12-2017

Prediction of Host-Microbe Interactions from Community High-Throughput Sequencing Data

Joshua Michael Stough
University of Tennessee, jstough@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Joshua Michael Stough entitled "Prediction of Host-Microbe Interactions from Community High-Throughput Sequencing Data." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Microbiology.

Steven W. Wilhelm, Major Professor

We have read this dissertation and recommend its acceptance:

Jill A. Mikucki, Andrew D. Steen, Erik R. Zinser

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Prediction of Host-Microbe Interactions from Community High-Throughput Sequencing Data

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

Joshua Michael Albert Stough

December 2017

Acknowledgements

First, I would like to thank my advisor and mentor, Dr. Steven Wilhelm, for giving me the freedom to pursue research the way I enjoy it most, and for applying pressure in just the right ways to keep me moving forward in spite of failure or lack of progress.

To my committee, thank you for your advice and patience over the years. Finishing this document and this degree would not be possible without the contributions each of you has made during my time here.

I would also like to extend my thanks to the many members of our department who have made this all possible. To Dr. Mohammed Moniruzzaman, Eric Gann, and Sam Coy, whose expertise in giant viruses has always informed and advised the direction of my research. To Dr. Gary LeCleir for the training, patience, and support in the lab. To Chelsi Cassilly and Joseph Jackson for their friendship and moral support.

Lastly, I would like to thank the members of my family that have provided encouragement, advice, and care over the ten years I have spent in schools preparing for this. None of this would be possible without the care, support, and patience of my wife. Likewise, my parents have always encouraged me to follow this path, one to which I have aspired since the first grade. Thank you all for standing with me through the best and worst of it.

Abstract

Microbial ecology is a diverse field, with a broad range of taxa, habitats, and trophic structures studied. Many of the major areas of research were developed independently, each with their own unique methods and standards, and their own questions and focus. This has changed in recent decades with the widespread implementation of culture-independent techniques, which exploit mechanisms shared by all life, regardless of habitat. In particular, high-throughput sequencing of environmentally isolated DNA and RNA has done much to expand our knowledge of the planet's microbial diversity and has allowed us to explore the complex interplay between community members. Additionally, metatranscriptomic data can be used to parse relationships between individual members of the community, allowing researchers to propose hypotheses that can be tested in a laboratory or field setting. However, use of this technology is still relatively young, and there is a considerable need for broader consideration of its pitfalls, as well as the development of novel approaches that allow those without a computational background or with fewer resources to navigate its challenges and reap its rewards. To address these needs, we have developed targeted computational approaches that simplify next-generation sequencing datasets to a more manageable size, and we have used these techniques to address specific questions in environmental ecosystems. In a dataset sequenced for the purpose of identifying ecological factors that drive *Microcystis aeruginosa* to dominate cyanobacterial harmful algal blooms worldwide, we used a targeted approach to predict replication and lysogenic dormancy in bacteriophage. We used RNA-seq data to characterize viral diversity in the *Sphagnum* peat bog microbiome, identifying a wealth of novel viruses and proposing several host-virus pairs. We were able to assemble and describe the genome of a freshwater giant virus as well as that of a virophage that may infect it, and we used our techniques to describe its activity in publicly available datasets. Lastly, we have extended our efforts into the realm of medicine where we showed the influence exerted by the mouse gut microbiome on the host immune response to malaria, identifying several genes that may play a key role in reducing disease severity.

Table of Contents

List of Tables

Chapter II

Chapter III

Chapter V

List of Figures

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

CHAPTER I:

INTRODUCTION

Part I: Literature Review: Microbial Ecology in the Age of Next-Generation Sequencing and the ‘omics data revolution

Abstract

Microbial ecology is a diverse field, with a broad range of taxa, habitats, and trophic structures studied. Many of the major areas of research were developed independently, each with their own unique methods and standards, and their own questions and focus. This has begun to change in recent decades with the widespread implementation of culture-independent techniques, which exploit mechanisms shared by all life, regardless of habitat. In particular, high-throughput sequencing of environmentally isolated DNA and RNA has done much to expand our knowledge of the planet’s microbial diversity and has allowed us to explore the complex interplay between community members. Additionally, metatranscriptomic data can be used to parse relationships between individual members of the community, allowing researchers to propose hypotheses that can be tested in a laboratory or field setting. However, use of this technology is still relatively young, and there is a considerable need for broader consideration of its pitfalls, as well as the development of novel approaches that allow those without a computational background or fewer resources to navigate its challenges and reap its rewards. In this review, I place microbial ecology in a historical context and explore the development of modern sequencing technologies and analysis software as they relate to the limitations and challenges in the field. I also elaborate on recently developed concepts used in the application of these technologies to build mathematical models for the prediction of dynamics in microbial communities.

Introduction

Microorganisms represent the most abundant form of life on this planet by far, both in number and in overall biomass, dominated primarily by prokaryotes (Whitman *et al.*, 1998). As the first form of cellular life to evolve, prokaryotes have come to colonize and thrive, in one form or another, in every known habitat on Earth, from the human gut to environments with extremes in temperature, pH, salinity, and radiation (Rothschild & Mancinelli, 2001, Hatzenpichler, 2012, Kruger *et al.*, 2013, Stevenson *et al.*, 2015). Their success, abundance and longevity is due in large part to their ability to rapidly adapt to change, enabled by high mutation rates and frequent interspecies transfer of genetic information. Upon encountering unfavorable conditions, many can form protective coatings, spores, or reduce metabolic processes to such an extent that they can survive for decades (Driks, 1999, Jones & Lennon, 2010). In order to take advantage of nearly any form of nutrition available, microbes have evolved a diverse array of metabolic functions, including catabolism of dozens of distinct glycan molecules in the human gut (Koropatkin *et al.*, 2012), degradation of complex branching carbohydrates such as lignocellulose (Tuomela *et al.*, 2000), and remediation of contaminants through heavy metal reduction (Lovley, 1993, Valls & De Lorenzo, 2002, Anderson *et al.*, 2003). Even the seemingly more mundane and well-studied microbial functions, such as photosynthesis, still have a profound impact on the planet, oxygenating the earth's atmosphere in eons past (Pufahl & Hiatt, 2012) and sustaining primary production in the oceans today (Longhurst *et al.*, 1995). Understanding this impact on their environment, their interactions with other microbes, and the drivers and trajectories of their evolution forms the basis of microbial ecology.

While microbes were initially studied as a direct result of their impact on human health, a subset of researchers examined their contribution to biogeochemical processes in the environment, recognizing their contribution to communities of larger organisms. Researchers during the early 19th century identified microbial hydrogen and ammonia oxidation that could be slowed or stopped experimentally by increasing the temperature or lowering the pH (de Saussure, 1804, Schloesing & Muntz, 1877). However, the true importance of microbes in large-scale chemical transformations would not be appreciated until the work of Dr. Sergei Winogradsky. Unlike previous efforts showing general processes exclusively in soil, Winogradsky was able to modify pure culture techniques for the growth of soil organisms to point out individual bacteria responsible for biochemical transformations. This work led to the discovery and isolation of *Beggiatoa*, a sulfur-oxidizing soil bacterium which would serve as Winogradski's model organism for many years (Winogradsky, 1887). Subsequent studies resulted in the discovery of the phototrophic purple-sulfur bacteria (Winogradsky & Brock, 1889), the elucidation of the two-step nitrification process (Winogradsky, 1890), the demonstration of nitrogen-fixation by a *Clostridium* species (Winogradsky, 1928), and iron cycling in soil. However, Winogradsky's contribution to the burgeoning field of microbial ecology is best realized in a simple invention. A glass tube containing a column of soil and water with a natural consortium of resident microbes that form complex, stratified communities based on oxygen concentration and nutrient flow, the "*Winogradsky column*" represents several tenets that its inventor believed were fundamental to studying environmental microbes:

1.) Microbes are best described residing in their natural habitat, for which artificial media is a poor and disruptive substitute that favors the growth of a small subset of generalist microbes.

2.) While pure culture is an extremely valuable technique, axenic culture removes competitive pressure and interspecies interaction from the organism which will have a profound impact on its physiology.

3.) Community dynamics, while complex and difficult to parse, are fundamental to the overall function of the community.

These ideas provide insight into microorganisms as populations, members of communities, and moving parts of ecosystems, and they have formed the core of microbial ecology as study as transitioned from the soil ecosystems studied by Winogradsky to marine, freshwater, deep sea, polar, and host-associated environments.

Early studies of microbial ecology incorporated a mix of *in situ* methods, such as the Winogradsky column, pure culture *in vitro* methods that were developed in parallel with studies of pathogenic bacteria, and enrichments, which use added or restricted nutrients and incubation conditions to select for microbes with desirable traits. These methods were often combined with the use of stains and microscopy to visualize microbes based on physiological criteria, such as cell wall composition, viability, and content. The development of the electron microscope in the 1930s massively improved the capacity to detect cell structure and together, visualizations enabled by microscopy would form the basis for microbial classification and early attempts at phylogenetic reconstruction. Subsequent development of more sophisticated technologies, such as flow cytometry (Fulwyler, 1965), would lead not only to the novel organisms like *Prochlorococcus*, but also allow researchers to begin quantifying their presence in the oceans and infer their impact on the environment (Chisholm *et al.*, 1988). Researchers quickly recognized the limitations in relying on abundance alone and simultaneously developed techniques to quantify ecosystem activities. The development of electrode probing technologies allowed researchers to measure physiochemical characteristics of sampling environments, determining pH, O_2 and CO_2 concentrations, and salinity (Kerridge, 1925). Another such method involves seeding *in situ* microbial samples with known quantities of radioactive or stable isotopes and determining the quantity of isotope either ends up as microbial biomass or metabolic byproduct, enabling researchers to quantify biochemical rates and, in some cases, identify which organisms carry out the process (Boschker *et al.*, 1998, Radajewski *et al.*, 2000).

While these methods have proven very powerful over the decades, allowing many researchers to lay the foundation for microbial ecology through the isolation of novel organisms and the elucidation of major nutrient cycles, they are not without their limitations. Even with the more advanced forms of isolation and culturing, such as single cell sorting and high-throughput screening of culture media by robots, it is predicted that only a small fraction of the global microbial diversity has been successfully cultured (Stewart, 2012). Inference of microbial evolutionary histories through morphology and functional capability is challenging, as the traditional perception of the biological species concept does not apply to organisms that reproduce asexually (Mayr, 1982). This also limits the information obtained from enzyme assays and stable isotope probing, as only net activity across the whole community is determined rather than the contribution of each member.

