Diversity of Viral Communities in a Northern Temperate Lake through Metagenomics

by

Cody Carrière Collis

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The samples that were used in this study were collected by Dr. Ellen Cameron in the Müller Lab at the University of Waterloo as part of her doctoral thesis and in collaboration with the *for*Water Research Network (NSERC Network for Forested Drinking Water Source Protection Technologies). DNA extraction for sample TL_JUL18 was performed by Dr. Ellen Cameron, and sequencing was completed at Metagenom Bio Inc. (Waterloo, Ontario).

For subsequent samples, DNA extraction and shotgun sequencing were performed by Dr. Pauline Wang, Dr. Yunchen Gong, and Jee Yeon Lee at the University of Toronto's Centre for the Analysis of Genome Evolution & Function.

This work was funded by an NSERC Discovery Grant (2022-03350), and a Phycological Society of America Fellowship awarded to Dr. Nissimov.

Abstract

Viruses in oceanic environments play pivotal roles in regulating nutrient transfer and increasing rates of carbon transport. They have also been implicated in the regulation of micro-organism community dynamics through infection of successful hosts, the stimulation of primary productivity, and the termination of algal blooms. Through auxiliary metabolic genes stolen from host cells, viruses are also thought to influence biogeochemical cycles through gene expression during infection. Viruses have been comparatively well studied in oceanic environments, but there is a dearth of knowledge in their roles in freshwater environments. As toxins produced by harmful algal blooms (HABs) can challenge conventional drinking water operations, it is critical to understand the role viruses may play in their proliferation and potentially their management. Examining viral communities using metagenomics will allow for the characterization of the diversity of viruses and their potential roles in freshwater environments. To accomplish this, metagenomic sequencing was carried out on water samples from Big Turkey Lake (Ontario) over a seasonal period from 2018 to 2020. Multiple viral classifiers were first combined with a viral binning approach, and subsequently generated viral metagenome-assembled genomes (vMAGs) were then evaluated for quality. All contigs identified as viral were assigned taxonomy based on amino acid alignment through BLAST, and the vMAGs were further analyzed through the IMG/VR database to determine if there was notable alignment with viruses that have known hosts. Auxiliary metabolic gene analysis was performed using the KEGG orthology database as a reference. The majority of viral contigs across all samples were unable to be assigned taxonomy, indicating that there is a large amount of unknown diversity of viruses within Big Turkey Lake. Identified contigs were either assigned as bacteriophages from the order *Caudovirales* or eukaryotic algal viruses from the family *Phycodnaviridae*, with the presence of virophages observed in low abundance. Bacteriophage abundance remained relatively stable throughout the year with a spike in winter, while eukaryotic algal viruses were most abundant in the summer. Auxiliary metabolic genes related to folate biosynthesis, carbon metabolism, nitrogen fixation, and photosynthesis were identified, and genes related to lipid metabolism were also found during winter. Findings from this study will provide a baseline foundation of viral communities in Big Turkey Lake, allowing for further research into potential hosts for these viruses, and how they could influence the formation and termination of harmful algal blooms in freshwater.

Acknowledgements

First, I would like to thank Dr. Jozef Nissimov, my supervisor. I'm sure doing an M.Sc during a global pandemic wasn't what either of us had in mind, but I appreciate all the support and understanding you gave me throughout this time. I hope I haven't been too bad of a student for you. Thank you for taking a chance on me all those years ago. I've thoroughly enjoyed what I've done here.

To my committee members, Dr. Kirsten Müller and Dr. Laura Hug, thank you for your feedback and advice, and for giving me the push to dive headfirst into bioinformatics.

To Dr. Nicolas Tromas for his advice and feedback, and for answering all the endless questions I've had about different viral analytic tools.

To Dr. Ellen Cameron and Nikhil George, both of whom helped me immensely in the first few months of starting this project.

A special thanks to Dr. Christine Dupont, for everything she has done for me, starting in my 4^{th} and 5^{th} year of my undergraduate studies. I do not think I would be the same person here today had you not set me on this path during my time in your Virology classes. It's hard to believe how an off-hand remark from my best friend to see you after class led me here.

Finally, to all of my friends and family who have supported me, to my mother who has ever been my rock, and to my father who will never get to see me present,

Thank you.

Table of Contents

Authors Declaration……ii

Statement of Contributions……iii

Abstract……...iv

Acknowledgements……..v

List of Figures……..…..…viii

List of Tables…….…..…..……...ix

Chapter 1: Introduction and Literature Review……...1

1.1 Phytoplankton and their Viruses: An Overview……1

1.2 Microalgal and Cyanobacterial Blooms……1

1.3 Anthropological and Environmental Impacts of Harmful Blooms……2

1.4 Aquatic Viruses in Oceanic and Freshwater Environments……..3

1.5 Metagenomic Analysis of Viruses……6

1.6 Auxiliary Metabolic Genes in Viruses…….....8

1.7 Sample Area: Turkey Lake Watershed……...……..9

1.8 Objectives……10

Chapter 2: Materials and Methods……12

2.1 Sample Collection……12

2.2 DNA Extraction and Assembly……...13

2.3 Bacterial Identification using Kraken2……15

2.4 Viral Identification, Binning, and Viral Metagenome-Assembled Genome Generation……..15

2.5 Taxonomic Identification of Viruses……...16

2.6 Viral Metagenome-Assembled Genome Analysis with IMG/VR……...17

2.7 Auxiliary Metabolic Gene Annotation……17

Chapter 3: Results……..17

3.1 Assembled Reads of Big Turkey Lake……18

3.2 Bacterial Community Composition……19

3.3 Viral Community Composition……...22

3.4 Constructing Bins for Viral Metagenome-Assembled Genomes……23

3.5 Viral Metagenome-Assembled Genome Analysis using IMG/VR……..26

3.6 Viral Auxiliary Metabolic Genes in Big Turkey Lake……29

Chapter 4: Discussion……30

4.1 Host Community Analysis of Big Turkey Lake……..32

4.2 Combining Viral-Identifying Tools to Analyze Metagenomes……...33

4.3 Viral Communities in Big Turkey Lake……..33

4.4 Classification of Viral Metagenome-Assembled Genomes from Big Turkey Lake…36

4.5 Viral Roles in Big Turkey Lake through Auxiliary Metabolic Gene Analysis…....37

4.6 Implications of Filtration on Results……39

4.7 State of Viral Taxonomy in the Future……..…..……...…40

Chapter 5: Conclusions and Further Research……...41

5.1 Diversity of Viruses in Freshwater Remains Largely Unknown……41

5.2 Viral Auxiliary Metabolic Genes within Big Turkey Lake Display Potential Viral Roles in Nitrogen Fixation, Carbon Fixation, and Photosynthesis……41

5.3 Future Research on Viruses in Big Turkey Lake Should be Focused on Virus-Host Relationships and Physicochemical Links to Seasonality……..…..…..…..…..…42

Bibliography…….……..……..44

Appendix……..……..57

List of Figures

Figure 1.1 Visualization of the viral shunt……..4

Figure 1.2 Visualization of the viral shuttle……5

Figure 2.1 The Turkey Lake Watershed, Ontario, Canada……...13

Figure 2.2 Overview of bioinformatics workflow used in this study……..…..……..…..14

Figure 3.1 Stacked bar chart of proportion of viral contigs from VIBRANT and VirSorter2…..19

Figure 3.2 Bacterial community composition in Big Turkey Lake as analyzed by Kraken2 and bracken……...20

Figure 3.3 Cyanobacterial community composition in Big Turkey Lake as analyzed by Kraken2 and bracken……21

Figure 3.4 Viral taxonomic families found across samples taken from Big Turkey Lake…..….22

Figure 3.5 Proportion of contigs binned by vRhyme in comparison to the entire sample……24

Figure 3.6 Quality Analysis of constructed bins using CheckV……...25

Figure 3.7 Viral Metagenome-Assembled Genomes in Big Turkey Lake as visualized using DNAPlotter……28

Figure 3.8 Individual Auxiliary Metabolic Gene composition from each sample…….…...30

List of Tables

Table 2.1 Water samples taken from Big Turkey Lake as used in this study……...12

Table 3.1 Assembled Contigs and Viral Contigs from Big Turkey Lake…….18

Table 3.2 Number of Viral Bins Constructed by vRhyme from Metagenomic Data……24

Table 3.3 BLAST analysis of Big Turkey Lake vMAGs using the IMG/VR database……26

Table 3.4 Number of Auxiliary Metabolic Genes found across samples in Big Turkey Lake….29

Table 3.5 Highlighted Auxiliary Metabolic Genes within Big Turkey Lake……31

Supplementary Table 1 Relative Abundance of Cyanobacterial Genera in Percent…….57

Supplementary Table 2 Auxiliary Metabolic Genes from Big Turkey Lake……..59

“When words become unclear, I shall focus with photographs −

When images become inadequate, I shall be content with silence.”

- Ansel Adams (1902-1984)

Chapter 1: Introduction and Literature Review

1.1 Phytoplankton and Their Viruses: An Overview

In aquatic environments, phytoplankton are ubiquitous organisms, and have a significant impact on the entire planet. They serve as primary producers at the base of the food web, providing a critical source of food for higher trophic levels (Glibert *et al.*, 2018), and represent 90% of the total biomass in marine systems (Wilhelm & Suttle, 1999). Even smaller than these micro-organisms are the viruses that infect them, who are considered major contributors to global mortality rates of phytoplankton (Middelboe, 2000; Brussaard *et al.*, 2008; Mojica *et al.*, 2016). Viruses in aquatic systems are vital for regulating the succession of phytoplankton (Müling *et al.*, 2005; Haaber & Middelboe, 2009), and influence nutrient and carbon cycling (Fuhrman, 1999; Suttle, 2007; Mojica & Brussaard, 2014). These viruses will be the focus of this research, examining their community composition over time and studying their potential effects on freshwater aquatic systems.

1.2 Microalgal and Cyanobacterial Blooms

An algal bloom or a cyanobacterial bloom occurs when conditions within the environment, such as temperature or nutrient availability, allow for a sharp increase in biomass of a given species leading to a change or distortion in the visibility or colour of a water column (Anderson *et al.*, 2012; Huisman *et al.*, 2018). Some of these colour changes are quite distinctive and vary depending on what organism is actively blooming. 'Red Tides' caused by dinoflagellates being, as expected from the name, a stark red colour (Anderson *et al.*, 2012), and 'Brown Tides', caused by *Aureococcus anophagefferens* are named for similar reasons, causing a murky brown colour in the water column (Gobler & Sunda, 2012).

Algal blooms and cyanobacterial blooms are not inherently harmful to their environments, nor are they a novel phenomenon, considering the presence of literature that references potential blooms throughout history (Huisman *et al.*, 2018). However, blooms can certainly become nuisances in areas that humans inhabit, leading to the development of the term 'Harmful Algal Bloom' (HABs). The significant issue with HABs is their increasing incidence rate globally, which has also been noted throughout

Canada (Pick, 2016; Winter *et al.,* 2011), attributed to increasing temperatures due to climate change (Paerl & Huisman, 2008) and anthropogenic nutrient run-off of nitrogen and phosphorus (Conley *et al.,* 2019).

1.3 Anthropological and Environmental Impacts of Harmful Blooms

The increasing incidence of HABs due to climate change means that their effects are felt more frequently by human populations. The primary issue, and the most well-known issue, that HABs generally bring about is the production of toxins. Some bloom-forming cyanobacteria, dinoflagellates, and others are capable of producing compounds that can be toxic and/or lethal to humans and other animals, like the toxins microcystin, cylindrospermopsin, and nodularin (Glibert *et al.,* 2018). For example, blooms of the cyanobacterium *Microcystis aeruginosa* occurring in Lake Taihu, China (Guo, 2007) and Lake Erie (Steffen *et al.,* 2017; McKindle *et al.,* 2020) have had devastating consequences on the drinking water supply of cities near them, due to the production of microcystin. In Canada, microcystin is currently the only toxin that has a Maximum Acceptable Concentration (MAC) in drinking water with an acceptable level being below 1.5 µg/L of water (Health Canada, 2021). Other toxins, specifically anatoxin-a and cylindrospermopsin, do not currently have MAC guidelines in Canada due to limited research on their effects and normal levels within the environment (Health Canada, 2021). Outside of toxin production by cyanobacteria, drinking water can also be impacted through the cyanobacterial production of taste and odour compounds such as geosmin, making drinking water unpleasant to consume (Li *et al.,* 2016; Österholm *et al.,* 2020).

Not all issues with HABs are related to human health; some blooms can also have a large impact on the economy. For example, blooms in Lake Erie have caused monetary losses due to decreases in property value, tourism, and recreation, with projected costs of uncontrolled blooms being upwards of $270 million CAD per year over the next 30 years (Smith *et al.,* 2019). In addition, when situations arise where toxins or taste and odour compounds are present, they must somehow be removed from drinking water for the community. Conventional practices in treatment plants are not always effective at removing taste and odour compounds like geosmin or 2-methylisoborneol (Ridal *et al.,* 2001; Srinivasan & Sorial, 2011). Effective treatment of drinking water containing these

compounds can cost communities millions of dollars due to treatment options being more expensive, and in lost revenue. (Dunlap *et al.*, 2015).

The impacts of HABs are not limited to humans; they can also be troublesome for organisms that live in a location that is prone to blooming. As stated before, what makes a bloom a bloom is the disruption of the water column's clarity through sheer biomass (Huisman *et al.*, 2018); this increased cell density can prevent sunlight from reaching aquatic macrophytes that need it (Paerl & Otten, 2013). In turn, this impacts organisms that rely on macrophytes, either as a source of food or as a source of shelter (Huisman *et al.*, 2018; Paerl & Otten, 2013). In addition, HABs can cause local anoxia within the water column, choking out and causing the death of other organisms in the environment, as seen with the mass fish kills in Lake Erie during the 1960's and 1970s (Allinger & Reavie, 2013; Watson *et al.*, 2016).

1.4 Aquatic Viruses in Oceanic and Freshwater Environments

Viruses in marine environments are much better characterized compared to viruses in freshwater, and their environmental roles are better understood. Currently, it is unclear if we can extrapolate how viruses in freshwater might function within an ecosystem or what role they may play, based on how viruses impact marine environments. In oceanic environments, viruses have been implicated in the cycling of organic material through a process known as the 'viral shunt' (Wilhelm & Suttle, 1999). In this process (Figure 1.1), viruses transform the cellular material of their host cells through viral lysis into dissolved organic matter (DOM) or particulate organic matter (POM) (Suttle, 2007; Zimmerman *et al.*, 2020). This transformation locks the cellular material into a form that higher trophic levels cannot utilize, and instead allows it to be utilized once again by the lower trophic levels or primary producers. This 'shunting' of the organic material away from higher trophic levels allows the ecosystem to have increased bacterial and phytoplankton respiration rates, resulting in a more productive environment at its base (Fuhrman, 1999; Suttle, 2007).

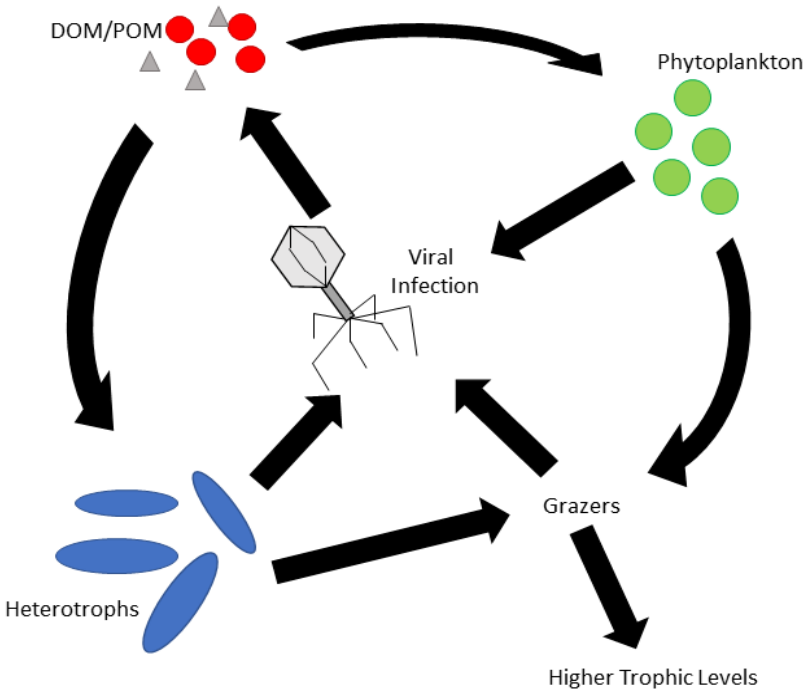


Figure 1.1 Visualization of the 'viral shunt'. Infection of heterotrophs, phytoplankton, and grazers by viruses release DOM/POM, preventing it from being utilized by higher trophic level organisms. Figure adapted from Suttle (2007).

Viruses have also been implicated in the transport of particulate organic carbon (POC) into the deep ocean through the 'viral shuttle' (Weinbauer, 2004; Sullivan *et al.*, 2017). Phytoplankton, when virally infected, are shown to have increased rates of sinking (Suttle, 2007), and have also been observed to increase the rate of Transparent Exopolymer Particles (TEPs) production during infection (Vardi *et al.*, 2012). These TEPs are 'sticky', in that they promote the clumping of organic material into marine snow, which subsequently sinks into the deep ocean (Mojica & Brussaard, 2014; Nissimov *et al.*, 2018; Zimmerman *et al.*, 2020). The 'shuttle' (Figure 1.2) is how viruses can tap into the biological carbon pump, which is responsible for sequestering carbon and preventing it from being used (Sanders *et al.*, 2014). Differences in depths between freshwater and oceanic environments means that it is unlikely the 'viral shuttle' would work exactly the same within the two environments; freshwater lakes are much shallower than the ocean and will have instances in the year where water layers will naturally intermix, usually due to factors such as depth, weather, and temperature (Yang *et al.*,

2018). This intermixing would cause the nutrients collected at the bottom of the lake to be brought back up to the surface, likely removing the possibility of sequestration. Nonetheless, particularly deep lakes such as Lake Baikal in Russia or Great Slave Lake in Canada are deep enough that the 'viral shuttle' could be a factor in carbon cycling and carbon sequestration in those specific environments. Meromictic lakes, which do not have period of the year where their layers intermix (Boehrer & Schultze, 2008), could also support carbon sequestration through the 'viral shuttle'.

