MICROBIAL COMMUNITY DYNAMICS OF A *MICROCYSTIS* BLOOM

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Helena L. Pound
December 2021

DEDICATION

To my brilliant mentor,

my loving parents,

my inspiring husband,

and

my bastardly cat.

ACKNOWLEDGEMENTS

I must acknowledge Dr. Bradley Fenwick, who was willing to dedicate an hour of his time to aimless college freshman and provide the advice that changed my life, “You should meet with a professor named Dr. Steven Wilhelm.”

Dr. Wilhelm is responsible for the researcher I am today. A talented scientist and a devoted mentor, Dr. Wilhelm taught me how to think critically, challenging myself and others to produce the best science we possibly can. He is a compassionate mentor who puts the needs of his students above his own and enjoys learning from his students just as much as we enjoy learning from him. He constantly fosters a community of inspiration and support in the lab, which has created a wonderfully accomplished lab family I am so grateful to be a part of.

The Wilhelm lab family has been a constant in my life for nearly 12 years. There are too many to name here, but each of them has brought me more joy than I could ever describe. I am so thankful for the hard research questions, the random data advice, the drinks shared, and the many adventures we experienced together. I must also acknowledge the many establishments that provided sustenance for our academic contemplations, research inspirations, and casual comradery.

Finally, I have the utmost appreciation for the love and constant support from my mother, father, and husband, through every woe and every success. They are the foundation from which I have built my life and I am so very grateful for them.

ABSTRACT

Harmful algal bloom events are notoriously associated with massive economic and environmental consequences, causing wildlife and human health risks. As these blooms increase in occurrence, duration, and severity around the world, it is essential to understand conditions leading to bloom formation and why they persist. Abiotic factors such as nutrients are commonly considered in bloom dynamics, but biotic interactions with co-occurring microbial species and viruses must also be taken into account. Harmful algal blooms dominated by the cyanobacterial genus *Microcystis* occur in bodies of water around the world and provide an ideal system in which to study top-down controls on bloom dynamics. Co-occurring cyanobacteria, photosynthetic eukaryotes, and heterotrophic bacteria can all directly influence *Microcystis* growth, through competitive or mutualistic interactions. Viruses are also often implicated in the top-down control of *Microcystis* bloom systems both directly and indirectly, and evidence suggests that viruses have potential to influence bloom dynamics. There have been observations of *Microcystis* phage infections in previous bloom events in Lake Erie, USA, and Lake Taihu, China, directly affecting *Microcystis* populations. Viruses can also act as indirect controls by infecting potential competitors such as photosynthetic eukaryotes, allowing for *Microcystis* to have the competitive advantage. The methods of studying *Microcystis* blooms are equally important as the hypotheses themselves. Molecular sequencing data can provide a wealth of knowledge not only on *Microcystis* physiology and activities, but also the identity and functions of the microbiome. Careful consideration of available analysis tools is required to accurately characterize these relationships and provide a more holistic view of a *Microcystis*-dominated harmful algal bloom. A full understanding

of these blooms cannot come from evaluating *Microcystis* in isolation, as it does not exist in isolation in the natural environment. This dissertation seeks to fill some of these gaps in the knowledge by characterizing the viral and host community in natural *Microcystis* blooms.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

Harmful Algal Blooms History 2
Microcystis Blooms 4
Community Controls on *Microcystis* Blooms 9
Methodological Advances in *Microcystis* Bloom Research 15
Dissertation Overview 20

CHAPTER 2: TRACING THE ACTIVE GENETIC DIVERSITY OF *MICROCYSTIS* AND *MICROCYSTIS* PHAGE THROUGH A TEMPORAL SURVEY OF *TAIHU* 22

Publication Note 23
Abstract 24
Introduction 26
Methods 29
Results 35
Discussion 38
Data Availability 46
Acknowledgements 47

CHAPTER 3: THE "NEGLECTED" VIRUSES OF *TAIHU*: ABUNDANT TRANSCRIPTS FOR VIRUSES INFECTING EUKARYOTES AND THEIR POTENTIAL ROLE IN PHYTOPLANKTON SUCCESSION 48

Publication Note 49
Abstract 50
Introduction 51
Methods 53
Results and Discussion 58
Data Availability 72
Acknowledgements 73

CHAPTER 4: A COMPARISON OF METATRANSCRIPTOMIC ASSESSMENT METHODS TO CHARACTERIZE *MICROCYSTIS* BLOOMS 74

Publication Note 75
Abstract 76
Introduction 77
Methods 80
Assessment 86
Discussion 93
Recommendations 95
Data Availability 96
Acknowledgements 96

CHAPTER 5: ENVIRONMENTAL STUDIES OF CYANOBACTERIAL HARMFUL ALGAL BLOOMS SHOULD INCLUDE INTERACTIONS WITH THE DYNAMIC MICROBIOME 97

Publication Note 98
Viewpoint Article 99
Acknowledgements 103

CHAPTER 6: CHANGES IN MICROBIOME ACTIVITYAND SPORADIC VIRAL INFECTIONS HELP EXPLAINOBSERVED "BOTTLE EFFECTS" IN AQUATIC SYSTEMS 104

Publication Note 105

ABSTRACT 106
INTRODUCTION 107
METHODS 110
RESULTS 116
DISCUSSION 126
DATA AVAILABILITY 132
ACKNOWLEDGEMENTS 132
CHAPTER 7: CONCLUSIONS 134
LIST OF REFERENCES 141
APPENDIX I: SUPPLEMENTAL FIGURES 157
VITA 175

LIST OF TABLES

Table 1. Marker Gene Sources..... 55

Table 2. Methods Analyzed..... 82

LIST OF FIGURES

Figure 1. *Microcystis* bloom 7

Figure 2. Viral Reproduction Cycle 12

Figure 3. *Microcystis* Bloom in China 30

Figure 4. Expression of *Microcystis* Genotypes 32

Figure 5. Virus and Host Correlations 37

Figure 6. Canonical Correspondence Analysis 39

Figure 7. NCLDV and RNA Virus Phylogenic Trees 59

Figure 8. Host Phylogenetic Tree 63

Figure 9. Viral Expression Heatmap 64

Figure 10. Virus Host Co-occurrence Clusters 66

Figure 11. Total *Microcystis* Reads 88

Figure 12. Microbiome Shifts 101

Figure 13. Chlorophyll *a* concentrations 117

Figure 14. Microbiome Expression Heatmap 119

Figure 15. *Microcystis* Phage Expression Heatmap 121

Figure 16. Lytic vs. Lysogenic Expression 123

Figure 17. Metabolite Correlation to Phage 125

LIST OF ATTACHMENTS

Supplemental Table 1. Table of environmental parameters measured with each sample collected (A) and the associated biplot scores for the *RpoB* canonical correspondence analysis (B).

Supplemental Table 2. Sequences of RpoB, McyA, and terminase genotypes identified and analyzed.

Supplemental Table 3. Blast databases with full sequences.

Supplemental Table 4. Reference proteins for base phylogenetic trees.

Supplemental Table 5. Raw transcript expression for viral candidates.

Supplemental Table 6. Table of virus-host correlation coefficients in co-occurrence clusters.

Supplemental Table 7. Table of Pearson's correlation coefficients between methods analyzed.

Supplemental Table 8. Table of total reads recruited to or classified as *Microcystis* (A) or to specific genes in *Microcystis* (B) for each method evaluated.

Supplemental Table 9. Table of environmental parameters measured with each sample collected, raw transcript expression for *rpo*B expression of the microbiome, *Microcystis* KEGG orthologies, *Microcystis* phage MaLMM01 genome, viral terminase and differential expression $\log_2$ fold changes and p-values.

CHAPTER 1: INTRODUCTION

Harmful Algal Blooms History

"Water, water, everywhere, nor any drop to drink" (Coleridge 1857). This line from the famous poem, *The Rime of the Ancient Mariner,* was originally used to describe salt water but has since been used to describe the global phenomena of harmful algal blooms. For centuries, there have been written and spoken accounts of discolored waters that cause sickness or death. Some believe that the Bible holds the first description of a harmful algal bloom, with the verses "all the waters that were in the river were turned to blood. And the fish that was in the river died; and the river stank, and the Egyptians could not drink of the water of the river" (Exodus 7: 20-21). Others point to reports from a Chinese General in ~1000 A.D., which described illnesses in troops that had drank from a green lake in the South of China (Chorus and Welker 2021). Still others point to Christopher Kirkby's account published in the Philosophical Transactions of the Royal Society in 1672, where he describes a "green, hairy efflorescence" that occurred in a lake in the summer months, and killed cattle, dogs, and poultry when they drank it. Countless other references have been made throughout history, mentioning native knowledge and personal accounts of seasonal changes to bodies of water that result in sickness or death if ingested or if seafood is ingested at that time (Dale and Yentsch 1978; Hallegraeff 2003). Regardless of when they were first noticed, harmful algal blooms have been occurring for a very long time, far before major anthropogenic influences.

Modern science defines harmful algal blooms as the overgrowth of algae to the point of nuisance, either through sheer biomass loads or through the presence of a toxin, or both (Glibert et al. 2005; Sellner et al. 2003). Blooms are commonly the result of

eutrophication, or the excess of nutrient inputs, either natural or human induced, which can lead to the overgrowth of algae (Burkholder 2000; Wetzel 1983). They are a global issue, occurring on every continent and in every ocean, even the Southern Ocean (Glibert et al. 2005; Ho et al. 2003). A variety of different types of algae are implicated in blooms, including eukaryotic phytoplankton, eukaryotic macroalgae, and cyanobacteria, among others (Sellner et al. 2003). Blooms occur in both fresh and salt waters and can be ephemeral in nature or persistent, depending on environmental conditions. Excessive biomass loads can cause a range of severe ecological impacts to a system, including death of submerged vegetation that provides a crucial fisheries habitat and/or hypoxia that leads to fish die-offs. Both mortality events can be attributed to the build-up of algae on the surface, first causing shading to the waters below, then causing massive amounts of oxygen depletion through respiration when the algae eventually die (Anderson et al. 2002).

The other primary ecological impact arises from the presence of toxic compounds, a characteristic of many harmful algal bloom events. These toxins, or other harmful metabolites, are commonly produced by the proliferating algae and can cause sickness or death to not only marine and aquatic animals, but also mammals and birds (Anderson et al. 2002). Human illness or death often occurs when contaminated water or seafood ingested, although some respiratory illnesses have been attributed to inhaled water droplets (Anderson et al. 2002; Glibert et al. 2005). Other symptoms have also been associated with the presence of blooms, including skin irritation and lesions, gastrointestinal issues, memory loss, and seizures (Anderson et al. 2002; Sellner et al. 2003). Many of the most notorious toxins associated with human illnesses are produced

by marine harmful agal blooms. These include ciguatoxin (ciguatera fish poisoning), okadaic acid (diarrhetic shellfish poisoning), brevetoxin (neurotoxic shellfish poisoning), and domoic acid (amnesic shellfish poisoning) (Grattan et al. 2016). However, there are many freshwater cyanobacterial harmful algal blooms that also produce deadly toxins, such as microcystins, cylindrospermopsins, and anatoxins (Chorus and Welker 2021).

In addition to ecological impacts, harmful algal blooms can have severe economic impacts. Commercial and recreational fisheries suffer during bloom conditions as do tourism-based communities and local water treatment municipalities (Hoagland et al. 2002). Tainted seafood not only causes billions of dollars in resource losses, but it also causes massive losses in regional incomes. Billions of dollars in commercial losses are associated with marine harmful algal blooms in the United States alone (Anderson et al. 2021). Losses associated with freshwater blooms are smaller, yet still significant. A single bloom event in 2014 in Lake Erie was estimated to cost approximately 65 million dollars (Bingham et al. 2015). Global costs are undoubtably higher, and likely to increase, with the spread of harmful algal blooms into new waterways. Recent increases in duration and severity of blooms are variable by region but increased human activities in all bloom regions will cause an increase in economic impacts (Hallegraeff et al. 2021).

Microcystis Blooms

While marine harmful algal blooms cause most of the ecological and economic challenges surrounding fisheries, freshwater harmful algal blooms pose a particular challenge to potable water resources. Freshwater blooms are primarily dominated by cyanobacteria, particularly during warmer, summer months (Hudnell 2010). Many

cyanobacterial bloom formers are capable to producing cyanotoxins, such as the potent hepatotoxin microcystin. The microcystin toxin was originally referred to as the "Fast Death Factor" due to the 1-3hr (LD_{50} = 50 μg kg^{-1}) lethal dose in white mice (Bishop et al. 1959; Rinehart et al. 1994). The World Health Organization (WHO) has set the following limits for general drinking water and short-term (up to 12 days) drinking water as 1μg L^{-1} and 12 μg L^{-1}, respectively (Chorus and Welker 2021; Organization 2020). The WHO also recommends suspending recreational activities when toxin loads exceed 24 μg L^{-1} (Chorus and Welker 2021; Organization 2020). The ingestion of contaminated water has been shown to cause liver poisoning in a variety of mammals, including dogs and sea otters (Miller et al. 2010; van der Merwe et al. 2012). In one tragic event, microcystin-contaminated water was used at a dialysis clinic in Brazil in 1996, causing 56 human deaths (Azevedo et al. 2002). Given the stringent limits on human ingestion, this toxic, cyclic heptapeptide poses a particular challenge to water treatment facilities. Toxin release primarily occurs when cells lyse and can persist in the water column for weeks (Harada 1996). Therefore, water treatment facilities primarily try to physically separate cells, while avoiding the disruption of cell integrity (Chorus and Welker 2021). To remove microcystins from the dissolved phase, water treatment facilities common apply powered activated carbon in the form of pulverized charcoal (Bruchet et al. 1998; Chorus and Welker 2021).

There are several genera of freshwater cyanobacteria capable of producing microcystin, including *Dolichospermum* (formerly *Anabaena*), *Nostoc, Planktothrix,* and *Microcystis.* While each of these has been shown to produce microcystin, heavy toxin loads are most commonly associated with *Microcystis*, hence the name (Botes et al.

1984). *Microcystis* is a buoyant, coccoid bacteria of the Order Chroococcales that can form colonies and accumulate on the surface of water bodies around the world (Figure 1). (Chorus and Welker 2021). There are records of *Microcystis* blooms in 108 countries that span every continent except Antarctica (Harke et al. 2016; Zurawell et al. 2005). There are approximately a dozen species "morphotypes," with *M. aeruginosa, M. flos-aquae, M. wesenbergii, M. viridis, M. botrys,* and *M. panniformis* among the most commonly observed (Crow 1923; Komárek and Komárková 2002). *M. aeruginosa* is typically the poster child for microcystin production, as the toxin was originally isolated in this species, but many other *Microcystis* species have also been shown to produce microcystin (Botes et al. 1984). However, it is important to note that not every *Microcystis* species or strain can produce microcystin and that the regulation of toxin production is still poorly understood (Harke et al. 2016).

Occurring in lakes and ponds around the world, *Microcystis* can rise to numerical dominance in a system when temperatures rise above 15º C (Jacoby et al. 2000; Reynolds et al. 1981). Anthropogenic nutrient loading, particularly nitrogen and phosphorus, has been largely attributed to the severity of bloom occurrence, although light and *Microcystis* is not able to fix nitrogen, it is able to utilize a variety of nitrogen sources, including ammonium and urea, which are commonly introduced to the watershed through agricultural fertilizers (Blomqvist et al. 1994; Chaffin et al. 2013; Krausfeldt et al. 2019a). Ecosystems with higher levels of biologically available forms of nitrogen tend to favor species like *Microcystis,* as opposed to nitrogen fixers like *Nostoc* or *Dolichospermum* (Harke et al. 2016; Hyenstrand et al. 1998). However, most bloom

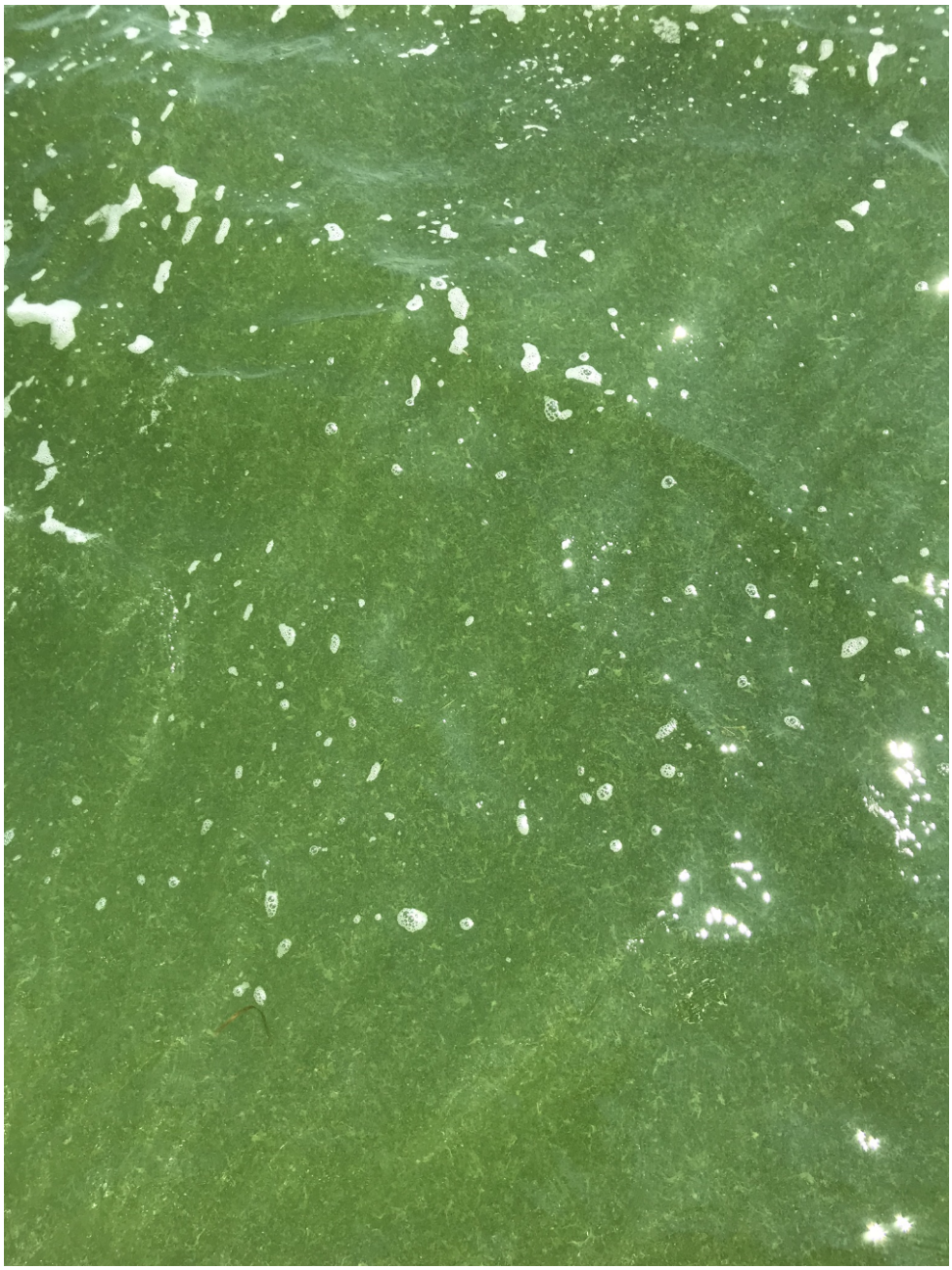


Figure 1. *Microcystis* bloom

Surface accumulation of *Microcystis* in a June 2019 bloom in Lake Erie, U.S.A. Photo credit: H.L. Pound.

management strategies have historically focus on controlling phosphorus input, as *Microcystis* is highly adept at utilizing and storing phosphorus, even when loads are minimal (Jacobson and Halmann 1982). Several studies have shown that *Microcystis* growth in culture and in the environment can be supported by both inorganic and organic forms of phosphorus (Davis et al. 2010; Harke and Gobler 2013).

While *Microcystis* may reach numerical dominance during a bloom event, studies in the past 20 years have shown that these systems are not homogenous (Cook et al. 2020; Eiler and Bertilsson 2004; Smith et al. 2021). Blooms are heterogenous systems comprised of photosynthetic eukaryotes like diatoms and other protists, other cyanobacteria, like *Nostoc* and *Planktothrix,* and heterotrophic bacteria, like Proteobacteria and Actinobacteria (Harke et al. 2016). While the specific interactions between *Microcystis* and these co-occurring community members are still unknown, there is evidence in both lab-based studies and natural systems that *Microcystis* grows better when heterotrophic bacteria are present (Kim et al. 2019). It is unclear if, or how, *Microcystis* selects for the co-occurring community. However, *Microcystis* stimulated shifts in the environmental conditions of the system are hypothesized to influence the mutualistic partners (Cook et al. 2020; Smith et al. 2021). For example, *Microcystis* has been shown to elevate the pH of its surroundings, reducing CO_2 through photosynthesis (Bullerjahn et al. 2016; Krausfeldt et al. 2019a) and decrease light levels through sheer biomass loads (Paerl and Otten 2013). Combined with the influence of environmental conditions, the community diversity in a bloom provides an exciting challenge for researchers.

Community Controls on *Microcystis* Blooms

Although other species are present, the lack of balanced community diversity in bloom events contradicts established theories concerning ecological checks and balances. Hutchinson's *Paradox of the Plankton* posits that the high diversity of algal species exists through environmental variability and symbiotic relationships, even under limiting nutrients (Hutchinson 1961). Harmful algal blooms somehow circumvent this concept, allowing the proliferation of a single or few dominant species. The mechanisms that allow this are poorly understood, as the drivers of bloom formation have not been fully elucidated. Most of the previous work on harmful algal blooms has pertained to abiotic controls such as nutrient or temperature triggers, commonly referred to as "bottom-up" controls (Steffen et al. 2015). Nitrogen and phosphorus concentrations are commonly implicated as primary controlling factors in algal blooms (Krausfeldt et al. 2017; Paerl and Otten 2013; Steffen et al. 2014b; Xu et al. 2010). These controls commonly examine the abiotic environmental factors or mechanisms that contribute to the function and behavior of a single species. However, recent studies have explored the role of "top-down" controls on algal growth and bloom formation (Steffen et al. 2017). "Top-down" controls examine the interaction networks within a community, considering predator/prey, competitive, and mutualistic relationships between species. These "top-down" controls are largely driven by the co-occurring members of the microbial community, including other cyanobacteria, heterotrophic bacteria, photosynthetic eukaryotes, and viruses that infect all of the above, including *Microcystis.*

The rest of the microbial community, including photosynthetic eukaryotes and heterotrophic bacteria can act as "top-down" controls, directly influencing *Microcystis* species in a competitive or mutualistic manner. The mutualistic relationship with heterotrophic bacteria mentioned above typically involves *Microcystis* providing carbon to the heterotrophs in exchange for micronutrients, vitamins, CO_2, and antioxidants (Amin et al. 2015; Basu et al. 2019). These heterotrophic bacteria are commonly closely associated with *Microcystis* in the area known as the phycosphere, the mucus region surrounding *Microcystis* cells (Bell and Mitchell 1972; Seymour et al. 2017). There is a lack of scientific consensus on the conservation of the heterotrophic community. Some studies have found no conserved community associated with *Microcystis* (Smith et al. 2021), while others have established high phylogenetic relatedness across various samples (Cook et al. 2020). However, most studies agree that the communities within the phycosphere differ from those without.

Meanwhile, a competitive relationship seems to exist between *Microcystis* and photosynthetic eukaryotes. While *Microcystis* forms blooms in the warmer, summer months, it is not the most numerically dominant species year-round. During the cooler, winter months, many lake systems are dominated by eukaryotic phytoplankton blooms, such as Chlorophyta or Bacillariophyta (Edgar et al. 2016; George and Edwards 1976; Ke et al. 2008; Niu et al. 2011). This seasonal succession indicates some form of competition between *Microcystis* and the photosynthetic eukaryotes that is not yet fully understood. The competitive relationships are hypothesized to exist primarily with the eukaryotic phytoplankton such as diatoms or protists. Commonly occupying a similar photosynthetic niche, the presence and success of such species can also influence nutrient and light

availability to *Microcystis*. Environmental shifts in nutrients, temperature, light or pH not only influence *Microcystis,* but these other community members. Regardless, the composition of the eukaryotic and heterotrophic bacterial community, influence the functional capacity of the community, thereby influencing their contributions to the success of *Microcystis.*

Another primary example of a "top-down" control is viral infection, where viruses can act in a predatory role by directly infecting *Microcystis* species. Viruses are the most numerous biological entities on Earth. A single drop of seawater can hold up to 10^7 viruses, while the entire Earth supports abundances that can total 10^{31} (Breitbart and Rohwer 2005; Cochlan et al. 1993; Suttle 2005). Viruses are both plentiful and globally diverse, occupying a huge variety of niches alongside their hosts. They are commonly correlated with host abundances, indicating the potential for coupling in observed succession patterns (Wommack and Colwell 2000). Viruses play an integral role in both macro and micro ecosystems, actively influencing population dynamics and community structure (Fuhrman 1999). By causing cell death, viruses can be directly responsible for bloom collapse, or indirectly responsible for bloom maintenance through nutrient and organic matter turnover (Wilhelm and Suttle 1999).

The impact of a virus on a microbial community can vary drastically based on the viral reproduction cycle. A lytic infection, the most studied, causes immediate cell death and release of viral progeny. This is thought to be the cause of "boom and bust" cycling associated with dominant species and potential toxin release (Thingstad and Lignell 1997). Some viruses, known as temperate viruses, are capable of both a productive lytic cycle and a dormant lysogenic cycle (Lwoff 1953) (Figure 2). By definition, lysogeny is

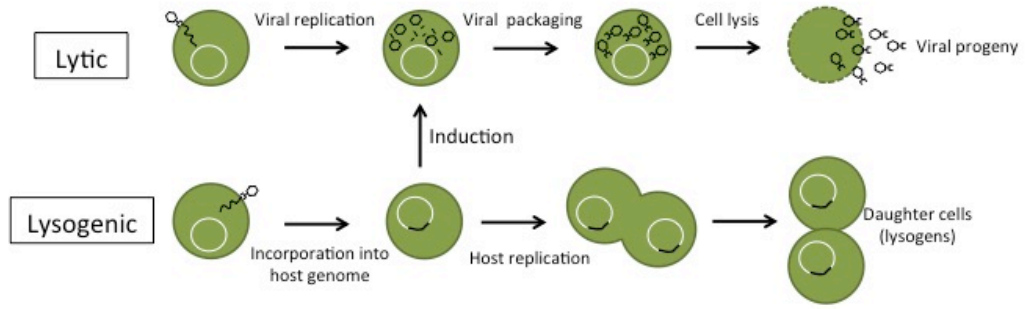


Figure 2. Viral Reproduction Cycle

When a virus encounters a cell, it either reproduces immediately via the lytic cycle or incorporates itself into the host genome via the lysogenic cycle. White circles indicate a host genome and black portions represent viral genetic material.

he “hereditary power to produce bacteriophage” (Lwoff 1953). A lysogenic infection causes a dormant infection in the host, where the virus inserts itself into the host genome as a prophage. As such, the virus can then be maintained within the host for many generations without producing viral progeny, acting as a potential viral reservoir.

For decades, researchers hypothesized that there were viruses capable of lysing *Microcystis*, as viruses had been previously implicated in bloom collapse in other species (Bratbak et al. 1998; Tomaru et al. 2004). Several *Microcystis* phage have been observed in environmental sequencing efforts, but only three have been isolated. In 2006, the first phage, known as Ma-LMM01, was originally isolated and a nearly identical phage, MaMV-DC was isolated in 2013 (Ou et al. 2013; Tucker and Pollard 2005; Yoshida et al. 2006). This virus was classified in the Myoviridae clade (due to its morphology) and has been implicated as a potential controlling agent in *M. aeruginosa* blooms (Paerl and Otten 2013; Tucker and Pollard 2005; Yoshida et al. 2006). Unlike other cyanophages though, the Ma-LMM01 myovirus does not contain many auxiliary metabolic genes, particularly those associated with photosynthetic machinery (Yoshida et al. 2008). Metatranscriptomic studies showed that genes similar to those in the *M. aeruginosa* phage Ma-LMM01 are actively expressed during *Microcystis* blooms, lending further support to their role as controlling agents in bloom dynamics and toxin release (Steffen et al. 2015; Stough et al. 2017). In fact, the presence of this phage was associated with the high dissolved toxin loads that caused the shutdown of the Toledo water treatment facility, leading to a water shortage crisis for 400,000 residents. A lytic phage infection was hypothesized to have lysed a large portion of the *Microcystis* bloom, releasing the internal toxin stores into the water column (Steffen et al. 2017). In this instance, the

phage was responsible for a partial bloom collapse. However, another study from a *Microcystis* bloom in Lake Tai, China observed the transcriptomic expression of the same virus, but described a seasonal shift from a lytic reproduction cycle to a lysogenic reproduction cycle (Stough et al. 2017). In this instance, the phage may have been responsible for bloom maintenance, through superinfection exclusion by preventing the infection of the host by another potentially lytic virus (Susskind and Botstein 1980). Scientists are not yet able to explain the cellular or viral mechanistic switch that determines a lytic versus lysogenic reproduction cycle in *Microcystis*, nor the environmental triggers that might drive this shift.

In addition to acting as a predatory control on *Microcystis* through direct infection, viruses can also act as an indirect "top-down" control by infecting other community members. Given the numerical dominance of *Microcystis* during a bloom event, most of the research has been focused on *Microcystis* and its phage. Much less consideration has been given to the role of the rest of viral community. These "neglected viruses" infect the co-occurring, but less abundant photosynthetic eukaryotes, heterotrophic bacteria, and other cyanobacteria. A viral-stimulated population reduction of these community members has the potential to influence *Microcystis* dynamics in a variety of ways, depending on the relationship *Microcystis* has with the community member. The potential for viral suppression of competitive species is reminiscent of the "kill the winner" hypothesis, although this theory was originally intended to describe a heterotrophic community in equilibrium dynamics (Thingstad and Lignell 1997). While the succession associated with harmful algal blooms do not display equilibrium dynamics, the concept presented may still hold value. It suggests that a viral infection and

subsequent population reduction of a competitive species could allow for the proliferation of the competitor (Winter et al. 2010). So, in the case of a *Microcystis* bloom, viral infection of photosynthetic eukaryotes may allow for *Microcystis* to reach the high densities observed during bloom events. However, a viral infection of a mutualistic heterotrophic bacterial population may hinder *Microcystis* growth and success. A lack of critical micronutrients, vitamins, or antioxidants provided by co-occurring heterotrophs could weaken a *Microcystis* bloom. Overall, it is very unclear what roles viruses may play in *Microcystis* bloom dynamics but is hypothesized to be highly dependent on the environmental conditions and community composition.

Methodological Advances in *Microcystis* Bloom Research

Methods to Evaluate Hosts

Just like any other field of research, the methods of studying harmful algal bloom research have vastly improved over the years. Early *Microcystis* studies employed the use of microscopes and identification keys of physical characteristics to classify rough species groups, a technique still used by current researchers (Crow 1923). It is unclear exactly when researchers began to attempt to isolate *Microcystis* cultures in the lab, however, reports of successful isolation began in the 1950's (Gerloff et al. 1950; Hughes et al. 1958). Once isolated, many culture-based studies were performed, examining a variety of topics from toxin production (Bishop et al. 1959) to nutrient utilization (McLachlan and Gorham 1962; Zehnder and Gorham 1960). Most of these early studies continued to use microscope counts, turbidity, or dry weights to evaluate growth of *Microcystis* and electrophoresis methods to isolate compounds like toxins. Among others,

chromatography and fluorometry methods continued to advance the state of the science for many years, until high throughput sequencing was introduced.

Heterotrophic bacteria were first suggested to form mutualistic relationships with cyanobacteria in 1950, when researchers attempting to culture the bloom formers found axenic isolation difficult (Gerloff et al. 1950). Studies in the 1970's began to isolate and identify *Microcystis* co-occurring species primarily using staining techniques, microscopy slides, and bacterial identification schemes (Ehlke 1974; Skerman 1960). Their presence and activity were later confirmed with bacterial production assays using tritiated leucine (Worm and Søndergaard 1998). Observations of diatoms in *Microcystis* blooms were first recorded in notes about a *Microcystis* bloom in North Sumatra in 1956 (Johnson 1956). These observations occurred sporadically for decades, primarily through microscopy and physiology-based identification schemes, although little research was explored on diatoms during summer *Microcystis* blooms for many years (Boström et al. 1989; Flower 1982).

A new method of species identification was established in 1977, with the sequencing of the universally conserved, prokaryotic 16S rRNA gene (Sanger et al. 1977; Woese and Fox 1977). DNA isolation and subsequent sequencing of this gene allows researchers to identify small nucleotide changes that distinguished individuals of one species from another. This method was used to characterize species present in *Microcystis* blooms beginning in the mid-1990s (Kondo et al. 2000; Neilan et al. 1994; Rudi et al. 1997). A comprehensive comparison of all known *Microcystis* species was performed using 16S rRNA gene sequencing in 2002 (Komárek and Komárková 2002). As all prokaryotic organisms have a 16S ribosomal subunit, 16S rRNA gene sequencing

is still a primary method used in current research to characterize the composition of *Microcystis* blooms. A similar ribosomal subunit, 18S, is commonly used to characterize the eukaryotic community. While the ribosomal small subunit provides an excellent way to characterize what species are present in a bloom, it can do little to inform researchers on the full genetic potential or functional activity of bloom species. Polymerase chain reactions further enabled researcher's ability to amplify DNA and better characterize specific genes within the *Microcystis* genome, such as the toxin producing operon (Mullis 1990; Ouellette et al. 2006; Rinta-Kanto et al. 2005).

Researchers subsequently built upon Sanger sequencing to develop more high-throughput methods of sequencing such as 454, Illumina, and Sequencing by Oligo Ligation Detection (SOLiD) sequencing, which allowed for more rapid sequencing of many genes (Margulies et al. 2005; Valouev et al. 2008; Wang et al. 2009). These advances allowed for the sequencing of the entire genome and transcriptome of an organism, providing insight on the genetic potential and transcriptional activity, respectively. *M. aeruginosa* NIES -843 was the first *Microcystis* isolate to be fully sequenced (Kaneko et al. 2007). To date, there are 176 draft genomes in the *Microcystis* genus, with 9 of those considered "closed", including *M. aeruginosa* NIES-843. Whole genome comparisons allow for more comprehensive evaluations of species relatedness, causing some debate amongst researchers on the number of unique *Microcystis* species (Harke et al. 2016; Otsuka et al. 2001). Regardless, genomic and transcriptomic sequencing of cultured *Microcystis* isolates has allowed for a much greater understanding of many cellular processes including toxin production, nutrient utilization, photosynthesis, and pigment production (Ouellette and Wilhelm 2003).

These sequencing technologies can then be expanded to encompass natural communities, not only characterizing natural *Microcystis* populations, but also the co-occurring microbial communities. This type of analysis tool is referred to as metagenomics or metatranscriptomics. While these tools were initially used to explore questions regarding toxin production in bloom events, they have rapidly expanded to encompass nutrient utilization, temperature and CO_2 influences, and ecological interactions within and among species present in a bloom (Steffen et al. 2014b; Tang et al. 2018). However, generating sequence data is only part of the challenge associated with these analyses. Not all sequence analysis tools and methods were created equal, therefore requiring researchers to carefully consider the rigor of the method chosen, as well as its suitability for the hypothesis tested. Many methods are based on the current knowledge of cultured *Microcystis* isolates, so special care must be given to account for the true diversity present in an environmental bloom sample.

Methods to Evaluate Viruses

Our current knowledge of *Microcystis* blooms accounts for a role of viruses in bloom dynamics, but during the early decades of *Microcystis* bloom research, viruses were largely ignored. They were hypothesized to be capable of lysing a *Microcystis* bloom, but there was very little evidence (Goriushyn and Chaplinskyaya 1968; Padan and Shilo 1973). The first virus capable of lysing *Microcystis* was not isolated until 2006 (Yoshida et al. 2006). Viruses associated with the other bloom community members was not actively investigated in the context of a bloom event during this time.

Unlike their hosts, viruses have no universally conserved markers like 16S or 18S rRNA genes. The advent of genomic and transcriptomic sequencing technologies that

benefited *Microcystis* research also benefited virus research, albeit at a slower pace. Challenges in culturing viral isolates mean that few *Microcystis* phage genomes exist, but the first isolated phage, MaLMM01 had its genome sequenced in 2008 and its transcriptome sequenced in 2018 (Morimoto et al. 2018; Yoshida et al. 2008). Metagenomic and metatranscriptomic studies have since provided exciting new observations regarding the role of phages in *Microcystis* bloom collapse and maintenance (Morimoto et al. 2019; Steffen et al. 2015; Stough et al. 2017). However, the viruses of other co-occurring community members have been largely overlooked, partly due to the lack of research on their hosts during bloom events, and partly due to their diverse genetic makeup. There are a variety of types of viral genomes, including, but not limited to, single-stranded DNA genomes, double-stranded DNA genomes, single-stranded RNA genomes, and double-stranded RNA genomes. Viruses infected eukaryotes are particularly diverse, encompassing all of these groups and ranging in size (Koonin et al. 2015). In the absence of a 16S or 18S rRNA gene, researchers must use marker genes such as the RNA-dependent RNA polymerase (RdRp) or major capsid protein (MCP), however each virus must be considered independently as no genes are universally conserved (Moniruzzaman et al. 2017; Stough et al. 2018). Many of these viruses, particularly RNA viruses, have previously been overlooked in DNA sequencing efforts of *Microcystis* blooms. While the technologies allow for incredible exploration of *Microcystis* blooms, researchers must be cognizant of the technical limitations and biases of various methods and must take care to account for all potential community members.

Dissertation Overview

This dissertation serves to characterize the microbiome of a *Microcystis* bloom, including both co-occurring hosts and viruses capable of influencing *Microcystis* dynamics. Most of the work presented here involves the analysis of natural bloom systems, fully describing the breadth of diversity present. It begins with a description of *Microcystis* host richness in a natural community and the influence of a *Microcystis* phage infection using a technique adept at characterizing genotypic host diversity. This analysis showed that *Microcystis* species or genotype was not correlated with the presence of lytic *Microcystis* phages, indicating that other environmental variables might be more responsible for viral dynamics within a bloom. The second chapter goes on to define the viral community associated with the non-*Microcystis* members of the bloom community, primarily those that infect photosynthetic eukaryotes. Conclusions from this study suggest that these "neglected viruses" may play a role in the suppression of their hosts, providing a competitive advantage for *Microcystis.* To better characterize the host community (both *Microcystis* and non-*Microcystis*), chapter three describes a methodological analysis of existing tools and methods used to classify and annotate host species activity in a *Microcystis* bloom metatranscriptome. This work improves upon previously used methods to characterize the host population and provides specialized recommendations for methods, depending on whether the study intends to focus primarily on *Microcystis,* or the larger bloom community. This dissertation then goes on to detail how these methods were used to describe shifts in the co-occurring community in both cultured lab isolates (chapter three) and in a natural bloom community (chapter four).

These studies enforce the idea that environmental changes such as nutrient additions or bottle effects can cause major shifts in the heterotrophic community, and in the case of the natural community analyzed in chapter four, the viral community. All of the work presented in this dissertation emphasizes the crucial role the community plays in *Microcystis* blooms. While they may not be the toxin producing, green sludge making headlines, the other organisms present in these communities cannot be ignored and deserve equal attention in bloom research.

CHAPTER 2: TRACING THE ACTIVE GENETIC DIVERSITY OF *MICROCYSTIS* AND *MICROCYSTIS* PHAGE THROUGH A TEMPORAL SURVEY OF *TAIHU*

Publication Note

This chapter is a version of a peer-reviewed article that was published in PLoS One by Helena L. Pound and Steven W Wilhelm. Helena L. Pound was responsible for primary data analysis and composition of the article with input from Steven Wilhelm.

