribosomal proteins for the backbone and the LT MAGs were imported to Geneious Prime V.2019.0.04 (https://www.geneious.com) separately for Bacteria and Archaea. For each ribosomal protein, we aligned the sequences using MAFFT (v7.388, with parameters: Automatic algorithm, BLOSUM62 scoring matrix, 1.53 gap open penalty, 0.123 offset value)[98]. Alignments were manually verified: in the case that more than one copy of the ribosomal protein was identified, we performed a sequence alignment of that protein using MAFFT (same settings) and compared the alignments for those copies. For example, if they corresponded exactly to a split protein, we concatenated them to obtain a full-length protein. If they were the same section (overlap) of the protein, but one was shorter than the other, the longer copy was retained. We applied a 50% gap masking threshold and concatenated the 16 (or 14) proteins. The concatenated alignment was exported into the fasta format and used as an input for RAxML-HPC, using the CIPRES server [99], with the following settings: datatype = protein, maximum-likelihood search = TRUE, no bfgs = FALSE, print br length = false, protein matrix spec: JTT, runtime = 168 h, use bootstrapping = TRUE. The resulting Newick format tree was visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The same procedure was followed to create the taxon-specific tree, for example for the Candidatus Tanganyikabacteria and Nitrospira, using phylum-specific references except with no gap masking

since sequences were highly similar and had few gaps. In addition, an amoA gene phylogeny was performed using UniProt amoA sequences with the two "predicted" comammox genomes with 90% masking (Supplementary Text).

Taxonomic assignment and comparison of manual vs. automated methods

Taxonomic classification of MAGs was performed manually by careful inspection of the RP16 gene phylogeny, bootstrap values of each group, and closest named representatives (Supplementary Material 1 and 2). In addition, we also assigned taxonomy using GTDB-tk [69], which uses ANI comparisons to reference genomes and 120 and 122 marker genes for Bacteria and Archea, respectively, using FastTree (**Supplementary Material 3 and 4**). We compared the results for taxonomic classification between the manually curated and automated approaches to check whether taxonomic assignment matched across phyla, and finer levels of resolution. Since the two trees (RP16 and gtdb-tk tree) are made using different sets of reference genomes, we are unable to do a direct comparison of the taxonomic position of each of the MAGs in our study. Therefore, we manually curated each manual vs. GTDB-tk classification and added a column stating whether the results matched (**Table S2**).

To enhance the MAGs relevance to freshwater microbial ecologists, we assigned taxonomic identities to 16S rRNA genes identified in our MAGs to names in the guide on freshwater microbial taxonomy [25]. 16S rRNA genes were identified in 313 out of 523 MAGs using CheckM's ssu_finder function [68]. The 16S rRNA genes were then used as an input in TaxAss [100], which is a tool to assign taxonomy to the identified 16S rRNA sequences against a freshwater-specific database (FreshTrain). This freshwater-specific database, albeit focused on epilimnia of temperate lakes, is

useful for comparable terminology between the LT genomes and the "typical" freshwater bacterial clades, lineages and tribes terminology defined previously [25]. We also compared taxonomic identification among the three methods and show the results in **Table S2**.

We chose to provide the information from all sources of evidence (RP16, GTDB-tk tree, manual curation, automated taxonomic classification, and 16S rRNA sequences), even if in instances some results might be inconsistent. It is important to note that each source of evidence and reference-based methods are biased by their database content, but we hope that providing several lines of evidence can help arrive to a consensus. For readability, the MAGs in our study are referred by their manually curated phylum or lineage name.

Support for Tanganyikabacteria, a monophyletic sister lineage to Sericytochromatia

We noted three MAGs from LT to be monophyletic, but initially were unrelated to any reference genomes. Since they were placed close to the Cyanobacteria, we created a detailed RP16 tree of 309 genomes of Cyanobacteria, and sister-lineages (Sericytochromatia/Melainabacteria)[101] , Blackallbacteria, WOR-1/Saganbacteria, and Margulisbacteria) including other freshwater lakes (Kopylova, Noé, and Touzet 2012)[24,102], groundwater marine, sediment, fecal, isolates, and other environments. We calculated pairwise genome-level average nucleotide identities (ANI) values between all 309 vs. 309 genomes using fastANI[103]. The 3 genomes from LT were closely related, yet distinct from, the recently defined Cyanobacterial class Sericytochromatia [101,104], using evidence from RP16 and GTDB-tk phylogeny of the MAGs, and ANI with closely related genomes in the literature (Table S3). The other existing Sericytochromatia used for comparison were isolated from Rifle acetate amendment columns (Candidatus Sericytochromatia bacterium S15B-

MN24 RAAC_196; GCA 002083785.1) and a coal bed methane well (Candidatus Sericytochromatia bacterium S15B-MN24 CBMW_12; GCA 002083825)

Comparison of MAG taxonomic diversity in LT vs. LB

To compare taxonomic diversity in two ancient deep lakes, LT and LB, we compare the ANI of metagenome-assembled genomes from the deepest samples from the two lakes. LB has 231 MAGs published[19]. These 231 MAGs were assembled from samples from depths of 1350 and 1250 m. We selected all MAGs that had a read abundance ≥0.5% of the KigC1200 sample (1200 m depth) microbial diversity, resulting in 260 MAGs. We used fastANI to compare all vs. all (491 vs. 491 MAGs) % ANI values. To assess the patterns, we generated histograms and mean ANI values and plotted them in R. We grouped the pairwise matches as Tanganyika vs. Tanganyika, Baikal vs. Baikal, and Tanganyika vs. Baikal (which is the same as Baikal vs. Tanganyika).

Metabolic potential analysis and comparison of metabolic potential and connection across three distinct depths in LT

Metabolic potential of LT MAGs was assessed using METABOLIC, which includes 143 custom HMM profiles[105], using hmmsearch (HMMER 3.1b2) (--use_tc option) and esl-reformat to export the alignments for the HMM hits[79]. We classified the number of genes involved in metabolism of sulfur, hydrogen, methane, nitrogen, oxygen, C1-compounds, carbon monoxide, carbon dioxide (carbon fixation), organic nitrogen (urea), halogenated compounds, arsenic, selenium, nitriles, and metals. To determine if an organism could perform a metabolic function, one copy of each representative gene of the pathway must have been present in the MAG, for which a value of 1 (presence) was written, as opposed to 0 (absence). To investigate heterotrophy associated with

utilization of complex carbohydrates, carbohydrate-degrading enzymes were annotated using hmmscan on the dbCAN2 [106] (dbCAN-HMMdb-V7 downloaded June 2019) database.

The sample profile collected on July 25, 2015 covered a vertical gradient ranging from surface samples to 1200 m. We analyzed samples collected near Kigoma from three different depths (0 m (KigCas0), 150 m (KigCas150), 1200 m (KigCas1200)) based on the general difference in abundance of key taxa. We used METABOLIC v4.0 [105] to identify and visualize organisms with genes involved in carbon, sulfur, or nitrogen metabolism. We combined the results into a single figure showing the number of genomes potentially involved in each reaction, and the community relative abundance of those organisms across three distinct depths. We performed some additional manual curation for differentiating between amo and pmo genes, and annotating nxr genes (Supplementary Text).

In addition to the functional gene annotations done by METABOLIC, which includes HMMs related to biogeochemical cycles, all 24 metagenomic assemblies were annotated by IMG/M, pipeline version 4.15.1 [107].

Results

We sampled a range of physicochemical measurements in LT, which are described in Figs. 1 and S1. We collected 24 metagenome samples from the LT water column spanning 0–1200 m, at two stations Kigoma and Mahale in 2015 (Figure 3.1), which consisted of 18 depth-discrete samples and six surface samples. Despite not having paired physicochemical profiles in 2015, temperature and DO from 2010 to 2013 were highly consistent and showed minimal interannual variability, particularly during our sampling period (July and October) (Figure 3.1and S2). In addition, the environmental profiles are similar to those collected in other studies[108–111]. Water column

temperatures ranged from 24 to 28 °C, and changes in DO were greatest at depths ranging from ~50 to 100 m during the time of sampling dropping to 0% saturation DO around 100 m. The thermocline depths shift vertically depending on the year (Fig S2). Secchi depth, a measure of how deep light penetrates through the water column, was on average 12.2 m in July and 12.0 m in October (Fig S3). Nitrate concentrations increased up to ~100 µg/L at 100 m deep, followed by rapid depletion with the onset of anoxia (Fig S4). A consistent chlorophyll-a peak was detected at a depth of ~120 m in 2010–2013 (Fig S5). Comparatively in 2018, the peak occurred around 50 m[111]. Other profiles collected in 2018 show a nitrate peak between 50 and 100 m in LT [111].

The microbiome of Lake Tanganyika

Metagenomic sequencing, assembly, and binning resulted in 3948 draft-quality MAGs that were dereplicated and quality-checked into a set of 523 non-redundant medium- to high-quality MAGs for downstream analyses. To assign taxonomic classifications to the organisms represented by the genomes, we combined two complimentary genome-based phylogenetic approaches: a manual RP16 approach, and GTDB-tk[69], an automated program which uses 120 concatenated protein coding genes. For the most part, we observed congruence between the two approaches (Fig S6) and the genomes represented 24 Archaea and 499 Bacteria from 34 phyla, 74 classes, and 118 orders (Table S2 and Figure 3.2). While manually curating the RP16 tree, and upon closer inspection and phylogenetic analysis of Cyanobacterial genomes and sister lineage genomes (Sericytochromatia, Melainabacteria, etc.), we noticed that three bacterial genomes formed a monophyletic freshwater-only clade within Sericytochromatia, sharing <75% genome-level ANI to other non-photosynthetic Cyanobacteria-like sequences (Figure 3.3 and Table S3). On this basis, we propose to name this lineage Candidatus Tanganyikabacteria (named after the lake).

Sericytochromatia is a class of non-photosynthetic Cyanobacteria-related organisms, that has recently gained attention along with related lineages such as Melainabacteria and Margulisbacteria due to the lack of photosynthesis genes, indicating phototrophy was not an ancestral feature of the Cyanobacteria phylum[101,104]. This is the first recovery of Sericytochromatia from freshwater lake environments, since others were found in glacial surface ice, biofilm from a bioreactor, a coal bed methane well, and an acetate amendment column from the terrestrial subsurface.

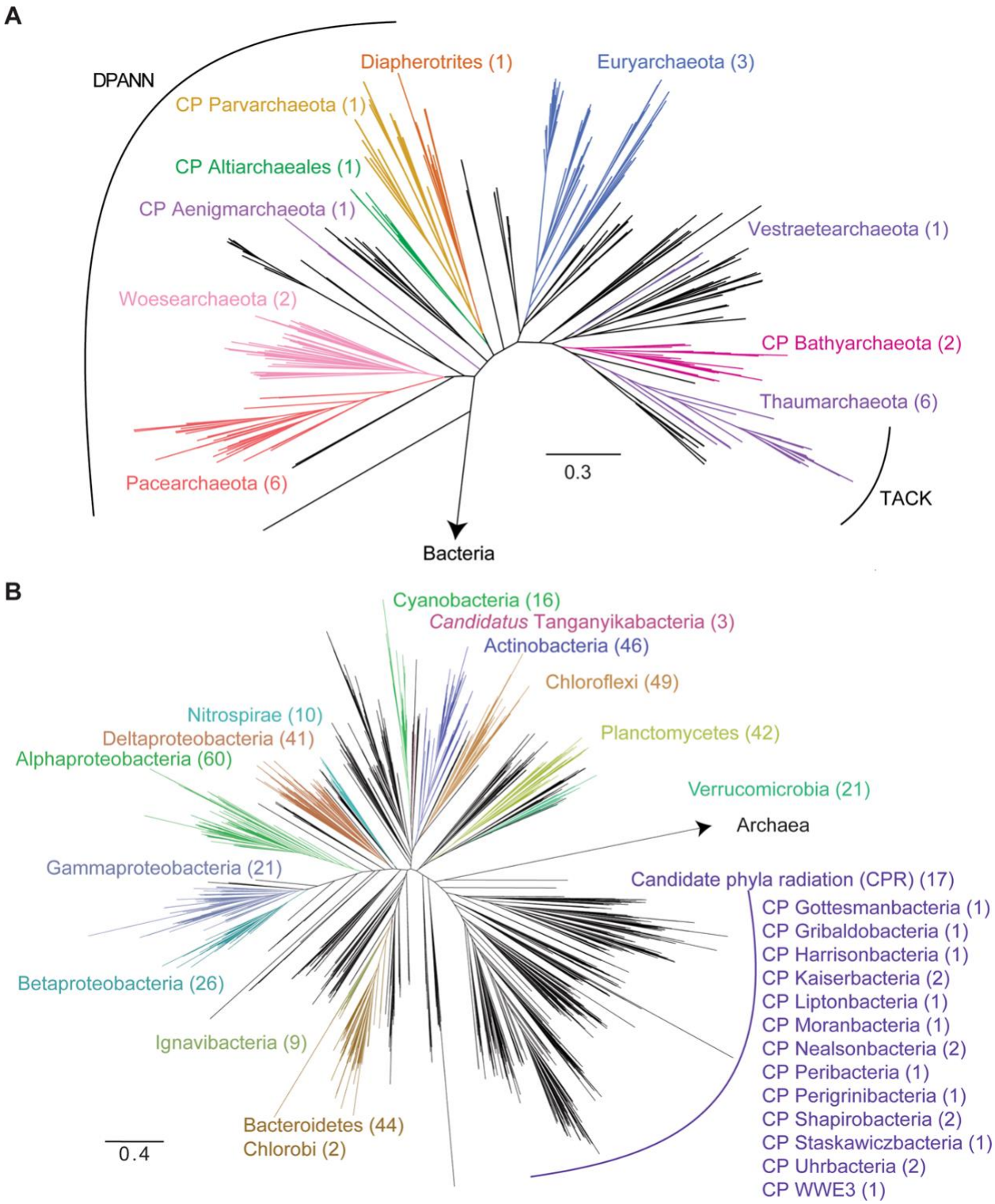

Figure 3.2 Phylogeny of Archaeal and Bacterial metagenome-assembled genomes (MAG) recovered from LT. Phylogeny of (**A)** Archaeal and (**B)** Bacterial metagenome-assembled

genomes (MAG) recovered from LT. The tree was constructed using 14 and 16 concatenated ribosomal proteins, respectively, and visualized using FigTree. Not all lineages are named on the figure. The groups DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota) and TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota) archaea, CPR (candidate phyla radiation) bacteria are labeled. The number of medium- and high-quality MAGs belonging to each group are listed in parentheses. Colored groups represent the most abundant lineages in LT. Tanganyikabacteria MAGs from this study are italicized. A more detailed version of the tree constructed in iTol is found in Supplementary Material 1 and 2 with bootstrap values and names of taxa.

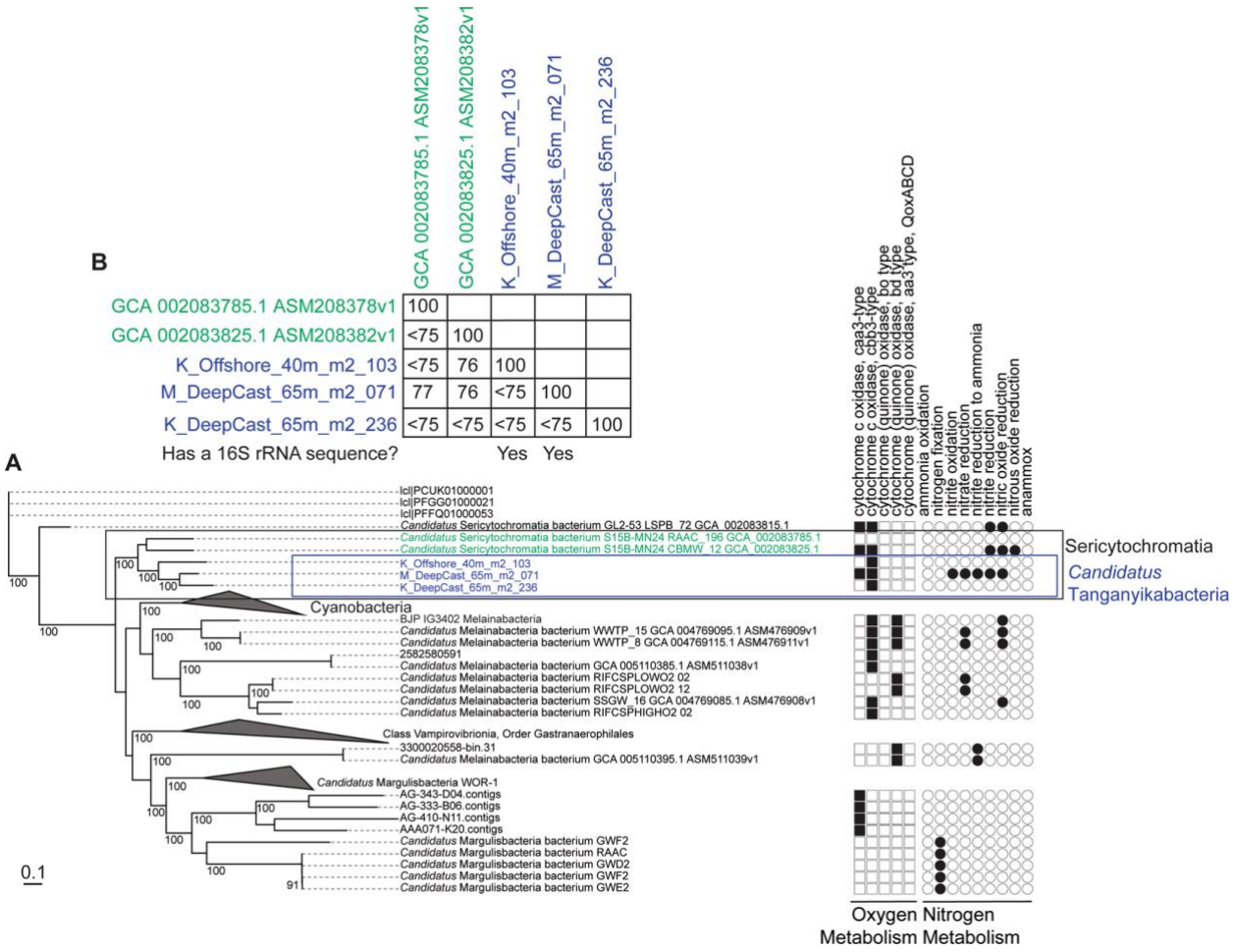

Figure 3.3 Concatenated gene phylogeny of Cyanobacteria and their non-photosynthetic sister-lineages, and comparison of average nucleotide identities. **A**. Concatenated genome phylogeny using 16 ribosomal proteins of metagenome-assembled genomes of Cyanobacteria and non-photosynthetic sister-lineages such as Margulisbacteria (WOR-1), Melainabacteria, and Sericytochromatia. The three MAGs from Lake Tanganyika are labeled in blue, and represent a high-support (100) monophyletic lineage among the known Sericytochromatia. The presence–absence plot shows genes involved in oxygen metabolism (squares) and nitrogen metabolism (circles). The MAG from Lake Tanganyika is the only one among all genomes to have genes for denitrification. M_DeepCast_65m_m2_071 has ~97% genome completeness (full uncollapsed

tree available in Supplementary Material). **B**. Genome-level average nucleotide identity (ANI) (%) are shown as pairwise matrix for the five Sericytochromatia MAGs (boxed).

To investigate the stratification of microbial populations and metabolic processes in the water column of LT, we identified three zones based on oxygen saturation. The microbial community composition and community relative abundance (calculated as relative abundance of reads mapped, RAR) across our samples (Fig S7) were distinct between the oxygenated upper layers and the deep anoxic layers (Figure 3.4 and Table S4). Notably, Archaea accounted for up to 30% of RAR in sub-oxic samples (Kigoma 80 and 120 m), and generally increased in abundance with depth. DPANN archaea were only found in sub-oxic (>50 m deep) samples. Candidate phyla radiation (CPR) organisms generally increased in abundance with depth, reaching up to ~2.5% RAR. Among bacteria, notably, common freshwater taxa such as Actinobacteria (e.g., acI, acIV) Alphaproteobacteria (LD12), and Cyanobacteria showed a ubiquitous distribution, whereas certain groups such as Chlorobi and Thaumarchaetota were most abundant below 100 m. The community structure, composition, and abundance of the surface samples in the five Mahale samples were similar throughout. Note that the naming (e.g., K for Kigoma, M for Mahale, depths) of the LT MAGs originate from the sample they were originally binned (Table S1), however, it does not necessarily reflect that the MAG had the highest % RAR in that sample, since all representative MAGs are the result of dereplication (see “Methods”).

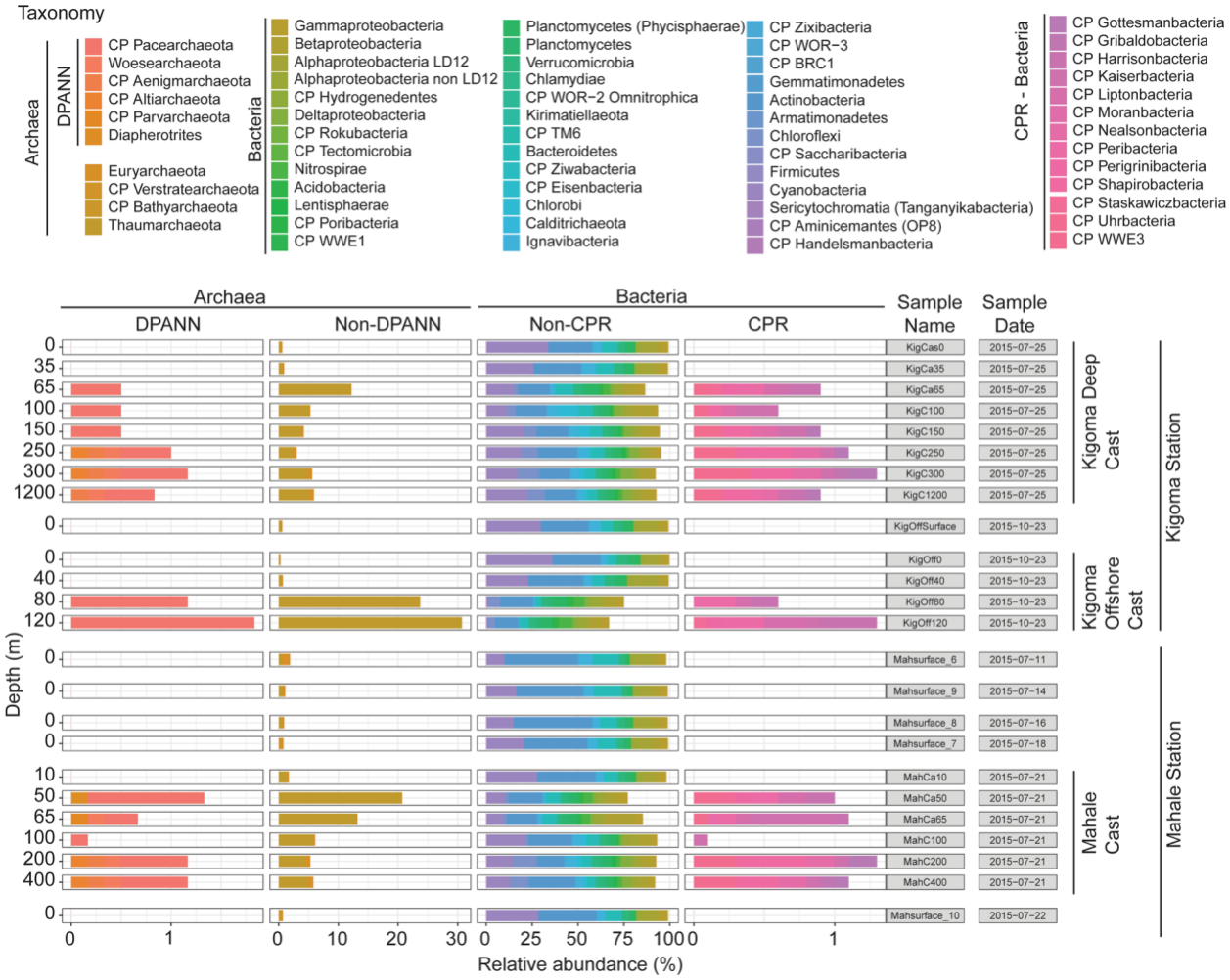

Figure 3.4 Bar plot showing the relative read abundance (RAR) per sample in the 24 metagenomes from Lake Tanganyika. MAGs are taxonomically ordered along the x-axis, with closely related taxa next to each other. In addition, information about whether they are Archaea (DPANN or not) and Bacteria (CPR or not) is specified. The three casts are identified with a vertical line on the right-hand side. Samples are organized by location, then sampling date, then depth.

How similar is the LT microbiome to that of other deep and ancient freshwater lakes?

While a comparison of LT's metagenomic diversity would be interesting to compare with other African Great Lakes (such as Lake Malawi, Lake Kivu, Lake Victoria), no published MAGs from those sites existed at the time of writing. We compared LT and LB's microbial diversity (based on MAGs) because they are both ancient, extremely deep lakes, and therefore might both have sufficient evolutionary time for a wide diversity of bacterial and archaeal lineages to evolve. Yet, both lakes differ drastically in terms of environmental settings (LB is seasonally ice-covered with a oxygenated hypolimnion, whereas LT is a tropical ice-free lake with an anoxic bottom water layer). To compare the microbial communities, we compared the MAGs recovered from LT and LB (Figure 3.5, S8 and Table S5).

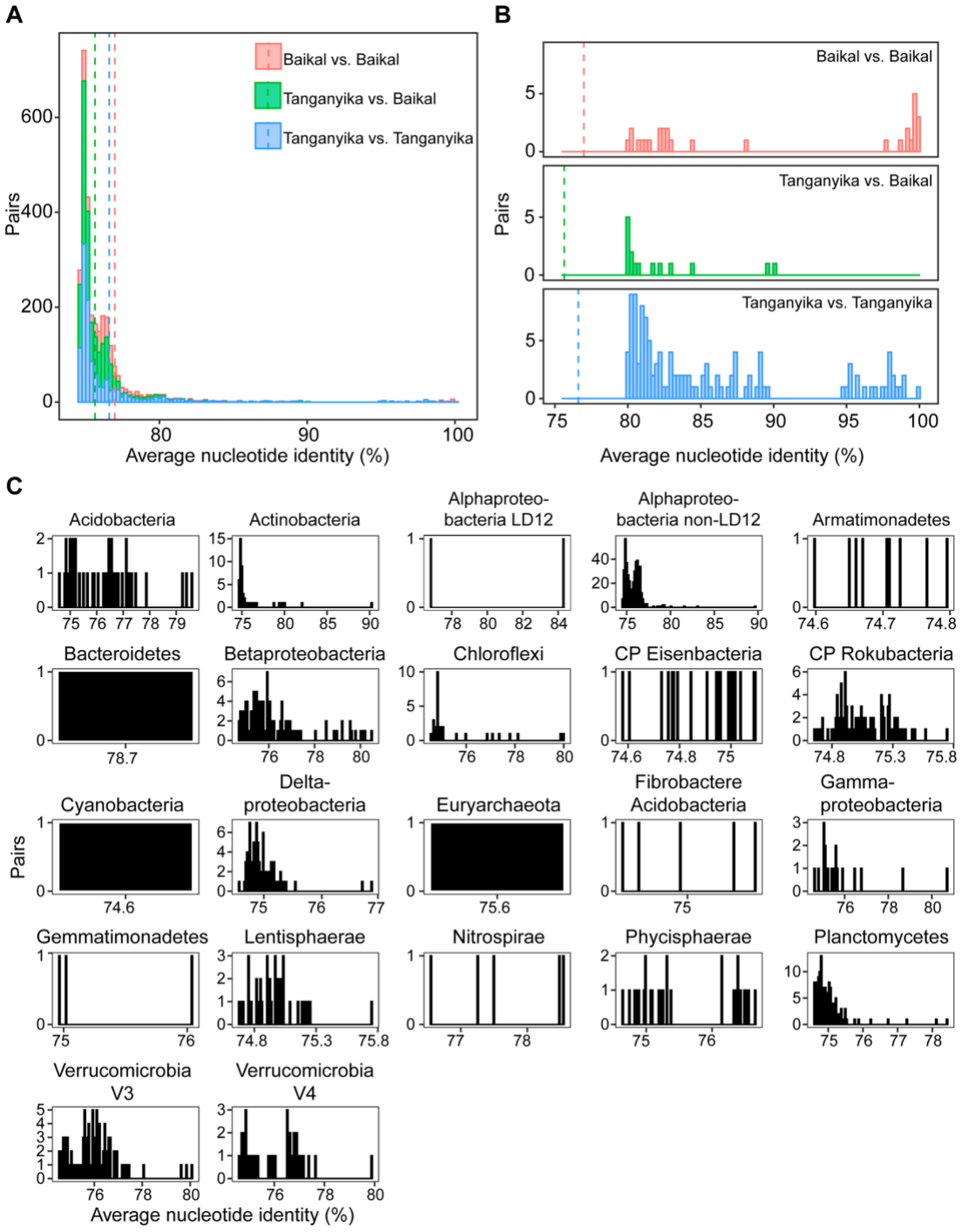

Figure 3.5 Comparison of metagenome-assembled genomes in Lake Tanganyika and Lake Baikal. A Histogram of the ANI values when doing pairwise comparison of full-genome MAGs from Lake Tanganyika vs. Lake Baikal. The vertical dashed lines correspond to the mean value

(Baikal vs. Baikal: 79.98%, Tanganyika vs. Baikal: 75.63%, and Tanganyika vs. Tanganyika 76.60%). B Zoom in ANI values above 80% only. There are slightly more ANI above the 97% ANI for Baikal vs. Baikal, than Tanganyika vs. Tanganyika. Two genomes from Baikal had 100% ANI, two genomes from Tanganyika (K_Offshore_surface_m2_005 and K_DeepCast_100m_m2_150) shared 99.91% ANI. The highest % ANI between a MAG from Tanganyika vs. Baikal was among M_surface_10_m2_136 and GCA_009694405.1_ASM969440v1_genomic (90.19% ANI), which were Candidatus Nanopelagicaceae bacterium, an Actinobacteria. C Of the pairs in "Tanganyika vs. Baikal" category, showing the number of pairwise comparisons for each % ANI value, organized by taxonomic group. In C, all the x-axes are "% ANI." For example, there is only one pair of Bacteroidetes, Cyanobacteria, and Euryarchaeota from the two lakes that have >75% ANI.