Evolution of Molecular Biology in Ecology

Much of this began to change in the 70s with the advent of DNA sequencing. The recognition of DNA as the universal hereditary molecule by Avery, McCarty, and MacCleod (Avery *et al.*, 1944), which was confirmed in 1952 (Hershey & Chase, 1952), and then quickly followed by the discovery of the molecular structure of DNA by James Watson and Francis Crick (Watson & Crick, 1953) enabled future researchers to begin work developing the technologies that could determine and manipulate the genetic sequence of an individual organism. A significant step forward was made when researchers used chemically modified nucleotides to generate DNA sequences as long as 2000 bases (Sanger & Coulson, 1975). This began the first steps towards characterizing the genetic material of full genes and genomes. The first full genome to be sequenced was that of the bacteriophage φX174 (Sanger *et al.*, 1977). The original incarnation of this technology incorporated radioactive nucleotide labels, which were later replaced with fluorescent labels and full automation (Smith *et al.*, 1986, Prober *et al.*, 1987), enabling the sequencing of much larger genomes, starting with *Saccharomyces cerevisiae* (Feldmann *et al.*, 1994). The power of these technologies would spark the interest that would later start the Human Genome Project.

However, through these methods, sequencing genomic content required a concentrated and pure source of the target length of DNA, as any contaminant disrupts base detection and renders the sequence unreadable. To generate enough genetic material for sequencing, Sanger sheared the original bacteriophage genome into smaller pieces using restriction endonucleases, inserted the fragments into bacterial plasmids, and transformed them into *Escherichia coli*, forming a genomic clone library. The plasmid could then be amplified inside the bacterium and extracted for sequencing later. As this technology was developed to overcome the low concentration of target DNA, it would later be adapted for the study of uncultured environmental organisms (Handelsman *et al.*, 1998). Genomic DNA would be extracted from whole soil samples, sheared, and transformed into *E. coli*, which could then be sequenced or screened for the production of novel metabolites (Rondon *et al.*, 2000, Gillespie *et al.*, 2002, Knietsch *et al.*, 2003). Subsequent modifications of the plasmids used in cloning would result in the development of the Bacterial Artificial Chromosome, which allowed for insertion and transformation of much larger pieces of genetic material, as much as 350,000 base pairs, and enabled researchers to examine the genetic causes of human disease (O'Connor *et al.*, 1989, Shizuya *et al.*, 1992, Stone *et al.*, 1996, Shizuya & Kouros-Mehr, 2001). The idea of collecting the community level genetic material as a “metagenome” would form the basis for later culture-independent studies of environmental microbes.

Much of the sequencing and cloning work done today would not be possible without the development of the Polymerase Chain Reaction (PCR) by Kary Mullis (Saiki *et al.*, 1988). PCR is an *in vitro* technique that mimics the natural process of DNA replication in an ordered and predictable series of cycles. A DNA sequence of interest is targeted using a pair of RNA primers specifically coded to amplify the intended region, a biochemical process which is carried out by a high temperature-stable DNA polymerase and a collection of spare ATP molecules and single nucleotides. The process generates millions of copies of the target DNA, enabling researchers to target specific regions of DNA from pure culture and generate enough genetic material for sequencing. This is frequently used in medicine for diagnosis of infectious disease and genetic disorders such as Sickle Cell Anemia (Saiki *et al.*, 1985, Pozio & La Rosa, 2003, Sachse & Hotzel, 2003). In many cases, PCR is used to amplify a gene conserved across multiple microbial taxa where Sanger sequencing would not be appropriate. The differences in sequence between individuals in a community can be separated *via* gel electrophoresis to predict ecosystem diversity and generate community fingerprints (Fischer & Lerman, 1979, Fischer & Lerman, 1980, Muyzer *et al.*, 1993).

Just as more powerful and refined microscopy enabled researchers to attempt to reconstruct the phylogeny of bacterial species using morphology and function, the combination of PCR and Sanger sequencing allowed researchers to look to genetic sequences to infer evolutionary history. Prior to the advent of whole genome sequencing, taxonomic classification of novel organisms and phylogenetic inference were both determined by calculating the inherited traits shared between individuals, and since physiology of prokaryotes is so radically distinct from eukaryotes, the two were not comparable by conventional means of taxonomy (Sneath & Sokal, 1973). However, as DNA is the material passed on from parent to progeny, regardless of sexual or asexual behavior, it stands as a strong candidate for inference of evolutionary relationships. In order to determine the phylogenetic relationships between all of the major branches of life, Carl Woese and George Fox constructed the first tree of life using the nucleotide sequence of the ribosomal small subunit RNA molecule (Woese & Fox, 1977). This gene proved an ideal candidate for the task, as it was highly conserved across all forms of cellular life, short enough to be Sanger sequenced, but long enough to provide large-scale phylogenetic resolution across major taxa. The result was the first full tree of life to be calculated from real biological data, which grouped all cellular organisms into three distinct domains: Eukaryota, Bacteria, and Archaea. Beyond its use in the full tree of life, the 16S small subunit ribosomal RNA molecule in prokaryotes would later be employed as a universal marker gene for their respective domains in community level studies by using PCR primers that hypothetically amplified any rRNA molecule in the sample (Eden *et al.*, 1991, Jiang *et al.*, 2006). Amplicons from these primer sets could then be inserted into plasmids to form clone libraries that contain information about microbial diversity and abundance from the amplified sample (Kowalchuk *et al.*, 1997, Rondon *et al.*, 2000, Manichanh *et al.*, 2006).

Currently Available Molecular Technologies

Sanger sequencing proved to be a very powerful technique, especially when the need for a high concentration of a single sequence was made much easier with PCR, allowing researchers to quickly determine a sequence from pure culture. However, even with the benefits of PCR and full automation, the process was slow and sequencing full genomes required extensive work mapping the genomic space before sequencing could begin. Its use in community level 16S rRNA sequencing and metagenomics was limited as well, as costs could be prohibitively expensive for hundreds of representatives. Indeed, early work by national laboratories on the Human Genome Project was quickly surpassed by privately funded labs using the recently developed paired-end shotgun whole genome sequencing (Roach *et al.*, 1995, Lander *et al.*, 2001). This approach involved shearing the extracted genomic DNA and cloning fragments into plasmids and yeast artificial chromosomes (Hsiao & Carbon, 1979), followed by sequencing and assembly *via* computer algorithm (Venter *et al.*, 2001). Despite major concerns that computational assembly would introduce too much potential for error in the output sequences, the approach rapidly became the new standard as the assembly algorithms improved (Myers *et al.*, 2000, Batzoglou *et al.*, 2002).

This success would be further bolstered by the development of high-throughput short read sequencing technologies, such as pyrosequencing, which determined DNA sequences as they were synthesized by pyrophosphate reactions (Nyren *et al.*, 1993, Ronaghi *et al.*, 1998, Margulies *et al.*, 2005). These new methods allowed researchers to sequence entire bacterial genomes in a single methodological run, whereas Sanger sequencing could require dozens of separate reactions preceded by the long process of whole genome mapping. They also revolutionized the process of metagenomic sequencing, as both 16S rRNA amplicon libraries and shotgun metagenomes could be sequenced directly without the need for one sequence per run, and eliminating the need for clone libraries in community-level metagenomics (Brulc *et al.*, 2009, Fierer *et al.*, 2012). While pyrosequencing is still in use, it has largely been replaced by Illumina™ dye sequencing, which operates under similar principles as Sanger sequencing by incorporating color labeled terminal nucleotides that are read by a laser-equipped computer, only on a much larger, 2-dimensional scale rather than in microfluidics (Canard & Sarfati, 1994). This modern iteration of high-throughput sequencing (often termed “next-generation sequencing”, regardless of actual generation) generates reads that are considerably shorter than Sanger’s method, but in such high volume that the limited length is hypothetically overcome. The sheer volume of information obtained from Illumina™ sequencing is often enough to cover the whole genomes of larger organisms in a single run over a few days, and has allowed some researchers to begin probing environmental ecosystems for the rarest community members (Caporaso *et al.*, 2011, Schloss *et al.*, 2011, Segata *et al.*, 2011).

In parallel with the development of high-throughput sequencing of metagenomic and community-level 16S rDNA libraries, sequencing of mRNA transcripts and small RNAs has seen a burst in activity with the widespread adoption of Illumina™ technology. As RNA is not sequenced directly, it was not possible until the discovery of the enzyme reverse-transcriptase from retroviruses (Baltimore, 1970, Temin & Mizutani, 1970). This enzyme transcribes an RNA template into DNA, and when adapted for use in molecular biology, allowed for the development of reverse transcriptase quantitative PCR, microarray technology, and RNAseq (Adams *et al.*, 1991, Freeman *et al.*, 1999, Wang *et al.*, 2009). However, it was not until the invention of pyrosequencing that RNAseq yielded enough reads to be considered reliable, allowing researchers to begin determining the transcriptional regulation patterns of whole populations of individual organisms. Still relatively young, the field of metatranscriptomics, sequencing transcripts from a complex microbial community, has yet to fully mature to incorporate the standards and practices that have been developed for metagenomic and 16S sequencing datasets (Schloss & Handelsman, 2005, Schloss *et al.*, 2011). However, as transcripts are very unstable *in vivo* and represent at least a precursor to true cellular activity, RNAseq data obtained from microbial communities can potentially ignore organisms that are abundant but dormant and do not contribute to community activity (Bernstein *et al.*, 2002).

Where metagenomes represent the metabolic and functional potential of a community, and the metatranscriptomes represent the transcriptional regulation and potential activity of a community, metaproteomes seek to describe the abundances of different proteins within a sample. In this technique, proteins are extracted from a sample and separated *via* 2D gel electrophoresis, after which they are sequenced and characterized using mass spectrometry (Anderson & Anderson, 1998, Blackstock & Weir, 1999). While this technology has the potential to more directly address the true activity of microbes in a community, limitations in the available tools and caveats in data interpretation hold it back. Proteomic methodologies often favor proteins in higher abundance, and proteins produced by the rare microbiota are frequently lost in the analysis (Wang *et al.*, 2016). In addition, the reduced phylogenetic resolution in amino acid sequences limits researchers' ability to distinguish between highly conserved proteins. Interpretation of protein abundance as potential enzymatic activity can often be misleading, as different proteins can have radically variable half-lives depending on the cell state and requirements for post-translational modifications influence active and inactive protein states (Petrov *et al.*, 2013). That being said, proteomes of individual organisms, often paired with genomic data, have done much to advance understanding of cell physiology (Washburn *et al.*, 2001, Anderson & Anderson, 2002, Rual *et al.*, 2005), and future developments in the technology and approaches have the potential to revolutionize the study of complex microbial communities.