Citation:

Pound HL and SW Wilhelm (2020). Tracing the active genetic diversity of *Microcystis* and *Microcystis* phage through a spatial and temporal survey of *Taihu.* PLoS One 15:12. DOI: 10.1371/journal.pone.0244482

Abstract

Harmful algal blooms are commonly thought to be dominated by a single genus, but they are not homogenous communities. Current approaches, both molecular and culture-based, often overlook fine-scale variations in community composition that can influence bloom dynamics. We combined homology-based searches (BLASTx) and phylogenetics to distinguish and quantify *Microcystis* host and phage members across a summer season during a 2014 *Microcystis*- dominated bloom that occurred in Lake Tai (*Taihu*), China. We found 47 different genotypes of the *Microcystis*-specific DNA-dependent RNA polymerase (*rpoB*), which included several morphospecies. *Microcystis flos-aquae* and *Microcystis wesenbergii* accounted for ~86% of total *Microcystis* transcripts, while the more commonly studied *Microcystis aeruginosa* only accounted for ~7%. *Microcystis* genotypes were classified into three temporal groups according to their expression patterns across the course of the bloom: early, constant and late. All *Microcystis* morphospecies were present in each group, indicating that expression patterns were likely dictated by competition driven by environmental factors, not phylogeny. We identified three primary *Microcystis*-infecting phages based on the viral terminase, including a novel *Siphoviridae* phage that may be capable of lysogeny. Within our dataset, *Myoviridae* phages consistent with those infecting *Microcystis* in a lytic manner were positively correlated to the early host genotypes, while the *Siphoviridae* phages were positively correlated to the late host genotypes, when the *Myoviridae* phages express putative genetic markers for lysogeny. The expression of genes in the microcystin-encoding *mcy* cassette was estimated using *mcyA*, which revealed 24

Microcystis-specific genotypes that were negatively correlated to the early host genotypes. Of all environmental factors measured, pH best described the temporal shift in the *Microcystis* community genotypic composition, promoting hypotheses regarding carbon concentration mechanisms and oxidative stress. Our work expounds on the complexity of HAB events, using a well-studied dataset to highlight the need for increased resolution of community dynamics.

Introduction

Bodies of water around the world are blighted by annual blooms of algal biomass. Known as harmful algal blooms (HABs), these events of economic and environmental concerns are often associated with the production of toxic compounds and/or excessive biomass accumulation (Anderson et al. 2002; Hallegraeff 2003). Most bloom events, be they summer or winter / freshwater or marine, occur when a single genus of algae evades normal biological constraints and achieves numerical dominance in a community (Glibert et al. 2005). As such, events are typically referred to by the name of the dominant genus: for example, in China's Lake Tai, (*Taihu* in Mandarin) these are generally referred to as *Microcystis* spp. blooms. The annual reoccurrence of similar species has been established in scientific literature, providing an excellent study system for genetic shifts in microbial communities (Tang et al. 2018).

Ecosystem processes are driven by the composition and function of the species present. Genotypic diversity of these species is of critical importance to community stability and performance (Latta et al. 2011). This diversity can be catalogued in several ways, including by the richness and evenness of species or subspecies present and/or the number and composition of functional roles present (Hooper and Vitousek 1997; Tilman et al. 1997). Genotypic diversity can arise in multiple ways, and is an ongoing process (Luria and Delbrück 1943). At the same time selective pressures are constant for microorganisms. Thus genetic drift occurs based on the stochastic effects of random selection (Wright 1931), niche partitioning and resource utilization efficacy (Tilman et al. 2001), environmental selection, and biotic interactions with fellow community members

including predation (Van Valen 1977) and symbiosis (Morris et al. 2012). Composite genotypic diversity in an ecosystem can influence the phenotypic community, which governs many ecosystem traits including biomass accumulation, the production of secondary compounds, and resiliency to abiotic and biotic stressors.

Genetic diversity may become particularly important during HAB events, given the presumed lack of community diversity associated blooms. Although blooms are typically dominated by only a few species, commonly members of the same genus, it is critical to remember that these communities are not homogenous. Not only are there many different genotypes or strains of the dominant taxa present, the spatial and temporal distribution of these genotypes can vary widely over the bloom duration. Indeed, there have been many observations of genotypic shifts in *Microcystis* spp. blooms, even over the course of a single bloom event (Carrillo et al. 2003; Kim et al. 2010; Rinta-Kanto et al. 2009; Yu et al. 2014), where the focus is often on the ability (or lack there-of) of these cyanobacteria to produce to potent hepatotoxin microcystin (*aka* "fast-death factor", (Hughes et al. 1958)). Most studies have used polymerase chain reaction (PCR) approaches to characterize various genotypes and their toxicity (Ouellette et al. 2006), which comes with the risk of some genotypes being excluded. As toxicity is not unique trait of a single species or strain, the community composition of co-occurring toxic organisms should be considered in toxin research (Zingone and Enevoldsen 2000).

The democratization in molecular biological tools has provided many new avenues of study in HABs. However, these advanced tools can often overlook the constant disconnect that is the dissimilarities between results from lab studies and measured environmental processes. While controlled laboratory experiments and

genomic sequencing of cultured isolates provide important information, they cannot represent the complexity of natural diversity and function present in an environment. For this reason, it may be less desirable to quantify the expression of genes in an environmental sample by recruiting sequences to the genome of a cultured isolate. This practice results in an underestimation of true diversity and can obscure species or strain level dynamics that may better reflect responses to environmental conditions.

Faced with the above, we sought to characterize the genotypic diversity of a spatially large and temporally extended *Microcystis* spp.-dominated bloom using metatranscriptomic sequencing. While previous efforts have examined specific subpopulations through recruitment to a model lab strain (Tang et al. 2018), nutrient cycling genes (Krausfeldt et al. 2017) and viruses as drivers of mortality (Pound et al. 2020; Stough et al. 2017), our efforts here were focused on the specific question of how diverse the bloom forming community was. We used a homology-based BLASTx approach accompanied by phylogenetics to identify the species and strains of *Microcystis* that were present and to quantify how their distribution varied across the course of this bloom. The analyses revealed patterns in genotype-specific expression that were linked to abiotic and biotic environmental factors, including pH and viral infection, respectively. We were also able to link changes in genotypic-marker expression to changes in toxin gene activity. Our analyses provide evidence that diversity, as well as small-scale patterns, are much broader than previously thought within the *Microcystis* community and present new hypotheses regarding the role of genotypic diversity in bloom dynamics.

Methods

Sampling and sequencing

Water samples were collected monthly from May to October from nine stations in Northwestern Lake Taihu in 2014 during daylight hours, as previously described (Krausfeldt et al. 2017; Pound et al. 2020; Stough et al. 2017; Tang et al. 2018). Based on anecdotal observations (*e.g.,* see Figure 3), we know that a *Microcystis*-dominated bloom had established biomass by the time our sample collection started in June, yet continued to accumulate through to October. Whole water samples were collected at the surface and passed through a 0.22-µm pore-size Sterivex™ filter, excess water removed, and the unit filled with RNAlater™ (Invitrogen) until extraction. Basic physical parameters were also measured at time of sampling using a multiparameter water quality sonde (YSI 6600 V2) and include dissolved oxygen, water temperature, pH, turbidity, electrical conductivity, and phycocyanin. Total nitrogen (TN), total dissolved nitrogen (TDN), ammonium (NH_4^+), total phosphorus (TP), total dissolved phosphorus (TDP), orthophosphate (PO_4^{3-}), and chlorophyll *a* (Chl-*a*) were measured using standard methods (Supplemental Table 1A). RNA extraction, sequencing: quality control details can be found in previous publications (Krausfeldt et al. 2017; Pound et al. 2020; Stough et al. 2017; Tang et al. 2018). A step by step protocol can be found at protocols.io describing the MoBio Powerwater DNA kit used to extract RNA and all modifications to the manufacturer protocol (Krausfeldt and Wilhelm 2017). The resulting material paired-end sequences of 125 bp were generated at Hudson Alpha (Huntsville, AL, USA). Sequences are available from the MG-RAST database ("Lake_ *Taihu*_metatranscriptome_project"). Trimmed

Figure 3. *Microcystis* Bloom in China

Image of the cyanobacterial community, dominated by *Microcystis* spp., from *Taihu*, China. Photo was taken during one of the sampling expeditions (October 7, 2014). Photo credit: S.W. Wilhelm.

sequences from each sample were then combined to create a single assembly using MegaHit (Li et al. 2015; Pound et al. 2020).

Host and virus detection

Genes of interest were identified and quantified using a hallmark gene approach previously described (Pound and Wilhelm 2019). Briefly, our combined assembly was queried against a BLASTx v.2.6.0+ database containing hallmark protein sequences from isolated reference genomes. *Microcystis* host species were identified using the DNA-dependent RNA polymerase (rpoB) and *Microcystis* phage were identified using the phage terminase (Supplemental Table 2). The toxin encoding protein mcyA was also used to characterize toxin production by *Microcystis* hosts and other cyanobacteria (Rinta-Kanto and Wilhelm 2006). Sequences with BLASTx hits were considered candidates if they had an e-value of less than e^{-30} or e^{-10} (for hosts and viruses respectively), were greater than 300 bp, and were less than 95% similar to any other candidate (Pound et al. 2020).

Host and virus taxonomy

Microcystis host (rpoB and mcyA) and virus candidate taxonomy was confirmed by placing the respective sequences on PhyML generated reference protein base phylogenetic trees using pplacer and visualized using iTOL v.4 (Letunic and Bork 2019; Matsen et al. 2010; Pound et al. 2020). All trees used to determine taxonomy are shown as cladograms to resolve taxonomy into distinct groups (Figure 4, Supplemental Figure 1 and 2). *Microcystis* spp. candidate sequences were first distinguished from other prokaryotes using a broad phylogenetic tree containing many prokaryotes and eukaryotes and then further characterized to species and genotype (Pound et al. 2020). The

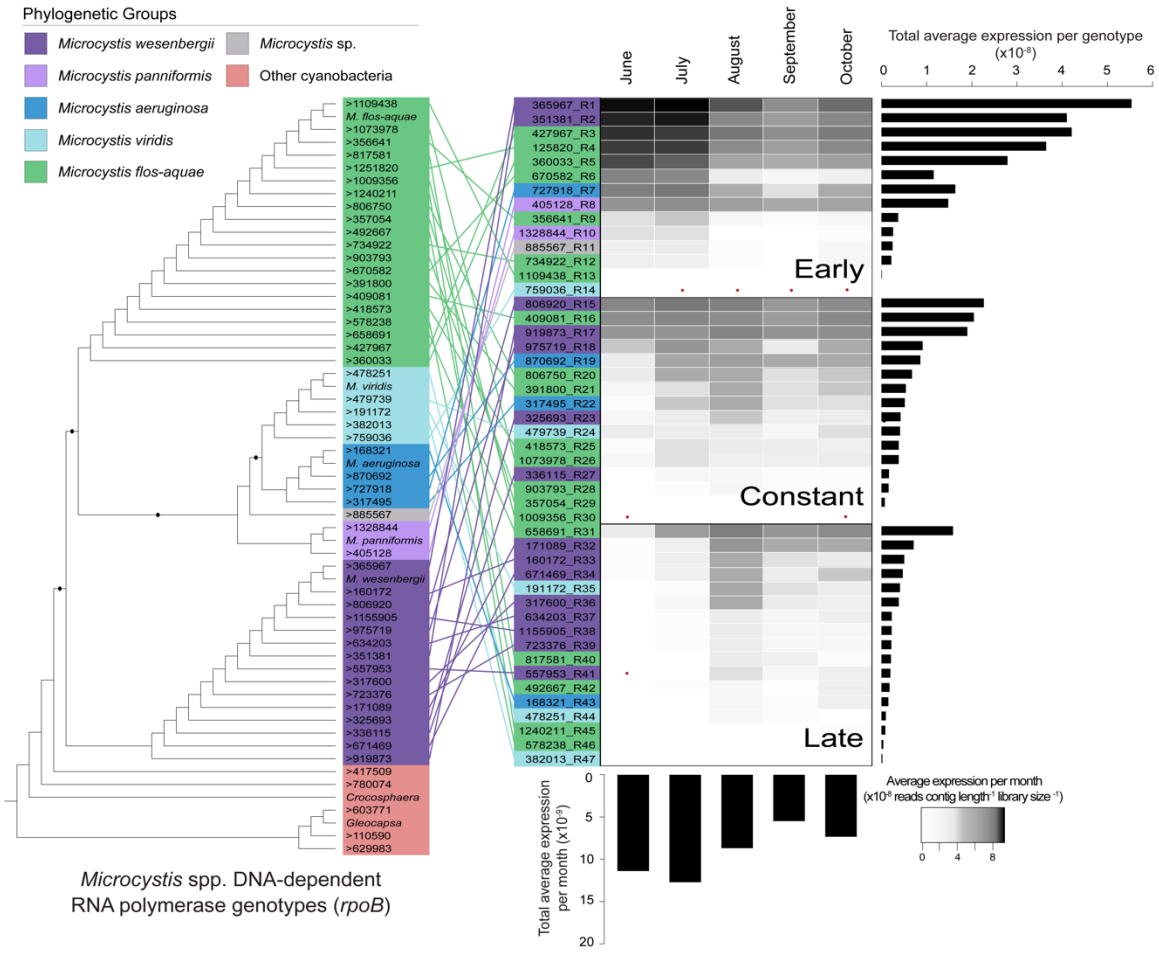


Figure 4. Expression of *Microcystis* Genotypes

Microcystis DNA-dependent RNA polymerase (rpoB) cladogram of genotypes, colored by phylogenetic group. Genotypes then rearranged based on temporal phase, with heatmap of the average total expression of each genotype. Bar charts represent the average total expression for each genotype and each month. Red dots indicate no expression.

Microcystis species-specific rpoB tree contained numerous isolates to increase genetic resolution. Approximate viral taxonomy was established by querying the candidates against the NCBI RefSeq non-redundant database, removing any false-positives or non-phage hits. The terminase tree for viruses contained many isolated viruses, both infecting and not infecting *Microcystis* and several virome sequences that originated from a *Microcystis* bloom and appeared in our BLASTx to the RefSeq database (Morimoto et al. 2019). Within these, three groups were considered specific to *Microcystis.* The first contains the *Myoviridae* strains Ma-LMM01, MaMV-DC, and Node 34. Ma-LMM01 and MaMV-DC are lab isolates known to infect *Microcystis* hosts, and Node 34 is an uncultured phage sequence from a *Microcystis* bloom virome that is 99% similar to Ma-LMM01 and MaMV-DC (Morimoto et al. 2019; Ou et al. 2013; Yoshida et al. 2006). The second group, referred to as Sipho I, is an unusual group with no isolated phage known to infect *Microcystis.* However, this group contains the *Siphoviridae* prophage-like sequence from *Microcystis aeruginosa* NIES-88 as well as Node 382, another uncultured phage from the Morimoto *et al*. virome (Morimoto et al. 2019). Because these viral terminase sequences are 98.5% similar and found both in the environment and in a *Microcystis* host genome, we assume that this is a previously overlooked phage capable of infecting *Microcystis* hosts. Node 331 was not included in this group, as it is only 56% similar. The third group, referred to as Sipho II, contains the recently discovered *Siphoviridae* phage, Mic1(Yang et al. 2020).

Host and virus activity

Transcript activity was quantified by recruiting trimmed reads to candidate contigs that were trimmed to the length of the aligned hallmark gene and normalized to

the trimmed contig length and library size (Gann et al. 2019; Pound et al. 2020). Monthly expression patterns were visualized using Heatmapper and the average expression of each genotype per month (Babicki et al. 2016). Genotypes were classified as early, constant, or late based on their patterns of average expression per month. Early genotypes displayed over 50% of their total expression during the months of June and July. Late genotypes displayed over 75% of their total expression during the months of August, September, and October. Constant genotypes did not have any month display higher than 25% of their total expression. The delineation between early and late months was originally proposed in Tang *et al*., based on changes in nutrient acquisition genes in the same dataset (Tang et al. 2018). Pearson correlations between host and virus genotypes were established and visualized using RStudio. Correlations were corrected for multiple comparisons using the Benjamini Hochberg procedure (Benjamini and Hochberg 1995). Environmental drivers of host genotypic patterns were analyzed using canonical correspondence analysis (CCA) in RStudio.

Novel virus detection

In our primary analysis, we noticed a high similarity between several phage transcripts (tail sheath, terminase, major capsid protein) and sequences including in a published *Microcystis* host genome, NIES-88 when performing BLASTx (Parajuli et al. 2016). Intrigued, we ran the host genome through Phaster (Arndt et al. 2016) and discovered a prophage-like portion. This portion of the genome was then annotated using OmicsBox v.1.2.4 (BioBam, Valencia, Spain) and visualized using CG View (Supplemental Figure 3) (Stothard and Wishart 2005).

Results

The Microcystis *community*

Our species-specific *Microcystis* rpoB cladogram indicated that 47 of the 52 candidates previously identified by Pound *et al*. (2020) were *Microcystis* (Figure 4). The other five contigs originated from other cyanobacterial species. The most abundant (total reads) *Microcystis* species were *Microcystis flos-aquae* and *Microcystis wesenbergii*, with 43.5% and 42.4%, respectively, of the total *Microcystis* spp. expression. These species also had the highest number of genotypes, with 20 *M. flos-aquae* and 15 *M. wesenbergii.* Only four genotypes of *M. aeruginosa* were observed, and these accounted for only ~7.3% of the total *Microcystis* spp. expression. The classification of early, constant, and late genotypes varies in species composition, with no obvious delineation or temporal shift in species (Figure 4). The early genotypes account for 59.7% of the total expression, while the constant and late genotypes account for 27.2% and 13.2%, respectively. In parallel, a total of 30 mcyA candidate contigs were identified: 24 most closely related to various *Microcystis* species, and six most closely related to other microcystin-producing cyanobacterial species (Supplemental Figure 1).

Microcystis-*infecting viruses*

Our method yielded 52 phage terminase candidates, 15 of which were determined to be *Microcystis* specific (Supplemental Figure 2). Seven genotypes were most closely related to the well-documented *Myoviridae* phages, Ma-LMM01 and Ma-MVDC (Ou et al. 2013; Yoshida et al. 2006). The *Myoviridae* group represented 22.4% of the total *Microcystis*-associated phage expression. Only one candidate was most closely related to

the Sipho I group containing our proposed novel phage, representing only 1.5% of the total *Microcystis* phage expression. The Sipho II group, containing Mic1, was the most abundant, with seven genotypes accounting for 76.1% of the total *Microcystis* phage expression. However, most of that expression came from just two genotypes and occurred in October.

Seasonal shifts in expression

Pearson correlations indicated distinct seasonal patterns in expression of *Microcystis*infecting phage as well as toxin production genes (Figure 5). Early *Microcystis rpoB* genotypes show a positive correlation to the *Myoviridae* phage group, but negative correlations to both *Siphoviridae* groups. The inverse is true in the late genotypes, when *rpoB* expression is negatively correlated to the *Myoviridae* group and positively correlated to both *Siphoviridae* groups. Correlation patterns with phage in the constantly expressed genotypes are less defined but show an inverse correlation to the Sipho I group and no significant correlation to the Sipho II group.

Correlation patterns of *Microcystis rpoB* and *mcyA* are similarly delineated by early, constant, and late *rpoB* genotypes (Figure 5). Early *rpoB* genotypes showed a negative relationship to toxin-encoding gene expression, while constant and late genotypes are positively correlated with toxin-encoding gene expression. *mcyA* expression is an average of 2.1x higher in the late months (August, September, and October) than it is in the early months, indicating that early genotypes display decreased expression of the microcystin encoding gene, while late genotypes display increased expression of the microcystin encoding gene.

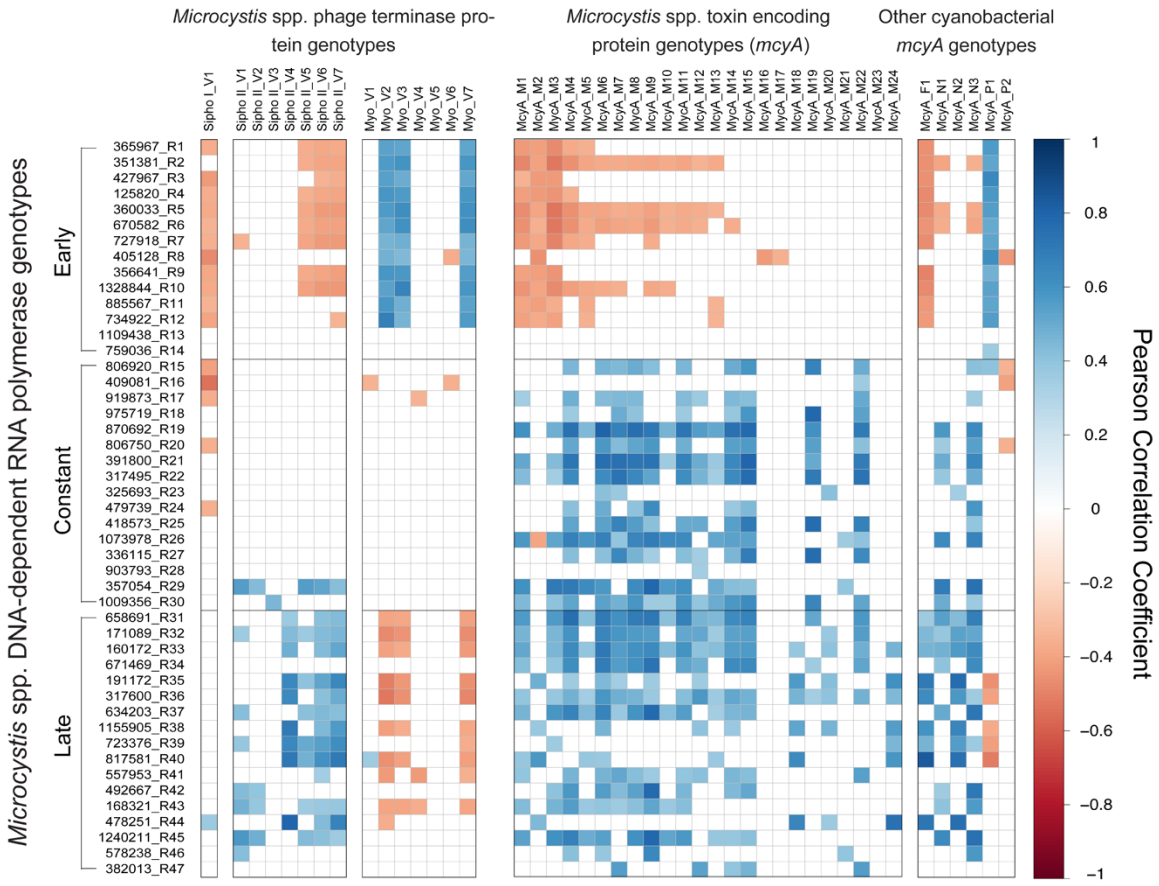


Figure 5. Virus and Host Correlations

Pearson correlation analysis heatmap of *Microcystis* DNA-dependent RNA polymerase (*rpoB*) genotypes correlated to *Microcystis* phage terminase genotypes and *Microcystis* toxin genotypes. Blue indicates a significant positive correlation between expression values and red indicates a significant negative correlation between expression values. White indicates no significant correlations.

Environmental variables

To determine why *rpoB* genotypes have resolved into early, constant, and late groups, we used a canonical correspondence analysis to orient the proportional expression of all *rpoB* genotypes in each of our 33 samples. CCA1 described 66.2% of the variation in our samples, while CCA2 described 6.2% of the variation. Of the environmental variables measured, sample month and pH were the most positively associated variables to CCA1 (Figure 6) (Supplemental Table 1B). Electrical conductivity, salinity, and total dissolved solids were also strongly associated with CCA1, in a negative direction. Total dissolved nitrogen and ammonium concentrations were the most highly correlated variables to the secondary axis, CCA2.

Discussion

Cyanobacterial harmful algal blooms are a growing concern around the globe, and thus how they manifest – both in terms of function(s) and who is carry out these functions (*i.e.,* which strains or species) is of critical importance to their management. The growing availability of genomic sequences has created opportunities for rapid assessment of environmental genomes and transcriptomes using approaches that involve "recruitment" (*i.e.,* mapping of the unknown environmental sequences to well-studied lab isolates). Yet, while these approaches have been valuable to date and taught us much about how blooms are constrained, they are dependent on the choice of microbial strain for comparison. To move beyond these limits, we have developed a workflow to characterize the subtle variations in diversity that are commonly overlooked within genera that cause these events. Given that lab strains of *Microcystis* represent a broad spectrum of genetic

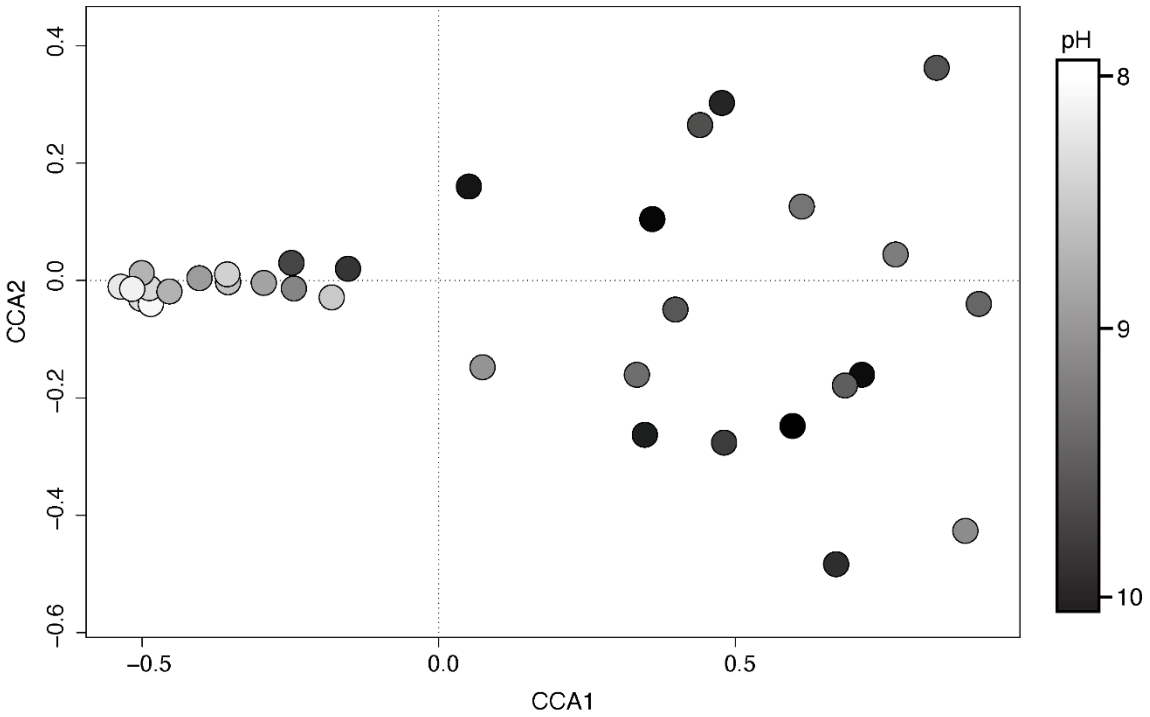


Figure 6. Canonical Correspondence Analysis

Canonical correspondence analysis of the proportional expression of all 47 *Microcystis* DNA-dependent RNA polymerase (*rpoB*) genotypes in each of our 33 samples. Color gradient indicates the pH of each sample.

potentials that could skew observations in the above approach, our approach creates an opportunity to more broadly capture spatial and temporal variation in *Microcystis* cell function *in situ* as multiple morphotypes and species are captured. When paired with parallel analyses of active viral infection of cells and toxin gene expression, a picture emerges that better resolves the variability that occurs in nature. The genotypic diversity present in Lake Tai reveals important features of *Microcystis* bloom dynamics in this system. Our findings suggest that the dominant species in the bloom are *M. flos-aquae* and *M. wesenbergii*, not *M. aeruginosa.* Our observations confirm other studies which employed PCR-based approaches to characterize toxic communities (Cai et al. 2012; Li et al. 2013; Otten and Paerl 2011) in concluding that *M. flos-aquae* and *M. wesenbergii* were common in Lake Tai. Our observations further show a relationship between the distribution of subsets of these strains, transcription of toxin-encoding genes in the *mcy* cassette and active infections by dsDNA-phage thought to target *Microcystis*. In examining these data, we demonstrate how environmental conditions - in this case pH - may play a role in promoting or constraining the interactions between bloom success, viral infection and the production of a potent hepatotoxin. Taken together our observation begin to shed light on the complex interactions that result in the proliferation of a different genera with a large cyanobacterial bloom.

Much knowledge in microbiology stems from the use of lab cultures to mimic conditions found in the environment. To date, most studies in both the lab and natural systems have been performed with *M. aeruginosa* due to the broad availability of isolates as well as its implicit role as a major cyanotoxin producer *in situ*. However, our data suggest these strains may not provide the most accurate representation of bloom biomass

in Lake Tai, nor bloom response to perturbations (Harke and Gobler 2015; Penn et al. 2014; Steffen et al. 2014b), as *M. aeruginosa* are only a subset (~ 7.3% of *Microcystis* spp. expression) of that community. *M. aeruginosa* NIES-843 was the first *Microcystis* species to have its genome fully-sequenced and closed (Kaneko et al. 2007), and therefore it has been the most widely applied genomic model for *Microcystis* blooms. As molecular biology becomes more broadly applied in natural systems, it is important to remember that an isolated reference organism may misrepresent the true diversity/function in a system. As our approach revealed 47 genotypes including at least 5 separate species of *Microcystis*, this work shows that going forward the use of a single reference genome could obscure the different patterns of seasonal expression we observed across genotypes.

Our analyses revealed that the *Microcystis* genotypes demonstrated different patterns of expression over the course of the bloom. There was a strict delineation between early and late genotypes, with additional genotypes occurring throughout the course of our sampling. These phases were initially characterized by Tang *et al*. (2018) as bloom "formation" and "maintenance". They suggested these temporal groups were associated with changes in nutrient utilization strategies, based on temporal changes in the expression of nitrogen and phosphorus transport genes that matched the temporal patterns we observed in genotype expression within the same dataset (Tang et al. 2018). From our observations, we now believe that the changes observed in nutrient utilization may have been, in part, a byproduct of the shifting genotypic composition. Unfortunately, it is impossible to separate cause and effect and it has become clear in recent years that nutrient concentrations measured in conjunction with biological parameters are as much

the residual (as opposed to the cause) of the biology that is present (Wilhelm et al. 2020). We note that the temporal shift in genotypic populations does not indicate a morphospecies succession pattern. All three temporal groups are comprised of multiple *Microcystis* species, with no definitive dominance of any particular morphospecies in a given temporal group (Figure 4). This suggests that the various species present might occupy parallel, yet different functional niches within the ecosystem, allowing for co-occurring species groups (Tilman et al. 1997). This further suggests that attributions of functions to morphospecies may be inexact.

In addition to the 47 *Microcystis* genotypes we observed, we also noted a wide diversity of *Microcystis* phages. In a parallel to efforts to recruit to *Microcystis* genomes, most literature pertaining to *Microcystis* phages refer only two isolated *Myoviridae* phages, Ma-LMM01 and MaMV-DC (Morimoto et al. 2018; Ou et al. 2013; Stough et al. 2017; Yoshida et al. 2006) and uses these as recruitment models. In the current study, our approach allowed us to detect these well-characterized *Myoviridae* phages, the newly characterized *Siphoviridae* phage, Mic1 (Yang et al. 2020), as well as a new *Siphoviridae* phage that we are confident infects *Microcystis* spp. This latter phage was originally "discovered" in our analyses of a genomic DNA scaffold of *M. aeruginosa* NIES-88, suggesting that it might have lysogenic potential (Supplemental Figure 3). Indeed, the genetic complement of this virus includes an integrase gene, which has not been observed in other *Microcystis* phages, although many have suggested the potential for lysogeny in *Microcystis* bloom systems (Stough et al. 2017; Yoshida et al. 2008). This NIES-88 phage portion does not appear to be a complete phage, indicating that it is likely remnant material. The presence of this phage in the host genome promotes questions regarding the

role of these viruses in horizontal gene transfer and the rate at which phage and host genomes exchange material: there is a definite precedent for this being a common occurrence in filamentous cyanobacteria (Martin et al. 2019).

Just as the gene expression of our host genotypes resolved into three temporal groups, our three main classes of *Microcystis* phages did the same. The early phase of the bloom (June and July) was characterized by and over-representation of the *Myoviridae* phage transcripts, which had been observed previously (Stough et al. 2017). Stough *et al*. used a single reference genome as bait to identify phage activity, yet in our effort we have shown that there were 7 genotypes present. The presence of multiple *Myoviridae* genotypes has been observed before *via* real-time PCR and might suggest rapid co-evolution with the host, or standing diversity (Kimura et al. 2013). While Ma-LMM01 and MaMV-DC were both originally isolated on *M. aeruginosa*, recent analysis of MaMV-DC indicates that it can limit the growth of *M. wesenbergii* and *M. flos-aquae*, although it did not form plaques (Wang et al. 2019). This suggests that the *Myoviridae* genotypes we observed may infect several *Microcystis* strains / species that were prominent in the spring. When the *Myoviridae* is highly expressed with the early host genotypes, infection by the two *Siphoviridae* groups are nearly absent. In fact, the single Sipho I genotype was not expressed at all during June or July, and four of the seven Sipho II genotypes were not expressed at all during June. Sipho II was much more highly correlated to the late host genotypes, when there was a decrease in lytic *Myoviridae* expression. This shift could be explained in a number of ways. The genotypic shift in hosts may reflect a shift in host susceptibility or viral specificity (Kimura et al. 2013) or a successional overturn, either through density-dependent infection or lysogenic

reproduction (Murray and Jackson 1992; Stough et al. 2017). However, it is important to note that Stough *et al.* observed an increase in genes associated with myocyanophage-lysogeny during the later months, when the late genotypes are expressed (Stough et al. 2017). It is possible that the *Siphoviridae* phages only become active once the *Myoviridae* phage have shifted into a lysogenic reproductive cycle. Or it might be that the shift to a lysogeny for the *Myoviridae* might prevent superinfection, causing a decrease in the expression of lytic *Myoviridae.* In any case, it is unclear whether the host or the viruses are driving the temporal patterns we observe.

The early host genotypes that strongly correlated to *Myoviridae* expression also strongly correlated (negatively) to the expression of the toxin coding *mcy*A gene. The constant and late genotypes are positively correlated with the expression of the toxin gene, even though actual toxin levels seem to decrease at the warmer temperatures typical of September and October (Peng et al. 2018). Many studies have attempted to characterize genotypes and how toxin genes change in copy abundance over the course of a bloom (Kim et al. 2010; Kurmayer and Kutzenberger 2003; Rinta-Kanto et al. 2009). Expression of *mcyA* occurs throughout the bloom, although there is an increase in the later months, providing a positive correlation to both the constant and late genotypes (Supplemental Figure 1). We note that again there is a disconnect between phylogeny and temporal patterns, given the diverse composition of *mcyA* taxonomic groups. Microcystin production is thought to rarely result from a single species or strain, but likely is a product of many co-occurring organisms (Rinta-Kanto et al. 2009; Yu et al. 2014; Zingone and Enevoldsen 2000). Indeed, Otten and Paerl warn against using *Microcystis* species type to estimate toxicity, because the co-occurring presence of many species of

various toxicities can obscure toxicity associations with particular species (Otten and Paerl 2011). Even the expression of the *mcyA* toxin-encoding gene (or any gene in that cassette) is not a guarantee that an organism is capable or actively producing microcystin, as often other components of the cassette can be missing (Tillett et al. 2001).

Although toxin gene expression and viral infection show similar temporal patterns, this does not suggest that they are directly related or regulated. More than likely, both observations are a function of cellular processes responding to similar external perturbations. Given that taxonomy is not capable of describing host activity, viral infection, or toxin gene expression, it is logical to hypothesize that the environment must be selecting for the genotypes we observed during various bloom phases. While previous studies have proposed that time of day could influence phage expression levels, we did not observe any pattern associated with diel cycling (vis a vis Morimoto et al. 2019). Our analyses revealed that one the primary environmental factor associated with the temporal shift in host *rpoB* expression was pH (Figure 6). System pH during cyanobacterial blooms is thought to increase due to the consumption of water column dissolved carbon dioxide. In parallel with this, pH is thought to influence *Microcystis* spp. carbon concentrating mechanisms, with evidence that *Microcystis* spp. are able to maintain growth at higher pH (Krausfeldt et al. 2019a; Visser et al. 2016). It is unclear how pH may be influencing toxicity or viral infection, although we hypothesize that linkages to photosynthesis, oxidative stress and nutrient acquisition are all potentially involved (Krausfeldt et al. 2019a). Electrical conductivity and salinity were also strongly associated with the temporal shift in host *RpoB* expression. These two environmental parameters are highly self-correlated, representing the presence of sodium ions in the

water. Combined with the strong association of total dissolved solids, we posit that this represents the presence of external terrestrial loading or resuspension, a hypothesis previously suggested by Wilhelm *et al.* (Wilhelm et al. 2011). This could also help explain why we do not observe a strong association with our nutrient parameters such as nitrogen or phosphorus. They are likely being turned over quickly by the biological population, particularly during the early months of the productive bloom, whereas sodium ions would remain in the water column.

Overall, our observations have allowed us to identify and quantify the abundant host and viral genotypes present in a freshwater cyanobacterial bloom system across a temporal profile. A major observation here is that the "bloom" does not appear to be a single genotype of one organism, but at least 47 different phylotypes that come from multiple species of the same genus. Temporal shifts in active viral infection (including transitions in the type of virus) and the expression of a key gene for toxin production confirm that the metabolism of *Microcystis* is likely regulated by a complex interaction with environmental drivers (Steffen et al. 2014b) and that these drivers influence the way researchers interpret bloom dynamics (Wilhelm et al. 2020). Our study highlights the complexity of bloom systems, and how environmental factors can vary and relate to similar organisms in different ways.

Data Availability

Raw sequencing data is publicly available in the MG-RAST database under the name "Lake_Taihu_metatranscriptome_project (Keegan et al. 2016)." Sequencing information and quality control details have been previously reported (Tang et al. 2018).

Assembled data is also in the MG-RAST database under the name “Taihu 2014 MegaHit Assembly 1.”

Acknowledgements

We thank Gary LeCleir, Robbie Martin, Eric Gann, Naomi Gilbert, Brittany Zepernick, Liz Denison and Gwen Stark for feedback and conversations on this topic. This work was supported by funds from the National Science Foundation (IOS- 1451528 to SWW), and funding from the NIEHS (1P01ES028939–01) and NSF (OCE-1840715) to SWW through the Great Lakes Center for Fresh Waters and Human Health at Bowling Green State University. We also acknowledge support from the Simons Foundation (735077) and the Kenneth & Blaire Mossman Endowment to The University of Tennessee. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CHAPTER 3: THE "NEGLECTED" VIRUSES OF *TAIHU*: ABUNDANT TRANSCRIPTS FOR VIRUSES INFECTING EUKARYOTES AND THEIR POTENTIAL ROLE IN PHYTOPLANKTON SUCCESSION

Publication Note

This chapter is a version of a peer-reviewed article that was published in Frontiers in Microbiology by Helena L. Pound, Eric R. Gann, Xiangming Tang, Lauren E. Krausfeldt, Matthew Huff, Margaret E. Staton, David Talmy and Steven W Wilhelm. Xiangming Tang and Lauren E. Krausfeldt were responsible for sample collection and processing. Eric R. Gann, Matthew Huff and Margaret E. Staton were responsible for assisting with bioinformatic analysis. Helena L. Pound was responsible for primary data analysis and composition of the article with input from David Talmy and Steven Wilhelm.