Comparing the MAGs from the 1200 m sample in LT (KigOff1200, >0.05 % RAR) and MAGs from deep samples (1200 and 1350 m) from Baikal, relatively high RAR of Thaumarchaeota were identified in both lakes. Thaumarchaeaota accounted for ~10% RAR in LT and ~20% in LB. Typical freshwater taxa such as the acI lineage of Actinobacteria and LD12 group of Alphaproteobacteria (Ca. *Fonsibacter*) were observed in higher abundance in surface samples but also at depths up to 50 m. CPR bacteria were found in both lakes. Archaea accounted for 6.3% of RAR in KigOff1200m, and about 2% in LB. Archaeal MAGs from KigOff1200 (and >0.05% RAR) belonged to the lineages Bathyarchaeota, Verstraetaerchaeota, Thaumarchaeota, Euryarchaeota, and DPANN (Pacearchaeota, Woesearchaeota, Aenigmarchaeota). The three DPANN MAGs from LB were closely related to Pacearchaeota and Woesearchaeota.

Interestingly, the majority of taxonomic groups observed in the deepest waters of LT (1200 m) was unique to LT (59%), whereas 68% of taxa richness recovered from LB that were bacterial MAGs were also identified in the deepest LT sample. *Nitrososphaerales* (Archaea) were identified in both LT and LB. Both LT and LB have a small abundance of Cyanobacteria. In LB, they account for ~1%, and are likely sourced from vertical mixing or sediment. However in LT, Cyanobacteria are more abundant, accounting for 14% RAR, but the lake is stratified with no mixing. Organisms from the lineage Desulfobacterota, with prominent roles in sulfur cycling (H2S generation) were identified in LT but not in LB. Overall, both lakes had high abundance and richness of Archaea and CPR, although the specific lineages (for example Desulfubacterota) likely reflect the differences in geochemistry in the lake, and to the different oxygen niches that these organisms may occupy.

Depth-dependent contrasts in microbial metabolism in LT: biogeochemical cycling of carbon, nitrogen, and sulfur

We investigated microbial metabolic potential and process-level linkages between MAGs in the surface, at 150 m, and at 1200 m, representing three distinct ecological layers within the lake (Figure 3.6 and Tables S6, 7). Individual metabolic pathways were identified in each MAG and their capacity to contribute to different (Fig S9 and Tables S6, 7), nitrogen (Fig S10 and Table S6, 7), and sulfur (Fig S11 and Table S6, 7) biogeochemical cycles were assessed (Fig S12 and Table S6, 7). Based on historical environmental data[108], light sufficient to support photosynthesis is available to ~70 m, and these surface waters are oxygen-rich and nutrient-depleted. The 150 m depth is subject to more interannual year variation in temperature and DO (Figs. S2–5). The 1200

m sample represents a relatively stable, dark, nutrient-rich, and anoxic environment, according to data collected from 2010 to 2013 in our study and in accordance with historical profiles.

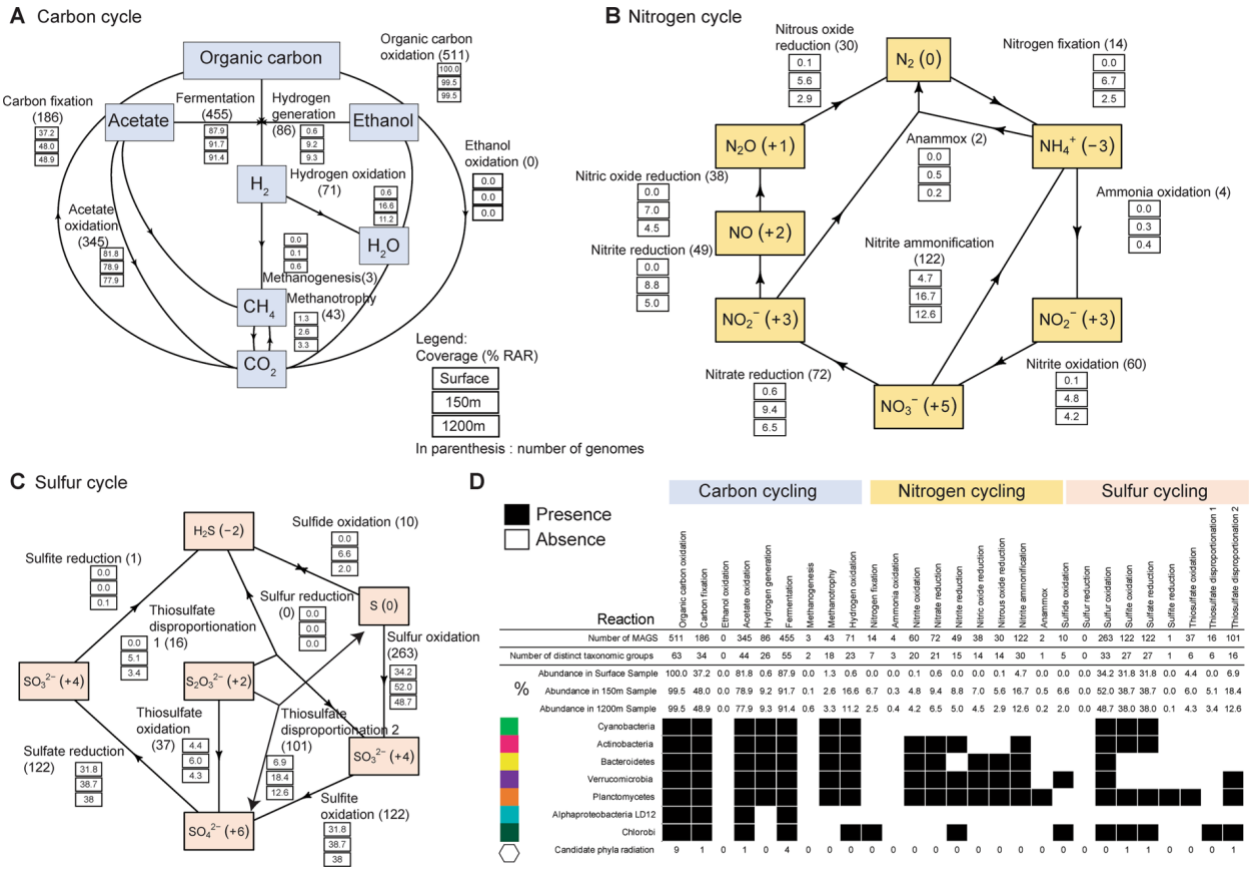

Figure 3.6 Microbial contributions to carbon, nitrogen and sulfur cycling in Lake Tanganyika Abundance of organisms involved in different **A** carbon, **B** nitrogen, and **C** sulfur cycling steps, in three depth-discrete samples from Kigoma (surface, 150 m, 1200 m).Oxidation states are shown in parentheses. **D** Presence and absence of individual steps in C, N, S cycling. Abundance of organisms is described in percentages. Only reactions in a subset of common freshwater taxa (Cyanobacteria, Actinobacteria, Bacteroidetes, Verrucomicrobia, Planctomycetes, Alphaproteobacteria (LD12), Chlorobi) and candidate phyla radiation are shown. Presence and absence of these pathways in taxonomic groups are represented by filled and open circles, respectively.

Microbial pathways associated with the carbon cycle included the use of organic carbon (e.g., sugars), ethanol, acetate, methanogenesis, methanotrophy, and CO_2-fixation (Figure 3.6). As would be expected, organisms capable of organic carbon oxidation were abundant comprising nearly 99% RAR community in the sub-oxic and anoxic samples. The metabolic potential for fermentation and hydrogen generation was also abundant in the sub-oxic and anoxic samples, represented by 91% RAR (455 MAGs) and 9% RAR (86 MAGs), respectively. Other anerobic processes including methanogenesis (3 MAGs) and methanotrophy (43 MAGs) were identified in a limited number of MAGs, but were observed to be more abundant in the 1200 m samples. Both bacteria and archaea encoded carbohydrate-degrading enzymes (CAZYmes) (Fig 4.S13 and Table S8). The highest densities of CAZYmes (when normalized by genome size) were identified in organisms from the lineages Verrucomicrobia, Lentisphaerae, and CP Shapirobacteria. Meanwhile, glycoside hydrolases (GHs) that participate in the breakdown of different complex carbohydrates were prominent in Verrucomicrobia and Planctomycetes, as has been observed in other studies of these common freshwater linneages in lakes[112,113].

We identified microorganisms involved in the inorganic nitrogen cycle, including oxidation and reduction processes (Figure 3.6 and Table S6, 7). The metabolic potential for nitrogen cycling was generally higher in the sub-oxic samples as compared to the anoxic samples. Nitrogen fixation capacity (14 MAGs) was found below the surface, and nitrogen-fixing organisms comprised up to 7% RAR at each depth. We identified a MAG belonging to the Thaumarchaeal class Nitrososphaeria that contains Archaeal *amoABC* genes for ammonia oxidation (Supplementary Material 5), and many non-CPR Bacteria with *nxr* genes involved in nitrite oxidation (Fig 4.S14).

Nitrospira bacteria were identified to be capable of complete ammonia oxidation (comammox) (Clade II-A), with highest RAR in the sub-oxic depths (Fig S15). Denitrification processes that remove nitrate and nitrite from the system were discovered across all depths, at approximately the same RAR in all depths. For example, the potential for nitrate reduction was identified in 72 MAGs that comprised 6.5–9% RAR and for nitrite reduction in 49 MAGs, at 5–9% RAR in anoxic depths. The nitrate/nitrite ammonification (DNRA) potential was identified in 122 MAGs accounting for between 5 and 17% RAR in each depth. Bacteria capable of anammox accounted for between 0.2 and 5% RAR at each depth.

Sulfur biogeochemical profiles exist in LT and generally show an increase in H_2S beginning in the sub-oxic zone into the anoxic zone with the highest concentration observed in the deepest waters[108]. H_2S can accumulate in the water column through sulfite reduction, thiosulfate disproportionation, and sulfur reduction. While only a few MAGs were potentially able to perform these processes (1, 16, and 10, respectively), we observed that all of them increased in RAR with depth and were more abundant in the anoxic depths (Figure 3.6). Sulfur oxidation (263 MAGs), sulfite oxidation (122 MAGs), and sulfate reduction were each prominently represented processes, with such organisms accounting for over 30% RAR at each depth.

To understand the behaviors of carbon, sulfur, and nitrogen biogeochemical cycles in the LT water column, we investigated the depth distribution of different microbially mediated processes. We observed differences amongst the carbon, sulfur, and nitrogen cycles. Processes that were more prominent at sub-oxic and anoxic depths, but still present at low (<5% RAR) from the surface include nitrogen cycling (nitrite oxidation, nitrous oxide reduction, nitrite ammonification, nitrate

reduction) and carbon cycling (methanotrophy, hydrogen oxidation, and hydrogen generation). Processes that were exclusive to the sub-oxic and anoxic depths include carbon cycling (methanogenesis), nitrogen cycling (nitrogen fixation, ammonia oxidation, nitrite oxide reduction, nitric oxide reduction, anammox), and sulfur cycling (sulfide oxidation, sulfite reduction, thiosulfate disproportionation H_2S and SO_3^{2-}). The potential for use of alternative electron acceptors such as chlorate, metals, arsenate, and selenite was also identified in organisms in sub-oxic and anoxic depths.

Given our earlier finding that the water column was populated with common freshwater taxa in the upper mixed layer, vs. less common organisms in the anoxic water column, we wanted to know if the differences were reflected in metabolism of carbon, nitrogen, and sulfur across these depths. We identified the metabolic potential of organisms from the lineages Cyanobacteria, Actinobacteria, Alphaproteobacteria (LD12), Verrucomicrobia, Planctomycetes, and Bacteroidetes (Figure 3.6). We counted the number of distinct phylogenetic taxa with the potential to conduct a biogeochemical transformation. The potential for nitrogen fixation was not identified in Cyanobacteria MAGs from the water column. The Cyanobacteria in LT were all identified as belonging to Synechoccales, which are known for not being nitrogen fixers [114–117]. Instead, a small group of microorganisms from seven distinct taxa, which were mostly abundant at sub-oxic and anoxic depths, was observed to be capable of nitrogen fixation (Deltaproteobacteria, Kirimatiellacea, Alphaproteobacteria (non-LD12), Chloroflexi, Chlorobi, Euryarchaeota, Gammaproteobacteria). The inferred involvement of microbes in biogeochemical cycling across an oxygen gradient such as in LT is summarized in Figure 3.7.

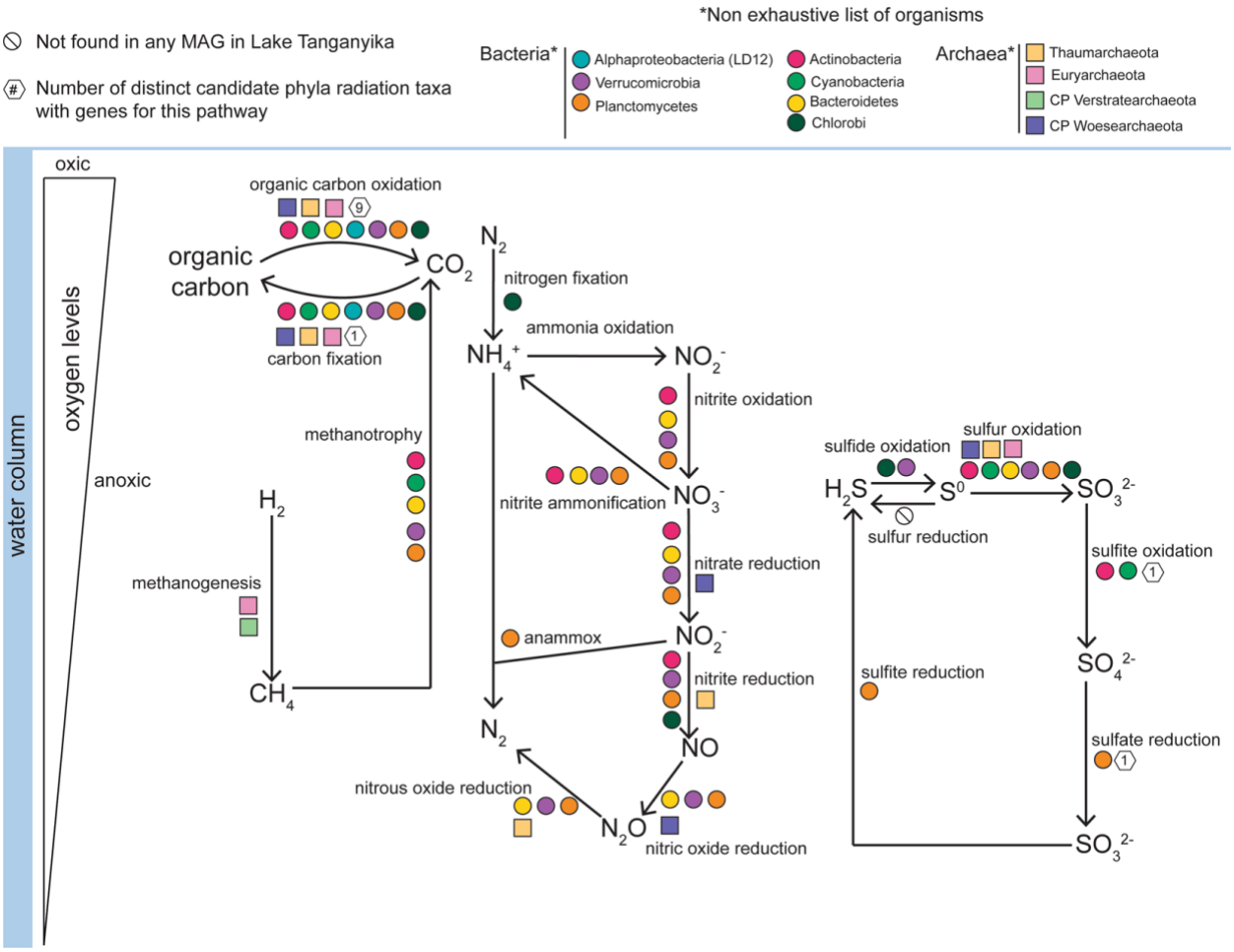

Figure 3.7 Summary figure showing the role of different microbial taxa in carbon, nitrogen, and sulfur cycling in Lake Tanganyika. The list of organisms shown and the reactions are not exhaustive.

Discussion

The natural history of LT, along with its environmental characteristics, makes it ideal for examining the role of microbial communities in nutrient and biogeochemical cycling in tropical freshwater lakes. Being an ancient, deep, stratified, and meromictic lake with varying oxygen concentrations, it also serves as a case study for microbial diversity and metabolism in permanently

anoxic environments. Here we highlight that microbial diversity and function in LT is distinct between its oxic mixed surface waters, and its anoxic deep waters. We note that although MAGs have key genes to participate in a given process, this does not necessarily mean that the corresponding organisms are actively using them all the time. As such, the purpose of our study was to provide an overview of the metabolic potential for microbially driven biogeochemical processes in LT and to guide future hypotheses for the study of tropical freshwater lakes under stress from a changing climate.

Comparison of microbiomes between two deep rift valley lakes, LT and Baikal: endemism vs. shared lineages

LT (1470 m) and LB (1642 m) are the world's two deepest freshwater lakes, yet are drastically different in their water column characteristics. In addition, both lakes are ancient lakes: LT is approximately ten million years, and LB is 25 million years. Since few metagenomic studies have been conducted on deep freshwater lakes, a comparison of LT and LB is informative in addressing questions about ancient microbial lineages, evolution, and endemic or shared lineages. As mentioned previously, LT is tropical and has an anoxic bottom layer. LB is located in Siberia, is seasonally ice-covered (twice a year), is dimictic (mixed twice a year), and has an oxygenated bottom layer. Both are ancient and formed geologically through rifts. LB's microbial community has been studied using metagenomics, from depths ranging from 0 to 70 m during the ice-cover period[85,118] and from 1250 to 1300 m deep collected in March 2018 also during an ice-covered period[19]. LT's microbiome has been studied via 16S rRNA amplicon sequencing in the past[86–88] and in this study via metagenomics. While LT is primarily anoxic below ~100–50 m, LB has deep-water currents near the coastal regions, which results in oxygenation of deep waters. The regular

mixing from surface waters from LB may imply lower endemism, as the "deep-water" ecological niche would be more frequently disturbed than in meromictic LT.

On one hand, because both LT and LB are ancient, deep, and rift valley freshwater lakes, we hypothesized that they might have a core microbiome common to both (shared lineages). However, given their contrasting environments (cold vs. tropical, oxic vs. anoxic), we expected that environmental characteristics driven by oxygen presence would be major drivers of microbial community structure and function, such as is the case in soils[119], sediments[120], marine oxygen-minimum zones[121] rather than ecosystem type, especially with the increased importance of nitrogen and sulfur cycling under low-oxygen conditions. In our study for example, we find taxa such as sulfur cycling Desulfobacterota only in LT and not in Baikal, likely reflecting that these H_2S producers are found in low-oxygen environments.

Metabolic connections across vertical gradients in LT

Our metabolic function analyses show the distribution of microbially driven carbon, nitrogen, and sulfur cycling across a depth gradient in LT. The carbon cycle of LT has mostly been investigated from the perspective of primary production, and methane cycling. It has been suggested that anaerobic methane oxidation contributes to decreasing the amount of methane in the anoxic zone of LT[82]. Methane is a greenhouse gas that is important in LT, particularly with respect to stratification. Because methane is more abundant below the oxycline, a shift in the oxycline could lead to methane release to the atmosphere via ebullition. Bacterial activity is thought to be a source of methane in nearby (~470 km away) Lake Kivu[14]. Major findings in methanogenesis in aquatic ecosystems now bring light to previously unrecognized biological sources and sinks of methane[122,123]. We and others[82] have found methanogenic Archaea in LT.

Nitrogen cycling has been well-studied in LT, mostly focused on nitrogen inputs and outputs, and from a chemical and biogeochemical perspective. For example, atmospheric deposition and nitrogen from river inflow contribute much of the nitrogen to LT[124]. In addition, it is shown that incoming nitrogen in the form of nitrate is quickly used by organisms in the surface waters, and the upper water column is often nitrate depleted. Profiles of nitrogen (NH_4^+, NO_2^-, NO_3^-) in LT show that NH_3 accumulated from 200 to 1200 m, NO_2^- is usually overall low throughout the water column, and NO_3^- is low at the surface, peaks at the oxycline, and declines where NH_4^+ increases [108]. Another way that NH_4^+ can be supplemented to the water column is through nitrogen fixation. In freshwater systems, it is typically thought that this process is conducted by Cyanobacteria, particularly the heterocystous *Anabaena flos-aquae*[125]. In Lake Mendota, a eutrophic temperate freshwater lake, metagenomic data have shown that a third of nitrogen fixation genes originate from Cyanobacteria, and the remaining from Betaproteobacteria and Gammaproteobacteria[102]. However, we were surprised to find that nitrogen fixation was not identified in any LT Cyanobacteria MAGs from the water column. Benthic nitrogen fixers (cyanobacteria and diatoms with endosymbionts) are dominant in the littoral zone of LT and can provide as much as 30% of N in nearby lake Malawi[126]. Rather, from the metagenomic data of free-living planktonic bacteria, nitrogen fixation was identified to potentially be performed in other groups such as Deltaproteobacteria, Kirimatiellacea, Alphaproteobacteria, Chloroflexi, Chlorobi, Euryarchaeota, and Gammaproteobacteria. Similar results have been observed in Trout Bog Lake, a stratified humic lake where the nitrogen fixers were more diverse and also included Chlorobi[102], like in LT. From the chemical evidence, we hypothesized that the oxycline is a hotspot for nitrite oxidation, and denitrification occurs below the oxycline. Nitrite oxidation pathways in LT MAGs

show that many organisms are potentially able to perform the reaction; however, the nxr enzyme may act reversibly (NO_2^- to NO_3^- and vice-versa). Denitrification processes (nitrate reduction, nitrite reduction, nitric oxide reduction, and nitrous oxide reduction) have similar representation across all three depths. A study of nitrogen cycling processes in LT[127] identified very low dissolved inorganic nitrogen concentrations in the euphotic zone, and proposed that very active N cycling must be occurring in the water column to prevent this accumulation. Another process leading to low dissolved inorganic nitrogen concentration is anammox. Incubations with 15N-labeled nitrate have showed that anammox was active in the 100–110 m water depth, and was comparable to rates observed in marine oxygen-minimum zones[88]. Our detection of Planctomycetes bacteria capable of anammox (Brocadiales) supports this conclusion.

Sulfur cycling in freshwater lakes has received less attention compared to carbon, nitrogen, and phosphorus. Profiles from LT show a clear increase in H_2S below the oxycline, stabilizing at a concentration of ~30-μM H_2S from 300 to 1200 m depth [108]. Biologically, H_2S is produced via sulfite reduction, thiosulfate disproportionation, sulfur reduction, all of which were identified exclusively in the sub-oxic and anoxic depths. Given the use of sulfur-based electron acceptors as alternatives to oxygen in anoxic systems, sulfur cycling has been documented to increase in importance with lower-oxygen availability. As one of the critical compounds for life, sulfur availability and processing are essential to support cellular functions. In LT, where a large portion of the water column is anoxic, sulfur cycling is expected to support primary productivity in the upper water column, and have an overall important effect on lake ecology and biogeochemistry[128]. While the abundance of sulfur cycling organisms suggests an active sulfur cycle in LT, intriguingly, the abundance of organisms mediating the oxidation of reduced sulfur compounds

(H_2S, S^0, $S_2O_3^{2-}$) far exceeded the abundance of organisms mediating the reduction of sulfur compounds (SO_3^{2-}, SO_4^{2-}, S^0). While this discrepancy may arise from increased enzymatic activity of sulfate/sulfite and sulfur reducing organisms, it may also be associated with the geochemistry of the deep waters in LT. Being a rift valley lake, it is home to hydrothermal vents, in which reduced compounds and gases from the Earth's crust including H_2, H_2S, and CH_4 may be produced[129,130]. Around these hydrothermal vents, white and brown microbials mats have previously been identified[130] and microbial life is abundant[131]. In marine environments, Beggiatoa mats are associated with sulfide oxidation[132], and Zetaproteobacteria are often the primary iron oxidizers in iron-rich systems, creating brown mats[133].

Conclusion

Our study provides genomic evidence for the capabilities of Bacteria and Archaea in tropical freshwater biogeochemical cycling and describes links between spatial distribution of organisms and biogeochemical processes. These processes are known to impact critical food webs that are renowned for their high biodiversity and that serve as important protein sources for local human populations, as the shoreline of LT is shared by four African countries. As a permanently stratified lake, LT also offers a window into the ecology and evolution of microorganisms in tropical freshwater environments. Our study encompassed the lake's continuous vertical redox gradient, and shows that microbial communities were dominated by core freshwater taxa at the surface and by relatively high abundances of Archaea and uncultured candidate phyla in the deeper anoxic waters compared with freshwater lake counterparts. The high prominence of Archaea making up to 30% of the community composition abundance at lower depths further highlighted this contrast. While freshwater lakes are often limited by phosphorus[134], tropical freshwater lakes are frequently

nitrogen-limited[135]. Our microbe-centric analyses reveal that abundant microbes in LT play important roles in nitrogen transformations that may remove fixed nitrogen from the water column, or fix nitrogen that may be upwelled and replenish the productive surface waters in bioavailable forms of fixed nitrogen.

Tropical lakes are abundant on earth, yet understudied both in the limnological literature and in the context of microbial metabolism and biogeochemistry. Yet, it is critically important to understand the biogeochemistry of tropical lakes, as lake ecosystem health is critical to the livelihood of millions of people, for example in LT[136,137]. With increasing global temperatures, extremely deep lakes such as LT will likely experience increased stratification, lower mixing, and increased anoxia[138]. We provide the most comprehensive metagenomics-based genomic study to date of microbial community metabolism in an anoxic tropical freshwater lake, which is an essential foundation for future work on environmental adaptation, food webs, and nutrient cycling. Overall, our study will enable the continued integration of aquatic ecology, "omics" data, and biogeochemistry in freshwater lakes, that is key to a holistic understanding of ongoing global change and its impact on surface freshwater resources.

Data availability

The MAGs can be accessed on NCBI BioProject ID PRJNA523022. The individual accession IDs for each genome are in Table S2. The genomes will be officially released on NCBI Genbank upon publication. All MAGs are also publicly available on the Open Science Framework: https://osf.io/pmhae/. The 24 raw, assembled, and annotated metagenomes are available on the Integrated Microbial Genomes & Microbiomes (IMG/M) portal using the following IMG Genome

IDs: 3300020220, 3300020083, 3300020183, 3300020200, 3300021376, 3300021093, 3300021091, 3300020109, 3300020074, 3300021092, 3300021424, 3300020179, 3300020193, 3300020204, 3300020221, 3300020196, 3300020190, 3300020197, 3300020222, 3300020214, 3300020084, 3300020198, 3300020603, 3300020578. An interactive version of the Archaeal and Bacteria trees can be accessed at iTOL at: https://itol.embl.de/shared/patriciatran. Code to generate the figures, and access to the custom HMM for the 16 ribosomal proteins and the metabolic genes is available at https://github.com/patriciatran/LakeTanganyika/ .

Chapter 4 Biogeochemistry Goes Viral: Towards a Multifaceted Approach To Study Viruses and Biogeochemical Cycling.

This chapter is published in mSystems. Citation:

Tran, Patricia Q., Karthik Anantharaman. (2021) "Biogeochemistry Goes Viral: towards a Multifaceted Approach to Study Viruses and Biogeochemical Cycling". mSystems. 6. No. 4. https://doi.org/10.1128/mSystems.01138-21

Abstract

Viruses are ubiquitous on Earth and are keystone components of environments, ecosystems, and human health. Yet, viruses remain poorly studied because most cannot be isolated in a laboratory. In the field of biogeochemistry, which aims to understand the interactions between biology, geology, and chemistry, there is progress to be made in understanding the different roles played by viruses in nutrient cycling, food webs, and elemental transformations. In this commentary, we outline current microbial ecology frameworks for understanding biogeochemical cycling in aquatic ecosystems. Next, we review some existing experimental and computational techniques that are enabling us to study the role of viruses in biogeochemical cycling, using examples from aquatic environments. Finally, we provide a conceptual model that balances limitations of computational tools when combined with biogeochemistry and ecological data. We envision meeting the grand challenge of understanding how viruses impact biogeochemical cycling by using a multifaceted approach to viral ecology.