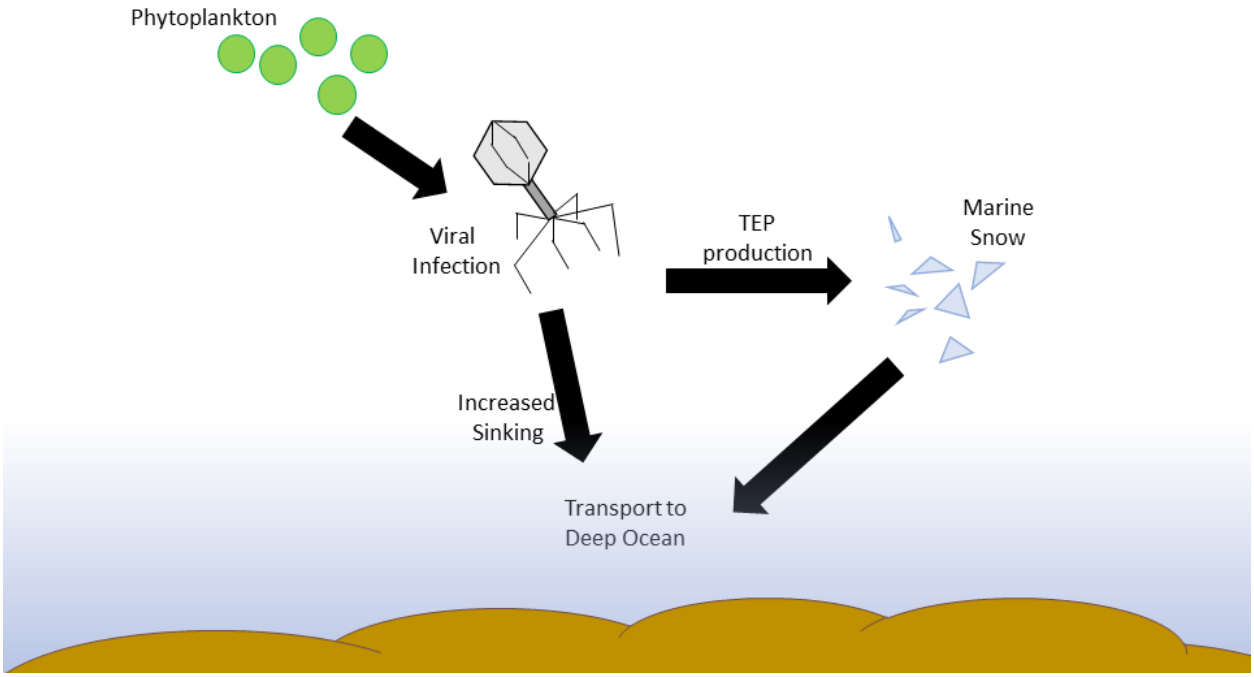


Figure 1.2 Visualization of the 'viral shuttle'. Viral infection can cause increased rates of sinking in virally infected cells, and an increase in the production of transparent exopolymer particles (TEP), both of which contribute to increased carbon transport and sequestration into the deep ocean.

Viruses are thought to be one of the main contributing factors towards the termination of blooms, both cyanobacterial and algal (Suttle, 2007). This is commonly seen in oceanic environments, where the level of viruses within a community increase as the bloom tapers off (Bratbak *et al.*, 1993). In addition, recent research into freshwater cyanobacterial blooms suggest that viruses are equally important during the duration of

the bloom, with multiple viral lysis events likely taking place over the course of a single bloom (Steffen *et al.,* 2015; McKindles *et al.,* 2020).

Mass lysis of blooms via viral predation does have a significant potential downside. Recent studies have found that viral lysis of toxin-producing cyanobacteria causes an increase in the amount of toxins dissolved in the water column. For example, a viral lysis event during a 2019 *M. aeruginosa* bloom in Lake Erie coincided with elevated levels of microcystin within the lake (McKindles *et al.,* 2020). Similarly, *in vitro* studies of the cyanobacterium *Nodularia* found increased levels of the toxin nodularin in samples that were lysed with viruses (Šulčius *et al.,* 2018).

1.5 Metagenomic Analysis of Viruses

Viruses differ from other microbes in that they do not have a comprehensive marker gene like the 16S rRNA gene in prokaryotes or the 18S rRNA gene in eukaryotes (Rohwer & Edwards, 2002). This lack of universal marker gene makes the discovery and classification of novel viruses comparably difficult. That is not to say there are no marker genes for specific groups of viruses; the HK97 protein fold is unique to members of *Herpesviridae* and double-stranded DNA bacteriophages, indicating that they have shared a lineage in the past (Baker *et al.,* 2005; Duda & Teschke, 2019). In theory, this fold could be used similarly to the 16S or 18S sequences and sequenced using amplicon sequencing, but this runs the risk of losing the vast majority of viral sequences that do not contain this gene. Shotgun sequencing, the primary way of sequencing for the analysis of viruses, mitigates losing diversity of viral communities by sequencing everything that is in a sample, rather than just the microbes that contain a specific gene (Mohiuddin & Schellhorn, 2015).

There are two major types of viruses that are found within freshwater environments, bacteriophages and algal viruses. Bacteriophages are viruses that infect bacteria, with three major families in aquatic habitats, split by morphology: *Myoviridae*, which have contractile tails, *Siphoviridae*, which have long non-contractile tails, and *Podoviridae*, which have short non-contractile tails (Maniloff & Ackermann, 1998). Other than morphology, these families do not inform much about the virus, such as their life cycles, or what host organism the virus infects; if you were to look at one of these

viral families holistically, you would find a variety of life cycles and incredibly broad host ranges (Xia *et al.*, 2013). Recent arguments have been made in favour of scrapping these three morphological families and changing to a system of taxonomy that more focused on genetics (Turner *et al.*, 2021), which will allow for better divisions within viral taxonomy that better informs on their potential function within their environment.

The other prevalent type of viruses found in freshwater lakes are Nucleocytoplasmic Large DNA Viruses (NCLDV) which infect eukaryotic algae. These viruses are among the largest viruses that are known to date, and include viruses such as *Pandoraviridae*, *Pithoviridae*, and *Poxviridae* (Koonin & Yutin, 2019). Genomes of NCLDV have been found to be over 1Mbp long (Koonin & Yutin, 2010). In freshwater, members of *Phycodnaviridae* and *Mimiviridae* are common (Wilson *et al.*, 2010; Palermo *et al.*, 2019). Like the bacteriophages, the taxonomy of NCLDV will probably change once more are discovered through metagenomic means, and a better understanding of their genomes has been elucidated (Wilson *et al.*, 2010).

A unique family of viruses that are not as common and poorly in databases are virophages from *Lavidaviridae*; viruses that 'infect' other viruses. There currently aren't many known virophages, with only 33 high quality sequences deposited in databases (Paez-Espino *et al.*, 2019; Fischer 2021). However, this number will likely increase as more metagenomic datasets are deposited into existing databases (Paez-Espino *et al.*, 2019). The virophages that are currently known and characterized target members of the NCLDV (Sobhy, 2018). In truth, 'infection' doesn't fully explain their relationship to the viruses that they parasitize – it appears that virophages use the replication machinery of both NCLDV and their hosts in order to replicate themselves, which has the added effect of inhibiting replication of its host virus (Fischer, 2021). Virophages differ from satellite viruses due to having much larger, double stranded DNA genomes, and genetic similarity to other DNA viruses (Fischer, 2021).

Other viruses exist in freshwater environments, but one challenge we face in trying to uncover them is the difficulty of finding them in shotgun sequencing data. The past decade has seen a multitude of various viral identifiers produced with different techniques. Many of these identifiers use database-driven methods in order to classify,

either through the use of alignment algorithms such as BLAST or by comparing genes within a query sequence to known viral genes (Andrade-Martinez *et al.*, 2022; Miao *et al.*, 2022). Using viral identifiers that rely on pre-built databases has the drawback of causing a positive feedback loop; viruses that are present within a database will cause similar viruses within an environmental sample to be observed. Conversely, viruses that are not found within a database are labeled as unidentified, uncharacterized, or potentially even not viral in origin, as they do not have similar sequences to which they can be matched (Miao *et al.*, 2022; Kieft & Anantharaman, 2022). Other identifiers have been created using methods that can limit the amount of bias coming from databases using techniques such as machine learning or k-mer analysis (Andrade-Martinez *et al.*, 2022; Miao *et al.*, 2022), although those can come at the cost of needing more computational power (Miao *et al.*, 2022). Programs such as VIBRANT (Kieft *et al.*, 2020) and VirSorter2 (Guo *et al.*, 2021) work by combining both database-driven and machine learning approaches, which can be useful in mitigating the blind spots of each technique.

1.6 Auxiliary Metabolic Genes in Viruses

The primary goal of viral infection is for the virus to force the host to produce its viral progeny, thus continuing the cycle of infection (Zimmerman *et al.*, 2020). That is not to say that all its genes are solely for the purpose of overtaking its host cell; many viruses have been found to contain genes that, ostensibly, do not have anything to do with viral replication. These are called ‘Auxiliary Metabolic Genes’ (AMGs) and are genes that viruses ‘steal’ from their hosts and potentially other more distantly related bacteria (Crummett *et al.*, 2016). AMGs as a class of genes are quite diverse, but what is most interesting about them is that many are conserved across different environments, such as genes involved in photosynthesis, carbon metabolism, nucleotide biosynthesis, and cofactor biosynthesis (Breitbart *et al.*, 2007; Hurwitz & U’ren, 2016; Crummett *et al.*, 2016). The inclusion of these genes across multiple viral genomes in different environments suggests that these genes were not simply stolen by chance to be discarded later down the line, but rather are key parts of the viral genome that universally increase viral fitness (Breitbart *et al.*, 2007).

Large scale metagenomics work in marine enviroments has greatly increased our awareness for AMGs (Williamson *et al.,* 2008; Hurwitz & Sullivan, 2013). These studies have uncovered AMGs from marine viruses that infect *Prochlorococcus* and *Synechococcus*. These viruses are most well known for their AMGs related to photosynthesis, specifically the photosystem II core reaction centre D1 protein (*psbA*) (Hurwitz & U'ren, 2016; Crummett *et al.,* 2016; Lindell *et al.,* 2005), although this is not the only example. Photosynthesis AMGs are thought to be particularly important for viruses to encode as viral replication will repress the host cell's protein synthesis (Fabricant & Kennel, 1970; Lindell *et al.,* 2005; Hurwitz & U'ren, 2016). As phage D1 proteins are increasingly expressed throughout phage infection (Lindell *et al.,* 2005), photosynthesis is supplemented and allows the phage to maintain enough energy for successful replication (Crummett *et al.,* 2016; Lindell *et al.,* 2005).

To say that AMGs are varied is an understatement and does not do justice to the impacts that they have on global biogeochemical cycling. To list some examples, viruses have been found to encode AMGs that are related to nitrogen cycling (Wang *et al.,* 2022), sulfur and thiosulfate oxidation (Anantharaman *et al.,* 2014; Kieft *et al.,* 2021), carbon metabolism through the TCA cycle and fixation through photosynthesis, methanol oxidation (Coutinho *et al.,* 2020), and many more. As the discovery of more AMGs continues, it becomes clear that only the tip of the iceberg has been scratched with regards to the role of viruses within the environment and how they impact global nutrient cycling.

1.7 Sample Area: Turkey Lake Watershed

The sample area for this study was Big Turkey Lake, a northern temperate lake located approximately 50 km north of the city of Sault Ste. Marie, Ontario (Jeffries *et al.,* 1988), currently being used as a location by the *forWater* network for research into cyanobacteria and water quality. This makes the location ideal for studying viruses in freshwater, due to the abundance of information that is currently available about this site. This lake is a part of a larger area known as the Turkey Lake Watershed and has been used throughout the past few decades as a location to study anthropogenic effects on the ecosystems of the Canadian shield (Jeffries & Foster, 2001). Initially chosen in 1980 as a

location to study the effects of acidic deposition from acid rain, numerous governmental agencies have contributed to the collaborative research efforts in this location, including Environment Canada, Natural Resources Canada, Fisheries and Oceans Canada, in addition to many universities (Jeffries & Foster, 2001). The entire Turkey Lakes Watershed composes an area of 10.5 km^2 and contains four distinct interconnected lakes with five separate basins: Batchwana Lake (two basins), Wishart Lake, Little Turkey Lake, and Big Turkey Lake (one basin) (Jeffries *et al.*, 1988). Of the five basins, four are considered dimictic, meaning they experience two periods of the year where they overturn (Jeffries *et al.*, 1988). Wishart Lake is the exception as the lake is too shallow to experience stratification (Jeffries *et al.*, 1988).

The cyanobacteria in the Turkey Lake Watershed have already been characterized in previous research studies (Jeffries *et al.*, 1988; Cameron, 2021). It has also been found that, consistent with trends of increasing toxic cyanobacterial blooms found globally and across Ontario (Winter *et al.*, 2011), the amount of cyanobacteria and their growing period found within the Turkey Lake Watershed has also been increasing (Cameron, 2021). In addition, water temperatures in the Turkey Lake Watershed have been increasing as well (Creed *et al.*, 2015), although it has also been noted that this likely is not the sole factor driving the increase in cyanobacterial abundance in this location (Cameron, 2021).

There is currently a lack of information about aquatic viruses that are observed in the Turkey Lake Watershed. As explained earlier, viruses can have a significant effect on the other organisms that live in their environment, thus studying the viruses in the Turkey Lake Watershed will allow for a more holistic image of the entire system of microbes and their ecological roles.

1.8 Objectives

Due to the limited knowledge on freshwater viruses and their potential impacts on bloom-forming cyanobacteria and their environments, we explored the viral communities in Big Turkey Lake, Ontario, using metagenomics. Specifically, the objectives were to study the overall composition of the viral communities and how they change over time, along with identifying potential AMGs encoded by the viral community and what they

may indicate about the role of viruses within the lake. To my knowledge, an exploration of the viral diversity within this lake has never been performed, meaning that the information gleaned through this study is hereby novel. As the Turkey Lakes Watershed has been the site of research on cyanobacterial communities in the past and present, this study will allow for a greater understanding of how viruses fit into the larger environment of freshwater lakes and will act as a springboard for further research into how viruses impact cyanobacterial populations.

Chapter 2: Materials and Methods

2.1 Sample Collection

The water samples that were used for this metagenomic study were collected by Dr. Ellen Cameron as part of her Doctoral Thesis in the Müller Lab (Cameron, 2021), studying cyanobacteria in the Turkey Lake Watershed (Figure 2.1). Of the samples that Dr. Cameron took, six were selected for further analysis as part of this study, taken from Big Turkey Lake of the Turkey Lake Watershed (47 02' 54.7"N, 84 25' 19.3"W), from different time points between July 2018 and January 2020 (Table 2.1). The samples were taken based on the secchi depth at the time of collection, which was determined using a secchi disk. The exception to this is the sample taken during January 2020, due to the lake experiencing ice cover – thus, only surface water was collected and filtered.

For each sample (Table 2.1), lake water was obtained using a Masterflex E/S portable peristaltic pump and was subsequently vacuum filtered on 47 mm 1.2 µm pore size Whatman GF/C filters (Whatman plc, Buckinghamshire, United Kingdom). After filtration, the GF/C filters were frozen and kept at a temperature of -20°C until further analysis by Dr. Cameron and in this study.

Table 2.1. Water samples taken from Big Turkey Lake as used in this study.

Sample	Collection Date	Depth (cat*)	Depth (m)	Volume Filtered (mL)
TL_JUL18	July 18th 2018	Secchi +1m	8	800
TL_MAY19	May 23rd 2019	Secchi	5.25	1000
TL_JUN19	June 28th 2019	Secchi	5.75	1000
TL_JUL19	July 24th 2019	Secchi	5.25	1000
TL_AUG19	August 21st 2019	Secchi	5	1000
TL_JAN20	January 23rd 2020	Surface	N/A	1000

*(cat) refers to the depth that the sample was taken in relation to the secchi depth at the time of sampling (a measure of the turbidity of the water column)

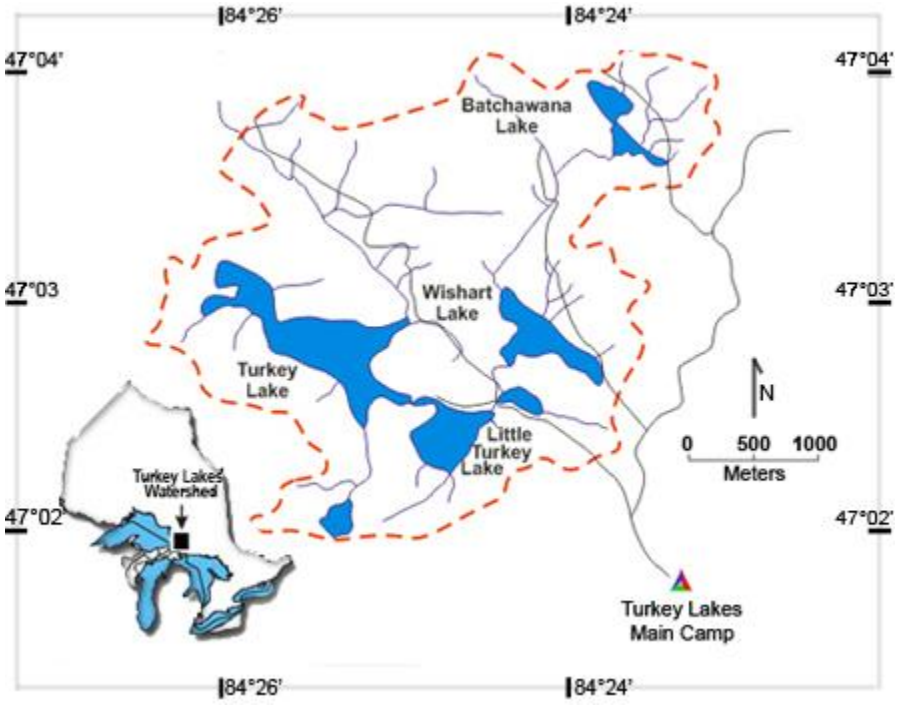


Figure 2.1 The Turkey Lake Watershed, Ontario, Canada**.** Samples were taken from Big Turkey Lake, which in this image is labeled simply as 'Turkey Lake'. Image used from the Government of Canada (Environment Canada, 2014).

2.2 DNA Extraction and Assembly

Sample TL_JUL18 was sequenced separately before the start of this project, while samples TL_MAY19 through TL_JAN20 were sequenced as part of it (Figure 2.2, Step 1). For sample TL_JUL18, preparation of the samples was completed by Dr. Ellen Cameron and sequencing was completed by the commercial laboratory Metagenom Bio Inc (Waterloo, Ontario, Canada). For samples TL_MAY19-TL_JAN20, DNA extraction and sequencing was done at the University of Toronto's Centre for the Analysis of Genome Evolution and Function (CAGEF) (Toronto, Ontario, Canada). Extraction of DNA from all samples were completed using the Qiagen DNeasy PowerSoil kit (QIAGEN Inc., Venlo, Netherlands) according to manufacturer protocol. For sequencing, libraries were prepared using the Illumina DNA prep kit, following the manufacturers protocol, while sequencing itself was performed using an Illumina MiSeq, 250x2 paired-ends sequencing (Illumina Inc., San Diego, California, USA).

The sequenced reads for sample TL_JUL18 were processed prior to the start of this project by Dr. Cameron, while the remaining samples were processed as part of this project. This was done using the Metagenome-Atlas pipeline (ver. 2.9.0) (Kieser *et al*., 2019), a pipeline which provides an all-in-one analysis including quality control, assembly, binning, and annotation of genes. For the purposes of this study, due to different tools being needed for analysis of viral sequences, only quality control and assembly was used in the Metagenome-Atlas pipeline. As part of the pipeline, quality control of the sequenced samples was performed using programs from the BBtools suite, a collection of bioinformatic programs designed for DNA and RNA analysis (Bushnell, 2019), while assembly was performed using the MetaSPAdes assembler. These assemblies were performed using servers run by the Doxey Lab at the University of Waterloo.