Citation:

Pound HL, ER Gann, X Tang, LE Krausfeldt, M Huff, ME Staton, D Talmy, and SW Wilhelm (2020). The "neglected viruses" of *Taihu:* Abundant transcripts for viruses infecting eukaryotes and their potential role in phytoplankton succession. Frontiers in Microbiology 11:338. DOI: 10.3389/fmicb.2020.00338

Abstract

Drivers of algal bloom dynamics remain poorly understood, but viruses have been implicated as important players. Research addressing bloom dynamics has generally been restricted to the virus-infection of the numerically dominant (*i.e.,* bloom forming) taxa. Yet this approach neglects a broad diversity of viral groups, limiting our knowledge of viral interactions and constraints within these systems. We examined hallmark virus marker genes in metatranscriptomic libraries from a seasonal and spatial survey of a *Microcystis aeruginosa* bloom in Lake Tai (*Taihu*) China to identify active infections by nucleocytoplasmic large DNA viruses (NCLDVs), RNA viruses, ssDNA viruses, bacteriophage, and virophage. Phylogenetic analyses revealed a diverse virus population with seasonal and spatial variability. We observed disproportionately high expression of markers associated with NCLDVs and ssRNA viruses (consistent with viruses that infect photosynthetic protists) relative to bacteriophage infecting heterotrophic bacteria or cyanobacteria during the height of the *Microcystis* bloom event. Under a modified kill-the-winner scheme, we hypothesize viruses infecting protists help suppress the photosynthetic eukaryotic community and allow for the proliferation of cyanobacteria such as *Microcystis*. Our observations provide a foundation for a little considered factor promoting algal blooms.

Introduction

Harmful algal blooms (HABs) are extreme biological events that have detrimental environmental and socioeconomic effects on both fresh and salt waters. In particular, freshwater HAB events during the last two decades have posed threats to potable water supplies and socioeconomic resources for population centers (Bullerjahn et al. 2016; Qin et al. 2010; Steffen et al. 2014a). Freshwater HABs commonly manifest as a summer bloom of cyanobacteria that follows from a winter/spring population of eukaryotic phytoplankton (Edgar et al. 2016; Ke et al. 2008; Niu et al. 2011). HAB research has historically focused on abiotic "bottom-up" controls, such as nutrient or temperature triggers that stimulate cyanobacterial growth (Tang et al. 2018). However, biological success can also arise during suppression of competitors, and there is less known about the decline of the eukaryotic population and factors that constrain these populations as cyanobacterial blooms form. And while grazing is often cited as the major driver of "top-down" regulation on communities (Vanderploeg et al. 2001), viruses also act as predators of algae.

Research over the past two decades has made it clear that viruses are central to ecosystem function (Wigington et al. 2016). There are many cases of direct viral modulation of phytoplankton populations in the literature, but these focus almost entirely on the lysis of dominant bloom-forming species (Brussaard et al. 2005; Hewson et al. 2001; Steffen et al. 2017; Tomaru et al. 2004; Wilson et al. 2002). In particular, there is less focus on numerically less-abundant protists that may be subject to virus-mediated suppression. In part this may be due to the general consensus that viruses of prokaryotes

are the most abundant and active in aquatic systems (Weinbauer 2004) or that many of the viruses infecting protists contain RNA genomes (Moniruzzaman et al. 2017) and are thus not detected by DNA-based metagenomics. These "neglected" viruses have the potential to control community dynamics, reducing the success of competitors. The "kill-the-winner" model invokes this important role of viral suppression of competitive hosts in microbial community assembly (Winter et al. 2010): however, the kill-the-winner model assumes equilibrium dynamics, and is historically thought of in the context of heterotrophic bacteria (Thingstad 2000). Nevertheless, suppression of competing hosts by viruses, consistent with kill-the-winner, could be important in the selection of dominant, bloom-forming phytoplankton.

To explore the role of the "neglected" viruses, we examined metatranscriptomic data from a Lake Tai (China, or *Taihu* in Mandarin) bloom event during 2014. Lake Tai experiences yearly successions from photosynthetic protists to cyanobacteria (dominated by *Microcystis*) and therefore presents an opportunity to study the impact of viruses during and after that transition (Supplemental Figure 4) (Ke et al. 2008; Niu et al. 2011). We assessed gene markers (Moniruzzaman et al. 2017; Stough et al. 2018) for active infections by NucleoCytoplasmic Large DNA Viruses (NCLDVs), ss/ds RNA viruses, ssDNA viruses, as well as bacteriophage that were infecting non-*Microcystis* populations. In parallel, we characterized the putative host community structure. We detected a broad spatial and temporal richness of viruses, including a shift between the two dominant virus types (*Microcystis*-infecting phage and ssRNA viruses). In addition to observations of active eukaryotic viruses, determinations of putative host diversity and statistical co-

occurrence analyses of "*who infects whom*" reveal potential advantages for *Microcystis* that account for part of its ability to form large bloom events in fresh waters.

Methods

Sample Collection

Samples were collected monthly (June to October) from a *Microcystis* spp.-dominated bloom in Lake Tai, China in 2014 at several stations across the northern parts of the lake as described previously (Krausfeldt et al. 2017; Stough et al. 2017). Biomass was collected by filtering between 25 and 180 mL (volume dependent on sample biomass density) of lake water through 0.2-µm nominal pore-size Sterivex™ filters (EMD Millipore Corporation, Burlington, Massachusetts) and immediately treated with ~2 mL of RNA*later* (Qiagen) for preservation. Samples were collected in a manner consistent with the separation of bacteria and phytoplankton from dissolved materials: free virus particles generally pass through these filters. Environmental parameters and nutrient data were collected with each sample and reported elsewhere (Tang et al. 2018). Total RNA was extracted, quality checked, ribosomally-reduced, and sequenced on the Illumina™ HiSeq platform at *Hudson Alpha Institute Genomic Services Laboratory* (Huntsville, AL) as previously described (Krausfeldt et al. 2017; Stough et al. 2017; Tang et al. 2018). Raw sequence data was retrieved from the Hudson Alpha servers and primarily handled in the CLC Genomics Workbench v. 10.1.1 suite (QIAGEN, Hilden, Germany). Quality scoring was used to remove reads with a score < 0.3. Trimmed reads from all 35 samples were combined into a single file and assembled using MEGAHIT to reduce redundancy that can be seen within different contig fragments from identical genes (Li et al. 2015).

Viral Detection

We employed a marker gene-based approach to discover viruses within the *Taihu* samples (Moniruzzaman et al. 2017) (Stough et al. 2018). This protocol is publicly available (Pound and Wilhelm 2019). A reference protein database was created from hallmark viral genes from a diverse range of taxonomic groups with sequences downloaded from the NCBI refseq database and Uniprot. Hallmark genes included the NCLDV major capsid protein (25 sequences), RNA virus RNA-dependent RNA polymerase (79 sequences), ssDNA viral replicase (11 sequences), and bacteriophage ribonucleotide reductase (1347 sequences) from known viral isolates. For virophage major capsid proteins, 10 sequences from virophage isolates as well as 6 from well-studied metagenomic assemblies were used. To classify hosts we employed 3,112 RpoB-Rpb1 sequences. All data used to establish base phylogenies are available (Table 1, Supplemental Table 3), and were selected to capture the known diversity in each sample (or as much known diversity as was available in databases). Several marker genes were used to discover bacteriophage including gp20, gp23, integrase, and ribonucleotide reductase, with the results of the latter presented in this paper. The other marker genes revealed no additional phage, nor did the phage finding tool, VirSorter (Roux et al. 2015). DNA-directed RNA polymerase beta subunit (Rpb1/RpoB) was used as a gene marker to identify potential eukaryotic and prokaryotic hosts (Table 1, Supplemental Table 3). Assembled libraries were queried with each virus-specific reference protein database in command line BLASTx with an e-value maximum of e^{-10} for viral candidates and e^{-30} for host candidates. A lower stringency was applied to potential viral candidates given the limited number of viral isolates in reference databases. Our separate stringency cutoffs

Table 1. Marker Gene Sources

Gene markers used in BLASTx for each virus type with reference. All sequences from pfam domains were viral.

	Marker gene	Reference used
NCLDVs	Major capsid protein	NCBI refseq manually curated (modified from Moniruzzaman *et al.* 2017).
RNA viruses	RNA-dependent RNA polymerase	PF00680 v. 31
ssDNA viruses	Replicase/helicase	PF00680 v. 31
Virophage	Major capsid protein	NCBI refseq manually curated (Moniruzzaman *et al.* 2017).
Bacteriophage	Gp20, gp23, integrase, ribonucleotide reductase	PF06810, PF07068, PF00589, PF00317
Hosts	DNA-directed RNA polymerase	k03043 and k03006 (Kegg orthology)

for virus and hosts attempt to eliminate the implicit bias towards broader diversity in the hosts, given the larger database. A hidden Markov model search was used to validate our BLASTx method but yielded 11 fewer viral candidates and provided no new candidates (data not shown). A custom python script was used to identify the coordinates of the aligned sequence within the query contig and extract only this region: these extracted components of the contigs were subsequently considered "viral candidates" (Gann et al. 2019). Only viral candidates greater than 150 nucleic acids were kept for further analyses. All remaining candidates were then queried against the NCBI refseq database, and any candidates with non-viral hits were removed from the analysis and considered false-positives. In preliminary work, we found that candidates less than 150 nucleic acids did not always produce positive BLASTx hits to viral sequences in the refseq BLASTx. Candidates were considered to represent "near complete" genes if they were larger than the smallest reference protein in length. Shorter candidates were considered and are discussed as "gene fragments".

RNA virus candidates were further analyzed using the pfam v. 29 database to identify ORFs in the full-length contigs (contigs before extraction of target gene sequence). Contigs containing both structural and non-structural components were considered "near complete genomes" (data not shown) (Moniruzzaman et al. 2017).

Viral Phylogeny and Expression

The hallmark gene reference proteins and "near complete" gene candidates were used to reconstruct maximum likelihood phylogenies as described in Moniruzzaman *et al*. 2017 (Supplemental Table 4). Gene fragment candidates were placed on trees using pplacer and subsequently visualized in iTOL v.4 (Letunic and Bork 2019; Matsen et al.

2010). All reference proteins used to reconstruct phylogenies came from viruses that have been isolated in labs and sequenced, with the exception of some virophage sequences. *Mavirus*, *Sputnik,* and *Zamilon* remain the only virophage that have been isolated, purified and sequenced to date (Fischer and Suttle 2011; Gaia et al. 2014; La Scola et al. 2008), while three *Chrysochromulina parva* Virus Virophages (CpVVs denoted Curly, Larry, and Moe) have been isolated but not purified beyond unialgal lysates (Stough et al. 2019). Several intact virophage have been assembled from metagenomic analyses and were employed here to make our phylogenetics more robust (Zhou et al. 2015; Zhou et al. 2013).

Relative virus and host activity in each sample was quantified by mapping trimmed reads from each sample library to the extracted viral or host candidates using a 90% similarity fraction over a 90% length fraction in CLC Genomics Workbench 10.0 (Supplemental Table 5). The abundances for each viral type therefore only represent the expression of the singular target gene (and not other components of the contig). Transcript abundance data were normalized by the length of the extracted viral portion of the candidate contig and by the library size of that sample. Viral transcript expression was visualized using Heatmapper (Babicki et al. 2016). A non-metric multidimensional scaling (nMDS) plot using Bray-Curtis dissimilarity was generated in Primer7 and used to describe the influence of environmental factors on the abundance data (Clarke and Gorley 2015).

Cluster Analyses

We worked from the assumption that virus transcripts in cells represent active infections and that viruses require a host population to be present to generate these

transcripts. Pearson correlation coefficients were established from square root transformed normalized transcript abundances using Primer 7 (Clarke and Gorley 2015). A SIMPROF test (α=0.5) was used to identify significant clusters of co-occurrences. Significant clusters containing both a putative host and putative virus candidate with correlation coefficients greater than 0.8 were visualized in Cytoscape (Smoot et al. 2011) with the nodes representing the viral or host candidate and the edges representing the strength of the correlation.

Results and Discussion

Our analyses revealed an active and diverse virus community in the *Microcystis*-dominated harmful algal bloom that occurred in Taihu during the 2014 summer sampling season. In total 8.42×10^8 transcripts across 35 samples collected during the bloom season passed QA/QC assessments. Across all stations ~70% of the transcripts were designated as of cyanobacterial origin (Tang et al. 2018), highlighting the "bloom" nature of this data set. Of the residual transcripts, $\sim 1.92 \times 10^5$ were associated with viruses. Phage assigned to groups known to infect *Microcystis* hosts comprised 47.9% of the total viral reads. The remaining 52.1% consisted of transcripts associated with a broad diversity of viruses. We assigned 827 unique viral candidate contigs to groups that included the NCLDVs, single-stranded and double-stranded RNA viruses, ssDNA viruses, non-*Microcystis* phage, and virophage.

The NCLDV group displayed the most richness (*i.e.,* number of unique viral contigs) amongst the identified viral groups, with 457 candidates assigned to families (Figure 7A). The majority of these candidates were assigned to the "Extended-

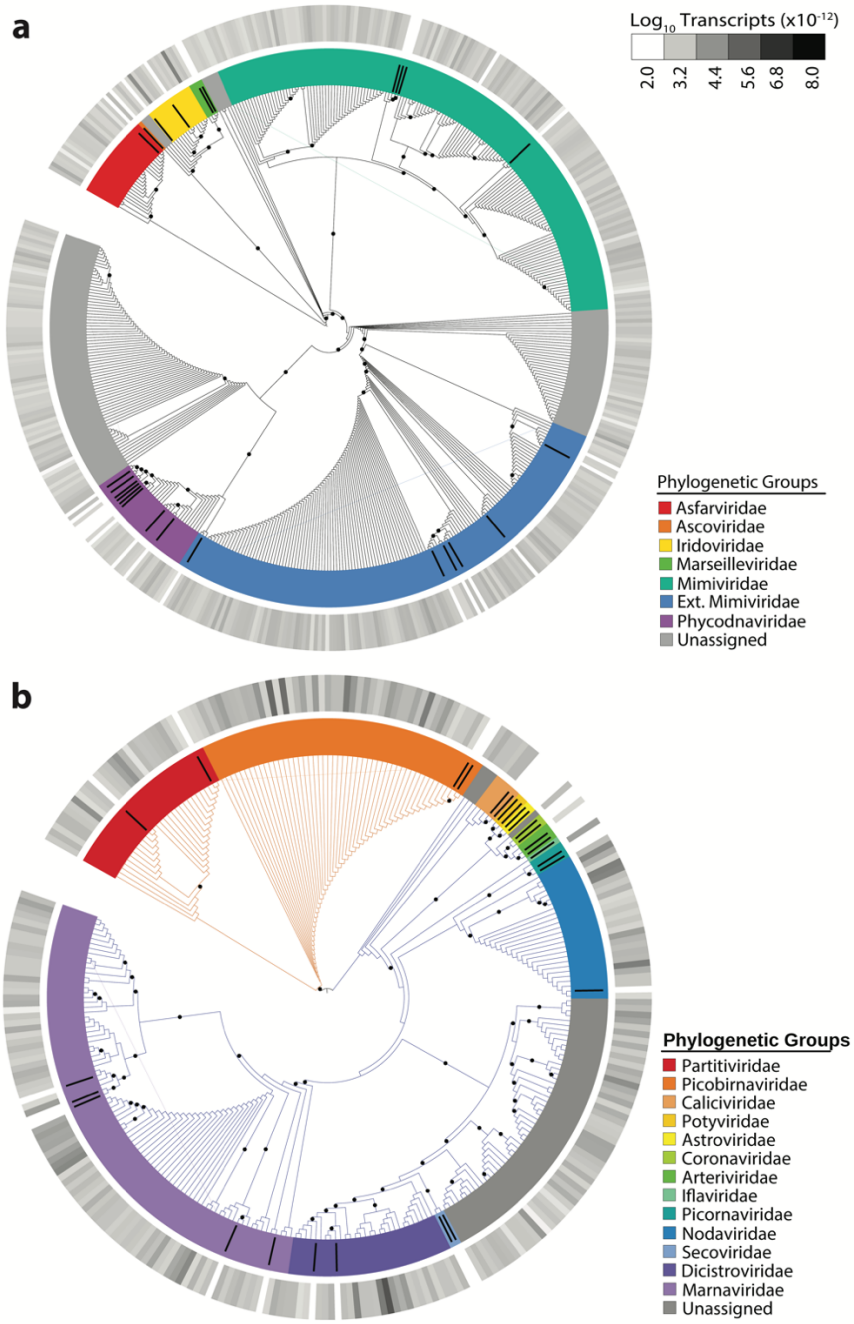


Figure 7. NCLDV and RNA Virus Phylogenic Trees

Maximum-likelihood phylogenetic placement of NCLDV candidates (A) and RNA virus candidates (B). Reference proteins are denoted by black bars with colors indicating viral family (Supplemental Material Table 2). Orange branches indicate dsRNA viruses and blue branches indicate ssRNA viruses (B). The outer heatmap ring indicates normalized transcript abundance for each candidate summed for all samples. Black dots represent bootstrap values >0.5.

Mimiviridae" or the *Mimiviridae* clades, which have been associated to date with many marine algae (Claverie and Abergel 2018). The most abundant NCLDV candidate contigs were assigned to the *Iridoviridae*, a NCLDV family whose members infect insects and cold-blooded vertebrates (Chinchar et al. 2009). Overall, NCLDVs made up 8.35% of the total reads in the virus dataset.

We also observed virophage signatures in our dataset, although with less richness (30 unique candidates) than the NCLDVs (Supplemental Figure 5). Most candidates were more closely related to an environmentally-derived sequence identified as the Dishui Lake Virophage (Gong et al. 2016). Transcripts associated with virophage were the least abundant in our dataset, making up only 0.25% of the total viral reads and <0.0001% of the total reads.

The use of a metatranscriptomic data from size-selected (> 0.2 μm) populations allowed us to detect infections by both DNA and RNA viruses, the latter of which should primarily be cell-associated. We detected 268 unique RNA virus candidates (Figure 7B). Overall, RNA viruses comprised 42.45% of the total virus transcripts, rivaling the 47.9% made up by *Microcystis* phage. Of the RNA virus candidates, 72% of the expression was assigned to ssRNA viruses and 28% to dsRNA viruses. The distribution between the number of single-stranded and double-stranded candidates was coincidentally identical, with 193 ssRNA virus candidates representing 72% of the richness and 75 dsRNA virus candidates representing 28% of the richness (Figure 7B). The dsRNA virus contigs belonged to *Partitiviridae* and *Picobirnaviridae.* Most of our ssRNA virus candidates belonged to the *Picornavirales* order, with the majority belonging to the *Marnaviridae* or *Dicistroviridae* families. *Marnaviridae* viruses*,* whose members include the *Chaetoceros*

tenuissimus virus and the *Rhizosolenia setigera* virus, commonly infect marine diatoms (Nagasaki et al. 2004; Shirai et al. 2008). In our dataset, *Marnaviridae*-like viruses represent 12% of the total RNA virus reads, while *Dicistroviridae*-like viruses, which are typically associated with insect hosts (Valles et al. 2017), comprise 43% of our total RNA virus reads. Interestingly, the majority (86%) of the *Dicistroviridae*-associated transcripts come from a single candidate, which was highly expressed in most of our samples.

We took the opportunity to examine the RNA virus contigs to determine if any represented near or full-length virus genomes. Of our original 268 contigs, 16 were determined to contain a "near full-length" genome based on our parameters (Supplemental Figure 6). All of these viruses were single-stranded, and most belonged to *Marnaviridae* or *Nodaviridae.* Each meta-assembled virus contained both an RNA-dependent RNA polymerase and a structural gene, most commonly a capsid protein specific to the phylogenetic group.

As part of our classification efforts, we also observed evidence for both ssDNA viruses and dsDNA bacteriophage not associated with *Microcystis* spp. ("other" phage). However, neither group contributed much to the overall transcript abundance. The ssDNA viruses had only 17 unique candidates, which comprised only 0.42% of the total viral reads or < 0.0001% of the total reads (Supplemental Figure 7). Eight candidates belonged to *Nanoviridae,* which typically infect plants. Ribonucleotide reductase markers indicated the presence of 14 unique candidates that were closely related to *Microcystis* phage. The remaining, "other," phage had only 41 unique candidates, representing 0.63% of the total viral reads and 0.001% of the total reads (Supplemental Figure 8). Most of the

"other" phage belonged to *Caudovirales*, with several candidates most closely related to other cyanophages, such as *Synechococcus* phage.

In addition to the virus community, we examined eukaryotic and bacterial members of the putative host community. With 1,946 unique candidates, we observed a vast richness that spanned cellular phylogeny (Figure 8). Host transcripts associated with DNA-dependent RNA polymerase beta-subunit made up 0.06% of the total reads in the dataset. Candidates closely related to *Microcystis* spp. were the most abundant, with 52 unique candidate sequences representing 46% of total host reads.

Our dataset also provided the opportunity to investigate variation in virus communities (based on transcript abundance for individual groups) over the course of the five-month bloom. Figure 9 details the read abundances of each viral group we targeted, with each column representing a different sample and each row representing a different category of virus (alternative color scheme Supplemental Figure 9). The colors indicate the sum of viral transcripts for each viral type in each sample, normalized to individual library size. In some samples, the ssRNA viruses showed similar proportions to the *Microcystis* lytic and lysogenic phage infections, suggesting that these viruses might be just as active. Of these 35 samples, markers for *Microcystis* phage lytically infecting their host had the most abundant representation in 10 samples, *Microcystis* phage in a lysogenic infection had the most abundant representation in 15, ssDNA viruses had the most abundant representation in 9, and dsRNA viruses had the most abundant representation in one sample (Supplemental Figure 10). According to a Pearson correlation analysis, this pattern was best explained from the environmental data sets (Tang et al. 2018) by pH and salinity (Supplemental Figure 10A).

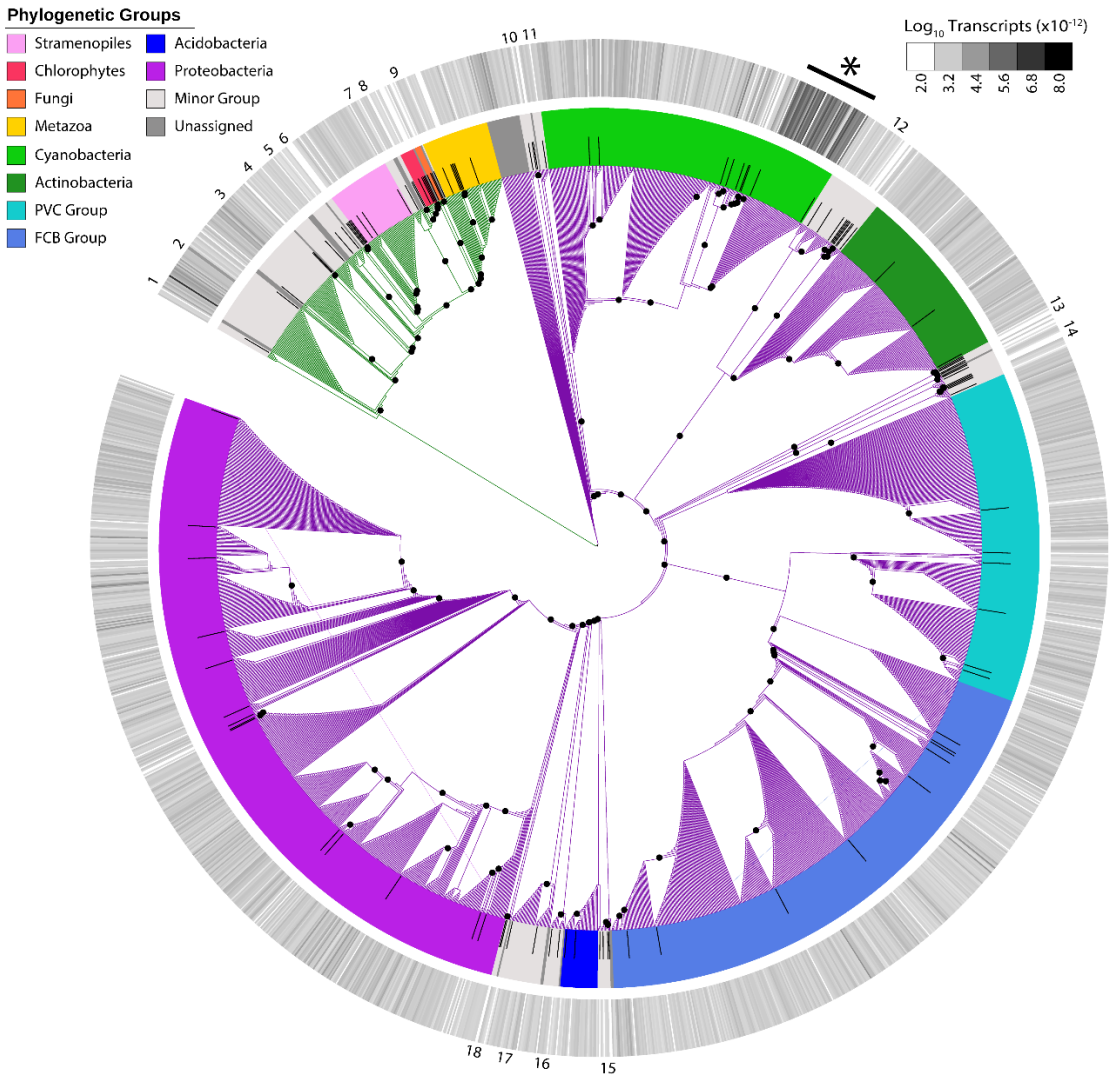


Figure 8. Host Phylogenetic Tree

Maximum-likelihood phylogenetic placement of host candidates. Reference proteins are denoted by black bars. Green branches indicate eukaryotic hosts and purple branches indicate bacterial hosts. Colored ring indicates viral family and the outer heatmap ring indicates normalized transcript abundance for each candidate summed for all samples. Numbers on outer edge represent collapsed reference groups (Supplemental Material Table 4). Black dots represent bootstrap values >0.5. **Microcystis*-like candidates.

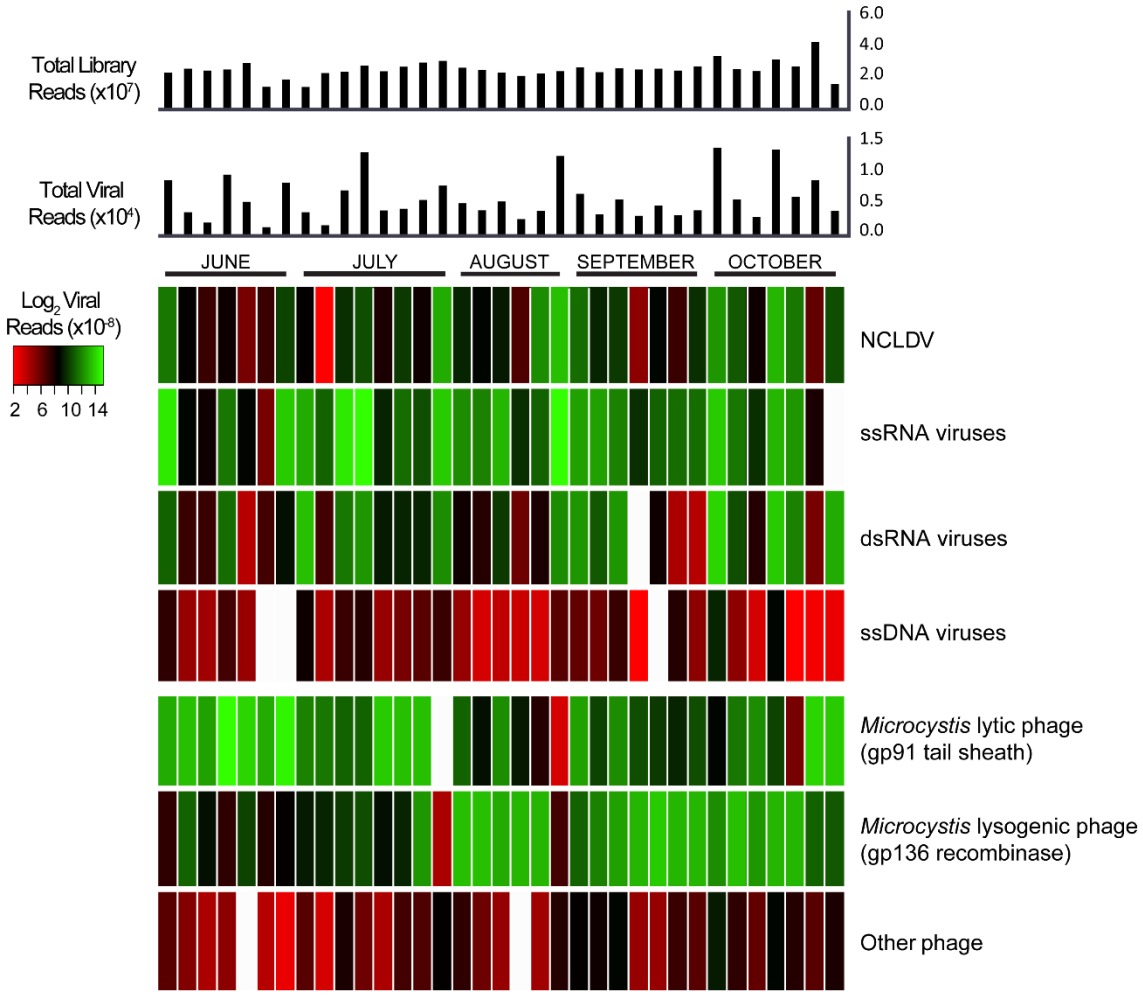


Figure 9. Viral Expression Heatmap

Heatmap of normalized viral abundances for each sample and virus type with histogram of total viral reads and total library reads in each sample. Each column describes a single sample. Color scale describes viral read abundance normalized to library size.

Co-occurrence networks described putative virus and host interactions across our 35 samples. We established 20 significant co-occurrence clusters that ranged in membership from two to eleven members, with each network containing at least one putative virus and putative host. Twelve clusters contained a bacterial host with an RNA virus or a NCLDV, while four clusters contained a eukaryotic host with an RNA virus, a NCLDV, or ssDNA virus (Supplemental Table 6). Three clusters had both bacterial and eukaryotic hosts with a combination of virus types (Figure 10). An additional three clusters contained a virophage but no NCLDV, which are typically thought to co-infect eukaryotic host cells. In all three cases, the virophage co-occurred with an RNA virus (Supplemental Table 6).

The gene marker approach used in this study allowed us to characterize the viral community within a cyanobacterium-dominated harmful algal bloom. This technique extended existing virus-finding programs, such as VirSorter or VirFinder, powerful tools which are currently limited to phage and metagenomic analysis (Ren et al. 2017; Roux et al. 2015). While a similar approach has been used to study marine eukaryotic phytoplankton systems and even *Sphagnum* peat bogs (Moniruzzaman et al. 2017; Stough et al. 2018), it has not been used to explore freshwater cyanobacterial blooms, to our knowledge. Our results are consistent with the *Aureococcus*-dominated HAB transcriptome analyzed by Moniruzzaman *et al.* in relation to the presence of both NCLDVs and RNA viruses (Moniruzzaman et al. 2017). Both studies showed abundant *Mimiviridae* and extended *Mimiviridae* group members, although Moniruzzaman *et al.* did not distinguish between the two groups. Our results also showed a similar, smaller representation of *Asfarviridae* and *Phycodnaviridae* (Moniruzzaman et al. 2017). Our

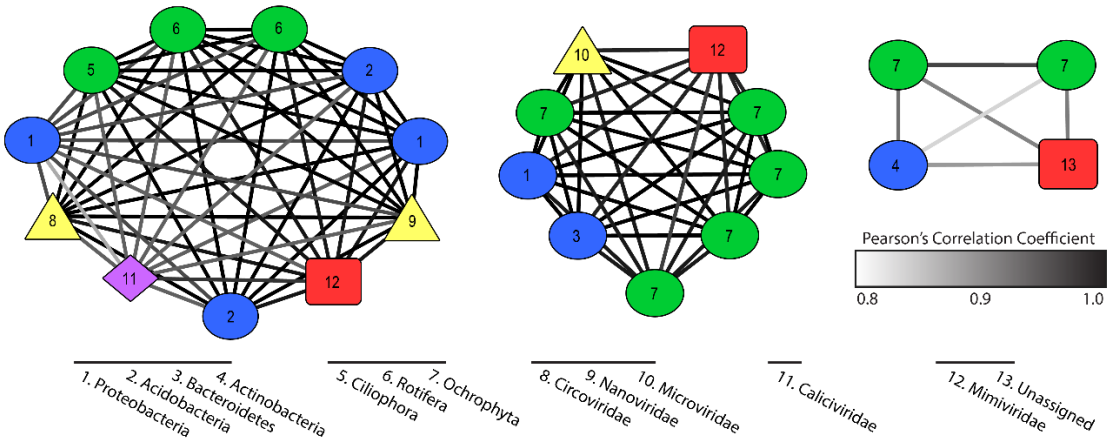


Figure 10. Virus Host Co-occurrence Clusters

Co-occurrence clusters that contain both eukaryotic and bacterial hosts with one or more viruses. Colored symbols indicate host or virus type and edges describe strength of Pearson's correlation. Numbers describe phylum identity of hosts and family identity of viruses.

ssRNA virus results are consistent with previous work that observed putative transcript / viruses in this group, although several of *Marnaviridae* viruses were labeled "unclassified marine viruses" in the previous publication (Moniruzzaman et al. 2017). That paper did not discuss dsRNA viruses, so it is unclear if our observed abundance (28% of RNA viral reads) of *Picobirnaviridae* and *Partitiviridae* is unusual. We also observed many virophage candidates, most of whom were closely related to the Dishui Lake virophage meta-assembly (a lake approximately 60 miles from our study site) (Gong et al. 2016).

Previously the viral component of metatranscriptomic analyses of *Microcystis* blooms have focused on the *Microcystis* phage (Steffen et al. 2017; Stough et al. 2017). However, our findings indicate that gene markers for RNA viruses, particularly ssRNA viruses, can reach similar transcript abundances during a bloom event. In many of our early season samples (June and July), we see higher relative abundances of ssRNA virus transcripts than the *Microcystis* lytic phage. These viruses are mostly members of *Marnaviridae,* whose members can infect photosynthetic eukaryotes including *Chaetoceros tenuissimus, Rhizosolenia setigera, and Heterosigma akashiwo* (Nagasaki et al. 2004; Shirai et al. 2008). This suggests that these "neglected" viruses might play a role in the transition from a winter diatom bloom to the summer cyanobacterial blooms. Seasonal successions of this nature have been well-recorded in Lake Erie, another shallow lake plagued by summer *Microcystis* blooms (Saxton et al. 2012). Other virus-driven community controls have been hypothesized to occur in Lake Erie, with viral infection of parasitic fungal hosts hypothesized to contribute to the maintenance of winter diatom blooms (Edgar et al. 2016). In both scenarios, viral suppression of a predator or competitor enables a particular species or group to thrive.

A virus-mediated transition from diatoms to cyanobacteria can be explained within the context of a modified version of the "kill-the-winner" framework. As originally proposed by Thingstad and Lignell (Thingstad and Lignell 1997), the "kill-the-winner" hypothesis posits that competition specialists – groups that would typically achieve high abundance due to abilities to assimilate resources and grow faster – are held at low densities when top-down pressure is strong (Thingstad and Lignell 1997). This theory has been invoked to explain why heterotrophic bacteria do not outcompete phytoplankton when mineral nutrients, but not carbon, are limiting (Thingstad and Lignell 1997). It has also been widely invoked as a mechanism to maintain diversity among bacteria (Kirchman 2016; Winter et al. 2010), and it has been suggested that this framework is not limited to heterotrophic bacteria and might be applied to other co-occurring species in aquatic systems (Våge et al. 2013). Yet, it is not widely thought of in the context of phytoplankton bloom dynamics.

Silica may act as the constraining nutrient that catalyzes the transition of the community. Reductions in silica deposition are hypothesized to occur in association with increased pH conditions during *Microcystis* blooms, constraining diatom growth (Krausfeldt et al. 2019a). In marine environments, silica limitation has been linked to increases in viral infection of diatoms (Kranzler et al. 2019). Acting together, nutrient limitation and viral infection could act as a negative feedback loop on diatoms that might provide an advantage for competitive species such as *Microcystis*.

The kill-the-winner hypothesis assumes equilibrium dynamics, and therefore is not implicated (or perhaps even appropriate) in selection during transient phytoplankton blooms. Bloom forming species are often thought of as opportunists, able to take

advantage of ephemeral increases in resource availability, predominately through rapid cell division (Grover 1990; Litchman and Klausmeier 2001). In classical quantitative ecology theories, feedbacks on the nutrient environment and seasonal changes in environmental conditions control successional bloom dynamics, from opportunists to other groups (Margalef 1978). In Taihu, diatoms are the opportunists. Abundant viruses consistent with those that infect dominant groups like *Chaetoceros tenuissimus* and *Rhizosolenia setigera* imply that early diatom dominance may be terminated and subsequently suppressed by virus-mediated lysis. Collapse of diatoms would lead to opportunities for other groups, especially groups able to resist predation and infection. In this interpretation of the "kill-the-winner" hypothesis (Thingstad and Lignell 1997), the potential for virus-induced cell lysis of the "neglected" community members could indirectly allow for cyanobacterial bloom species to achieve dominance.

The latter season dominance of *Microcystis* depends on an ability to both consume the available nutrients and itself resist viral infection. One mechanism of *Microcystis* resistance is through high number of restriction modification systems (Zhao et al. 2018), but *Microcystis* may also minimize losses by forming a lysogenic state (Stough et al. 2017). In other bacterial systems, lysogeny can cause homoimmunity and allow microbes to escape secondary infection (*aka* super-infection) by other phage, ultimately prolonging the existence of the host (Donnelly-Wu et al. 1993). In the case of the *Microcystis* populations, this generates an ability for cells to escape (at least temporarily) the significant top-down pressure from free *Microcystis*-infecting phage.

While we observed transcripts for viruses potentially capable of controlling a variety of protists, we saw a surprisingly limited richness and abundance in phage. The

majority of phage transcripts were associated with *Microcystis* phage, which was not a surprise, given the dominance of the *Microcystis* host. However, *Microcystis* blooms are commonly accompanied by an abundant heterotrophic bacterial microbiome, particularly proteobacteria, who can play a large role in nitrogen cycling in this system (Krausfeldt et al. 2017; Wilhelm et al. 2011). We expected to observe a corresponding virus community, yet non-*Microcystis* phage accounted for less than 1% of total virus reads. This contradicts previous findings that phage are the most active viruses in aquatic microbial systems (Weinbauer 2004) although it is consistent with transcriptomic observations in at least one other environment (Stough et al. 2019). This disparity may be the result of underestimation of phage sequences, our methodology or a disproportionate expression of RNA virus transcripts. We note that a number of phage candidate sequences were disqualified as false positives during the BLASTx check against the non-redundant database. We identified these sequences as putative prokaryotes and removed them from further analysis and quantification. However, even if every disqualified candidate was considered incorrectly annotated and included in our quantifications, there were still ~5x more transcripts associated with the non-phage "neglected" viruses than with the non-*Microcystis* phage. Other possibilities, including much larger burst sizes associated with RNA viruses or increased expression of the Rdrp gene marker are also possible explanations. However, the literature on this subject is patchy, with little information available focusing on insect viruses. To normalize for variation in expression of RNA virus genes during infection, an isolated infection transcriptome would be required for each algal virus we found.

Regardless, there is precedent to our findings; a similar gene marker approach also revealed higher richness and transcript abundance associated with NCLDVs and RNA viruses in a metatranscriptomic study of the microbiome of *Sphagnum* peat moss relative to the diversity of markers for phage gene expression (Stough et al. 2018). Another nucleic acid weighting approach also indicated a higher abundance of RNA viruses compared to DNA viruses in aquatic systems (Steward et al. 2013). It is unclear if our results show contradicting biology or capture previously overlooked and abundant virus types such as NCLDVs and RNA viruses.

Our co-occurrence analyses revealed populations that behave (in terms of persistence and abundance) in similar manners over the course of the bloom. We resolved 23 clusters that contained both host and virus transcripts with a statistically significant co-occurrence of expression across all 35 samples. While this is not directly indicative of a viral infection of the host in the cluster, it does imply these groups were responding to similar pressures in the system, either abiotic or biotic, and to some degree, their ecologies are thus "linked". For example, we discovered several virophage that clustered significantly with RNA viruses and ssDNA viruses with no obvious NCLDV candidate. As virophage have only ever been identified co-infecting a host with a NCLDV, the observed clusters likely do not represent active co-infections, but hypothetically two different viruses benefitting from similar (or parallel) scenarios. We also observe clusters that contain only bacterial hosts and NCLDVs, the latter of which are to date only known to infect eukaryotes. Again, these co-occurrence observations are likely describing community dynamics (*i.e.,* microbiomes) as opposed to direct infections. These smaller

clusters of viruses and potential hosts can be used to generate hypotheses regarding predator-prey cascades in a community and interspecies viral interactions.