The importance of viruses in aquatic biogeochemistry

Microbial communities are central to biogeochemical cycling, as observed in marine[139], soil[140], and freshwater[141] environments. Over the past decades, technological advances have led to the increase of genomic sequencing, resulting in discoveries about the roles of microbes, particularly bacteria and archaea. However, few studies in aquatic microbial ecology transcend the domains of life to the realm of viruses. This lack of understanding of viruses prevents their inclusion in next-generation models that are being used to inform long-term predictions of metabolism, ecosystems, and biogeochemistry.

Most studies either focus on bacteria, archaea, or viruses individually. When combined, studies can explain how sudden shifts in biogeochemical processes in otherwise stable communities are driven by viruses[142]. Microorganisms form complex communities that interact with each other through predation mechanisms such as cell lysis, grazing, and competition for resources (Figure 4.1). Therefore, studying how all these drivers interact with each other may provide a mechanistic understanding that goes beyond descriptive ecology.

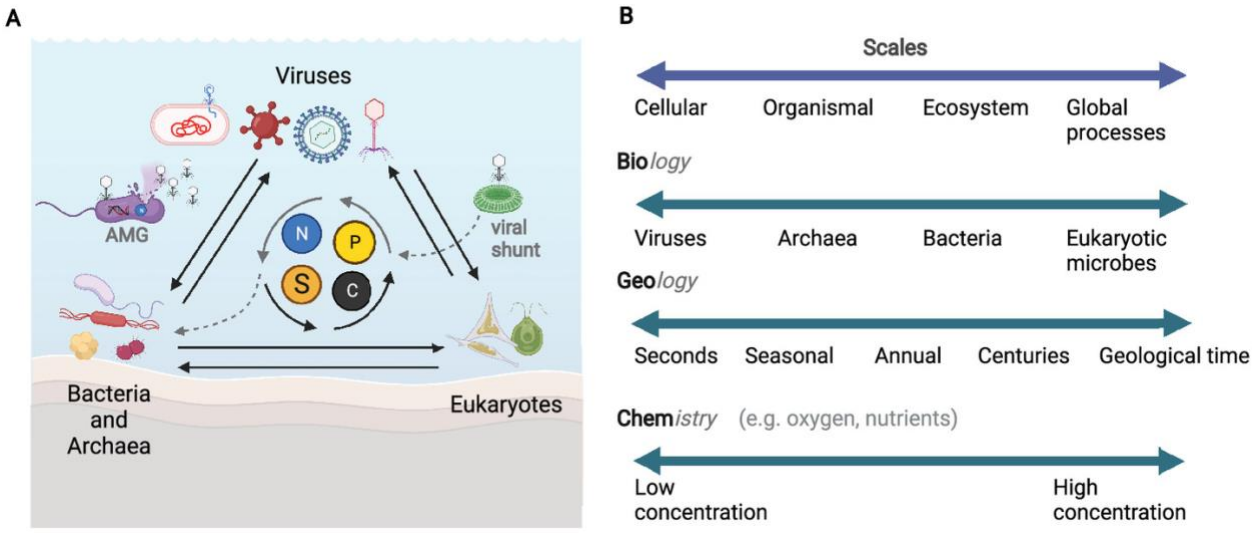

Figure 4.1 Interactions between microbial members and scales of interactions. A. Complex microbial communities made up of viruses, bacteria, archaea, and eukaryotes interact with each other and their environment through mechanisms such as predation and competition for resources. Two specific examples are shown including an example of an auxiliary metabolic gene (AMG)-containing phage that acts on the metabolism of a bacterium to produce more viral progeny (left) and an example of a virus infecting a phytoplankton to influence C, N, P, and S availability, via the viral shunt (right). **B.** Different levels of organization contribute to a holistic understanding of ecology and are associated with challenges of studying viruses. Each of the biology, geology, and chemistry components can be studied across a range of scales, from cellular to global processes.

Viral ecology studies have demonstrated that viral roles in ecosystems cannot be ignored. For example, lytic viruses can target microbes, release carbon that fuels the microbial food web (the viral shunt)[143], and have direct effects on the microbial community composition[144]. Additionally, viruses encoding auxiliary metabolic genes (AMGs) can manipulate their hosts and impact microbial metabolism and processes such as carbon, nitrogen, sulfur, and iron cycling[145]. These biogeochemical pathways are often tightly associated with environmental conditions such as oxyclines or chemoclines in aquatic ecosystems (Figure 4.1). Most viral genomic studies that specifically address biogeochemical pathways have assessed marine environments. Other aquatic environments, including inland lakes, coastal regions, streams, and rivers, also have dynamic spatiotemporal patterns that are related to microbial (bacterial, archaeal, eukaryotic) roles in biogeochemical cycling but remain understudied in the context of viruses. Evidence points to similarly prominent viral communities in ecosystems such as lakes, where AMG-containing viruses are potentially involved in biogeochemical cycling[142,146–148]. With growing evidence of viral roles in biogeochemical cycling, obtaining a more holistic understanding of functional roles, interactions, and effects of these communities can be achieved by bridging across the bacterium-archaeon-eukaryote-virus boundaries.

Techniques to study the roles of the environmental viruses

Experimental and laboratory techniques exist and provide an initial set of tools to begin integrating different scales of biology (Figure 4.1). Some methods rely on the ability to culture viruses with their host, whereas others can be performed without a cultivated host. Enumeration of viruses by phage plaque assays show that virus counts vary within an aquatic ecosystem [149]. In a global

analysis of virus morphology in the oceans, researchers used microscopy to observe that nontailed viruses dominated surface ocean microbial communities [149]. By incorporating ecological context, follow-up studies have showed that nontailed viruses in marine environments are a major predation mechanism on bacteria [150]. Yet, most viruses studied in culture are tailed, thereby showing the importance of both cultivation-based and cultivation-independent lines of evidence for understanding ecological relevance. Dilution to extinction, another laboratory method, involves filtering water, followed by enrichment, purification, and isolation to finally obtain a virus-host system[151]. Model host-virus systems are useful to explore targeted biogeochemical pathways and host-virus interactions since the controlled environment provides higher reproducibility. For example, carbon regeneration could be addressed by changing the abundance of viruses and measuring the host growth rate and biomass over time. Similarly, a host known to be involved in denitrification can be measurably impaired or improved upon the addition of a virus that targets it, by tracking host, viral, and chemical characteristics over time.

One step toward a more holistic understanding of biogeochemical processes in ecosystems is to move beyond studying model organisms to learn about other components of an ecosystem. Additionally, biogeochemistry relies on biology, geology, and chemistry, all of which have various techniques that can help understand the overall impact of viral ecology. Whereas there is a generalized recognition of the need to study uncultured microorganisms (archaea, bacteria, eukaryotes) to understand ecosystem processes, this concept is not as common in the field of virology. Since viruses are dependent on a host for cultivation and most microorganisms in nature cannot be cultivated, few environmentally relevant viruses have been cultured thus far.

To circumvent the limitations of culture-dependent viral ecology, the ongoing development of computational techniques that address the interpretative challenges of viral "omics" data will facilitate environmental virus analysis in complex environmental ecosystems. In the past years, the field of microbial metagenomics (mostly bacteria and archaea) has seen a shift from bulk read-based metagenome characterization toward functional understanding at the scale of metagenome-assembled genomes, and even at strain-level understanding of evolutionary processes and ecological patterns. The improved ability to leverage information from metagenomics is in part due to computational advances like high-throughput sequence processing, genome binning, improved algorithmic efficiency, and standardization of data. Such computational advances may be possible in the future for viral omics. Viral genomic tools are being written, tested, compared, and used to gain ecological insights [152–154], and information is becoming standardized [155]. In time, these tools will facilitate a better understanding of viruses and their complex biogeochemical interactions.

Transcending laboratory-only and genomics-only boundaries can lead to novel methods for studying viral ecology that take advantage of both strengths. Single-cell viral tagging and sequencing, analogous in some ways to single-cell genome sequencing of bacteria, rely on tagging viruses with fluorescent dyes and then cell sorting and sequencing to identify actual host-virus interactions without the need for culturing. Single-cell viral tagging and sequencing were developed for the human gut[156]. Another technique, epicPCR, consists of linking phylogenetic genes to functional genes, and then uses sequencing to obtain high-throughput ecologically relevant information about cells, such as their roles in sulfate reduction [157]. This technique is powerful in identifying virus-host interactions, especially for when viruses are not represented in

databases. EpicPCR has been adapted to study virus-host interactions without cultivation in estuarine environments, and differences have been observed in viral lifestyles and strategies between viruses with narrow and broad host ranges in the environment[158]. Overall, these examples of technological advances which combine the strengths of laboratory and computation (sequencing and its interpretation) could be applied to viral biogeochemical cycling in aquatic environments. For example, we imagine collecting a time series of microbial samples in an environment that is subject to environmental change (either biological or abiotic) using any of the aforementioned methods to sequence and study virus-host interactions. This would typically be followed by genomic investigation of viruses and hosts and interpretation of their involvement in carbon, nitrogen, or sulfur biogeochemical cycling over time, while being certain that the physical interaction between the viruses and their hosts exists. All these techniques highlight the future of viral ecology and the potential for their application across aquatic ecosystems.

Looking forward: Combining multifaceted approaches is important to obtain a holistic understanding of ecosystem ecology

The amount of genomic data generated has exponentially increased in recent years, and their interpretation benefits from a thorough understanding of the historical and ecological context and of future challenges that the ecosystem may encounter (Figure 4.2). We believe that the ability to interpret viral ecology data, particularly omics based, will be facilitated by collecting metadata and contextualizing the study system. For example, one could study the impact of carbon on bacterial growth at various resolutions ranging from simple studies focused on positive feedback at an organismal or community level (Figure 4.2) to increasing complexity of interactions (Figure 4.2). Moving toward more integrative studies, the incorporation of multiple species, multiple realms of

life, and comprehensive metadata about biogeochemistry and the environment will allow us to determine complex positive and negative feedback in the system (Figure 4.2). Specifically in the case of viral ecology, we suggest that standard virus sampling methods be coupled with detailed metadata collection of biogeochemistry and microbial communities (bacteria, archaea, and eukaryotes), which could greatly increase the ability to interpret and synthesize results.

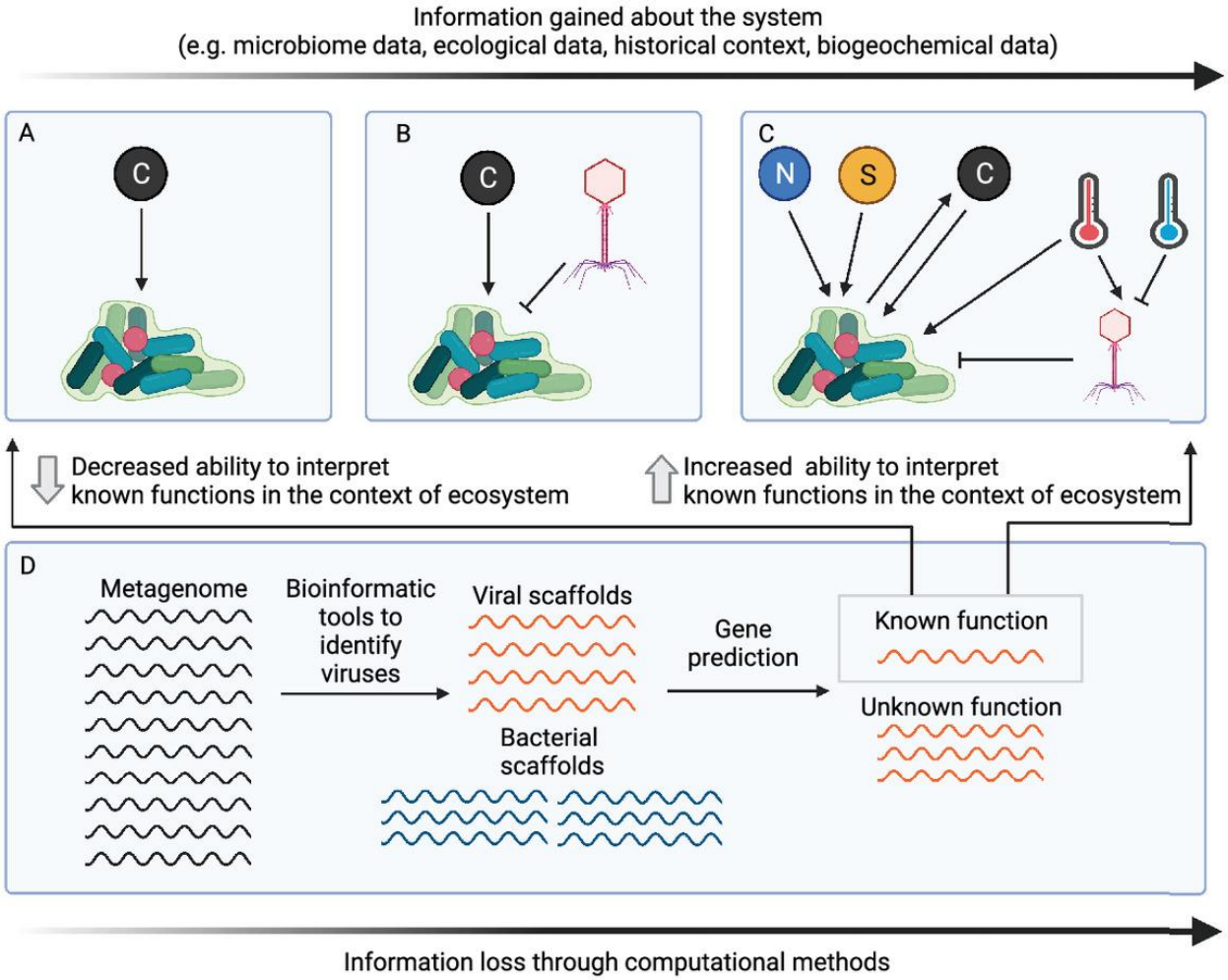

Figure 4.2 Conceptual framework for maximizing information about viral ecology and biogeochemistry in nature. Along the axis at the top of the figure are ways to gain more information about a system. **A.** Example showing the positive feedback of carbon on bacterial growth. **B.** Addition of viruses increases complexity over that shown in panel A. **C.** Further adding

detailed biogeochemical and environmental metadata such as carbon, nitrogen, sulfur, and temperature can relate complex environmental conditions to ecology but increases complexity over the complexity shown in panels A and B. Examples of positive (arrow tip) and negative (inhibitor tip) interaction are shown. **D.** Loss of information across various steps of computational analyses in viral ecology. The loss of information from computational analysis can be balanced by information gained from biogeochemical and environmental metadata.

Figure 4.2D demonstrates how computational techniques and their results, while offering a glimpse into viral ecology, remain challenging to interpret. In the simplified example, a metagenome generated from a given sample is used as a starting point to computationally identify viruses. Along each step of the pipeline, context is lost because a relatively low percentage of viruses are identified, of which most viral genomes are partial, and even fewer of the identified viruses have an identified ecological function or role. The analysis of viral genomics can be challenging on its own, especially where viral bioinformatic methods remain in constant development and have their own shortcomings. Given the same genomic data set and outcome, the ability to interpret ecological functions is significantly increased with the availability of comprehensive metadata and biogeochemical data (Figure 4.2C and D) compared to without this data (Figure 4.2A).

Finally, we envision that full integration of viral ecology into measurable and predictable outcomes would involve its integration into biogeochemical and ecosystem models. Substantial efforts have been made to integrate metagenomic and metatranscriptomic data of microorganisms (bacteria and archaea) in predicting biogeochemical processes such as carbon, nitrogen, and sulfur cycling

across redox gradients[159,160]. Realistically, it has taken over a decade of work for the field of (bacterial and archaeal) metagenomics to move on from descriptive studies of biodiversity toward mechanistic and predictive models that integrate multiple lines of experimental and genomic evidence. Even so, these integrative studies are not the norm. While challenges and opportunities in viral ecology will involve overcoming resource limitations and cross-disciplinary learning curves, we envision gaining the ability to closely couple viral ecology and biogeochemistry through these multifaceted efforts

Chapter 5 Viral impacts on microbial activity and biogeochemical cycling in a seasonally anoxic freshwater lake

Patricia Q. Tran[1,2], Samantha C. Bachand[1], Benjamin Peterson[3], Shaomei He[1], Katherine McMahon[1,4], Karthik Anantharaman[1]

1. Department of Bacteriology, University of Wisconsin-Madison
2. Freshwater and Marine Sciences Doctoral Program, University of Wisconsin-Madison
3. Department of Environmental Toxicology, University of California – Davis
4. Department of Civil & Environmental Engineering, University of Wisconsin-Madison

This chapter has been posted as a pre-print, and will be submitted to a journal for consideration.

Patricia Q Tran, Samantha C Bachand, Benjamin Peterson, Shaomei He, Katherine McMahon, Karthik Anantharaman. "Viral impacts on microbial activity and biogeochemical cycling in a seasonally anoxic freshwater lake". bioRxiv 2023.04.19.537559; doi: https://doi.org/10.1101/2023.04.19.537559

Supplementary material for this chapter is here:

https://www.biorxiv.org/content/10.1101/2023.04.19.537559v1.supplementary-material

Abstract

Microbial biogeochemical cycling relies on alternative electron acceptors when oxygen is unavailable, yet the role of viruses (bacteriophages) in these processes is understudied. We investigated how seasonal anoxia impacts viral and microbial biogeochemical cycling, by using paired total metagenomes, viromes, and metatranscriptomes, that were collected weekly. Stratification and anoxia drove microbial community composition, but dataset origin impacted the interpretation of viral community structure, activity, and function. Importantly, taxa abundance did not correlate with activity for both microbes and viruses. We identified virus-host linkages for 116 phages across 55 distinct hosts, many of which expressed genes for aerobic methane oxidation, nitrogen fixation, denitrification, and sulfate reduction. Overall, this work demonstrates the breadth and dynamics of virus-host interactions in mediating biogeochemistry. Additionally, we propose that viral community detection, functional potential, and activity are sensitive to pre-sequencing decisions, which must be kept in mind when interpreting genomic data in a biologically meaningful way.

Introduction

Anoxic zones are globally distributed and span oceans, coastal zones, and lakes. Lakes are expected to become more anoxic in coming decades, as warming water surface temperatures will increase temperature/density differences between the top and bottom layers of lakes, making it harder for water to mix throughout[4,161,162]. The increase in the duration and extent of anoxic zones is expected to transform biogeochemical cycling. Microbes play critical roles in carbon, nitrogen, sulfur, and phosphorus cycling under anoxia[102]. For example, under anoxic conditions, microbes transform inorganic mercury to toxic methylmercury[163,164], reduce sulfate to toxic hydrogen

sulfide, reduce nitrate by denitrification to form the potent greenhouse gas, nitrous oxide, and conduct methanogenesis to form the greenhouse gas, methane.

In seasonally anoxic lakes, the upper mixed water column is a habitat for fish, photosynthetic organisms, and aerobic heterotrophic microbes. In contrast, anoxic habitats are home to microbes with more diverse metabolic strategies such as chemolithotrophy, which is the use of inorganic compounds to generate energy and synthesize organic carbon. In addition, in the absence (or under the lower pressure) of top-down control of grazers and eukaryotic bacterivores on the bacterial community in anoxic hypolimnion, viral influences in controlling microbial communities can be heightened[165]. When considering the whole microbial community (viruses, bacteria, archaea, and eukaryotes), culture-independent metagenomics has contributed to the discovery and inference of virus-host relationships. Metagenomics can hint to which processes might be happening in the environment, with the caveat that genes must be transcribed into RNA, modified based on the cellular environment, and form proteins. For example, cyanophages in marine habitats have been shown to control photosynthesis[166]. The use of viral genomics also resulted in insights about the viral communities of freshwater lakes[72,146,167–169]. These approaches can be supplemented by (meta)transcriptomics (RNA-seq) and cultivation-based studies to test these predicted viral functions in the environment.

Besides the need for further lines of evidence to support metagenomic data (e.g., transcriptomics, field-based measurements, or physiology studies), is the need for benchmarking viral omics methods. Two major methodologies are commonly applied to identify viral genomes depending on where/how the viral fraction is obtained: 1) bioinformatically mining “total” metagenomes

(microbial cells plus viruses associated with cells) to extract viral genomes (presumably recovering viruses attached to or inside host cells, also referred to as cellular metagenomes in this study), and 2) physically collecting and sequencing the cell-free fraction (presumably largely consisting of free viral-like particles, referred to as viromes in this study). Each method has its own theoretical biases, but the extent to which these biases can affect the interpretation of viral-host interactions in freshwater lakes is unclear because previously published freshwater studies mainly rely on bioinformatically extracting viral genomes from the total metagenomic fraction. Computationally extracting viruses from total metagenomes is a common way to "discover" viruses and make ecological interpretations. However, it is unknown how this methodological choice influences ecological interpretation specifically in freshwater environments. For example, in agricultural soils, viromes outperform total metagenomes in assessing the viral community, but this conclusion may or may not be applied to freshwaters, as environmental factors such as moisture, pH, availability of organic matter or recalcitrant material may affect DNA extraction may be specific to each environment[170]. Given that total metagenomes are often used to identify viral metagenomic scaffolds and to annotate functions (e.g., auxiliary metabolic genes (AMGs)), it is important to understand how methodology choice affects ecological interpretations (e.g., AMGs involved in photosynthesis and biogeochemical cycling of nitrogen, sulfur, and carbon).

Lake Mendota (Madison, Wisconsin, USA) is a eutrophic freshwater lake that has a long-term continuous monitoring program focused on microbial communities in the upper mixed layer[11,171] and some studies of the microbial community in the anoxic hypolimnion (deep waters)[164]. Prior work shows that the microbial community in hypolimnion is different compared to the epilimnion, and has distinct biogeochemical functions[164,172]. Here, we seek to understand how environmental

heterogeneity (defined as spatiotemporal patterns of anoxia, and pre-post fall mixing) impacted virus-host interactions, and how these interactions in turn impacted their activity in carbon, sulfur, and nitrogen cycling.

Results

We collected depth-discrete samples from Lake Mendota between July 16, 2020, to October 19, 2020, spanning the period of thermal stratification and fall mixing. Lake Mendota was thermally stratified from late May until October 18, 2020 (Figure 5.1A). Anoxic water, which is 0 mg/L of dissolved oxygen) was detected in July, as is historically typical for Lake Mendota[161,173]. We chose 16 water samples that captured the stratified and mixed period for sequencing (16 total metagenomes, 14 viromes, 16 metatranscriptomes) (**Supplementary Table 1)**. Samples captured a range of environmental data such as oxygen (including oxic, oxycline, anoxic layers), pH, phycocyanin, turbidity, and water temperature.

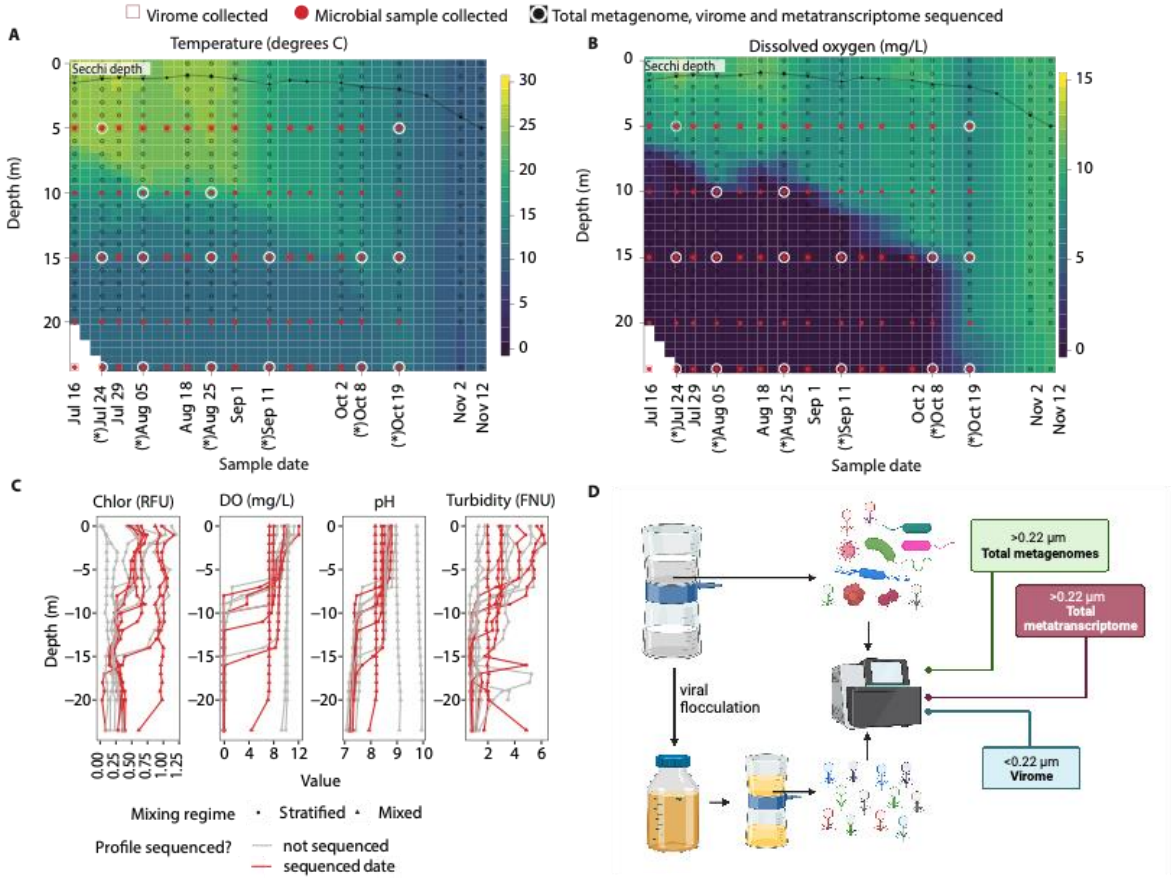

Figure 5.1 Samples and environmental data. Weekly samples were collected at 5, 10, 15, 20, and 23.5m from July 16, 2020, to October 19, 2020. Viral samples were obtained for depths 5, 15, and 23.5m each week. Samples were collected during the day. A. and B. Temperature (A) and dissolved oxygen profiles (B) in Lake Mendota. The lake was stratified until fall mixing on October 18, 2020. 16 samples representing the oxic, anoxic, and oxycline layers were selected for total metagenome, virome, and metatranscriptome sequencing. Asterix represent dates where samples were sequenced. C. Chlorophyll (Chlor), dissolved oxygen (DO), pH, and turbidity profile in 2020. D. From each sample, water was filtered through 0.22um filters for total metagenomes and metatranscriptomes and the filtrate was subject to viral precipitation with iron chloride to generate subsequent viromes.

While we did not collect sulfide data in 2020, data collected in 2021 showed that sulfide began accumulating in the water column in mid-summer and accumulated in the anoxic hypolimnion during the stratification period. Lake Mendota typically has sulfide concentrations higher than 150μM in the anoxic hypolimnion (**Supplementary Figure 1**) and substantial oxygen demand due to high epilimnion primary productivity[174]. Lake Mendota is a low-iron and high-sulfate lake (sulfate is of geological origin) and prior studies have documented vigorous sulfate-reduction[31], based on measurements in the sediment[175], and from the profile of sulfate-sulfur in the anoxic hypolimnion[164,172]. While we did not collect methane and nitrate and nitrite in 2020, prior studies have shown the accumulation of methane [in the hypolimnion?] during summer stratification[176] and nitrate/nitrite in the water column.

Microbial community composition and expression change as a function of anoxia, but abundance does not correlate to activity

We assembled metagenome-assembled genomes (MAGs) from the 16 total metagenomes and identified viral genomes in both the total metagenome and virome fractions, and used bioinformatic analyses to link phages to their potential bacterial hosts. Finally, we used metatranscriptomes to assess the activity of these phage-host pairs. We obtained a final set of 431 MAGs (429 bacteria, 2 archaea) of medium to high quality for downstream analyses. Among them, 349 MAGs were assigned to 176 unique genera and 82 MAGs could not be assigned to any genera (**Supplementary Table 2, Supplementary Figure 2**). Of those 82 MAGs without a genus

assignment, most of them were associated with the phyla Verrucomicrobiota, Bacteroidota, and Patescibacteria. 133 MAGs shared ≥ 97% ANI (average nucleotide identity) to MAGs previously published from Lake Mendota[102,164]. Notably, previous MAGs were generated from different water samples, namely water column integrated from the surface down to 12m[102], and two depth-discrete anoxic profiles collected in 2017[164] (**Supplementary Table 3**). Besides common freshwater phyla, we recovered 18 Patescibacteria, 81 Gammaproteobacteria, 42 Kirimatiellaeota, and 20 Verrucomicrobiota MAGs.

The hypolimnion harbored more diverse taxa over time during the stratified period, as indicated by the increase in the Shannon diversity indices (Figure 5.2**A**). Among site categories, beta-diversity was 0.16 in the oxycline & stratified sample, 0.18 in the anoxic & stratified sample, and 0.20 in the oxic & mixed sample.