Just as chemists have employed mass spectrometry in the identification of proteins in a given culture or environmental sample, others have used a similar method to characterize and quantify the total available pools of nutrients and metabolic products, called the metabolome. The metabolic profile can be used to gain insight into ongoing metabolic processes inside a pure culture or a complex community. As with proteomes, this method suffers from difficulty in data interpretation, as it can be difficult to distinguish whether observed metabolite abundances are due to changes in influx or efflux. These studies appear to work best when used in time course experiments that track metabolites during the course of the cell cycle, usually in combination with genomic or transcript sequencing (Oliver *et al.*, 1998, Tweeddale *et al.*, 1998, Fiehn, 2001, Fiehn *et al.*, 2001, Goodacre *et al.*, 2004). Used together, the collection of metagenomics, metatranscriptomics, proteomics, and metabolomics form a unique field within cell biology known as systems biology, which models the abundance of genes, transcripts, proteins, and metabolites as part of a complete system in an attempt to understand how it functions.

Evolution of Information Technologies

The discovery of the structure of DNA and the subsequent development of sequencing technology produced a new series of challenges to researchers interested in genetics. While researchers would eventually determine the nature of codons and the genetic code in translation (Crick *et al.*, 1961, Gardner *et al.*, 1962, Wahba *et al.*, 1963), and the use of DNA sequence in phylogenetic inference is described above, analysis of sequences by hand was challenging and time-consuming. The field of genetics quickly resorted to modern computers to assist in large-scale sequence analysis, leading to the development of a number of software tools, the most important of which in microbial ecology are covered here.

Determining the similarity or relatedness of two DNA or amino acid sequences requires that they be aligned, identifying a corresponding start point between them and aligning shared nucleotide bases or amino acids, marking gaps or polymorphisms where they appear. The first quantitative attempt to accomplish this was proposed by Needleman and Wunsch in 1970 (Needleman & Wunsch, 1970). This method for global alignment of two sequences was quickly replaced by a more sensitive and general method for local alignment, incorporating calculations for gaps (Smith & Waterman, 1981). However, both of these methods calculate optimal alignments between two sequences, and as such were very slow and computationally intensive for the time, making larger scale alignments against databases of reference sequences impractical. In order to address this problem, other researchers designed a compromise; an algorithm that operated under a significantly reduced number of steps, and thus much faster, but that produced less than optimal alignments and was prone to false positives. This was referred to as the Basic Local Alignment Search Tool (BLAST), which quickly became a staple of modern bioinformatics, now forming the base of the National Center for Biotechnology Information (NCBI) (Altschul *et al.*, 1990). The simplicity and speed of this tool has led many to rely on it exclusively for homology determination of unknown sequences, but as it was only designed for sequence alignment based on similarity, homology can often be inferred incorrectly. While more reliable alignments can be calculated for smaller collections of sequences using software like MUSCLE (Edgar, 2004), the need for fast alignments to probe the rapidly growing sequence databases makes them impractical. One method rising in popularity is the Hidden Markov Model algorithm implemented in the HMMER software package (Finn *et al.*, 2015), which builds alignments and detects homology by comparing scores from query sequences with a null model. The result is a more reliable homology assignment that is considerably quicker than the Smith & Waterman method, and which has been employed for searches on both the Pfam and InterPro sequence databases (Finn *et al.*, 2016, Finn *et al.*, 2017).

Another key piece of software developed soon after sequencing was the sequence assembler, which first reached widespread use with the development of pairwise shotgun genomic sequencing mentioned above (Myers *et al.*, 2000, Venter *et al.*, 2001). These algorithms take sequencing reads and align them together to form longer contiguous sequences, also known as contigs, based on overlap, usually by constructing a DeBrujin graph to score alignments and optimize contig building (Good, 1946). The most recent iterations of these algorithms are designed primarily to deal with one of the two major problems with traditional assembly: 1.) reduce the size, complexity, or length of time necessary to assemble sequences, or 2.) improve the quality of assembly by compensating for assembler's tendency to assemble reads that should not be connected. These incorrect assemblies, known as chimeras, are especially problematic in metagenomic sequencing datasets where multiple organisms possess genes with highly conserved regions, leading to frequent chimeric assembly. Recent assemblers, such as SPAdes and SOAPdenovo, have been developed with quality control checks to account for single nucleotide polymorphisms (SNPs) that can confound assemblies and contig quality checks to reduce the frequency of chimeric assemblies (Bankevich *et al.*, 2012, Xie *et al.*, 2014).

The use of high-throughput sequencing to generate large 16S and 18S rRNA datasets has proven to be a remarkably useful tool for characterizing the diversity of microbial communities and the abundance of individual taxa. However, the sheer number of reads generated from sequencing runs, in addition to some methodological concerns has posed a number of challenges for researchers using this method. As the 16S rRNA molecule is conserved amongst all bacteria and exhibits high similarity across even the most distantly related taxa, phylogenetic identification of each individual sequence is not necessary when many sequences are almost entirely identical. This has led to the development of multiple computational algorithms, such as mothur and QIIME, that cluster rRNA reads into Operational Taxonomic Units (OTUs) based on similarity (Sneath & Sokal, 1973), picking a representative sequence from the bin for identification (Schloss *et al.*, 2009, Caporaso *et al.*, 2010). However, since the determination of similarity needed to distinguish between different clusters is largely up to the user, this has brought into focus the scale of similarity needed to distinguish between different taxa, all the way down to individual species and strains (Gevers *et al.*, 2005, Doolittle & Papke, 2006, Koeppel *et al.*, 2008). A recent survey examining OTU clustering across multiple environments determined that OTUs are largely consistent, regardless of ecosystem, and that the method had a much greater impact on results (Schmidt *et al.*, 2014). Today, improvements in DNA extraction and sequencing depth have allowed researchers to begin to explore the diversity and activity of the rare microbiome, low-abundance microbes in the community that appear to have a profound impact on overall ecosystem function (D Ainsworth *et al.*, 2015, Jousset *et al.*, 2017).

Even with the speed and capacity of modern computers and development of advanced algorithms that improve the computational efficiency of bioinformatic tasks, the sheer amount of sequencing data produced per unit cost has grown exponentially, whereas progress in the cost and capability in computing has remained largely the same (Figure 1.1). Some researchers have resorted to high-performance computing, or supercomputing, which massively scales up conventional computing processes to run dozens of processes simultaneously. However, supercomputers can be prohibitively expensive to both purchase and maintain, and thus they are often shared across a broad user base, limiting the time and resources each individual user has access to. These tools often also require a significant degree of skill in navigating complex computational tasks, in which biologists are not often trained. The need for user-friendly, but

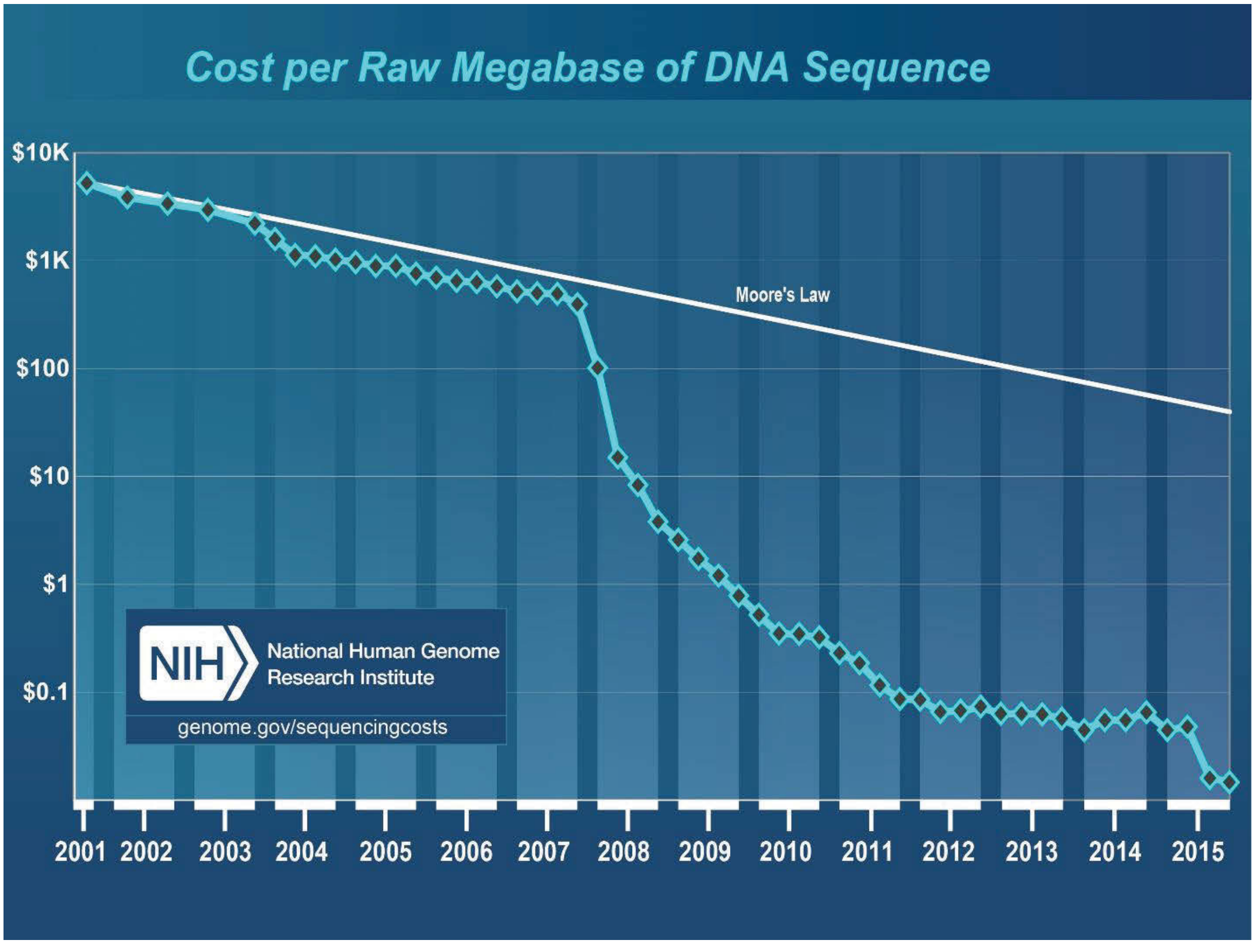

Figure 1.1. Cost of Sequencing versus Moore’s Law. The cost of determining one megabase (Mb; a million bases) of DNA sequence of a specified quality versus the hypothetical data reflecting Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years. Technology improvements that 'keep up' with Moore's Law are widely regarded to be doing exceedingly well, making it useful for comparison (Wetterstrand K. Data available at www.genome.gov/sequencingcostsdata).

powerful, computational resources have led to the development of server-based software tools, such as the Galaxy portal, Rapid Annotations using Subsystems Technology (RAST), and metagenomics RAST (MG-RAST). RAST in particular enables users to annotate entire genomes over the course of 1-2 days without sacrificing accuracy to save computational power (Aziz *et al.*, 2008). MG-RAST is an adaptation of the RAST server-based model of gene annotation applied to focus on short read shotgun metagenomic sequencing datasets, and has recently been reworked to allow analysis of short read metatranscriptome datasets as well (Meyer *et al.*, 2008).