Given the above information, we looked for data and relationships consistent with the hypothesis that non-*Microcystis* viruses were involved in the establishment and/or maintenance of the *Microcystis* bloom. We observed two statically significant co-occurrences including cyanobacterial species and a *Marnaviridae*-like virus (consistent with viruses that infect *Chaetoceros tenuissimus* and *Rhizosolenia setigera*) (Supplemental Table 6). By constraining competing members of the microbial community, these neglected viruses have the potential to create opportunities for *Microcystis* populations to compete for nutrient resources. Our findings highlight the activity of protist-infecting viruses - particularly ssRNA viruses - that have been overlooked. The evidence described here provides the rationale for future studies pertaining to the role of indirect infections influencing bloom-forming species, and the complicated collection of drivers that result in broad global persistence of *Microcystis*.

Data Availability

Raw sequencing data is publicly available in the MG-RAST database (Keegan et al. 2016) under the name "Lake_Taihu_metatranscriptome_project." Sequencing information and quality control details have been previously reported (Tang et al. 2018). Assembled data is also in the MG-RAST database under the name "Taihu 2014 MegaHit Assembly 1." The python script used to extract the aligned portion of BLASTx hits is available in GitHub at https://github.com/Wilhelmlab/general-scripts/blob/master/Pound2019_Extract_aligned.py.

Acknowledgements

This work was supported by grants from the *National Science Foundation* (DEB 1240870, IOS 1451528 and DBI 1530975 to SWW), The *BGSU Great Lakes Center for Fresh Waters and Human Health* (NSF Grant OCE-1840715 and NIH Grant 1P01ES028939-01), the Key Research Program of Frontier Sciences, CAS (QYZDJ-SSW-DQC008), and the *Kenneth and Blaire Mossman Endowment* to the University of Tennessee. We thank the Taihu Laboratory for Lake Ecosystem Research (TLLER) for their help in sample collection.

CHAPTER 4: A COMPARISON OF METATRANSCRIPTOMIC ASSESSMENT METHODS TO CHARACTERIZE *MICROCYSTIS* BLOOMS

Publication Note

This chapter is a version of a peer-reviewed article that was submitted to *Limnology and Oceanography: Methods* by Helena L. Pound, Eric R. Gann, and Steven W. Wilhelm in July 2021. Helena Pound was responsible for experimental design, sample collection and processing, and bioinformatic analysis. Eric Gann provided coding assistance, particularly in the creation of Frankenstein's *Microcystis* genome. The article was written by Helena Pound with input from Eric Gann and Steven Wilhelm.

Abstract

Harmful algal blooms are increasing in duration and severity globally, resulting in increased research interest. The use of genetic sequencing technologies has provided a wealth of opportunity to advance knowledge, but also poses a risk to that knowledge if handled incorrectly. The vast numbers of sequence processing tools and protocols provide a method to test nearly every hypothesis, but each method has inherent strengths and weaknesses. Here, we tested six methods to classify and quantify metatranscriptomic activity from a harmful algal bloom dominated by *Microcystis* spp. Three online tools were evaluated (Kaiju, MG-RAST, and GhostKOALA) in addition to three local tools that included a command line BLASTx approach, recruitment of reads to individual *Microcystis* genomes, and recruitment to a combined *Microcystis* composite genome generated from sequenced isolates with complete, closed genomes. Based on the analyses of each tool presented in this study, two recommendations are made that are dependent on the hypothesis to be tested. For researchers only interested in the function and physiology of *Microcystis* spp., read recruitments to the composite genome, referred to as "Frankenstein's *Microcystis*", provided the highest total estimates of transcript expression. However, for researchers interested in the entire bloom microbiome, the online GhostKOALA annotation tool, followed by subsequent read recruitments, provided functional and taxonomic characterization, in addition to transcript expression estimates. This study highlights the critical need for careful evaluation of methods before data analysis.

Introduction

The ecological and economic ramifications of harmful algal blooms draw significant attention from researchers and the public. Toxic algae cause major challenges for recreational and commercial fisheries as well as municipal water supplies (Bullerjahn et al. 2016). In fact, blooms of the cyanobacteria genus *Microcystis* have caused problems for water treatment facilities around the world, leading to water shortages that included significant recent crises in both the USA and China (Qin et al. 2010; Steffen et al. 2017). As blooms continue to increase in duration and severity, there is an associated increase in research to understand how blooms function (Harke et al. 2016; Tang et al. 2018). For many years, researchers have investigated the environmental conditions that stimulate blooms and attempted to mimic them in the laboratory (Kaebernick et al. 2000; Orr and Jones 1998; Sandrini et al. 2014; Steffen et al. 2014b). Studies have relied on water quality metrics such as pH, toxin estimates, nutrient concentrations, chlorophyll *a* estimates, and basic microscopy until recent advances in DNA/RNA sequencing (Krüger and Eloff 1978; Paerl et al. 2016; Reynolds et al. 1981; Seitzinger 1991). The democratization of new sequencing technologies has opened new opportunities in harmful algal bloom research. The ability to explore the genetic potential or activity of a single organism or entire community ((meta)genomics or (meta)transcriptomics, respectively) within the natural background allows scientists to investigate both novel and historical ecology and physiology (Hennon and Dyhrman 2020; Steffen et al. 2014b).

As technologies continue to advance, so does the volume of data they generate, and all at decreasing cost (Shakya et al. 2019). As an example, recent metatranscriptomic

studies have readily reached 20 million reads per sample, as compared to the 1 million reads per sample only a few years before (Steffen et al. 2015; Stough et al. 2017). While sample collection, processing, and sequencing must be carefully planned and executed, a parallel challenge in sequencing lies in the analysis of the data generated. The opportunity for error in analysis can range from inaccurate estimations of gene expression to misidentified taxonomy. Every researcher is faced with the challenge of deciding which informatics method(s) to use to test their hypotheses, while also balancing needs for efficiency and accuracy. For example, online analysis tools may be more approachable to computationally limited labs, as these web-based tools can handle large datasets using remote servers and rarely possess a steep learning curve to implement (Kanehisa et al. 2016; Keegan et al. 2016; Menzel et al. 2016). Meanwhile, in-house coding methods can incorporate personally curated databases and allow for multiple tools to be selected and combined into a highly specialized and project specific workflow (Moniruzzaman et al. 2017; Pound et al. 2020; Stough et al. 2017). Tools also vary in the type of input data required, as some require just the libraries of trimmed reads, while others require the libraries to be assembled into contigs. Completely analyzing sequencing data may involve some combination of these or other techniques. All have their strengths and weaknesses that must be carefully considered, as these idiosyncrasies can easily influence the data and conclusions made.

The use of molecular tools to study *Microcystis* bloom events has exploded in recent years as early work demonstrated how PCR/qPCR/sequencing could be used to distinguish toxic from non-toxic populations and do so in a quantitative manner (Kurmayer and Kutzenberger 2003; Rinta-Kanto et al. 2005). Water treatment facilities,

government agencies tasked with monitoring, and private commercial groups all have engaged molecular biological approaches to study *Microcystis* in recent years (Chorus and Welker 2021). Currently, most transcriptomic based *Microcystis* research characterizes activity in the environment by recruiting reads back to a single genome strain. Many studies use the reference genome *Microcystis aeruginosa* NIES-843 (Harke and Gobler 2013; Steffen et al. 2017; Tang et al. 2018) since it was the first to be sequenced, but other strains have also been used (Davenport et al. 2019; Morimoto et al. 2019). However, it has been demonstrated that there is a great diversity of genetic potential encoded by different *Microcystis* sp. strains, and in the rest of the microbiome, within a single bloom (Cook et al. 2020; Meyer et al. 2017; Pound and Wilhelm 2020b).

This study sought to compare multiple techniques and tools in the analyses of RNA sequencing data generated from natural harmful algal bloom communities in fresh waters. Metatranscriptomic libraries were generated from a 2019 *Microcystis* spp.-dominated bloom in Lake Erie, USA. A variety of techniques were used to characterize the functional activity and taxonomic composition, using both trimmed reads and assembled contigs. Techniques were evaluated for three primary metrics. The first was the number of reads recruited, either to all *Microcystis* genes or to specific *Microcystis* marker genes important to central metabolism (resolved by phylogeny). It is assumed that methods capable of detecting and estimating the natural diversity in an environmental sample will recruit more reads than methods that are unable to resolve the fine scale genetic variation present in the environment. The second was the ability for a tool to classify sequence data by function and taxonomy, independent of any recruitments. Some tools classify the reads themselves, while others classify assembled contigs, that then

require reads to be recruited back to them to generate relative quantitative results. The third was to determine how well each method correlated, which provides support that the variation between samples is reflective of the ecology of this bloom and not the methods themselves. Based on our analyses we make two recommendations that are dependent on the hypotheses being tested. The following study details how we came to those conclusions and discusses the various strengths and weaknesses of each approach examined.

Methods

Sampling and Sequencing

Samples were collected from the surface of a 2019 *Microcystis* spp.- dominated bloom in Lake Erie, USA on July 21, 2019 and incubated in 1.0 L acid-washed polycarbonate bottles for 48 hours. Whole water was then passed through a 0.2 µm Sterivex filter (Millipore) and flash frozen. RNA was extracted using an acid-phenol based extraction protocol with an added DNase treatment to remove any residual DNA (Pound and Wilhelm 2020a). RNA quality was checked using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific) and quantified using a Qubit hsRNA assay (Invitrogen). Extracted RNA had the ribosomal rRNA reduced and then 100 bp paired-end reads were generated on the Illumina NovaSeq platform at Hudson Alpha Discovery Life Sciences (Huntsville, Alabama). Sequence processing and assembly was described in detail in online protocol (Pound and Wilhelm 2021). Briefly, sequences were quality controlled and trimmed in the CLC Genomics workbench version 20.0.4 (Qiagen). Residual ribosomal rRNA reads were removed *in silico* using SortMeRNA version 4

(Kopylova et al. 2012). The non-ribosomal reads classified by SortMeRNA from all samples were jointly assembled in MegaHit version 1.2.9 (Li et al. 2015). This was done to reduce redundancy among identical sequences in various samples as done previously (Pound et al. 2020). Trimmed non-ribosomal reads have been uploaded to and are available on MG-RAST (Keegan et al. 2016) under project name "LE2019MT" and raw reads are available on the NCBI SRA database under BioProject number PRJNA737197.

Background on Approaches

We explored three widely used online platforms for sequencing analysis: Kaiju (Menzel et al. 2016), MG-RAST (Keegan et al. 2016), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology and Links Annotation (GhostKOALA) (Kanehisa et al. 2016). GhostKOALA and MG-RAST both provide functional and taxonomic classifications, while Kaiju provides only a taxonomic classification (Table 2). Kaiju and MG-RAST accept trimmed sequencing reads as input, while GhostKOALA requires coding sequences to be identified from assembled contigs. We also explored in-house approaches, including variations of a customizable BLASTx database approach (Pound et al. 2020; Pound and Wilhelm 2020b), read recruitments to coding sequences of individual *Microcystis* genomes (Davenport et al. 2019; Harke and Gobler 2013; Krausfeldt et al. 2019b), and recruitments to a *Microcystis* composite genome created for this analysis. The latter is referred to as "Frankenstein's *Microcystis*" in reference to Mary Shelley's fictional monster comprised of parts from many individuals (Shelley 1818). The BLASTx approach taxonomically classifies a gene of interest within assembled contigs, while the recruitments to the individual and composite genomes use

Table 2. Methods Analyzed

Comparison of methods analyzed in this study. Y indicates a requirement or function provided, while N indicates a lack of requirement or output provided.

	BLASTx	Kaiju	MG-RAST	Ghost KOALA	Individual genomes	Frankenstein genome
Community taxonomy	**Y**	**Y**	**Y**	**Y**	N	N
Gene function	**Y**	N	**Y**	**Y**	**Y**	**Y**
User-defined database	**Y**	N	N	N	**Y**	**Y**
Input	**Contigs**	**Reads**	**Reads**	**Contigs**	**Reads**	**Reads**
Additional read recruitment	**Y**	N	N	**Y**	N	N
Online platform	**N**	**Y**	**Y**	**Y**	N	N
Example References	(Moniruzzaman et al. 2017; Pound et al. 2020)	(Chen et al. 2019)	(Krausfeldt et al. 2019a; Zhang et al. 2016)	(Cook et al. 2020; Li et al. 2018; Xie et al. 2016)	(Harke and Gobler 2015; Steffen et al. 2017)	This study.

trimmed reads to functionally classify reads against the coding sequences of a single organism (Table 2).

BLASTx Approach

Identification and transcript quantification for genes of interest was performed using a BLASTx method previously described (Pound and Wilhelm 2019; Pound and Wilhelm 2020b). In short, protein databases were established for the following genes of interest using the KEGG orthology number: DNA-directed RNA polymerase, beta subunit (*rpo*B, K03043), ribulose bisphosphate carboxylase large chain (*rbc*L, K01601), ribose- phosphate pyrophosphokinase 1 (*prp*S, K00948), bifunctional purine biosynthesis protein (*pur*H, K00602), and orotidine 5'-phosphate decarboxylase (*pyr*F, K01591). All reference sequences were downloaded from UniProt KB database version 2020_03 (see data availability). The assembled contigs from all samples was then queried using the command-line BLASTx against each database specific to a gene of interest, retaining sequences with a minimum contig length of 300 bp and an e-value < 1 $x10^{-30}$ (Camacho et al. 2009). Those with BLASTx hits were considered "candidates" and only the aligned portion of the gene sequence was used for recruitments (Pound and Wilhelm 2019). Trimmed reads from each sample were recruited to each trimmed candidate contig in CLC Genomics workbench version 20.0.4 (Qiagen) using a 90% length fraction and 90% identity (Pound and Wilhelm 2020b). Any reads that could be recruited to more than one contig with equal identity was randomly assigned, to mimic MG-RAST and Kaiju data handling approaches and ensure all reads were quantified. The reference trees were established with maximum likelihood phylogenies based on reference proteins from

isolated organisms using PhyML version 3.0 (Guindon et al. 2010). To determine taxonomy, candidate sequences were placed on reference phylogenetic trees using the pplacer algorithm (Matsen et al. 2010).

Kaiju

Trimmed reads were uploaded to the online Kaiju platform and default parameters were used to taxonomically classify reads (Menzel et al. 2016). Reads were classified as *Microcystis* spp. by summing all the reads that were identified to the *Microcystis* genus.

MG-RAST

Trimmed reads were uploaded to the online MG-RAST platform (Keegan et al. 2016). All parameters were default except for no dereplication and the percent identity (> 90%). Reads were classified as *Microcystis* spp. by summing all the reads identified to the *Microcystis* genus based on RefSeq annotations. The number of reads classified as five genes important in central metabolism (*rpo*B, *rbc*L, *prp*S, *pur*H, and *pyr*F) annotated by the KEGG orthology database within MG-RAST were also collected.

GhostKOALA

The nucleotide and translated amino acid sequences of coding sequences were predicted within the assembled contigs using MetaGeneMark version 3.25 (Besemer and Borodovsky 1999; Zhu et al. 2010). Amino acid sequences were functionally and taxonomically classified *via* KEGG orthology using the prokaryote + eukaryote + virus database using GhostKOALA (Kanehisa et al. 2016; Pound et al. 2021b). Predicted coding sequences were assigned KEGG orthology numbers (KOs) and trimmed reads were recruited to the coding sequences using a 90% similarity fraction over a 90% length fraction in CLC Genomic Workbench. Reads that could be recruited to more than one

contig with equal identity were randomly assigned (to mimic MG-RAST and Kaiju data handling approaches and ensure all reads were quantified). The number of reads recruited to the five genes important in central metabolism mentioned above were also pulled from the read recruitments.

Recruitment to Individual Genomes

Coding sequences from closed *Microcystis* spp. genomic assemblies were used in this study. The coding sequences from the following genomes were used: *Microcystis aeruginosa* FD4 (accession: NZ_CP046973.1), *Microcystis* sp. MC19 (accession: NZ_CP020664.1), *Microcystis viridis* NIES-102 (accession: NZ_AP019314.1), *Microcystis aeruginosa* NIES-298 (accession: NZ_CP046058.1), *Microcystis aeruginosa* NIES-843 (accession: NC_010296.1), *Microcystis aeruginosa* NIES-2481 (accession: NZ_CP012375.1), *Microcystis aeruginosa* NIES-2549 (accession: NZ_CP011304.1), *Microcystis panniformis* FACHB-1757 (accession: CP011339.1), and *Microcystis aeruginosa* PCC7806SL (accession: NZ_CP020771.1). Reads were recruited using a 90% similarity fraction over a 90% length fraction to the coding sequences. Any reads that could be recruited to more than one coding sequence with equal identity were randomly assigned. The number of reads recruited to the five markers of central metabolism mentioned above were also pulled from the read recruitments to the individual genomes.

Frankenstein's Microcystis: Construction and Recruitment

Coding sequences from all closed *Microcystis* genomes (see list above) were combined into a single FASTA file and clustered at different nucleotide identities using CD-HIT-EST (Fu et al. 2012; Pound et al. 2021a). Coding sequences from the individual

genomes were clustered at the nucleotide identities of 0.95, 0.90, 0.85, and 0.80 with the word size set to 10, 8, 6, and 5, respectively, to establish composite genomes of various stringencies. Trimmed reads were recruited using a 90% similarity fraction over a 90% length fraction to each combined genome in CLC Genomics Workbench. Any reads that could be recruited to more than one coding sequence with equal identity were randomly assigned (to mimic MG-RAST and Kaiju data handling approaches and ensure all reads were quantified). After comparing the various stringencies (Supplemental Figure 11), the 0.95 identity cluster showed the greatest number of coding sequences, while reducing redundancy, and was used for all subsequent analyses. This synthetic library is referred to as "Frankenstein's *Microcystis*" and is available on GitHub (see data availability). The number of reads recruited to five genes important in central metabolism mentioned above were also pulled from the read recruitments to the individual genomes.

Methods Correlation

The total number of reads recruited to, or classified as, *Microcystis* per sample, and those recruited to, or classified as, specific *Microcystis* genes important in central metabolism per sample were normalized by the total number of trimmed reads in each sample library. Pearson's correlations were then established between methods using Prism version 8 (GraphPad). The Pearson's *r* coefficients are reported in Supplemental Table 7.

Assessment

We used a metatranscriptomic dataset generated from a *Microcystis* spp.-dominated harmful algal bloom in Lake Erie in August 2019 to compare the

methods/tools described above. RNA was extracted from samples collected from bottle incubations (where nutrients were being manipulated) as well as *in situ* samples and then sequenced. A total of ~923 million processed reads were generated across twenty metatranscriptomes (~ 46 million reads per library, see Supplemental Table 8A). These reads assembled into 2,335,243 contigs that varied in length from 157 to 66,630 nucleotides. Total *Microcystis* reads are reported per sample, while reads recruited to or classified as individual genes reads are reported as the average across all samples.

Total Reads recruited to* Microcystis *spp.

One of the primary metrics used to rate the performance of the methods tested was the number of total reads recruited to or classified as *Microcystis* spp. All but one of our methods (the BLASTx approach) provided an estimate of total *Microcystis* reads, but not all methods performed equally (Figure 11). The Kaiju method recruited the fewest reads, as total reads classified as *Microcystis* ranged from 4.08×10^6 to 1.18×10^7, which was between 11.5 and 22.6% of the total reads per sample library (Supplemental Table 8A). Total *Microcystis* read estimates from MG-RAST (13.8 and 28.9% of library), individual *Microcystis* genome recruitment (17.0 and 37.7% of library), and Frankenstein's *Microcystis* genome recruitment (21.4 - 43.8% of library) surpassed Kaiju's estimates but did not provide the largest estimates. GhostKOALA recruited the most reads, as total reads recruited to *Microcystis* annotated genes ranged from 8.38×10^6 to 2.36×10^7, which was between 24.2 and 45.2% of the total available reads per sample library.

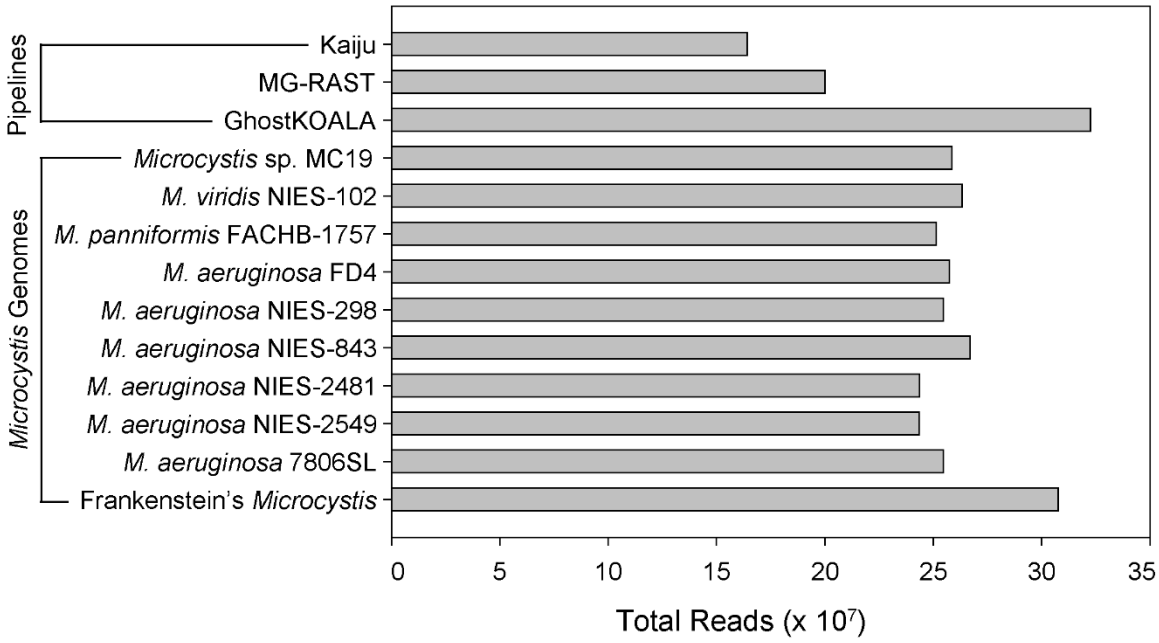


Figure 11. Total *Microcystis* Reads

The sum of all reads in all samples that were recruited to or classified as *Microcystis* in each method.

Reads recruited to* Microcystis-*specific marker genes

It was also important that the methods tested be evaluated for the number of reads recruited to or classified as individual genes, as this is important for future studies concerned with the dynamics of individual genes, as opposed to total shifts in the genome. All but one of our methods (the Kaiju approach) provided read estimates of individual *Microcystis*-specific genes, and most methods performed relatively equally (Supplemental Figure 12). MG-RAST had the fewest reads classified to any of the genes we tested. On average, 1.77×10^4 reads per sample were classified as *rpo*B, 2.39×10^4 reads per sample were classified as *rbc*L, 6.57×10^3 reads per sample were classified as *prp*S, 4.30×10^3 reads per sample were classified as *pur*H, and 3.72×10^2 reads per sample were classified as *pyr*F (Supplemental Table 8B). The other methods evaluated were remarkably similar, including the Ghost KOALA annotated gene recruitment, individual *Microcystis* genome recruitment, and Frankenstein's *Microcystis* genome recruitment, although the BLASTx method outperformed the others in four of the five genes. On average, the BLASTx method recruited 2.86×10^4 reads per sample to *rpo*B, 3.64×10^4 reads per sample to *rbc*L, 8.95×10^3 reads per sample to *prp*S, 4.20×10^3 reads per sample to *pur*H, and 6.22×10^2 reads per sample to *pyr*F (Supplemental Table 8B).

Methods Evaluation

The BLASTx method involved the use of in-house curated databases that each contains a single marker gene from many species; it thus does not allow for a summary of all pathways as the other methods do. However, this approach provides the user full control over the database used and allows for greater confidence in taxonomic predictions based on the subsequent placement of predicted coding sequences on phylogenetic trees

(Matsen et al. 2010). These curated databases can be updated by the user at any point and can be accessed locally. This approach characterizes the assembled contigs, which allows for a more accurate estimate of true community diversity, as opposed to recruiting genes to a reference genome. The number of reads recruited for each gene were >99% correlated to the number of reads recruited by any of the other methods (Supplemental Table 7). The primary weakness of this method is the lack of efficiency. This method is restricted to individual gene investigations, which may work well for specific questions, or small genomes (*e.g.,* RNA viruses, (Moniruzzaman et al. 2017) but is impractical for the complete characterization of the approximately 3,600 genes in a *Microcystis* genome.

While the Kaiju method was the easiest to execute, by our measures it was one of the poorest performers. The program did not provide an efficient way to summarize functional data, as each read was individually characterized to a specific organism, few of which were annotated in the same manner. Therefore, this tool was primarily used to classify taxonomic data, by summing all reads to the *Microcystis* genus. Kaiju does excel in characterizing the entire community, as opposed to recruiting reads to a single species genome. However, it classified fewer reads to *Microcystis* than all other methods. It is unclear if the "missing" reads were not annotated at all or mistakenly annotated as a different organism. It is worth noting that this tool and its databases have not been updated since 2017. This method was the only method tested that was not as strongly correlated to the other methods, with correlation coefficients ranging from 0.788 to 0.831, depending on the method compared (Supplemental Table 7). This suggests that Kaiju may not be consistent in the way it analyzes samples, reducing how well the read estimates correlate to other methods, regardless of the magnitude of reads recruited.

MG-RAST was like Kaiju in that it provided an easy online interface that accepted trimmed reads and functionally characterized the entire community, not just *Microcystis*. We note that MG-RAST out-performed Kaiju, but none of the other methods tested in this study. This method did provide the opportunity to characterize individual genes, including our five marker genes. The number of reads recruited for each gene and total *Microcystis* were >98% correlated to the number of reads recruited by any of the other methods (Supplemental Table 7). The decrease in total, and gene specific, reads recruited is suspected to arise from many probable *Microcystis* reads being incorrectly annotated as *Cyanothece* spp., a species that was not regularly detected in our other methods that were capable of characterizing the entire community, such as the Kaiju or BLASTx methods.

The GhostKOALA method provided both community taxonomy and gene functional characterization, all from an online database. The annotation is based on assembled data and therefore required the recruitment of trimmed reads to genes of interest to make the results quantitative. In our hands this method identified more total reads as *Microcystis* than any other method. The number of reads recruited for each gene and total *Microcystis* were >98% correlated to the number of reads recruited by any of the other methods (Supplemental Table 7). The only notable disadvantage to GhostKOALA methodology is that annotations are limited to KEGG orthologies, so some transcripts within samples may not be annotated.

To assess whether there was a difference in the number of reads that recruited to an individual *Microcystis* genome we recruited reads to coding sequences from all nine complete genomes in NCBI including six *Microcystis aeruginosa* strains, one *Microcystis*

panniformis strain, one *Microcystis viridis* strain, and one *Microcystis* sp. strain (Supplemental Table 8A). This method does not rely on assembled contigs, and all genomes were easily downloaded from the NCBI database, which is updated regularly to include new genomes and annotations. Each of the five specific genes analyzed in this study showed the highest recruitments in a different genome strain, although the variation between each strain was minimal (1-2 %) (Supplemental Figure 12). The number of reads recruited for each gene and all coding sequences in *Microcystis* were >98% correlated to the number of reads recruited by any of the other methods (Supplemental Table 7). The primary disadvantage of this method is the lack of taxonomic or functional data on the rest of the microbial community within a sample.

While the Frankenstein's genome method is primarily a read recruitment method, it has an additional step as a composite genome must first be generated. However, the approach allows the user to customize and update "Frankenstein's *Microcystis*" as other genomes are completed and published. There were a total of 44,950 coding sequences between the nine strains (average = 4,994). Clustering all the coding sequences at decreasing nucleotide identity reduced the total number of clusters from 13,600 to 8,920, when clustered at a nucleotide identity of 0.95 or 0.80, respectively. When all coding sequences were clustered, regardless of nucleotide identity, 20.7 – 43.4% of the total library reads recruited (Supplemental Figure 11). The 95%-identity composite genome performed well in both total reads recruited and reads recruited to the individual genes. The number of reads recruited for each gene and total *Microcystis* were >98% correlated to the number of reads recruited by any of the other methods (Supplemental Table 7). As

with the individual genomes though, this method only allows for the characterization of the *Microcystis* community, not the rest of the microbiome.

Discussion

While the preparation of (meta)transcriptomic samples before sequencing can shape the overall outcome of an analysis for microbial communities (Gann et al. 2021), our analyses have indicated that the way sequences are evaluated can have equally large impacts on the conclusions reached. As the volume of sequencing data generated grows and more researchers turn to molecular tools to address environmental questions, researchers may be tempted to use publicly available, automated pipelines or easily downloaded single-strain genomes. Each of these methods has strengths and weaknesses, and it falls upon the researcher to establish best practices. This study however provides guidelines to the growing community of *Microcystis* researchers but can also serve as advisory to those interested in other organisms or communities.

One of the primary concerns for any methodology for metatranscriptomic sequencing analysis is the ability to efficiently and accurately characterize the true diversity present in a sample, regardless of whether it is from a lab culture or an environmental sample. For many years, the common practice has been to quantify activity by recruited reads back to a single reference genome (Davenport et al. 2019; Harke and Gobler 2013; Steffen et al. 2017). Here we have tested a composite of available complete genomes, Frankenstein's *Microcystis*, where coding sequences from different isolates of the same species were clustered together to reduce redundancy: this provides for a database containing both the common core genes associated with

Microcystis as well as the unique coding potentials associated with different isolates. We note that this approach can be extended to any microorganism of interest if multiple, well-curated genomes are available for clustering. However, it is important to note that even a composite genome is limited to the diversity of sequenced, well-annotated isolates of microorganisms, and still may not represent a natural community's true diversity. Moving forward researchers will need to update Frankenstein's *Microcystis* with newly sequenced isolates as they become available. While the BLASTx method we adopted from Pound et al. (2020) and Moniruzzaman et al. (2017) was also capable of more fully characterizing community diversity, the restriction to individual gene databases makes it cumbersome and less ideal for broader hypotheses.

A key detail to consider *a priori* with sequencing data is the resolution to which a researcher may want to characterize the community. As mentioned above, recruitments to Frankenstein's *Microcystis* provide an efficient and comprehensive way to analyze the *Microcystis* community. However, it is well-known that the rest of the microbiome is likely important to how harmful algal blooms function (Cook et al. 2020). While the BLASTx method mentioned above can characterize the entire microbial community, marker-gene database inefficiencies are still present. Many tools such as Kaiju, or even 16S rRNA gene sequencing, can provide information on what organisms are present, but few tools also provide information on the functional genes present.

Many other factors should also be considered when choosing a method, including the frequency the tool is updated. Even though GhostKOALA performed well in our analysis, the KEGG Gene database will require regular updates to stay relevant. The only way to have full control over a database is to curate it manually, although we recognize

that this can be inefficient and can lead to biases from individual laboratory groups based on annotations. Computational power and coding expertise should also be considered, as some of methods discussed are extremely user friendly while other require some knowledge of command line coding. Many of the online tools use remote servers, while many of the command line tools would likely require local computational power.

Recommendations

For researchers sequencing *Microcystis*-dominated harmful algal blooms, we make two suggestions based on our analyses and the resolution of a proposed study. For those wishing to focus solely on *Microcystis* function and activity, we would recommend using a regularly updated Frankenstein's *Microcystis* composite genome as it does not require the transcriptomes being assembled into contigs and can provide detailed data on every potential gene currently sequenced in *Microcystis* genus. It is also important to note that the combined genome approach used to establish Frankenstein's *Microcystis* could easily be applied to other organisms and study systems. However, for researchers wishing to focus on microbiome ecology and interactions between species, we recommend using the GhostKOALA approach, which provides both functional and taxonomic characterization of the entire community, although it does require additional read recruitments to estimate transcriptional activity. Regardless of the study system, we highly stress the critical importance of taking great care in selecting an appropriate method when processing sequence data.

Data Availability

Trimmed non-ribosomal reads have been uploaded to and are available on MG-RAST (Keegan et al. 2016) under project name "LE2019MT," and raw reads are on NCBI SRA database under BioProject number PRJNA737197. Reference protein sequences used to characterize individual genes and the Frankenstein's *Microcystis* genome are publicly available as fasta files at https://github.com/Wilhelmlab/PoundGannWilhelm2021.

Acknowledgements

We would like to acknowledge Justin Chaffin and the Ohio State SeaGrant Stone Lab for their assistance in sample collection. Robbie Martin and Brittany Zepernick were instrumental in the sampling efforts. Funding was provided by a grant to the *Great Lakes Center for Fresh Waters and Human Health* (NSF Grant OCE-1840715 and NIH Grant 1P01ES028939-01).

CHAPTER 5: ENVIRONMENTAL STUDIES OF CYANOBACTERIAL HARMFUL ALGAL BLOOMS SHOULD INCLUDE INTERACTIONS WITH THE DYNAMIC MICROBIOME

Publication Note

This chapter is a version of a peer-reviewed viewpoint article that was published in Environmental Science and Technology by Helena L. Pound, Robbie M. Martin, Cody S. Sheik, Morgan M. Steffen, Silvia E. Newell, Gregory J. Dick, Robert A. McKay, George S. Bullerjahn, and Steven W. Wilhelm. Morgan Steffen was responsible for sample collection and processing, Helena L. Pound was responsible for primary data analysis, and all authors participated in drafting the article.

Citation:

Pound HL, RM Martin, CS Sheik, MM Steffen, SE Newell, GJ Dick, RA McKay, GS Bullerjahn, and SW Wilhelm (2021). Environmental studies of cyanobacterial harmful algal blooms should include interactions with the dynamic microbiome. Environmental Science and Technology. DOI: 10.1021/acs.est.1c04207

Viewpoint Article

Biology is complicated. Nowhere might this be more true than in aquatic systems. Lakes, especially those in temperate regions, commonly undergo seasonal dynamics in the background of constant anthropogenic insult. Among the ecosystem level responses are cyanobacterial blooms (cHABs), which render water bodies unusable and potentially toxic. High-profile interruptions of access to potable water affecting > 400,000 residents of Toledo, OH in 2014 and more than 2,000,000 residents of Wuxi, China in 2007 highlight this problem (Bullerjahn et al. 2016). Indeed, global-scale observations report an increase in the scale and frequency of cHABs on six of the seven continents (Harke et al. 2016). While eutrophication is clearly a primary driving force, climate change and invasive species are also factors. Ultimately, research into the specific drivers of cHABs continues to provide unclear, and often contradictory, mechanisms of bloom formation: an example of this may be the ongoing debate on the roles of nitrogen and phosphorus as bloom promoters(Wilhelm et al. 2020). Meanwhile, cyanobacteria continue to dominate large freshwater systems despite decades of nutrient control, albeit these controls have been largely phosphorus-focused. Scientifically, there is tremendous focus on both the physiology and ecology of key genera (*e.g.*, *Microcystis* and *Planktothrix*) which produce the toxic secondary metabolite microcystin, a compound originally known as "*Fast Death Factor*" (Hughes et al. 1958). But while tremendous progress has been made, observations remain complicated and often contradictory: this includes (but is not limited to) the roles of pH, temperature and viruses in constraining or promoting cyanobacterial harmful algal blooms (Wilhelm et al. 2020) Despite all efforts,

a potential cause of variability in both lab and field experiments is frequently overlooked: the co-occurring microbes which numerically represent a majority of the microbial community.

In nature, heterotrophic bacteria persist in fresh waters at densities ranging from 10^5 to 10^9 per L of water. Yet in the lab, scientists have worked for decades to reduce or even remove these microbial "contaminants" to create axenic cultures. It is clear, however, that observations without these microbiomes are limited in their relationship to processes that occur in nature. Yet while co-culture experiments with algae may seem more ecologically relevant when bacteria are present, there is no guarantee that the composition or metabolism of microbiomes in lab samples are representative of what one finds *in situ*. As an example, we re-examined the efforts of Steffen et al. (Steffen et al. 2014b) which compared transcripts from *Microcystis aeruginosa* NIES843 cultures in light-limited (control), N-reduced (with different N sources) and P-reduced conditions after three generations. Looking beyond the *Microcystis* cells, changes in both the identity and activity of the microbiome in each treatment were unique (Figure 12). This raises questions as to how data on emergent processes (carbon and nutrient cycling, metabolism, potable water safety, *etc.*) can be correctly interpreted with respect to the bloom-causing agents *vs* the co-occurring community. How do changing microbiomes influence biogeochemical processes (carbon and nutrient pools, metabolites, etc.) and the activity of the bloom-causing organism itself? Are results of experiments an indirect or even direct result of this co-occurring community that is likely sharing metabolites and affording protection from environmental stress, among other functions? To what degree can a microbes' microbiome influence gene expression of the individual or the

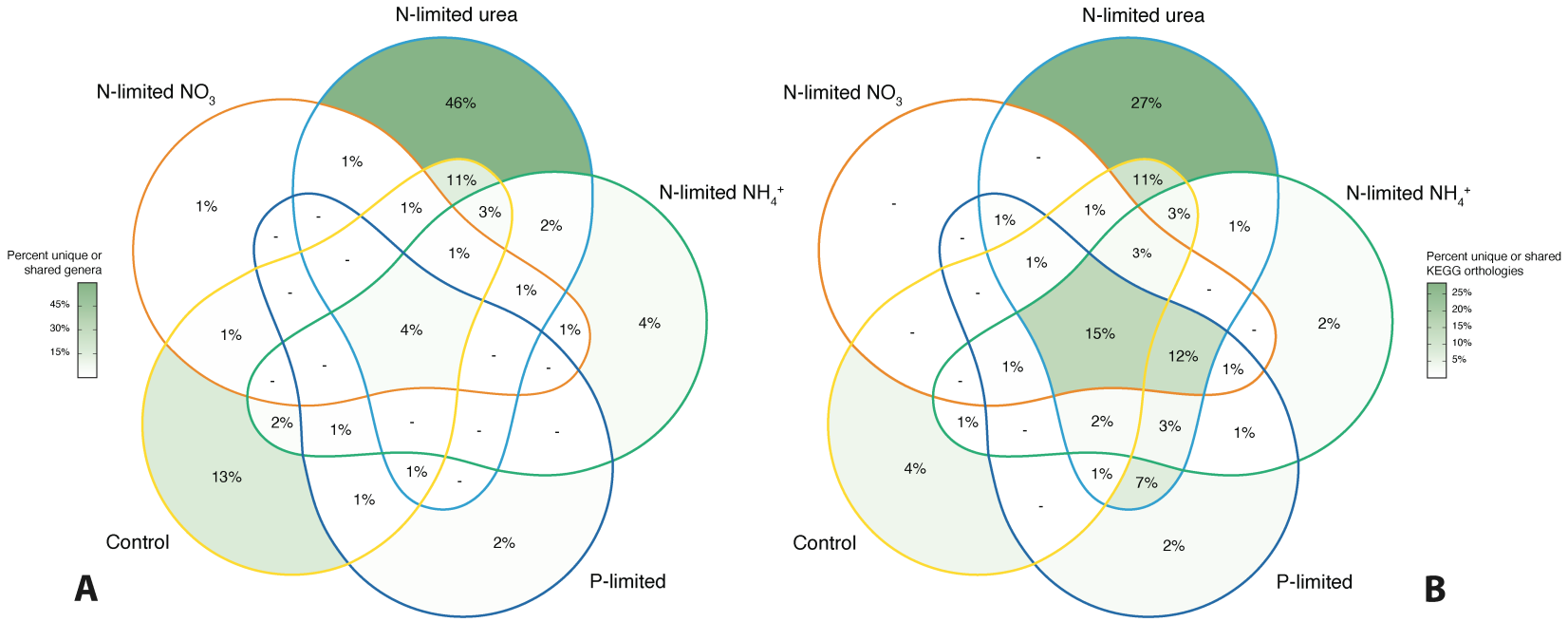


Figure 12. Microbiome Shifts

Nutrient availability shifts microbiomes. *Microcystis aeurginosa* NIES 843 cultures started from the same stock were transferred for 3 generations to unique nitrogen (N) and phosphorus (P) conditions to determine how the cyanobacterium would respond (Steffen et al. 2014). We re-examined metatranscriptomes (*via* Pound et al. 2021, protocols.io, doi = 10.17504/protocols.io.buvbnw2n) to demonstrate shifts in the A) community richness (*rpo*B phylogenies), and B) functional richness (unique KEGG orthology assignments) of the co-occurring microbiome. Results demonstrate that only small components of community (4%) and function (15%) remained common after 3 generations. Note that unions denoted with "-" are for transcripts when phylogeny (<2) and function (< 10) were < 1%.

subsequent production of compounds in complex natural systems? What science observes as a biological pattern in nature is a combination of physics and chemistry at the level of the individual cells which interacts with the physics and chemistry of every other organism, all within the background of a changing environment. We cannot move forward with a clearer understanding of how these blooms are constrained without acknowledging the cyanobacteria are not alone.