To determine which organisms were active, we compared MAG abundances to their transcriptional activity. We used the MAG relative abundance, expression, and the ratio of expression to abundance to compare who is present, active, and overexpressed respectively (Figure 5.2**B**). From metagenome data, we calculated the relative abundance of MAGs within each sample based on their metagenome read coverage, and the top 10 most abundant MAGs in each metagenome are shown in the left panel of Figure 5.2**B**. From the metatranscriptome data, after mapping, the read counts were first normalized by the internal standard read count (see Methods) to obtain an $RPKM_{normalized}$ value for each MAG, allowing the comparison of expression/activities among different samples. The top 10 most active MAGs in each metatranscriptome are shown in the middle panel of Figure 5.2**B**. The $RPKM_{normalized}$ value of each MAG was then divided by its

relative abundance within the corresponding metagenome, to obtain a ratio of RPKM to abundance to see if a MAG is overexpressed compared to other expressed MAGs. The top 10 overexpressed MAGs are shown in the right panel of Figure 5.2**B**).

Overall, active taxa were not correlated with abundant taxa. For example, differences in microbial community composition existed among zones of different oxygen availabilities, and the community composition homogenized upon fall mixing as indicated by the similar community composition throughout the water column post-mixing (**Supplementary Figure 3**). In the surface oxygenated samples, the microbial community was dominated by Actinobacteriota, Cyanobacteria, and Planctomycetota, but Cyanobacteria and Actinobacteriota were the most active. At the oxycline, *Burkholderiales* (Gammaproteobacteria) was both abundant and active, although *Methylococcales* (also a Gammaproteobacteria) were among the most active and most overexpressed. In the anoxic hypolimnion, the most abundant taxa were initially *Burkholderiales*, but over time, Verrucomicrobiota became consistently highly ranked. However, Desulfobacteriota was most often highest expressed in the hypolimnion. Additionally, when considering overexpression (high RPKM to abundance ratio), Campylobacterota (f__*Sulfurimonadaceae*, g__*Sulfuricurvum*) and *Thiomicrospirales* (Gammaproteobacteria) were notable in the anoxic hypolimnion. Both these groups of organisms are known to participate in sulfur cycling and are commonly found in oxygen-minimum zones in marine systems[177–179]. *Sulfuricurvum* is a facultative anaerobic chemolithoautotrophic sulfur oxidizer that has been known to use nitrate as an electron acceptor[180]. *Thiomicrospirales* have been shown to increase with depth in oxygen minimum zones[179]. After fall mixing, the microbial community "homogenized" across layers, with Alphaproteobacteria becoming highly ranked in terms of abundance but in comparison,

Methylococcales remained a highly expressed taxa, along with some taxa from the hypolimnion (e.g. *Desulfobacteria*). Many of the top-ranked active taxa in the anoxic hypolimnion were associated with phages, as shown by the asterix in Figure 5.2**B**.

Viral community composition

The microbial community composition and expression patterns were distinct after fall mixing and differed between the epilimnion and hypolimnion, but the viral community was mostly distinguished by the origin of the dataset analyzed – in other words, whether the viruses were recovered from the total metagenomes or viromes (Figure 5.3). *Caudoviricetes* (*Myoviridae*, *Siphoviridae*, and *Podoviridae*) dominated each sample (**Supplementary Figure 4**). In the viromes, a higher number of viral genomes were of unknown taxonomy, compared to those in the total metagenomes. *Microviridae* contributed to a higher percentage of the total metagenome viral community but were seldom observed in the viromes.

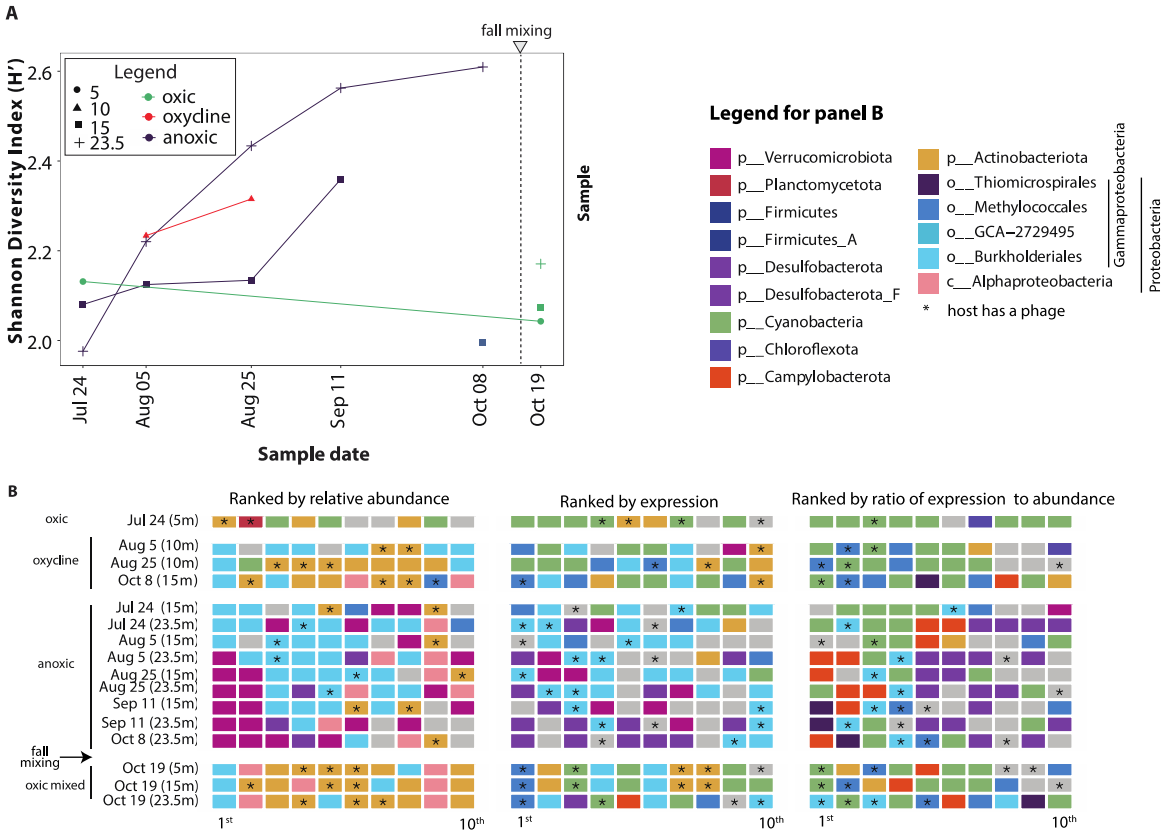

Figure 5.2 Diversity and the top 10 MAGs ranked by abundance, expression, and overexpression. A. The Shannon Diversity indices were calculated in each of the 16 total metagenomes. B. Top 10 MAGs were ranked by the abundance (read coverage within a metagenome), expression (RPKM divided by internal standard count), and overexpression (the ratio of expression to abundance), respectively, and were colored by their taxonomy (with Grey indicating taxa not listed in the legend). MAGs with at least one associated phage (Figure 5.7) are indicated with an asterisk.

Comparison of viruses in "total metagenomes" vs. "viromes"

Total metagenomes are the fraction that is retained on 0.22 μm filters, and viromes are the fraction that passed through 0.22 μm filters and were chemically precipitated. We investigated the viral

community in both the virome and total metagenome fractions to identify differences or similarities in viral community composition, function, and activity.

For any pair of samples, significantly more viral scaffolds were identified in the viromes than in the total metagenomes (paired t-test, t = -11.236, df = 13, p-value = 4.581e-08) (Figure 5.4**A**). For example, about 60-80% of scaffolds were viral in the viromes, whereas only ~20% of scaffolds were viral in the total metagenomes. Viromes yielded more phage genomes in most sample pairs, except in 2 sample pairs (August 25th 23.5m and October 10th 23.5m) (Figure 5.3**B**). Additionally, there tended to be more lytic phages than lysogenic phages in both total metagenomes and viromes (Figure 5.4**C**).

Additionally, we found that the number of complete and high-quality viral genomes was higher in viromes than in the total metagenomes (1,466 vs. 780). The number of complete genomes was higher in total metagenomes than in viromes (176 complete genomes vs. 132). We found 6 mega-phages (>200kpb) and 469 small phages (<15kbp) (all are complete or high-quality genomes) (**Supplementary Table 4**). With 2,246 complete and high-quality phage genomes considered in this analysis, 84% of them were never expressed over the course of the metatranscriptomics time series. 354 (or 16%) of the phages were expressed at least once. On any given day, between 1.34 to 4.41% of the viral genomes showed positive RNA expression (Figure 5.4**D**). For any given metatranscriptome, a variety of proportions of sample origin existed, but the oxycline (e.g. Oct 08 at 15m) and post-mixing samples tended to have over 50% of active phage originating from the viromes (Figure 5.4**D**).

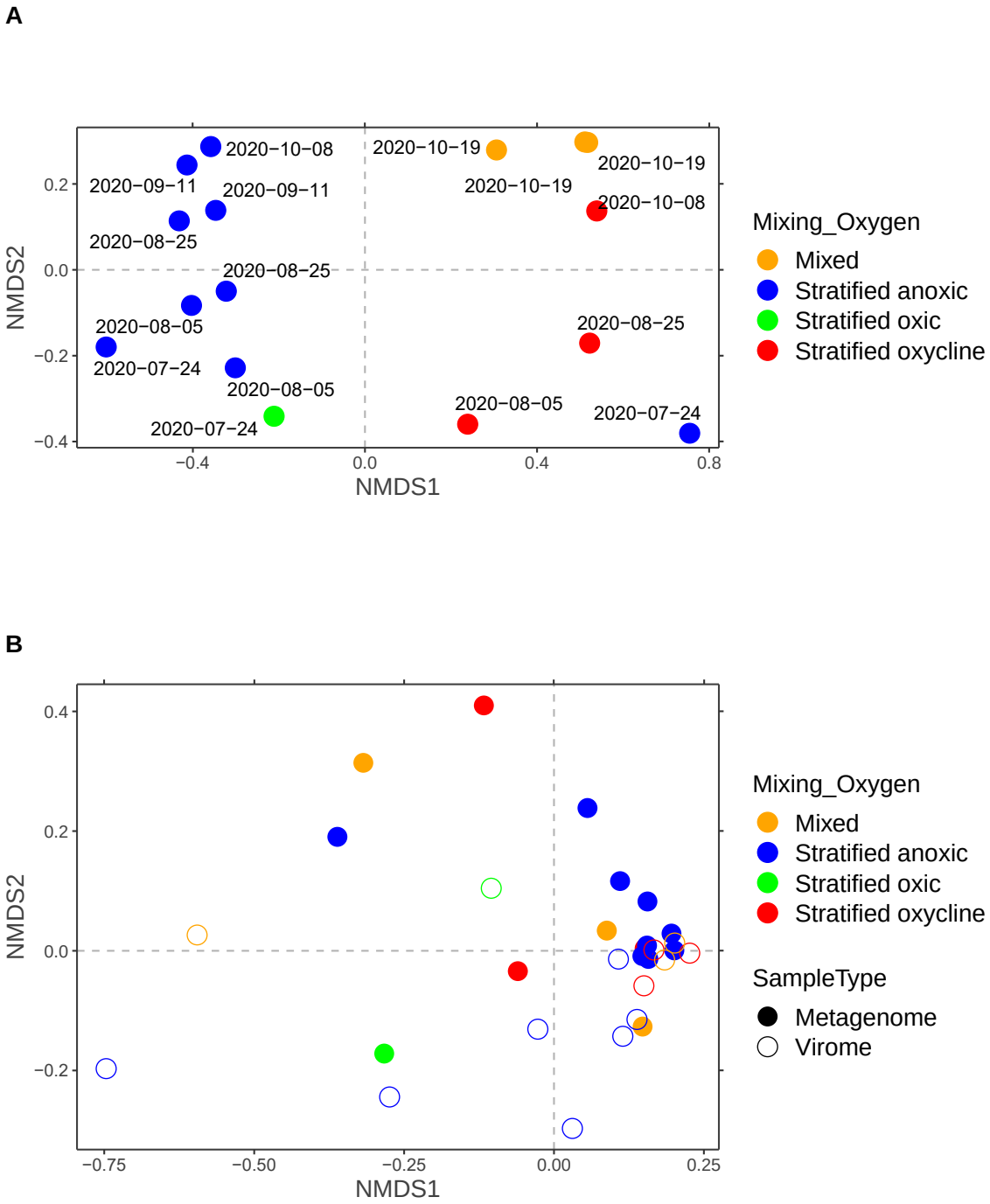

Figure 5.3 Non-metric dimension scale plot of (A) the prokaryotic community (B) and the viral community. Environmental type (season and stratification) was a driver of the prokaryotic community (clustered by color), but not the viral community. Rather, the viral community was

broadly clustered by sample type, that is whether they were originating from a total metagenome or viromes.

Phage activity and impact on nutrient cycles

Auxiliary metabolic genes (AMGs) are genes found on phage genomes, that are not involved in critical functions like phage replication. While they have different definitions, they are generally considered to augment their hosts' metabolism, by enhancing certain metabolic pathways that result in more viral progeny being released upon host-cell lysis[145]. 91 different AMG families were identified (distinct KO identifiers) and comprised 1,287 proteins distributed on 494 phage genomes. 135 phages were from identified metagenomes (18 lysogenic, 117 lytic) and 359 were from viromes (28 lysogenic and 331 lytic). We calculated the abundance and RNA expression of those AMG-encoding phages (Figure 5.5). 59 AMGs were found on phages that were inactive (no RPKM expression during any sampling point), and 27 AMGs were on phages that were expressed at least once. Among those that were expressed, 13 AMGs were on phages recovered from the total metagenome fractions, and 6 were on phages recovered only from the virome fractions, and 8 from both fractions. (**Supplementary Table 5**). Notably, photosynthesis AMGs (*psbAD*) were found on phages from both fractions but were transcriptionally inactive. They were identified on phage genomes that could be attributed to 8 closely related species of *Synechococcus* reference hosts. The *pmoC* AMG associated with aerobic methane oxidation was found on an inactivate phage, for which the host could not be identified in our study, but is likely a *Methylococcales* based on a previous study using Lake Mendota data[148]. Organosulfur metabolism AMG (*cysH*) were found and active in both fractions. Finally, a sulfur oxidation AMG (*soxB*) was found in viral genomes originating from the viromes but transcriptionally inactive.

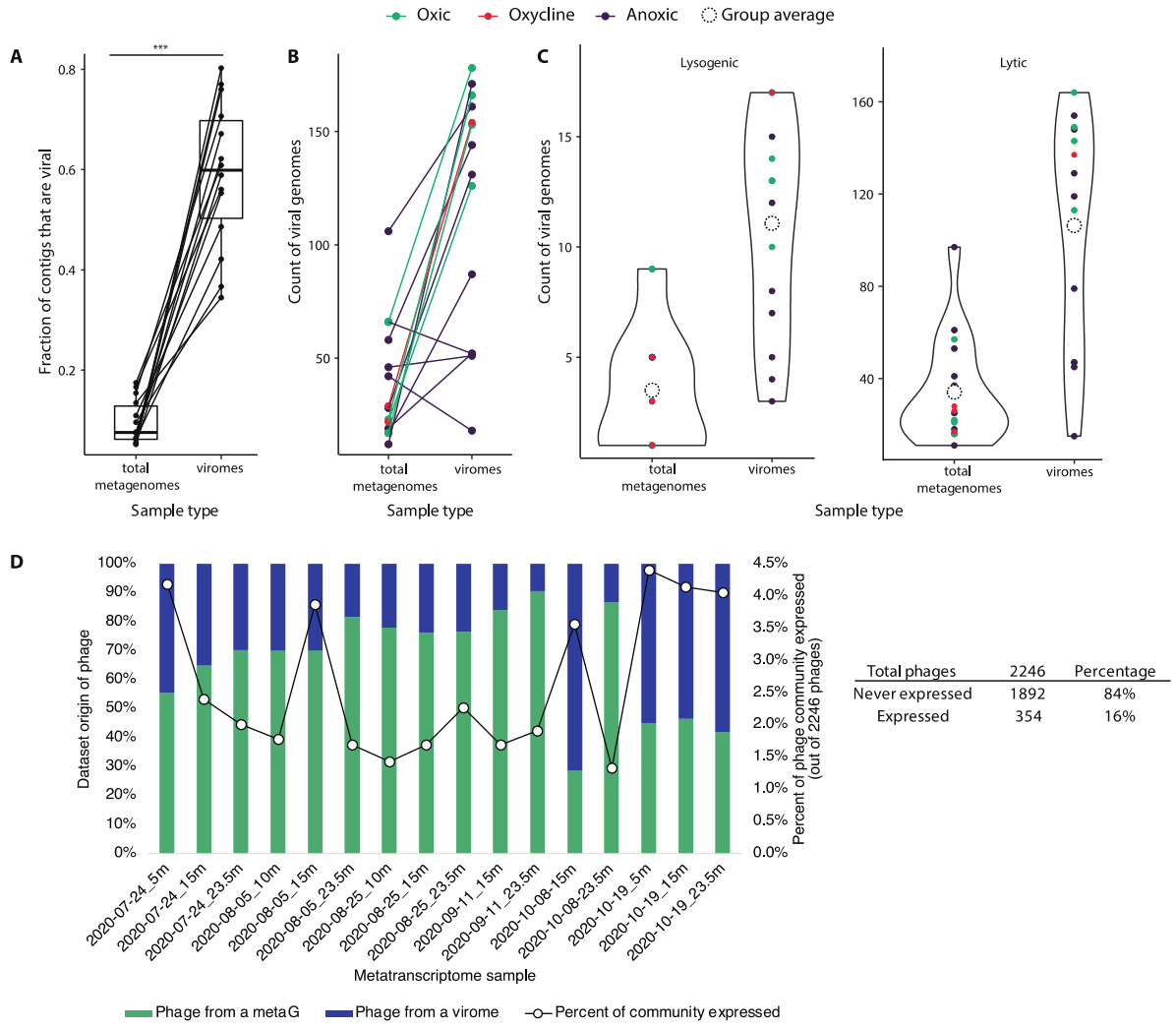

Figure 5.4 Comparison of viruses based on dataset origin. A. A significantly higher percentage (paired t-test, t = -11.236, df = 13, p-value = 4.581e-08) of assembled scaffolds are identified as "viral" from the virome fraction for any given physical sample pair. B. The number the putative complete or high-quality phage genomes is higher in the virome than in the total metagenome, except in 2 samples. C. The number of lytic and lysogenic phage genomes (complete or high-quality) found in viromes is higher than when computationally extracted from total metagenomes, and the number of recovered lytic phage genomes is higher than the number of lysogenic genomes in both total metagenomes and viromes (note the difference in y-axis scale). D. Origin and activity

of phages. Of the 2246 phages with complete and high-quality genomes, 16% of them were expressed at least once in the metatranscriptomes. Per sample point, at most ~4.5% of the 2246 phages were active.

Phage-host interaction diversity and specificity

We predicted phage-host interactions by identifying the CRISPR sequences in host genomes, keeping in mind that approximately 22% of bacteria in a population maintain the signature of previous phage infection[181]. We identified 113 CRISPR cassette arrays in 68 MAGs of the 431 prokaryote MAGs, and 11 of those CRISPR arrays (in 9 MAGs) had a 100% match to phages (0 mismatches) with a Lake Mendota phage dataset (genomes of all qualities included), which assigned them to 22 phages. We inferred that an array that matched a phage corresponded to the susceptibility of that host to phage infection, at least at some point in the recent past. The inferred infected groups include Gammaproteobacteria, Bacteroidota, Cyanobacteria (*Dolichospermum flosaquae* and *Microcystis*), Verrucomicrobiota (Verrucomicrobiota and *Kirimatiellae*), Chloroflexota, and UBA2361. These phage were taxonomically assigned to *Caudoviricetes* (*Myoviridae*, *Siphoviridae*, and *Podoviridae*), although several could not be assigned. Two of the Chloroflexota phages encoded the *cysK* AMG, involved in cysteine metabolism. Most identified phages were obtained from virome samples as opposed to total metagenome samples. Each identified MAG appeared to be susceptible to a range of phages that are not closely related, except *Dolicospermum flosaquaea* which was infected by 5 taxonomically unknown phages sharing high ANI among themselves. In contrast, we found that phages displayed a high host-specificity (infected only 1 MAG) (**Supplementary Table 6**, **Supplementary Table 7**).

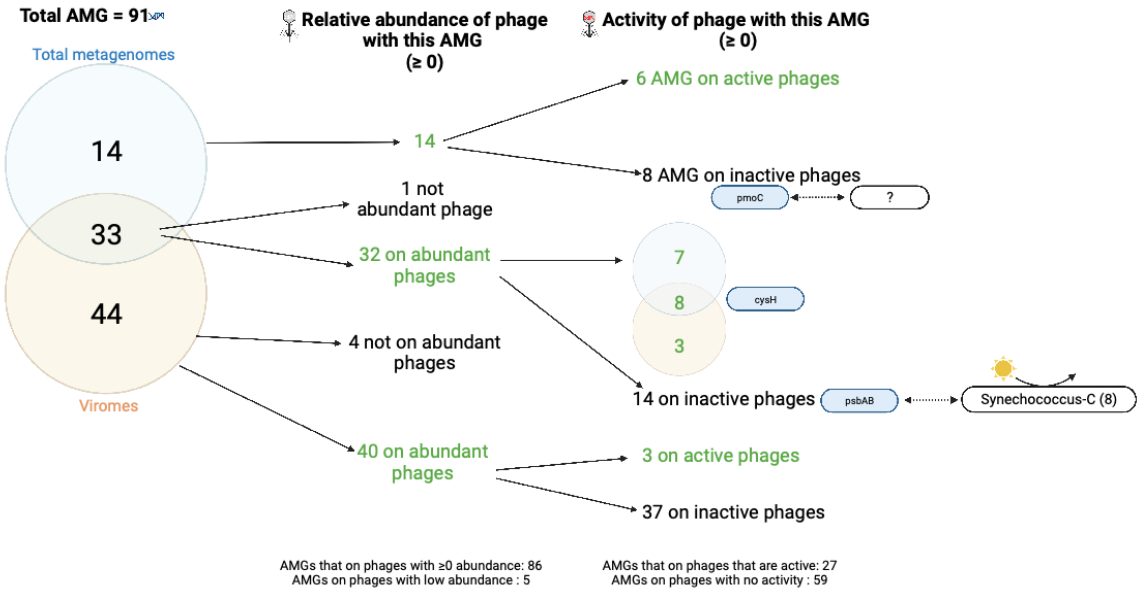

Figure 5.5 Phage and AMG expression, and examples of phage-host relationships. 91 auxiliary metabolic genes (AMGs) found on high-quality and complete phage genomes and their distribution between "expressed" and "not expressed", and between total metagenomes and viromes. Of those, 59 AMGs were on phages that were not expressed, including ecologically relevant AMGs pmoC (aerobic methane oxidation), and psbA and psbD (photosynthesis). The psbAD AMGs were found on 2 phages whose hosts were 8 closely related species of Cyanobacteria, specifically Synechococcus-C. The pmoC AMG was on a phage that was not predicted to be linked to a specific prokaryote MAG in our dataset or to previously published bacterial and archaeal reference genomes.

Additional ways to predict phage-host interactions rely on the host-virus similarity in sequences (e.g. tRNAs, prophages) and sequence composition (e.g. kmer) and viral marker genes indicative

of hosts[181]. Machine learning approaches relying on features learned from reference databases of viruses with their known hosts was developed to assign phages to potential hosts, even without having host genomes. Therefore, to further identify phage-host relationships, we used iPHoP[182], which integrated the above prediction methods to assign phages to putative hosts that may or may not have a MAG available. Among the 431 MAGs, 50 had a putative phage associated with them, 55 when including the additional manual CRISPR cassette method. When looking at the expression of those 50 MAGs across the 16 metatranscriptomes, 15 of the MAGs were never expressed. Among those expression categories, generally fewer MAGs were found to be associated with a phage across all categories. While a greater diversity of organisms was active in the hypolimnion, only those belonging to Gammaproteobacteria, Bacteroidota, and Actinobacteriota were directly impacted by phages.

Phages impact biogeochemical processes driven by active microorganisms

To estimate the importance of these phage-host interactions on biogeochemical cycling, we predicted metabolism of the phage-impacted bacteria and mapped RNA-seq data onto these to determine activity (**Supplementary Table 8**). Among the 55 phage-associated MAGs, many hosts had key genes for aerobic methane oxidation, denitrification, nitrogen fixation, and sulfur cycling, many of which were also expressed (Figure 5.6). While we did recover a *Methylococcales* MAG that encoded all *pmoABC* genes for methane oxidation, its phage did not encode any homologous AMGs. Phage-associated *Burkholderiales*, Cyanobacteria, Fibrobacteria (inactive), and Verrucomicrobiota encoded genes for nitrogen fixation. *Burkholderiales*, especially *Rhodoferax*

organisms had the full set of genes for complete denitrification. Additionally, several *Burkholderiales* organisms also possessed genes for sulfur oxidation and expressed them.

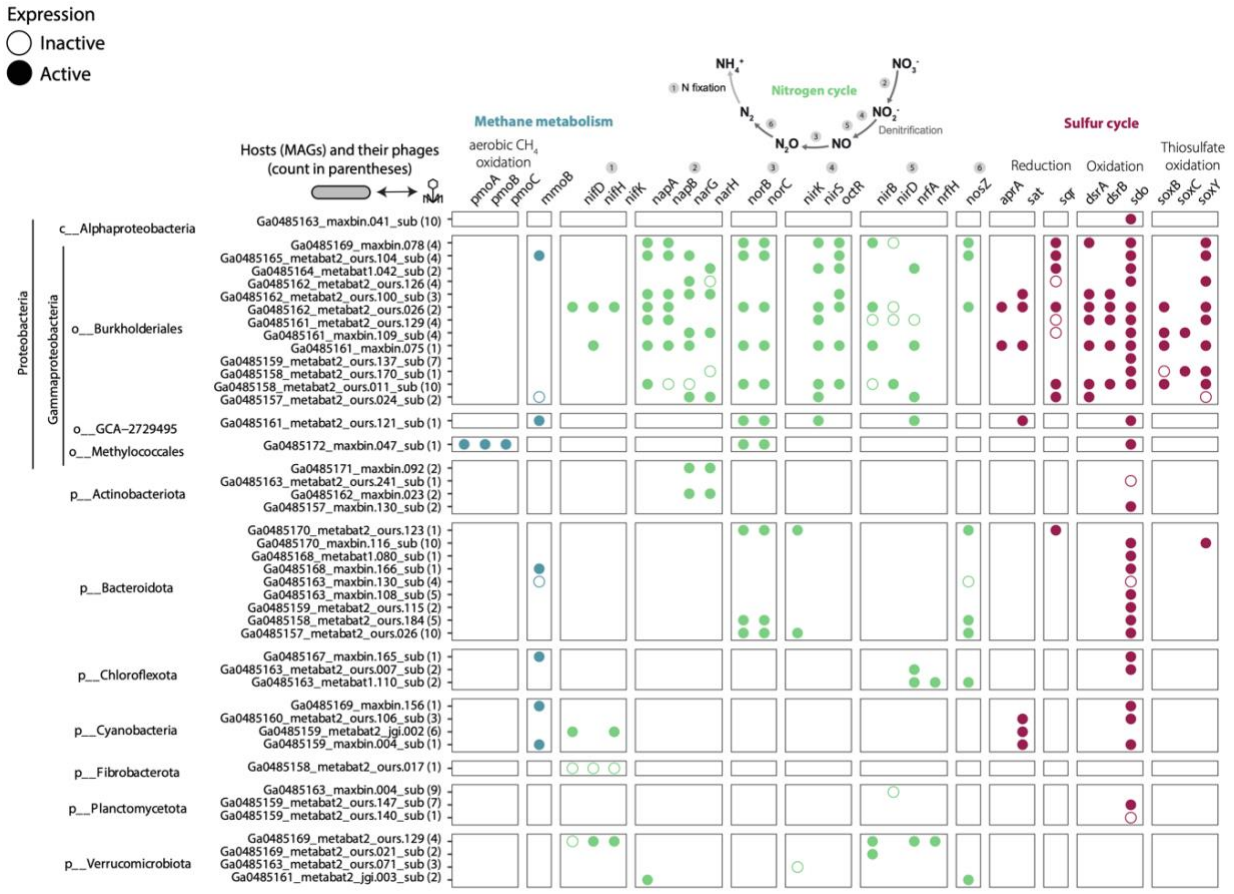

Figure 5.6 Methane, nitrogen, and sulfur metabolism and gene expression of 55 MAGs with at least one identified phage.44 out of 55 MAGs had N, S or CH_4 genes to be plotted. Phage susceptibility ranged from 1 to 10 phages per MAG. 1 MAG (Methylococcales) had all pmoABC genes in the aerobic methane oxidation pathway (the associated phage did not encode a pmo auxiliary metabolic gene. Burkholderiales (Gammaproteobacteria), Fibrobacteria, and Verrucomicrobiota MAGs had genes in nitrogen fixation. Phage-associated Gammaproteobacteria were involved in multiple steps of the denitrification and sulfur transformation pathways. No phage-associated host had genes in nitrification (aerobic ammonia

oxidation nor nitrite oxidation). Note that some dsrD genes were found on some MAGs, but the gene was not active. Data from Supplementary Table 8 was used for plotting.