Further adaptations of this model also exist for more specific purposes, such as VIROME and MetaVir. As the sequences of viruses are often radically different from cellular organisms as described below, and since they are generally poorly represented in reference databases, the server-based tool VIROME queries short read shotgun metagenomic sequencing datasets against a specialized database using predicted open reading frame translations to reduce the sequence divergence often observed in viral nucleotide sequences (Wommack *et al.*, 2012). MetaVir, on the other hand analyses contig sequences submitted by the user to identify viruses from much longer and more informative stretches of genetic material. MetaVir examines the orientation of open reading frames and kmer frequencies to evaluate the quality of viral metagenome contigs, identify viral sequences, and compare abundances between libraries (Roux *et al.*, 2014).

All of these technologies have been used in one form or another to advance the field of microbial ecology by discovering novel microbes, characterizing the functional potential of old ones, and by describing community diversity and expression patterns. Additionally, the combination of multiple techniques can be used to describe individual organisms and ecosystems in a more holistic context, allowing researchers to gain a more complete picture as to how microbial communities function. However, it must be noted that these methods are ultimately tools whose primary benefit in describing microbial communities is to propose novel hypotheses and make predictions that can be subsequently tested in laboratory and field settings.

Part II: Linking Sequencing data to Microbial interactions

Life after sequencing: how do we interpret results from 'omics data?

The techniques and technologies for determining the functional potential, transcriptional regulation, protein abundance, and metabolite pools within microbial communities, holistically termed Systems Biology, has caused radical shifts in the way the ecology of microbes is studied. The individual fields of microbial ecology, generally separated by environment, have seen traditional techniques that were originally developed for the unique requirements of their ecosystems either replaced or augmented by systems biology approaches, breaking down the boundaries between disciplines. The universality of the central dogma of biology across all life has allowed researchers to apply analyses that bridge the gaps between ecosystems and explore the ecological factors that govern all microbial communities. However, while researchers can generate molecular sequencing data very quickly and cheaply, and many tools exist to process the information such that analysis is possible, best practices in interpretation and application of the results are an open question. The first wave of studies incorporating large sequencing datasets in unexplored microbial communities generally focus on the low-hanging fruit, namely description of community composition and functional potential. In many cases this is a necessary step to lay a strong foundation for future research and has led to important discoveries, including the SAR11 clade, one of the most abundant bacterial populations in the global oceans (Morris *et al.*, 2002). It should be noted, as mentioned above, that ultimately the methods mentioned here are tools built for the purpose of proposing new hypotheses describing the facets of microbial ecosystems, and thus descriptive studies must serve to address relevant ecological questions. The purpose of this review is to discuss applications of systems biology approaches, primarily focusing on 16S, metagenomics, and metatranscriptomics datasets, to the study of host-microbe interactions.

To move beyond cataloguing microbial community members and into the realm of predicting relationships between them, there is a need to define their activity using molecular sequencing so that they can be compared between organisms, samples, and datasets. As sequencing datasets are generated from PCR amplification reactions and sequences are read from within lanes on a sequencing flow cell with a maximum red capacity, abundance of genes or transcripts within a dataset must be treated as relative, where each represents a fraction of the total number generated (Margulies *et al.*, 2005). As a result, sequencing depth has become a primary concern in all applications because highly abundant oligonucleotides in the sample can dominate results, obscuring unique patterns in abundance or expression (Hewson *et al.*, 2009). In general, this problem limits comparison of organism or transcript abundance only between samples within a sequencing run, so long as data is normalized to account for differences in sequencing depth. One proposed solution is to introduce internal standards to sequencing that improve the reproducibility and allow researchers to estimate absolute abundance. Building a mock community sample containing 16S rRNA genes at known concentration has enabled researchers to monitor and quantify the impact of amplification to output sequencing datasets (Parada *et al.*, 2016). A similar approach in metatranscriptomics involved the addition of a plasmid expressing a gene at known quantities within samples, which could then be measured and used to calculate the absolute abundance of the environmentally isolated transcripts (Gifford *et al.*, 2011). While these methods have yet to be applied broadly across microbial community analyses, they pose important considerations in the interpretation of data, which must be considered when attempting to elucidate relationships between microbes.

Network analysis: statistical and mathematical prediction of host-microbe interactions

Assuming organismal or transcript abundance can be determined, how can this information be used to predict relationships? One potential route is co-occurrence, which stands as an ecologically relevant set of patterns that reflects coexistence and diversity maintenance within a community (HilleRisLambers *et al.*, 2012). Studies of co-occurrence in microbial communities can be applied in multiple ways, including early presence-absence studies, which have been used to determine whether population distribution is random or subject to species interaction (Stone & Roberts, 1990). The developments in high-throughput sequencing methods allow researchers to apply co-occurrence theory using correlation coefficients to predict coexistence and competitive exclusion (Kittelmann *et al.*, 2013). Correlations have similarly been used to develop network analyses, drawing maps that connect different microbes based on symbiotic or predatory relationships (Fuhrman & Steele, 2008, Williams *et al.*, 2014). More refined implementations of these methods have incorporated mathematical modeling elements to account for delayed responses to changes between organismal abundances (Parada & Fuhrman, 2017). However, every case mentioned here studying microbial communities relies on metagenomes, and compares the relative abundance of genes and organisms, rather than activity. While this approach can potentially yield meaningful microbial relationships, microbial activity represents a more realistic predictor of physiological response. A recent study incorporated time course metatranscriptomes to predict virus host pairs in marine systems with significant success, posing a method that may better represent the innate physiology for such analyses (Moniruzzaman *et al.*, 2017).

Lysogeny: prediction of alternative viral survival strategy

One of the confounding factors representing a major gap in attempts to model viral dynamics is lysogeny, where an infecting virus eschews replication and lysis of the host cell in favor or integrating into the host genetic material and remaining dormant, called a prophage. An intact prophage replicates along with the host cell, providing protection from infection by other lytic phage in order to ensure its own survival. The components that facilitate lysogeny and dictate the decision to integrate, rather than lyse the host, have been studied primarily in the Lambda phage, which infects *Escherichia coli* (Oppenheim *et al.*, 2005). Additionally, a recent study was able to determine population-level peptide communication in *Bacillus* SPbeta group phage that regulates the lytic-lysogenic decision (Erez *et al.*, 2017). However, outside of these well studied systems, lysogeny is poorly understood despite mounting evidence that large portions of environmental and host-associated bacteria are lysogenized (Beres & Musser, 2007, McDaniel *et al.*, 2008, Waller *et al.*, 2012, Waller *et al.*, 2014). This is in part due to the difficulty inherent to the study of lysogeny, which is compounded by the incredible diversity in phage populations. We believe that the answer to some of these questions lies in metatranscriptomics data, as expression of prophage genes is necessary to maintain a stable lysogenic relationship and protect against superinfection (Kourilsky, 1975, Abedon, 1992, Abedon, 1999). In addition, viral transcripts are strong evidence of activity, as free virus particles possess no metabolism and do not transcribe genes. Further elucidation of lytic-lysogenic decision making strategies shared amongst broad environmental taxa will aid in future model development.

Chrysoschromulina parva Virus: a case study in culture-based viral ecology

While modeling and statistical methods for linking viruses to hosts have proven useful, ultimately all relationships predicted will require experimental testing, either *in situ* or culture-based, to reject or refine the proposed hypotheses. One specific example of this concept in action is the recently discovered giant virus infecting the freshwater haptophyte *Chrysochromulina parva* (CpV), the genome of which is presented in the following chapters (Mirza *et al.*, 2015). Until recently, giant viruses have been frequently observed in electron micrographs, but were assumed to be bacteria due to their size (Wilhelm *et al.*, 2016, Wilhelm *et al.*, 2017). The discovery of Mimivirus infecting *Acanthamoeba*, and the close relatives discovered since, show that giant viruses are radically different from the conventional model of viruses, possessing hundreds of genes, many of which are responsible for functions previously only found in cellular life (Filee *et al.*, 2008), including translational machinery and auxiliary metabolic functions. As giant viruses have been implicated both in human disease (Popgeorgiev *et al.*, 2013, Yolken *et al.*, 2014) and the collapse of harmful algal blooms (Schroeder *et al.*, 2003, Gastrich *et al.*, 2004), there is a desperate need for further expansion of known physiology and diversity. As the first freshwater representative of the algal Mimiviruses to be isolated and maintained in culture, CpV stands as an important model virus for the future study of Mimiviruses ecology and physiology in freshwater ecosystems.

Conclusion

The following chapters present the application of high-throughput transcript sequencing and statistical approaches to predict relationships between microbes, viruses, and their respective hosts in a broad array of environmental ecosystems. While environments, microbial community constituents, represented taxa, and specific conclusions between chapters are largely unrelated, the work described here is linked by the following goals to:

1.) develop targeted approaches that allow researchers to reduce the size and complexity of the notoriously large high-throughput sequencing datasets in order to propose hypotheses that can be tested in a field or laboratory setting and,

2.) discover or predict the relationships between microbes and their hosts, with a particular focus on viruses.