Going forward, a paradigm shift is needed to address interactions between ecosystem-threatening freshwater cyanobacteria and "the others" that coexist. We know that heterotrophic bacteria are constant companions of algae in lakes around the world(Cook et al. 2020) and may serve to facilitate or repress cyanobacterial growth, and that bacterial functions during blooms vary in time and space. Indeed, while it is quite likely that core metabolism is conserved in cyanobacteria, the metabolic processes of co-occurring microbes are potentially exchanging primary or secondary metabolites with these cyanobacteria to otherwise mollify or inflame environmental insults. Additionally, it is not just bacteria that may influence cyanobacteria: viruses and fungal chytrids likely play a role in both the repression and success of cyanobacterial blooms.

To this end, we see an urgent need for researchers to report on their toxin-producing phototroph of interest as well as the diversity (*i.e.,* richness and evenness) of the co-occurring microbes in experiments. Deep sequencing of rRNA genes and entire metagenomes/metatranscriptomes has become a standard lab and field tool in microbial ecology: the scientific community should strive to report the identity and function of the entire microbial consortia that persist in lab cultures or during field research efforts.

Researchers are encouraged to consider the functional potential of co-occurring organisms and how these roles might influence cyanobacterial growth dynamics. Moreover, beyond the scope of the initial project, when researchers provide data on the identity and/or function/functional potential they will allow informaticians and modelers an opportunity to address microbiome effects in subsequent analyses of that data. The interactions within the microbial consortia, and between environmental cues and this group, are what drive the phenotypes we observe and must be considered during interpretation of experiments. Furthermore, as microbiome analysis tools continue to improve, especially for shotgun sequencing efforts, insight into the metabolic functions of these microbes will improve our ability to decipher the underlying drivers of undesirable cyanobacterial blooms. What remains certain is that future studies across ecology should no longer ignore the identities and actions of other members of experimental systems, and document such shifts in community structure and function as they occur in laboratory manipulations.

Acknowledgements

Ideas in this paper have evolved from group discussions within the *Great Lakes Center for Fresh Waters and Human Health* (awards 1P01ES028939-01 from NIEHS and NSF OCE 1840715). We thank the NSF, NIH, NOAA and the funding agencies supporting individual labs.

CHAPTER 6: CHANGES IN MICROBIOME ACTIVITYAND SPORADIC VIRAL INFECTIONS HELP EXPLAINOBSERVED "BOTTLE EFFECTS" IN AQUATIC SYSTEMS

Publication Note

This chapter is a version of a peer-reviewed article that will be submitted to a journal by Helena L. Pound, Robbie M. Martin, Brittany N. Zepernick, Courtney J. Christopher, Sara M. Howard, Hector F. Castro, Shawn R. Campagna, Gregory L. Boyer, George S Bullerjahn, Justin D. Chaffin, and Steven W. Wilhelm. Helena L. Pound was responsible for experimental design, sample collection and processing, and bioinformatic analysis. Robbie Martin and Brittany Zepernick assisted with sample collection and processing. Courtney Christopher, Sara Howard, Hector Castro, and Shawn Campagna assisted with metabolomic analyses. George Bullerjahn and Justin Chaffin provided facilities, research support, and experimental guidance. The article was co-written by Helena Pound and Steven Wilhelm with feedback from the other authors.

Abstract

The environmental conditions experienced by microbial communities are rarely fully simulated in the laboratory. Researchers use experimental containers ("bottles"), where natural samples can be manipulated and evaluated. However, container-based methods are subject to "bottle effects"; changes that occur when enclosing the plankton and microbial community. We noted variability consistent with these phenomena in a short-term, nutrient amendment experiment during a 2019 Lake Erie, U.S.A. *Microcystis* spp. bloom. We observed community activity (transcription) in heterotrophic bacteria in a time-frame consistent with a response linked to changes in nutrient availability, yet not to changes in phototroph activity, demonstrating how the often overlooked microbiome of cyanobacterial blooms can be altered. Samples processed at the time of collection contained abundant transcripts from Bacteroidetes, which were reduced in all bottles - including controls - after 48 hours. Significant biological variability was also observed between replicates in the expression of a *Microcystis*-infecting phage, with phosphate-amended replicates showing a 10-fold variation. These phage expression patterns were correlated with ~35% of *Microcystis* spp. functional genes and ~45% of the cellular-metabolites measured across the entire microbial community, suggesting phage activity not only influenced *Microcystis* dynamics, but the entire microbiome. Our observations demonstrate that "bottle effects" may amplify natural heterogenity among replicates, and show how this variation can be harnessed to provide insight on virus and host ecology.

Introduction

Research in environmental microbial ecology is often driven by observations and survey-style data where samples are gathered from natural systems during transects or time series to provide important insight into the spatial and temporal diversity and function of microbial communities (e.g., Venter et al. 2004). Yet, in many cases, the ability to test hypotheses (Popper 1959) requires experimental manipulation of the entire community under natural conditions. To do this, aquatic researchers often employ “bottle experiments” where water samples are collected and distributed across a series of incubation vessels. Bottle experiments have been used for many decades (Schelske 1984) and are advantageous because the duration of the experiments is similar to the generation time of the organisms being studied (Sterner 2008). Incubation vessels, or bottles, are manipulated experimentally by introducing biological, chemical, or physical variation to a subset of the bottles that can be compared to unamended “control” bottles. Many experiments of this nature have been completed and have been used to document processes in nature that range from the role of iron (Fe) as a growth-limiting element in marine systems (Martin and Fitzwater 1988) to the rate at which virus populations are actively produced within microbial communities (Wilhelm et al. 2002). Despite their detractors (Schindler 1998), incubation experiments remain fundamental to the aquatic microbial ecologists' toolset.

The process of executing a well-designed experiment in an enclosure, such as the common bottle-based experiments, is challenging and requires that attention be paid to geographical spatial and temporal scales that could influence the outcome of the

experiment and the interpretation of observations. The simple act of placing lake water into an enclosed bottle can result in differences which are often considered to be "environmental uncertainty" or "bottle effects" (Stewart et al. 2012). Bottle effects due to the volume of the container have been well documented (Spivak et al. 2011). Additionally, unlike natural systems, regenerative processes for nutrients and dissolved gases are often artificially restricted in a microcosm (Zobell and Anderson 1936), with cascading nutrient limitations as a result (Davidson and Howarth, 2007). The presence of a few large grazers in bottle experiments can readily introduce noise into the data (e.g., Urrutia-Cordero et al. 2015; Vanni and Temte 1990). Hypothetically, when water containing natural plankton communities are incubated in several bottles under identical conditions, the response metric measured at the end of the experiment should be consistent among replicates; however, there is always variation among replicates. Therefore, the enclosed nature of bottle experiments allows for independent successional patterns and uneven distribution of biological processes from communities with levels of heterogeneity, even amongst replicate bottles of the same treatment.

While many researchers focus only on phytoplankton responses in bottle experiments, it is well-established that heterotrophic bacteria and viruses in the phycosphere profoundly impact phytoplankton (Cook et al. 2020; Jankowiak and Gobler 2020). Understanding the responses and variabilities associated with heterotrophic bacteria and viruses placed in a bottle may help explain bottle effects, because slight changes to the microbiome could lead to notable differences in microbial diversity, including phytoplankton, and function over time: even when an experiment starts from a consistent microbiome, it can quickly change identity and function (Pound et al. 2021c).

Perhaps more challenging, however, are the influence of viruses. Viruses are ubiquitous in natural systems and key drivers of processes at the biogeochemical, population, and ecosystem levels (Breitbart 2012; Sullivan et al. 2017; Wilhelm and Suttle 1999). Unlike copepods and other grazers, viruses are often "invisible" to researchers and thus overlooked as possible drivers of biological variability. Yet when viruses in a system actively infect a population, the exponential amplification due to viral bursts can have strong effects.

To investigate the effects of short-term nutrient pulses on *Microcystis* in the Laurentian Great Lakes, we conducted a series of *in situ* bottle incubation experiments comparing unamended controls to nutrient amendments over 48 hours. We carefully examined the transcriptome and metabolome of the entire community in each replicate of each nutrient treatment with the aim of identifying community responses to episodic nutrient introductions. During our analyses, we not only observed variability between treatments, but also between replicates. We carefully examined this variation in an attempt to categorize outlier observations as either "unexplained random error" or "biological activity". Our observations demonstrated not only how nutrient additions influenced transcriptional and metabolic landscapes of a toxic cyanobacterial bloom, but also how the confluence of experimental effects and natural variation - including the impact of viruses on microbial communities - can be resolved. We present this work in the context of providing mechanisms for variability that might commonly be attributed to bottle effects, and as a path forward to understanding environmental viral infections.

Methods

Experimental setup for incubations and in situ measurements

Lake water was collected in 20 L carboys from the surface of a *Microcystis* spp.-dominated bloom in Lake Erie, (USA) on July 21st, 2019 (41º44.946 N 83º06.448 W). Water column physiochemistry was recorded prior to sampling using an EXO multiparameter sonde (YSI xylem). Homogenized water was aliquoted into 18 x 1.2 L polycarbonate bottles and subject to four nutrient addition treatments in biological triplicate: 1 □M phosphate□as KH_2PO_4:K_2HPO_4 (Belisle et al. 2016; Wilhelm et al. 2003), and 180 □M nitrogen as either stable-isotope-labeled nitrate ($^{15}NO_3$), stable-isotope-labeled ammonia ($^{15}NH_3$), or stable-isotope dual-labeled urea ($^{13}CH_4{}^{15}N_2O$). Non-replicated bottles of non-labeled nitrate, ammonia, and urea additions were also included to act as standards for metabolome normalization (see metabolite methods below). Six bottles were included as controls with no nutrient additions, three of which were sampled immediately and considered time zero samples. The remaining three control bottles and all nutrient-amended bottles were incubated *in situ* in Lake Erie off docks at Stone Laboratory (41°39.467 N 82°49.600 W) for 48 hours.

Samples were collected at 48 h for metabolomics analysis and RNA sequencing, as well as to measure pH, nutrients in the dissolved fraction, chlorophyll *a,* and toxins. Approximately 150 mL of water was collected on 0.2 □m GTTP filters (EMD Millipore Corporation, Burlington, Massachusetts), stored in a 2mL cryovial and flash-frozen in liquid nitrogen for metabolomics analysis (see below). Approximately 150 mL of water from each bottle was filtered through a 0.2-µm nominal pore-size Sterivex™ filter (EMD

Millipore Corporation, Burlington, Massachusetts) and flash-frozen in liquid nitrogen for subsequent RNA extraction and sequencing (see RNA extraction methods below). pH was measured *via* immediate readings of 15 mL subsamples using a pH probe (Mettler Toledo Seven Compact™ pH/Ion meter S220, fitted with a Mettler InLab Expert Pro-ISM electrode with a temperature range of up to 100° C). Approximately 50 mL of 0.2 µm- filtered water was captured from the Sterivex filtrate for dissolved nutrient analysis (nitrate, ammonia, and phosphate) and stored at 4°C until processing on a SEAL Analytical (Mequon, WI, USA) QuAAtro 5-Channel continuous segmented flow auto-analyzer (Chaffin et al. 2019). Approximately 100 mL of water was filtered onto 0.2-µm nominal pore-size 47-mm diameter polycarbonate filters, and chlorophyll *a* extracted in 90% acetone for 24 h at 4° C prior to measurement on a Turner Designs 10-AU Field Fluorometer (Welschmeyer 1994). Intracellular microcystin concentrations were determined using HPLC coupled with single quadrupole mass spectroscopy and photodioarray spectroscopy as described by Tang et al. (Tang et al. 2018) and in *protocol.io* (Boyer 2020). Standards (microcystin-RR, −LR and -LF) were analyzed at the beginning and end of each run to ensure that the retention times did not drift, and individual toxin congeners (RR, YR, and LR) were identified based on their retention time, their characteristic absorbance spectrum in the photodiode array detector and their characteristic molecular ions.

RNA extraction, processing, and sequencing

RNA extraction, sequencing, and quality control details can be found in previous publications (Pound et al. 2020; Tang et al. 2017). A step-by-step protocol can be found at *protocols.io* describing the acid-phenol based RNA extraction method with an added

DNase treatment to remove any residual DNA (Pound and Wilhelm 2020a). RNA quality was checked using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific) and quantified using the Qubit hsRNA assay (Invitrogen). Extracted RNA was processed using a Illumina® Stranded Total RNA Prep, Ligation with Ribo-Zero Plus (for 96 Samples) and then 50-million 100-bp paired-end reads were generated on the Illumina NovaSeq platform at Hudson Alpha Discovery Life Sciences (Huntsville, Alabama). Raw reads are available on the NCBI SRA database under BioProject number PRJNA737197. Sequence processing and assembly details can be found in an online, step-by-step protocols.io protocol (Pound and Wilhelm 2021). Briefly, sequences were quality controlled and trimmed in the CLC Genomics workbench version 20.0.4 (Qiagen). Lingering ribosomal rRNA reads were removed *in silico* using SortMeRNA version 4 (Kopylova et al. 2012). The non-ribosomal trimmed reads from all samples were jointly assembled in MegaHit version 1.2.9 (Li et al. 2015). As previously described (Pound et al. 2020), this was done to reduce redundancy among identical sequences in various samples.

Characterization of microbiome in metatranscriptome assemblies

Gene coding sequences were predicted from nucleotide and translated amino acid sequences using MetaGeneMark version 3.25 (Besemer and Borodovsky 1999; Zhu et al. 2010). Amino acid sequences were functionally and taxonomically annotated *via* KEGG orthology using the prokaryote + eukaryote + virus database in GhostKOALA (Kanehisa et al. 2016; Pound et al. 2021b). KEGG orthology numbers (KOs) were assigned to predicted coding sequences and transcript expression was estimated by recruiting trimmed reads to the coding sequences using a 90% similarity fraction over a 90% length

fraction in CLC Genomic Workbench. Reads that could be recruited to more than one contig with equal identity were randomly assigned. The KEGG orthology annotations provided by GhostKOALA allow for estimates of individual taxa and/or individual annotated genes. DNA-dependent RNA polymerase (*rpoB*) expression was used as a hallmark gene to quantify active transcription, as a proxy for organism abundance. Differential expression analysis of the expression of *Microcystis* specific functional genes annotated with KOs was performed using DeSeq2 (Love et al. 2014).

Characterization of virus signatures in metatranscriptome assembly

Identification and transcript quantification for viruses of interest was performed using a BLASTx method previously described (Pound and Wilhelm 2019; Pound and Wilhelm 2020b). Phage protein databases were established for the following genes of interest: terminase, major capsid protein, and tail sheath with all reference sequences downloaded from UniProt KB database version 2020_03 (see data availability). The assembled contigs from all samples were then queried using the command-line BLASTx against each database specific to a gene of interest, retaining sequences with a minimum contig length of 300 bp and an e-value < 1 $x10^{-20}$ (Camacho et al. 2009). Contigs with BLASTx hits were considered "candidates" and only the aligned portion of the gene sequence was used for subsequent expression analyses (Pound and Wilhelm 2019). Non-ribosomal trimmed reads from each sample were recruited to aligned portion of each candidate contig in CLC Genomics workbench version 20.0.4 (Qiagen) using a 90% length fraction and 90% identity (Pound and Wilhelm 2020b). Non-specific recruitments (reads that could be recruited to more than one contig with equal identity) were randomly assigned, to ensure all reads were quantified. Phylogenetic trees were established with

maximum likelihood phylogenies based on reference proteins from isolated organisms using PhyML version 3.0 (Guindon et al. 2010). To determine taxonomy, candidate sequences were placed on reference phylogenetic trees using the pplacer algorithm (Matsen et al. 2010). Transcript expression for the entire *Microcystis* phage genome was estimated by recruiting reads to the *Microcystis* phage Ma-LMM01 genome (Accession AB231700.1) using a 90% length fraction and 90% identity (Pound and Wilhelm 2020b; Yoshida et al. 2008).

Metabolite extraction and assessment

Samples for metabolomics analysis were processed in a manner similar to Krausfeldt et al. (Krausfeldt et al. 2019a). In brief, filters with biomass were exposed to pre-chilled extraction solvent (1.3 mL of 40:40:20 HPLC grade acetonitrile/methanol/water with 0.1M formic acid) at -20° C for 20 minutes before water soluble metabolites were extracted from filters at 4° C. Extracted metabolites were dried under a stream of N_2 and resuspended in sterile water. Samples were immediately placed in the Ultimate 3000 RS autosampler (Dionex, Sunnyvale, California) and an injection volume of 10 μL was separated through a reverse-phase C18 (Synergi 2.5 µm Hydro- RP, 100 Å, 100 x 2.00 mm) liquid chromatography column (Phenomenex, Torrance, California) kept at 25° C. Solvent A consisted of 97:3 water:methanol, 10mM tributylamine, and 15mM acetic acid. Solvent B was 100% methanol. The solvent gradient from 0 to 5 min was 100% A: 0% B, from 5 to 13 min was 80% A: 20% B, from 13 to 15.5 min was 45% A: 55% B, from 15.5 to19 min was 5% A: 95% B, and from 19 to 25 min was 100% A: 0% B with a flow rate of 200 µL/min. The mass spectrometer was operated in full scan mode following an adopted protocol (Lu et al. 2010). The

chromatographic eluent was ionized *via* an electrospray ionization (ESI) source in negative mode and coupled to an Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, Massachusetts) through a 0.1-mm internal diameter fused silica capillary tube. The samples were run with a spray voltage of 3 kV, N_2 sheath gas of 10 (arbitrary units), capillary temperature of 320° C, and automatic gain control (AGC) target set to 3 x 10^6 ions. Samples were analyzed at a resolution of 140,000 and a scan window of 85–800 *m/z* from 0 to 9 min, and 110–1000 *m/z* from 9 to 25 min. Files generated by Xcalibur (RAW) were converted to the open-source mzML format (Martens et al. 2011) *via* the open-source msconvert software as part of the ProteoWizard package (Chambers et al. 2012). MAVEN (mzroll) software, Princeton University (Clasquin et al. 2012; Melamud et al. 2010), which uses a peak grouping algorithm and retention time alignment, was used for visualization of extracted ion chromatograms. Metabolites, validated from an in-house library, were first manually identified using samples from the unlabeled nutrient bottles and integrated based on exact mass (± 5 ppm mass tolerance) and retention time ($\Delta \leq 1.5$ min). Peak intensities for a given sample were normalized by the sum of all metabolites detected for that sample. However, only metabolites that showed stable isotope incorporation were considered in further analysis, ensuring that the metabolites considered were being actively used or produced by the cell at the time of sampling.

Statistical analyses

Statistical significance between treatments was determined using ordinary two-way ANOVAs with a Tukey' HSD test for chlorophyll *a*, toxin concentrations, transcript expression, and metabolite peak intensities. Shifts in microbial communities were

characterized using canonical correspondence analyses that incorporated dissolved nutrient metadata. Pearson correlation coefficients were established with Benjamini Hochberg corrections for multiple comparisons. All statistical analyses were carried out in R studio.

Results

Basic bloom characterization

Water column physiochemistry at the onset of the experiment included dissolved oxygen = 8.73 mg L^{-1}; pH = 8.89, turbidity = 0.63 NTU; temperature = 26.2° C; and chlorophyll *a* = 0.23 RFU. After 48 h of incubation. chlorophyll *a* concentrations were significantly higher in the bottles with phosphate added (p-value < 0.001) than in the control bottles (Figure 13, Supplemental Data 9). Dissolved nutrients revealed a significant drawdown in nitrate in the phosphate-amended bottles (p-value < 0.001), as compared to the control bottles (Supplemental Data 9). These data indicate the growth of the phototrophic community was P-limited. Microcystin concentrations were not significantly different between treatments (p-value > 0.1), nor were there differences in congener profile (Supplemental Figure 13).

Microcystis ***nutrient transport gene expression***

As expected, *Microcystis* was the most represented genus in the sequencing dataset, with normalized DNA-dependent RNA polymerase (*rpoB*) expression ranging from 5.2 to 27 per 10^7 basepair (Supplemental Figure 14). Overall, ammonium (*amt*) and urea transporters (*urt* family) were significantly upregulated from the control at 48 hours in the phosphate-amended bottles, while phosphate transporters (*pst* family) were

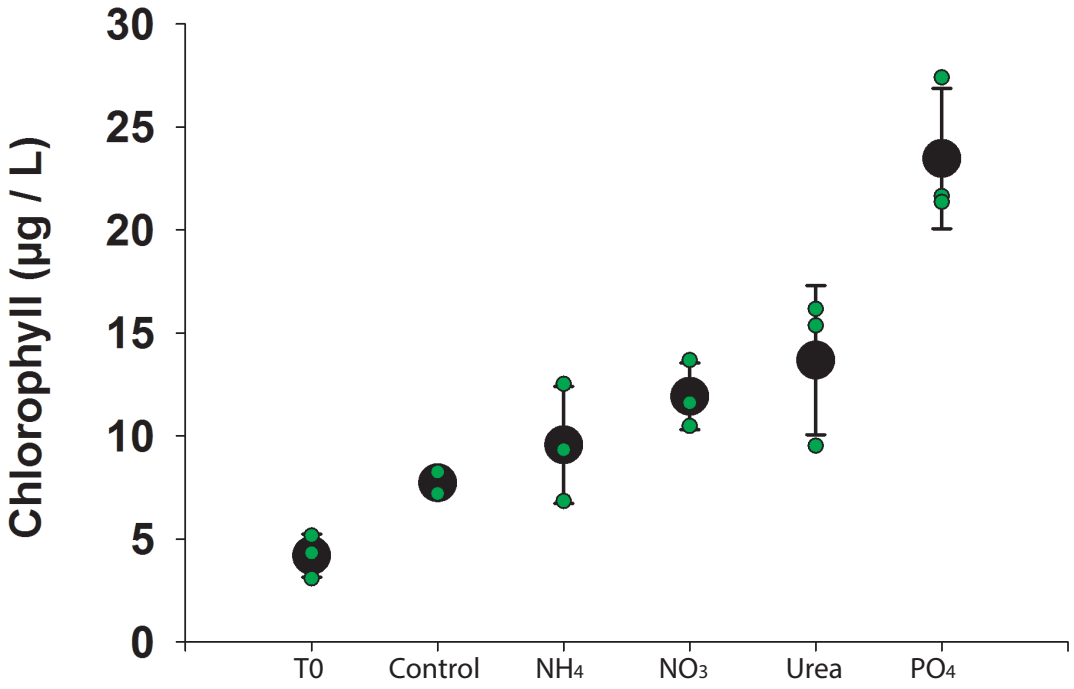


Figure 13. Chlorophyll *a* concentrations

Chlorophyll *a* concentration across experimental manipulations. Individual bottles are denoted in green and the average is denoted in black. Error bars represent standard deviation.

significantly downregulated (p-value < 0.01) (Supplemental Data 9). Relative to control, at 48 h nitrate/nitrite transporters (nrt family) were downregulated in the ammonia- and urea-amended bottles (p-value < 0.001). Within control bottles, there were significant differences in nitrate and urea transporter expression between T0 and T48 h. Both were significantly upregulated at 48 h, relative to T0 (p-values < 0.001) (Supplemental Data 9).

Co-occurring host community gene expression

The sum of *rpoB* expression for all other co-occurring, non-viral organisms in the community ranged from 5.6 to 13 per 10^7 basepair (Supplemental Figure 14). In most treatments, *Microcystis* alone showed higher transcript expression levels than the rest of the community combined, particularly in the phosphate treatments. Ammonia was the only treatment in which the combined community transcript expression of *rpoB* was higher than *Microcystis*. Changes in the co-occurring community appear to be primarily a result of shifts in transcription within the heterotrophic community. This is particularly apparent in the transcriptional shifts in the heterotrophic community observed between T0 and T48 controls (Supplemental Figure 15). This shift was reversed in the T48 ammonia bottles, with the heterotrophic community demonstrating a transcriptional composition more similar to the T0 control. Characterization of the taxonomy within these shifts confirmed that they primarily occurred in the heterotrophic community, with lesser change in the cyanobacterial species. The majority of active expression (75-85%) in the co-occurring community, based on the sum of all samples, originated from Actinobacteria, Bacteroidetes, Proteobacteria, and other Cyanobacteria species (Figure 14). Excluding *Microcystis,* Bacteroidetes showed the highest levels of active transcription in the time zero bottles while Betaproteobacteria was the most

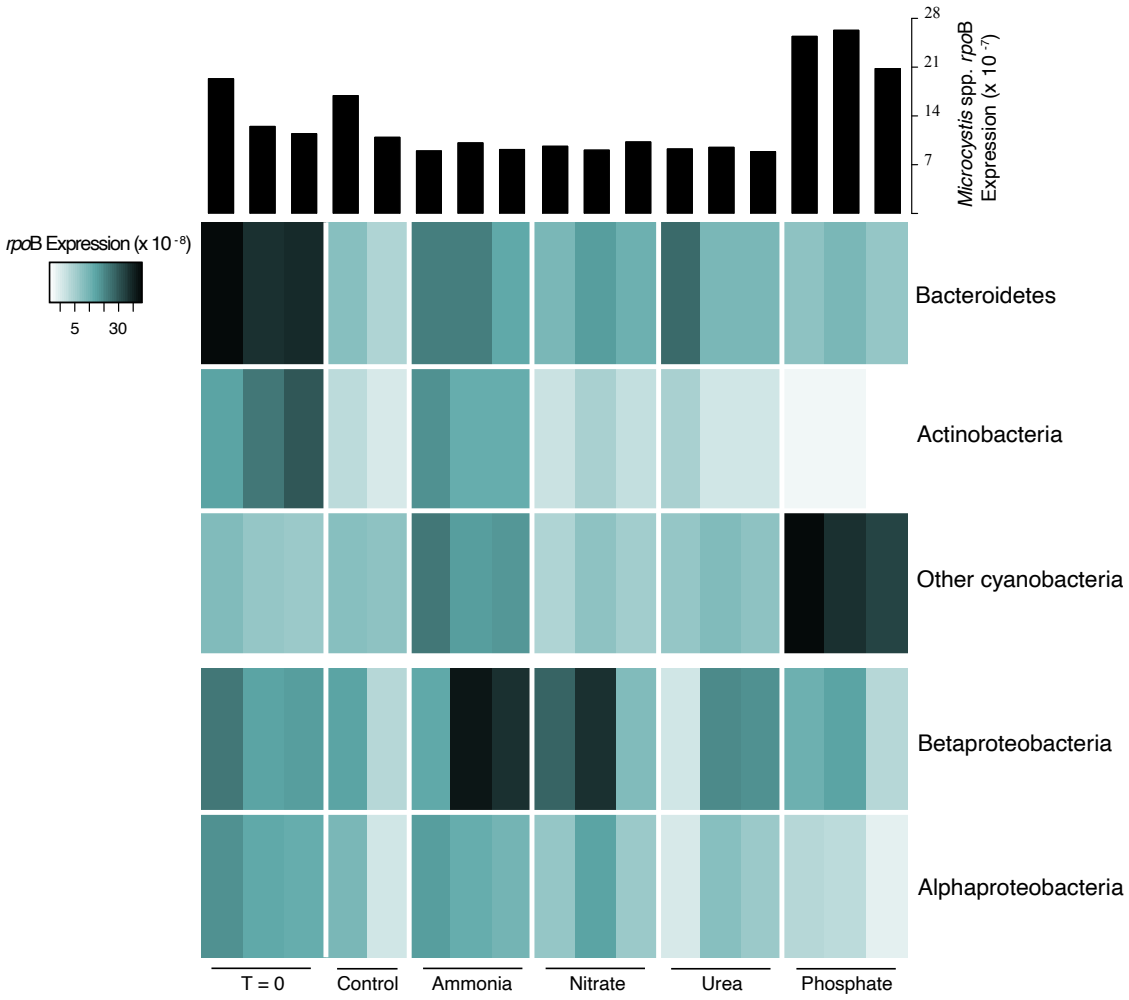


Figure 14. Microbiome Expression Heatmap

Heatmap of normalized transcript expression of the DNA-dependent RNA polymerase (*rpo*B) of dominant phyla (top three) and classes (bottom two) in co-occurring heterotrophic microbial community. Transcript expression was normalized to the library size for each sample. Histograms at the top describe the transcriptional response of *Microcystis* spp. *rpoB* for comparison.

transcriptionally active in the ammonia and nitrate bottles. Within the cyanobacterial group (excluding *Microcystis*), *Pseudanabaena* was the most transcriptionally active in all but the ammonia bottles, where *Synechococcaceae* were the most represented (Supplemental Data 9). Variations in dominant co-occurring taxa occur not only between nutrient treatments, but among bottles within the same treatment. For example, there is notable variation in Betaproteobacteria. Even when *Microcystis* remains fairly constant between replicates of the same nutrient treatment, the co-occurring microbial community does not (Figure 14).

Viral community characterization

An evaluation of the viral community using a BLASTx approach revealed that *Microcystis* phage were the most transcriptionally active viruses and that viruses infecting other co-occurring community members were either largely absent or were not being actively transcribed. Normalized expression computed from either *Microcystis* phage contigs (data not shown) or from the entire *Microcystis* phage Ma-LMM01 genome indicated low levels of phage activity (2.2 to 57 per 10^8 basepair) in most bottles regardless of the nutrient treatment, excluding the phosphate-amended bottles. However, across phosphate treatments, phage expression was highly variable, ranging from 4.7 to 48 per 10^7 basepair (Figure 15). The third replicate had a 10-fold increase in phage transcript expression compared to the first replicate. These increases were not associated with increased host expression, as *Microcystis rpoB* transcription remained largely unchanged across phosphate treatments.

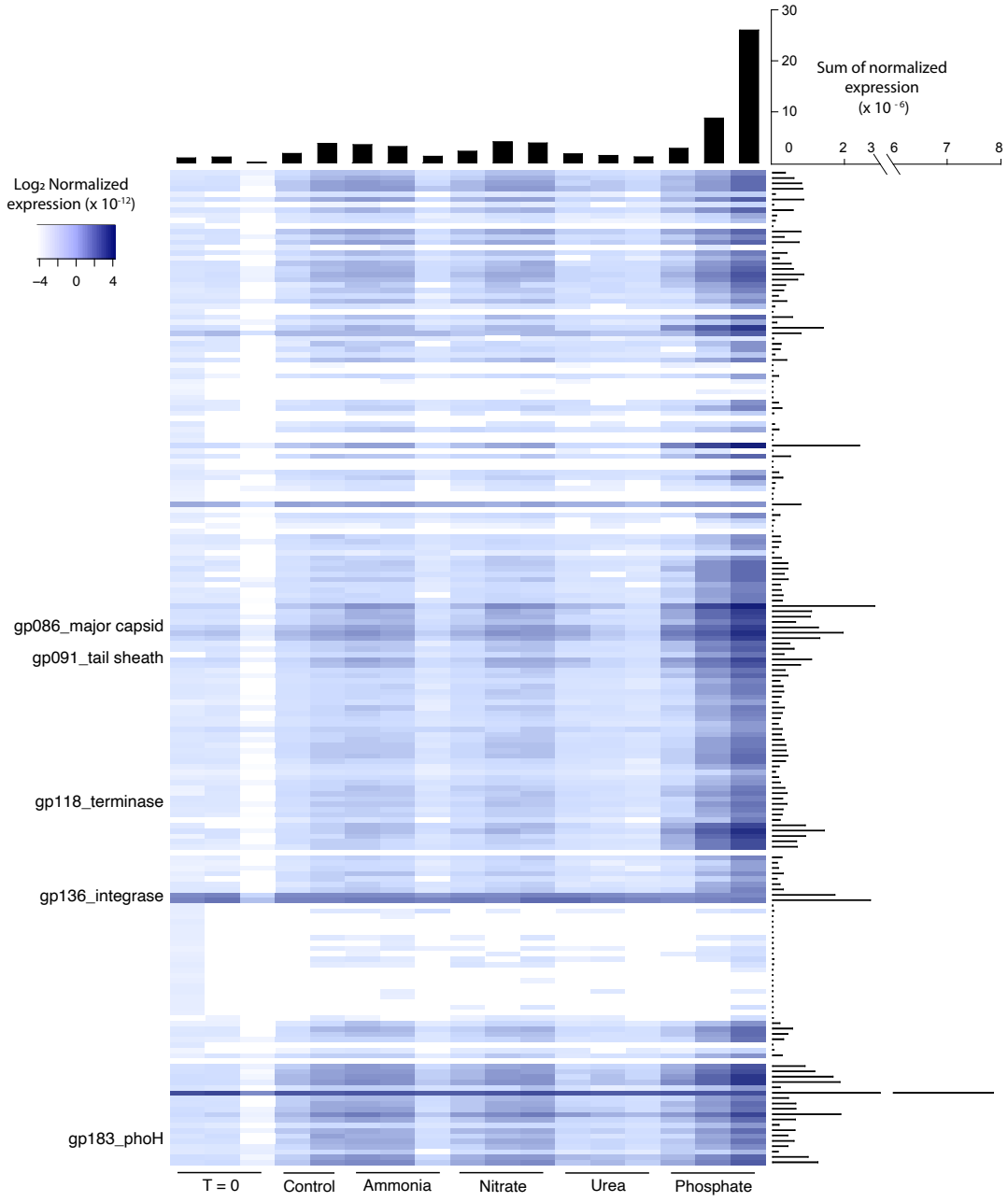


Figure 15. *Microcystis* Phage Expression Heatmap

Heatmap of normalized transcript expression for *Microcystis* phage Ma-LMM01 genes in individual experimental bottles. Histograms describe the sum of normalized transcript expression for each bottle (horizontal histogram) and each gene (vertical histogram). Transcript expression was normalized to the library size of each sample and the length of each gene.

The pattern of gene transcript expression across the *Microcystis* phage Ma-LMM01 genome held fairly constant, with most genes following a similar dramatic increase in the third phosphate replicate. This includes genes that serve as typical markers for lytic activity, such as the tail sheath gene (gp_91), the major capsid gene (gp_86), and the terminase gene (gp_118) (Stough et al. 2017), (Figure 15). A hypothesized auxiliary metabolic gene thought to be induced by phosphate starvation, *pho*H, also showed the dramatic increase in transcription in the third phosphate replicate relative to all others. In contrast, expression of genes associated with lysogeny, such as the integrase/recombinase (gp_136) and transposase (gp_135), displayed consistent and relatively low transcription across all bottles, including all of the phosphate bottles (Figure 16).

Overall, the BLASTx analysis indicated that trancription of non-*Microcystis* virus genes, both prokaryotic and eukaryotic infecting viruses, were nearly absent (data not shown). An exception included a *Synechococcus*-like myovirus terminase gene that was more transcriptionally active than *Microcystis* phage in the T0 and ammonia-amended bottles (Supplemental Figure 16).

This pattern *Microcystis* phage activity held fairly constant across the genome, with most genes following a similar dramatic increase in the last phosphate bottle, including the typical markers for lytic activity, such as the tail sheath gene (gp_91), the major capsid gene (gp_86), and the terminase gene (gp_118) (Figure 15). A hypothesized auxiliary metabolic gene thought to be induced by phosphate starvation, *pho*H, also showed the dramatic increase observed in the rest of the genome. However, not every gene in the phage genome showed this dramatic increase in the last phosphate bottles.

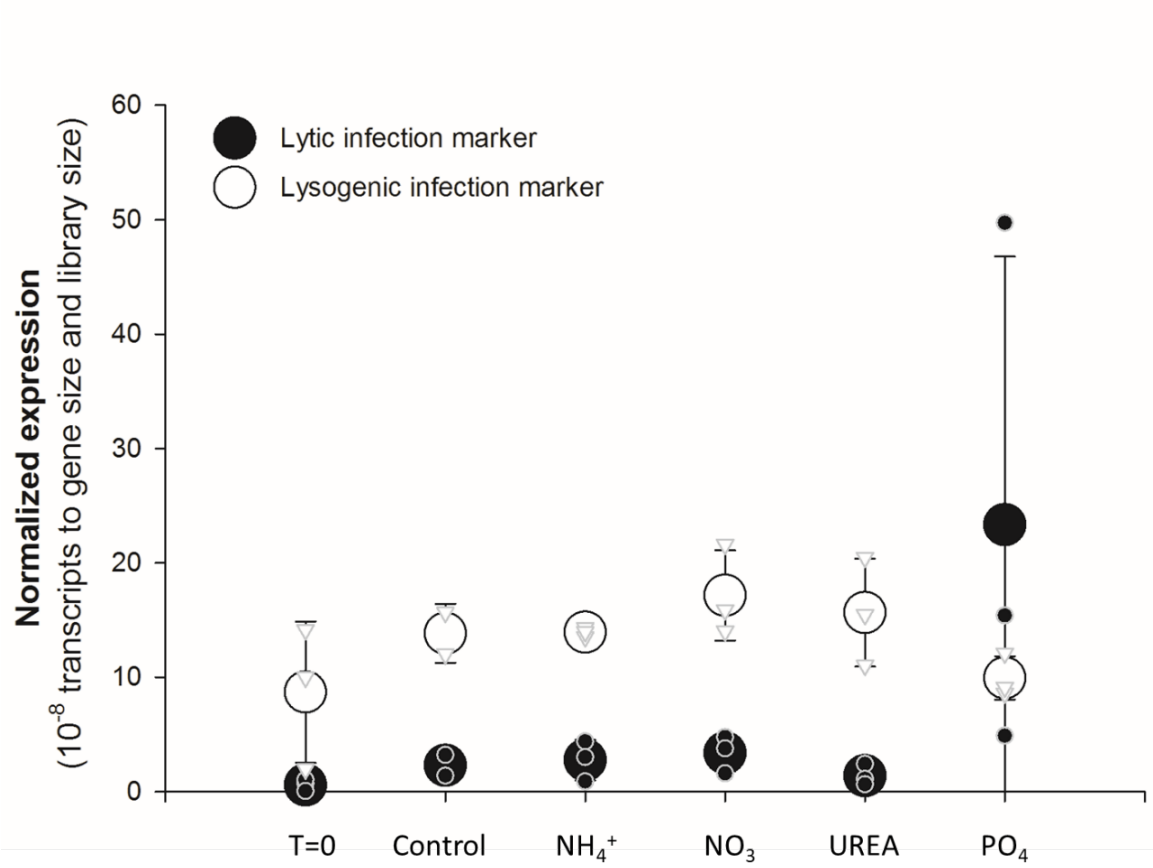


Figure 16. Lytic vs. Lysogenic Expression

Average (large symbols) and individual bottle (small symbols) normalized transcript expression of a *Microcystis* phage marker of lytic activity (gp91 tail sheath, black circles) and a marker of lysogenic activity (gp136 recombinase/integrase, open triangles). Transcript expression was normalized to both library size of each sample and the length of each gene.

Common gene markers associated with lysogeny, such as the integrase/ recombinase (gp_136) and transposase (gp_135), displayed nearly constant expression across all bottles, even the phosphate bottles with high expression levels of the lytic markers (Figure 16).