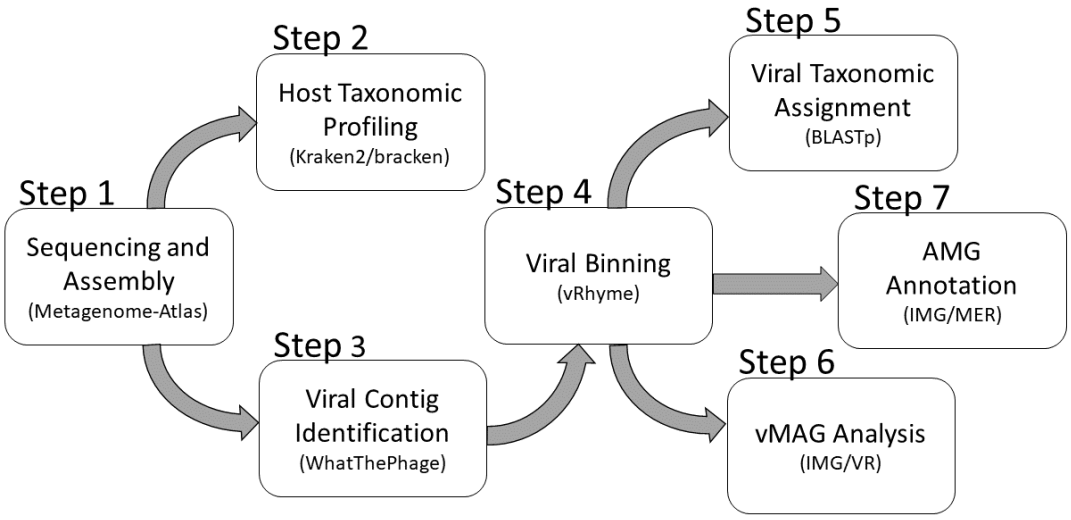


Figure 2.2 Overview of bioinformatics workflow used in this study.

2.3 Bacterial Identification using Kraken2

Once assembled using the Metagenome-Atlas pipeline, the contigs were sent through the Kraken2 analysis tool (Wood *et al.,* 2019) in order to determine the general makeup of the bacterial community at the time of sampling (Figure 2.2, Step 2). Kraken2 was used on the Graham cluster of the Digital Research Alliance of Canada with the standard database taxonomy, consisting of bacteria, archaea, viral, and human sequences.

Bracken (Lu *et al.,* 2017) was used on the analysis results from Kraken2 to more accurately measure abundance of the organisms found in the samples. Bracken was used at the phylum level and the genus level to measure abundance of bacterial phyla and genera within the phylum cyanobacteria. Results from the Bracken tool was analyzed using the Pavian web application through R (Breitweiser & Salzberg, 2020).

2.4 Viral Identification, Binning, and Viral Metagenome-Assembled Genome Generation

To identify contigs which were viral in origin, a combination of VirSorter2 and VIBRANT (Guo *et al.,* 2021; Kieft *et al.,* 2020) was used in order to widen the search while remaining relatively conservative (Figure 2.2, Step 3). To do this, the assembled contigs from each sample were put through the pipeline WhatThePhage (Marquet *et al.,* 2022) on servers from the Doxey Lab, after which the VIBRANT and VirSorter2 output files were concatenated (merged), and all duplicate contigs from the resulting file being removed.

VIBRANT and VirSorter2 are excellent tools for the purposes of this study's viral identification, as they are not specialized in only bacteriophages; both VIBRANT and VirSorter2 were shown to have higher accuracy in also identifying NCLDV while VirSorter2 also has high accuracy in detecting virophages and RNA viruses (Guo *et al.,* 2021). Combining the output of different viral prediction tools has also been proposed as a way of mitigating the blind spots of a specific tool (Kieft & Anantharaman, 2022).

Once the identification was complete, the virus-specific binning tool vRhyme (Kieft *et al.,* 2021) was used to create viral bins and potential viral metagenome-assembled genomes (vMAGs) (Figure 2.2, Step 4). vRhyme was installed on the Graham

cluster of Compute Canada (now the Digital Research Alliance of Canada). The completed bins were then linked together to form potential vMAGs using auxiliary vRhyme scripts that were included as part of the installation of the program.

In order to analyze the linked bins for potential vMAG candidacy, the program CheckV (Nayfach *et al.,* 2021) was used to test for the bin quality, again using the Graham cluster on Compute Canada. Bins that had a completeness score of >90% were considered high-quality by CheckV and thus candidates for vMAG status (Nayfach *et al,*. 2021).

2.5 Taxonomic Identification

To assign taxonomy to our identified viral contigs and bins, the resulting linked bins and unbinned contigs from vRhyme were analyzed using BLASTp (Altschul *et al.,*1990) (Figure 2.2, Step 5). First, amino acid sequences and open reading frames predicted for the linked bins were determined using Prodigal (Hyatt *et al.,* 2010), a software that can predict protein-coding genes from genomic data. The resulting sequences were then compared to the nr database through BLASTp, specifically being compared to only viral taxa. The nr database was chosen as the reference database due to the inclusion of environmental sequences that were not included in the refseq_protein database.

For a viral contig to be assigned to a specific family, methodology from previous literature sources were used (Jian *et al.,* 2021). Briefly, each Open Reading Frame (ORF) had to have a minimum bit score of 50 and an e value of $<1e^{-5}$ for it to be considered an ORF from a specific taxonomic family. In total, 50% or more of the ORFs of a specific contig had to be assigned to the same family for it to be considered a candidate for that taxonomic family. For viral bins, the same methodology was used for assigning each ORF a family. However, instead of >50% of total ORFs for candidacy into a taxonomic family, a cutoff of >33% was used instead. As binning classifies contigs that it deems as likely taxonomically similar, it was decided that taxonomic assignment of bins did not need to be as strict as the 50% cut-off that was used when examining singular contigs.

2.6 Viral Metagenome-Assembled Genome Analysis using IMG/VR

Specific vMAGs were analyzed using BLAST analysis of the IMG/VR database (Roux *et al.,* 2021) located on the Joint Genome Institute (JGI) website (Figure 2.2, Step 6). The IMG/VR database was used for this purpose as it has information about a close hit's potential host species, allowing for the comparison of hosts between the query sequences and the hit sequences.

Each vMAG had its genome annotated using PROKKA, a software tool used to quickly annotate prokaryotic genomes (Seemann, 2014). PROKKA was used with the -k viruses flag, which indicates to the program to only use viral databases to annotate genes. Visualization of the resulting vMAGs and their annotated genomes was done using DNAPlotter (Carver *et al.,* 2014).

2.7 Auxiliary Metabolic Genes (AMG) annotation

To identify which viral genes could potentially encode AMGs, the viral contigs were analyzed using JGI's Integrated Microbial Genomes with Microbiome Samples – Expert Review (IMG/MER) resources (Chen *et al.,* 2021) (Figure 2.2, Step 7). Sequences were first uploaded to the JGI Genomes Online Database (GOLD) (Mukherjee *et al.,* 2022) (Study ID: Gs0159196), where they were additionally annotated through IMG/MER. The KEGG Orthology (KO) Database (Kanehisa *et al.,* 2016) was chosen for functional annotation of the proteins found within the viral contigs, as KEGG is used by VIBRANT, one of the viral identifiers used in this study, for the functional annotation of AMGs. Once the annotation was completed, the resulting annotations were manually curated to remove genes that were categorized as involved with DNA replication and reproduction, while highlighting the genes that are involved in pathways related to different metabolic processes. These include, but are not limited to, nutrient acquisition, photosynthesis, and amino acid synthesis.

Chapter 3: Results

3.1 Assembled Reads of Big Turkey Lake

The amount of reads and the DNA concentration from each sample taken from Big Turkey Lake were varied between samples, indicating that the level of biomass, both viral and non-viral, found within the lake during the period of sampling was variable (Table 3.1). The largest difference between the number of assembled contigs between samples comes from the TL_JUL18 sample, which was sequenced separately from the other samples. TL_JUL18 had an assembled contig count of 82 018, which was 3.5× the number of contigs of the sample with the next highest sample (TL_JUN19). The proportion of viral reads in a sample also varied between samples, with a range of 0.14% viral contigs (TL_JUL19) to 0.72% viral contigs (TL_MAY19).

The assembled contigs that were identified as viral by VIBRANT and VirSorter2 were 'adjusted' to remove the duplicate contigs that both programs marked as viral. Apart from the sample from TL_JUL19, all samples had some degree of overlap of identified viral contigs between the two programs. In addition, across samples, the amount of overlap between the two programs was relatively similar (Figure 3.1), with 14% overlap (TL_JUN19) to 28% overlap (TL_JAN20). Neither VIBRANT nor VirSorter2 were more adept at identifying viral contigs across samples in this study – i.e. both programs had samples where one identified more contigs than the other.

Table 3.1 Assembled Contigs and Viral Contigs from Big Turkey Lake.

Sample	DNA concentration	Assembled Contigs	Adjusted Viral Contigs*	% Viral Contigs
TL_JUL18	11.7 ng/µl	82018	581	0.708
TL_MAY19	1.6 ng/µl	15221	110	0.723
TL_JUN19	5.1 ng/µl	23257	135	0.580
TL_JUL19	3.7 ng/µl	5547	8	0.144
TL_AUG19	4.9 ng/µl	19214	42	0.219
TL_JAN20	3.5 ng/µl	20765	130	0.626

*Adjusted contigs indicates that duplicate contigs identified by both software have been removed.

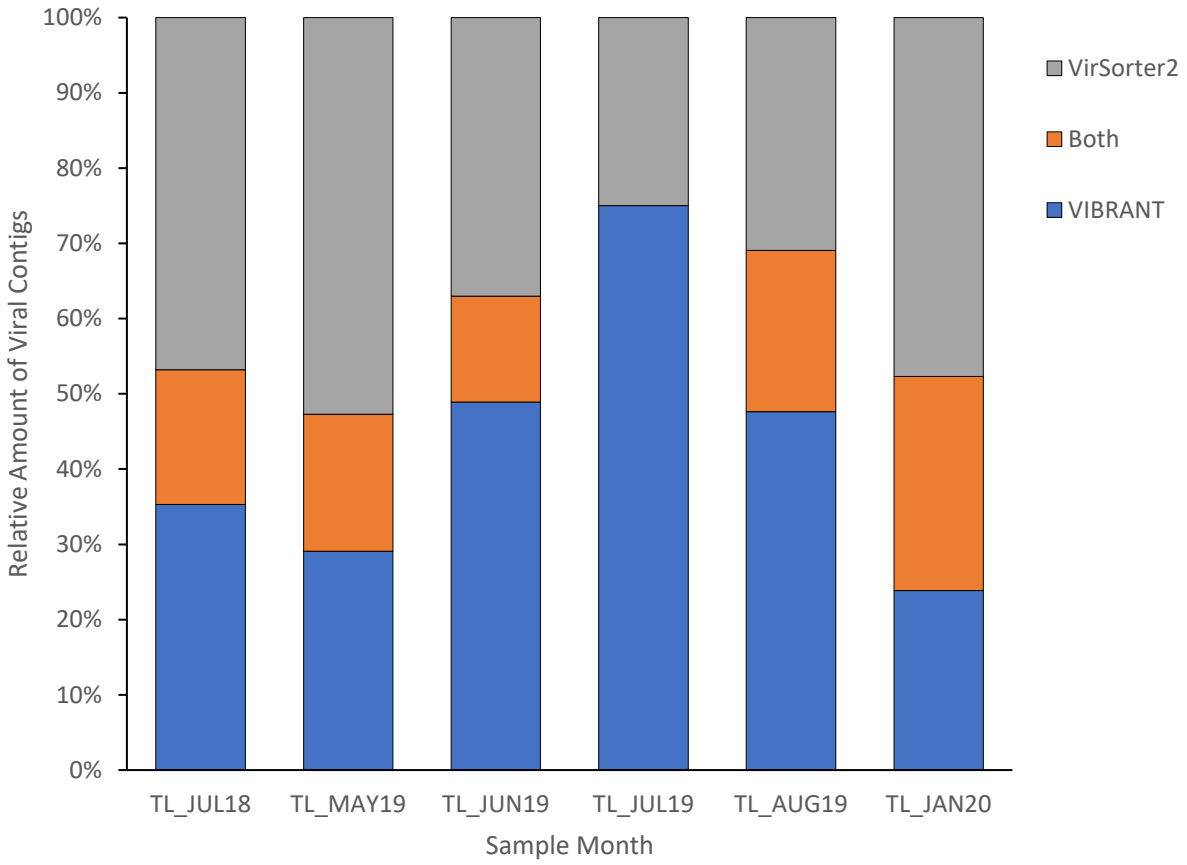


Figure 3.1 Relative proportion of viral contigs from VIBRANT and VirSorter2. Contigs that were identified as viral by both software programs were grouped together under 'Both'.

3.2 Bacterial Community Composition

Kraken2 analysis of bacterial (Figure 3.2) communities found that the phylum *Proteobacteria* (Figure 3.2, blue) was the most abundant across all six samples, with the phylum *Actinobacteria* close behind (Figure 3.2, red). *Cyanobacteria* (Figure 3.2, green) were most numerous in the summer months (TL_JUL18, TL_JUN19, TL_AUG19), while less abundant in the spring and winter months of May and January.

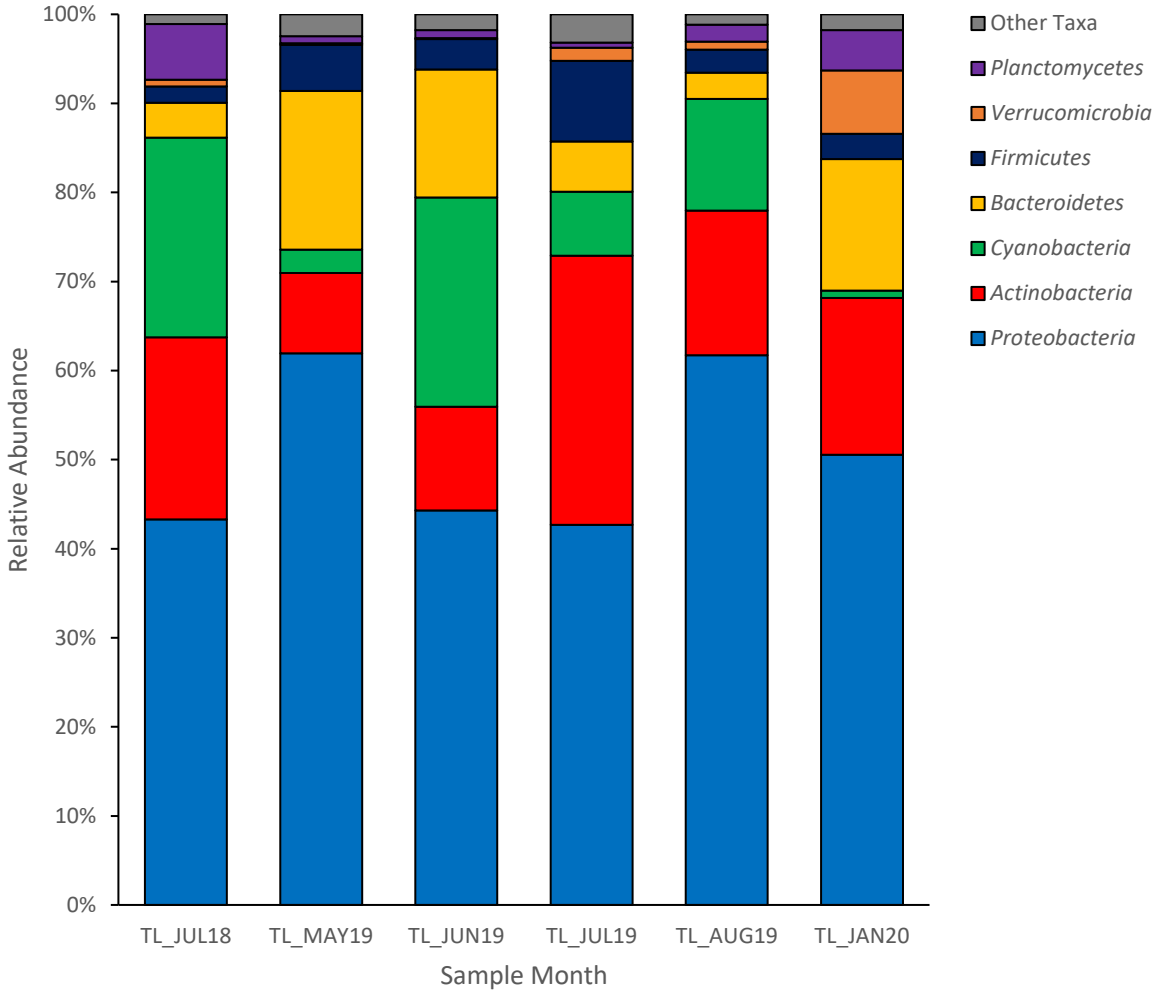


Figure 3.2 Bacterial community composition in Big Turkey Lake as analyzed by Kraken2 and bracken. Taxa that had <1% relative abundance were listen under 'Other Taxa' (gray).