Within this body of work, we discovered transcriptional patterns, consistent across temporal and geographic scales, that suggested rampant bacteriophage lysogeny occurs during *Microcystis aeruginosa* blooms in the Chinese hypereutrophic Lake Tai. As viral mRNA molecules for DNA viruses are only produced during active infections, lytic or lysogenic, the data shown here represents strong evidence for a relationship between virus and host. The results we obtained further suggest a series of viral expression markers that could be used to further predict lytic and lysogenic activity in *Microcystis* phage. As lysogeny often protects the host from subsequent infection by other lytic phage, these observations may provide an explanation for *Microcystis* success as a bloom former and its ability to defy Hutchinson’s paradox of the plankton and the kill-the-winner-hypothesis (Hutchinson, 1961, Thingstad & Lignell, 1997). We used the same technology to explore the viral diversity and activity in the microbiome of *Sphagnum* peat bog environments. We applied a pipeline in development to characterize viruses from multiple taxa and we were able to identify a broad diversity of both DNA and RNA viruses, and predict the hosts of many of the identified viruses. We also sequenced, assembled, and annotated the genomes of the *Chrysochromulina parva* Virus and its virophage, in which we observed a versatile giant virus clearly originating in an emerging clade within the NCLDV group. Subsequent examination of CpV activity using transcript sequences from freshwater ecosystems revealed significant activity in Lake Tai, China during the *Microcystis* bloom in 2013 and a strong correlation between virus and virophage expression. Lastly, to determine the contribution of the gut microbiome to malaria resistance in mice, we isolated and sequenced the bacterial community and mouse metatranscriptomes. During our analysis, we were able to identify multiple genes potentially involved in the interface between gut microbes and their host that may contribute to resistence to infection by *Plasmodium*. Altogether this body of work establishes a collection of powerful methods for targeting specific organisms and activities in diverse microbial ecosystems, and proposes hypotheses that advance the understanding of the environments studied herein.

CHAPTER II:

MOLECULAR PREDICTION OF LYTIC *VS* LYSOGENIC STATE FOR *MICROCYSTIS* PHAGE: METATRANSCRIPTOMIC EVIDENCE OF LYSOGENY DURING LARGE BLOOM EVENTS

Publication Note

This chapter is a version of a peer-reviewed article previously published in PLoS ONE 12(9): e0184146 by Joshua M.A. Stough, Xiangming Tang, Lauren E. Krausfeldt, Morgan M. Steffen, Guang Gao, Gregory L. Boyer, and Steven W. Wilhelm.

My contribution to this work was the experimental conceptualization, data processing and analysis, and primary authorship and editing of the manuscript.

Abstract

Microcystis aeruginosa is a freshwater bloom-forming cyanobacterium capable of producing the potent hepatotoxin, microcystin. Despite increased interest in this organism, little is known about the viruses that infect it and drive nutrient mobilization and transfer of genetic material between organisms. The genomic complement of sequenced phage suggests these viruses are capable of integrating into the host genome, though this activity has not been observed in the laboratory. While analyzing RNA-sequence data obtained from *Microcystis* blooms in Lake Tai (*Taihu*, China), we observed that a series of lysogeny-associated genes were highly expressed when genes involved in lytic infection were down-regulated. This pattern was consistent, though not always statistically significant, across multiple spatial and temporally distinct samples. For example, samples from Lake Tai (2014) showed a predominance of lytic virus activity from late July through October, while genes associated with lysogeny were strongly expressed in the early months (June – July) and toward the end of bloom season (October). Analyses of whole phage genome expression shows that transcription patterns are shared across sampling locations and that genes consistently clustered by co-expression into lytic and lysogenic groups. Expression of lytic-cycle associated genes was positively correlated to total dissolved nitrogen, ammonium concentration, and salinity. Lysogeny-associated gene expression was positively correlated with pH and total dissolved phosphorous. Our results suggest that lysogeny may be prevalent in *Microcystis* blooms and support the hypothesis that environmental conditions drive switching between temperate and lytic life cycles during bloom proliferation.

Introduction

Viruses are one of the most potent drivers of nutrient cycles, horizontal gene transfer, and microbial evolution in aquatic ecosystems (Brussaard *et al.*, 2008, Weitz & Wilhelm, 2012).

Bacteriophage play an important role in microbial communities by lysing primary producers and heterotrophic bacteria, releasing nutrients from biomass (Wilhelm & Suttle, 1999). Moreover, due to their density-dependent infection, viruses are thought to reduce the competitive advantages of some of the most prolific organisms – the "*kill-the-winner*" hypothesis (Thingstad & Lignell, 1997). Phage genomes also can encode auxiliary metabolic genes that serve to augment host metabolism during infection, considerably altering the functional potential of entire populations within the microbial community (Thompson *et al.*, 2011, Roux *et al.*, 2016). Despite their recognized importance, much of the potential of viruses remains uncharacterized, highlighting a crucial need for examination of the role they play across ecosystems.

Microcystis aeruginosa has repeatedly been identified as a nuisance bloom-former in freshwater systems over the last several decades (Harke *et al.*, 2016). It has come to the forefront of public attention as the primary agent in blooms worldwide and for its ability to produce a potent hepatotoxin, originally known as "Fast-Death Factor" (Bishop *et al.*, 1959), but now known as microcystin (Carmichael, 1996, Brittain *et al.*, 2000). Recent impacts include the shutdown of the public water supply to the City of Toledo (Ohio) during the *Microcystis* bloom in 2014 (Steffen *et al.*, 2017), and the considerable accumulation of toxic algal biomass in Lake Tai, China (*Taihu* in Chinese) (Qin *et al.*, 2010, Krausfeldt *et al.*, 2017). While significant strides have been made describing the ecology (Brunberg, 1999, Steffen *et al.*, 2012, Steffen *et al.*, 2015), physiology (Kromkamp *et al.*, 1988, Shen *et al.*, 2011, Harke *et al.*, 2017), and genetics (Kaneko *et al.*, 2007, Steffen *et al.*, 2014, Yamaguchi *et al.*, 2015) of *Microcystis*, little is known about the effect of phage on *Microcystis* ecology. To date, only 11 viruses infecting *M. aeruginosa* have ever been brought into culture (Tucker & Pollard, 2005, Yoshida *et al.*, 2006, Hargreaves *et al.*, 2013, Ou *et al.*, 2013, Watkins *et al.*, 2014, Mankiewicz-Boczek *et al.*, 2016), of which only 2 have sequenced genomes (Yoshida *et al.*, 2008, Ou *et al.*, 2015), and each of these isolates has subsequently been lost to science. *Microcystis* phage Ma-LMM01, classified as an unassigned myovirus, has been the best studied. The availability of Ma-LMM01's full genome sequence has led to analyses of distribution (*via* PCR and qPCR-based techniques) and some characterization of its genetic regulation (Yoshida-Takashima *et al.*, 2012, Rozon & Short, 2013).

Ma-LMM01 appears to have been host specific in lab studies, targeting *M. aeruginosa* at the strain level (Yoshida *et al.*, 2006). This has led to the hypothesis that phage play a role in modulating dominant strains during blooms (Yoshida *et al.*, 2008). Ecologically, one gene from this virus (*gp91*), encoding a viral tail sheath and present in the genomes of both *Microcystis* phages Ma-LMM01 and MaMV-DC (Yoshida *et al.*, 2008, Ou *et al.*, 2015), has been used *via* qPCR to suggest *Microcystis*-specific phage particles can be present at concentrations >10,000 mL^{-1} of lake water (Takashima *et al.*, 2007, Rozon & Short, 2013). These virus densities and a projected high level of host specificity suggest the potential for long-term predator-prey coevolution between virus and host, a trait generally associated with temperate phage (Bobay *et al.*, 2013). They also suggest that bloom events of susceptible *Microcystis* cells should quickly succumb to phage infection (Thingstad & Lignell, 1997).

Beyond an ability to infect and lyse *Microcystis*, the Ma-LMM01 genome encodes machinery necessary for lysogeny and induction, including 3 transposases, a serine recombinase, and 2 prophage anti-repressors. In addition, one transposase (*gp135*) and the recombinase (*gp136*) make up a 2-gene mobile genetic element called IS607, originally identified in *Helicobacter pylori*, and has led some to hypothesize that these genes further act independently as a transposon (Kersulyte *et al.*, 2000, Kuno *et al.*, 2010). Although there is an absence of lysogenic activity with *Microcystis* observed in the laboratory, expression of these genes has been documented in environmental samples (although they were not tied to lysogeny, Steffen *et al.*, 2015). Taken together, the presence of lysogeny-associated genes within *Microcystis* and the implied protection against superinfection might explain how this genus can come to dominate freshwater ecosystems and escape Hutchinson's *Paradox of the Plankton* (Hutchinson, 1961) or the "*kill-the-winner"* phenomenon (Thingstad & Lignell, 1997).

During analyses of metatranscriptomic data from *Microcystis* blooms in Lake Tai, we observed expression of phage-encoded lysogeny-associated genes that negatively correlated with expression of genes consistent with lytic infection and phage replication. Regulation of these putative lysogenic genes appears to be strongly associated with specific environmental conditions in the water column. Based on these observations, we hypothesize that phage lysogenize the *Microcystis* bloom community in a manner that is constrained by nitrogen and phosphorus availability.

Materials and Methods

Sample collection and Survey of Environmental Conditions

Samples were obtained from Lake Tai over the course of five months during the *M. aeruginosa* bloom in 2014 and have been used in conjunction with several other experiments (e.g., Krausfeldt *et al.*, 2017). Surface water samples were collected monthly from June to October from 11 different locations across the lake (Table 2.1). From all stations and dates 35 samples were selected (based on the quality and quantity of extracted RNA) were submitted for RNA-seq. Samples from Lake Tai (25-180 mLs) were collected on 0.2-µm nominal pore-size Sterivex™ (EMD Millipore Corporation, Darmstadt, Germany) and preserved for transport by adding ~ 2 mL of RNA*later* (ThermoFisher Scientific, Waltham, MA).