Overall, the BLASTx analysis revealed that non-*Microcystis* viruses, both prokaryotic and eukaryotic, were nearly absent (data not shown). However, expression patterns based on the phage terminase gene revealed that a *Synechococcus*-like phage was more transcriptionally active than the *Microcystis* phage in the time zero bottles and in the ammonia amended bottles (Supplemental Figure 16, Supplemental Data 9).

Correlation between phage gene expression and metabolites

A total of 58 metabolites were detected in the community, all of which were being actively cycled at the time of sampling (Figure 17A). Almost half (25 of 58) of the metabolites were correlated to expression of the Ma-LMM01 tail sheath gene (gp_91), an indicator of lytic infection activity (false discovery rate (FDR) corrected p-value < 0.05) (Figure 17B). One metabolite (orotate) was inversely correlated to gp_91 expression. Many of significantly correlated metabolites have been linked to phage expression in previous studies of other systems, including UDP sugars and metabolites involved in the pentose phosphate pathway (Ankrah et al. 2014; Thompson et al. 2011).

It is unclear what proportion of the metabolite pools can be attributed to *Microcystis* activity. To gain insight into linkages between *Microcystis* and *Microcystis* phage detected, we correlated all *Microcystis* transcript expression to the phage tail sheath transcript expression across all of our samples. Supplemental Figure 17 shows the expression of each KEGG orthology group annotated as *Microcystis,* and many show the

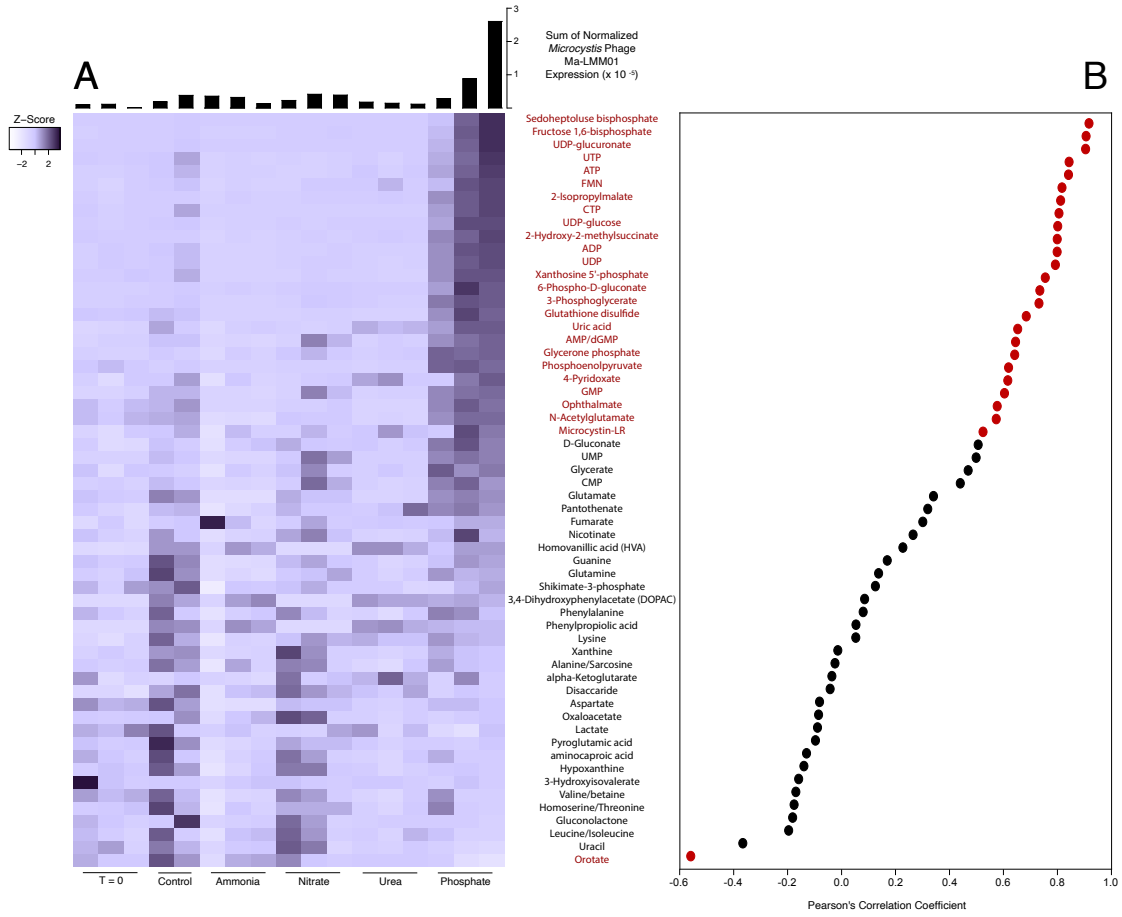


Figure 17. Metabolite Correlation to Phage

Heatmap of metabolite peak intensities (A). Color scale has been scaled to each metabolite (z-score). Histograms describe the sum of normalized transcript expression for *Microcystis* phage Ma-LMM01 in each experimental bottle. Transcript expression was normalized to the library size of each sample and the length of each gene. Scatter plot shows the Pearson's correlation coefficient between each metabolite detected and the *Microcystis* phage Ma-LMM01 tail sheath normalized transcript expression (B). Red dots and metabolite labels indicate a significant correlation ($p < 0.05$), while black dots and metabolite labels indicate correlations deemed to be non-significant statistically ($p > 0.05$).

same pattern of increased expression in the third phosphate replicate. In fact, 497 of 1414 total detected (35.1%) of *Microcystis* KEGG orthoglogies are significantly correlated to the *Microcystis* Ma-LMM01 tail sheath gene (FDR p-value < 0.05). All of these significant correlations were directly correlated, with no KEGG orthology group significantly inversely correlated.

Discussion

Experiments in microbial ecology often use container-bound populations to explore responses to perturbation. These experiments commonly involve replicate bottles or micro/mescocosms to increase confidence in the observations by reproducing results. When identical treatment bottles, or replicates, do not show identical responses, many scientists chalk the discrepancies up as "experimental error" or "unexplained variability." Upon closer inspection, however, at least some of this variability can often be attributed to subtle shifts in the microbiome of a target community either in nature or a non-axenic lab culture (Pound et al. 2021c).

In harmful algal bloom research, blooms are often characterized by the amount of chlorophyll *a* present. While this can provide a useful baseline for bloom biomass, it serves as a non-specific estimate of phototrophs in the community. It does not provide specific information on the diversity of phototrophic community nor the heterotrophic community. In this study, measurements of chlorophyll *a* responses to nutrient additions revealed that the phototrophic community was constrained by available phosphate. RNA sequence analysis confirmed that transcription in the phototrophic community was dominated by *Microcystis*, particularly in the phosphate-amended bottles. Phosphate

constraints on *Microcystis* biomass has been well-documented in many natural bloom systems (Davis et al. 2010; Wilhelm et al. 2003). There was no change in chlorophyll *a* associated with the various nitrogen amendments, indicating that the phototroph biomass was not constrained by nitrogen at the time of sampling.

However, as hypothesized nearly two decades ago (Wilhelm et al. 2003), nitrogen-amendments did cause a noted change in the community transcription. Our canonical correspondence analysis revealed that the microbial community transcriptional shifts occurred primarily within the heterotrophic bacterial community. This was particulatly evident in the control bottles after 48 h as well as in the ammonia-amended bottles. The initial heterotrophic bacterial community was dominated by Bacteroidetes transcription, but after 48 h of incubation, the active heterotrophic microbial community in the control bottles had shifted, while the phototrophic community remained largely unchanged. The addition of ammonia appeared to maintain the transcriptional activity dominance of Bacteroidetes and Betaproteobacteria. This observation suggested that ammonia recycling was decoupled when the *in situ* heterotrophic community was constrained to a bottle. To phrase this another way, it suggests that at the time of sampling the *in situ* bacterial community structure may have been dependent on ammonia as a main N source. Heterotrophic bacteria generally have faster growth rates than many planktonic phototrophs, which could allow for a more rapid change in that community because of changing environmental pressures (DeBruyn et al. 2004). This shift in community activitiy in the control bottles after 48 h is but one example of how bottle effects can alter community dynamics before consequences of experimental treatments are ever considered. We note that this response is not unexpected: recently we (Pound et

al. 2021c) demonstrated that a single batch culture of non-axenic *Microcystis aeruginosa* undergoes dramatic shifts in the transcriptional activity of co-occurring heterotrophic bacteria after only three batch culture transfers in growth media containing different nitrogen sources.

The hypothesis that the *in situ* community was utilizing ammonia regeneration as a nitrogen source, even in high concentrations of nitrate (initial nitrate concentration was 30µmol/L), finds support within *Microcystis* functional gene expression data. The decrease in expression of nitrate/nitrite transporters in the ammonia- and urea amended-bottles suggested these communities were not relying on nitrate or nitrite as a primary nitrogen source. This is further supported by the increase in urea and ammonia transporters observed in the control after 48 h, indicating that *Microcystis* was preferentially using those nitrogen sources when enclosed in a bottle. When phosphate was added to some bottles, we saw a similar increase in urea and ammonia transporters, indicating a phosphate-constrained community that showed preferential uptake of urea and ammonia as a nitrogen source. This conclusion is supported by several recent studies. The utilization of regenerated ammonium has been shown to support non-N-fixing blooms in Lake Erie (Hampel et al. 2019), and a stable isotope study showed that *Microcystis* would preferentially assimilate low concentrations of ammonium in the presence of high nitrate concentrations (Chaffin and Bridgeman 2014). A preference for urea has been observed in many other *Microcystis* bloom events: urea is a common agricultural-based anthropogenic input, serving as both a nitrogen and carbon source to bloom organisms (Davis et al. 2010; Krausfeldt et al. 2019a).

Characterization of the microbial community, based on *rpoB* transcriptional levels, revealed that bottle effects not only caused shifts between treatments, but can also cause shifts within treatments, resulting in variability between replicate bottles. Once a community is isolated in a container, it undergoes succession independent of other communities/containers regardless of treatment. Heterogeneity in natural microbial systems provides initial variation in bottled communities, and through independent succession, communities diverge further when contained. This phenomenon is apparent in several of the observations made from our experiments. On the simplest level, we recorded variable chlorophyll *a* levels between bottles of the same treatment. While such variation is common, the cause is rarely considered nor explained. It could be the result of a physical or chemical change, such as a shift in light levels, or a biological change in the non-target (*i.e.,* in this case the heteotrophic microbial) community. It could either mean a shift in the type of phototroph present, or a shift in the photosynthetic activity of a single taxa, or some combination thereof. While there are a variety of potential sources of varibiality, shifts in the microbial community, either in function or taxonomy, could play a crucial role in observed variation. We also see bottle effects influencing intra-treatment variability in community composition. We documented variations in the heterotrophic community between bottles of the same treatment, where expression levels of the same group of organisms varied dramatically between bottles.

Perhaps our most dramatic example of intra-treatment variability was observed in *Microcystis* phage transcriptional activity within the phosphate-amended bottles. The three replicate bottles showed different levels of phage gene expression, with the third replicate displaying a 10-fold higher representation of *Microcystis*-infecting phage

relative to the first replicate. The viral expression levels observed in the third phosphate replicate suggested that a significant lytic event took place. Phage infection in the third replicate was associated with a ~25% decrease in *Microcystis* abundance (as measured by *rpoB*). In contrast, expression of the genetic marker for lysogeny was consistent and relatively low across all bottles, including the bottle with high transcriptional levels of the marker gene for lytic activity. This suggests that this lytic event was not the result of lysogenic induction, or if so, induction was incomplete. Such variable virus expression has recently been observed in other experiment systems, including an *Emiliania huxleyi* bloom induced in mesocosms, where researchers noted highly variable virus expression among replicate samples (Kuhlisch et al. 2021). In a likewise manner, their observations were tied to variability in host metrics, including chlorophyll *a* and metabolites.

The occurrence of a large-scale virus-mediated lytic event, even when a host population does not fully lyse, can have massive effects on host dynamics. We have shown that ~35% of KO functional groups in *Microcystis* were significantly correlated to the phage transcript expression, indicating that the cells were responding to the viral activity. Many of these KO functional groups seemed to be associated with nucleotide recycling and urea utilization. Given that nucleotide recycling is a major component of virus replication, this seems like a rational relationship. We have also shown that ~45% of the detected cellular metabolites from the entire community were significantly correlated with virus marker expression. The high levels of UDP sugars observed have been documented before during viral infection of a marine heterotroph. Those authors suggested that these sugars may reflect changes in cell wall integrity as result of virus activity (Ankrah et al. 2014). Other metabolites, such as sedoheptulose bisphosphate and

fructose-1,6 bisphosphate, are involved in the pentose phosphate pathways which has been hypothesized to provide extra energy and nucleotide precursors during viral infections of other cyanobacterial species (Thompson et al. 2011). Although we cannot determine the abundance of metabolites produced solely by *Microcystis*, it is possible that the community, including *Microcystis,* is responding to the viral activity generated "virocell" state (*cf.* Forterre 2011) within this cyanobacterium: as Microcystis biochemistry changes due to infection the co-occuring community undergoes some parallel change. We note that not all viruses detected within these experiments demonstrated this variability between replicates. For example, myoviruses infecting *Synechococcus* provided very consistent, reproducible data across replicates, with elevated transcriptional activity observed in the ammonia-amended bottles that also had a higher *Synechocccaceae* host transcriptional activity (Supplemental Figure 16).

Without the analysis of viral transcription and metabolite pools, the presence and variation in the *Microcystis* viral infection could have been overlooked and variability might go unexplained. Indeed any resulting phage influence on *Microcystis* transcription might be misconstrued, leading to invalid conclusions on the dynamics of the effect of P. While not as dramatic in the current study, the same conclusions could be erroneously arrived at due to shifts in the function of the heterotrophic community. Functional relationships between *Microcystis* and its associated heterotrophic counterparts are still unclear, but many studies acknowledge a possible mutualistic relationship that can alter *Microcystis* growth dynamics (Cook et al. 2020; Kim et al. 2019). The "minor" community shifts that occur due to bottle effects have the potential to cause functional changes in *Microcystis,* introducing variability that might otherwise be ignored. Indeed,

the apparent reliance on ammonium regeneration within the heterotrophic community could be a source of such variability within just a 48 h timescale. To this end researchers need to careful consider the dynamics of any incubation experiments they undertake, as it is highly likely that this divergence due to "unintended biology" increases with length of time over any incubation period.

The complexity of biological systems cannot be overstated. Each organism (including viruses, bacteria, cyanobacteria and fungi) can play a pivotal role in cyanobacterial bloom dynamics. Going forward, we encourage all researchers to fully consider the presence, activity, and potential influence of each entity in the system in which they study. Moreover, beginning to cast a light on varability or "bottle effects" within basic experimental methods will not only provide a better understanding of the community of interes, but of biology (and all it possibilities) in general.

Data Availability

Raw reads are available on the NCBI SRA database under BioProject number PRJNA737197 and trimmed, non-ribosomal reads are available on MG-RAST (Keegan et al. 2016) under project name "LE2019MT."

Acknowledgements

We are grateful for the assistance of Cathy Buckingham and Eric Gann for their assistance in sample processing and analysis. This work was supported by funding from the NIEHS (1P01ES028939–01) and NSF (OCE-1840715) through the Great Lakes Center for Fresh Waters and Human Health at *Bowling Green State University*. The

metabolomic analysis was supported by funding from the NSF (CHE-1710175). We also acknowledge support from the Simons Foundation (735077) and the *Kenneth & Blaire Mossman* Endowment at The University of Tennessee.

CHAPTER 7: CONCLUSIONS

This body of work seeks to emphasize the complexity of heterogenous systems like *Microcystis* spp.- dominated harmful algal blooms. *Microcystis* spp. blooms can have severe ecological and economic ramifications, especially when microcystin toxin loads are high. Microcystins produced by *Microcystis* have caused regular summer challenges for municipal water treatment facilities around the globe and numerous water shortage crises. Combined with the unsightly biomass accumulation, *Microcystis* spp. have attracted heavy research interests, especially surrounding toxin production. While these blooms are known to have other organisms present, the majority of the research has focused on *Microcystis* alone, either through lab manipulations of a cultured isolate or field-based experiments of a natural water sample. This is understandable, as there are still so many unknowns regarding bloom occurrence, biomass accumulation, toxin production, and seasonal succession, just to name a few topics. However, *Microcystis* does not exist in isolation in nature, nor in many of the isolated lab cultures. Although some axenic strains exist, most studies have shown that *Microcystis* grows more efficiently when grown with a community of heterotrophic bacteria, that are naturally present in nature. The natural microbial community not only contains heterotrophic bacteria, but also other phototropic bacteria, phototrophic eukaryotes, and viruses. These other members are occasionally acknowledged but are rarely investigated in *Microcystis* research. This dissertation seeks to characterize this understudied community and how it can influence *Microcystis* spp. dynamics, while also improving the methods to do so.

Only by considering all members and functional activities present can researchers hope to fully understand how and why a bloom occurs.

Our investigation of the community began with the viral community, specifically the phage that can directly infect *Microcystis.* Using a massive metatranscriptomic dataset from a *Microcystis* spp.- dominated bloom in Lake Tai, China, we sought to further characterize a viral lysis event that had been previously detected (Stough et al. 2017). The original publication observed that this virus was closely related to the Ma-LMM01 *Microcystis* phage and noted a shift in viral transcript expression, with earlier bloom months showing higher levels of a gene associated with lytic reproduction and later bloom months showing higher levels of a gene associated with lysogenic reproduction. However, the original paper determined this by recruiting reads to the genome of the closest known relative, the Ma-LMM01 phage. While this method is legitimate, it tends to obscure any variations in the natural diversity present. We improved on this method using a BLASTx approach (described in detail in earlier chapters) to identify and quantify the true richness present in both *Microcystis* viruses and in the *Microcystis* hosts. We discovered several things regarding the relationship between the *Microcystis* species (or genotypes) present and their viruses.

To begin, we confirmed the findings from the original publication regarding the presence of a Ma-LMM01-like phage and the shift from lytic to lysogenic expression. We characterized many different genotypes of this virus; however they were all expressed in a similar seasonal pattern. In addition, we characterized two additional viruses, both classified as Siphoviridae. One has already been observed in other Chinese lakes, while the other has not yet been characterized, and appears to have once been a lysogenic

phage, as remnants of its genome can be found in a sequenced *Microcystis* isolate genome. This discovery provided more evidence that *Microcystis* spp. may indeed be capable of lysogeny, although that has yet to be proven. The expression of both Siphoviridae virus types detected was heaviest in the later months, when the Ma-LMM01-like phage showed heavy lysogenic expression. This suggests that there may have been some shift in the viral community, allowing for increased lysis by other viruses with the dominant Ma-LMM01 phage was not actively lysing the community. In addition to the variation detected in the viral community, there were also many species of *Microcystis* present, including *M. flos-aquae, M. wesenbergii, M. panniformis, M. viridis,* and *M. aeruginosa.* However, these species did not shift seasonally with the expression of the viral reproductive genes, but rather maintained a mixed composition of all species in all months. This indicated that phylogeny was not controlled by or controlling viral expression or lysis events. Yet investigations of toxin producing genes did show a shift in expression, similar to the Ma-LMM01-like phage. We do not conclude that the viruses are controlling the production of toxin, but that both viruses and toxin are likely being acted upon by a common environmental force. By correlating the environmental parameters collected during initial sampling, we were able to identify pH as that common environmental parameter. Further experimentation and investigation would be required to confirm and characterize that relationship. However, the findings from this paper concluded that *Microcystis* viruses do directly influence *Microcystis* populations through lysis events, but it may be that environmental variables influence viral dynamics more than *Microcystis* phylogeny.

This work was expanded on by the investigation of all of the other viruses present in bloom, those that do not directly infect *Microcystis*, but rather, the other co-occurring community members present. Using the same BLASTx approach used in the first study, we characterized the nucleocytoplasmic large DNA viruses (NCLDVs), RNA viruses, ssDNA viruses, virophage, and other DNA phage present in the same samples from the Lake Tai, China dataset. This analysis revealed high expression of RNA viruses and NCLDVs closely related to viruses that commonly infect photosynthetic eukaryotes. Many water systems that are dominated by *Microcystis* spp. in the summer are dominated by photosynthetic eukaryotes like diatoms in the winter. While there was little expression of other phage, the presence and expression of eukaryotic viruses suggests that these viruses may be playing a role in the suppression of their hosts. This promotes hypotheses regarding a modified "kill the winner" strategy, in which viruses infect and reduce the population of the dominant competitor, providing a competitive advantage to the weaker competitor, which would be *Microcystis* spp. in this case. There are likely many other factors at work here, such as pH, which was already implicated in the control of community dynamics in this system. However, this work clearly supports the continued investigation of the microbial community that exists alongside *Microcystis* in bloom events.

In addition to the known importance of heterotrophic bacteria in *Microcystis* spp. growth dynamics, the potential role for virus-mediated community succession further prompted our investigation of the co-occurring host community. While our methods for detecting and quantifying viral activity is the best available given the dearth of viral sequences and tools in existence, we saw an opportunity to improve upon our

characterization of the hist community. The methods discussed so far either did a good job at characterizing the phylogeny and expression of a single gene (the BLASTx) approach or did a good job at charactering the expression of all genes in a single organism (the genome recruitment approach). Neither of these would suffice to characterize the functional expression of an entire community. So, we undertook an evaluation of a many methods available to establish a "best practice" method for continued analysis of the host community present. We assessed both online, user-friendly sequence analysis tools as well as locally based, coding pipelines, evaluating everything from data input type to transcript expression estimates. This assessment included the evaluation of an existing method using a novel component, where we recruited reads to combined "Frankenstein" genome, comprised of many *Microcystis* strains. Our comparative analysis revealed two superior methods for characterizing a bloom community. For researchers primarily focused on *Microcystis* spp., we recommended read recruitment to Frankenstein's *Microcystis*, where the included genetic heterogeneity from the multiple strains allows for increased detection and enumeration of *Microcystis* spp. reads present. For researchers who intend to focus on the entire microbial community, we recommended the GhostKOALA online platform, which allows for functional and taxonomic characterization of sequences, followed by read recruitment which then quantifies the expression of those annotated sequences.

The GhostKOALA approach to characterizing the community provided an excellent means to re-evaluate existing datasets and the opportunity to answer new questions. We sought to emphasize the importance of the co-occurring microbially community using an existing dataset from simple, lab-based nutrient amendment

experiment that might commonly occur in labs around the world (Steffen et al. 2014b). Even though the experiment used a single, non-axenic *Microcystis* strain we were able to show large shifts in community taxonomy and function after three generations, when cultures were exposed to different nitrogen sources. These results prompted additional questions regarding the how these changes in the community might influence *Microcystis* response, and whether the observed functional shifts in *Microcystis* response in the initial publication were truly the result of the nutrient amendment or the result in the shift in the co-occurring community. These questions highlight the crucial need for a full community characterization when evaluating *Microcystis* spp. blooms.

After first characterizing the entire community in a "simple" cultured isolate, we set out to characterize the natural community in an environmental experiment. For this expansion, we described both the co-occurring microbial host community and the co-occurring viral community. While basic water quality parameters suggested that the system was phototrophically constrained by phosphorus bioavailability, we observed shifts in the functions and taxonomies of the microbial community associated with all nutrient additions (urea, ammonia, nitrate, and phosphate) after 48 hours and even noticed variable shifts within replicate bottles. We also observed remarkable shifts in the control bottles after 48 hours. This indicates that merely putting bloom water into a bottle can influence the community present, potentially skewing conclusions. Bottle based experiments are common, therefore this observation is of notable importance and suggests that researchers must consider this shift when interpreting experimental results. The co-occurring community was primarily dominated by heterotrophic bacteria, with little expression of eukaryotes or their viruses. We found that Ma-LMM01-like

Microcystis phages were the most expressed viruses in the system, although their expression remained fairly constant, with the exception of a single bottle. One phosphate addition bottle showed significantly more phage, indicating a potential lysis event that occurred during the 48 hours. The single replicate highlights the incredible heterogeneity of a natural system and how this variability can be observed in experimental results. Indeed, many of the *Microcystis* genes were differentially expressed in the bottle and we even seen significant correlations to a variety of community metabolites, indicating that the variability in the *Microcystis* phage alone can shape the entire community. Without a full characterization of the host and viral community surrounding *Microcystis*, many *Microcystis* observations may be overlooked as "error" in an experiment.

Overall, this dissertation provides robust justification for the inclusion of the entire community when researching *Microcystis* spp.-dominated harmful algal blooms. The work presented here not only provides the rationale for this avenue of study but also provides the methodological approaches to do so. Careful evaluation of the co-occurring microbial and viral community must be integrated into *Microcystis* bloom research as the critical answers needed to address the global challenges these blooms pose cannot be answered through the study of *Microcystis* alone.

LIST OF REFERENCES

Amin, S. and others 2015. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. Nature **522:** 98-101.
Anderson, D. M. and others 2021. Marine harmful algal blooms (HABs) in the United States: history, current status and future trends. Harmful Algae **102:** 101975.
Anderson, D. M., P. M. Glibert, and J. M. Burkholder. 2002. Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. Estuaries **25:** 704-726.
Ankrah, N. Y. D. and others 2014. Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. The ISME Journal **8:** 1089-1100.
Arndt, D. and others 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research **44:** W16-W21.
Azevedo, S. M. and others 2002. Human intoxication by microcystins during renal dialysis treatment in Caruaru—Brazil. Toxicology **181:** 441-446.
Babicki, S. and others 2016. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Research **44:** W147-W153.
Basu, S., M. Gledhill, D. de Beer, S. P. Matondkar, and Y. Shaked. 2019. Colonies of marine cyanobacteria Trichodesmium interact with associated bacteria to acquire iron from dust. Communications biology **2:** 1-8.
Belisle, B. S. and others 2016. Urea in Lake Erie: Organic nutrient sources as potentially important drivers of phytoplankton biomass. Journal of Great Lakes Research **42:** 599-607.
Bell, W., and R. Mitchell. 1972. Chemotactic and growth responses of marine bacteria to algal extracellular products. The Biological Bulletin **143:** 265-277.
Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57:** 289-300.
Besemer, J., and M. Borodovsky. 1999. Heuristic approach to deriving models for gene finding. Nucleic acids research **27:** 3911-3920.
Bingham, M., S. Sinha, and F. Lupi. 2015. Economic benefits of reducing harmful algal blooms in Lake Erie. Environmental Consulting & Technology, Inc., Report **66**.
Bishop, C., E. Anet, and P. Gorham. 1959. Isolation and identification of the fast-death factor in Microcystis aeruginosa NRC-1. Canadian journal of biochemistry and physiology **37:** 453-471.
Blomqvist, P., A. Pettersson, and P. Hyenstrand. 1994. Ammonium-nitrogen: a key regulatory factor causing dominance of non-nitrogen-fixing cyanobacteria in aquatic systems. Archiv für Hydrobiologie**:** 141-164.
Boström, B., A.-K. Pettersson, and I. Ahlgren. 1989. Seasonal dynamics of a cyanobacteria-dominated microbial community in surface sediments of a shallow, eutrophic lake. Aquatic Sciences **51:** 153-178.
Botes, D. P. and others 1984. The structure of cyanoginosin-LA, a cyclic heptapeptide toxin from the cyanobacterium Microcystis aeruginosa. Journal of the Chemical Society, Perkin Transactions 1**:** 2311-2318.

Boyer, G. L. 2020. LCMS-SOP Determination of microcystins in water samples by high performance liquid chromatography (HPLC) with single quadrupole mass spectrometry (MS).

Bratbak, G., A. Jacobsen, M. Heldal, K. Nagasaki, and F. J. A. M. E. Thingstad. 1998. Virus production in Phaeocystis pouchetii and its relation to host cell growth and nutrition. **16:** 1-9.

Breitbart, M. 2012. Marine viruses: truth or dare. Ann Rev Mar Sci **4:** 425-448.

Breitbart, M., and F. Rohwer. 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol **13:** 278-284.

Bruchet, A., F. Bernazeau, I. Baudin, and P. Pieronne. 1998. Algal toxins in surface waters: analysis and treatment. Water Supply **16:** 619-623.

Brussaard, C. P. D., B. Kuipers, and M. J. W. Veldhuis. 2005. A mesocosm study of Phaeocystis globosa population dynamics. Harmful Algae **4:** 859-874.

Bullerjahn, G. S. and others 2016. Global solutions to regional problems: collecting global expertise to address the problem of harmful cyanobacterial blooms. A Lake Erie case study. Harmful Algae **54:** 223-238.

Burkholder, J. 2000. Critical needs in harmful algal bloom research. Opportunities for environmental applications of marine biotechnology national academy of sciences. National Research Council, Washington, DC**:** 126-149.

Cai, Y., F. Kong, L. Shi, and Y. Yu. 2012. Spatial heterogeneity of cyanobacterial communities and genetic variation of Microcystis populations within large, shallow eutrophic lakes (Lake Taihu and Lake Chaohu, China). Journal of environmental sciences **24:** 1832-1842.

Camacho, C. and others 2009. BLAST+: architecture and applications. BMC bioinformatics **10:** 1-9.

Carrillo, E. and others 2003. Interstrain variability in toxin production in populations of the cyanobacterium Microcystis aeruginosa from water-supply reservoirs of Andalusia and lagoons of Doñana National Park (southern Spain). Phycologia **42:** 269-274.

Chaffin, J. D., and T. B. Bridgeman. 2014. Organic and inorganic nitrogen utilization by nitrogen-stressed cyanobacteria during bloom conditions. Journal of Applied Phycology **26:** 299-309.

Chaffin, J. D., T. B. Bridgeman, and D. L. Bade. 2013. Nitrogen constrains the growth of late summer cyanobacterial blooms in Lake Erie. Advances in Microbiology **2013**.

Chaffin, J. D. and others 2019. Cyanobacterial blooms in the central basin of Lake Erie: Potentials for cyanotoxins and environmental drivers. Journal of Great Lakes Research **45:** 277-289.

Chambers, M. C. and others 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology **30:** 918-920.

Chen, H., L. Jing, Z. Yao, F. Meng, and Y. Teng. 2019. Prevalence, source and risk of antibiotic resistance genes in the sediments of Lake Tai (China) deciphered by metagenomic assembly: a comparison with other global lakes. Environment international **127:** 267-275.

Chinchar, V., A. Hyatt, T. Miyazaki, and T. Williams. 2009. Family *Iridoviridae*: poor viral relations no longer, p. 123-170. Lesser known large dsDNA viruses. Springer.
Chorus, I., and M. Welker. 2021. Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management. Taylor & Francis.
Clarke, K. R., and R. N. Gorley. 2015. PRIMER V7: user manual/tutorial. PRIMER-E.
Clasquin, M. F., E. Melamud, and J. D. Rabinowitz. 2012. LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine. Current protocols in bioinformatics **37:** 14.11. 11-14.11. 23.
Claverie, J.-M., and C. Abergel. 2018. *Mimiviridae*: an expanding family of highly diverse large dsDNA viruses infecting a wide phylogenetic range of aquatic eukaryotes. Viruses **10:** 506.
Cochlan, W. P., J. Wikner, G. F. Steward, D. C. Smith, and F. J. M. E.-P. S. Azam. 1993. Spatial distribution of viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments. **92:** 77-77.
Coleridge, S. T. 1857. Rime of the Ancient Mariner. Appleton & Company.
Cook, K. V. and others 2020. The global Microcystis interactome. Limnology and oceanography **65:** S194-S207.
Crow, W. B. 1923. The taxonomy and variation of the genus Microcystis in Ceylon. New Phytologist **22:** 59-68.
Dale, B., and C. M. Yentsch. 1978. Red Tide and Paralytic Shellfish Poisoning. Oceanus.
Davenport, E. J. and others 2019. Metatranscriptomic analyses of diel metabolic functions during a Microcystis bloom in Western Lake Erie (United States). Frontiers in microbiology **10:** 2081.
Davis, T. W. and others 2010. Effects of nitrogenous compounds and phosphorus on the growth of toxic and non-toxic strains of Microcystis during cyanobacterial blooms. **61:** 149-162.
DeBruyn, J. M., J. A. Leigh-Bell, R. M. L. McKay, R. A. Bourbonniere, and S. W. Wilhelm. 2004. Microbial distributions and the impact of phosphorus on bacterial activity in Lake Erie. Journal of Great Lakes Research **30:** 166-183.
Donnelly-Wu, M. K., W. R. Jacobs Jr, and G. F. Hatfull. 1993. Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria. Molecular Microbiology **7:** 407-417.
Edgar, R. and others 2016. Adaptations to photoautotrophy associated with seasonal ice cover in a large lake revealed by metatranscriptome analysis of a winter diatom bloom. Journal of Great Lakes Research **42:** 1007-1015.
Ehlke, T. A. 1974. Comparison of Bacterial and Phytoplankton Populations Under Natural and Laboratory Conditions. US Department of the Interior **2:** 533-539.
Eiler, A., and S. Bertilsson. 2004. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. Environmental microbiology **6:** 1228-1243.
Fischer, M. G., and C. A. Suttle. 2011. A virophage at the origin of large DNA transposons. Science **332:** 231-234.
Flower, R. J. 1982. The occurrence of an epiphytic diatom on Microcystis aeruginosa. The Irish Naturalists' Journal**:** 553-555.

Forterre, P. 2011. Manipulation of cellular synthases and the nature of viruses: the virocell concept. Compets Rendus Chimie **14:** 392-399.
Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28:** 3150-3152.
Fuhrman, J. A. 1999. Marine viruses and their biogeochemical and ecological effects. Nature **399:** 541.
Gaia, M. and others 2014. Zamilon, a novel virophage with Mimiviridae host specificity. PLoS One **9:** e94923.
Gann, E. R., Y. Kang, S. Dyhrman, C. J. Gobler, and S. W. Wilhelm. 2021. Metatranscriptome library preparation influences analyses of viral community activity during a brown tide bloom. Frontiers in Microbiology **12:** 1126.
Gann, E. R., H. L. Pound, and S. W. Wilhelm. 2019. Python Script: Extracting aligned portion of viral hallmark genes from blastx sequence.
George, D., and R. Edwards. 1976. The effect of wind on the distribution of chlorophyll a and crustacean plankton in a shallow eutrophic reservoir. Journal of Applied Ecology**:** 667-690.
Gerloff, G. C., G. P. Fitzgerald, and F. Skoog. 1950. The isolation, purification, and culture of blue-green algae. American Journal of Botany**:** 216-218.
Glibert, P. M., D. M. Anderson, P. Gentien, E. Granéli, and K. G. Sellner. 2005. The global, complex phenomena of harmful algal blooms. Oceanography.
Gong, C. and others 2016. Novel virophages discovered in a freshwater lake in China. Frontiers in Microbiology **7:** 5.
Goriushyn, V., and S. Chaplinskyaya. 1968. The discovery of viruses lysing blue-green algae in the Dneprovsk reservoirs. Tsvetenie Vody. Dumka, Kiev.
Grattan, L. M., S. Holobaugh, and J. G. Morris Jr. 2016. Harmful algal blooms and public health. Harmful algae **57:** 2-8.
Grover, J. P. 1990. Resource competition in a variable environment: phytoplankton growing according to Monod's model. The American Naturalist **136:** 771-789.
Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology **59:** 307-321.
Hallegraeff, G. 2003. Harmful algal blooms: a global overview. Manual on harmful marine microalgae **33:** 1-22.
Hallegraeff, G., H. Enevoldsen, and A. Zingone. 2021. Global harmful algal bloom status reporting. Elsevier.
Hampel, J. J., M. J. McCarthy, M. Neudeck, G. S. Bullerjahn, R. M. L. McKay, and S. E. Newell. 2019. Ammonium recycling supports toxic Planktothrix blooms in Sandusky Bay, Lake Erie: Evidence from stable isotope and metatranscriptome data. Harmful Algae **81:** 42-52.
Harada, K.-I. 1996. Trace analysis of microcystins. Phycologia **35:** 36-41.
Harke, M. J., and C. J. Gobler. 2015. Daily transcriptome changes reveal the role of nitrogen in controlling microcystin synthesis and nutrient transport in the toxic cyanobacterium, Microcystis aeruginosa. BMC genomics **16:** 1068.

Harke, M. J., and C. J. J. P. O. Gobler. 2013. Global transcriptional responses of the toxic cyanobacterium, Microcystis aeruginosa, to nitrogen stress, phosphorus stress, and growth on organic matter. **8:** e69834.
Harke, M. J. and others 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. Harmful Algae **54:** 4-20.
Hennon, G. M., and S. T. Dyhrman. 2020. Progress and promise of omics for predicting the impacts of climate change on harmful algal blooms. Harmful algae **91:** 101587.
Hewson, I., J. M. O'Neil, and W. C. Dennison. 2001. Virus-like particles associated with *Lyngbya majuscula* (Cyanophyta; Oscillatoriacea) bloom decline in Moreton Bay, Australia. Aquatic Microbial Ecology **25:** 207-213.
Ho, K.-C., S.-H. Kang, and I. H. Lam. 2003. Distribution of Alexandrium tamarense in Drake Passage and the threat of harmful algal blooms in the Antarctic Ocean. Ocean and Polar Research **25:** 625-631.
Hoagland, P., D. M. Anderson, Y. Kaoru, and A. W. White. 2002. The economic effects of harmful algal blooms in the United States: estimates, assessment issues, and information needs. Estuaries **25:** 819-837.
Hooper, D. U., and P. M. Vitousek. 1997. The effects of plant composition and diversity on ecosystem processes. Science **277:** 1302-1305.
Hudnell, H. K. 2010. The state of US freshwater harmful algal blooms assessments, policy and legislation. Toxicon **55:** 1024-1034.
Hughes, E. D., P. R. Gorham, and A. Zhender. 1958. Toxicity of a unialgal culture of *Microcystis aeruginosa*. Can J Microbiol **4:** 225-236.
Hutchinson, G. E. 1961. The paradox of the plankton. The American Naturalist **95:** 137-145.
Hyenstrand, P., P. Blomqvist, and A. Pettersson. 1998. Factors determining cyanobacterial success in aquatic systems: a literature review. Archiv für Hydrobiologie Spec. Iss. Advances in Limnology **51:** 41-62.
Jacobson, L., and M. Halmann. 1982. Polyphosphate metabolism in the blue-green alga Microcystis aeru-ginosa. Journal of Plankton Research **4:** 481-488.
Jacoby, J. M., D. C. Collier, E. B. Welch, F. J. Hardy, and M. Crayton. 2000. Environmental factors associated with a toxic bloom of Microcystis aeruginosa. Canadian Journal of Fisheries and Aquatic Sciences **57:** 231-240.
Jankowiak, J. G., and C. J. Gobler. 2020. The composition and function of microbiomes within Microcystis colonies are significantly different than native bacterial assemblages in two North American Lakes. Frontiers in Microbiology **11:** 1016.
Johnson, D. 1956. Systematic and ecological notes on the Cladocera of Lake Toba, and the surrounding country, North Sumatra. Zoological Journal of the Linnean Society **43:** 72-91.
Kaebernick, M., B. A. Neilan, T. Börner, and E. Dittmann. 2000. Light and the transcriptional response of the microcystin biosynthesis gene cluster. Applied and environmental microbiology **66:** 3387-3392.