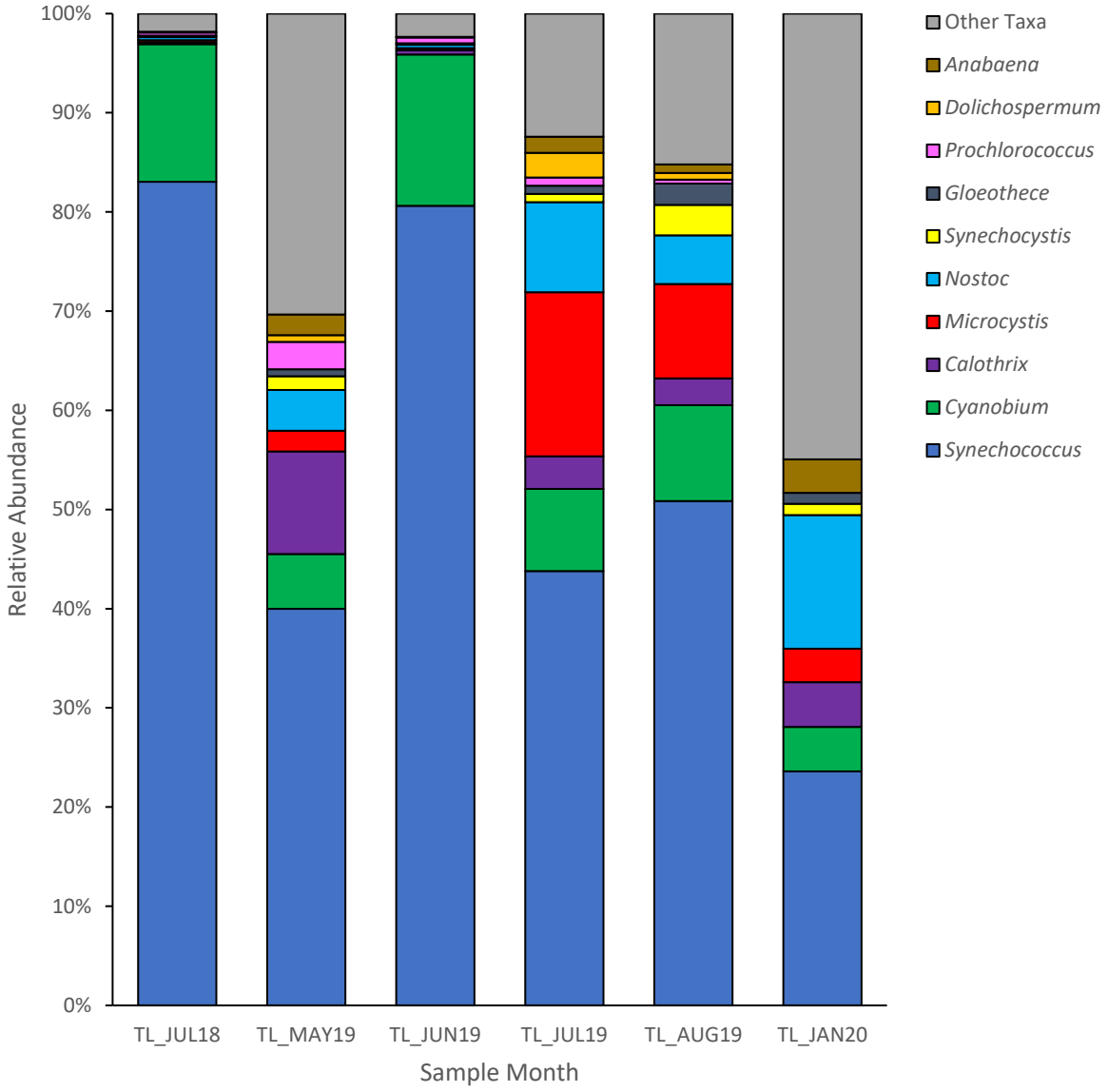


Figure 3.3 Cyanobacterial community composition in Big Turkey Lake as analyzed by Kraken2 and bracken. Genera that were <2% relative abundance were listed under 'Other Taxa' (gray).

Cyanobacteria in Big Turkey Lake (Figure 3.3) was primarily dominated by picocyanobacteria, namely those of the genera *Synechococcus* (Figure 3.3, blue) and *Cyanobium* (Figure 3.3, green). Members of the genera *Nostoc* (Figure 3.3, cyan) and *Microcystis* (Figure 3.3, red), were present within multiple samples, although not as dominant as the picocyanobacteria. Cyanobacteria that were found in <2% abundance include, but are not limited to, members of the genera *Cylindrospermopsis*, *Oscillatoria*,

and *Nodularia*. A full list of the cyanobacterial genera found within Big Turkey Lake can be found within the appendix.

3.3 Viral Community Composition

The viral community composition of Big Turkey Lake was primarily composed of reads that were unable to be assigned a taxonomic family through BLASTp analysis, with all six samples showing unassigned reads with a relative abundance of >50% (Figure 3.4, light gray).

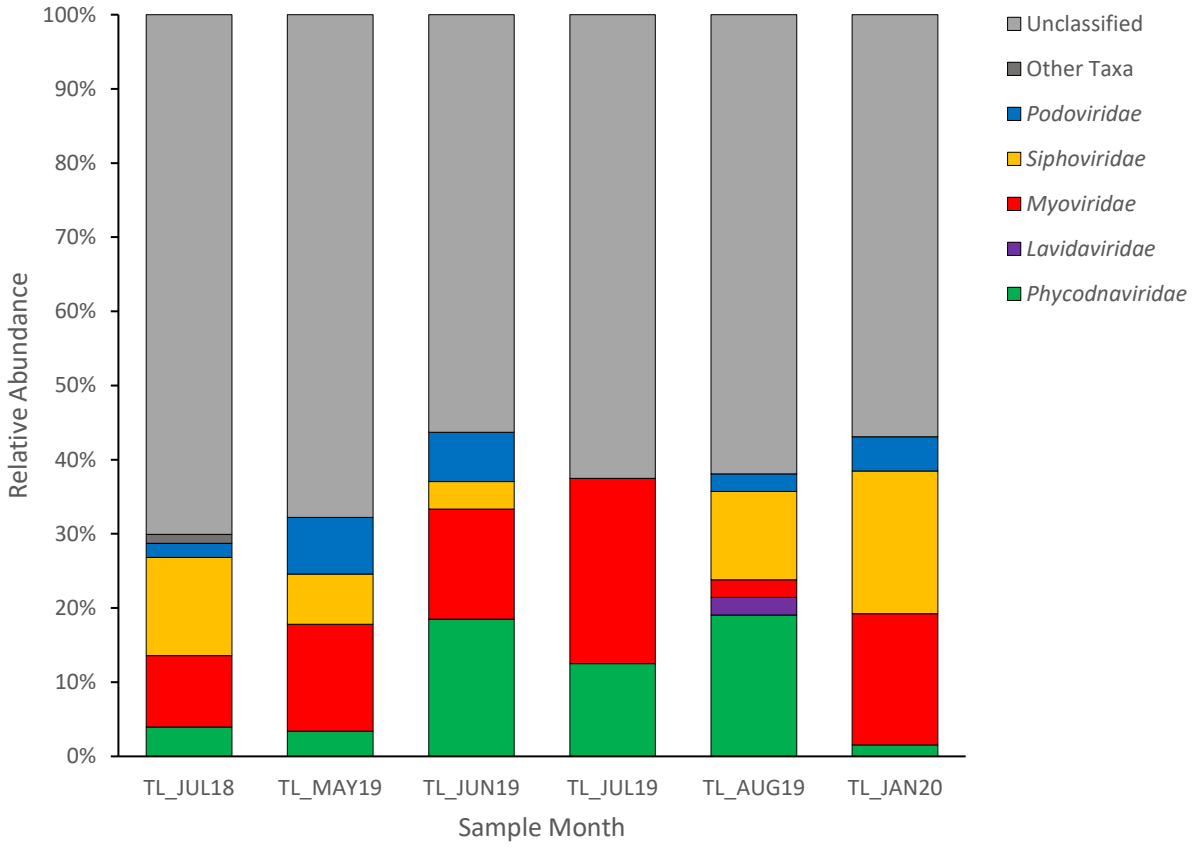


Figure 3.4 Viral taxonomic families found across samples taken from Big Turkey Lake. Families that are <1% relative abundance were listed under 'Other Taxa' (dark gray).

Phycodnaviridae (Figure 3.4, green) were found across all six samples, with a distinct pattern of seasonality across samples. Spring 2019 and winter 2020, represented by the TL_MAY19 (3.4%) and TL_JAN20 (1.5%) samples, had markedly lower relative

abundance of *Phycodnaviridae* reads than those from the summer samples of TL_JUN19 (18.5%), TL_JUL19 (12.5%), and TL_AUG19 (19.1%).

Bacteriophages comprised the majority of assigned reads across all six samples. *Myoviridae* is the first major family of bacteriophage that was found in Big Turkey Lake, with sequences found in all six samples (Figure 3.4, red). The highest relative abundance of *Myoviridae* was found in the sample from TL_JUL19 (25.0%), while the lowest was in TL_AUG19 (2.4%). *Podoviridae* (Figure 3.4, blue) and *Siphoviridae* (Figure 3.4, yellow) were found in all samples with the exception of TL_JUL19. When present, *Podoviridae* had its highest relative abundance in TL_MAY19 (7.6%), and its lowest relative abundance in TL_JUL18 (1.9%). *Siphoviridae*, when present, had its highest relative abundance in TL_JAN20 (19.2%) and the lowest in TL_JUN19 (3.7%).

When looking at the bacteriophage sequences holistically, their relative abundance remains more consistent across most samples. Samples TL_JUL18 through to TL_JUL19 all have bacteriophage abundance around 25%, while the other two samples are more varied, with 16.7% and 41.5% for TL_AUG19 and TL_JAN20, respectively.

The final taxonomic family of importance that was identified are the *Lavidaviridae*, which was found in both the TL_JUL18 (0.17%) and the TL_AUG19 (2.4%) samples. This family represents virophages. Other viral taxonomic families classified that had a <1% abundance were found in the TL_JUL18 sample, and consisted solely of the viral families *Herelleviridae*, and *Autographiviridae,* which infect bacteria, and *Mimiviridae*, which infects eukaryotic algae.

3.4 Constructing Bins for Viral Metagenome-Assembled Genomes (vMAGs)

Viral bins were able to be constructed for each of the six samples. Across all six samples, a total of 104 bins were created, representing a total of 363 viral contigs (Table 3.2).

Table 3.2 Number of Viral Bins Constructed by vRhyme from Metagenomic Data

Sample	# of Bins	# of contigs represented
TL_JUL18	61	189
TL_MAY19	9	34
TL_JUN19	23	86
TL_JUL19	1	3
TL_AUG19	3	12
TL_JAN20	7	39

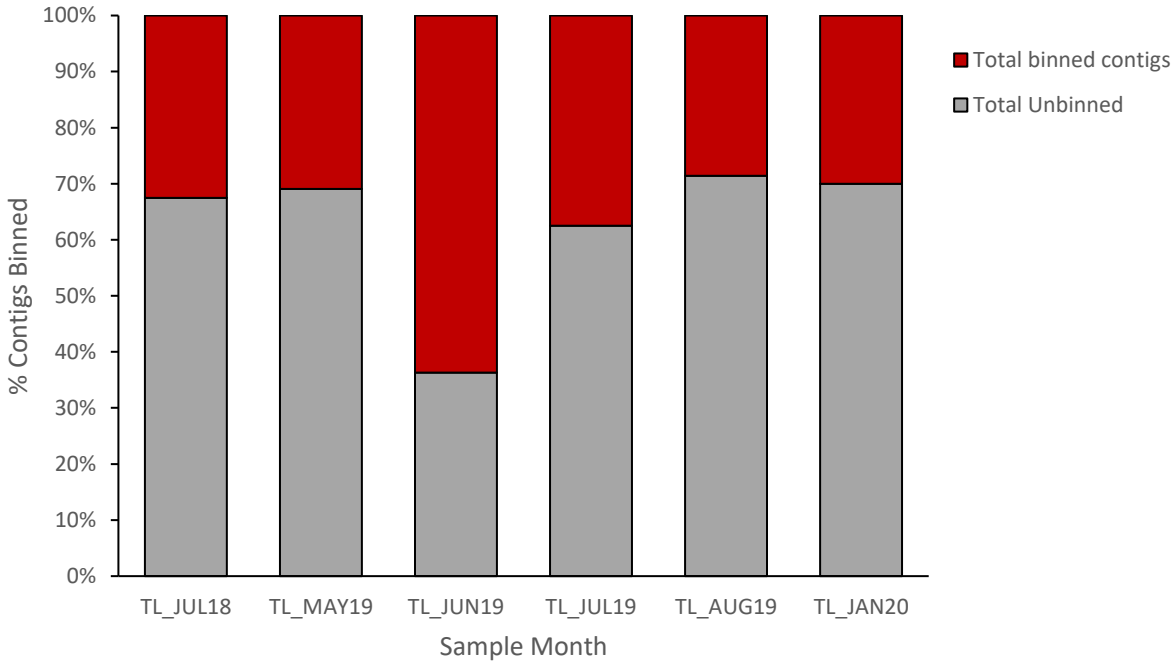


Figure 3.5 Proportion of contigs binned by vRhyme in comparison to the entire sample.

The proportion of binned contigs across all samples remained relatively consistent, with most samples having 30% of their contigs placed in a bin (Figure 3.5). The exception to this was the TL_JUN19 sample, which had over 60% of its viral contigs placed in a bin.

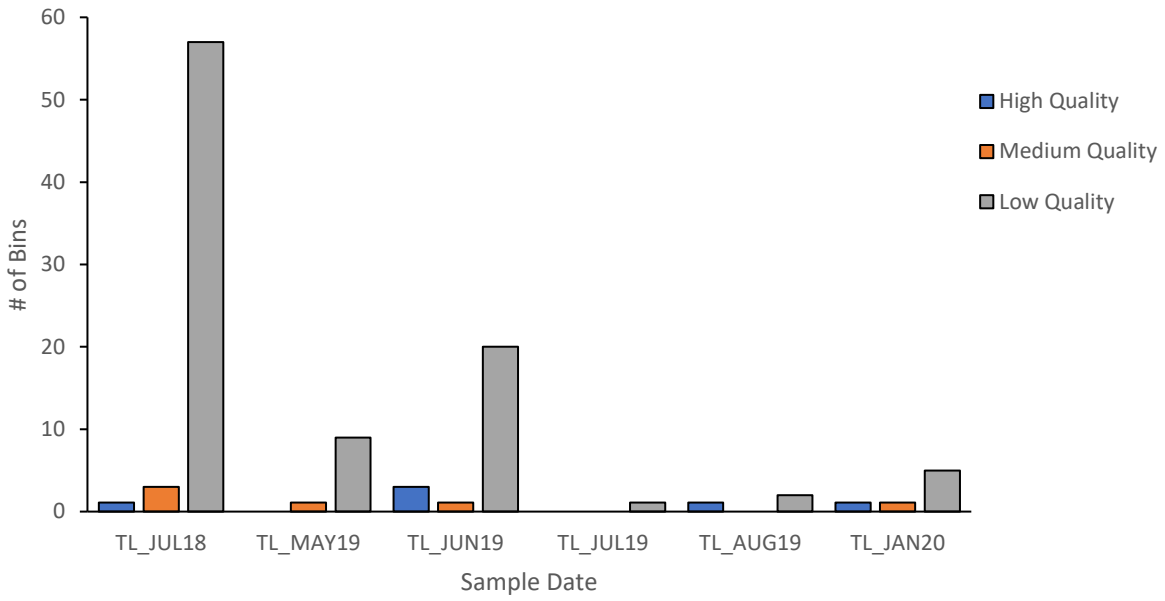


Figure 3.6 Quality Analysis of constructed bins using CheckV. High quality bins were considered those that had >90% completeness rating, medium quality indicates a completeness rating from 50%-90%, and <50% completeness was considered low quality.

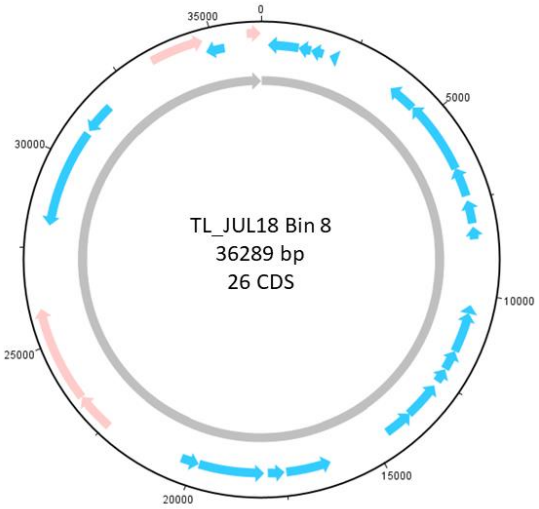
CheckV analysis of the bins (Figure 3.6) showed that the majority of the bins were considered 'Low-Quality', which meant that the completeness level was rated as being <50%. By contrast, six bins across all samples were considered by CheckV to be 'Medium-Quality', and six bins were considered to be 'High Quality', indicating the bins had a completeness of >90%. These six bins indicated to be 'High Quality' were considered candidates for vMAG status.

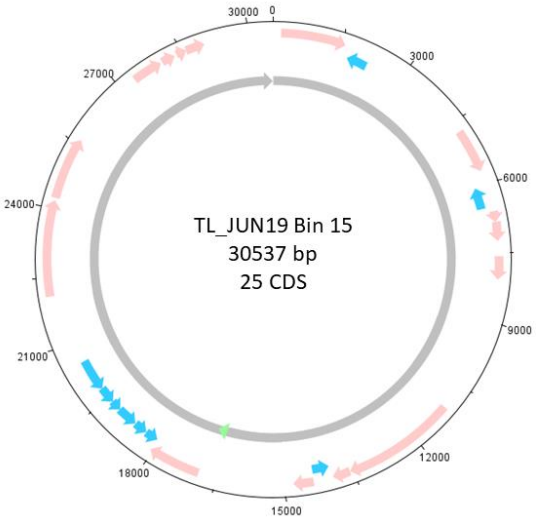
3.5 vMAG analysis using IMG

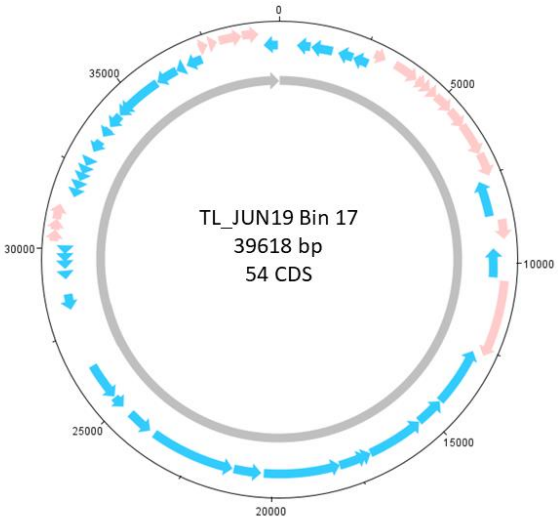
Using the IMG/VR database, the vMAGs that were constructed were analyzed to determine if they could be identified further, and if there are were any predicted hosts that the vMAGs aligned to (Table 3.3). Out of the six, only the vMAG constructed from bin 15 aligned to a virus that infected a known host, which was to a *Synechococcus* sp. However, all of the vMAGs aligned to sequences that were categorized as from *Caudovirales*, indicating all vMAGs were bacteriophages. The vMAGs were unable to be resolved into lower taxonomy. All alignments performed using the IMG/VR database had an e-value of 0.0 and well exceeded the bit score cutoff of 50 used in the previous BLASTp analysis.