Water column depth and Secchi depth (SD) were measured using a water depth gauge (Uwitec, Austria) and Secchi disk, respectively. Water temperature, electrical conductivity (EC), pH, dissolved oxygen (DO) and phycocyanobilin (PC) were measured *in situ* using a multiparameter water quality sonde (YSI 6600 V2, Yellow Springs Instruments Inc., USA). Total nitrogen (TN), total dissolved nitrogen (TDN), ammonium (NH_4), nitrate (NO_3), total phosphorus (TP), total dissolved phosphorus (TDP), orthophosphate (PO_4), total dissolved solids (TDS), and chlorophyll *a* (chl *a*) were all measured according to standard methods (Jin & Tu, 1990).

Cyanobacterial toxins were determined using liquid chromatography coupled with mass spectroscopy as previously described (Boyer, 2007). Fourteen common microcystin congeners were determined by reverse phase liquid chromatography (microcystins RR, dRR, mRR, hYR, YR, LR, mLR, dLR, AR, FR, LA, LW, LF, WR and R-NOD) using a Waters ZQ4000 mass spectrometer coupled with a photodiode array spectrometer. Microcystins were all quantified against a microcystin-LR standard, and their presence confirmed using diagnostic ADDA UV signatures. We also looked for anatoxin-a (ATX), homoanatoxin-a, cylindrospermopsin (CYL) and deoxycylindrospermopsin in these extracts using HPLC coupled with mass selective (LCMS) or tandom mass (LC-MS/MS: Waters TQD) detection, and quantified against respective standards. Method detection limits were dependent on the volume filtered, ranging from 0.1-0.3 µg MC-LR / L and were less than 0.01 µg/L for anatoxin-a, cylindrospermopsin, and their variants.

RNA Extraction and Sequencing

Total RNA was extracted using the MOBIO PowerWater (now Qiagen DNeasy PowerWater) DNA isolation kit for Sterivex (Qiagen, San Diego, CA) modified and optimized for RNA isolation. RNA concentration and purity were determined using a NanoDrop™ ND-1000 spectrophotometer. Extracted RNA was tested for DNA contamination by running a polymerase chain reaction using universal bacterial 16S rDNA primers 27F and 1522R (sensitivity ~ 10 gene copies per sample). The On-Spin Column DNase I Kit (MO BIO Laboratories) was used for DNA removal, with the modification that DNase was allowed to sit for up to 30 min to increase the efficiency of DNA removal. Purified RNA samples were shipped to the Hudson Alpha Institute Genomic Services Laboratory (Huntsville, AL) for rRNA reduction, using the Ribo-Zero Gold Epidemiology rRNA removal kit, and sequencing on the Illumina HiSeq™ platform using a paired-end 125 bp flow cell.

RNA-seq Data Processing

Raw sequences were processed using the CLC Genomics Workbench v. 9.5.4 suite (QIAGEN, Hilden, Germany). Bases below 0.03 error score cutoff were trimmed. Samples were subjected to a subsequent *in silico* rRNA reduction using the SortmeRNA 2.0 software package (Kopylova *et al.*, 2012). Filtered paired-reads were competitively mapped to cyanobacterial and phage genomes (S2) with a 0.9 read-length fraction and 0.9 identity-fraction cutoffs. Transcripts were enumerated as read pairs mapped within the open reading frames of individual genes, and counts normalized by library size (unless noted). Paired reads with ends mapping to different genomes were not included in downstream analyses or counts. Sequence information has been deposited in MG-RAST database under the study Lake_Taihu_metatranscriptome_project (sample IDs in Table 2.1).

Phylogenetic Analysis

Reference sequences from Proteobacteria, Cyanobacteria, and phage identified by sequence alignment as IS607 regions in (Kuno *et al.*, 2010) were downloaded from NCBI (

Table 2.3). IS607 reference sequences were aligned in MEGA 7.0.14 software (Kumar *et al.*, 2016) using the MUSCLE algorithm (Edgar, 2004) and this alignment was then used to generate a maximum likelihood tree with a Shimodaira-Hasegawa-like approximate likelihood ratio test branch validation using PhyML (Guindon *et al.*, 2010). The reference sequences were then aligned with RNA-seq reads mapping to the Ma-LMM01 IS607 region in HMMER v. 3.1 (hmmer.org). Reads from the alignment were placed the reference tree using pplacer (Matsen *et al.*, 2010). Quantity of reads placed on the tree was visualized as branch width using the guppy software package (Matsen & Evans, 2013).

Statistical Analysis

Microcystis phage Ma-LMM01 gene read counts were $\log_2(x+10)$ transformed and Pearson correlation values were calculated in R Statistics (R Core Team, 2015) using the Hmisc R package (Harrell Jr., 2016). Mapped read counts per gene were normalized to expression of *M. aeruginosa* NIES-843 *rpoB* (as a proxy for host cell density) and plotted using the SigmaPlot software package (Systat Software, Chicago, IL). Whole genome expression was determined by counting reads mapped within gene regions on the Ma-LMM01 reference genome, which were normalized by library size, square root transformed, and used to generate a Bray-Curtis dissimilarity matrix and non-metric multidimensional scaling (nMDS) plots in the PRIMER7 software suite (Clark & Gorley, 2015). Associated environmental variables were correlated with Bray-Curtis dissimilarity distributions and plotted as vectors on the nMDS. The relationship between environmental variables and expression of the phage genome was determined using the BEST analysis (Clark & Gorley, 2015). The co-occurrence of expression of whole genome expression was grouped using the CLUSTER function using the Pearson correlation coefficient as the index of association with a 0.1 p-value cutoff. The results of this analysis were visualized in a dendrogram, all in PRIMER7.

Results

Differential Expression of Genes from *Microcystis*-infecting phage

Normalized expression of the Ma-LMM01-like tail sheath (*gp091*), transposase (*gp135*), and site-specific recombinase (*gp136*) observed in Lake Tai are shown in Figure 2.1. Of the 35 samples, 2 (T07_9 and T08_9) exhibited negligible expression of phage and host genes and have been removed from subsequent analyses. In the remaining 33 samples, 16 showed more abundant expression of *gp091* relative to *gp135*, with a ratio ranging from 1.21 to 79-fold, implying that lytic infection was dominant. These samples were collected during the earlier months (June and July) of the bloom season, with the exception of T09_1, T10_7, and T10_9, which were collected during September and October. The remaining samples showed expression of the *gp135* and *gp136* to be greater than the expression of *gp091*, implying the *Microcystis* community was, at least to some degree, lysogenized. These samples primarily occurred during the months of August, September, and October. Statistically, sample location within the lake did not relate to expression patterns, with each station exhibiting periods with dominance of lytic or putative-lysogenic transcripts almost in equal measure across all five months. Tail sheath expression was significantly and negatively correlated with both transposase (Figure 2.2, Pearson's ρ= -0.53, p = 0.0017) and recombinase abundance (not shown, ρ= -0.57, p = 0.0001). Transposase and recombinase were very highly correlated (ρ= 0.98, $p > 10^{-8}$, R = 0.986 on a linear function fit), suggesting tightly coordinated co-expression.

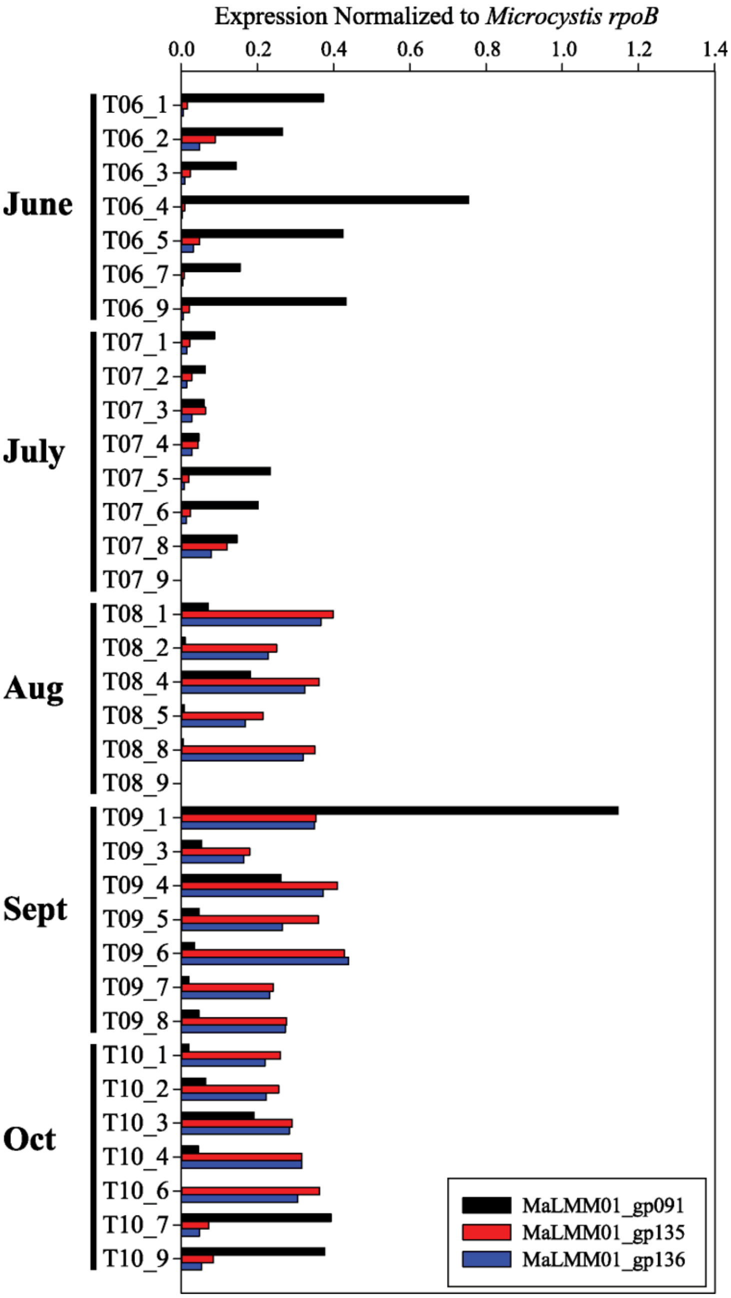

Figure 2.1. Lytic and lysogenic gene expression by station. Spatial and temporal gene expression of lytic and lysogenic genes from *Microcystis*-phage in Lake Tai. Expression of the *Microcystis* phage Ma-LMM01 phage viral tail sheath (*gp091*, black), transposase (*gp135*, red), and recombinase (*gp136*, blue) normalized by expression of *Microcystis aeruginosa* RNA polymerase B (*rpoB*) observed in the Lake Tai dataset.