Kanehisa, M., Y. Sato, and K. Morishima. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of molecular biology **428:** 726-731.
Kaneko, T. and others 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium Microcystis aeruginosa NIES-843. DNA research **14:** 247-256.
Ke, Z., P. Xie, and L. Guo. 2008. Controlling factors of spring–summer phytoplankton succession in Lake Taihu (Meiliang Bay, China). Hydrobiologia **607:** 41-49.
Keegan, K. P., E. M. Glass, and F. Meyer. 2016. MG-RAST, a metagenomics service for analysis of microbial community structure and function, p. 207-233. Microbial Environmental Genomics (MEG). Springer.
Kim, M., B. Shin, J. Lee, H. Y. Park, and W. Park. 2019. Culture-independent and culture-dependent analyses of the bacterial community in the phycosphere of cyanobloom-forming Microcystis aeruginosa. Scientific reports **9:** 1-13.
Kim, S.-G., S.-H. Joung, C.-Y. Ahn, S.-R. Ko, S. M. Boo, and H.-M. Oh. 2010. Annual variation of Microcystis genotypes and their potential toxicity in water and sediment from a eutrophic reservoir. FEMS microbiology ecology **74:** 93-102.
Kimura, S., Y. Sako, and T. Yoshida. 2013. Rapid microcystis cyanophage gene diversification revealed by long- and short-term genetic analyses of the tail sheath gene in a natural pond. Appl Environ Microbiol **79:** 2789-2795.
Kirchman, D. L. 2016. Growth rates of microbes in the oceans. Annual review of marine science **8:** 285-309.
Komárek, J., and J. Komárková. 2002. Review of the European Microcystis morphospecies (Cyanoprokaryotes) from nature. Fottea **2:** 1-24.
Kondo, R., G. Kagiya, S. Hiroishi, and M. Watanabe. 2000. Genetic typing of a bloom-forming cyanobacterial genus Microcystis in Japan using 16S rRNA gene sequence. Plankton Biol. Ecol **47:** 1-6.
Koonin, E. V., V. V. Dolja, and M. Krupovic. 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology **479:** 2-25.
Kopylova, E., L. Noé, and H. Touzet. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28:** 3211-3217.
Kranzler, C. F. and others 2019. Silicon limitation facilitates virus infection and mortality of marine diatoms. Nature microbiology**:** 1-8.
Krausfeldt, L. E., A. T. Farmer, H. C. Gonzalez, B. N. Zepernick, S. R. Campagna, and S. W. Wilhelm. 2019a. Urea is both a carbon and nitrogen source for *Microcystis aeruginosa*: tracking ^{13}C incorporation at bloom pH conditions. Frontiers in Microbiology **10:** 1064.
Krausfeldt, L. E., M. M. Steffen, R. M. McKay, G. S. Bullerjahn, G. L. Boyer, and S. W. Wilhelm. 2019b. Insight into the molecular mechanisms for microcystin biodegradation in Lake Erie and Lake Taihu. Frontiers in microbiology **10:** 2741.
Krausfeldt, L. E. and others 2017. Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu. FEMS Microbiology Ecology **93**.
Krausfeldt, L. E., and S. W. Wilhelm. 2017. RNA extraction from Sterivex filters.
Krüger, G., and J. Eloff. 1978. The effect of temperature on specific growth rate and activation energy of Microcystis and Synechococcus isolates relevant to the onset

of natural blooms. Journal of the Limnological Society of Southern Africa **4:** 9-20.
Kuhlisch, C., G. Schleyer, N. Shahaf, F. Vincent, D. Schatz, and A. Vardi. 2021. Viral infection of algal blooms leaves a unique metabolic footprint on the dissolved organic matter in the ocean. Science Advances **7:** eabf4680.
Kurmayer, R., and T. Kutzenberger. 2003. Application of real-time PCR for quantification of microcystin genotypes in a population of the toxic cyanobacterium Microcystis sp. Applied and Environmental Microbiology **69:** 6723-6730.
La Scola, B. and others 2008. The virophage as a unique parasite of the giant mimivirus. Nature **455:** 100.
Latta, L. C. and others 2011. Species and genotype diversity drive community and ecosystem properties in experimental microcosms. Evolutionary Ecology **25:** 1107-1125.
Letunic, I., and P. Bork. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research.
Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31:** 1674-1676.
Li, M., W. Zhu, L. Gao, J. Huang, and L. Li. 2013. Seasonal variations of morphospecies composition and colony size of Microcystis in a shallow hypertrophic lake (Lake Taihu, China). Fresen Environ Bull **22:** 3474-3483.
Li, Q. and others 2018. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming Microcystis-epibiont communities. Frontiers in microbiology **9:** 746.
Litchman, E., and C. A. Klausmeier. 2001. Competition of phytoplankton under fluctuating light. The American Naturalist **157:** 170-187.
Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology **15:** 1-21.
Lu, W., M. F. Clasquin, E. Melamud, D. Amador-Noguez, A. A. Caudy, and J. D. Rabinowitz. 2010. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Analytical Biochemistry **82:** 3212-3221.
Luria, S. E., and M. Delbrück. 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics **28:** 491.
Lwoff, A. J. B. R. 1953. Lysogeny. **17:** 269.
Margalef, R. 1978. Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanologica Acta **1:** 493-509.
Margulies, M. and others 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376-380.
Martens, L. and others 2011. mzML—a community standard for mass spectrometry data. Molecular & Cellular Proteomics **10**.
Martin, J. H., and S. E. Fitzwater. 1988. Iron deficiency limits phytoplankton growth in the north-east pacific subarctic. Nature**:** 331:341-343.

Martin, R. M. and others 2019. Cylindrospermopsis raciborskii Virus and host: Genomic characterization and ecological relevance.
Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics **11:** 538.
McLachlan, J., and P. R. Gorham. 1962. Effects of pH and Nitrogen Sources on Growth of Microcystis aeruginosa Kütz. Canadian Journal of Microbiology **8:** 1-11.
Melamud, E., L. Vastag, and J. D. Rabinowitz. 2010. Metabolomic analysis and visualization engine for LC− MS data. Analytical chemistry **82:** 9818-9826.
Menzel, P., K. L. Ng, and A. Krogh. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications **7:** 1-9.
Meyer, K. A., T. W. Davis, S. B. Watson, V. J. Denef, M. A. Berry, and G. J. Dick. 2017. Genome sequences of lower Great Lakes Microcystis sp. reveal strain-specific genes that are present and expressed in western Lake Erie blooms. PLoS One **12:** e0183859.
Miller, M. A. and others 2010. Evidence for a novel marine harmful algal bloom: cyanotoxin (microcystin) transfer from land to sea otters. PLoS One **5:** e12576.
Moniruzzaman, M., L. L. Wurch, H. Alexander, S. T. Dyhrman, C. J. Gobler, and S. W. Wilhelm. 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nature Communications **8:** 16054.
Morimoto, D., S. Kimura, Y. Sako, and T. J. F. i. m. Yoshida. 2018. Transcriptome Analysis of a Bloom-Forming Cyanobacterium Microcystis aeruginosa during Ma-LMM01 Phage Infection. **9:** 2.
Morimoto, D. and others 2019. Cooccurrence of broad-and narrow-host-range viruses infecting the bloom-forming toxic cyanobacterium Microcystis aeruginosa. Applied and environmental microbiology **85:** e01170-01119.
Morris, J. J., R. E. Lenski, and E. R. Zinser. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio **3:** e00036-00012.
Mullis, K. B. 1990. The unusual origin of the polymerase chain reaction. Scientific American **262:** 56-65.
Murray, A. G., and G. A. Jackson. 1992. Viral dynamics: a model of the effects of size, shape, motion and abundance of single-celled planktonic organisms and other particles. Marine Ecology Progress Series**:** 103-116.
Nagasaki, K., Y. Tomaru, K. Nakanishi, N. Hata, N. Katanozaka, and M. Yamaguchi. 2004. Dynamics of *Heterocapsa circularisquama* (*Dinophyceae*) and its viruses in Ago Bay, Japan. Aquatic Microbial Ecology **34:** 219-226.
Neilan, B. A., P. T. Cox, P. R. Hawkins, and A. E. Goodman. 1994. 16S ribosomal RNA gene sequence and phylogeny of toxic Microcystis sp.(cyanobacteria). DNA Sequence **4:** 333-337.
Niu, Y. and others 2011. Phytoplankton community succession shaping bacterioplankton community composition in Lake Taihu, China. Water Research **45:** 4169-4182.
Organization, W. H. 2020. Cyanobacterial toxins: microcystins. World Health Organization.

Orr, P. T., and G. J. Jones. 1998. Relationship between microcystin production and cell division rates in nitrogen-limited Microcystis aeruginosa cultures. Limnology and oceanography **43:** 1604-1614.
Otsuka, S., S. Suda, S. Shibata, H. Oyaizu, S. Matsumoto, and M. M. Watanabe. 2001. A proposal for the unification of five species of the cyanobacterial genus Microcystis Kützing ex Lemmermann 1907 under the rules of the Bacteriological Code. International Journal of Systematic and Evolutionary Microbiology **51:** 873-879.
Otten, T. G., and H. W. Paerl. 2011. Phylogenetic inference of colony isolates comprising seasonal Microcystis blooms in Lake Taihu, China. Microbial ecology **62:** 907-918.
Ou, T., S. Li, X. Liao, and Q. Zhang. 2013. Cultivation and characterization of the MaMV-DC cyanophage that infects bloom-forming cyanobacterium Microcystis aeruginosa. Virologica Sinica **28:** 266-271.
Ouellette, A. J., S. M. Handy, and S. W. Wilhelm. 2006. Toxic Microcystis is widespread in Lake Erie: PCR detection of toxin genes and molecular characterization of associated cyanobacterial communities. Microbial ecology **51:** 154-165.
Ouellette, A. J., and S. W. Wilhelm. 2003. Toxic cyanobacteria: the evolving molecular toolbox. Frontiers in Ecology and the Environment **1:** 359-366.
Padan, E., and M. Shilo. 1973. Cyanophages-viruses attacking blue-green algae. Bacteriological Reviews **37:** 343-370.
Paerl, H. W., and T. G. Otten. 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol **65:** 995-1010.
Paerl, H. W. and others 2016. It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. **50:** 10805-10813.
Parajuli, A. and others 2016. A Unique Tryptophan C-Prenyltransferase from the Kawaguchipeptin Biosynthetic Pathway. Angewandte Chemie **128:** 3660-3663.
Peng, G. and others 2018. Seasonally relevant cool temperatures interact with N chemistry to increase microcystins produced in lab cultures of Microcystis aeruginosa NIES-843. Environmental science & technology **52:** 4127-4136.
Penn, K., J. Wang, S. C. Fernando, and J. R. Thompson. 2014. Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. The ISME journal **8:** 1866-1878.
Popper, K. 1959. The logic of scientific discovery. Routledge.
Pound, H., E. R. Gann, and S. Wilhelm. 2021a. Creating a Frankenstein's Genome.
---. 2021b. Functional and taxonomic characterization of sequence data using GhostKOALA.
Pound, H., and S. Wilhelm. 2019. Metatranscriptomic screening for genes of interest.
---. 2020a. RNA extraction from Sterivex using phenol:chloroform.
---. 2021. Sequence processing and assembly workflow using CLC Workbench, SortMeRNA, and MegaHit.
Pound, H. L. and others 2020. The "Neglected Viruses" of Taihu: Abundant Transcripts for Viruses Infecting Eukaryotes and Their Potential Role in Phytoplankton Succession. Frontiers in Microbiology **11:** 338.

---. 2021c. Environmental Studies of Cyanobacterial Harmful Algal Blooms Should Include Interactions with the Dynamic Microbiome. Environmental Science & Technology.

Pound, H. L., and S. W. Wilhelm. 2020b. Tracing the active genetic diversity of Microcystis and Microcystis phage through a temporal survey of Taihu. Plos one **15:** e0244482.

Qin, B. and others 2010. A drinking water crisis in Lake Taihu, China: linkage to climatic variability and lake management. Environmental Management **45:** 105-112.

Ren, J., N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome **5:** 69.

Reynolds, C. S., G. Jaworski, H. Cmiech, and G. Leedale. 1981. On the annual cycle of the blue-green alga Microcystis aeruginosa Kütz. emend. Elenkin. Philosophical Transactions of the Royal Society of London. B, Biological Sciences **293:** 419-477.

Rinehart, K. L., M. Namikoshi, and B. W. Choi. 1994. Structure and biosynthesis of toxins from blue-green algae (cyanobacteria). Journal of applied phycology **6:** 159-176.

Rinta-Kanto, J. M., E. A. Konopko, J. M. DeBruyn, R. A. Bourbonniere, G. L. Boyer, and S. W. Wilhelm. 2009. Lake Erie Microcystis: relationship between microcystin production, dynamics of genotypes and environmental parameters in a large lake. Harmful algae **8:** 665-673.

Rinta-Kanto, J. M., A. J. A. Ouellette, G. L. Boyer, M. R. Twiss, T. B. Bridgeman, and S. W. Wilhelm. 2005. Quantification of toxic *Microcystis* spp. during the 2003 and 2004 blooms in western Lake Erie using quantitative real-time PCR. Environ Sci Technol **39:** 4198-4205.

Rinta-Kanto, J. M., and S. W. Wilhelm. 2006. Diversity of microcystin-producing cyanobacteria in spatially isolated regions of Lake Erie. Applied and environmental microbiology **72:** 5083-5085.

Roux, S., F. Enault, B. L. Hurwitz, and M. B. Sullivan. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ **3:** e985.

Rudi, K., O. M. Skulberg, F. Larsen, and K. S. Jakobsen. 1997. Strain characterization and classification of oxyphotobacteria in clone cultures on the basis of 16S rRNA sequences from the variable regions V6, V7, and V8. Applied and environmental microbiology **63:** 2593-2599.

Sandrini, G., H. C. Matthijs, J. M. Verspagen, G. Muyzer, and J. Huisman. 2014. Genetic diversity of inorganic carbon uptake systems causes variation in CO 2 response of the cyanobacterium Microcystis. The ISME journal **8:** 589-600.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences **74:** 5463-5467.

Saxton, M. A., N. A. D'souza, R. A. Bourbonniere, R. M. L. McKay, and S. W. Wilhelm. 2012. Seasonal Si: C ratios in Lake Erie diatoms—evidence of an active winter diatom community. Journal of Great Lakes Research **38:** 206-211.

Schelske, C. 1984. In situ and natural phytoplankton assemblage bioassays. Algae as ecological indicators**:** 15-47.
Schindler, D. W. 1998. Whole-ecosystem experiments: replicatrion versus realism: the need for ecosystem-scale experiments. Ecosystems **1:** 323-334.
Seitzinger, S. P. 1991. The effect of pH on the release of phosphorus from Potomac Estuary sediments: Implications for blue-green algal blooms. Estuarine, Coastal and Shelf Science **33:** 409-418.
Sellner, K. G., G. J. Doucette, and G. J. Kirkpatrick. 2003. Harmful algal blooms: causes, impacts and detection. Journal of Industrial Microbiology and Biotechnology **30:** 383-406.
Seymour, J. R., S. A. Amin, J.-B. Raina, and R. Stocker. 2017. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. Nature microbiology **2:** 1-12.
Shakya, M., C.-C. Lo, and P. S. Chain. 2019. Advances and challenges in metatranscriptomic analysis. Frontiers in genetics **10:** 904.
Shelley, M. W. 1818. Frankenstein; or, The modern Prometheus. Lackington, HUghes, Harding, Mavor, and Jones.
Shirai, Y., Y. Tomaru, Y. Takao, H. Suzuki, T. Nagumo, and K. Nagasaki. 2008. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus* Meunier. Applied and Environmental Microbiology **74:** 4022-4027.
Skerman, V. B. 1960. A Guide to the Identification of the Genera of Bacteria. Academic Medicine **35:** 92.
Smith, D. J., J. Y. Tan, M. A. Powers, X. N. Lin, T. W. Davis, and G. J. Dick. 2021. Individual Microcystis colonies harbour distinct bacterial communities that differ by Microcystis oligotype and with time. Environmental Microbiology.
Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27:** 431-432.
Spivak, A. C., M. J. Vanni, and E. M. Mette. 2011. Moving on up: can results from simple aquatic mesocosm experiments be applied across broad spatial scales? Freshwater Biology **56:** 279-291.
Steffen, M. M., B. S. Belisle, S. B. Watson, G. L. Boyer, R. A. Bourbonniere, and S. W. Wilhelm. 2015. Metatranscriptomic evidence for co-occurring top-down and bottom-up controls on toxic cyanobacterial communities. Appl Environ Microbiol **81:** 3268-3276.
Steffen, M. M., B. S. Belisle, S. B. Watson, G. L. Boyer, and S. W. Wilhelm. 2014a. Review: Status, causes and consequences of cyanobacterial blooms in Lake Erie. J Gt Lakes Res **40:** 215-225.
Steffen, M. M. and others 2017. Ecophysiological examination of the Lake Erie *Microcystis* bloom in 2014: linkages between biology and the water supply shutdown of Toledo, OH. Environmental Science and Technology **51:** 6745-6755.
---. 2014b. Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in *Microcystis*. The ISME Journal **8:** 2080.

Sterner, R. W. 2008. On the phosphorus limitation paradigm for lakes. International Review of Hydrobiology **93:** 433-445.
Steward, G. F., A. I. Culley, J. A. Mueller, E. M. Wood-Charlson, M. Belcaid, and G. Poisson. 2013. Are we missing half of the viruses in the ocean? The Isme Journal **7:** 672.
Stewart, F. J. and others 2012. Experimental incubations elicit profound changes in community transcription in OMZ bacterioplankton. PLoS One **7:** e37118.
Stothard, P., and D. S. Wishart. 2005. Circular genome visualization and exploration using CGView. Bioinformatics **21:** 537-539.
Stough, J. M., M. Kolton, J. E. Kostka, D. J. Weston, D. A. Pelletier, and S. W. Wilhelm. 2018. Diversity of active viral infections within the *Sphagnum* microbiome. Applied and Environmental Microbiology **84:** e01124-01118.
Stough, J. M. and others 2017. Molecular prediction of lytic vs lysogenic states for Microcystis phage: Metatranscriptomic evidence of lysogeny during large bloom events. PloS One **12:** e0184146.
---. 2019. Genome and Environmental Activity of a *Chrysochromulina parva* Virus and its Virophages. Frontiers in Microbiology **10**.
Sullivan, M. B., J. S. Weitz, and S. W. Wilhelm. 2017. Viral ecology comes of age. Environmental Microbiology Reports **9:** 33-35.
Susskind, M. M., and D. J. V. Botstein. 1980. Superinfection exclusion by λ prophage in lysogens ofSalmonella typhimurium. **100:** 212-216.
Suttle, C. A. 2005. Viruses in the sea. Nature **437:** 356-361.
Tang, X., J. Chao, Y. Gong, Y. Wang, S. W. Wilhelm, and G. Gao. 2017. Spatiotemporal dynamics of bacterial community composition in large shallow eutrophic Lake Taihu: High overlap between free-living and particle-attached assemblages. Limnology and Oceanography **62:** 1366-1382.
Tang, X. and others 2018. Seasonal gene expression and the ecophysiological implications of toxic *Microcystis aeruginosa* blooms in Lake Taihu. Environmental Science and Technology **52:** 11049-11059.
Thingstad, T., and R. Lignell. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology **13:** 19-27.
Thingstad, T. F. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography **45:** 1320-1328.
Thompson, L. R. and others 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences **108:** E757-E764.
Tillett, D., D. L. Parker, and B. A. Neilan. 2001. Detection of toxigenicity by a probe for the microcystin synthetase A gene (mcyA) of the cyanobacterial genus Microcystis: comparison of toxicities with 16S rRNA and phycocyanin operon (phycocyanin intergenic spacer) phylogenies. Applied and environmental microbiology **67:** 2810-2818.

Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie, and E. Siemann. 1997. The influence of functional diversity and composition on ecosystem processes. Science **277:** 1300-1302.
Tilman, D., P. B. Reich, J. Knops, D. Wedin, T. Mielke, and C. Lehman. 2001. Diversity and productivity in a long-term grassland experiment. Science **294:** 843-845.
Tomaru, Y., K. Tarutani, M. Yamaguchi, and K. Nagasaki. 2004. Quantitative and qualitative impacts of viral infection on a *Heterosigma akashiwo* (Raphidophyceae) bloom in Hiroshima Bay, Japan. Aquatic Microbial Ecology **34:** 227-238.
Tucker, S., and P. Pollard. 2005. Identification of cyanophage Ma-LBP and infection of the cyanobacterium Microcystis aeruginosa from an Australian subtropical lake by the virus. Appl Environ Microbiol **71:** 629-635.
Urrutia-Cordero, P., M. K. Ekvall, and L. A. Hansson. 2015. Responses of cyanobacteria to herbivorous zooplankton across predator regimes: who mows the bloom? Freshwater Biology **60:** 960-972.
Valles, S. and others 2017. ICTV virus taxonomy profile: Dicistroviridae. The Journal of General Virology **98:** 355.
Valouev, A. and others 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome research **18:** 1051-1063.
van der Merwe, D., L. Sebbag, J. C. Nietfeld, M. T. Aubel, A. Foss, and E. Carney. 2012. Investigation of a Microcystis aeruginosa cyanobacterial freshwater harmful algal bloom associated with acute microcystin toxicosis in a dog. Journal of Veterinary Diagnostic Investigation **24:** 679-687.
Van Valen, L. 1977. The red queen. The American Naturalist **111:** 809-810.
Vanderploeg, H. A. and others 2001. Zebra mussel (*Dreissena polymorpha*) selective filtration promoted toxic *Microcystis* blooms in Saginaw Bay (Lake Huron) and Lake Erie. Can J Fish Aquat Sci **58:** 1208-1221.
Vanni, M. J., and J. Temte. 1990. Seasonal patterns of grazing and nutrient limitation of phytoplankton in a eutrophic lake. Limnology and Oceanography **35:** 697-709.
Venter, J. C. and others 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:** 66-74.
Visser, P. M. and others 2016. How rising CO2 and global warming may stimulate harmful cyanobacterial blooms. Harmful Algae **54:** 145-159.
Våge, S., J. E. Storesund, and T. F. Thingstad. 2013. Adding a cost of resistance description extends the ability of virus–host model to explain observed patterns in structure and function of pelagic microbial communities. Environmental Microbiology **15:** 1842-1852.
Wang, J. and others 2019. Interaction between cyanophage MaMV-DC and eight Microcystis strains, revealed by genetic defense systems. Harmful algae **85:** 101699.
Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10:** 57-63.
Weinbauer, M. G. 2004. Ecology of prokaryotic viruses. FEMS Microbiol Rev **28:** 127-181.

Welschmeyer, N. A. 1994. Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. Limnology and oceanography **39:** 1985-1992.
Wetzel, R. G. 1983. Recommendations for future research on periphyton, p. 339-346. Periphyton of freshwater ecosystems. Springer.
Wigington, C. H. and others 2016. Re-examination of the relationship between marine virus and microbial cell abundances. Nature Microbiology **1:** 15024.
Wilhelm, S. W., S. M. Brigden, and C. A. Suttle. 2002. A dilution technique for the direct measurement of viral production: a comparison in stratified and tidally mixed coastal waters. Microbial Ecology **43:** 168-173.
Wilhelm, S. W., G. S. Bullerjahn, and R. M. McKay. 2020. The complicated and confusing ecology of *Microcystis* blooms. mBIO **11:** e00529-00520.
Wilhelm, S. W. and others 2003. Effect of phosphorus amendments on present day plankton communities in pelagic Lake Erie. **32:** 275-285.
---. 2011. The relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu (Lake Tai), China. Harmful Algae **10:** 207-215.
Wilhelm, S. W., and C. A. Suttle. 1999. Viruses and nutrient cycles in the sea. BioScience **49:** 781-788.
Wilson, W. H., G. A. Tarran, D. Schroeder, M. Cox, J. Oke, and G. Malin. 2002. Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. Journal of the Marine Biological Association of the United Kingdom **82:** 369-377.
Winter, C., T. Bouvier, M. G. Weinbauer, and T. F. Thingstad. 2010. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. Microbiol. Mol. Biol. Rev. **74:** 42-57.
Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences **74:** 5088-5090.
Wommack, E. K., and R. R. Colwell. 2000. Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews **64:** 69-114.
Worm, J., and M. Søndergaard. 1998. Dynamics of heterotrophic bacteria attached to Microcystis spp.(Cyanobacteria). Aquatic Microbial Ecology **14:** 19-28.
Wright, S. 1931. Evolution in Mendelian populations. Genetics **16:** 97.
Xie, M. and others 2016. Metagenomic analysis reveals symbiotic relationship among bacteria in Microcystis-dominated community. Frontiers in Microbiology **7:** 56.
Xu, H., H. W. Paerl, B. Qin, G. Zhu, and G. Gaoa. 2010. Nitrogen and phosphorus inputs control phytoplankton growth in eutrophic Lake Taihu, China. Limnology and Oceanography **55:** 420-432.
Yang, F. and others 2020. Genomic Analysis of Mic1 Reveals a Novel Freshwater Long-Tailed Cyanophage. Frontiers in Microbiology **11:** 484.
Yoshida, T. and others 2008. Ma-LMM01 infecting toxic Microcystis aeruginosa illuminates diverse cyanophage genome strategies. **190:** 1762-1772.
---. 2006. Isolation and characterization of a cyanophage infecting the toxic cyanobacterium *Microcystis aeruginosa*. Applied Environmental Microbiology **72:** 1239-1247.

Yu, L., F. Kong, M. Zhang, Z. Yang, X. Shi, and M. Du. 2014. The dynamics of Microcystis genotypes and microcystin production and associations with environmental factors during blooms in Lake Chaohu, China. Toxins **6:** 3238-3257.

Zehnder, A., and P. R. Gorham. 1960. Factors influencing the growth of Microcystis aeruginosa Kütz. Emend. Elenkin. Canadian Journal of Microbiology **6:** 645-660.

Zhang, Z., X.-X. Zhang, B. Wu, J. Yin, Y. Yu, and L. Yang. 2016. Comprehensive insights into microcystin-LR effects on hepatic lipid metabolism using cross-omics technologies. Journal of hazardous materials **315:** 126-134.

Zhao, L., Y. Song, L. Li, N. Gan, J. J. Brand, and L. Song. 2018. The highly heterogeneous methylated genomes and diverse restriction-modification systems of bloom-forming *Microcystis*. Harmful Algae **75:** 87-93.

Zhou, J., D. Sun, A. Childers, T. R. McDermott, Y. Wang, and M. R. Liles. 2015. Three novel virophage genomes discovered from Yellowstone Lake metagenomes. Journal of Virology **89:** 1278-1285.

Zhou, J. and others 2013. Diversity of virophages in metagenomic data sets. Journal of Virology **87:** 4225-4236.

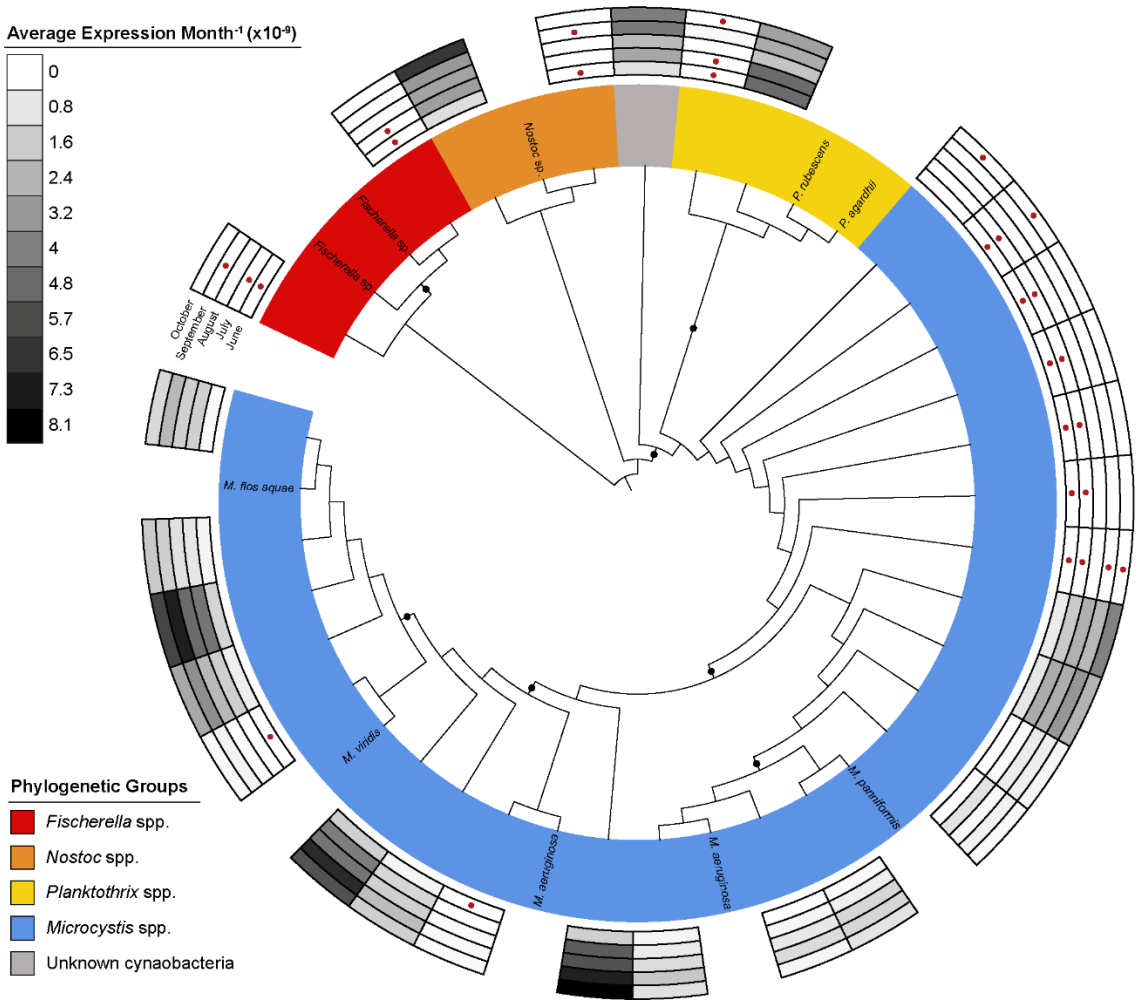
Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. Ab initio gene identification in metagenomic sequences. Nucleic acids research **38:** e132-e132.

Zingone, A., and H. O. Enevoldsen. 2000. The diversity of harmful algal blooms: a challenge for science and management. Ocean & Coastal Management **43:** 725-748.

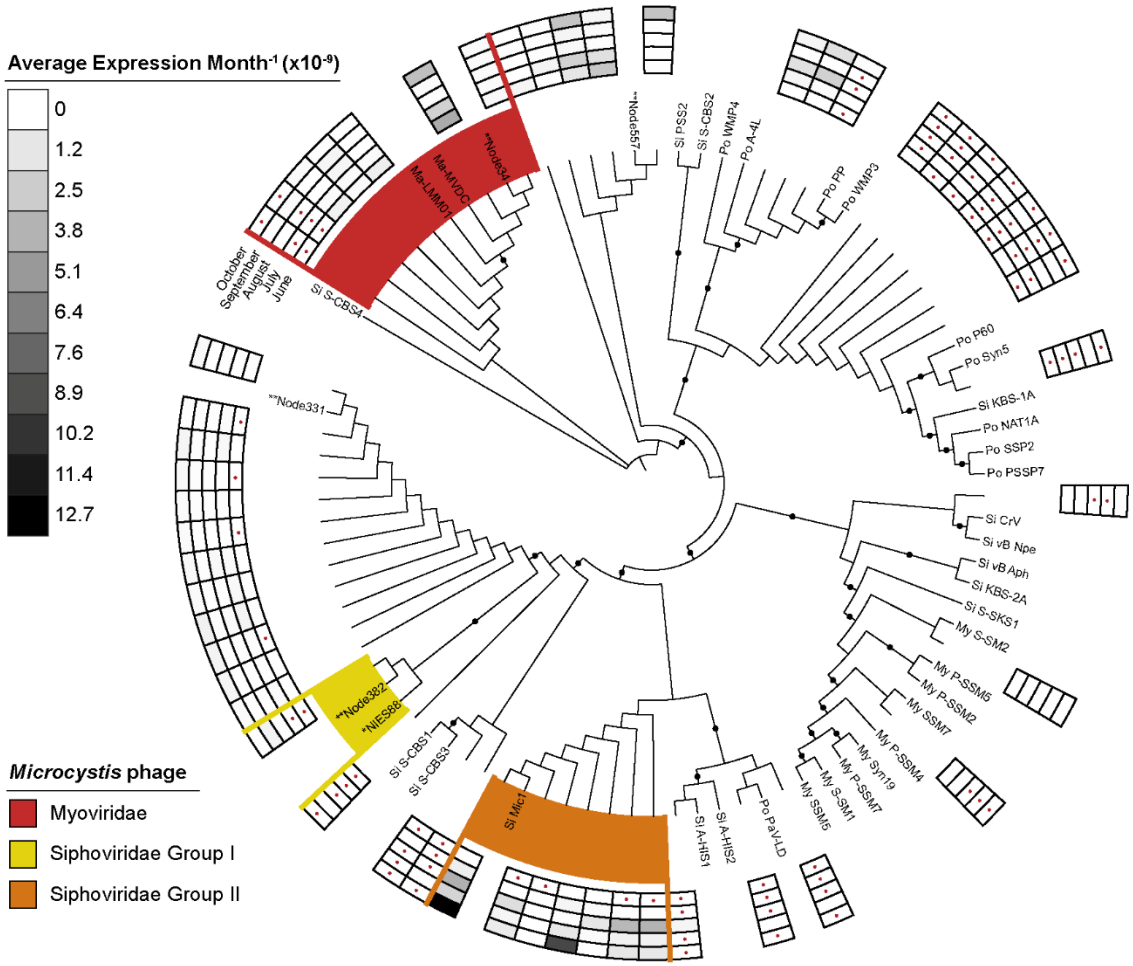
Zobell, C. E., and D. Q. Anderson. 1936. Observations on the multiplication of bacteria in different volumes of stored sea water and the influence of oxygen tension and solid surfaces. The biological bulletin **71:** 324-342.

Zurawell, R. W., H. Chen, J. M. Burke, and E. E. Prepas. 2005. Hepatotoxic cyanobacteria: a review of the biological importance of microcystins in freshwater environments. Journal of Toxicology and Environmental Health, Part B **8:** 1-37.

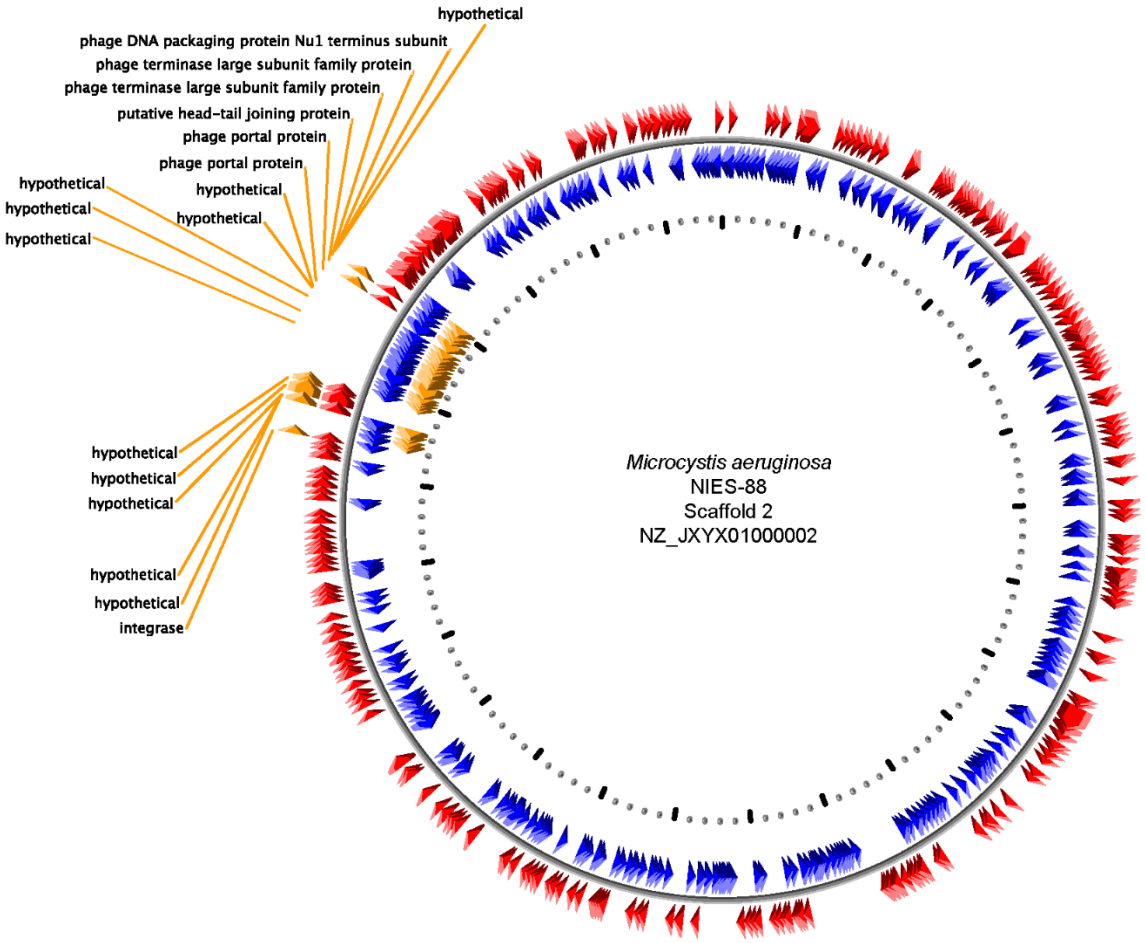
APPENDIX I: SUPPLEMENTAL FIGURES



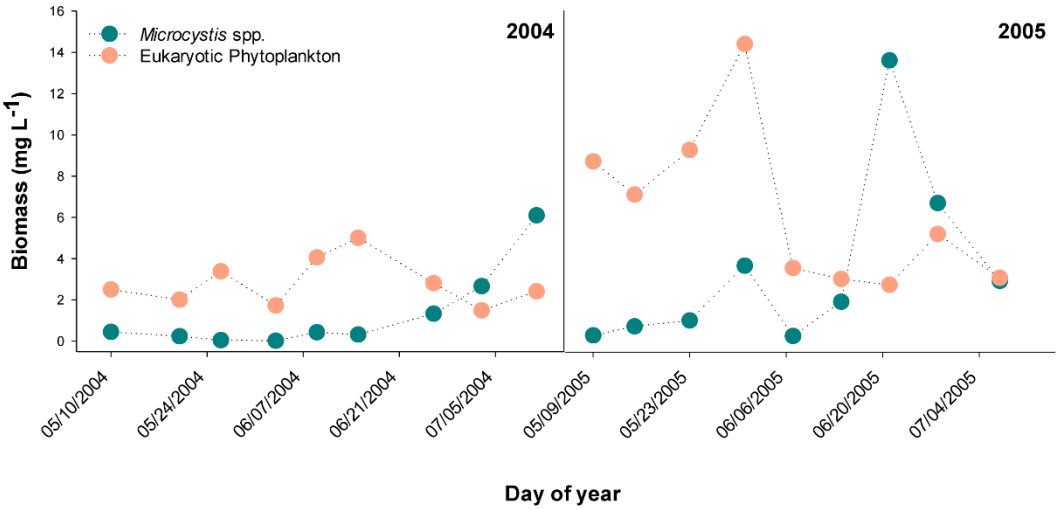
Supplemental Figure 1. Cladogram of toxin-encoding candidate contigs (*mcyA*). Inner color ring indicates taxonomic group and outer heatmap rings indicate the average expression of each candidate per month. Black dots indicate bootstrap values greater than 0.5 and red dots indicate no expression.