Phage-host abundance and expression dynamics

Any given phage was found between 1 and 16 samples and expressed (active) between 0 and 9 samples. Cyanophages were among the most abundant and active phages in the dataset. (**Supplementary Figure 5**) and included hosts from *Microcystis* (genera *Microcystis* and *Dolichospermum*) and *Nostocera* (10 different genera). Among all phages that were active, only 6 were present *and* active *and* had a MAG as a host (2 Cyanobacteria, 3 Bacteroidota, and 1 Gammaproteobacteria). The 3 Bacteroidota (Ga0485170_maxbin.116_sub, Ga0485159_maxbin.015, Ga0485168_metabat1.080_sub) were not among top-ranked abundant, active nor overexpressed MAGs. The phage-host abundance and expression patterns of the 2 Cyanobacteria MAGs and their cyanophages, and of the *Thiobacillacea* (Gammaproteobacteria) and its phage are shown in Figure 5.7. The first cyanobacteria MAG, the *Dolichospermum*, was

abundant and active pre- and post-fall mixing, and phage-host dynamics were coordinated.

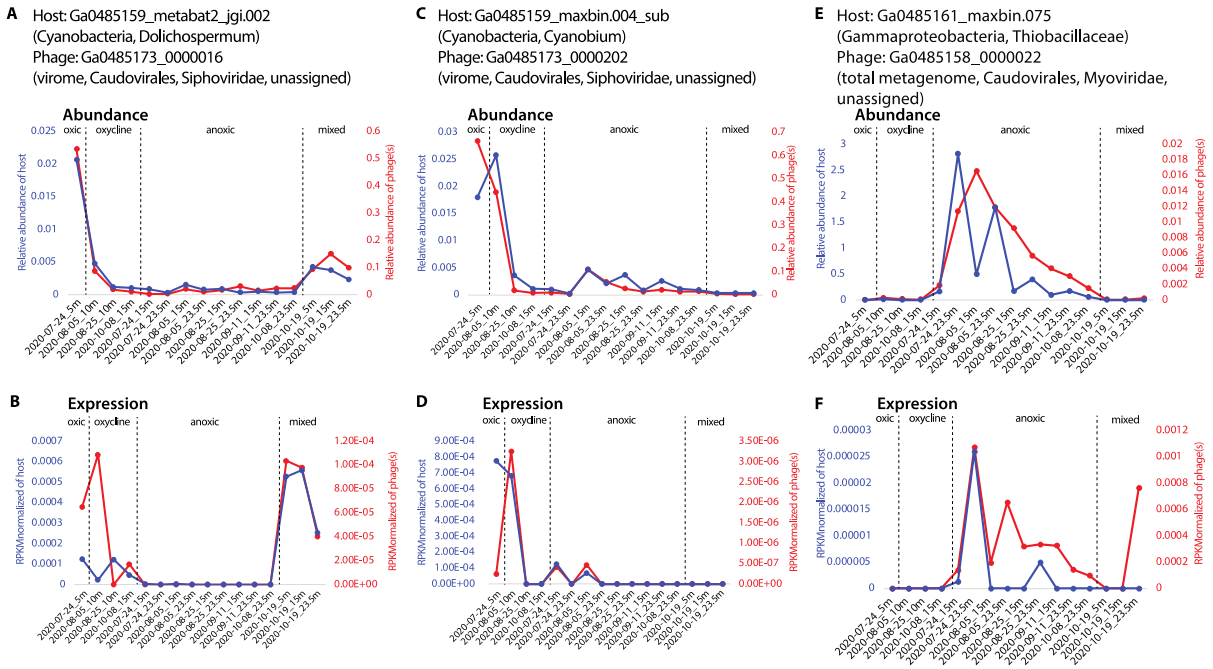

Figure 5.7 Phage-host abundance and activity patterns of Cyanobacteria and Thiobacillaceae hosts and their phages. A and B. The Dolichospermum phage-host pair's abundance (A) and activity (B) show a higher abundance during stratification in the oxic epilimnion, and post mixing. In comparison, the phage was highly active in the oxycline and post mixing. This MAG is not among the most abundant ones but is in the top 10 ranked most active in 4 out of 16 samples, which correspond to the oxic samples (July 25 5m, and the 3 samples from Oct 19). C. and D. The Cyanobium phage-host pair's abundance (C) and activity (D) show that these organisms are abundant in the upper water column and throughout the water column but were only active in the oxycline and in the upper anoxic zone (15m). This MAG is in the 7th most active taxa in the July 5m (oxic) sample. E and F. This Thiobacillaceae ranked in the top 10 most abundant taxa of 5 anoxic samples and in the top 10 most active MAG in all the 9 anoxic samples and of the oxic-anoxic 23.5m sample (Figure 2B and Supplementary Table 7).

Discussion

Here, we study how spatiotemporal impacts of stratification and anoxia influence microbial and viral communities and activity in a mid-size eutrophic lake, that is representative of temperate lakes worldwide. We investigated how biogeochemical cycles in seasonally anoxic freshwater environments were correlated with viruses and prokaryotes involved in those functions. Our analyses revealed differences in the abundance and expression levels of the prokaryotic community over weekly scales, mainly differentiated by stratification and oxygen levels, but this was not observed in the viral community. Additionally, our comparison of viromes and metagenomes revealed that viromes resulted in more high-quality and complete phage genomes and a higher number of previously undiscovered "unknown" phage taxa. Finally, we identified 9 MAGs with putative phages just based on CRISPR arrays, and a total of 55 MAGs with phages based on the integration of multiple phage-host prediction methods. These 55 MAGs represented 7 phyla. These bacterial hosts were inferred to be involved in methane oxidation, denitrification, nitrogen fixation, sulfur oxidation, and sulfate reduction, and many of these genes were indeed expressed. Next, we contextualize what is known about microbial communities of stratified anoxic lakes, freshwater phages, and the implications of this work beyond freshwater phage ecology.

Increased microbial diversity over time and space is not directly correlated to activity

Prokaryotic diversity increased with time and water column depth. This trend might reflect the emergence of diverse and heterogenous microniches corresponding to the dynamics of carbon/nutrient availabilities and redox gradients with the succession of preferred terminal electron acceptors since the stratification. As the fall mix started, physicochemical "homogenization" of lake water occurred. For example, the entire water column became oxygenated, and as oxygen was the preferred electron acceptor thermodynamically, microbial processes with alternative electron acceptors were suppressed. As a result, such homogenization could "simplify" the microbial community, indicated by the decrease in Shannon diversity index post-mixing.

Environmental stratification and redox gradients result in strong differences in microbial community structures between the oxic epilimnion and the anoxic hypolimnion, and fall mixing can "reset" the water column's microbial community[171]. Lake Mendota did not have many archaeal groups, unlike in anoxic freshwater lakes with deep hypolimnions[183,184]. As for the epilimnion, in Lake Mendota, a historical time-series of 16S rRNA gene sequences revealed that the upper 12m of the water column, mostly oxygenated but sometimes spanning the oxycline, mostly contains Cyanobacteria, Actinobacteriota, Proteobacteria, and Bacteroidota[185]. In contrast, reconstructed genomes of Bacteroidota and the PVC superphylum (Planctomycetota, Verrucomicrobiota, Chlamydiota)[164] were recovered from the hypolimnion. 16S rRNA gene-amplicon-based studies of the microbial community across depths (2016-2017) revealed that taxonomic groups were strongly influenced by overall redox status, rather than the onset of anoxia[172]. In concordance with this, the changes in the oxycline depths across the season were related to the microbial community, and even in the hypolimnion, differences were observed between the 15m and 23.5 m layers.

We noted that the abundance of taxa (based on metagenomic sequencing) did not necessarily correspond to the expression. For example, while the metagenomic data supported abundant groups commonly found in Lake Mendota and other freshwater lakes, our results highlight the importance of *Methylococcales*, Desulfobacteriota, and *Burkholderiales* as highest expreession in the lake, especially in the oxycline and anoxic hypolimnion, which constitute a major ecological space during summer stratification.

Viral impacts on biogeochemistry

Here, we specifically related phages to hosts (binned MAGs) from the same environment, in addition to comparing with viral host databases to identify putative hosts (not represented by our MAGs). The CRISPR-spacer similarity between hosts and viruses was widely used to link these pairs across environments, namely because these are natural bacterial and archaeal immune machinery against viral infections[186]. Consistent with the literature, phages displayed high host specificity, with one phage infecting one group of bacteria[187]. In other systems, ~2000 phages were associated with 90 MAGs across 25 phyla, in a community with a total of 29 phyla[188]. Here, our initial overall community is less diverse (27 phyla). We found phages infecting Cyanobacteria, Gammaproteobacteria, Alphaproteobacteriota, Actinobacteriota, Fibrobacteria, Verrucomicrobiota, Bacteroidoita, Chloroflexota, and Planctomycetota – all are common freshwater phyla. Cyanophages have been reported in freshwater lakes, and other commonly found freshwater taxa such as Actinobacteria, Alphaproteobacteria, Planctomycetes, and Verrucomicrobiota have been previously reported to be infected by phages[189,190]. We also found phages infecting Chloroflexota and Gammaproteobacteria (orders *Burkholderiales*, *Methylococcacea*, and GCA_2729495). In lakes, Chloroflexota is often associated with

hypolimnion[191] and can make up to abundance up to ~25% in lakes[192]. Gammaproteobacteria phages have also been found in the Baltic Sea[193], and *Methylococcaccea* phages have been found in Lake Baikal[146].

Another case is when we found phages, but no host could be found. Methodological reasons may explain this, as unbinned scaffolds (up to 50% of the metagenome sometimes) could contribute to incomplete host genome(s). In such cases, leveraging public databases enabled us to identify potential hosts: for example, we found photosynthesis AMGs on phages infecting *Synechococcus* species when expanding our search to the reference database. Additionally, while we reconstructed multiple Cyanobacteria genomes, we did not resolve a *Synechococcus* genome, but did recover closely-related *Cyanobium* species (9 MAGs). *Synechococcus* spp. are known to be found in freshwater lakes though, for example in oligotrophic reservoirs[194], along with their cyanophages[195]. In these cases, we can only infer the potential effect of the phage on the biogeochemistry of the lake – *Synechococcus* is among the most well-studied oxygenic photosynthetic bacteria, but this inference about the ecological role based on the host might be less obvious for other taxa.

Methane, nitrogen and sulfur cycling MAGs were found in the dataset, and were infected by phages. Additionally, *Methylococcales* (Ga0485172_maxbin.047_sub) encoding *pmoABC* genes for aerobic methane oxidation, were among the most expressed MAGs in our dataset. Denitrification in the anoxic hypolimnion results in a decrease of NO_3^- and an accumulation of intermediates. As for sulfate, about 45% of the sulfide in the sediment is derived from sulfate reduction in Lake Mendota[175], and prior studies have shown that sulfate reduction could occur in

the water column based on MAGs and biogeochemical profiles[164]. Our results support these findings, as many sulfate-reducing bacteria were active based on metatranscriptomic data, for example, the *Desulfobacula* genus within the *Desulfobacteraceae* family which are known to be sulfate-reducers.

Capturing active MAGs and active phages also depends on the timing of sampling. It has been shown that environmental factors such as salinity, temperature, UV-light[196], and the host growth rate[197] can influence phage communities and lifestyles, including the switch from lysogenic to lytic. In freshwater lakes, these environmental parameters change over time and space, sometimes following daily, weekly, seasonal, annual, or even decadal patterns.

Collecting viromes in conjunction with total metagenomes allows for recovering more diversity of phages

Compared to cell-associated (total) metagenomes, a higher number of phages were recovered from viromes, supporting similar observation in soils[170]. In freshwater lakes, sequencing depth[189] and different size fractions (based on the filter pore size)[198] have been shown to influence which viruses are detected, as such it was beneficial to pursue paired metagenome-viromes in this study to capture the entire microbial community. Here, we found more lytic phage genomes than lysogenic. This goes against our expectation that metagenomes would recover more lysogenic phages, and this might have been caused by free lytic viral particles that would otherwise pass through the filter and could accumulate on the filter as it was clogged with biomass accumulated during the filtration process.

The impact of collecting and analyzing viromes has been demonstrated in oceans[199,200] and laboratory procedures have been optimized[201]. In freshwaters, for example, viruses were extracted from total cellular metagenomes in multiple lakes [72,146,202,203]. Viromes have been generated for freshwater lakes[204], with varying sequencing methods, and paired with either metagenomes or transcriptomes using the technologies that were current at the time of those studies. Many physical separation methods of viruses have also been used in the past, including tangential flow filtration[205], $FeCl_3$ precipitation[206] and PEG[207,208]. Our combined total cellular metagenomes, viromes, and metatranscriptomes provide further evidence of not only the community structure and composition but also the activity of those populations in freshwater lakes.

Knowing that phages are generally highly specific to their hosts and that their hosts are influenced by environmental factors (e.g stratification and mixing), we also hypothesized that the viral community in the hypolimnion would be distinct from that of the epilimnion. This was not observed, instead, sampling methodology was a greater contributor to community differences. Recently, phages have been suggested to have a broader range of potential hosts, even domains[209], potentially explaining why we observed less-specific phage-host impact than anticipated. Another implication of this work was the high proportion of "viral dark matter"[210] in the virome fraction. The identifiable viral fraction of Lake Mendota was dominated by the *Caudoviricetes* order (*Podoviridae*, *Myoviridae*, and *Siphoviridae*), as are many other freshwater lakes[165,205,206,211–214]. The high-quality, complete viral genomes reported here will expand the repertoire of freshwater phage genomes[211], which we anticipate becoming a genomic resource for the microbial ecology research community, enabling cross-site comparisons in the future.

Careful interpretation of phage AMG and their actual activity is needed

While several phage genomes were obtained, and many AMGs were found with functions ranging from core cellular metabolism to biogeochemically relevant pathways, improved ways to interpret the data remain an urgent need. Here, we observed up to 4.5% of all recovered complete and high-quality phage genomes were expressed at any given point in time. Even fewer were captured as expressing AMGs using RNA-seq. For example, while many AMGs were encoded by phages from both total metagenome and virome fractions (e.g. *psbA, psbD* in photosynthesis), many AMGs were sometimes only found in the total metagenomes (e.g. *pmoC* involved in methane and ammonia oxidation) or only in viromes (e.g. *soxB*). Further, capturing phage activities from metatranscriptomes could be a matter of luck, to sample and sequence the exact "time snapshot" when the lytic phase occurs. Few AMGs were expressed at any given sampling time, with most phages encoding core metabolism AMGs being the most expressed. These findings imply that depending on the research question and the biogeochemical cycle of interest, ecological interpretation from total metagenomes can over- or under-estimate true activity. This must be kept in mind when analyzing viral scaffolds from total metagenomes and inferring activity. For example, the lack of AMGs in a cycle of interest, and the lack of viruses with an identifiable host may be a result of a methodological bias (false negative), rather than true absence, and similarly, the presence of an AMG of interest may be in an inactive phage and inactive host.

Unlike for prokaryotic community identification, viral identification remains strongly impacted by laboratory methods. Methods prior to sequencing DNA impacted the viral community composition observed, functional genes found, and therefore reference genomes onto which transcriptomes were mapped onto to determine activity. These findings are applicable beyond freshwater phage

ecologists, as anyone working in viromics data and using bioinformatics tools to bring biological insight should be careful about introduced biases or acknowledge interpretation caveats. Second, we found that both prokaryotic and phage activity were not correlated with abundance, which implies that abundance-only or genomic potential methods can lead to overestimation or underestimation of activity. Finally, as environmental phage detection becomes applied to water quality, for example in the context of public health[215], and bioinformatics tools are developed to make sense of complex genomic data, careful attention should be given to the way that available datasets (mostly bulk metagenomes, not viral-fraction specific, or reference-based viral inference) impact detection and interpretation.

Methods

Field sampling

Lake Mendota, Wisconsin, USA was sampled at the "Deep Hole" location (Latitude 43°05'54", Longitude 89°24'28"), during the open-water period from July 7, 2020, to October 19, 2020. The samples were collected approximately weekly at 5 depths (5, 10, 15, 20, and 23.5m) and spanned the shift from a stratified water column with oxic, oxycline, and anoxic layers to a mixed oxygenated water column on October 18, 2020 (**Figure 1**).

Water was filtered back in the lab (less than 1-5 hours after sampling) through a serial vacuum pump through a 0.22μm filter to capture cells for total metagenome or metatranscriptome sequencing. The filters were stored at -80ºC until DNA or RNA could be extracted. The filtrate from 5, 15, and 23.5m was further amended with $FeCl_3$, a flocculation technique that concentrates virions for the viromes. The $FeCl_3$ protocol was adapted from[216,217] . In short, $FeCl_3$ stock was

made, and 1000uL was added to ~500mL of lake water filtrate. Bottles were manually mixed for 5 minutes and left at room temperature to incubate for 1 hour. Then $FeCl_3$-amended lake was filtered through a 0.8μM filter to collect the viral biomass.

DNA extraction

We sequenced 30 samples (16 total metagenomes and 14 viromes) through the US Department of Energy Joint Genome Institute (JGI). We chose those 30 samples out of the 200+ filters collected based on their position in the water column (i.e. whether they were in oxic, oxycline, or anoxic portions of the lake) (**Supplementary Table 1**), as well as to capture the pre- and post-fall mixing microbial community. The DNeasy PowerSoil kit (Qiagen) was used to extract DNA from both the >0.22 μm and <0.22 μm fractions. The DNA was quantified using the QuBit DNA.

RNA extraction

RNA extraction was performed using a phenol-chloroform protocol on 16 filters that corresponded to the 16 total metagenomes[164,218] . In summary, frozen filters were put into a bead-beating tube (e.g. PowerSoil (Qiagen)), and 1mL of TRIzol (Invitrogen) was added to each filter. After the filters were bead beaten for 1 minute at the medium speed, an internal standard (i.e. an *in vitro* transcribed pFN18A of a known concentration) was added[164,218] and inverted to mix. The samples were centrifuged for 1 min. The supernatant was transferred and 300uL of chloroform was added, mixed by inversion for 1 min, and incubated at room temperature for 3 minutes. After 15 min of centrifugation at 4^{o}C (13200xg), the supernatant was transferred to a new tube, and cold 70% ethanol was added and mixed well. Next, the RNeasy extraction kit (Qiagen) was used to recover the extracted RNA. DNase treatment was performed using the TURBO DNA-free™ Kit

(Invitrogen), which contains the DNase inactivation reagent. RNA was quantified using the QuBit RNA kit (Invitrogen).

Biogeochemical field data collection and laboratory analysis

Due to the COVID-19 pandemic, sampling and collecting of field data were limited. While we initially planned on collecting weekly sulfate, sulfide, nitrate, and nitrite data for 2020, we were unable to collect all biogeochemical data. Therefore, in 2021, we collected sulfate and sulfide. To collect sulfide in the field, 9 mL of lake water was fixed with 1 mL of 10% Zinc Acetate, mixed, then stored in a cooler with dry ice for sulfide measurements. Zinc acetate addition prevents the oxidation of sulfide with oxygen. In the lab, we used the Cline assay to measure sulfide.

For sulfate data collection, we collected ~10 mL of unfiltered lake water and froze it with dry ice in the field. Samples were brought back to the lab and immediately stored at -20°C until processing. Sulfate data was analyzed using ion-chromatography (IC) (Dionex 4mm ICS-2100 anion analyzer). Five standards were made with concentrations ranging from 0.8ppm - 80ppm of sulfate ($MgSO_4$ was used as the source of sulfate). 7 ml of each standard and sample were used form measurements. The loop size was 25μL, and the ion-chromatography columns used were AGII-HC and ASII-HC, the mobile phase was 30mM NaOH, and the suppressor was ASRS-4mm. One standard was checked every 10 samples to ensure proper measurements throughout. Results were analyzed in the software Chromeleon.

DNA sequencing

16 total metagenomes and 14 viromes were sequenced at the Joint Genome Institute (JGI) (**Supplementary Table 1)**. The JGI Gold Project ID is Gs0151897, under the name "Freshwater bacterial and viral communities from Lake Mendota, Wisconsin, United States". Details of sequencing for each sample can be found on the JGI portal, using the "Report" tab.

For the 16 total metagenomes, plate-based DNA library preparation for Illumina sequencing was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using the Kapa Biosystems library preparation kit. ~2 ng of sample DNA was sheared to 506 bp using a Covaris LE220 focused-ultrasonicator. The sheared DNA fragments were size selected by double-SPRI and then the selected fragments were end-repaired, A-tailed, and ligated with Illumina compatible sequencing adaptors from IDT containing a unique molecular index barcode for each sample library. The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The flowcell was sequenced on the Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2x151 indexed run recipe.

For the 14 viromes, ~0.02 ng of DNA was denatured and sheared to ~300 bp using the Covaris LE220, and >300 bp was size selected using SPRI beads (Beckman Coulter). The fragments were denatured again and a library was created using Accel-NGS 1S Plus DNA Library Kit (Swift Biosciences) to ligate Illumina-compatible adapters (IDT, Inc). qPCR was used to determine the concentration of the libraries which were sequenced on the Illumina platform. The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit

and run on a Roche LightCycler 480 real-time PCR instrument. The flowcell was sequenced on the Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2x151 indexed run recipe.

RNA sequencing and processing

RNA sequencing was performed at SeqCenter (PA). Illumina Stranded RNA library preparation with RiboZero Plus rRNA depletion to offer RNA Sequencing. Metatranscriptomes were quality-trimmed using fastp with a quality score cutoff of 20. The following non-coding rRNA reads were removed using sortmerna (v4.2.0)[219] : 5.8S and 5S based on RFAM database[220] ; archaeal and bacterial 16S and 23S based on SILVA database (v.119); and eukaryotic 18S and 28S based on SILVA database (v.119)[221]. Reads mapping to the internal standard were then identified using sortmerna and quantified. Reads not associated with the internal standard or rRNA were used for further analysis.

Metagenome assembly, binning, and MAG quality filtering

The quality of the assembled scaffolds, assembled with SPADES v.3.13.0[74] with the following options: -m 2000 -o spades3 --only-assembler -k 33,55,77,99,127 –meta) was assessed by 1) quantifying the percentage of original reads mapping back to the scaffolds, with a value of >80% meaning a good assembly quality 2) the distribution of the lengths of the scaffolds in the assembly and 3) the N50 value of the assembly. This information is found in the PDF report from JGI available on IMG/M.

A complete workflow for the metagenomic binning is shown in **Supplementary Figure 6**. Binning was performed using Metawrap v1.3.2[222], a binning wrapper that relies on Metabat1 v0.32.5[1690], Metbat2 v2.12.1[91], Maxbin2[92], using default settings, but using differential coverage generated by mapping reads from each of the 16 metagenomes to each of the 16 assemblies. Cross-mapping here means that all 16 interleaved FASTQ files were used to bin each 16 total metagenomes. We compared 3 different binning methods: Binning using reads from itself, binning using reads from samples from the same ecosystem (oxic, anoxic-oxic, and anoxic depths) "within layer", and cross-mapping across all samples. We chose cross-mapping for all samples because compared to binning using the respective assembly only *or* binning with reads from the same environmental layer, cross-mapping resulted in the highest number of bins *and* the highest average bin completeness (**Supplementary Figure 7**). Further, to improve bin quality and completeness, we next performed bin refinement.

Bin refinement was performed using DAS Tool v.1.1.3[93]. Bin refinement involves clustering bins that may be from the same population and uses algorithms to build "better", more complete bins[93]. The bins provided for each sample came from metabat1, metabat2, maxbin (ran through Metawrap) using differential coverage, and a set of independently generated bins from JGI, which uses metabat2 but without the differential coverage (that is 1 fastq file was used with each 1 assembly during the metabat2 command). These differences in metabat2 versions and settings used are why we provided all "four" binning results as input for bin refinement. After bin refinement, on average, 15 to 20% of bins were kept (1,445 MAGs).

We then performed a taxonomic assignment using GTDB-tk v.1.7.0[78]. This was done early, as opposed to after getting the final bin set, because the dereplication step relies on genome quality estimate calculation by CheckM v1.1.3[68]. CheckM can be run without default marker genes, or with a CPR (i.e. Patescibacteria) marker gene set, which results in more accurate estimates for CPR genomes. By running the taxonomic assignment first, we will divide the refined bin set into CPR genomes and non-CPR genomes, and then run CheckM with the appropriate marker gene set accordingly. In total, 43 CPR bins, and 1,402 non-CPR bins (17 of which were Archaea).

To reduce redundancy in the MAGs across the 16 metagenome samples, dereplication was done using dRep v.3.2.2[77]. Unlike bin refinement which improves upon existing bins and may create "new" bins from the inputs, dereplication clusters bins and selects one representative bin from each cluster. As we move through the processing of the metagenomic data, it is always possible to go back, and even though only 1 representative genome is selected after dereplication, it is possible to know which other very closely related genomes existed, which can be useful for projects outside the scope of this paper. The completeness ≥50% and contamination <=10% thresholds were used as they generally represent MIMAG standards[94]. After dereplication, we obtained 18 CPR MAGs and 413 non-CPR MAGs (2 were Archaea) for a total of 431 MAGs. These 431 MAGs consisted of the "final bin set".

MAG naming convention

MAGs from the final bin set are described in **Supplementary Table 2**. The naming pattern is as follows: "GaXXXXXXX" represents the GaID of the metagenome or viromes, searchable on the JGI IMG/M website, from which the bin was originally generated. Following are "metabat1",

"metabat2_ours", or "maxbin2" representing the binning tool used, all with differential coverage. If the name has "_jgi", the MAG was obtained using metabat2 without differential coverage, as originating from the standard JGI sample processing pipeline. Following is a number with 3 digits, representing an arbitrary bin number within each binning run. If the bin name has "_sub" in it, it means the MAG results from bin refinement (and improved upon the original binning". Because the final bin set represents closely related MAGs (dereplicated set), the GaID of the MAG does not mean the MAG is from that sample only, instead that the best representative MAG originated from it.

Assessing genome coverage and transcript abundance per MAG

We used CoverM[223] to obtain genome-level relative abundance and expression for all 431 MAGs and phages. We used CoverM with the option -m rpkm to assess the expression of the 431 MAGs across all the samples. Because we used internal standards, we were able to get a read count of the internal standard for each of the 16 samples. We divided the rpkm by the respective internal standard read count to obtain the "rpkm normalized" value. To calculate the genome-level coverage, we used Bowtie2 v.2.2.5[224] with default settings to build an index of the 431 MAGs, sorted it with samtools sort v.1.14[225], and then used inStrain v.1.5.7[226] quick_profile to obtain per-genome depth and breadth of the coverage.

Assessing the metabolic potential of the MAGs

We used METABOLIC v.4.0 [105] on all 431 MAGs to get an overview of genes involved in carbon, nitrogen, sulfur cycling, and in general core metabolism (KEGG[227], Pfam[228], CAZyme[229], MEROPS[230]).

Phylogenetic tree

A concatenated gene phylogeny of 16 ribosomal proteins from the 431 MAGs was created to gain an overview of the taxonomic diversity in Lake Mendota, and visualize the distribution of the organisms over time, and any metabolic genes of interest. The function “create-gene-phylogeny” in MetabolisHMM v. 2.21[231] was used to create a bacterial tree of 429 MAGs (the 2 archaeal genomes were not included as different marker genes are used when building an archaeal tree and a bacterial tree, and because a 2-genome tree is not very informative). The *fasta* alignment file from metabolisHMM was then used as an input to generate a maximum-likelihood phylogenetic tree using RAxML Blackbox HPC v.8.2.11[232], through the CIPRES server v.3.3. iTOL (interactive Tree of Life)’s interface was used to visualize and annotate the tree.

Diversity Index Calculations

The Shannon Diversity Index and the beta-diversity index were calculated in R using the package "vegan” v.2.5.7 with an abundance matrix grouped at the class level.

Viral identification and AMG annotation

VIBRANT[233] was used on the assembled scaffolds of the 16 metagenomes and the 14 viromes to identify viral genomes, and to identify whether the phage was likely lytic (non-integated) or lysogenic. The -virome option was used on the viromes. In total, about 500,000 phages were found, representing about 4 million protein-coding sequences. Phage genome quality was assessed with CheckV[153].

Phage-host interactions

The method of host homology was initially used to assess sequence similarity between the 431 MAGs and all phages that we recovered from Lake Mendota using DIAMOND blastp[234]. We first identified CRISPR sequences in the 431 MAGs to narrow down interactions, as CRISPR sequences in bacterial hosts represent the phages that may infect them. CRISPR sequences were identified with piler-cr[235], to identify repeats and spacers. To link MAGs containing spacer sequences to phages that may infect, we used BLAST homology of identified CRISPR spacers vs. all the Lake Mendota phages, allowing for at most 1 mismatch. To further attempt to assign phage-host interactions, especially for phages whose hosts are not represented by MAGs, we selected the 2,246 complete and high-quality phages in our study and ran iPHoP[182] to assign putative hosts using the regular iPHoP database, with confidence scores greater or equal to 90. We also ran the search on a custom database by adding the 431 MAGs from our study to the regular iPHoP (*add_custom_db* option).

Taxonomic assignment of phages

We used Diamond blastp against the NCBI taxonomic database to assign taxonomy to the phages.