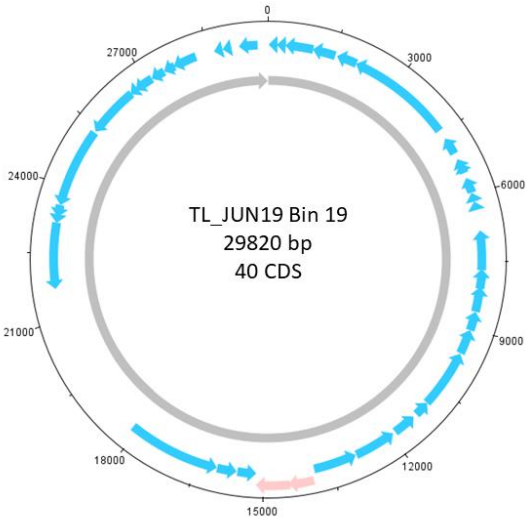
Table 3.3 BLAST analysis of Big Turkey Lake vMAGs using the IMG/VR database

vMAG	Date	Predicted Completeness	Assigned BLASTp Family	IMG/VR Lineage	IMG/VR Predicted Host
TL_JUL18, Bin 8	July '18	100.0	Unassigned	Caudovirales	Unassigned
TL_JUN19, Bin 15	June '19	92.64	Unassigned	Caudovirales	*Synechococcus* sp.
TL_JUN19, Bin 17	June '19	100.0	Unassigned	Caudovirales	Unassigned
TL_JUN19, Bin 19	June '19	100.0	Siphoviridae	Caudovirales	Unassigned
TL_AUG19, Bin 3	Aug '19	93.94	Unassigned	Caudovirales	Unassigned
TL_JAN20, Bin 2	Jan '20	100.0	Unassigned	Caudovirales	Unassigned


TL_JUL18 Bin 8
36289 bp
26 CDS
0
5000
10000
15000
20000
25000
30000
35000


TL_JUN19 Bin 15
30537 bp
25 CDS
0
3000
6000
9000
12000
15000
18000
21000
24000
27000
30000


TL_JUN19 Bin 17
39618 bp
54 CDS
0
5000
10000
15000
20000
25000
30000
35000


TL_JUN19 Bin 19
29820 bp
40 CDS
0
3000
6000
9000
12000
15000
18000
21000
24000
27000

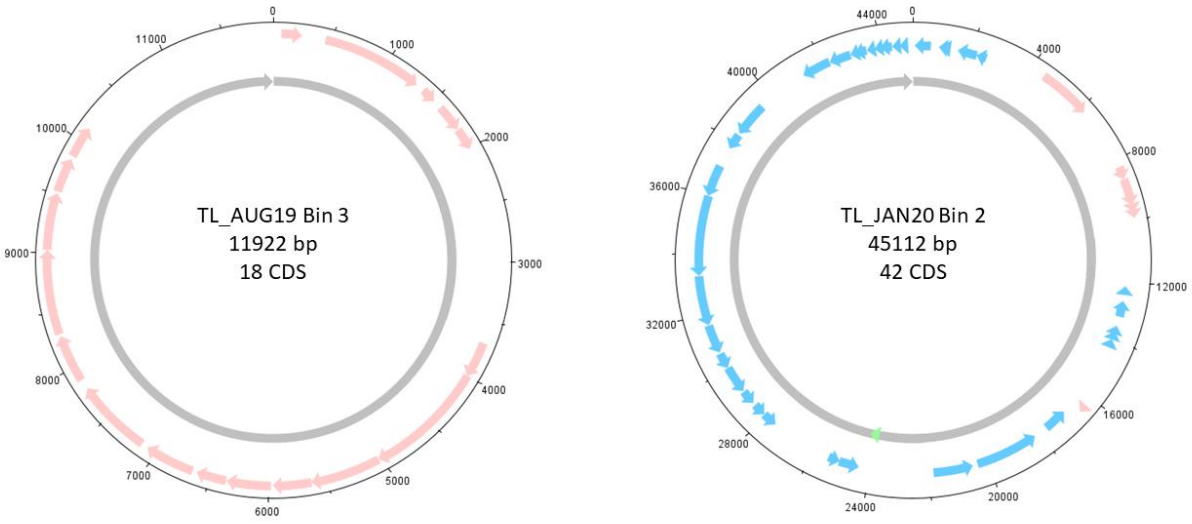


Figure 3.7 Viral Metagenome-Assembled Genomes in Big Turkey Lake as visualized using DNAPlotter. Genes located on the forward strand were coloured pink, while genes on the reverse strand were coloured blue. tRNA genes were coloured green, located on the central track (gray). vMAGs are circularized for visual clarity.

Six vMAGs were visualized using DNAPlotter (Figure 3.7). Across all six samples, only three coding sequences were annotated by PROKKA (1.46% out of 205 CDSs). Two of these were annotated as tRNA coding sequences for tRNA-Ser and tRNA-His, found in TL_MAY19 and TL_JAN20 respectively. The third annotation aligned with a viral exonuclease. The remaining genes were annotated as ‘hypothetical proteins’ by PROKKA, indicating that their function is still unknown.

3.6 Auxiliary Metabolic Genes in Big Turkey Lake

The number of AMGs within each sample varied, although were generally proportional to the amount of viral reads in a sample; i.e. samples that had fewer viral contigs contained fewer AMGs (Table 3.4). There were seven categories that were found across samples (Figure 3.8). The most common AMG categories represented were amino acid metabolism and biosynthesis of secondary metabolites, which were found in five out of the six samples – amino acid metabolism was not found in TL_JUN19, while biosynthesis of secondary metabolites were not found in TL_AUG19. AMGs involved with nucleotide metabolism and cofactor biosynthesis were found in four of the six samples, while energy metabolism genes were found in three samples (.

AMGs involved with carbohydrate and lipid metabolism were the most infrequently found AMG category, with carbohydrate metabolism found in two samples (TL_JUL18, TL_JAN20), and lipid metabolism only found in one sample (TL_JAN20). Lipid metabolism was the only category of AMG that was only represented in the winter (TL_JAN20). Nucleotide metabolism was the most abundant AMG category, with twenty-seven genes across all six samples. Notable AMGs include those involved in pathways for folate biosynthesis, nitrogen fixation, and photosynthesis (Table 3.5).

Table 3.4 Number of Auxiliary Metabolic Genes found across samples in Big Turkey Lake

Sample	# of AMGs	# of viral contigs
TL_JUL18	34	581
TL_MAY19	9	110
TL_JUN19	16	135
TL_JUL19	3	8
TL_AUG19	3	42
TL_JAN20	17	130

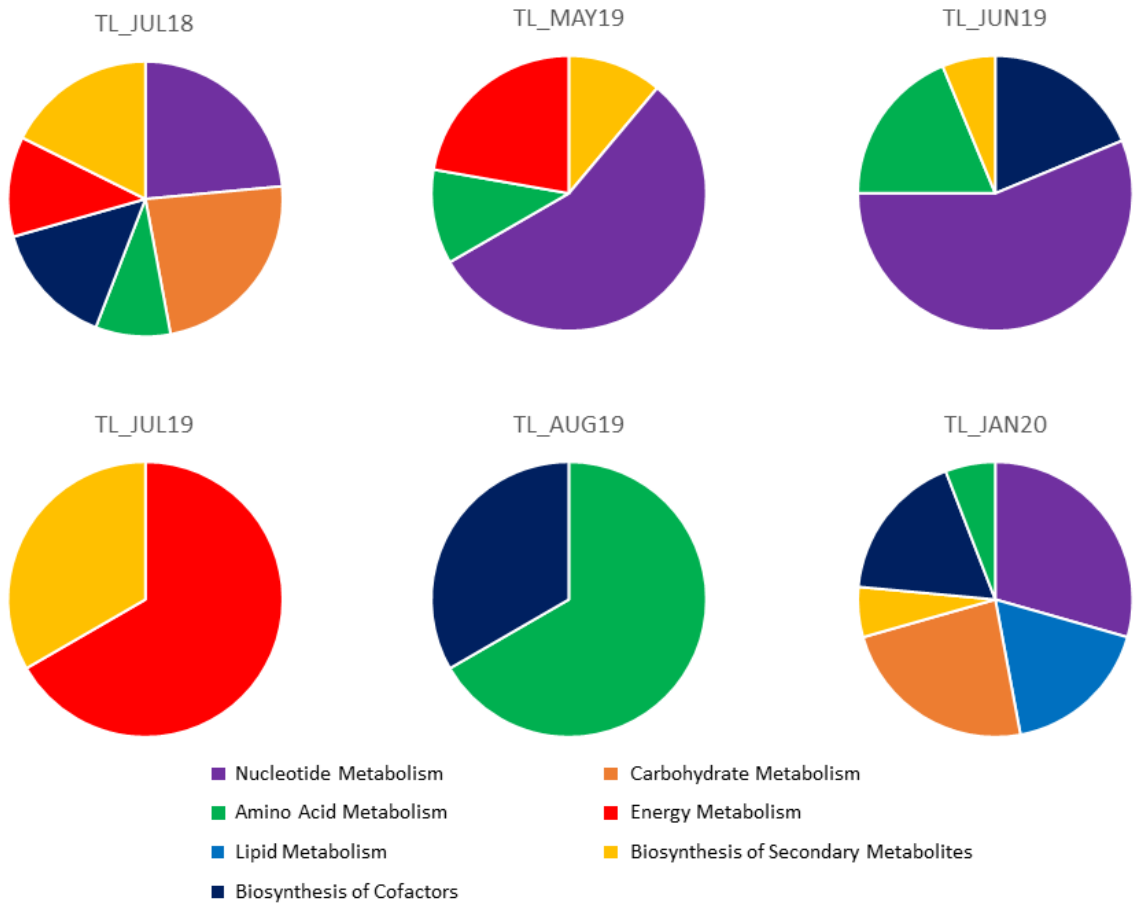


Figure 3.8 Individual Auxiliary Metabolic Gene composition from each sample.

Table 3.5 Highlighted Auxiliary Metabolic Genes within Big Turkey Lake

Name	Function	# of Hits	Pathway
***fol*A, DHFR**	dihydrofolate reductase [EC:1.5.1.3]	1	Biosynthesis of cofactors
***fol*E, GCH1**	GTP cyclopyrrolone IA [EC:3.5.4.16]	1	Biosynthesis of Cofactors
mdh	malate dehydrogenase [EC:1.1.1.37]	2	Energy Metabolism
***nif*U, *isc*U**	nitrogen fixation protein NifU and related proteins	1	Energy Metabolism
***suc*D**	succinyl-CoA synthetase alpha subunit [EC:6.2.1.5]	2	Energy Metabolism
***psb*K**	photosystem II PsbK protein	1	Energy Metabolism
***gua*C, E1.7.1.7**	GMP reductase [EC:1.7.1.7]	2	Nucleotide metabolism
thy*A, *TYMS	thymidylate synthase [EC:2.1.1.45]	2	Nucleotide metabolism
***thy*X, *thy*1**	thymidylate synthase (FAD) [EC:2.1.1.148]	3	Nucleotide metabolism
***gmk*, E2.7.4.8,**	guanylate kinase [EC:2.7.4.8]	1	Nucleotide metabolism
***nrd*A, *nrd*E, E1.17.4.1A,**	ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]	4	Nucleotide Metabolism
***fab*G, *OAR*1**	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	1	Lipid Metabolism
***fab*F, *OXSM*, *CEM*1**	3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	1	Lipid Metabolism
***ugt*P**	processive 1,2-diacylglycerol beta-glucosyltransferase [EC:2.4.1.315]	1	Lipid Metabolism

An extensive list of all AMGs found within Big Turkey Lake over the study period can be found within the Appendix.

Chapter 4: Discussion

4.1 Host Community Analysis of Big Turkey Lake

Using Kraken2 to analyze the assembled metagenomes, potential host taxa were assigned to both the phylum and genus level. On the phylum level across all samples, Proteobacteria and Actinobacteria were the most abundant across samples, with Cyanobacteria and Bacteroidetes also being present in lower abundance. Previous studies in temperate freshwater lakes have shown similar community structures, showing taxa that are known to be relatively common bacterial phyla in freshwater environments (Newton *et al.*, 2011). While all phyla in the bacterial community of Big Turkey Lake do display changes in their relative abundance over time, Cyanobacteria displays a very pronounced seasonal shift, with relative abundance dropping off sharply in the winter compared to the rest of the year. This aligns with current knowledge on Cyanobacterial community dynamics, as it has been shown that the general trend of Cyanobacterial abundance over time is an increase in biomass in the summer and a decrease in biomass in the winter (Agawin *et al.*, 1998).

The genera *Synechococcus* and *Cyanobium* were the genera that displayed the consistently highest abundances in Big Turkey Lake; notably in TL_JUL18 and TL_JUN19 samples, both genera were the only ones to have a >1% abundance. Cameron (2021) previously noted the prevalence of picocyanobacteria within the Turkey Lake Watershed, so the dominance of these genera within these samples aligns with previous studies. Both genera have been reported to carry genes that allow them to produce toxins and taste and odour compounds (Jakubowska & Szelag-Wsielewska, 2015; Journey *et al.*, 2012), so their prevalence in freshwater lakes, is something that should be monitored. Another key cyanobacterial genus found within Big Turkey Lake is *Microcystis*, which had hits across all six samples, indicating it can be found throughout the year. This genus is also particularly noteworthy as toxin-producing strains have been the cause of severe HABs globally (Huisman *et al.*, 2018; Guo, 2007).

Previous work on Big Turkey Lake by Dr. Cameron used amplicon sequencing of the 16S rRNA gene, while this study used shotgun sequencing. Both techniques delivered

similar results in the relative abundance of the phylum Cyanobacteria*;* however, amplicon sequencing of the 16S rRNA gene showed higher relative abundance of cyanobacteria in comparison to the analysis of shotgun sequencing using Kraken2. Recent articles comparing the two methods has shown that neither method is inherently better than the other, as they both have their unique biases (Peterson *et al.,* 2021). Comparative studies between the two techniques leads to mixed results, with some seeing increased representation of low abundance taxa using shotgun sequencing (Durazzi *et al.,* 2021; Brumfield *et al.,* 2020), and others see increased diversity using amplicon sequencing (Peterson *et al.,* 2021). Combining the results of both techniques could allow for covering the other techniques biases, allowing for more confidence in reporting of the viral host communities.

4.2 Combining Viral-Identifying Tools to Analyze Metagenomes

Previous literature has stressed the importance of not relying on one specific identifier as the sole method to identify viral contigs from a metagenomic sample (Kieft & Anantharaman, 2022). Excluding sample TL_JUN19, which did not contain any overlapping contigs that were identified as viral by both programs (VIBRANT & VirSorter2), there was some overlap between the viral contigs that were identified by the two programs used in this study. The range of overlap between the contigs identified was between 14% to 28%, which indicates that the majority of contigs in each sample were unique to one of the two programs used in this study. This supports the current recommendations that were stated by Kieft & Anantharaman (2022) of combining the results of viral identifiers rather than simply relying on one sole viral identifier for the entire sample or study. This broadens the net that is used to identify the contigs as viral and allows for analysis of contigs that would have otherwise been overlooked if using a different method.

4.3 Viral Communities in Big Turkey Lake

The viral communities found in Big Turkey Lake were diverse, with a variety of different bacteriophage families, eukaryotic virus families, and virophages. However, most of the viral sequences that were found from each sample were unable to be assigned taxonomy through BLASTp analysis. This is not overly surprising, as viruses are poorly

represented in databases compared to prokaryotes (Palermo *et al.,* 2019), although it does show that there is still a large amount of diversity that is not being properly seen. The usage of NCBI's nr database mitigates this underrepresentation due to the inclusion of both high-quality and low-quality sequences and environmental samples, which greatly increases the size of the reference database (Bagheri *et al.,* 2020). However, even with this increased reference database, the majority of sequences from each sample were taxonomically unidentified; only the sample TL_JUN19 had more than 40% of the total viral contigs identified after analysis.

Of the contigs that were able to be assigned taxonomy, bacteriophages made up the bulk across all samples. Bacteriophage sequences remained relatively equal in abundance, although TL_JAN20 did have a higher relative abundance compared to the other sample dates. TL_JUL19 was unique because it only contained bacteriophage sequences that aligned with *Myoviridae*. However, this could easily be attributed to sampling bias, as the sample contained only eight viral contigs, which is fewer in comparison to the rest of the samples.

The Kraken2 analysis of bacterial taxa in Big Turkey Lake shows that cyanobacteria are much less abundant in TL_JAN20, while BLASTp analysis shows relative abundance of bacteriophages is at its highest. This might suggest that most of the bacteriophages during the winter are ones who parasitize heterotrophic bacteria rather than cyanobacteria, although there currently isn't enough information to confirm this hypothesis. In future studies, this could be further examined by looking at the relationship between chlorophyll-a and the abundance of bacteriophage in a sample; chlorophyll-a and cyanobacterial viruses tend to be positively correlated within an environment (Vrede *et al.,* 2003; Tijdens *et al.,* 2008; Palermo *et al.,* 2019). This line of thought also reflects other studies in Canadian lakes, which show that viral sequences collected from winter and summer periods do not represent overlapping taxa (Girard *et al.,* 2020). Ultimately, this could explain the difference in relative bacteriophage abundance between samples taken in the summer months and TL_JAN20 that was noted previously.

Also of note from these samples were the eukaryotic algal viruses, of which the vast majority that were found were part of the *Phycodnaviridae* family. These displayed

strong seasonality, with samples from the summer months of June, July, and August having the highest relative abundance compared to the winter and spring months. Because the work by Dr. Cameron did not look at eukaryotic algae and their presence in the lake, direct links between potential virus-host interactions for eukaryotic viruses cannot be drawn for this specific time period, so not many conclusions can be drawn about the seasonality of *Phycodnaviridae* and if it relates back to their host organisms. However, Jeffries *et al.* (1988) did note in their original work that Big Turkey Lake contained members of Chlorophyta, Chrysophyceae, and Pyrrophyta, so it is likely that members of those taxa are hosts for the viruses found in this study and could be candidates for further examination in the future.

Virophages, which co-infect with NCLDVs like those from *Phycodnaviridae*, do not appear to be a significant member of the community in Big Turkey Lake, with only two contigs across all six samples being assigned to *Lavidaviridae*. That being said, virophages are poorly represented in databases, less so than other viruses in the first place (Paez-Espino *et al.*, 2019), so it is possible that the true abundance of virophages could be concealed within the unassigned contigs of this study.

The viral community found in Big Turkey Lake is most interesting when compared to viral communities found in other lakes. Palermo *et al.* (2019) studied the viral communities within Hamilton Harbour, an embayment in Lake Ontario which has been highly contaminated through industrial waste (Giglia, 2015), and in essence, represents a habitat that is the opposite of the Turkey Lake Watershed (which is relatively untouched by human activity). The communities in Hamilton Harbour were markedly different than those found within this study, most notably being in the increased abundance and diversity of virophages, which were the dominant viruses in some samples. In addition, bacteriophages were less abundant than expected. This is much different than the results in the Big Turkey Lake study, where bacteriophages were the dominant identified virus across all samples. There are also key differences in the characterized NCLDV sequences found within Hamilton Harbour, with *Mimiviridae* being the dominant family of algal viruses over *Phycodnaviridae* (Palermo *et al.*, 2019), which was, again, the opposite of what was found in this study. It is possible that these

difference could be due to geographical site variations, as the two sites have drastically different levels of human activity and anthropogenic impacts associated with them. This could lead to interesting research questions in the future regarding the impacts of human activity on viral communities.

4.4 Classifications of vMAGs from Big Turkey Lake

Of the six vMAGs that were identified, only one was able to be resolved to the family level through BLASTp analysis in the nr database, which was assigned to the family *Siphoviridae*. Further analysis using the IMG/VR database from JGI did not resolve any other vMAGs to the family level, although it did show that all six of the vMAGs aligned with other *Caudovirales* viruses indicating that all vMAGs are bacteriophages. In addition, one vMAG aligned with a virus that is known to infect the cyanobacteria *Synechococcus*, which is one of the most common genera of cyanobacteria in Big Turkey Lake (Cameron, 2021; this study). Notably, the cyanobacteria population within the sample that this vMAG was identified in, TL_JUN19, was dominated by *Synechococcus*.