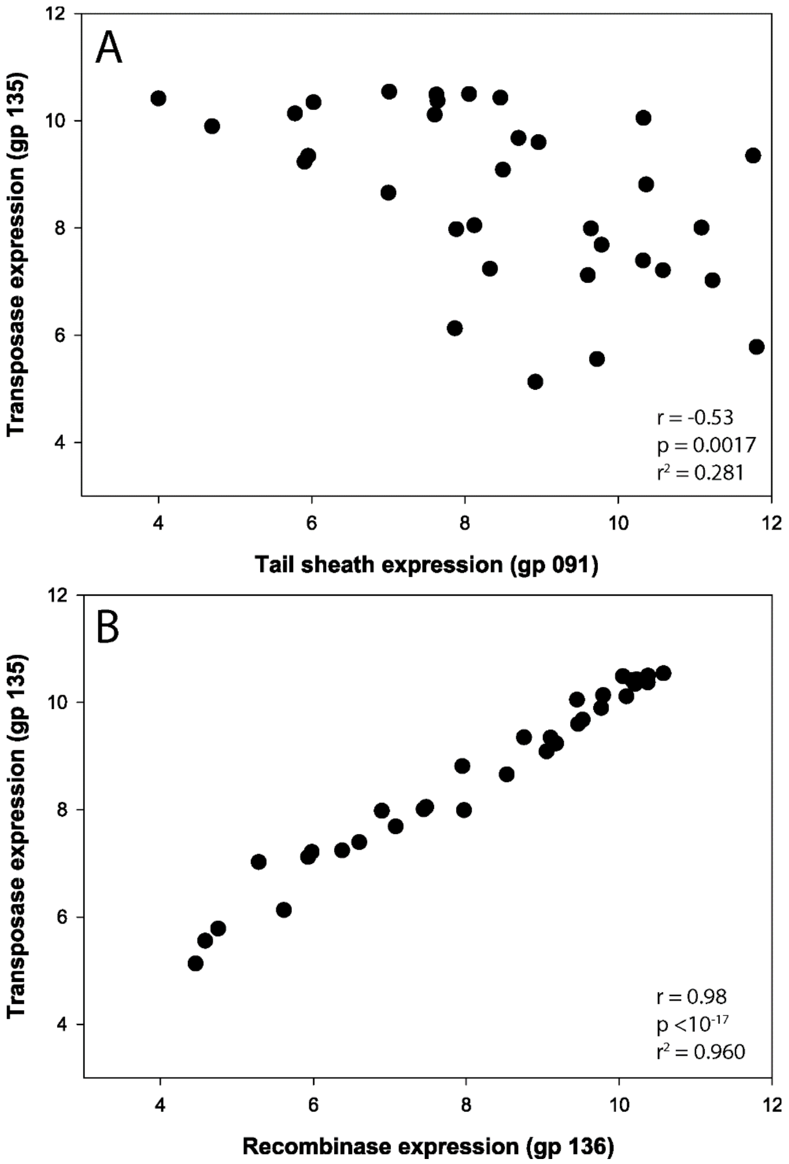

Figure 2.2. Tail sheath, transposase, and recombinase coexpression. Co-expression of genes associated with putative lytic and lysogenic infections in Lake Tai. A. Scatterplot comparing expression of Ma-LMM01 viral tail sheath (*gp091*, *x*-axis) to viral transposase (*gp135*, *y*-axis). Expression values are absolute read abundance $\log_2$ normalized and demonstrate the negative relationship between the putative lytic (*gp091*) and lysogenic (*gp135*) infection markers. B. Scatterplot comparing expression of Ma-LMM01 recombinase gene (*gp136*, x-axis) to viral transposase (*gp135*, y-axis), both putative markers of lysogenic infection of *Microcystis*.

Microcystis-phage genome expression

As a proxy for *in situ* expression of all *Microcystis* phage genes, we recruited environmental transcripts to the Ma-LMM01 genome. Results observed from samples collected in Lake Tai, and organized by hierarchical clustering, are represented in Figure 2.3. Each of the genes for both datasets generally fell into one of three major clusters. The first cluster includes all the genes potentially involved in lysogeny, including all three transposases (*gp031* and *gp032*– collapsed in branch A, *gp135*), the serine recombinase (*gp136*), and two hypothetical proteins (*gp171*, *gp067*).

The second cluster is predominantly made up of genes involved in phage packaging and cell lysis. It contains 60 genes, including 2 encoding lysozymes (*gp069* – collapsed in branch W, and *gp095* – collapsed in branch X) and the genes for DNA terminase (*gp118* - collapsed in branch DD), DNA primase (*gp134* - collapsed in branch AA), and a putative Fe/S oxidoreductase (*gp128* - collapsed in branch AA), which are the only ORFs with functions assigned. These genes exhibit high correlation values ($\rho \geq 0.7$), of which 48 are significantly co-expressed ($p \leq 0.1$) with at least one other gene.

The third cluster is the largest, and is made up of genes whose products are associated with nucleotide metabolism, DNA replication, and the structural components of the phage. It is made up of 112 genes including the viral tail sheath (*gp091*, collapsed in branch T), phage-encoded RecA (collapsed in branch S), the phycobilisome degradation protein NblA (collapsed in branch N), and a rIIA-like protein (collapsed in branch P). Viral tail sheath expression was highly correlated with genes *gp088* and *gp092*, which were predicted by protein size to encode viral tail tube proteins. Genes *gp086* and *gp087* also clustered with the tail sheath, which are believed to encode major head proteins for the phage particle.

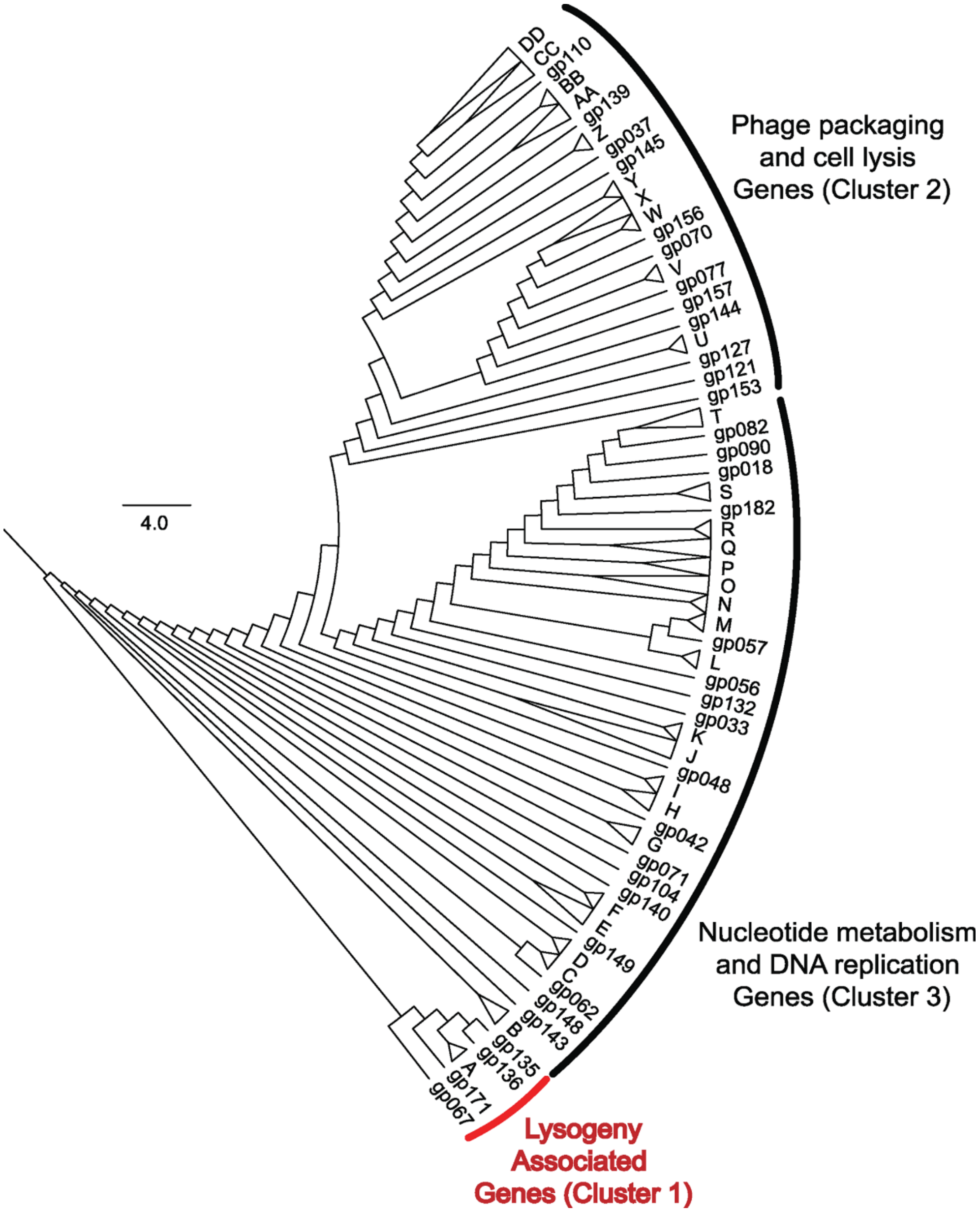

Figure 2.3. Ma-LMM01 whole genome coexpression. Cluster analysis of statistically co-expressed *Microcystis*-phage gene expression (based on Ma-LMM01 genome) in Lake Tai. Individual branches represent genes correlated with the expression of others. Transcript sets are collapsed and labeled with a letter where expression patterns were statistically indistinguishable (see Table 2.4 for genes contained in collapsed branches).