Data availability

The data is available at NCBI BioProject PRJNA758276, which contains all 431 prokaryotic metagenome-assembled genomes and the processed 16 RNA-seq samples. The phage genomes can be accessed at https://doi.org/10.6084/m9.figshare.22213846 . The original and assembled metagenomes were sequenced through the JGI Gold Gs0151897, and the sequenced data and

assembled scaffolds can be downloaded on JGI IMG/M. The R code is available at: https://github.com/patriciatran/LakeMendota-PairedOmics

Acknowledgments

We are grateful to the Center for Limnology and the Long-Term Ecological Research – North Temperate Lakes group for supporting field sampling. PQT received funding from the Natural Science and Engineering Research Council of Canada (NSERC) doctoral fellowship, the Anna Grant Birge Memorial Award from the Center for Limnology, and the Baldwin Distinguished Graduate Fellowship from the Department of Bacteriology at the University of Wisconsin-Madison. DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via a Community Science Program New Investigator award to KA and PQT (award number 506328). KA acknowledges support from the USDA National Institute of Food and Agriculture (NIFA) under grant: Hatch project 1025641, and from the National Science Foundation grant DBI-2047598.

Chapter 6 Conclusions and Perspectives

Concluding remarks

In this dissertation, I presented a comprehensive analysis of microbial communities, specifically bacteria, archaea, and phages, and their impact on biogeochemical cycling in anoxic lakes. My approach involved a combination of cultivation-dependent and -independent methods, providing a holistic understanding of microbial interactions that are highly dynamic in time and space. By leveraging interdisciplinary strengths, I have demonstrated that microbial communities are interacting members of an ecosystem, and their impact extends beyond their immediate environment. In the next sections, I detail examples of how the work here has impacted the field of microbial ecology & genomics and present opportunities for future research.

Implications for global microbial ecology

Global Patterns of Biodiversity in Deep Lakes

This work expanded the knowledge of deep-lake microbiology, especially with the genomic characterization of Lake Tanganyika's microbiome. Since starting this chapter in 2018, work in Lake Baikal[19,146,169,202,236–239], has also brought insight into the microbiome of this ultra-deep lake. We found that despite the two lakes being ancient and deep lakes, they each have endemic organisms, for example, Lake Tanganyika has a high percentage of archaea and CPR taxa. Recent studies have investigated the global freshwater CPR through ecogenomics perspectives[240], and since 2018, have gained recognition in freshwater environments[241]. In some ways, Lake Tanganyika is more reminiscent of marine counterparts. Other deep lakes around the world, such as the Laurentian Great Lakes[242], could provide insight about common biogeographical patterns of biodiversity in these deep lakes. When writing **Chapter 4,** one theme

was to connect microbes across biogeochemical functions. Since the development of one of the projects I co-led, METABOLIC[105], the tool which was used to annotate carbon, nitrogen and sulfur cycling in Lake Tanganyika, has been applied to multiple environments. Scientists now have a better understanding of the role of microbial communities in these cycles in different environments such as groundwater[243], disturbed ecosystems[244], coral microbiomes[245], and in mangrove sediments[246], as well as several other published work. The ability to scale up these analyses, for example to support simultaneous laboratory experiments, alleviates the computational and biological bottleneck of microbiome interpretation.

Lake Mendota as a model ecosystem

Lake Mendota is known as the longest-standing freshwater lake monitoring site, the basecamp of the Long-Term Ecological Research North Temperate Lakes (LTER-NTL) since 1983 (https://lter.limnology.wisc.edu/). This dissertation contributes to two major aspects of the LTER-NTL and Microbial Observatory which were absent before:

1. Weekly samples and analysis of the hypolimnetic microbial community and function.
2. Incorporating viruses in environmental monitoring, which intersects with human and public health dimensions.

Additional samples archived and available to be used for future projects.

Since 2018, I collected hundreds of microbial samples (Figure 6.1), some were analyzed as part of this dissertation, and many remain as archival samples or to be used for future projects. Most samples are on filters, have replicates, are stored in a –80ºC freezer, are collected over various depths, and cover both ice-covered and open-water periods of the year. Potential applications

including generation of metagenomes and performing qPCR to quantify genes of interest. A list of all samples is found in **Appendix 1.**

The already sequenced data is available on the Joint Genome Institute IMG/M portal and the National Center for Biotechnology Information (NCBI)'s GenBank Repository (see publication data available in each chapter). The data represents a great starting point for someone wanting to learn computational microbiology research using real-world datasets, while at the same time focusing on a research question of interest. For example, they may serve as preliminary data to answer questions beyond the scope of this dissertation for example, the viral community of Lake Tanganyika, comparative genomics of ecologically relevant taxa (phylogenetics), or comparative genomics and transcriptomics of functions of interest (e.g. aerobic methane oxidizers).

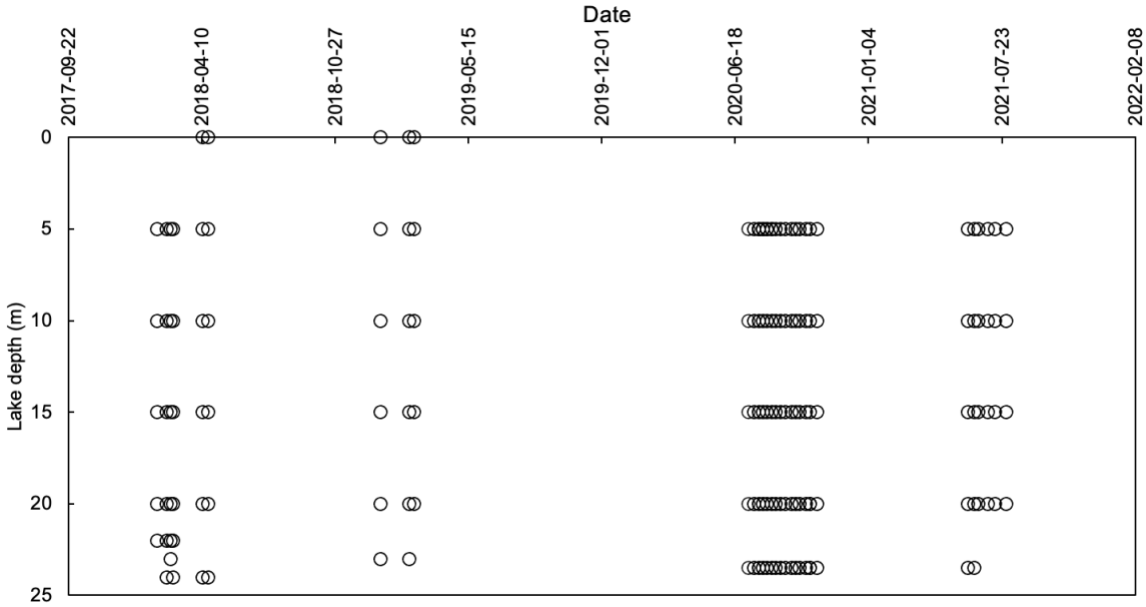

Figure 6.1 Filters collected in Lake Mendota (WI, USA) during this dissertation. Some have been extracted for this dissertation, and many are stored as archived for future projects.

Environmental change and responses across multi-kingdom microbial community

Prior work[247,248] highlighted that microbes are resilient to environmental change, but that different members might respond differently (e.g. bacteria vs phytoplankton). Here, I add the layer of bacteriophages, which did not respond to environmental factors (stratification and anoxia) the same way the bacteria did. In a 2008 thesis, Jones observed that the microbial community was rapidly mixed after a typhoon event, but this was not observed in the phytoplankton community. In 2010, Shade's dissertation work replicated these full water column mixing experiments in Wisconsin lakes and showed that microbial communities (bacterial) responded differently in the hypolimnion and the epilimnion, and showed that the communities were resilient to the change. Given that the phytoplankton and viral communities responded differently to environmental change, a future experiment could test whether sudden disturbance results in the same resiliency in the viral community. This brings an interesting future area of research, as environmental drivers of community composition are not the same for all microbes.

Prioritizing depth-discrete hypolimnetic sampling

Anoxia is relevant to the theme of how lakes will respond to climate change. In recent years the decline in oxygen has been an area of study in Lake Mendota. For example, in the 2020 LTER proposal[249], Q.1.2 seeks to understand how climatic factors will lead and impact anoxia. One of the consequences of climate change is the decline in oxygen in Wisconsin lakes, which leads to oxy-thermal squeeze[250]. This is of interest too, because the anoxic hypolimnion responds differently to climatic drivers than the oxic epilimnion, which is of concern when estimating metabolism. For example, anoxia impacts carbon, phosphorus and nitrogen concentration,

availability, and retention, making lakes "sinks" less[251]. Anoxia is also intricately linked to various climate change and biological dynamics. For example, after the 2020 spiny water flea invasion of Lake Mendota, the anoxia lag (time to anoxia post-summer start) is shortened, the bloom expands earlier into the season, sinks into the anoxic layers, and potentially would increase oxygen demand[11]. The relationship between anoxia, cyanobacteria, and sulfate-reducing bacteria has been shown elsewhere as well[252].

While there is an interest in moving away from process-based (and mechanistic) models of metabolism due to the substantial amount of biological and chemical data needed to calibrate them, future approaches such as machine learning-based models require data to be trained. Years ago, the need for process-based models was highlighted[253]. Given the underrepresentation of hypolimnetic data for Lake Mendota, when compared to measurements in the upper water column, focusing on hypolimnetic processes in sampling strategies is essential to acquiring better data for models. For example, we start with expanding the microbial data for the hypolimnion and collected YSI profiles of temperature, DO, etc. all the way down to 23.5m deep (the buoy stops at 20m and the LTER Microbial Observatory stops at 20m). Our data collected in 2020 and 2021 is readily available as part of the Core Datasets of LTER[254]. Our chemical and biological data showed that the hypolimnion is a heterogeneous environment with strong gradients, and even over meter-long distances, there are differences in the microbial community and chemical concentrations (e.g. sulfide).

Anoxic and oxic-anoxic interface hypolimnion microbial community

Prior work in Lake Mendota has highlighted the importance of the oxic-anoxic interface in understanding microbially-mediated processed. For example, Hale's 2022 work[255] showed that

Cyanobacteria, typically studied in the context of bloom-forming surface microbes, exist beneath the surface, and those lower-down produced toxins. Peterson showed that the hypolimnetic member Kiritimatiellia was found to be a group of interest based on the prior profile of the hypolimnion[256] . This group found in the water column of freshwater lakes, encoded the *hgcA* gene responsible for mercury methylation. At the time, the lack of representative full genomes could not allow for full genomic reconstruction. Here, we found that those groups were largely inactive, at least at the transcriptional level. In fact, we mapped RNA-seq data onto the Lake Mendota MAGs, from **Chapter 5**, which contained the *hcgA* genes to determine their activity. While the Verrucomicrobia/Kiritimatiellia that contained hgcA genes were abundant in the metagenome, those belonging to Bacteroidetes and Desulfobacteriota were those that were the most transcriptionally active. These groups also happen to be notable because they have many phages infecting them.

Cryptic Biogeochemical Cycling

Having the isolates from **Chapter 2** will provide the opportunity to build a model system to further characterize oxic hydrogen sulfide production. When presenting this work at seminars and conferences, two major questions often came up: *why* would Chryseobacterium produce hydrogen sulfide from cysteine when theoretically cysteine could be used for biomass instead, and what is the environmental significance of cysteine? To answer the first question, detailed physiological characterization is needed, for example, one could isotopically label cysteine and use GC-MS to track how the cysteine is processed by the cell. This metabolic flux experiment would first require method development to measure cysteine, along with nitrogen and hydrogen sulfide products. Second, while kits are readily available to measure cysteine in blood, previous ways to measure

this amino acid in environmental samples were either expensive or impractical in the field. Hotvedt developed a cheap and fast way to measure cysteine in aquatic samples using HPLC available at UW-Madison[71]. It would be interesting to collect field samples and apply this method to quantify environmental cysteine to characterize the dynamics of cysteine in a natural system. Similarly, it is uncertain the speed at which H_2S would be produced in an environmental setting, as the oxidation or uptake of H_2S by other microbial organisms could explain the lack of H_2S measured in the lake. After all, the preliminary data that sparked our investigation was because we observed that H_2S *had* been produced, whereas field-based measurements only capture the net amount at a specific point in time.

Environmental viruses: detection, ecology, and implications

Viruses, which are known to impact all realms of life, had previously been shown to impact microbial ecology in marine environments[139,150,257,258]. These bacteriophages were known to control bacterial community structure, composition, and biogeochemical function – from an ecological standpoint, the interest was warranted.

Viral detection

In the recent months and years, the field of environmental viral detection has blossomed. Viral metagenomics shows promise for characterizing viral dark matter[210,259], which enables scientists to identify and characterize viruses. Without prior need to isolate and maintain both the virus and the host isolated in the lab, cultivation-independent methods can help identify them, identify hosts, and track them in the environment. Our work, although in freshwater lakes, is applicable to our understanding and applications of environmental monitoring of viruses and pathogens. Existing

technologies to monitor viruses include flow-cytometry (pass a virus sample through a "beam" and viruses which "match" will be reported), qPCR (viruses whose DNA, or RNA, match specific genomes of known viruses will be counted). These methods are often faster and cheaper than metagenomic sequencing. However, one major limitation of those methods is that they do not detect what we do not know. Environmental metagenomics enables us to discover this unknown diversity using *de novo* methods. Our work has 2 practical warnings however:

1. Pre-sequencing methods to generate environmental metagenomes of viruses impact the viruses detected.
2. There is currently a lack of environmental viromes in public databases, which biases some new bioinformatics tools being developed to identify viruses.

Some of the active work building upon this dissertation is figuring out if the patterns we observe in freshwater lakes are applicable to other systems. For example, we observed a difference in the number, quality, and auxiliary metabolic genes of the viruses when comparing metagenomes vs. viromes. Researchers in our lab are now trying to apply similar analysis in the human gut, marine, and their ecosystems, and how applicable and transferrable these findings are across systems. Additionally, we found that viromes had more unclassified viruses, and further studies could investigate these viral sequences. For example, viral binning[260] could reconstruct more complete genomes, which will enable better taxonomic classification against reference databases.

Viral ecology

The field of environmental virology is rapidly progressing in recent years. In ecology, viruses can impact dozens of taxa in a single ecosystem[188], infect archaea[261], and display various broad phage specificity ranges[187]. In this dissertation, we showed that hundreds of viruses were associated with

at least 55 bacterial genomes, and RNA-seq data showed that the viruses and their bacterial hosts displayed dynamic activity patterns. If one were to use the filters (**Appendix 1**) to generate new weekly data, one could get more highly resolved time- and spatial-dynamics of these phage-host interactions.

An aspect of viral ecology that was not addressed in this dissertation are RNA viruses. RNA viruses are a whole group of viruses that can infect eukaryotes and bacteria. RNA viruses include human pathogens such as the human immunodeficiency virus 1 (HIV1), influenza (IAV), the Dengue virus (DV) and the Measles virus (MV) and the SARS-CoV2. Recently, re-analyses of global RNA viromes datasets have expanded viral diversity, uncovering novel phyla[262]. Future studies could investigate the RNA viral diversity and ecology in Lake Mendota using the existing transcriptomics data generated in this dissertation (available on NCBI).

Applications in public health

The field of microbiology was revolutionized when scientists were able to identify and infer functions from sequencing data. Wastewater-based epidemiology (WBE)[263] involves collecting environmental samples from wastewater treatment plants, as the populations shed DNA and RNA in their stool, which eventually makes their way to these treatment plants. The idea is that sequencing the RNA and DNA and comparing them against the reference of known pathogens (e.g. SARS-CoV, *E. coli*, influenza virus, etc.), could help public health agencies in monitoring and tracking trends and outbreaks. In the US, the National Wastewater System Surveillance Initiative by the Center for Disease Control (CDC)[264], and in State public health laboratories' monitoring programs[265], are examples of this effort. The technologies, methods, and themes in this dissertation: viral detection from environmental samples, time-series analyses, and bioinformatics

approaches to uncover viral diversity, predict hosts from genomic data, and phylogenetics could be applied to help address these future public health challenges.

References (whole thesis)

1. Richard W. Healy, Thomas C. Winter, James W. LaBaugh, & O. Lehn Franke. *Water Budgets: Foundations for Effective Water-Resources and Environmental Management*. (2007).
2. Tranvik, L. J., Cole, J. J. & Prairie, Y. T. The study of carbon in inland waters—from isolated ecosystems to players in the global carbon cycle. *Limnol Oceanogr Lett* **3**, 41–48 (2018).
3. Adrian, R. *et al.* Lakes as sentinels of climate change. *Limnol Oceanogr* **54**, 2283–2297 (2009).
4. Jane, S. F. *et al.* Widespread deoxygenation of temperate lakes. *Nature* **594**, 66–70 (2021).
5. Woolway, R. I. & Merchant, C. J. Worldwide alteration of lake mixing regimes in response to climate change. *Nat. Geosci.* **12**, 271–276 (2019).
6. Bush, T. *et al.* Oxic-anoxic regime shifts mediated by feedbacks between biogeochemical processes and microbial community dynamics. *Nature Communications* **8**, (2017).
7. Understanding the Mineral Resources of the Midcontinent Rift | U.S. Geological Survey. https://www.usgs.gov/news/science-snippet/understanding-mineral-resources-midcontinent-rift.
8. Takahashi, T. & Koblmüller, S. The Adaptive Radiation of Cichlid Fish in Lake Tanganyika: A Morphological Perspective. *International Journal of Evolutionary Biology* **2011**, 1–14 (2011).
9. Branchu, Ph. & Bergonzini, L. Chloride concentrations in Lake Tanganyika: an indicator of the hydrological budget? *Hydrol. Earth Syst. Sci.* **8**, 256–265 (2004).

10. Duffy, C. J., Dugan, H. A. & Hanson, P. C. The age of water and carbon in lake-catchments: A simple dynamical model. *Limnology and Oceanography Letters* **3**, 236–245 (2018).
11. Rohwer, R. R. *et al. The aftermath of a trophic cascade: Increased anoxia following species invasion of a eutrophic lake*. http://biorxiv.org/lookup/doi/10.1101/2023.01.27.525925 (2023) doi:10.1101/2023.01.27.525925.
12. Huttula, T. & Sarvala, J. Tanganyika Lake: Strong in Hydrodynamics, Diverse in Ecology. in *Encyclopedia of Lakes and Reservoirs* (eds. Bengtsson, L., Herschy, R. W. & Fairbridge, R. W.) 776–782 (Springer Netherlands, 2012). doi:10.1007/978-1-4020-4410-6_168.
13. Baxter, P. J., Kapila, M. & Mfonfu, D. Lake Nyos disaster, Cameroon, 1986: the medical effects of large scale emission of carbon dioxide? *BMJ* **298**, 1437–1441 (1989).
14. Schoell, M., Tietze, K. & Schoberth, S. M. Origin of methane in Lake Kivu (East-Central Africa). *Chemical Geology* **71**, 257–265 (1988).
15. Jones, N. How dangerous is Africa's explosive Lake Kivu? *Nature* **597**, 466–469 (2021).
16. Faber, M. D. Microbial degradation of recalcitrant compounds and synthetic aromatic polymers. *Enzyme and Microbial Technology* **1**, 226–232 (1979).
17. Azam, F. *et al.* The Ecological Role of Water-Column Microbes in the Sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).
18. The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0048969719335946.

19. Cabello-Yeves, P. J. *et al.* Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. *Limnology and Oceanography* **65**, 1471–1488 (2020).
20. Linz, A. M., He, S., Stevens, S. L. R., Anantharaman, K. & Robin, R. Connections between freshwater carbon and nutrient cycles revealed through. **221**, (2018).
21. Linz, A. M. *et al.* Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog Lakes. *mSphere* **2**, e00169-17 (2017).
22. Rohwer, Robin R. Temporal and Taxonomic Scales of Change in Freshwater Microbial Community Composition. (University of Wisconsin - Madison, 2019).
23. Capo, E. *et al.* Landscape Setting Drives the Microbial Eukaryotic Community Structure in Four Swedish Mountain Lakes over the Holocene. *Microorganisms* **9**, 355 (2021).
24. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal* **10**, 1589–1601 (2016).
25. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14 (2011).
26. Bastviken, D., Persson, L., Odham, G. & Tranvik, L. Degradation of dissolved organic matter in oxic and anoxic lake water. *Limnology and Oceanography* **49**, 109–116 (2004).
27. Diao, M., Huisman, J. & Muyzer, G. Spatio-temporal dynamics of sulfur bacteria during oxic--anoxic regime shifts in a seasonally stratified lake. *FEMS Microbiology Ecology* **94**, (2018).
28. Diao, M., Sinnige, R., Kalbitz, K., Huisman, J. & Muyzer, G. Succession of Bacterial Communities in a Seasonally Stratified Lake with an Anoxic and Sulfidic Hypolimnion. *Frontiers in Microbiology* **8**, (2017).

29. Lau, M. P. & del Giorgio, P. Reactivity, fate and functional roles of dissolved organic matter in anoxic inland waters. *Biology Letters* **16**, 20190694 (2020).

30. Foti, M. *et al.* Diversity, Activity, and Abundance of Sulfate-Reducing Bacteria in Saline and Hypersaline Soda Lakes. *Appl. Environ. Microbiol.* **73**, 2093–2100 (2007).

31. INGVORSEN, K. & BROCK, T. ELECTRON FLOW VIA SULFATE REDUCTION AND METHANOGENESIS IN THE ANAEROBIC HYPOLIMNION OF LAKE MENDOTA. *LIMNOLOGY AND OCEANOGRAPHY* **27**, 559–564 (1982).

32. Tchong, S.-I., Xu, H. & White, R. H. L-cysteine desulfidase: an [4Fe-4S] enzyme isolated from Methanocaldococcus jannaschii that catalyzes the breakdown of L-cysteine into pyruvate, ammonia, and sulfide. *Biochemistry* **44**, 1659–1670 (2005).

33. Berg, J. S. *et al.* Microbial diversity involved in iron and cryptic sulfur cycling in the ferruginous, low-sulfate waters of Lake Pavin. *PLOS ONE* **14**, e0212787 (2019).

34. Vigneron, A. *et al.* Genomic evidence for sulfur intermediates as new biogeochemical hubs in a model aquatic microbial ecosystem. *Microbiome* **9**, 46 (2021).

35. Göbbels, L. *et al.* Cysteine: an overlooked energy and carbon source. *Sci Rep* **11**, 2139 (2021).

36. Petersen, M., Schreiner, K. M., Hyde, E., Sheik, C. & Katsev, S. Characterization of Organic Sulfur in Lake Superior, and its Contribution to Sulfur Biogeochemical Cycling. **2019**, B11L-2247 (2019).

37. Kamyshny, A., Druschel, G., Mansaray, Z. F. & Farquhar, J. Multiple sulfur isotopes fractionations associated with abiotic sulfur transformations in Yellowstone National Park geothermal springs. *Geochem Trans* **15**, 7 (2014).

38. Eckert, W. & Conrad, R. Sulfide and Methane Evolution in the Hypolimnion of a Subtropical Lake: A Three-Year Study. *Biogeochemistry* **82**, 67–76 (2007).

39. Peterson, B. D. *et al.* Mercury Methylation Genes Identified across Diverse Anaerobic Microbial Guilds in a Eutrophic Sulfate-Enriched Lake. *Environ. Sci. Technol.* **54**, 15840–15851 (2020).

40. Hine, C. & Mitchell, J. R. Calorie restriction and methionine restriction in control of endogenous hydrogen sulfide production by the transsulfuration pathway. *Exp Gerontol* **68**, 26–32 (2015).

41. Stipanuk, M. H. Metabolism of Sulfur-Containing Amino Acids: How the Body Copes with Excess Methionine, Cysteine, and Sulfide. *The Journal of Nutrition* **150**, 2494S-2505S (2020).

42. Reese, B. K., Anderson, M. A. & Amrhein, C. Hydrogen sulfide production and volatilization in a polymictic eutrophic saline lake, Salton Sea, California. *Sci Total Environ* **406**, 205–218 (2008).

43. Zhou, Y. & Imlay, J. A. Escherichia coli K-12 Lacks a High-Affinity Assimilatory Cysteine Importer. *mBio* **11**, e01073-20.

44. Pal, V. K., Bandyopadhyay, P. & Singh, A. Hydrogen Sulfide in Physiology and Pathogenesis of Bacteria and Viruses. *IUBMB Life* **70**, 393–410 (2018).

45. An, S. & Berg, G. Stenotrophomonas maltophilia. *Trends in Microbiology* **26**, 637–638 (2018).

46. Chen, M.-X., Li, H.-Y., Ye, X.-S. & He, X.-Y. Draft Genome Sequence of an Extracellular Protease-Producing Bacterium, Stenotrophomonas bentonitica VV6, Isolated from Arctic Seawater. *Genome Announcements* **6**, e01610-17.

47. Sánchez-Castro, I., Bakkali, M. & Merroun, M. L. Draft Genome Sequence of Stenotrophomonas bentonitica BII-R7T, a Selenite-Reducing Bacterium Isolated from Spanish Bentonites. *Genome Announcements* **5**, e00719-17.

48. de Beer, H. *et al.* Chryseobacterium piscium sp. nov., isolated from fish of the South Atlantic Ocean off South Africa. *International Journal of Systematic and Evolutionary Microbiology* **56**, 1317–1322.

49. Sun, Q. *et al.* Structural Basis for the Inhibition Mechanism of Human Cystathionine γ-Lyase, an Enzyme Responsible for the Production of H2S *. *Journal of Biological Chemistry* **284**, 3076–3085 (2009).

50. Hatzenpichler, R., Krukenberg, V., Spietz, R. L. & Jay, Z. J. Next-generation physiology approaches to study microbiome function at single cell level. *Nat Rev Microbiol* **18**, 241–256 (2020).

51. Hubberten, U., Lara, R. J. & Kattner, G. Amino acid composition of seawater and dissolved humic substances in the Greenland Sea. *Marine Chemistry* **45**, 121–128 (1994).

52. Cindy & Lee. Amino Acid and Amine Biogeochemistry in Marine Particulate Material and Sediments. https://www.semanticscholar.org/paper/Amino-Acid-and-Amine-Biogeochemistry-in-Marine-and-Cindy.-Lee/8f4bba4498ee7eeebd720492a938fe792ecb91cf (2005).

53. Bagiyan, G. A., Koroleva, I. K., Soroka, N. V. & Ufimtsev, A. V. Oxidation of thiol compounds by molecular oxygen in aqueous solutions. *Russian Chemical Bulletin* **52**, 1135–1141 (2003).

54. Korshunov, S., Imlay, K. R. C. & Imlay, J. A. Cystine import is a valuable but risky process whose hazards Escherichia coli minimizes by inducing a cysteine exporter. *Molecular Microbiology* **113**, 22–39 (2020).

55. Brosnan, J. T. & Brosnan, M. E. The Sulfur-Containing Amino Acids: An Overview. *The Journal of Nutrition* **136**, 1636S-1640S (2006).

56. Thume, K. *et al.* The metabolite dimethylsulfoxonium propionate extends the marine organosulfur cycle. *Nature* **563**, 412–415 (2018).

57. Yamamoto, K., Oshima, T., Nonaka, G., Ito, H. & Ishihama, A. Induction of the Escherichia coli cysK gene by genetic and environmental factors. *FEMS Microbiol Lett* **323**, 88–95 (2011).

58. Sharma, S. *et al.* A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. *Sci Data* **2**, 150008 (2015).

59. Chen, M. *et al.* Increasing sulfate concentrations result in higher sulfide production and phosphorous mobilization in a shallow eutrophic freshwater lake. *Water Res* **96**, 94–104 (2016).

60. Brock, T. D., Brock, T. D. & Brock, W. A. *A Eutrophic Lake: Lake Mendota, Wisconsin*. (Springer-Verlag, 1985).

61. Jørgensen, N. O. G. Free amino acids in lakes: Concentrations and assimilation rates in relation to phytoplankton and bacterial production1. *Limnology and Oceanography* **32**, 97–111 (1987).

62. Gardner, W. & Lee, G. The role of amino acids in the nitrogen cycle of Lake Mendota. *Limnology and Oceanography - LIMNOL OCEANOGR* **20**, 379–388 (1975).

63. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. USA* 5 (1985).

64. Nucleobytes. 4Peaks: For peaks, four peaks. The DNA sequence trace viewer for OS X.

65. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

66. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13**, e1005595 (2017).

67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

68. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–55 (2015).

69. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996–1004 (2018).

70. Zhou, Z., Tran, P., Liu, Y., Kieft, K. & Anantharaman, K. METABOLIC: A scalable high-throughput metabolic and biogeochemical functional trait profiler based on microbial genomes. *bioRxiv* 761643 (2019) doi:10.1101/761643.

71. Hotvedt, Jacob C & Anantharaman, Karthik. Rapid and cheap quantification of Cysteine viaoxidationto Cystine by DMSO and Ortho-phthalaldehyde pre-column derivatization. (2021).

72. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* **8**, (2017).

73. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

74. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

75. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner (2014).

76. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

77. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* 1–5 (2017) doi:10.1038/ismej.2017.126.

78. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

79. Eddy, S. R. Accelerated profile HMM searches. *PLoS Computational Biology* **7**, (2011).

80. Aramaki, T. *et al.* KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *bioRxiv* (2019) doi:10.1101/602110.

81. Alin, S. R. & Johnson, T. C. Carbon cycling in large lakes of the world: A synthesis of production, burial, and lake-atmosphere exchange estimates. *Global Biogeochemical Cycles* **21**, GB3002 (2007).

82. Durisch-Kaiser, E. *et al.* What prevents outgassing of methane to the atmosphere in Lake Tanganyika? *Journal of Geophysical Research* **116**, (2011).

83. Salzburger, W. Understanding explosive diversification through cichlid fish genomics. *Nat Rev Genet* **19**, 705–717 (2018).

84. Corman, J. R. *et al.* Upwelling couples chemical and biological dynamics across the littoral and pelagic zones of Lake Tanganyika, East Africa. *Limnology and Oceanography* **55**, 214–224 (2010).

85. Cabello-Yeves, P. J. *et al.* Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. *Applied and Environmental Microbiology* AEM.02132-17 (2017) doi:10.1128/AEM.02132-17.

86. De Wever, A. Spatio-temporal dynamics in the microbial food web in Lake Tanganyika. *University of Gent* 1–169 (2006).

87. Pirlot, S., Unrein, F., Descy, J.-P. & Servais, P. Fate of heterotrophic bacteria in Lake Tanganyika (East Africa): Fate of bacteria in Lake Tanganyika. *FEMS Microbiology Ecology* **62**, 354–364 (2007).

88. Schubert, C. J. *et al.* Anaerobic ammonium oxidation in a tropical freshwater system (Lake Tanganyika). *Environmental Microbiology* **8**, 1857–1863 (2006).

89. Shade, A. *et al.* Interannual dynamics and phenology of bacterial communities in a eutrophic lake. *Limnology and Oceanography* **52**, 487–494 (2007).

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

91. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

92. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

93. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).

94. The Genome Standards Consortium *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725–731 (2017).

95. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).

96. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).

97. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 1–6 (2016).

98. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

99. Miller, M. A., Pfeiffer, W. & Schwartz, Terri. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. in *Proceedings of the Gateway Computing Environments Workshop* 1–8 (2010).

100. Rohwer, R. R., Hamilton, J. J., Newton, R. J. & McMahon, K. D. TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **3**, (2018).

101. Soo, R. M., Hemp, J., Parks, D. H., Fischer, W. W. & Hugenholtz, P. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. *Science* **355**, 1436–1440 (2017).

102. Linz, A. M. *et al.* Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ* **6**, e6075 (2018).

103. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, (2018).

104. Soo, R. M. *et al.* An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution* **6**, 1031–1045 (2014).

105. Zhou, Z. *et al.* METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022).

106. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95–W101 (2018).

107. Mukherjee, S. *et al.* Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Res* **47**, D649–D659 (2019).

108. Edmond, J. M. *et al.* Nutrient chemistry of the water column of Lake Tanganyika. *Limnology and Oceanography* **38**, 725–738 (1993).

109. Verburga, P. & Hecky, R. E. The physics of the warming of Lake Tanganyika by climate change. *Limnology and Oceanography* **54**, 2418–2430 (2009).

110. Järvinen, M., Salonen, K., Sarvala, J., Vuorio, K. & Virtanen, A. The stoichiometry of particulate nutrients in Lake Tanganyika – implications for nutrient limitation of phytoplankton. *Hydrobiologia* **407**, 81–88 (1999).

111. Ehrenfels, B. *et al.* Thermocline depth and euphotic zone thickness regulate the abundance of diazotrophic cyanobacteria in Lake Tanganyika. *Biogeosciences* (2020) doi:10.5194/bg-2020-214.

112. Tran, P. *et al.* Microbial life under ice: Metagenome diversity and *in situ* activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environmental Microbiology* **20**, 2568–2584 (2018).

113. Martinez-Garcia, M. *et al.* Capturing single cell genomes of active polysaccharide degraders: An unexpected contribution of verrucomicrobia. *PLoS ONE* **7**, 1–11 (2012).

114. Damrow, R., Maldener, I. & Zilliges, Y. The Multiple Functions of Common Microbial Carbon Polymers, Glycogen and PHB, during Stress Responses in the Non-Diazotrophic Cyanobacterium Synechocystis sp. PCC 6803. *Front. Microbiol.* **7**, (2016).

115. Paerl, H. W. & Otten, T. G. Duelling 'CyanoHABs': unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-N2-fixing harmful cyanobacteria. *Environmental Microbiology* **18**, 316–324 (2016).

116. Raymond, J., Siefert, J. L., Staples, C. R. & Blankenship, R. E. The Natural History of Nitrogen Fixation. *Mol Biol Evol* **21**, 541–554 (2004).

117. Berman-Frank, I., Lundgren, P. & Falkowski, P. Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Research in Microbiology* **154**, 157–164 (2003).

118. Cabello-Yeves, P. J. *et al.* Reconstruction of diverse verrucomicrobial genomes from metagenome datasets of freshwater reservoirs. *Frontiers in Microbiology* **8**, (2017).

119. Hansel, C. M., Fendorf, S., Jardine, P. M. & Francis, C. A. Changes in Bacterial and Archaeal Community Structure and Functional Diversity along a Geochemically Variable Soil Profile. *Appl. Environ. Microbiol.* **74**, 1620–1633 (2008).

120. Edlund, A., Hårdeman, F., Jansson, J. K. & Sjöling, S. Active bacterial community structure along vertical redox gradients in Baltic Sea sediment. *Environmental Microbiology* **10**, 2051–2063 (2008).

121. Beman, J. M. & Carolan, M. T. Deoxygenation alters bacterial diversity and community composition in the ocean's largest oxygen minimum zone. *Nature Communications* **4**, 2705 (2013).

122. Bogard, M. J. *et al.* Oxic water column methanogenesis as a major component of aquatic CH4 fluxes. *Nature Communications* **5**, (2014).

123. Vanwonterghem, I. *et al.* Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature Microbiology* **1**, (2016).

124. Gao, Q. *et al.* Wet deposition of atmospheric nitrogen contributes to nitrogen loading in the surface waters of Lake Tanganyika, East Africa: a case study of the Kigoma region. *Environmental Science and Pollution Research* **25**, 11646–11660 (2018).

125. Chale, F. M. M. Inorganic Nutrient Concentrations and Chlorophyll in the Euphotic Zone of Lake Tanganyika. *Hydrobiologia* **523**, 189–197 (2004).

126. Higgins, S. N., Hecky, R. E. & Taylor, W. D. Epilithic nitrogen fixation in the rocky littoral zones of Lake Malawi, Africa. *Limnology and Oceanography* **46**, 976–982 (2001).

127. Brion, N. *et al.* Inorganic Nitrogen Uptake and River Inputs in Northern Lake Tanganyika. *Journal of Great Lakes Research* **32**, 553–564 (2006).

128. Norici, A., Hell, R. & Giordano, M. Sulfur and primary production in aquatic environments: an ecological perspective. *Photosynth Res* **86**, 409–417 (2005).

129. Botz, R. W. & Stoffers, P. Light hydrocarbon gases in Lake Tanganyika hydrothermal fluids (East-Central Africa). *Chemical Geology* **104**, 217–224 (1993).

130. Tiercelin, J.-J. *et al.* Hydrothermal vents in Lake Tanganyika, East African, Rift system. *Geology* **21**, 499–502 (1993).

131. Elsgaard, L. & Prieur, D. Hydrothermal vents in Lake Tanganyika harbor spore-forming thermophiles with extremely rapid growth. *Journal of Great Lakes Research* **37**, 203–206 (2011).

132. Preisler, A. *et al.* Biological and chemical sulfide oxidation in a Beggiatoa inhabited marine sediment. *The ISME Journal* **1**, 341–353 (2007).

133. McAllister, S. M. *et al.* The Fe(II)-oxidizing Zetaproteobacteria: historical, ecological and genomic perspectives. *FEMS Microbiol Ecol* **95**, (2019).

134. Carpenter, S. R. Phosphorus control is critical to mitigating eutrophication. *Proceedings of the National Academy of Sciences* **105**, 11039–11040 (2008).

135. Jr, W. M. L. Causes for the high frequency of nitrogen limitation in tropical lakes. *SIL Proceedings, 1922-2010* **28**, 210–213 (2002).

136. De Keyzer, E. L. R. *et al.* Local perceptions on the state of the pelagic fisheries and fisheries management in Uvira, Lake Tanganyika, DR Congo. *Journal of Great Lakes Research* (2020) doi:10.1016/j.jglr.2020.09.003.

137. MÖLSÄ, Hannu. Management of Fisheries on Lake Tanganyika Challenges for Research and the Community. (University of Kuopio, 2008).

138. Foley, B., Jones, I. D., Maberly, S. C. & Rippey, B. Long-term changes in oxygen depletion in a small temperate lake: effects of climate change and eutrophication. *Freshwater Biology* **57**, 278–289 (2012).

139. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).

140. Jin, M. *et al.* Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 58 (2019).

141. Zimmerman, A. E. *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol* **18**, 21–34 (2020).

142. Okazaki, Y. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. 28.

143. Suttle, S. W. Viruses and Nutrient Cycles in the Sea. **49**, 8 (1999).

144. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol Rep* **4**, (2012).

145. Mara, P. *et al.* Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *ISME J* **14**, 3079–3092 (2020).

146. Coutinho, F. H. *et al.* New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* **8**, 163 (2020).

147. Kieft, K. Metagenomics-Enabled Viral Ecology to Advance Our Understanding of Human and Environmental Microbiomes. (2022).

148. Chen, L.-X. *et al.* Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat Microbiol* **5**, 1504–1515 (2020).

149. Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* **7**, 1738–1751 (2013).

150. Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118–122 (2018).

151. Buchholz, H. H. *et al.* Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. *ISME J* **15**, 1585–1598 (2021).

152. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).

153. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578–585 (2021).

154. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* **5**, e3243 (2017).

155. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **37**, 29–37 (2019).

156. Džunková, M. *et al.* Defining the human gut host–phage network through single-cell viral tagging. *Nat Microbiol* **4**, 2192–2203 (2019).

157. Spencer, S. J. *et al.* Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J* **10**, 427–436 (2016).

158. Sakowski, E. G. *et al.* Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nat Microbiol* (2021) doi:10.1038/s41564-021-00873-4.

159. Louca, S. *et al.* Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *PNAS* **113**, E5925–E5933 (2016).

160. Reed, D. C., Algar, C. K., Huber, J. A. & Dick, G. J. Gene-centric approach to integrating environmental genomics and biogeochemical models. *PNAS* **111**, 1879–1884 (2014).

161. Snortheim, C. A. *et al.* Meteorological drivers of hypolimnetic anoxia in a eutrophic, north temperate lake. *Ecological Modelling* **343**, 39–53 (2017).

162. Schwefel, R., Gaudard, A., Wüest, A. & Bouffard, D. Effects of climate change on deepwater oxygen and winter mixing in a deep lake (Lake Geneva): Comparing observational findings and modeling: CLIMATE CHANGE EFFECTS IN A DEEP LAKE. *Water Resources Research* **52**, 8811–8826 (2016).

163. Jones, D. S. *et al.* Molecular evidence for novel mercury methylating microorganisms in sulfate-impacted lakes. *ISME J* **13**, 1659–1675 (2019).

164. Peterson, B. D. *et al.* Mercury Methylation Genes Identified across Diverse Anaerobic Microbial Guilds in a Eutrophic Sulfate-Enriched Lake. *Environ. Sci. Technol.* **54**, 15840–15851 (2020).

165. Sime-Ngando, T. *et al.* Short-term variations in abundances and potential activities of viruses, bacteria and nanoprotists in Lake Bourget. *Ecol Res* **23**, 851–861 (2008).

166. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nature Microbiology* **3**, 754–766 (2018).

167. Djikeng, A., Kuzmickas, R., Anderson, N. G. & Spiro, D. J. Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**, e7264 (2009).

168. Langlois, V., Girard, C., Vincent, W. F. & Culley, A. I. A Tale of Two Seasons: Distinct Seasonal Viral Communities in a Thermokarst Lake. *Microorganisms* **11**, 428 (2023).

169. Potapov, S. A. *et al.* Metagenomic Analysis of Virioplankton from the Pelagic Zone of Lake Baikal. *Viruses* **11**, (2019).

170. Santos-Medellin, C. *et al.* Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J* **15**, 1956–1970 (2021).

171. Shade, A., Jones, S. E. & McMahon, K. D. The influence of habitat heterogeneity on freshwater bacterial community composition and dynamics. *Environmental Microbiology* **10**, 1057–1067 (2008).

172. Marick, R. A., Peterson, B. D. & McMahon, K. D. *Stratification in Microbial Communities with Depth and Redox Status in a Eutrophic Lake Across Two Years*. http://biorxiv.org/lookup/doi/10.1101/2021.10.15.464574 (2021) doi:10.1101/2021.10.15.464574.

173. Magee, M. R. & Wu, C. H. Response of water temperatures and stratification to changing climate in three lakes with different morphometry. *Hydrology and Earth System Sciences* **21**, 6253–6274 (2017).

174. Brock, T. D. Long-term Change in Lake Mendota. in *A Eutrophic Lake: Lake Mendota, Wisconsin* (ed. Brock, T. D.) 203–215 (Springer, 1985). doi:10.1007/978-1-4419-8700-6_9.

175. Nriagu, J. O. SULFUR METABOLISM AND SEDIMENTARY ENVIRONMENT: LAKE MENDOTA, WISCONSIN. *Limnology and Oceanography* **13**, 430–439 (1968).

176. FALLON, R., HARRITS, S., HANSON, R. & BROCK, T. THE ROLE OF METHANE IN INTERNAL CARBON CYCLING IN LAKE MENDOTA DURING SUMMER STRATIFICATION. *LIMNOLOGY AND OCEANOGRAPHY* **25**, 357–360 (1980).

177. Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front Microbiol* **6**, 469 (2015).

178. Callbeck, C. M. *et al.* Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. *Nat Commun* **9**, 1729 (2018).

179. Muck, S. *et al.* Niche Differentiation of Aerobic and Anaerobic Ammonia Oxidizers in a High Latitude Deep Oxygen Minimum Zone. *Frontiers in Microbiology* **10**, (2019).

180. Kodama, Y. & Watanabe, K. Sulfuricurvum kujiense gen. nov., sp. nov., a facultatively anaerobic, chemolithoautotrophic, sulfur-oxidizing bacterium isolated from an underground crude-oil storage cavity. *Int J Syst Evol Microbiol* **54**, 2297–2300 (2004).

181. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* **40**, 258–272 (2016).

182. Roux, S. *et al.* iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. 2022.07.28.501908 Preprint at https://doi.org/10.1101/2022.07.28.501908 (2022).

183. Callieri, C. *et al.* The mesopelagic anoxic Black Sea as an unexpected habitat for Synechococcus challenges our understanding of global “deep red fluorescence”. *The ISME Journal* **13**, 1676–1687 (2019).

184. Tran, P. Q. *et al.* Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J* **15**, 1971–1986 (2021).

185. Newton, R. J. & McMahon, K. D. Seasonal differences in bacterial community composition following nutrient additions in a eutrophic lake. *Environmental Microbiology* **13**, 887–899 (2011).

186. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181–186 (2008).

187. Koskella, B. & Meaden, S. Understanding Bacteriophage Specificity in Natural Microbial Communities. *Viruses* **5**, 806–823 (2013).

188. Amundson, K. K., Roux, S., Shelton, J. L. & Wilkins, M. J. *Long-term CRISPR array dynamics and stable host-virus co-existence in subsurface fractured shales*. http://biorxiv.org/lookup/doi/10.1101/2023.02.03.526977 (2023) doi:10.1101/2023.02.03.526977.

189. Kavagutti, V. S., Andrei, A.-Ş., Mehrshad, M., Salcher, M. M. & Ghai, R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome* **7**, 135 (2019).

190. Wilhelm, S. W. *et al.* Marine and Freshwater Cyanophages in a Laurentian Great Lake: Evidence from Infectivity Assays and Molecular Analyses of g20 Genes. *Appl Environ Microbiol* **72**, 4957–4963 (2006).

191. Denef, V. J., Mueller, R. S., Chiang, E., Liebig, J. R. & Vanderploeg, H. A. 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 (2015).

192. Mehrshad, M. *et al.* Hidden in plain sight—highly abundant and diverse planktonic freshwater Chloroflexi. *Microbiome* **6**, 176 (2018).

193. Genomic and Seasonal Variations among Aquatic Phages Infecting the Baltic Sea Gammaproteobacterium Rheinheimera sp. Strain BAL341 | Applied and Environmental Microbiology. https://journals.asm.org/doi/10.1128/AEM.01003-19.

194. Cabello-Yeves, P. J. *et al.* Novel Synechococcus Genomes Reconstructed from Freshwater Reservoirs. *Front Microbiol* **8**, 1151 (2017).

195. Chénard, C. & Suttle, C. A. Phylogenetic Diversity of Sequences of Cyanophage Photosynthetic Gene psbA in Marine and Freshwaters. *Applied and Environmental Microbiology* **74**, 5317–5324 (2008).

196. Zhang, M., Zhang, T., Yu, M., Chen, Y.-L. & Jin, M. The Life Cycle Transitions of Temperate Phages: Regulating Factors and Potential Ecological Implications. *Viruses* **14**, 1904 (2022).

197. Williamson, S. J. & Paul, J. H. Environmental factors that influence the transition from lysogenic to lytic existence in the phiHSIC/Listonella pelagia marine phage-host system. *Microb Ecol* **52**, 217–225 (2006).

198. Palermo, C. N., Shea, D. W. & Short, S. M. Analysis of Different Size Fractions Provides a More Complete Perspective of Viral Diversity in a Freshwater Embayment. *Applied and Environmental Microbiology* **87**, e00197-21 (2021).

199. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).

200. Genomic analysis of uncultured marine viral communities | PNAS. https://www.pnas.org/doi/full/10.1073/pnas.202488399.

201. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**, 470–483 (2009).

202. Butina, T. V. *et al.* Extended Evaluation of Viral Diversity in Lake Baikal through Metagenomics. *Microorganisms* **9**, 760 (2021).

203. Gu, C. *et al.* Saline lakes on the Qinghai-Tibet Plateau harbor unique viral assemblages mediating microbial environmental adaption. *iScience* **24**, 103439 (2021).

204. Potapov, S. *et al.* The Viral Fraction Metatranscriptomes of Lake Baikal. *Microorganisms* **10**, 1937 (2022).

205. Roux, S. *et al.* Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* **7**, e33641 (2012).

206. Moon, K., Kim, S., Kang, I. & Cho, J.-C. Viral metagenomes of Lake Soyang, the largest freshwater lake in South Korea. *Sci Data* **7**, 349 (2020).

207. Mohiuddin, M. & Schellhorn, H. E. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**, 960 (2015).

208. Kim, Y., Aw, T. G., Teal, T. K. & Rose, J. B. Metagenomic Investigation of Viral Communities in Ballast Water. *Environ. Sci. Technol.* **49**, 8396–8407 (2015).

209. Hwang, Y., Roux, S., Coclet, C., Krause, S. J. E. & Girguis, P. R. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol* 1–12 (2023) doi:10.1038/s41564-023-01347-5.

210. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490 (2015).

211. Elbehery, A. H. A. & Deng, L. Insights into the global freshwater virome. *Frontiers in Microbiology* **13**, (2022).

212. Chopyk, J. *et al.* Seasonal dynamics in taxonomy and function within bacterial and viral metagenomic assemblages recovered from a freshwater agricultural pond. *Environmental Microbiome* **15**, 18 (2020).

213. Comte, J. *et al.* Microbial community structure and dynamics in the largest natural French lake (Lake Bourget). *Microbial Ecology* **52**, 72–89 (2006).

214. Jaffe, A. L. *et al.* Variable impact of geochemical gradients on the functional potential of bacteria, archaea, and phages from the permanently stratified Lac Pavin. 2022.07.18.500538 Preprint at https://doi.org/10.1101/2022.07.18.500538 (2022).

215. Bibby, K. *et al.* Metagenomics and the development of viral water quality tools. *npj Clean Water* **2**, 1–13 (2019).

216. Langenfeld, K., Chin, K., Roy, A., Wigginton, K. & Duhaime, M. B. Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**, e11111 (2021).

217. John, S., Poulos, B. & Schirmer, C. Iron Chloride Precipitation of Viruses from Seawater. (2015).

218. Linz, A. M., Aylward, F. O., Bertilsson, S. & McMahon, K. D. Time-series metatranscriptomes reveal conserved patterns between phototrophic and heterotrophic microbes in diverse freshwater systems. *Limnology and Oceanography* **65**, S101–S112 (2020).

219. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

220. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41**, D387-395 (2013).

221. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188–7196 (2007).

222. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

223. Woodcroft, B. J. CoverM. (2022).

224. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

225. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

226. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).

227. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361 (2017).

228. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

229. Yin, Y. *et al.* DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **40**, 445–451 (2012).

230. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* **46**, D624–D632 (2018).

231. McDaniel, E. A., Anantharaman, K. & McMahon, K. D. metabolisHMM: Phylogenomic analysis for exploration of microbial phylogenies and metabolic pathways. 2019.12.20.884627 Preprint at https://doi.org/10.1101/2019.12.20.884627 (2020).

232. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. *PLoS ONE* **6**, e27731 (2011).

233. Kieft, K., Zhou, Z. & Anantharaman, K. *VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences*. http://biorxiv.org/lookup/doi/10.1101/855387 (2019) doi:10.1101/855387.

234. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).

235. Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007).

236. Bashenkhaeva, M. V. *et al.* Comparative analysis of free-living and particle-associated bacterial communities of Lake Baikal during the ice-covered period. *Journal of Great Lakes Research* **46**, 508–518 (2020).

237. Butina, T. V., Khanaev, I. V., Kravtsova, L. S., Maikova, O. O. & Bukin, Y. S. Metavirome datasets from two endemic Baikal sponges Baikalospongia bacillifera. *Data Brief* **29**, 105260 (2020).

238. Cabello-Yeves, P. J. *et al.* Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. *Limnology and Oceanography* (2019) doi:10.1002/lno.11401.

239. Potapov, S. *et al.* Characteristics of the viromes in the pelagic zone of Lake Baikal. *LFWB* 1013–1014 (2020) doi:10.31951/2658-3518-2020-A-4-1013.

240. Chiriac, M.-C. *et al.* Ecogenomics sheds light on diverse lifestyle strategies in freshwater CPR. *Microbiome* **10**, 84 (2022).

241. Vigneron, A. *et al.* Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnology and Oceanography Letters* **5**, 212–220 (2020).

242. Paver, S. F., Newton, R. J. & Coleman, M. L. Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environmental Microbiology* **22**, 433–446 (2020).

243. He, C. *et al.* Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat Microbiol* **6**, 354–365 (2021).

244. Chen, Y.-J. *et al.* Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J* **15**, 2986–3004 (2021).

245. Cárdenas, A. *et al.* Greater functional diversity and redundancy of coral endolithic microbiomes align with lower coral bleaching susceptibility. *ISME J* **16**, 2406–2420 (2022).

246. Qian, L. *et al.* Vertically stratified methane, nitrogen and sulphur cycling and coupling mechanisms in mangrove sediment microbiomes. *Microbiome* **11**, 71 (2023).

247. Jones, S. E. External drivers of limnetic bacterial community composition and function. (University of Wisconsin - Madison, 2008).

248. Shade, Ashley.

AQUATIC BACTERIA, LAKES, AND WATER COLUMN OVERTURN: A MODEL MI

CROBIAL SYSTEM FOR DISTURBANCE ECOLOGY. (University of Wisconsin - Madison, 2010).

249. Comparative Studies of a Suite of Lakes in Wisconsin. (2020).
250. Magee, M. R., McIntyre, P. B., Hanson, P. C. & Wu, C. H. Drivers and Management Implications of Long-Term Cisco Oxythermal Habitat Decline in Lake Mendota, WI. *Environmental Management* **63**, 396–407 (2019).
251. Carey, C. C. *et al.* Anoxia decreases the magnitude of the carbon, nitrogen, and phosphorus sink in freshwaters. *Global Change Biology* **28**, 4861–4881 (2022).
252. Bhatnagar, S. *et al.* Microbial community dynamics and coexistence in a sulfide-driven phototrophic bloom. *Environmental Microbiome* **15**, 3 (2020).
253. Cuddington, K. *et al.* Process-based models are required to manage ecological systems in a changing world. *Ecosphere* **4**, art20 (2013).
254. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North Temperate Lakes LTER: Chemical Limnology of Primary Study Lakes: Nutrients, pH and Carbon 1981 - current ver 52.
255. Hale, R. J. CHARACTERIZATION OF THE CYANOBACTERIAL COMMUNITY AND CYANOTOXINS IN LAKE MENDOTA ACROSS TIME, SPACE, AND THE WATER COLUMN. (University of Wisconsin - Madison, 2022).
256. Peterson, Benjamin D. Ecophysiology of mercury-methylating microorganisms in freshwater ecosystems. (University of Wisconsin - Madison, 2021).
257. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *The ISME Journal* 1 (2018) doi:10.1038/s41396-018-0289-4.

258. Coutinho, F. H. *et al.* Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**, 15955 (2017).

259. Santiago-Rodriguez, T. M. & Hollister, E. B. Unraveling the viral dark matter through viral metagenomics. *Frontiers in Immunology* **13**, (2022).

260. Johansen, J. *et al.* Genome binning of viral entities from bulk metagenomics data. *Nat Commun* **13**, 965 (2022).

261. Pietilä, M. K., Demina, T. A., Atanasova, N. S., Oksanen, H. M. & Bamford, D. H. Archaeal viruses and bacteriophages: comparisons and contrasts. *Trends Microbiol* **22**, 334–344 (2014).

262. Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e18 (2022).

263. Wastewater monitoring comes of age. *Nat Microbiol* **7**, 1101–1102 (2022).

264. How Wastewater Surveillance Works | National Wastewater Surveillance System | CDC. https://www.cdc.gov/nwss/how-wws-works.html (2023).

265. COVID-19 wastewater | Wisconsin State Laboratory of Hygiene. https://www.slh.wisc.edu/environmental/covid-19-wastewater/.