Three sequences annotated with PROKKA were identified, with two being tRNA (tRNA-Ser and tRNA-His). The purpose of tRNA within bacteriophages has not been fully elucidated, although current research seems to link viral tRNA genes with increased viral fitness, with tRNA deletions from viral genomes causing a decrease in both burst size and protein synthesis (Albers & Czech, 2016). One possible explanation for the decreased burst size and protein synthesis is that tRNAs in viral genomes are included because they represent codons that are rare within their host, but common within the phage genome – by including these tRNA genes, a virus is then able to replicate more efficiently (Bailly-Bechet, Vergassola, & Rocha, 2007). The third sequence identified by PROKKA was an exonuclease, which are commonly found in bacteriophage genomes as part of recombination machinery. This aids the virus in overcoming host defenses, and in its own replication (Brewster & Tolun, 2020). The remaining coding sequences in the vMAGs were annotated as hypothetical proteins, indicating they did not have similar protein representatives in the PROKKA database. While many of these genes were annotated as hypothetical, that doesn't mean that these vMAGs are truly novel; instead,

these vMAGs are most likely bacteriophages that are poorly represented within the conventional databases that most programs use for their annotation.

4.5 Viral Roles in Big Turkey Lake through Auxiliary Metabolic Gene Analysis

Auxiliary Metabolic Genes (AMGs) were found across all samples with more AMGs being found in samples that had more viral contigs. Most categories of AMGs did not display clear trends of seasonality between samples, although this is hard to fully determine as two samples only contained three AMGs, causing a strong likelihood of sampling bias. The exception to this, however, are AMGs that are related to lipid metabolism, which were only found in the winter sample (TL_JAN20). These genes were 1,2-diacylglycerol beta-glucosyltransferase (*ugt*P) and 3-oxoacyl-[acyl-carrier protein] reductase and synthase (*fab*F, *fab*G), both of which are important in the biosynthesis of membrane lipids (Matsuoka *et al.,* 2016; Guo *et al.,* 2019). In addition, the *fab*G gene is highly conserved among bacterial species, emphasizing its importance (Guo *et al.,* 2019). These AMGs in particular are quite interesting, as bacteria are known to adapt to cold temperatures by increasing production of membrane lipids and fatty acids (Hassan *et al.,* 2020). It is possible that viruses that infect organisms during this period of the year have adapted to colder weather conditions by forcing their hosts to produce more fatty acids. It is also possible that, if the virus integrates itself into the host genome through lysogeny, these genes would be helping the host survive the winter through the production of these lipid metabolism genes, until a more appropriate time for replication arrives. After all, it would be in the best interest of the virus that its host does not die before its viral progeny can be successfully replicated.

The most abundant pathway represented in these AMGs were those related to nucleotide metabolism, which had multiple genes like guanine reductase (*gua*C), guanylate kinase (*gmk*), and ribonucleoside-diphosphate reductase (*nrd*A) which were represented across multiple samples. AMGs in this category would be important for helping to bolster the biosynthesis of nucleic acids, which is an important step in viral replication (Gao *et al.,* 2016; Crummett *et al.,* 2016). Viruses with these genes would be able to produce more viral progeny and have higher fitness than those that do not.

One pathway that was represented across samples was related to folate biosynthesis, with genes present such as thymidylate synthase (*thy*X, *thy*A), and dihydrofolate reductase (*fol*A), all of which have been found in viruses previously (Luo *et al.*, 2022). As viruses downregulate production of host proteins (Fabricant & Kennel, 1970; Hurwitz & U'ren, 2016), and folate is an essential nutrient for DNA and protein synthesis (Bermingham & Derrick, 2002), it is possible that viruses in Big Turkey Lake have evolved to force their hosts to produce more folic acid to boost their own replication. Other key AMGs that were found within this study were those related to energy metabolism pathways. Specifically, succinyl-CoA synthetase (*suc*D) and malate dehydrogenase (*mdh*) were represented across multiple samples, both of which are genes involved in the TCA cycle.

There were notable AMGs found that have potential impact on the environment in Big Turkey Lake, for example a protein involved in photosynthesis: photosystem II (*psbK*). Oceanic viruses have been found to be common carriers of photosystem II genes, and it has been suggested that viruses may be key players in global photosynthesis rates (Lindell *et al.*, 2005; Heyerhoff *et al.*, 2022), with metatranscriptome studies showing that viruses can contribute 40% of *psb*A expression within an environment (Sieradzki *et al.*, 2019). This implies viruses in marine environments are key players in global oxygen production. The presence of these genes in freshwater viruses suggest that they could be playing a similar role to oceanic viruses in increasing the global rates of photosynthesis in lakes.

Another key AMG that could imply another potential role for viruses was in nitrogen fixation, specifically in the gene *nif*U. The presence of this gene implies that viruses might play a role in the nitrogen cycle within Big Turkey Lake. This is not the first time that the *nif*U gene has been found in viruses, with the gene showing up in studies sampling from oceanic environments (Williamson *et al.*, 2008; Mara *et al.*, 2020). The specific pathway being nitrogen fixation is also notable; infamous bloom-former *M. aeruginosa* is unable to fix nitrogen and relies on external sources to get the nitrogen it needs (Paerl *et al.*, 2014). While purely speculative, AMGs related to the nitrogen cycle and specifically to nitrogen fixation may cause higher amounts of nitrogen within a body

of water as they infect their hosts, which could contribute to a future bloom of *M. aeruginosa*. While certainly this would not be the only cause of a potential bloom, viruses may be another key piece of the puzzle that future work on bloom mitigation should consider.

Of course, it is vital to state that this is not something that can be concluded from a single gene found within a sample from Big Turkey Lake. In fact, the gene *nif*U has been shown to be not required for nitrogen fixation (Lyons & Thiel, 1995), and similar *nif*U-like genes have been found in non-nitrogen fixing organisms (Hwang *et al.,* 1996). While speculation about viruses and links to nitrogen fixation can certainly be interesting, there is not currently enough evidence to do anything but that: speculate.

4.6 Implications of Sample Filtration on Results

The original samples taken by Dr. Cameron were filtered through 1.2 µm pore size filters, with the resulting DNA extraction being done on the biomass recovered from the filters. With a pore size of 1.2 µm, even the largest viruses such as members of the NCLDV clade are too small to be caught by the filter, as their particle sizes generally range from 0.1 µm (Colson *et al.,* 2013) to 1.2 µm in length (Legendre *et al.,* 2014). Logically, bacteriophages that are smaller than members of the NCLDV would also not be captured by the filter. Due to the loss of viruses that pass through the pores during filtration, the diversity of viruses that are recovered after DNA extraction and analysis do not truly reflect the diversity of viruses within Big Turkey Lake at the time of sampling. In addition, viruses are known to be present within the sediment of lakes (Mei & Danovaro, 2004), and would also not be captured and represented within this study. This limitation as a result of pore size would also extend to the diversity of bacteria found within each sample, as picocyanobacteria range from 0.2 µm to 2.0 µm (Śliwińska-Wilczewska *et al.,* 2018), meaning that it is likely that not all bacterial diversity has been captured.

The implication of this limitation by pore size is that the majority of the viruses captured by this study are those that were associated with cells. This includes viruses that were either actively infecting cells and replicating or have attached to cells and were preparing for infection. While this does mean that the study performed here does not fully

encapsulate the entire scope of viral diversity in Big Turkey Lake, it allows for the examination of viruses that are active in the environment, and these are the viruses that are most likely to have a larger impact on the environment through AMGs and through their effects on cyanobacterial lysis.

4.7 State of Viral Taxonomy in the Future

One of the issues with assigning taxonomy to viral contigs at the current point in time is the current state of flux that we find ourselves in. Historically, the taxonomy of bacteriophages has been separated through morphology, with the three major families of *Myoviridae, Siphoviridae*, and *Podoviridae* all found within the order *Caudovirales* (Turner *et al.*, 2021). With the increased prevalence of genetic analysis within the past two decades, there has been a general shift away from morphology-based taxonomy; in the past two years, it has been suggested that the separation into the three major morphology-based families is an outdated practice, and that the families should be abolished in total (Turner *et al.*, 2021; Adriaenssens, 2021). This also coincides with an expansion of the taxonomic hierarchies that are used to classify viruses, from the previous five ranks to the new fifteen ranks (International Committee on Taxonomy of Viruses Executive Committee, 2020). With this taxonomic rank expansion and potential changes to abolish previously used bacteriophage families, it can be expected that the entire system of viral taxonomy will soon be upended.

As stated previously, the morphology-based taxonomy that is currently used does not lend itself well to identifying potential host organisms for viruses (Xia *et al.*, 2013). However, some more recently identified viral families are based in genetics rather than morphology and have defined host ranges. E.g., *Herelleviridae,* which currently only infects bacteria from the phylum Firmicutes (Barylski *et al.*, 2020). While purely speculative, it is possible that a reorganization of viral families will emphasize relationships between viruses that share hosts, or what family of organism the virus most likely infects.

Chapter 5: Conclusions and Further Research

5.1 Diversity of Viruses in Freshwater Remains Largely Unknown

While the diversity of viruses within Big Turkey Lake is quite complex, most viruses within the lake are unidentified at the time of this study's completion. This is mostly likely due to the underrepresentation of viruses within the databases that were used to assign taxonomy. This study is limited with regards to both the number of samples and the replication of samples taken, which does not permit for in-depth statistical analysis. While current data does imply seasonality among viruses in Big Turkey Lake, there is not enough information to conclude with confidence on diversity differences among viral communities, and if that seasonality is related to their hosts seasonality. *Phycodnaviridae,* eukaryotic algal viruses, seem to be most abundant during the summer. Bacteriophages, on the other hand, seem to be more consistently abundant across samples and most seasons, although they were more abundant in the winter.

These observations and their interpretation will likely change as the taxonomy of viruses shift from being based on morphology to being based on genetics. It is currently unclear how, exactly, this will influence the results of this study, but it is possible that more information can be gleaned about the viral community and their potential hosts after the taxonomy has been updated. These observations, which are based on individual samples across seasons, will also likely change with further sampling in the future; this includes the sequencing of replicates, and covering a larger temporal and spatial scale with increased sampling across many years.

5.2 Viral Auxiliary Metabolic Genes within Big Turkey Lake Display Potential Viral Roles in Nitrogen Fixation, Carbon Fixation, and Photosynthesis

AMGs were found across all samples taken from Big Turkey Lake; most categories of AMGs were consistent across all samples. Lipid metabolism genes were only found during the winter, which could be a potential adaptation for increased fitness during period of low temperatures. Genes that were for pathways such as amino acid metabolism, nucleotide metabolism, and the biosynthesis of secondary metabolites were among the most common AMGs within this environment. Individual AMGs that are

found across samples were important for folate metabolism (*thy*X, *fol*E, *fol*A) and for energy production and carbon metabolism through the TCA cycle (*suc*D, *mdh*). In addition, the presence of *psb*K suggests that viruses in freshwater may also have a large impact on the rate of photosynthesis within freshwater environments, similar to the use of photosynthesis AMGs in marine environments (Lindell *et al.*, 2005; Heyerhoff *et al.*, 2022). Other AMGs have also suggested that viruses in Big Turkey Lake might have a hand in the nitrogen cycle, due to the presence of nitrogen fixation gene *nif*U, within one sample. This may have significant implications for harmful algal blooms, as excess bioavailable nitrogen can be a key factor in their proliferation.

The presence of lipid metabolism AMGs during the winter period as a possible adaptation to colder temperatures is interesting, and to my knowledge has not been highlighted in previous research. Further research into viral AMGs could consider examining seasonality within these genes to see if lipids are, in fact a seasonal adaptation, or if there are other genes that provide similar advantages during different points of the year.

5.3 Future Research on Viruses in Big Turkey Lake Should be Focused on Virus-Host Relationships and Physicochemical Links to Seasonality

This study presents baseline knowledge of the type of microbial cell-associated viruses present within Big Turkey Lake, the taxonomic families that these viruses belong to, and their potential function within the wider ecosystem. One thing that is missing from this study is an exploration of potential hosts for these viruses. This would shed light on which viruses infect which hosts, how viruses potentially influence the abundance of cyanobacteria, and ultimately could be linked to important questions surrounding cyanobacteria in freshwater environments, like the production of toxins and taste and odour compounds. As the overall goal within the research at the Turkey Lakes Watershed is related to impacts of cyanobacteria on water quality, this would be an important addition to our knowledge on this subject.

There are numerous tools that could be used to explore this avenue of research, such as, but not limited to, HoPhage (Tan *et al.*, 2021), HostPhinder (Villarroel *et al.*, 2016), and VirHostMatcher (Ahlgren *et al.*, 2017). While all of these tools rely on

different methods to identify hosts from viral metagenomic data, some of them do run into the issue of underrepresentation in databases preventing accurate prediction (Tan *et al.*, 2021). Another avenue to explore would be the use of CRISPR spacers to assign hosts to a virus based on the similarity between the two (Sanguino *et al.*, 2015). Ultimately, these approaches for the identification of viral hosts will allow for further understanding of viruses in freshwater environments and should be explored in the future.

Another important aspect to focus on in the future is whether seasonality within viral communities in Big Turkey Lake is truly occurring, as this is currently the only study to examine these trends. In addition, this study has a low number of samples, so definitive conclusions are hard to declare – as stated before, more replicates and a longer study duration will help tease out more information about viral communities and their link to seasonal patterns. Once this has been completed, and if future studies reinforce the findings of this current study, linking these communities to various physicochemical factors like temperature, pH, nutrient concentrations, etc. will allow us to understand how viral communities change with their environment, and in turn, tell us more about how these viruses interact with their host species.

Bibliography

Adriaenssens, E. M. (2021). Phage Diversity in the Human Gut Microbiome: a Taxonomist's Perspective. *MSystems*, *6*(4), e00799-21. https://doi.org/10.1128/mSystems.00799-21

Agawin, N. S., Duarte, C., & Agustí, S. (1998). Growth and abundance of Synechococcus sp. in a Mediterranean Bay: Seasonality and relationship with temperature. *Marine Ecology-Progress Series*, *170*, 45–53. https://doi.org/10.3354/meps170045

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research*, *45*(1), 39–53. https://doi.org/10.1093/nar/gkw1002

Albers, S., & Czech, A. (2016). Exploiting tRNAs to Boost Virulence. *Life*, *6*(1), 4. https://doi.org/10.3390/life6010004

Allinger, L. E., & Reavie, E. D. (2013). The ecological history of Lake Erie as recorded by the phytoplankton community. *Journal of Great Lakes Research 39*(3), 365–382. https://doi.org/10.1016/j.jglr.2013.06.014

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science (New York, N.Y.)*, *344*(6185), 757–760. https://doi.org/10.1126/science.1252229

Anderson, D. M., Cembella, A. D., & Hallegraeff, G. M. (2012). Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annual Review of Marine Science*, *4*, 143–176. https://doi.org/10.1146/annurev-marine-120308-081121

Andrade-Martínez, J. S., Valera, L. C. C., Cárdenas, L. A. C., Laura, F.-J., Gamaliel, L.-L., Leonardo, M.-G. J., Guillermo, R.-P., & Alejandro, R. (2022). Computational Tools for the Analysis of Uncultivated Phage Genomes. *Microbiology and Molecular Biology Reviews*, *86*(2), e00004-21. https://doi.org/10.1128/mmbr.00004-21

Bagheri, H., Dyer, R., Severin, A., Rajan, H. (2020). Comprehensive Analysis of Non Redundant Protein Database. *Research Square*. https://doi.org/10.21203/rs.3.rs-54568/v1

Bailly-Bechet, M., Vergassola, M., & Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome research*, *17*(10), 1486–1495. https://doi.org/10.1101/gr.6649807

Baker, M. L., Jiang, W., Rixon, F. J., & Chiu, W. (2005). Common ancestry of herpesviruses and tailed DNA bacteriophages. *Journal of Virology*, *79*(23), 14967–14970. https://doi.org/10.1128/JVI.79.23.14967-14970.2005

Barylski, J., Kropinski, A. M., Alikhan, N. F., Adriaenssens, E. M., & ICTV Report Consortium (2020). ICTV Virus Taxonomy Profile: *Herelleviridae*. *The Journal of general virology*, *101*(4), 362–363. https://doi.org/10.1099/jgv.0.001392

Bermingham, A., & Derrick, J. P. (2002). The folic acid biosynthesis pathway in bacteria: evaluation of potential for antibacterial drug discovery. *BioEssays*, *24*(7), 637–648. https://doi.org/https://doi.org/10.1002/bies.10114

Boehrer, B., & Schultze, M. (2008). Stratification of lakes. *Reviews of Geophysics*, *46*(2). https://doi.org/https://doi.org/10.1029/2006RG000210

Bratbak, G., Egge, J. K., & Heldal, M. (1993). Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series*, *93*(1/2), 39–48. http://www.jstor.org/stable/24832806

Breitbart, M., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography*, *20*(SPL.ISS. 2), 135–139. https://doi.org/10.5670/oceanog.2007.58

Breitwieser, F. P., & Salzberg, S. L. (2020). Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics (Oxford, England)*, *36*(4), 1303–1304. https://doi.org/10.1093/bioinformatics/btz715

Brewster, J. L., & Tolun, G. (2020). Half a century of bacteriophage lambda recombinase: In vitro studies of lambda exonuclease and Red-beta annealase. *IUBMB Life*, *72*(8), 1622–1633. https://doi.org/https://doi.org/10.1002/iub.2343

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., & Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLOS ONE*, *15*(2), e0228899-. https://doi.org/10.1371/journal.pone.0228899

Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K., Paul, J. H., Schroeder, D. C., Suttle, C. A., Vaqué, D., & Wommack, K. E. (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *The ISME Journal*, *2*(6), 575–578. https://doi.org/10.1038/ismej.2008.31

Bushnell, B. (2019). *BBTools*. Sourceforge. https://sourceforge.net/projects/bbmap/

Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics (Oxford, England)*, *25*(1), 119–120. https://doi.org/10.1093/bioinformatics/btn578

Cameron, E. S. (2021). Spatiotemporal Shifts in Cyanobacterial Communities in a Northern Temperate Watershed – Applications of Next-Generation Sequencing and Implications for Monitoring and Climate Change Adaptation. UWSpace. http://hdl.handle.net/10012/17185

Chen, I.-M. A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., Roux, S., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research*, *49*(D1), D751–D763. https://doi.org/10.1093/nar/gkaa939

Colson, P., de Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D. K., Cheng, X.-W., Federici, B. A., van Etten, J. L., Koonin, E. v, la Scola, B., & Raoult, D. (2013). "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archives of Virology*, *158*(12), 2517–2521. https://doi.org/10.1007/s00705-013-1768-6

Conley, D. J., Paerl, H. W., Howarth, R. W., Boesch, D. F., Seitzinger, S. P., Havens, K. E., Lancelot, C., & Likens, G. E. (2009). Controlling eutrophication: Nitrogen and phosphorus. *Science 323*(5917), 1014–1015. https://doi.org/10.1126/science.1167755