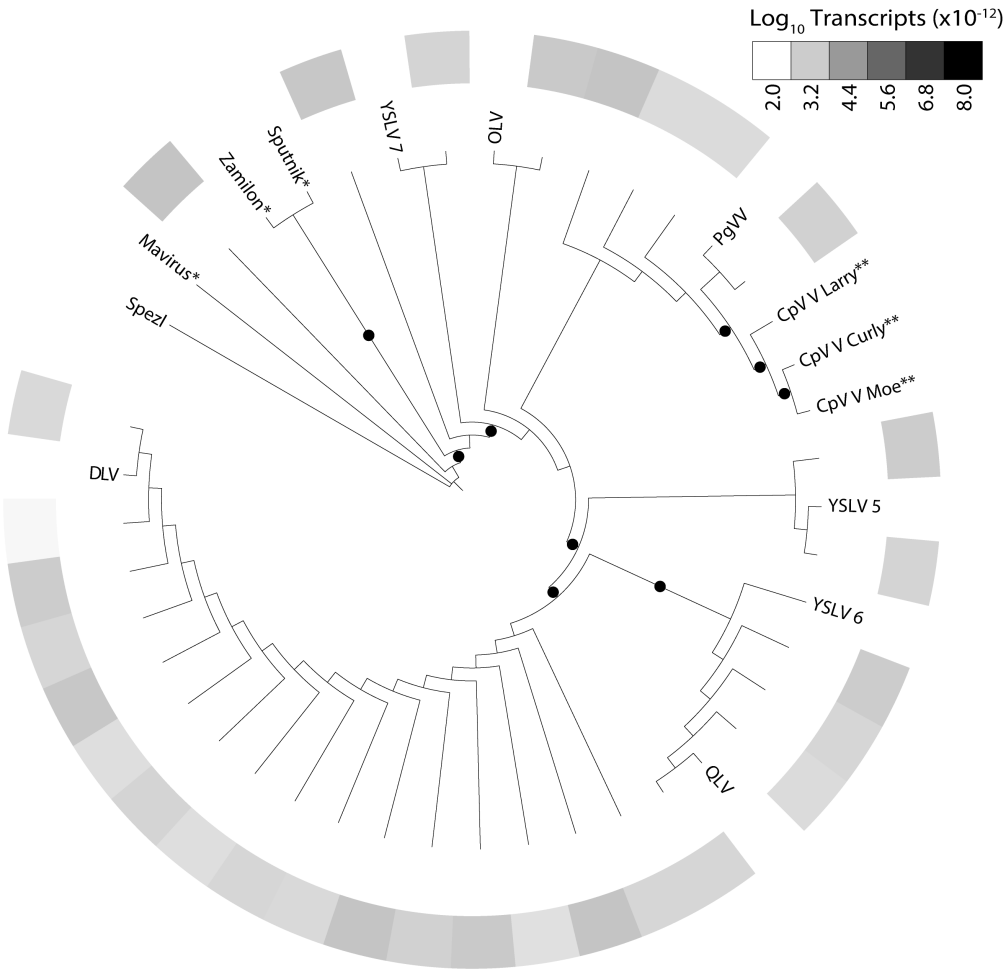
Supplemental Figure 2. Cladogram of phage terminase candidate contigs. Inner color ring indicates *Microcystis* phage group and outer heatmap rings indicate the average expression of each candidate per month. Black dots indicate bootstrap values greater than 0.5 and red dots indicate no expression.



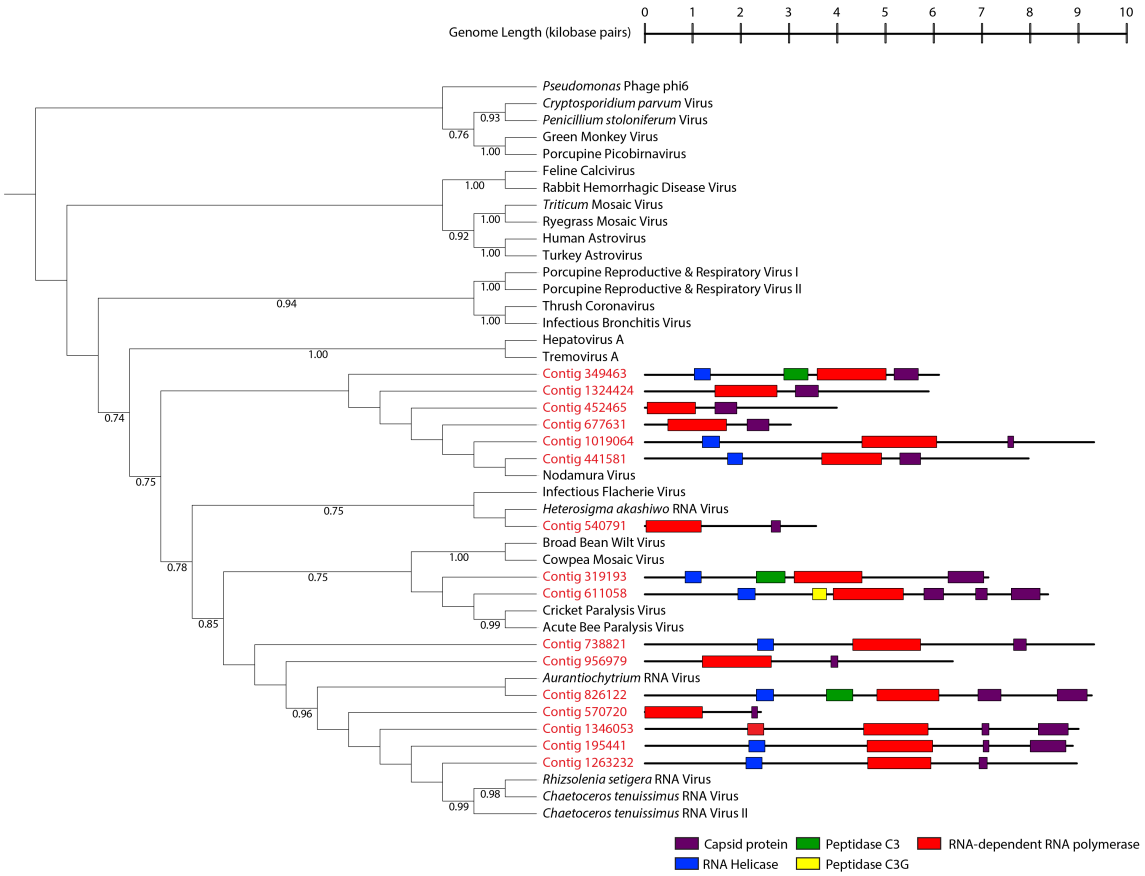
Supplemental Figure 3. Genome scaffold map of NIES-88 (Accession number NZ_JXYX010000002). Orange open reading frames indicate phage-like genes.



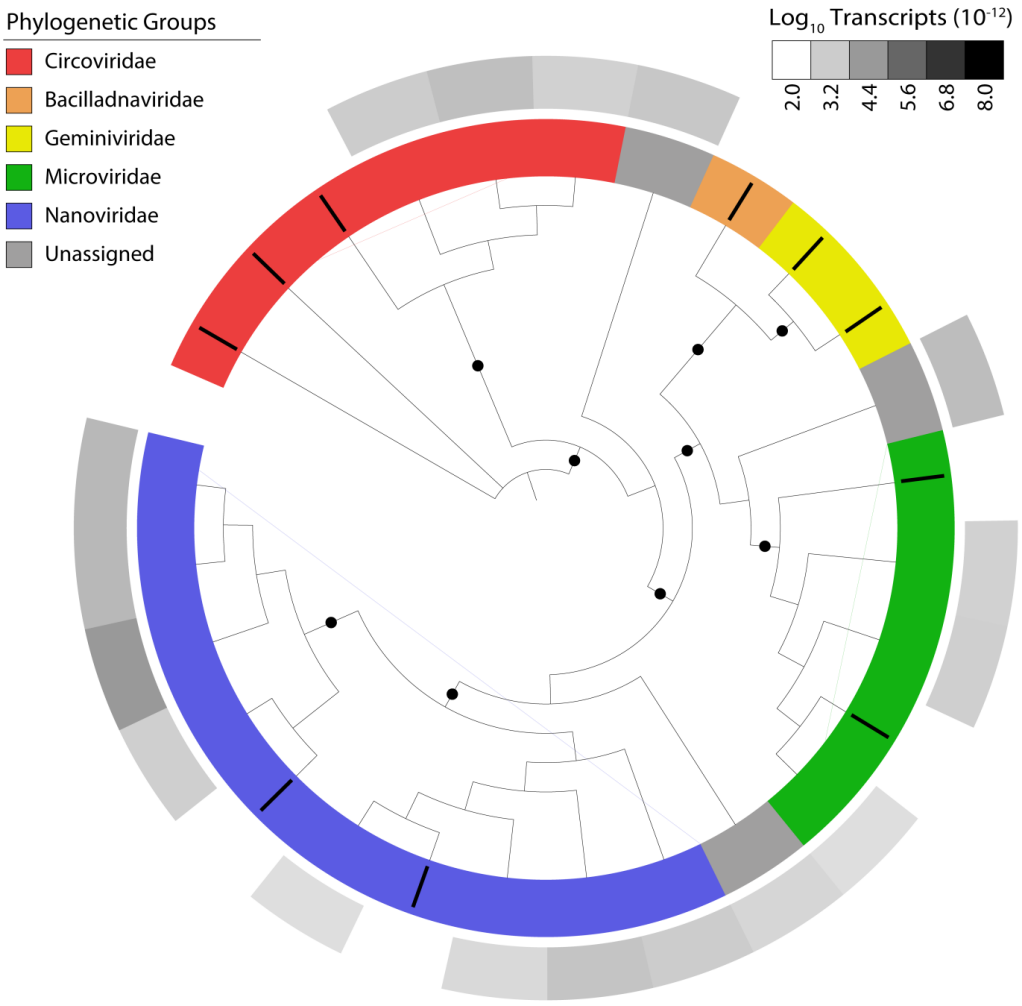
Supplemental Figure 4. Changes in biomass of *Microcystis* spp. And eukaryotic phytoplankton in Lake Taihu in 2004 and 2005. Adapted from Ke *et al.* 2008.



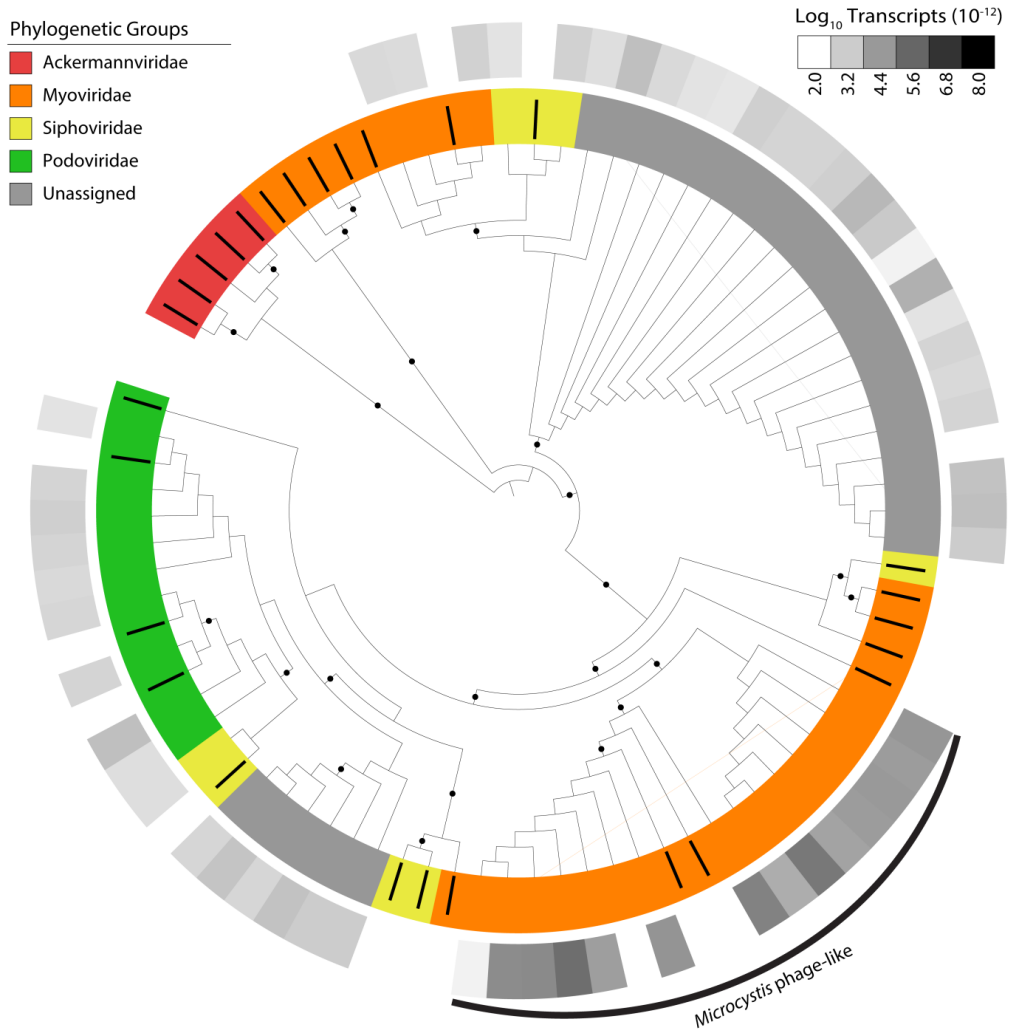
Supplemental Figure 5. Maximum-likelihood phylogenetic placement of virophage candidates. Reference proteins are denoted by abbreviated names. Outer heatmap ring indicates normalized transcript abundance for each candidate summed for all samples. Black dots represent bootstrap values >0.5. *Isolated reference with genome sequenced. **Isolated with only environmental sequencing.



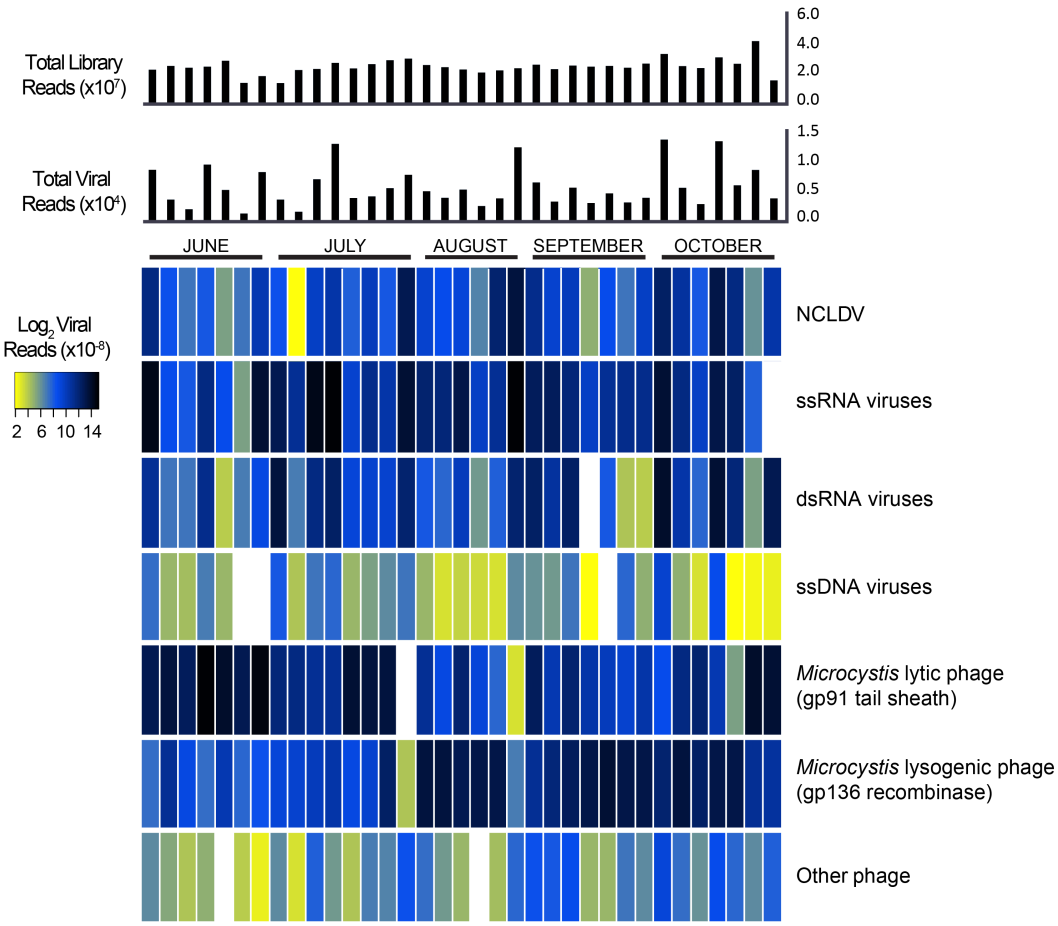
Supplemental Figure 6. Maximum-likelihood phylogenetic placement of near full-length RNA virus genomes with genomic content (B). Reference proteins are denoted by abbreviated names in black. Candidates with near full-length genomes are denoted in red (B).



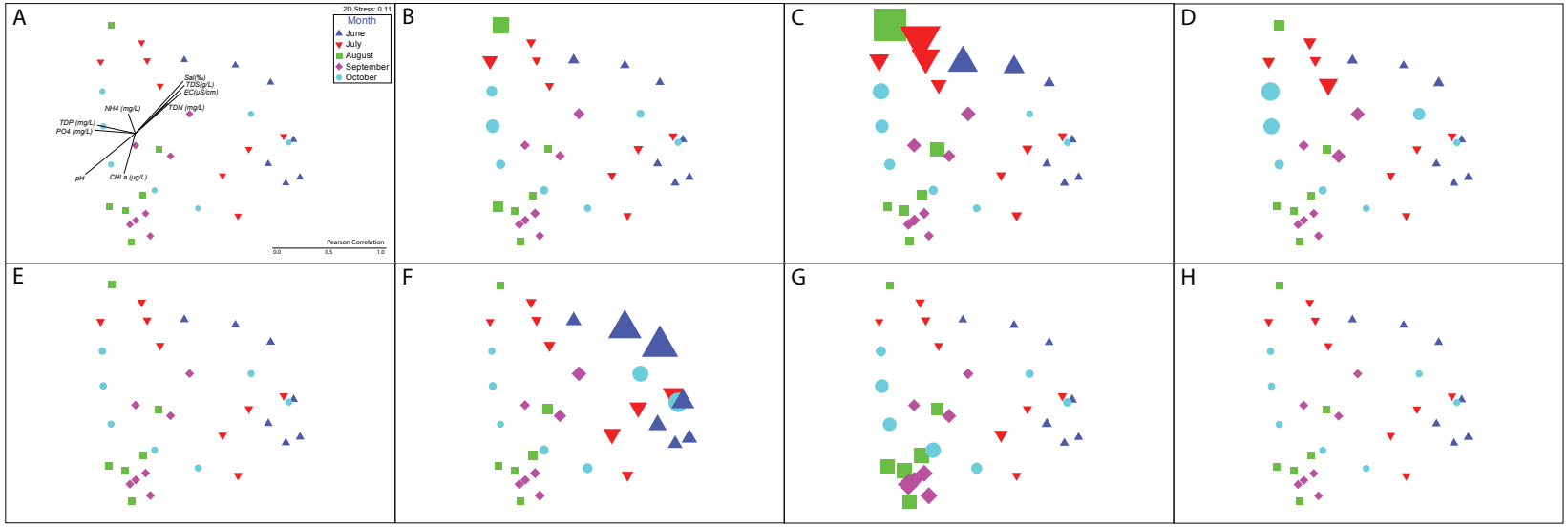
Supplemental Figure 7. Maximum-likelihood phylogenetic placement of ssDNA virus candidates. Reference proteins are denoted by black bars with colors indicating viral family. The outer heatmap ring indicates normalized transcript abundance for each candidate summed for all samples. Black dots represent bootstrap values >0.5.



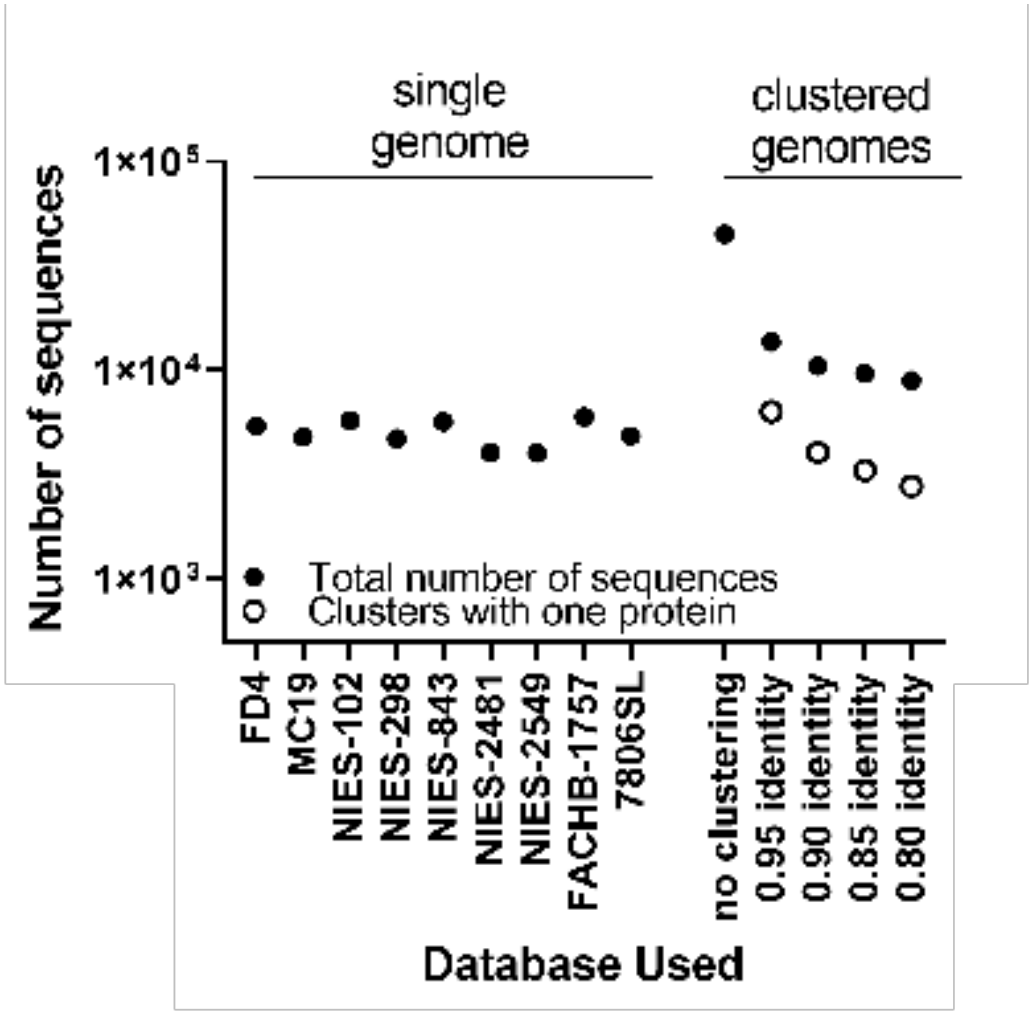
Supplemental Figure 8. Maximum-likelihood phylogenetic placement of phage candidates. Reference proteins are denoted by black bars with colors indicating viral family. The outer heatmap ring indicates normalized transcript abundance for each candidate summed for all samples. Black dots represent bootstrap values >0.5.



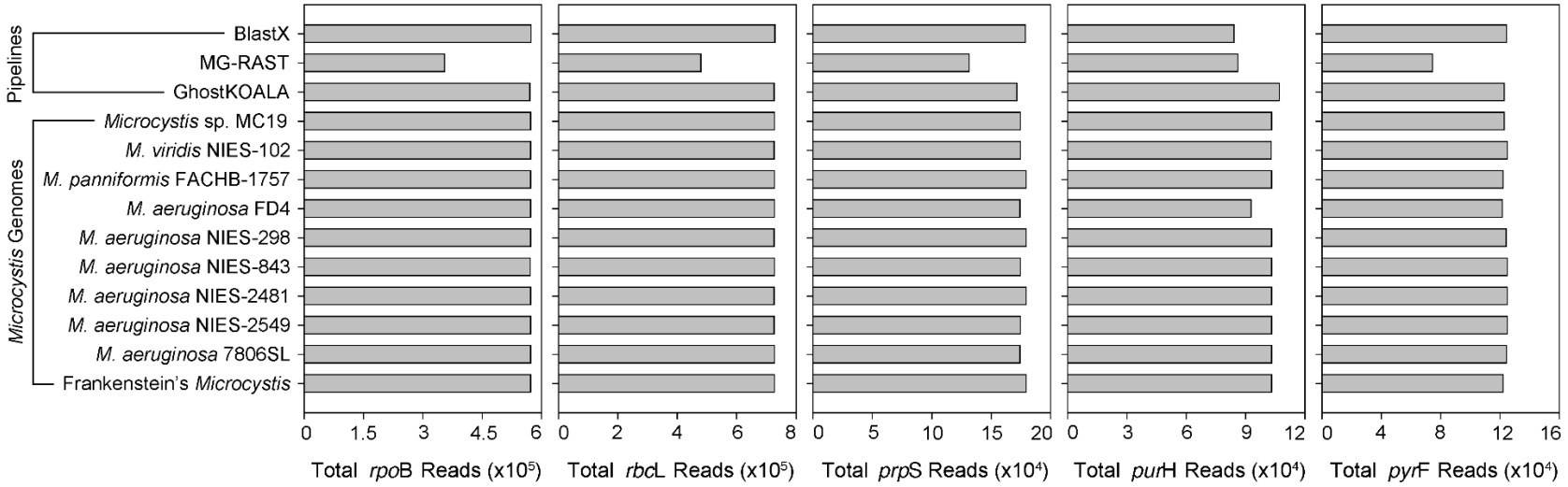
Supplemental Figure 9. Heatmap of normalized viral abundances for each sample and virus type with histogram of total viral reads and total library reads in each sample. Each column describes a single sample. Color scale describes viral read abundance normalized to library size.



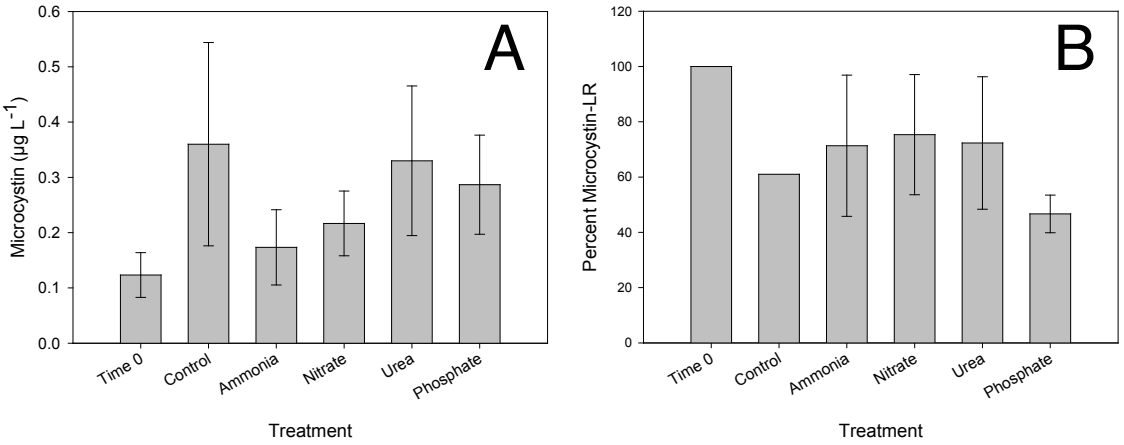
Supplemental Figure 10. Non-metric multidimensional scaling plot describing Bray-Curtis dissimilarity between 35 samples. Colors indicate month of sampling. Vectors indicate environmental variables that explain nMDS positioning (A). B-H show the same nMDS with each sample scaled to show the abundance of a particular virus type in each sample: NCLDV (B), ssRNA virus (C), dsRNA virus (D), ssDNA virus (E), *Microcystis* lytic phage (F), *Microcystis* lysogenic phage (G), "other" phage (H).



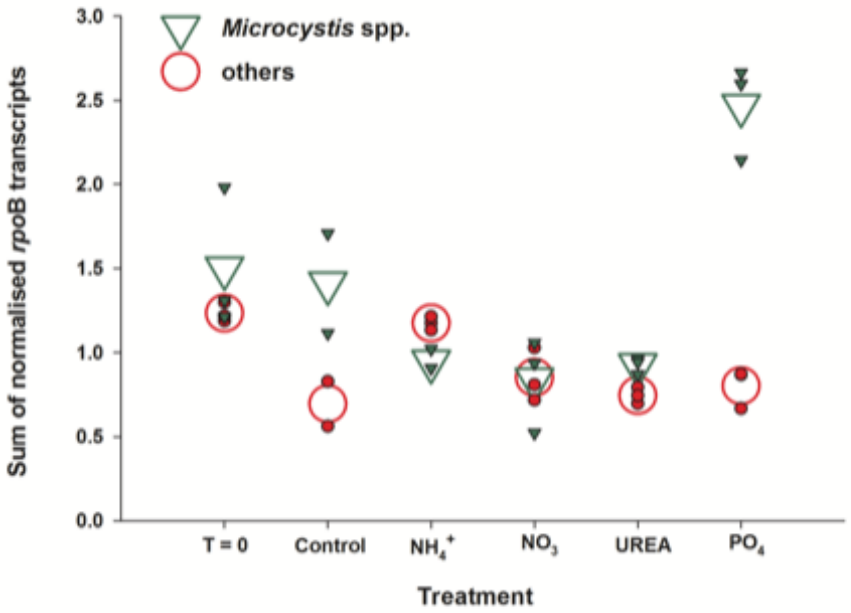
Supplemental Figure 11. Closed circles indicate the total number of coding sequences in each of the individual genomes and combined genomes of various stringencies. Open circles indicate the number of clusters that only contained a single-coding sequence in the combined genomes, indicating a lack of redundancy in those genes.



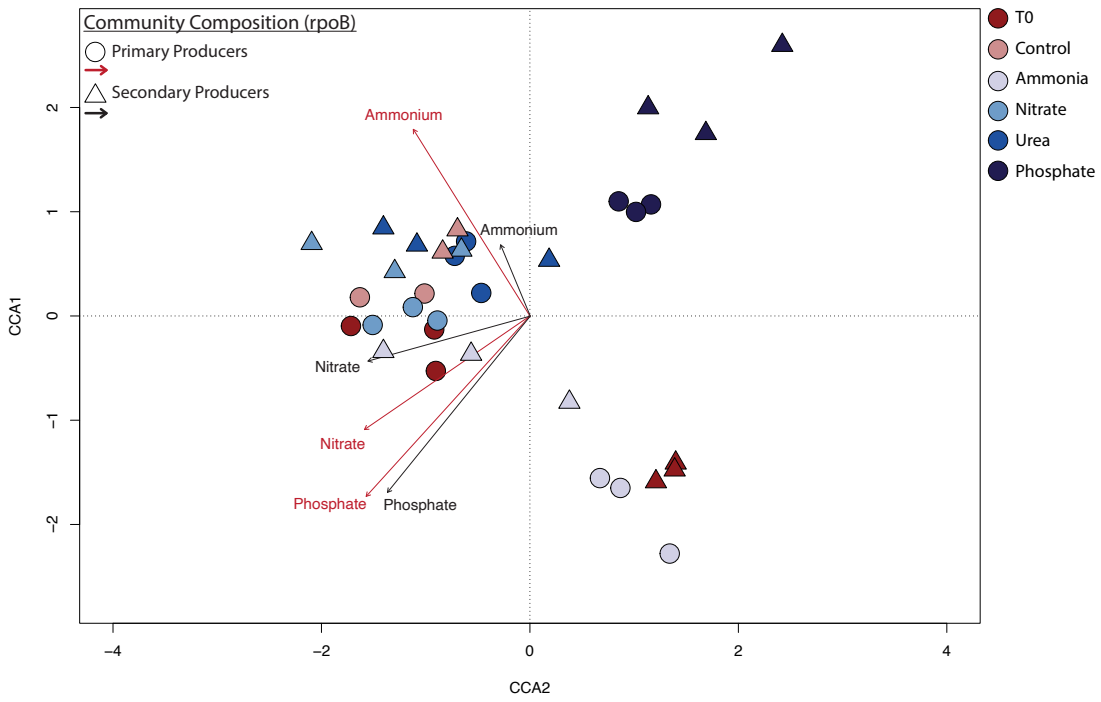
Supplemental Figure 12. The sum of all reads in all samples that were recruited to or classified as a specific *Microcystis* gene in each method.



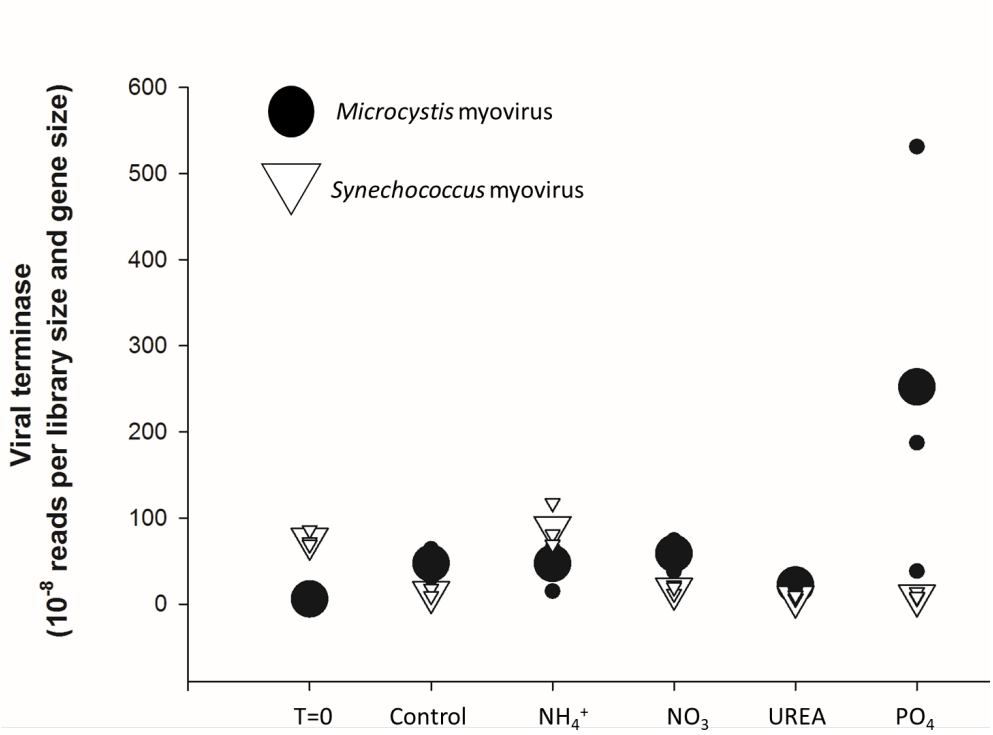
Supplemental Figure 13. Average micrograms of microcystin per liter in each treatment (A). Average percentage of microcystin-LR in each treatment (B). Error bars represent standard deviation.



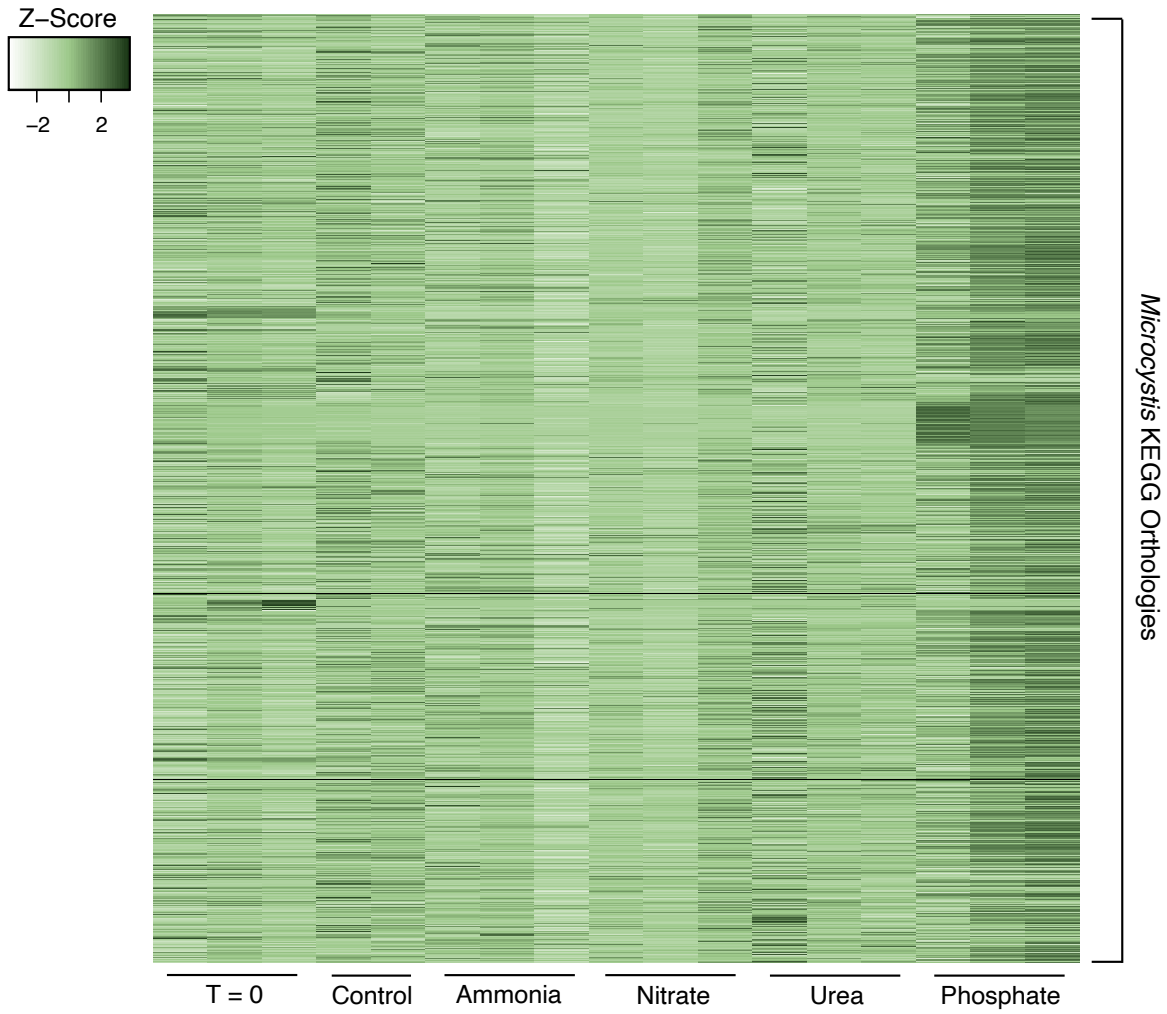
Supplemental Figure 14. Average (large markers) and individual bottle (small markers) expression of normalized transcript expression of the DNA-dependent RNA polymerase (*rpo*B) of *Microcystis* spp. (blue triangles) and the sum of all other co-occurring community members (red circles).



Supplemental Figure 15. Canonical correspondence analysis of *rpo*B expression, separated by primary producers (circles) and secondary producers (triangles). Nutrient treatment is distinguished by color.



Supplemental Figure 16. Average (large markers) and individual bottle (small markers) normalized transcript expression of a *Microcystis* phage marker (gp118 terminase, black circles) and a *Synechococcus* phage marker (gp 118 terminase, open triangles). Transcript expression was normalized to the library size of each sample and the length of each gene.



Supplemental Figure 17. Heatmap of normalized transcript expression of all *Microcystis* spp. KEGG orthology (KO) categories. Color scale has been scaled to each KO (z-score). The transcript expression for each annotated sequence was summed to each KO in each experimental bottle. Transcript expression was normalized to the library size of each sample and the length of each sequence. The list of Kos included, in the order presented here, can be found in Supplemental Data 1, Sheet "*Microcystis*_allKO_sums."

VITA

Helena L. Pound was born in Knoxville, TN and raised in Walland, TN. She attended Alcoa City Schools and graduated valedictorian in 2009. She attended the University of Tennessee Knoxville and graduated with a B.S. in Wildlife and Fisheries Management in 2014. While at the University of Tennessee she worked for 4.5 years as an undergraduate research assistant in the Aquatic Microbial Ecology Research Group under the direction of Dr. Steven Wilhelm. She also attended an NSF-sponsored Research Experience for Undergraduates at Bigelow Laboratory for Ocean Sciences in 2011 under the direction of Dr. Willie Wilson. She attended the College of Charleston and graduated with a M.S. in Marine Biology in 2017. She returned to the University of Tennessee Knoxville to pursue her PhD in 2017 under the direction of Dr. Steven Wilhelm. Her work has primarily focused on harmful algal bloom ecology and viral ecology. She has participated in 13 research expeditions and spent approximately 140 days at sea. She received the *NOAA Seas Grant Knauss Marine Policy Fellowship* in 2020 and was placed in the Department of Energy Wind Energy Technology Office on the offshore wind portfolio. She currently lives in Knoxville, TN, where she enjoys a quiet life with her husband, Justin Burch, and their cat, Diablo.