Appendix

Filter ID	Sample Name	Biol rep	Tech rep	Volume filtered (mL)	Date (m/d/y)	Depth (meters)
MEWINTER1801	ME02Feb2018D5	A	1	250	2/2/18	5
MEWINTER1802	ME02Feb2018D5	A	2	250	2/2/18	5
MEWINTER1803	ME02Feb2018D10	A	1	250	2/2/18	10
MEWINTER1804	ME02Feb2018D10	A	2	250	2/2/18	10
MEWINTER1805	ME02Feb2018D15	A	1	250	2/2/18	15
MEWINTER1806	ME02Feb2018D15	A	2	250	2/2/18	15
MEWINTER1807	ME02Feb2018D20	A	1	250	2/2/18	20
MEWINTER1808	ME02Feb2018D20	A	2	250	2/2/18	20
MEWINTER1809	ME02Feb2018D22	A	1	250	2/2/18	22
MEWINTER1810	ME02Feb2018D22	A	2	250	2/2/18	22
MEWINTER1811	ME16Feb2018D5	A	1	250	2/16/18	5
MEWINTER1812	ME16Feb2018D5	A	2	250	2/16/18	5
MEWINTER1813	ME16Feb2018D10	A	1	250	2/16/18	10
MEWINTER1814	ME16Feb2018D10	A	2	250	2/16/18	10
MEWINTER1815	ME16Feb2018D15	A	1	250	2/16/18	15
MEWINTER1816	ME16Feb2018D15	A	2	250	2/16/18	15
MEWINTER1817	ME16Feb2018D20	A	1	250	2/16/18	20
MEWINTER1818	ME16Feb2018D20	A	2	250	2/16/18	20

MEWINTER1819	ME16Feb2018D22	A	1	250	2/16/18	22
MEWINTER1820	ME16Feb2018D22	A	2	250	2/16/18	22
MEWINTER1821	ME16Feb2018D24	A	1	250	2/16/18	24
MEWINTER1822	ME16Feb2018D24	A	2	250	2/16/18	24
MEWINTER1823	ME23Feb2018D5	A	1	250	2/23/18	5
MEWINTER1824	ME23Feb2018D5	A	2	250	2/23/18	5
MEWINTER1825	ME23Feb2018D10	A	1	250	2/23/18	10
MEWINTER1826	ME23Feb2018D10	A	2	250	2/23/18	10
MEWINTER1827	ME23Feb2018D15	A	1	250	2/23/18	15
MEWINTER1828	ME23Feb2018D15	A	2	250	2/23/18	15
MEWINTER1829	ME23Feb2018D20	A	1	250	2/23/18	20
MEWINTER1830	ME23Feb2018D20	A	2	250	2/23/18	20
MEWINTER1831	ME23Feb2018D22	A	1	250	2/23/18	22
MEWINTER1832	ME23Feb2018D22	A	2	250	2/23/18	22
MEWINTER1833	ME23Feb2018D23	A	1	250	2/23/18	23
MEWINTER1834	ME23Feb2018D23	A	2	250	2/23/18	23
MEWINTER1835	ME26Feb2018D5	A	1	250	2/26/18	5
MEWINTER1836	ME26Feb2018D5	A	2	250	2/26/18	5
MEWINTER1837	ME26Feb2018D10	A	1	250	2/26/18	10
MEWINTER1838	ME26Feb2018D10	A	2	250	2/26/18	10
MEWINTER1839	ME26Feb2018D15	A	1	250	2/26/18	15
MEWINTER1840	ME26Feb2018D15	A	2	250	2/26/18	15
MEWINTER1841	ME26Feb2018D20	A	1	250	2/26/18	20
MEWINTER1842	ME26Feb2018D20	A	2	250	2/26/18	20

MEWINTER1843	ME26Feb2018D22	A	1	250	2/26/18	22
MEWINTER1844	ME26Feb2018D22	A	2	250	2/26/18	22
MEWINTER1845	ME26Feb2018D24	A	1	250	2/26/18	24
MEWINTER1846	ME26Feb2018D24	A	2	250	2/26/18	24
MEWINTER1847	ME11Apr2018D0	A	1	250	4/11/18	0
MEWINTER1848	ME11Apr2018D0	A	2	250	4/11/18	0
MEWINTER1849	ME11Apr2018D5	A	1	250	4/11/18	5
MEWINTER1850	ME11Apr2018D5	A	2	250	4/11/18	5
MEWINTER1851	ME11Apr2018D10	A	1	250	4/11/18	10
MEWINTER1852	ME11Apr2018D10	A	2	250	4/11/18	10
MEWINTER1853	ME11Apr2018D15	A	1	250	4/11/18	15
MEWINTER1854	ME11Apr2018D15	A	2	250	4/11/18	15
MEWINTER1855	ME11Apr2018D20	A	1	250	4/11/18	20
MEWINTER1856	ME11Apr2018D20	A	2	250	4/11/18	20
MEWINTER1857	ME11Apr2018D24	A	1	250	4/11/18	24
MEWINTER1858	ME11Apr2018D24	A	2	250	4/11/18	24
MEWINTER1859	ME20Apr2018D0	A	1	250	4/20/18	0
MEWINTER1860	ME20Apr2018D0	A	2	250	4/20/18	0
MEWINTER1861	ME20Apr2018D5	A	1	250	4/20/18	5
MEWINTER1862	ME20Apr2018D5	A	2	250	4/20/18	5
MEWINTER1863	ME20Apr2018D10	A	1	250	4/20/18	10
MEWINTER1864	ME20Apr2018D10	A	2	250	4/20/18	10
MEWINTER1865	ME20Apr2018D15	A	1	250	4/20/18	15
MEWINTER1866	ME20Apr2018D15	A	2	250	4/20/18	15

MEWINTER1867	ME20Apr2018D20	A	1	250	4/20/18	20
MEWINTER1868	ME20Apr2018D20	A	2	250	4/20/18	20
MEWINTER1869	ME20Apr2018D24	A	1	250	4/20/18	24
MEWINTER1870	ME20Apr2018D24	A	2	250	4/20/18	24
MEWINTER1923	ME01Mar2019D0	A	1	250	1/3/19	0
MEWINTER1924	ME01Mar2019D0	A	2	250	1/3/19	0
MEWINTER1925	ME01Mar2019D5	A	1	250	1/3/19	5
MEWINTER1926	ME01Mar2019D5	A	2	250	1/3/19	5
MEWINTER1927	ME01Mar2019D10	A	1	250	1/3/19	10
MEWINTER1928	ME01Mar2019D10	A	2	250	1/3/19	10
MEWINTER1929	ME01Mar2019D15	A	1	250	1/3/19	15
MEWINTER1930	ME01Mar2019D15	A	2	250	1/3/19	15
MEWINTER1931	ME01Mar2019D20	A	1	250	1/3/19	20
MEWINTER1932	ME01Mar2019D20	A	2	250	1/3/19	20
MEWINTER1933	ME01Mar2019D23	A	1	250	1/3/19	23
MEWINTER1934	ME01Mar2019D23	A	2	250	1/3/19	23
MEWINTER1901	ME15Feb2019D0	A	1	250	2/15/19	0
MEWINTER1902	ME15Feb2019D0	A	2	250	2/15/19	0
MEWINTER1903	ME15Feb2019D5	A	1	250	2/15/19	5
MEWINTER1904	ME15Feb2019D5	A	2	250	2/15/19	5
MEWINTER1905	ME15Feb2019D10	A	1	250	2/15/19	10
MEWINTER1906	ME15Feb2019D10	A	2	250	2/15/19	10
MEWINTER1907	ME15Feb2019D15	A	1	250	2/15/19	15
MEWINTER1908	ME15Feb2019D15	A	2	250	2/15/19	15

MEWINTER1909	ME15Feb2019D20	A	1	250	2/15/19	20
MEWINTER1910	ME15Feb2019D20	A	2	250	2/15/19	20
MEWINTER1921	ME15Feb2019D23	A	1	250	2/15/19	23
MEWINTER1922	ME15Feb2019D23	A	2	250	2/15/19	23
MEWINTER1911	ME22Feb2019D0	A	1	250	2/22/19	0
MEWINTER1912	ME22Feb2019D0	A	2	250	2/22/19	0
MEWINTER1913	ME22Feb2019D5	A	1	250	2/22/19	5
MEWINTER1914	ME22Feb2019D5	A	2	250	2/22/19	5
MEWINTER1915	ME22Feb2019D10	A	1	250	2/22/19	10
MEWINTER1916	ME22Feb2019D10	A	2	250	2/22/19	10
MEWINTER1917	ME22Feb2019D15	A	1	250	2/22/19	15
MEWINTER1918	ME22Feb2019D15	A	2	250	2/22/19	15
MEWINTER1919	ME22Feb2019D20	A	1	250	2/22/19	20
MEWINTER1920	ME22Feb2019D20	A	2	250	2/22/19	20
ME_2020-07-08_5m_Bacteria_A	ME_2020-07-08_5	1	A	500	7/8/20	5
ME_2020-07-08_10m_Bacteria_A	ME_2020-07-08_10	1	A	500	7/8/20	10
ME_2020-07-08_15m_Bacteria_A	ME_2020-07-08_15	1	A	500	7/8/20	15
ME_2020-07-08_20m_Bacteria_A	ME_2020-07-08_20	1	A	500	7/8/20	20
ME_2020-07-08_23.5m_Bacteria_A	ME_2020-07-08_23.5	1	A	500	7/8/20	23.5

ME_2020-07-16_5m_Bacteria_A	ME_2020-07-16_5	1	A	500	7/16/20	5
ME_2020-07-16_5m_Bacteria_B	ME_2020-07-16_5	1	B	500	7/16/20	5
ME_2020-07-16_10m_Bacteria_A	ME_2020-07-16_10	1	A	500	7/16/20	10
ME_2020-07-16_10m_Bacteria_B	ME_2020-07-16_10	1	B	500	7/16/20	10
ME_2020-07-16_15m_Bacteria_A	ME_2020-07-16_15	1	A	500	7/16/20	15
ME_2020-07-16_15m_Bacteria_B	ME_2020-07-16_15	1	B	500	7/16/20	15
ME_2020-07-16_20m_Bacteria_A	ME_2020-07-16_20	1	A	500	7/16/20	20
ME_2020-07-16_20m_Bacteria_B	ME_2020-07-16_20	1	B	500	7/16/20	20
ME_2020-07-16_23.5m_Bacteria_A	ME_2020-07-16_23.5	1	A	500	7/16/20	23.5
ME_2020-07-16_23.5m_Bacteria_B	ME_2020-07-16_23.5	1	B	500	7/16/20	23.5
ME_2020-07-24_5m_Bacteria_A	ME_2020-07-24_5	1	A	250	7/24/20	5
ME_2020-07-24_5m_Bacteria_B	ME_2020-07-24_5	1	B	250	7/24/20	5

ME_2020-07-24_5m_Bacteria_C	ME_2020-07-24_5	1	C	250	7/24/20	5
ME_2020-07-24_5m_Bacteria_D	ME_2020-07-24_5	1	D	250	7/24/20	5
ME_2020-07-24_10m_Bacteria_A	ME_2020-07-24_10	1	A	500	7/24/20	10
ME_2020-07-24_10m_Bacteria_B	ME_2020-07-24_10	1	B	500	7/24/20	10
ME_2020-07-24_15m_Bacteria_A	ME_2020-07-24_15	1	A	500	7/24/20	15
ME_2020-07-24_15m_Bacteria_B	ME_2020-07-24_15	1	B	500	7/24/20	15
ME_2020-07-24_20m_Bacteria_A	ME_2020-07-24_20	1	A	500	7/24/20	20
ME_2020-07-24_20m_Bacteria_B	ME_2020-07-24_20	1	B	500	7/24/20	20
ME_2020-07-24_23.5m_Bacteria_A	ME_2020-07-24_23.5	1	A	500	7/24/20	23.5
ME_2020-07-24_23.5m_Bacteria_B	ME_2020-07-24_23.5	1	B	500	7/24/20	23.5
ME_2020-07-29_5m_Bacteria_A	ME_2020-07-29_5	1	A	250	7/29/20	5
ME_2020-07-29_5m_Bacteria_B	ME_2020-07-29_5	1	B	250	7/29/20	5

ME_2020-07-29_5m_Bacteria_C	ME_2020-07-29_5	1	C	250	7/29/20	5
ME_2020-07-29_5m_Bacteria_D	ME_2020-07-29_5	1	D	250	7/29/20	5
ME_2020-07-29_10m_Bacteria_A	ME_2020-07-29_10	1	A	500	7/29/20	10
ME_2020-07-29_10m_Bacteria_B	ME_2020-07-29_10	1	B	500	7/29/20	10
ME_2020-07-29_15m_Bacteria_A	ME_2020-07-29_15	1	A	500	7/29/20	15
ME_2020-07-29_15m_Bacteria_B	ME_2020-07-29_15	1	B	500	7/29/20	15
ME_2020-07-29_20m_Bacteria_A	ME_2020-07-29_20	1	A	500	7/29/20	20
ME_2020-07-29_20m_Bacteria_B	ME_2020-07-29_20	1	B	500	7/29/20	20
ME_2020-07-29_23.5m_Bacteria_A	ME_2020-07-29_23.5	1	A	500	7/29/20	23.5
ME_2020-07-29_23.5m_Bacteria_B	ME_2020-07-29_23.5	1	B	500	7/29/20	23.5
ME_2020-08-05_5m_Bacteria_A	ME_2020-08-05_5	1	A	250	8/5/20	5
ME_2020-08-05_5m_Bacteria_B	ME_2020-08-05_5	1	B	250	8/5/20	5

ME_2020-08-05_5m_Bacteria_C	ME_2020-08-05_5	1	C	250	8/5/20	5
ME_2020-08-05_5m_Bacteria_D	ME_2020-08-05_5	1	D	250	8/5/20	5
ME_2020-08-05_10m_Bacteria_A	ME_2020-08-05_10	1	A	500	8/5/20	10
ME_2020-08-05_10m_Bacteria_B	ME_2020-08-05_10	1	B	500	8/5/20	10
ME_2020-08-05_15m_Bacteria_A	ME_2020-08-05_15	1	A	500	8/5/20	15
ME_2020-08-05_15m_Bacteria_B	ME_2020-08-05_15	1	B	500	8/5/20	15
ME_2020-08-05_20m_Bacteria_A	ME_2020-08-05_20	1	A	500	8/5/20	20
ME_2020-08-05_20m_Bacteria_B	ME_2020-08-05_20	1	B	500	8/5/20	20
ME_2020-08-05_23.5m_Bacteria_A	ME_2020-08-05_23.5	1	A	500	8/5/20	23.5
ME_2020-08-05_23.5m_Bacteria_B	ME_2020-08-05_23.5	1	B	500	8/5/20	23.5
ME_2020-08-12_5m_Bacteria_A	ME_2020-08-12_5	1	A	250	8/12/20	5
ME_2020-08-12_5m_Bacteria_B	ME_2020-08-12_5	1	B	250	8/12/20	5

ME_2020-08-12_5m_Bacteria_C	ME_2020-08-12_5	1	C	250	8/12/20	5
ME_2020-08-12_5m_Bacteria_D	ME_2020-08-12_5	1	D	250	8/12/20	5
ME_2020-08-12_10m_Bacteria_A	ME_2020-08-12_10	1	A	500	8/12/20	10
ME_2020-08-12_10m_Bacteria_B	ME_2020-08-12_10	1	B	500	8/12/20	10
ME_2020-08-12_15m_Bacteria_A	ME_2020-08-12_15	1	A	500	8/12/20	15
ME_2020-08-12_15m_Bacteria_B	ME_2020-08-12_15	1	B	500	8/12/20	15
ME_2020-08-12_20m_Bacteria_A	ME_2020-08-12_20	1	A	500	8/12/20	20
ME_2020-08-12_20m_Bacteria_B	ME_2020-08-12_20	1	B	500	8/12/20	20
ME_2020-08-12_23.5m_Bacteria_A	ME_2020-08-12_23.5	1	A	500	8/12/20	23.5
ME_2020-08-12_23.5m_Bacteria_B	ME_2020-08-12_23.5	1	B	450	8/12/20	23.5
ME_2020-08-18_5m_Bacteria_A	ME_2020-08-18_5	1	A	500	8/18/20	5
ME_2020-08-18_5m_Bacteria_B	ME_2020-08-18_5	1	B	500	8/18/20	5

ME_2020-08-18_10m_Bacteria_A	ME_2020-08-18_10	1	A	500	8/18/20	10
ME_2020-08-18_10m_Bacteria_B	ME_2020-08-18_10	1	B	500	8/18/20	10
ME_2020-08-18_15m_Bacteria_A	ME_2020-08-18_15	1	A	500	8/18/20	15
ME_2020-08-18_15m_Bacteria_B	ME_2020-08-18_15	1	B	500	8/18/20	15
ME_2020-08-18_20m_Bacteria_A	ME_2020-08-18_20	1	A	500	8/18/20	20
ME_2020-08-18_20m_Bacteria_B	ME_2020-08-18_20	1	B	500	8/18/20	20
ME_2020-08-18_23.5m_Bacteria_A	ME_2020-08-18_23.5	1	A	500	8/18/20	23.5
ME_2020-08-18_23.5m_Bacteria_B	ME_2020-08-18_23.5	1	B	350	8/18/20	23.5
ME_2020-08-25_5m_Bacteria_A	ME_2020-08-25_5	1	A	500	8/25/20	5
ME_2020-08-25_5m_Bacteria_B	ME_2020-08-25_5	1	B	350	8/25/20	5
ME_2020-08-25_10m_Bacteria_A	ME_2020-08-25_10	1	A	500	8/25/20	10
ME_2020-08-25_10m_Bacteria_B	ME_2020-08-25_10	1	B	500	8/25/20	10

ME_2020-08-25_15m_Bacteria_A	ME_2020-08-25_15	1	A	500	8/25/20	15
ME_2020-08-25_15m_Bacteria_B	ME_2020-08-25_15	1	B	500	8/25/20	15
ME_2020-08-25_20m_Bacteria_A	ME_2020-08-25_20	1	A	500	8/25/20	20
ME_2020-08-25_20m_Bacteria_B	ME_2020-08-25_20	1	B	500	8/25/20	20
ME_2020-08-25_23.5m_Bacteria_A	ME_2020-08-25_23.5	1	A	500	8/25/20	23.5
ME_2020-08-25_23.5m_Bacteria_B	ME_2020-08-25_23.5	1	B	350	8/25/20	23.5
ME_2020-09-01_5m_Bacteria_A	ME_2020-09-01_5	1	A	500	9/1/20	5
ME_2020-09-01_5m_Bacteria_B	ME_2020-09-01_5	1	B	500	9/1/20	5
ME_2020-09-01_10m_Bacteria_A	ME_2020-09-01_10	1	A	500	9/1/20	10
ME_2020-09-01_10m_Bacteria_B	ME_2020-09-01_10	1	B	350	9/1/20	10
ME_2020-09-01_15m_Bacteria_A	ME_2020-09-01_15	1	A	500	9/1/20	15
ME_2020-09-01_15m_Bacteria_B	ME_2020-09-01_15	1	B	500	9/1/20	15

ME_2020-09-01_20m_Bacteria_A	ME_2020-09-01_20	1	A	500	9/1/20	20
ME_2020-09-01_20m_Bacteria_B	ME_2020-09-01_20	1	B	300	9/1/20	20
ME_2020-09-01_23.5m_Bacteria_A	ME_2020-09-01_23.5	1	A	500	9/1/20	23.5
ME_2020-09-01_23.5m_Bacteria_B	ME_2020-09-01_23.5	1	B	350	9/1/20	23.5
ME_2020-09-11_5m_Bacteria_A	ME_2020-09-11_5	1	A	450	9/11/20	5
ME_2020-09-11_5m_Bacteria_B	ME_2020-09-11_5	1	B	400	9/11/20	5
ME_2020-09-11_10m_Bacteria_A	ME_2020-09-11_10	1	A	450	9/11/20	10
ME_2020-09-11_10m_Bacteria_B	ME_2020-09-11_10	1	B	350	9/11/20	10
ME_2020-09-11_15m_Bacteria_A	ME_2020-09-11_15	1	A	500	9/11/20	15
ME_2020-09-11_15m_Bacteria_B	ME_2020-09-11_15	1	B	650	9/11/20	15
ME_2020-09-11_20m_Bacteria_A	ME_2020-09-11_20	1	A	400	9/11/20	20
ME_2020-09-11_20m_Bacteria_B	ME_2020-09-11_20	1	B	400	9/11/20	20

ME_2020-09-11_23.5m_Bacteria_A	ME_2020-09-11_23.5	1	A	400	9/11/20	23.5
ME_2020-09-11_23.5m_Bacteria_B	ME_2020-09-11_23.5	1	B	400	9/11/20	23.5
ME_2020-09-17_5m_Bacteria_A	ME_2020-09-17_5	1	A	250	9/17/20	5
ME_2020-09-17_5m_Bacteria_B	ME_2020-09-17_5	1	B	250	9/17/20	5
ME_2020-09-17_10m_Bacteria_A	ME_2020-09-17_10	1	A	350	9/17/20	10
ME_2020-09-17_10m_Bacteria_B	ME_2020-09-17_10	1	B	350	9/17/20	10
ME_2020-09-17_15m_Bacteria_A	ME_2020-09-17_15	1	A	400	9/17/20	15
ME_2020-09-17_15m_Bacteria_B	ME_2020-09-17_15	1	B	400	9/17/20	15
ME_2020-09-17_20m_Bacteria_A	ME_2020-09-17_20	1	A	400	9/17/20	20
ME_2020-09-17_20m_Bacteria_B	ME_2020-09-17_20	1	B	350	9/17/20	20
ME_2020-09-17_23.5m_Bacteria_A	ME_2020-09-17_23.5	1	A	500	9/17/20	23.5
ME_2020-09-17_23.5m_Bacteria_B	ME_2020-09-17_23.5	1	B	400	9/17/20	23.5

ME_2020-09-23_5m_Bacteria_A	ME_2020-09-23_5	1	A	350	9/23/20	5
ME_2020-09-23_5m_Bacteria_B	ME_2020-09-23_5	1	B	350	9/23/20	5
ME_2020-09-23_10m_Bacteria_A	ME_2020-09-23_10	1	A	400	9/23/20	10
ME_2020-09-23_10m_Bacteria_B	ME_2020-09-23_10	1	B	400	9/23/20	10
ME_2020-09-23_15m_Bacteria_A	ME_2020-09-23_15	1	A	400	9/23/20	15
ME_2020-09-23_15m_Bacteria_B	ME_2020-09-23_15	1	B	500	9/23/20	15
ME_2020-09-23_20m_Bacteria_A	ME_2020-09-23_20	1	A	350	9/23/20	20
ME_2020-09-23_20m_Bacteria_B	ME_2020-09-23_20	1	B	400	9/23/20	20
ME_2020-09-23_23.5m_Bacteria_A	ME_2020-09-23_23.5	1	A	300	9/23/20	23.5
ME_2020-09-23_23.5m_Bacteria_B	ME_2020-09-23_23.5	1	B	350	9/23/20	23.5
ME_2020-10-02_5m_Bacteria_A	ME_2020-10-02_5	1	A	450	10/2/20	5
ME_2020-10-02_5m_Bacteria_B	ME_2020-10-02_5	1	B	450	10/2/20	5

ME_2020-10-02_10m_Bacteria_A	ME_2020-10-02_10	1	A	450	10/2/20	10
ME_2020-10-02_10m_Bacteria_B	ME_2020-10-02_10	1	B	450	10/2/20	10
ME_2020-10-02_15m_Bacteria_A	ME_2020-10-02_15	1	A	450	10/2/20	15
ME_2020-10-02_15m_Bacteria_B	ME_2020-10-02_15	1	B	450	10/2/20	15
ME_2020-10-02_20m_Bacteria_A	ME_2020-10-02_20	1	A	450	10/2/20	20
ME_2020-10-02_20m_Bacteria_B	ME_2020-10-02_20	1	B	450	10/2/20	20
ME_2020-10-02_23.5m_Bacteria_A	ME_2020-10-02_23.5	1	A	450	10/2/20	23.5
ME_2020-10-02_23.5m_Bacteria_B	ME_2020-10-02_23.5	1	B	450	10/2/20	23.5
ME_2020-10-08_5m_Bacteria_A	ME_2020-10-08_5	1	A	350	10/8/20	5
ME_2020-10-08_5m_Bacteria_B	ME_2020-10-08_5	1	B	350	10/8/20	5
ME_2020-10-08_10m_Bacteria_A	ME_2020-10-08_10	1	A	350	10/8/20	10
ME_2020-10-08_10m_Bacteria_B	ME_2020-10-08_10	1	B	350	10/8/20	10

ME_2020-10-08_15m_Bacteria_A	ME_2020-10-08_15	1	A	350	10/8/20	15
ME_2020-10-08_15m_Bacteria_B	ME_2020-10-08_15	1	B	350	10/8/20	15
ME_2020-10-08_20m_Bacteria_A	ME_2020-10-08_20	1	A	250	10/8/20	20
ME_2020-10-08_20m_Bacteria_B	ME_2020-10-08_20	1	B	250	10/8/20	20
ME_2020-10-08_20m_Bacteria_C	ME_2020-10-08_20	1	C	250	10/8/20	20
ME_2020-10-08_23.5m_Bacteria_A	ME_2020-10-08_23.5	1	A	250	10/8/20	23.5
ME_2020-10-08_23.5m_Bacteria_B	ME_2020-10-08_23.5	1	B	250	10/8/20	23.5
ME_2020-10-08_23.5m_Bacteria_C	ME_2020-10-08_23.5	1	C	250	10/8/20	23.5
ME_2020-10-19_5m_Bacteria_A	ME_2020-10-19_5	1	A	500	10/19/20	5
ME_2020-10-19_5m_Bacteria_B	ME_2020-10-19_5	1	B	500	10/19/20	5
ME_2020-10-19_10m_Bacteria_A	ME_2020-10-19_10	1	A	550	10/19/20	10
ME_2020-10-19_10m_Bacteria_B	ME_2020-10-19_10	1	B	500	10/19/20	10

ME_2020-10-19_15m_Bacteria_A	ME_2020-10-19_15	1	A	550	10/19/20	15
ME_2020-10-19_15m_Bacteria_B	ME_2020-10-19_15	1	B	500	10/19/20	15
ME_2020-10-19_20m_Bacteria_A	ME_2020-10-19_20	1	A	550	10/19/20	20
ME_2020-10-19_20m_Bacteria_B	ME_2020-10-19_20	1	B	500	10/19/20	20
ME_2020-10-19_23.5m_Bacteria_A	ME_2020-10-19_23.5	1	A	400	10/19/20	23.5
ME_2020-10-19_23.5m_Bacteria_B	ME_2020-10-19_23.5	1	B	500	10/19/20	23.5
ME_2021-6-02-5A	ME_2021-6-02-5m	1	A	250	6/2/21	5
ME_2021-6-02-5B	ME_2021-6-02-5m	1	B	250	6/2/21	5
ME_2021-06-02-10A	ME_2021-06-02-10m	1	A	250	6/2/21	10
ME_2021-6-02-10B	ME_2021-6-02-10m	1	B	250	6/2/21	10
ME_2021-6-02-15A	ME_2021-6-02-15m	1	A	250	6/2/21	15
ME_2021-6-02-15B	ME_2021-6-02-15m	1	B	250	6/2/21	15
ME_2021-6-02-20A	ME_2021-6-02-20m	1	A	250	6/2/21	20

ME_2021-6-02-20B	ME_2021-6-02-20m	1	B	250	6/2/21	20
ME_2021-6-02-23.5A	ME_2021-6-02-23.5m	1	A	250	6/2/21	23.5
ME_2021-6-02-23.5B	ME_2021-6-02-23.5m	1	B	250	6/2/21	23.5
ME_2021-06-11-5A	ME_2021-06-11-5m	1	A	400	6/11/21	5
ME_2021-06-11-5B	ME_2021-06-11-5m	1	B	400	6/11/21	5
ME_2021-06-11-10A	ME_2021-06-11-10m	1	A	400	6/11/21	10
ME_2021-06-11-10B	ME_2021-06-11-10m	1	B	400	6/11/21	10
ME_2021-06-11-15A	ME_2021-06-11-15m	1	A	400	6/11/21	15
ME_2021-06-11-15B	ME_2021-06-11-15m	1	B	400	6/11/21	15
ME_2021-06-11-20A	ME_2021-06-11-20m	1	A	400	6/11/21	20
ME_2021-06-11-20B	ME_2021-06-11-20m	1	B	400	6/11/21	20
ME_2021-06-11-23.5A	ME_2021-06-11-23.5m	1	A	400	6/11/21	23.5

ME_2021-06-11-23.5B	ME_2021-06-11-23.5m	1	B	400	6/11/21	23.5
ME_2021-06-18-5m_Bacteria_A	ME_2021-06-18-5m	1	A	400	6/18/21	5
ME_2021-06-18-5m_Bacteria_B	ME_2021-06-18-5m	1	B	400	6/18/21	5
ME_2021-06-18-10m_Bacteria_A	ME_2021-06-18-10m	1	A	400	6/18/21	10
ME_2021-06-18-10m_Bacteria_B	ME_2021-06-18-10m	1	B	400	6/18/21	10
ME_2021-06-18-15m_Bacteria_A	ME_2021-06-18-15m	1	A	400	6/18/21	15
ME_2021-06-18-15m_Bacteria_B	ME_2021-06-18-15m	1	B	400	6/18/21	15
ME_2021-06-18-20m_Bacteria_A	ME_2021-06-18-20m	1	A	400	6/18/21	20
ME_2021-06-18-20m_Bacteria_B	ME_2021-06-18-20m	1	B	400	6/18/21	20
ME_2021-07-2-5m_Bacteria_A	ME_2021-07-2-5m	1	A	350	7/2/21	5
ME_2021-07-2-5m_Bacteria_B	ME_2021-07-2-5m	1	B	350	7/2/21	5
ME_2021-07-2-10m_Bacteria_A	ME_2021-07-2-10m	1	A	350	7/2/21	10

ME_2021-07-2-10m_Bacteria_B	ME_2021-07-2-10m	1	B	350	7/2/21	10
ME_2021-07-2-15m_Bacteria_A	ME_2021-07-2-15m	1	A	350	7/2/21	15
ME_2021-07-2-15m_Bacteria_B	ME_2021-07-2-15m	1	B	350	7/2/21	15
ME_2021-07-2-20m_Bacteria_A	ME_2021-07-2-20m	1	A	350	7/2/21	20
ME_2021-07-2-20m_Bacteria_B	ME_2021-07-2-20m	1	B	350	7/2/21	20
ME_2021-07-13-5m_Bacteria_A	ME_2021-07-13-5m	1	A	400	7/13/21	5
ME_2021-07-13-5m_Bacteria_B	ME_2021-07-13-5m	1	B	400	7/13/21	5
ME_2021-07-13-10m_Bacteria_A	ME_2021-07-13-10m	1	A	400	7/13/21	10
ME_2021-07-13-10m_Bacteria_B	ME_2021-07-13-10m	1	B	400	7/13/21	10
ME_2021-07-13-15m_Bacteria_A	ME_2021-07-13-15m	1	A	400	7/13/21	15
ME_2021-07-13-15m_Bacteria_B	ME_2021-07-13-15m	1	B	400	7/13/21	15
ME_2021-07-13-20m_Bacteria_A	ME_2021-07-13-20m	1	A	400	7/13/21	20

ME_2021-07-13-20m_Bacteria_B	ME_2021-07-13-20m	1	B	400	7/13/21	20
ME_2021-07-29-5m_Bacteria_A	ME_2021-07-29-5m	1	A	400	7/29/21	5
ME_2021-07-29-5m_Bacteria_B	ME_2021-07-29-5m	1	B	400	7/29/21	5
ME_2021-07-29-10m_Bacteria_A	ME_2021-07-29-10m	1	A	400	7/29/21	10
ME_2021-07-29-10m_Bacteria_B	ME_2021-07-29-10m	1	B	400	7/29/21	10
ME_2021-07-29-15m_Bacteria_A	ME_2021-07-29-15m	1	A	400	7/29/21	15
ME_2021-07-29-15m_Bacteria_B	ME_2021-07-29-15m	1	B	400	7/29/21	15
ME_2021-07-29-20m_Bacteria_A	ME_2021-07-29-20m	1	A	400	7/29/21	20
ME_2021-07-29-20m_Bacteria_B	ME_2021-07-29-20m	1	B	400	7/29/21	20

Appendix 1 Samples collected in Lake Mendota from 2018-2021 as part of this dissertation.