Coutinho, F. H., Cabello-Yeves, P. J., Gonzalez-Serrano, R., Rosselli, R., López-Pérez, M., Zemskaya, T. I., Zakharenko, A. S., Ivanov, V. G., & Rodriguez-Valera, F. (2020). New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome*, *8*(1), 163. https://doi.org/10.1186/s40168-020-00936-4

Creed, I. F., Hwang, T., Lutz, B., & Way, D. (2015). Climate warming causes intensification of the hydrological cycle, resulting in changes to the vernal and autumnal windows in a northern temperate forest. *Hydrological Processes*, *29*(16), 3519–3534. https://doi.org/https://doi.org/10.1002/hyp.10450

Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F., & Martiny, J. B. H. (2016). The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology*, *499*, 219–229. https://doi.org/10.1016/j.virol.2016.09.016

Duda, R. L., & Teschke, C. M. (2019). The amazing HK97 fold: versatile results of modest differences. *Current opinion in virology*, *36*, 9–16. https://doi.org/10.1016/j.coviro.2019.02.001

Dunlap, C. R., Sklenar, K. S., & Blake, L. J. (2015). A Costly Endeavor: Addressing Algae Problems in a Water Supply. *Journal (American Water Works Association)*, *107*(5), E255–E262. https://www-jstor-org/stable/jamewatworass.107.5.e255

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., & de Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, *11*(1), 3030. https://doi.org/10.1038/s41598-021-82726-y

Environment Canada. (2014). *Map depicting the Turkey Lakes Watershed with study site boundaries and main camp location* [Image]. Government of Canada. https://www.canada.ca/en/environment-climate-change/services/turkey-lakes-watershed-study/site.html

Fabricant, R., & Kennell, D. (1970). Inhibition of host protein synthesis during infection of Escherichi coli by bacteriophage T4. 3. Inhibition by ghosts. *Journal of Virology*, *6*(6), 772–781. https://doi.org/10.1128/JVI.6.6.772-781.1970

Fischer, M. G. (2020). The virophage family lavidaviridae. *Current Issues in Molecular Biology*, *40*, 1–24. https://doi.org/10.21775/cimb.040.001

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature 399*(6736). 541–548. https://doi.org/10.1038/21119

Gao, E.-B., Huang, Y., & Ning, D. (2016). Metabolic Genes within Cyanophage Genomes: Implications for Diversity and Evolution. *Genes*, *7*(10). https://doi.org/10.3390/genes7100080

Giglia, S.N. (2015). From Man vs. Nature to Environment vs. Budget – The Shifting Battles in the History of Pollution and Toxicity in Hamilton Harbour. *The Great Lakes Journal of Undergraduate History. 3*(1) 20–35. https://scholar.uwindsor.ca/gljuh/vol3/iss1/4

Girard, C., Langlois, V., Vigneron, A., Vincent, W. F., & Culley, A. I. (2020). Seasonal Regime Shift in the Viral Communities of a Permafrost Thaw Lake. *Viruses*, *12*(11), 1204. https://doi.org/10.3390/v12111204

Glibert, P. M., Al-Azri, A., Icarus Allen, J., Bouwman, A. F., Beusen, A. H. W., Burford, M. A., Harrison, P. J., & Zhou, M. (2018). *Key Questions and Recent Research Advances on Harmful Algal Blooms in Relation to Nutrients and Eutrophication BT - Global Ecology and Oceanography of Harmful Algal Blooms*, 229–259. https://doi.org/10.1007/978-3-319-70069-4_12

Gobler, C. J., & Sunda, W. G. (2012). Ecosystem disruptive algal blooms of the brown tide species, Aureococcus anophagefferens and Aureoumbra lagunensis. *Harmful Algae*, *14*, 36–45. https://doi.org/https://doi.org/10.1016/j.hal.2011.10.013

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome, 9*(1), 37. https://doi.org/10.1186/s40168-020-00990-y

Guo, L. (2007). Ecology: Doing battle with the green monster of Taihu Lake. *Science, 317*(5842), 1166. https://doi.org/10.1126/science.317.5842.1166

Guo, Q.-Q., Zhang, W.-B., Zhang, C., Song, Y.-L., Liao, Y.-L., Ma, J.-C., Yu, Y.-H., & Wang, H.-H. (2019). Characterization of 3-Oxacyl-Acyl Carrier Protein Reductase Homolog Genes in Pseudomonas aeruginosa PAO1. *Frontiers in Microbiology*, *10*. https://www.frontiersin.org/articles/10.3389/fmicb.2019.01028

Haaber, J., & Middelboe, M. (2009). Viral lysis of Phaeocystis pouchetii: Implications for algal population dynamics and heterotrophic C, N and P cycling. *The ISME Journal*, *3*(4), 430–441. https://doi.org/10.1038/ismej.2008.125

Hassan, N., Anesio, A. M., Rafiq, M., Holtvoeth, J., Bull, I., Haleem, A., Shah, A. A., & Hasan, F. (2020). Temperature Driven Membrane Lipid Adaptation in Glacial Psychrophilic Bacteria. *Frontiers in microbiology*, *11*, 824. https://doi.org/10.3389/fmicb.2020.00824

Health Canada. (2021). *Guidelines for Canadian Drinking Water Quality: Guideline Techincal Document – cyanobacterial Toxins.* Retrieved from Government of Canada website: https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidelines-canadian-drinking-water-quality-guideline-technical-document-cyanobacterial-toxins-document.html

Heyerhoff, B., Engelen, B., & Bunse, C. (2022). Auxiliary Metabolic Gene Functions in Pelagic and Benthic Viruses of the Baltic Sea. *Frontiers in microbiology*, *13*, 863620. https://doi.org/10.3389/fmicb.2022.863620

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., & Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology*, *16*(8), 471–483. https://doi.org/10.1038/s41579-018-0040-1

Hurwitz, B. L., & Sullivan, M. B. (2013). The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLOS ONE*, *8*(2), e57355-. https://doi.org/10.1371/journal.pone.0057355

Hurwitz, B. L., & U'Ren, J. M. (2016). Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology*, *31*, 161–168. https://doi.org/10.1016/j.mib.2016.04.002

Hwang, D. M., Dempsey, A., Tan, K.-T., & Liew, C.-C. (1996). A modular domain of NifU, a nitrogen fixation cluster protein, is highly conserved in evolution. *Journal of Molecular Evolution*, *43*(5), 536–540. https://doi.org/10.1007/BF02337525

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

International Committee on Taxonomy of Viruses Executive Committee, Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., … Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, *5*(5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Jakubowska, N., & Szeląg-Wasielewska, E. (2015). Toxic picoplanktonic cyanobacteria--review. *Marine drugs*, *13*(3), 1497–1518. https://doi.org/10.3390/md13031497

Jeffries, D. S., & Foster, N. W. (2001). The Turkey Lakes Watershed Study: Milestones and Prospects. *Ecosystems*, *4*(6), 501–502. https://doi.org/10.1007/s10021-001-0023-2

Jeffries, D. S., Kelso, J. R. M., & Morrison, I. K. (1988). Physical, chemical, and biological characteristics of the Turkey Lakes Watershed, central Ontario, Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, *45*(Suppl.1), 3–13. https://doi.org/10.1139/f88-262

Jian, H., Yi, Y., Wang, J., Hao, Y., Zhang, M., Wang, S., Meng, C., Zhang, Y., Jing, H., Wang, Y., & Xiao, X. (2021). Diversity and distribution of viruses inhabiting the deepest ocean on Earth. *The ISME Journal*, *15*(10), 3094–3110. https://doi.org/10.1038/s41396-021-00994-y

Journey, C. A., Beaulieu, K. M., & Bradley, P. M. (2013). Environmental Factors that Influence Cyanobacteria and Geosmin Occurrence in Reservoirs. *Current Perspectives in Contaminant Hydrology and Water Resources Sustainability* Ch. 2. https://doi.org/10.5772/54807

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Kieft, K., Adams, A., Salamzade, R., Kalan, L., & Anantharaman, K. (2022). vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Research*, *50*(14), e83–e83. https://doi.org/10.1093/nar/gkac341

Kieft, K., & Anantharaman, K. (2022). Virus genomics: what is being overlooked? *Current Opinion in Virology*, *53*, 101200. https://doi.org/https://doi.org/10.1016/j.coviro.2022.101200

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1), 90. https://doi.org/10.1186/s40168-020-00867-0

Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., Hess, M., Sullivan, M. B., Walsh, D. A., Roux, S., & Anantharaman, K. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nature communications*, *12*(1), 3503. https://doi.org/10.1038/s41467-021-23698-5

Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M., & McCue, L. A. (2020). ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics*, *21*(1), 257. https://doi.org/10.1186/s12859-020-03585-4

Koonin, E., & Yutin, N. (2010). Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. *Intervirology*, *53*(5), 284–292. https://doi.org/10.1159/000312913

Koonin, E. V., & Yutin, N. (2019). Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Advances in virus research*, *103*, 167–202. https://doi.org/10.1016/bs.aivir.2018.09.002

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., & Claverie, J.-M. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences*, *111*(11), 4274–4279. https://doi.org/10.1073/pnas.1320670111

Li, X., Dreher, T. W., & Li, R. (2016). An overview of diversity, occurrence, genetics and toxin production of bloom-forming Dolichospermum (Anabaena) species. *Harmful Algae 54*, 54–68. Elsevier B.V. https://doi.org/10.1016/j.hal.2015.10.015

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., & Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, *438*(7064), 86–89. https://doi.org/10.1038/nature04111

Lu, J., Breitwieser, F. P., Thielen, P., Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* 3:e104 https://doi.org/10.7717/peerj-cs.104

Luo, X.-Q., Wang, P., Li, J.-L., Ahmad, M., Duan, L., Yin, L.-Z., Deng, Q.-Q., Fang, B.-Z., Li, S.-H., & Li, W.-J. (2022). Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome*, *10*(1), 190. https://doi.org/10.1186/s40168-022-01384-y

Lyons, E. M., & Thiel, T. (1995). Characterization of nifB, nifS, and nifU genes in the cyanobacterium Anabaena variabilis: NifB is required for the vanadium-dependent nitrogenase. *Journal of Bacteriology*, *177*(6), 1570–1575. https://doi.org/10.1128/jb.177.6.1570-1575.1995

Maniloff, J., & Ackermann, H.-W. (1998). Taxonomy of bacterial viruses: establishment of tailed virus genera and the other Caudovirales. *Archives of Virology*, *143*(10), 2051–2063. https://doi.org/10.1007/s007050050442

Matsuoka, S., Seki, T., Matsumoto, K., & Hara, H. (2016). Suppression of abnormal morphology and extracytoplasmic function sigma activity in Bacillus subtilis ugtP mutant cells by expression of heterologous glucolipid synthases from Acholeplasma laidlawii. *Bioscience, Biotechnology, and Biochemistry*, *80*(12), 2325–2333. https://doi.org/10.1080/09168451.2016.1217147

Mara, P., Vik, D., Pachiadaki, M. G., Suter, E. A., Poulos, B., Taylor, G. T., Sullivan, M. B., & Edgcomb, V. P. (2020). Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *The ISME Journal*, *14*(12), 3079–3092. https://doi.org/10.1038/s41396-020-00739-3

Marquet, M., Hölzer, M., Pletz, M. W., Viehweger, A., Makarewicz, O., Ehricht, R., & Brandt, C. (2022). What the Phage: A scalable workflow for the identification and analysis of phage sequences. *BioRxiv*, 2020.07.24.219899. https://doi.org/10.1101/2020.07.24.219899

McKindles, K. M., Manes, M. A., DeMarco, J. R., McClure, A., McKay, R. M., Davis, T. W., & Bullerjahn, G. S. (2020). Dissolved Microcystin Release Coincident with Lysis of a Bloom Dominated by Microcystis spp. in Western Lake Erie Attributed to a Novel Cyanophage. *Applied and Environmental Microbiology*, *86*(22). https://doi.org/10.1128/AEM.01397-20

Mei, M. L., & Danovaro, R. (2004). Virus production and life strategies in aquatic sediments. *Limnology and Oceanography*, *49*(2), 459–470. https://doi.org/https://doi.org/10.4319/lo.2004.49.2.0459

Miao, Y., Liu, F., Hou, T., & Liu, Y. (2022). Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics*, *38*(5), 1216–1222. https://doi.org/10.1093/bioinformatics/btab845

Middelboe, M. (2000). Bacterial Growth Rate and Marine Virus–Host Dynamics. *Microbial Ecology*, *40*(2), 114–124. https://doi.org/10.1007/s002480000050

Mohiuddin, M., & Schellhorn, H. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology*, *6*. https://www.frontiersin.org/articles/10.3389/fmicb.2015.00960

Mojica, K. D. A., & Brussaard, C. P. D. (2014). Factors affecting virus dynamics and microbial host-virus interactions in marine environments. *FEMS Microbiology Ecology*, 89(3), 495–515. Blackwell Publishing Ltd. https://doi.org/10.1111/1574-6941.12343

Mojica, K. D. A., Huisman, J., Wilhelm, S. W., & Brussaard, C. P. D. (2016). Latitudinal variation in virus-induced mortality of phytoplankton across the North Atlantic Ocean. *The ISME Journal*, *10*(2), 500–513. https://doi.org/10.1038/ismej.2015.130

Mukherjee, S., Stamatis, D., Li, C. T., Ovchinnikova, G., Bertsch, J., Sundaramurthi, J. C., Kandimalla, M., Nicolopoulos, P. A., Favognano, A., Chen, I.-M. A., Kyrpides, N. C., & Reddy, T. B. K. (2022). Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Research*, gkac974. https://doi.org/10.1093/nar/gkac974

Mühling, M., Fuller, N. J., Millard, A., Somerfield, P. J., Marie, D., Wilson, W. H., Scanlan, D. J., Post, A. F., Joint, I., & Mann, N. H. (2005). Genetic diversity of marine Synechococcus and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environmental Microbiology*, *7*(4), 499–508. https://doi.org/https://doi.org/10.1111/j.1462-2920.2005.00713.x

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews*, *75*(1), 14–49. https://doi.org/10.1128/MMBR.00028-10

Nissimov, J. I., Vandzura, R., Johns, C. T., Natale, F., Haramaty, L., & Bidle, K. D. (2018). Dynamics of transparent exopolymer particle production and aggregation during viral infection of the coccolithophore, *Emiliania huxleyi*. *Environmental Microbiology*, *20*(8), 2880–2897. https://doi.org/10.1111/1462-2920.14261

Österholm, J., Popin, R. v., Fewer, D. P., & Sivonen, K. (2020). Phylogenomic Analysis of Secondary Metabolism in the Toxic Cyanobacterial Genera Anabaena, Dolichospermum and Aphanizomenon. *Toxins*, *12*(4), 248. https://doi.org/10.3390/toxins12040248

Paerl, H. W., Gardner, W. S., McCarthy, M. J., Peierls, B. L., & Wilhelm, S. W. (2014). Algal blooms: Noteworthy nitrogen. *Science*, *346*(6206), 175. https://doi.org/10.1126/science.346.6206.175-a

Paerl, H. W., & Huisman, J. (2008). Climate: Blooms like it hot. *Science*, *320*(5872), 57–58. https://doi.org/10.1126/science.1155398

Paerl, H. W., & Otten, T. G. (2013). Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. *Microbial Ecology*, *65*(4), 995–1010. https://doi.org/10.1007/s00248-012-0159-y

Paez-Espino, D., Zhou, J., Roux, S., Nayfach, S., Pavlopoulos, G. A., Schulz, F., McMahon, K. D., Walsh, D., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Tringe, S. G., & Kyrpides, N. C. (2019). Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome*, *7*(1), 157. https://doi.org/10.1186/s40168-019-0768-5

Palermo, C. N., Fulthorpe, R. R., Saati, R., & Short, S. M. (2019). Metagenomic Analysis of Virus Diversity and Relative Abundance in a Eutrophic Freshwater Harbour. *Viruses*, *11*(9), 792. https://doi.org/10.3390/v11090792

Peterson, D., Bonham, K. S., Rowland, S., Pattanayak, C. W., Resonance, C., Klepac-Ceraj, V., Deoni, S. C. L., D'Sa, V., Bruchhage, M., Volpe, A., Beauchemin, J., Wallace, C., Rogers, J., Cano, R., Fernandes, J., Walsh, E., Rhodes, B., Huentelman, M., Lewis, C., … Braun, J. (2021). Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.670336

Pick, F. R. (2016). Blooming algae: A Canadian perspective on the rise of toxic cyanobacteria. *Canadian Journal of Fisheries and Aquatic Sciences*, *73*(7), 1149–1158. https://doi.org/10.1139/cjfas-2015-0470

Ridal, J., Brownlee, B., McKenna, G., & Levac, N. (2001). Removal of Taste and Odour Compounds by Conventional Granular Activated Carbon Filtration. *Water Quality Research Journal*, *36*(1), 43–54. https://doi.org/10.2166/wqrj.2001.003

Rohwer, F., & Edwards, R. (2002). The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. *Journal of Bacteriology*, *184*(16), 4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002

Roux, S., Páez-Espino, D., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T. B. K., Nayfach, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., & Kyrpides, N. C. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, *49*(D1), D764–D775. https://doi.org/10.1093/nar/gkaa946

Sanders, R., Henson, S. A., Koski, M., de La Rocha, C. L., Painter, S. C., Poulton, A. J., Riley, J., Salihoglu, B., Visser, A., Yool, A., Bellerby, R., & Martin, A. P. (2014). The Biological Carbon Pump in the North Atlantic. *Progress in Oceanography*, *129*, 200–218. https://doi.org/https://doi.org/10.1016/j.pocean.2014.05.005

Sanguino, L., Franqueville, L., Vogel, T. M., & Larose, C. (2015). Linking environmental prokaryotic viruses and their host through CRISPRs. *FEMS Microbiology Ecology*, *91*(5), fiv046. https://doi.org/10.1093/femsec/fiv046

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B., & Fuhrman, J. A. (2019). Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nature Communications*, *10*(1), 1169. https://doi.org/10.1038/s41467-019-09106-z

Śliwińska-Wilczewska, S., Maculewicz, J., Barreiro Felpeto, A., & Latała, A. (2018). Allelopathic and Bloom-Forming Picocyanobacteria in a Changing World. *Toxins*, *10*(1), 48. https://doi.org/10.3390/toxins10010048

Smith, R. B., Bass, B., Sawyer, D., Depew, D., & Watson, S. B. (2019). Estimating the economic costs of algal blooms in the Canadian Lake Erie Basin. *Harmful Algae*, *87*, 101624. https://doi.org/10.1016/j.hal.2019.101624

Sobhy H. (2018). Virophages and Their Interactions with Giant Viruses and Host Cells. *Proteomes*, *6*(2), 23. https://doi.org/10.3390/proteomes6020023

Srinivasan, R., & Sorial, G. A. (2011). Treatment of taste and odor causing compounds 2-methyl isoborneol and geosmin in drinking water: A critical review. *Journal of Environmental Sciences*, *23*(1), 1–13. https://doi.org/https://doi.org/10.1016/S1001-0742(10)60367-1

Steffen, M. M., Belisle, S. B., Watson, S. B., Gregory, B. L., Richard, B. A., & Wilhelm, S. W. (2015). Metatranscriptomic Evidence for Co-Occurring Top-Down and Bottom-Up Controls on Toxic Cyanobacterial Communities. *Applied and Environmental Microbiology*, *81*(9), 3268–3276. https://doi.org/10.1128/AEM.04101-14

Steffen, M. M., Davis, T. W., McKay, R. M. L., Bullerjahn, G. S., Krausfeldt, L. E., Stough, J. M. A., Neitzey, M. L., Gilbert, N. E., Boyer, G. L., Johengen, T. H., Gossiaux, D. C., Burtner, A. M., Palladino, D., Rowe, M. D., Dick, G. J., Meyer, K. A., Levy, S., Boone, B. E., Stumpf, R. P., … Wilhelm, S. W. (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*(12), 6745–6755. https://doi.org/10.1021/acs.est.7b00856

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., & Nikolski, M. (2022). VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Frontiers in Bioinformatics*, *2*. https://www.frontiersin.org/articles/10.3389/fbinf.2022.867111

Šulčius, S., Mazur-Marzec, H., Vitonytė, I., Kvederavičiūtė, K., Kuznecova, J., Šimoliūnas, E., & Holmfeldt, K. (2018). Insights into cyanophage-mediated dynamics of nodularin and other non-ribosomal peptides in Nodularia spumigena. *Harmful Algae*, *78*, 69–74. https://doi.org/10.1016/j.hal.2018.07.004

Sullivan, M. B., Weitz, J. S., & Wilhelm, S. (2017). Viral ecology comes of age. *Environmental Microbiology Reports*, *9*(1), 33–35. https://doi.org/10.1111/1758-2229.12504

Suttle, C. A. (2007). Marine viruses - Major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(10), 801–812. https://doi.org/10.1038/nrmicro1750

Tan, J., Fang, Z., Wu, S., Guo, Q., Jiang, X., & Zhu, H. (2022). HoPhage: an ab initio tool for identifying hosts of phage fragments from metaviromes. *Bioinformatics*, *38*(2), 543–545. https://doi.org/10.1093/bioinformatics/btab585

Tijdens, M., Hoogveld, H. L., Kamst-van Agterveld, M. P., Simis, S. G. H., Baudoux, A.-C., Laanbroek, H. J., & Gons†, H. J. (2008). Population Dynamics and Diversity of Viruses, Bacteria and Phytoplankton in a Shallow Eutrophic Lake. *Microbial Ecology*, *56*(1), 29–42. https://doi.org/10.1007/s00248-007-9321-3

Vardi, A., Haramaty, L., van Mooy, B. A. S., Fredricks, H. F., Kimmance, S. A., Larsen, A., & Bidle, K. D. (2012). Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19327–19332. https://doi.org/10.1073/pnas.1208895109

Villarroel, J., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., & Larsen, M. V. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses*, *8*(5), 116. https://doi.org/10.3390/v8050116

Vrede, K., Stensdotter, U., & Lindström, E. S. (2003). Viral and Bacterioplankton Dynamics in Two Lakes with Different Humic Contents. *Microbial Ecology*, *46*(4), 406–415. https://doi.org/10.1007/s00248-003-2009-4

Wang, S., Yang, Y., & Jing, J. (2022). A Synthesis of Viral Contribution to Marine Nitrogen Cycling. *Frontiers in Microbiology*, *13*. https://www.frontiersin.org/articles/10.3389/fmicb.2022.834581

Watson, S. B., Miller, C., Arhonditsis, G., Boyer, G. L., Carmichael, W., Charlton, M. N., Confesor, R., Depew, D. C., Höök, T. O., Ludsin, S. A., Matisoff, G., McElmurry, S. P., Murray, M. W., Peter Richards, R., Rao, Y. R., Steffen, M. M., & Wilhelm, S. W. (2016). The re-eutrophication of Lake Erie: Harmful algal blooms and hypoxia. *Harmful Algae*, *56*, 44–66. https://doi.org/10.1016/j.hal.2016.04.010

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, *28*(2), 127–181. https://doi.org/https://doi.org/10.1016/j.femsre.2003.08.001

Wilhelm, S. W., & Suttle, C. A. (1999). Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, *49*(10), 781–788. https://doi.org/10.2307/1313569

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., & Venter, J. C. (2008). The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLOS ONE*, *3*(1), e1456-. https://doi.org/10.1371/journal.pone.0001456

Wilson, W. H., Van Etten, J. L., & Allen, M. J. (2009). The Phycodnaviridae: the story of how tiny giants rule the world. *Current topics in microbiology and immunology*, *328*, 1–42. https://doi.org/10.1007/978-3-540-68618-7_1

Winter, J. G., Desellas, A. M., Fletcher, R., Heintsch, L., & Morley, A. (2011). Algal blooms in Ontario, Canada: Increases in reports since 1994. *Lake and Reservoir Management*, *27*(2), 107–114. https://doi.org/10.1080/07438141.2011.557765

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Xia, H., Li, T., Deng, F., & Hu, Z. (2013). Freshwater cyanophages. *Virologica Sinica*, *28*(5), 253–259. https://doi.org/10.1007/s12250-013-3370-1

Yang, Y., Wang, Y., Zhang, Z., Wang, W., Ren, X., Gao, Y., Liu, S., & Lee, X. (2018). Diurnal and Seasonal Variations of Thermal Stratification and Vertical Mixing in a Shallow Fresh Water Lake. *Journal of Meteorological Research*, *32*(2), 219–232. https://doi.org/10.1007/s13351-018-7099-5

Zimmerman, A. E., Howard-Varona, C., Needham, D. M., John, S. G., Worden, A. Z., Sullivan, M. B., Waldbauer, J. R., & Coleman, M. L. (2020). Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nature Reviews Microbiology*, *18*(1), 21–34. https://doi.org/10.1038/s41579-019-0270-x

Appendix

Supplementary Table 1. Relative Abundance of Cyanobacterial Genera in in Percent

Name	TL_JUL18	TL_MAY19	TL_JUN19	TL_JUL19	TL_AUG19	TL_JAN20
Synechococcus	83.057	40.000	80.619	43.802	50.845	23.596
Cyanobium	13.858	5.517	15.251	8.264	9.677	4.494
Microcystis	0.198	2.069	0.233	16.529	9.524	3.371
Nostoc	0.323	4.138	0.333	9.091	4.916	13.483
Synechocystis	0.115	1.379	0.100	0.826	3.072	1.124
Calothrix	0.198	10.345	0.433	3.306	2.688	4.494
Prochlorococcus	0.313	2.759	0.599	0.826	0.384	0.000
Gloeothece	0.031	0.690	0.033	0.826	2.151	1.124
Parasynechococcus	0.250	0.000	0.233	0.000	0.077	0.000
Rippkaea	0.000	0.690	0.000	0.000	1.152	1.124
Stanieria	0.042	2.069	0.100	0.000	0.998	1.124
Sphaerospermopsis	0.031	2.759	0.133	0.826	0.922	4.494
Leptolyngbya	0.125	2.069	0.200	0.826	0.922	3.371
Anabaena	0.042	2.069	0.033	1.653	0.845	3.371
Planktothrix	0.063	0.000	0.133	1.653	0.845	1.124
Tolypothrix	0.021	0.690	0.033	0.000	0.768	1.124
Geminocystis	0.104	0.690	0.167	2.479	0.691	2.247
Crocosphaera	0.021	0.000	0.000	0.826	0.768	0.000
Dolichospermum	0.031	0.690	0.033	2.479	0.691	0.000
Gloeobacter	0.094	0.690	0.000	0.000	0.154	3.371
Pseudanabaena	0.052	0.000	0.167	0.000	0.614	2.247
Acaryochloris	0.073	0.690	0.067	0.000	0.230	1.124
Fischerella	0.052	1.379	0.067	0.826	0.461	3.371
Oscillatoria	0.021	0.690	0.067	0.826	0.461	1.124
Thermosynechococcus	0.063	2.759	0.000	0.000	0.077	2.247
Scytonema	0.063	0.000	0.067	0.000	0.154	3.371
Richelia	0.021	0.690	0.067	0.000	0.384	0.000
Nodularia	0.042	0.690	0.033	0.000	0.384	0.000
Kovacikia	0.052	1.379	0.100	0.000	0.230	2.247
Thermoleptolyngbya	0.052	0.690	0.000	0.000	0.077	0.000
Cylindrospermopsis	0.042	0.690	0.000	0.000	0.154	0.000
Chondrocystis	0.010	0.690	0.033	0.826	0.307	0.000
Moorena	0.042	1.379	0.033	0.826	0.307	0.000

Name	TL_JUL18	TL_MAY19	TL_JUN19	TL_JUL19	TL_AUG19	TL_JAN20
Crinalium	0.021	0.690	0.067	0.000	0.307	0.000
Microcoleus	0.042	0.690	0.067	0.000	0.000	1.124
Halomicronema	0.042	0.690	0.000	0.000	0.077	0.000
Anabaenopsis	0.031	0.000	0.033	0.000	0.307	2.247
Leptodesmis	0.042	0.000	0.100	0.000	0.230	0.000
Euhalothece	0.000	0.000	0.000	0.000	0.307	0.000
Pleurocapsa	0.021	0.000	0.000	0.000	0.307	0.000
Trichormus	0.021	0.690	0.000	0.000	0.230	1.124
Rivularia	0.021	0.690	0.067	0.000	0.230	0.000
Chamaesiphon	0.031	0.000	0.033	0.000	0.077	0.000
Koinonema	0.031	0.000	0.033	0.000	0.077	0.000
Dactylococcopsis	0.010	0.000	0.000	0.000	0.230	0.000
Halothece	0.000	0.000	0.000	0.000	0.230	0.000
Allocoleopsis	0.010	0.000	0.000	0.000	0.230	0.000
Geitlerinema	0.031	0.000	0.000	0.000	0.154	0.000
Cylindrospermum	0.021	0.690	0.033	0.000	0.154	0.000
Brasilonema	0.021	1.379	0.033	0.000	0.077	1.124
Gloeocapsa	0.010	0.690	0.033	0.000	0.154	1.124
Cyanobacterium	0.010	1.379	0.033	0.000	0.154	2.247
Oxynema	0.010	0.000	0.033	1.653	0.077	0.000
Limnospira	0.021	0.000	0.000	0.000	0.154	0.000
Chroococcidiopsis	0.010	0.000	0.000	0.000	0.154	0.000
Anthocerotibacter	0.021	0.000	0.000	0.000	0.000	1.124
Candidatus Atelocyanobacterium	0.000	0.690	0.000	0.000	0.077	1.124
Leptothermofonsia	0.000	0.690	0.033	0.826	0.000	0.000
Thermostichus	0.010	0.000	0.033	0.000	0.077	0.000
Gloeomargarita	0.010	0.000	0.000	0.000	0.000	0.000

Supplementary Table 2. Auxiliary Metabolic Genes from Big Turkey Lake

TLW43			
Name	**Definition**	**Gene Count**	**Pathway**
speD, AMD1	S-adenosylmethionine decarboxylase [EC:4.1.1.50]	1	Amino acid metabolism
glnA, GLUL	glutamine synthetase [EC:6.3.1.2]	1	Amino acid metabolism
mhpC	2-hydroxy-6-oxonona-2,4-dienedioate hydrolase [EC:3.7.1.14]	1	Amino acid metabolism
DHFR, folA	dihydrofolate reductase [EC:1.5.1.3]	1	Biosynthesis of cofactors
pyrF	orotidine-5'-phosphate decarboxylase [EC:4.1.1.23]	1	Biosynthesis of cofactors
hemH, FECH	protoporphyrin/coproporphyrin ferrochelatase [EC:4.99.1.1 4.99.1.9]	1	Biosynthesis of cofactors
iscS, NFS1	cysteine desulfurase [EC:2.8.1.7]	1	Biosynthesis of cofactors
GAE, cap1J	UDP-glucuronate 4-epimerase [EC:5.1.3.6]	1	Biosynthesis of cofactors
rfbD, rmlD	dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133]	1	Biosynthesis of secondary metabolites
rfbA, rmlA, rffH	glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24]	1	Biosynthesis of secondary metabolites
trpC	indole-3-glycerol phosphate synthase [EC:4.1.1.48]	1	Biosynthesis of secondary metabolites
E2.2.1.6L, ilvB, ilvG, ilvI	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	1	Biosynthesis of secondary metabolites
rfbB, rmlB, rffG	dTDP-glucose 4,6-dehydratase [EC:4.2.1.46]	1	Biosynthesis of secondary metabolites
rfbC, rmlC	dTDP-4-dehydrorhamnose 3,5-epimerase [EC:5.1.3.13]	1	Biosynthesis of secondary metabolites

gmd, GMDS	GDPmannose 4,6-dehydratase [EC:4.2.1.47]	3	Carbohydrate Metabolism
galE, GALE	UDP-glucose 4-epimerase [EC:5.1.3.2]	1	Carbohydrate Metabolism
pgmB	beta-phosphoglucomutase [EC:5.4.2.6]	1	Carbohydrate Metabolism
TSTA3, fcl	GDP-L-fucose synthase [EC:1.1.1.271]	2	Carbohydrate Metabolism
glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16]	1	Carbohydrate Metabolism
mdh	malate dehydrogenase [EC:1.1.1.37]	1	Energy Metabolism
sucD	succinyl-CoA synthetase alpha subunit [EC:6.2.1.5]	1	Energy Metabolism
psbK	photosystem II PsbK protein	1	Energy Metabolism
ndhB	NAD(P)H-quinone oxidoreductase subunit 2 [EC:7.1.1.2]	1	Energy Metabolism
E1.7.1.7, guaC	GMP reductase [EC:1.7.1.7]	1	Nucleotide metabolism
thyA, TYMS	thymidylate synthase [EC:2.1.1.45]	1	Nucleotide metabolism
APRT, apt	adenine phosphoribosyltransferase [EC:2.4.2.7]	1	Nucleotide metabolism
E2.7.4.8, gmk	guanylate kinase [EC:2.7.4.8]	1	Nucleotide metabolism
dcd	dCTP deaminase [EC:3.5.4.13]	1	Nucleotide metabolism
dut, DUT	dUTP pyrophosphatase [EC:3.6.1.23]	1	Nucleotide metabolism
pyrG, CTPS	CTP synthase [EC:6.3.4.2]	1	Nucleotide metabolism
thyX, thy1	thymidylate synthase (FAD) [EC:2.1.1.148]	1	Nucleotide metabolism
TLW233			
rtpR	ribonucleoside-triphosphate reductase (thioredoxin) [EC:1.17.4.2]	1	Nucleotide Metabolism
dut, DUT	dUTP pyrophosphatase [EC:3.6.1.23]	1	Nucleotide Metabolism
E4.6.1.1	adenylate cyclase [EC:4.6.1.1]	1	Nucleotide Metabolism
thyX, thy1	thymidylate synthase (FAD) [EC:2.1.1.148]	1	Nucleotide Metabolism

HDDC3	guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase [EC:3.1.7.2]	1	Nucleotide Metabolism
iscU, nifU	nitrogen fixation protein NifU and related proteins	1	Energy Metabolism
erpA	iron-sulfur cluster insertion protein	1	Energy Metabolism
K23144	glucosamine-1-phosphate N-acetyltransferase	1	Biosynthesis of Secondary Metabolites
map	methionyl aminopeptidase [EC:3.4.11.18]	1	Amino Acid Metabolism
TLW278			
E1.7.1.7, guaC	GMP reductase [EC:1.7.1.7]	1	Nucleotide Metabolism
E1.17.4.1A, nrdA, nrdE	ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]	2	Nucleotide Metabolism
rtpR	ribonucleoside-triphosphate reductase (thioredoxin) [EC:1.17.4.2]	1	Nucleotide Metabolism
thyA, TYMS	thymidylate synthase [EC:2.1.1.45]	1	Nucleotide Metabolism
tdk, TK	thymidine kinase [EC:2.7.1.21]	2	Nucleotide Metabolism
thyX, thy1	thymidylate synthase (FAD) [EC:2.1.1.148]	2	Nucleotide Metabolism
acpP	acyl carrier protein	1	Biosynthesis of Secondary Metabolites
pdxA	4-hydroxythreonine-4-phosphate dehydrogenase [EC:1.1.1.262]	1	Biosynthesis of cofactors
gpt	xanthine phosphoribosyltransferase [EC:2.4.2.22]	1	Biosynthesis of cofactors
trpC	indole-3-glycerol phosphate synthase [EC:4.1.1.48]	1	Amino Acid Metabolism
speD, AMD1	S-adenosylmethionine decarboxylase [EC:4.1.1.50]	2	Amino Acid Metabolism
TLW323			
mdh	malate dehydrogenase [EC:1.1.1.37]	1	Energy Metabolism
sucD	succinyl-CoA synthetase alpha subunit [EC:6.2.1.5]	1	Energy Metabolism

galE, GALE	UDP-glucose 4-epimerase [EC:5.1.3.2]	1	Carbohydrate Metabolism
TLW368			
DNMT1, dcm	DNA (cytosine-5)-methyltransferase 1 [EC:2.1.1.37]	1	Amino Acid Metabolism
glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16]	1	Amino Acid Metabolism
queF	7-cyano-7-deazaguanine reductase [EC:1.7.1.13]	1	Biosynthesis of cofactors
TLW410			
E1.17.4.1A, nrdA, nrdE	ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]	2	Nucleotide Metabolism
E1.17.4.1B, nrdB, nrdF	ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]	1	Nucleotide Metabolism
comEB	dCMP deaminase [EC:3.5.4.12]	1	Nucleotide Metabolism
dut, DUT	dUTP pyrophosphatase [EC:3.6.1.23]	1	Nucleotide Metabolism
fabG, OAR1	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	1	Lipid Metabolism
ugtP	processive 1,2-diacylglycerol beta-glucosyltransferase [EC:2.4.1.315]	1	Lipid Metabolism
fabF, OXSM, CEM1	3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	1	Lipid Metabolism
glmS, GFPT	glutamine---fructose-6-phosphate transaminase (isomerizing) [EC:2.6.1.16]	1	Carbohydrate Metabolism
gmd, GMDS	GDPmannose 4,6-dehydratase [EC:4.2.1.47]	1	Carbohydrate Metabolism
TSTA3, fcl	GDP-L-fucose synthase [EC:1.1.1.271]	1	Carbohydrate Metabolism
UXS1, uxs	UDP-glucuronate decarboxylase [EC:4.1.1.35]	1	Carbohydrate Metabolism

acpP	acyl carrier protein	1	Biosynthesis of Secondary Metabolites
GCH1, folE	GTP cyclohydrolase IA [EC:3.5.4.16]	1	Biosynthesis of Cofactors
queD, ptpS, PTS	6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase [EC:4.2.3.12 4.1.2.50]	1	Biosynthesis of Cofactors
ribBA	3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II [EC:4.1.99.12 3.5.4.25]	1	Biosynthesis of Cofactors
dapB	4-hydroxy-tetrahydrodipicolinate reductase [EC:1.17.1.8]	1	Amino Acid Metabolism

Labels in yellow indicate AMG was already highlighted within the results section.