Environmental Drivers of Phage Gene Expression

A non-metric multidimensional scaling (nMDS) plot of the Bray-Curtis dissimilarity analysis of phage genome expression in Lake Tai is shown in Figure 2.4. Samples were distributed in a continuum across the *x*-axis, forming two primary clusters where phage gene expression was at least 60% similar. The position of samples along the *x*-axis corresponds significantly to the ratio of tail sheath to transposon (*gp135*) expression (a similar trend was observed when the ratio of tail sheath to recombinase (*gp136*) expression is plotted on the samples, data not shown). Vectors for environmental variables are plotted on the nMDS, showing that pH (towards lysogenic) and concentration of total dissolved solids (towards lytic) contributed most significantly to position along the x-axis. Total dissolved nitrogen and phosphorous also contributed to position along the x-axis, driving the position of samples towards greater expression of lytic genes or putative lysogenic genes, respectively. The dissolved oxygen concentration and water temperature also contributed, though more significantly to position along the *y*-axis. The BEST analysis of environmental variable contribution to expression of the entire phage genome confirmed these associations, and determined that water temperature, pH, and concentration of total dissolved solids, phosphorous, nitrogen, and oxygen concentrations were responsible for 33% of the variation in gene expression ($p = 0.05$). Phage gene expression was not correlated to toxin concentration (*gp091*: microcystin µg/L, $\rho = -0.19$, $p = 0.2956$; *gp135*: microcystin, $\rho = 0.29$, $p = 0.099$; *gp136*: microcystin, $\rho = 0.25$, $p = 0.1623$).

Discussion

We surveyed community metatranscriptomes from natural populations of *M. aeruginosa* at “bloom densities” to describe the physiology and ecology of *Microcystis*, and in the process identified active phage infections by the *Microcystis* phage Ma-LMM01. We have analyzed

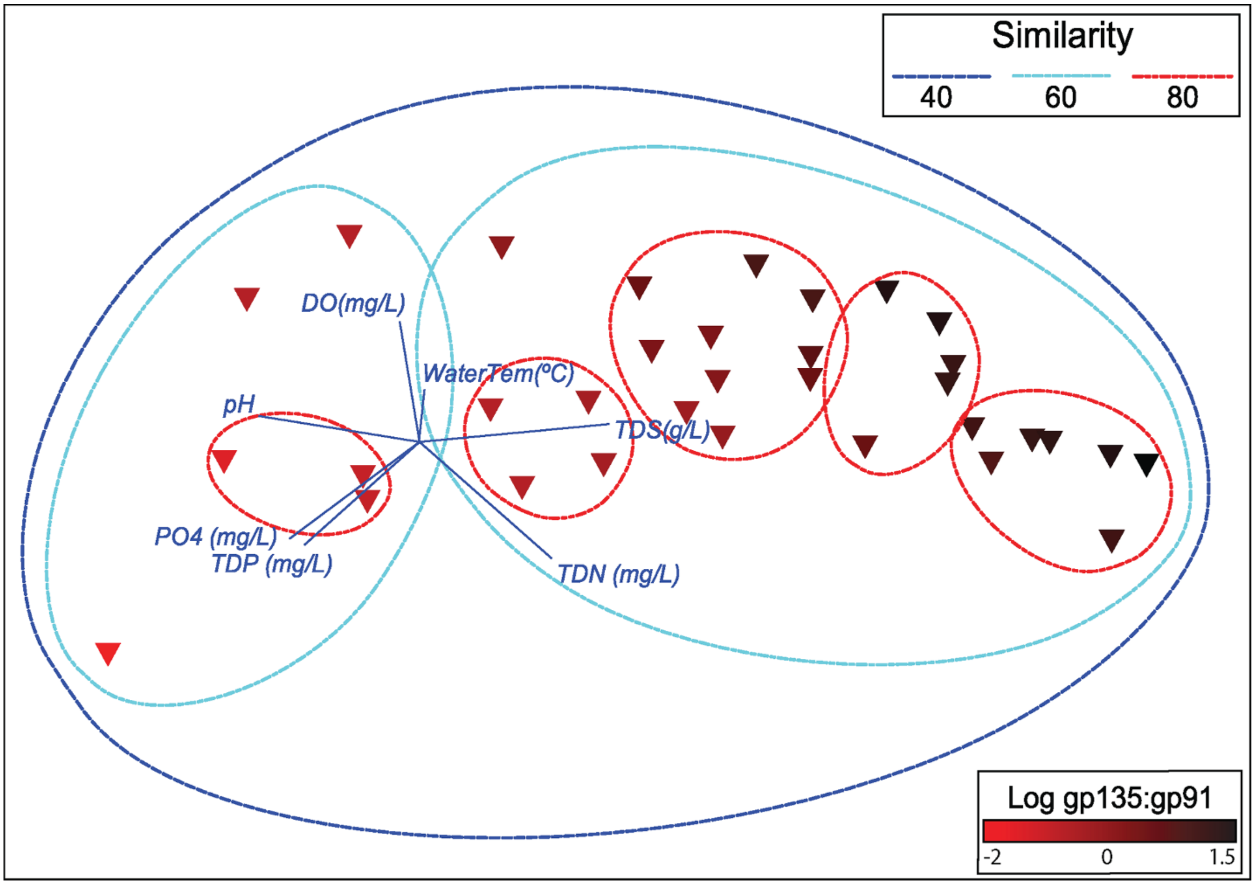

Figure 2.4. Environmental contribution to whole genome expression. A. Non-metric multidimensional scaling plot of Bray-Curtis dissimilarity between *Microcystis*-infecting phage whole genome expression for Lake Tai. Read abundance was normalized by library size and square root transformed. Ellipses represent minimum similarity between samples at the 40%, 60%, and 80% levels. Symbols have been colored based on the $\log_2$ transformed *gp091*:*gp135* expression ratio to denote lytic (black) vs lysogenic (red) dominated states. Environmental variables identified in the BEST analysis have been correlated (Pearson) with similarity between samples and plotted as vectors, indicating the direction on the 2-dimensional plane with which they correlated.

these data in light of available nutrient concentrations, toxin levels, and environmental conditions to predict how lake chemistry and climate influenced *Microcystis* phage gene expression. Our observations suggest that expression across the entire phage genome appears to have switched between the expression of genes involved in active viral replication (*i.e.,* the lytic cycle), and the expression of genes that have been proposed to allow the phage to integrate into the host genome (*i.e.,* lysogeny). Lastly, we found that the expression of phage genes appears to have been strongly associated with total dissolved solids and pH as well as the availability of nutrients, specifically the relative abundance of nitrogen and phosphorous. These observations have given rise to three distinct hypotheses: 1.) These correlations and co-occurrences are the product of random chance; 2.) The pattern of gene expression represents a novel physiological interaction (the purpose of which is currently unclear) between this phage and its host and was independent of lysogeny; 3.) The results indicate that *Microcystis* phage were actively switching between lytic and lysogenic cycles. We address these conclusions below within the context of factors that drive freshwater microbial communities.

The possibility that observed patterns in phage gene expression were the result of random chance is not supported by our analyses. The observation of similar expression patterns across Lake Tai suggests the mechanism by which *Microcystis*-infecting phage regulate gene expression has been largely conserved and is important for this virus's survival. Previous attempts to describe Ma-LMM01 transcriptional regulation in the laboratory relied on ORF orientation in the virus genome sequence, which yielded two general groups of genes: an "early" gene region containing 144 genes that were suggested to be responsible for nucleotide metabolism and genome replication, and a late gene region, encoding the remaining 40 genes, believed to encode phage structural components (Yoshida *et al.*, 2008). A subsequent study used q-rtPCR in culture to measure transcripts of the viral tail sheath (*gp091*), a putative late gene, and the gene for the phycobilisome degradation protein (NblA), a putative early region gene (Yoshida-Takashima *et al.*, 2012). They observed a temporal separation of expression between the two genes and hypothesized that the larger gene regions were consistent with early/late phage gene expression, a regulation strategy observed in other cyanomyoviruses (Clokie *et al.*, 2006). The disconnect between the regions identified by Yoshida and colleagues with our clustering is not surprising: in dealing with natural populations (unlike lab studies), we were most likely dealing with non-synchronous infections. Indeed, that there are statistically relevant relationships within the expression data suggests there are strong environmental controls on lytic *vs* lysogenic decisions.

A second potential explanation for our observations, that switching between expression states is unrelated to lysogeny, remains plausible. Much of the gene expression we attributed to genome integration originates in the virus' three putative-transposases (*gp031*, *gp032*, *gp135*) and the recombinase (*gp136*), all of which have some homologues in different strains of *M. aeruginosa* and other cyanobacteria. Transposase *gp135* belongs to a potential family of mobile elements, IS607, which was originally identified in *Helicobacter pylori* (Kersulyte *et al.*, 2000). IS607 representatives encode a corresponding serine recombinase (*gp136* in Ma-LMM01) and together, this gene pair is widespread amongst sequenced cyanobacteria (Kuno *et al.*, 2010). While these genes can be phylogenetically resolved across the length of the insertion sequence, determining the genomic origin of short sequencing reads is more challenging. Our pplacer phylogenetic tree (Figure 2.5) demonstrates the majority of reads were identified as viral in origin, but the dearth of sequenced phage genomes related to Ma-LMM01 makes it difficult to evaluate the consistency of IS607 in viruses. Additionally, the IS607 encoded serine recombinase is atypical in structure amongst other similar enzymes. The DNA-binding and catalytic domains are flipped in orientation, resulting in a recombinase that acts *via* a modified mechanism, leading to a significant reduction in insertion site specificity (Boocock & Rice, 2013). At the outset, it is not known how this would influence activity of the insertion sequence in the context of viral infection, nor how it could play a role in lysogeny, but we speculate that decreased binding specificity might better allow integration of the virus into the notoriously plastic *M. aeruginosa* genome (Kaneko *et al.*, 2007, Steffen *et al.*, 2014). It should also be noted that the presence of insertion sequences in phage genomes are very rare, as they can negatively impact virus survival (Sakaguchi *et al.*, 2005).

That observed shifts in *Microcystis*-phage gene expression represent active genome integration (lysogeny) are the most consistent with our observations and those in other systems. Moreover, that this process is tied to nutrient availability in the water column gives this observation significant ecological relevance. The formation of a lysogen would explain why putative lysogenic genes are conserved in the phage genome in a variety of geographic locations (Steffen *et al.*, 2015). There is a broad literature suggesting that phage have adapted to replicate or integrate depending on the conditions that favor the growth or senescence of their particular host (Miller & Day, 2008, Paul, 2008, Payet & Suttle, 2013, Brum *et al.*, 2016). Nutrient availability has long been associated with the formation of prophage in environmental systems, though it is generally thought to inhibit induction indirectly by limiting the material available to produce viral progeny, rather than by direct sensing for lysis-lysogeny decision making (McDaniel & Paul, 2005). In better characterized phage-host systems, such as Lambda phage, the richness of the growth medium modulates signals in host metabolism that influence the lysis-lysogeny decision (Wilson *et al.*, 2002). Unfortunately, our ability to determine the mechanism

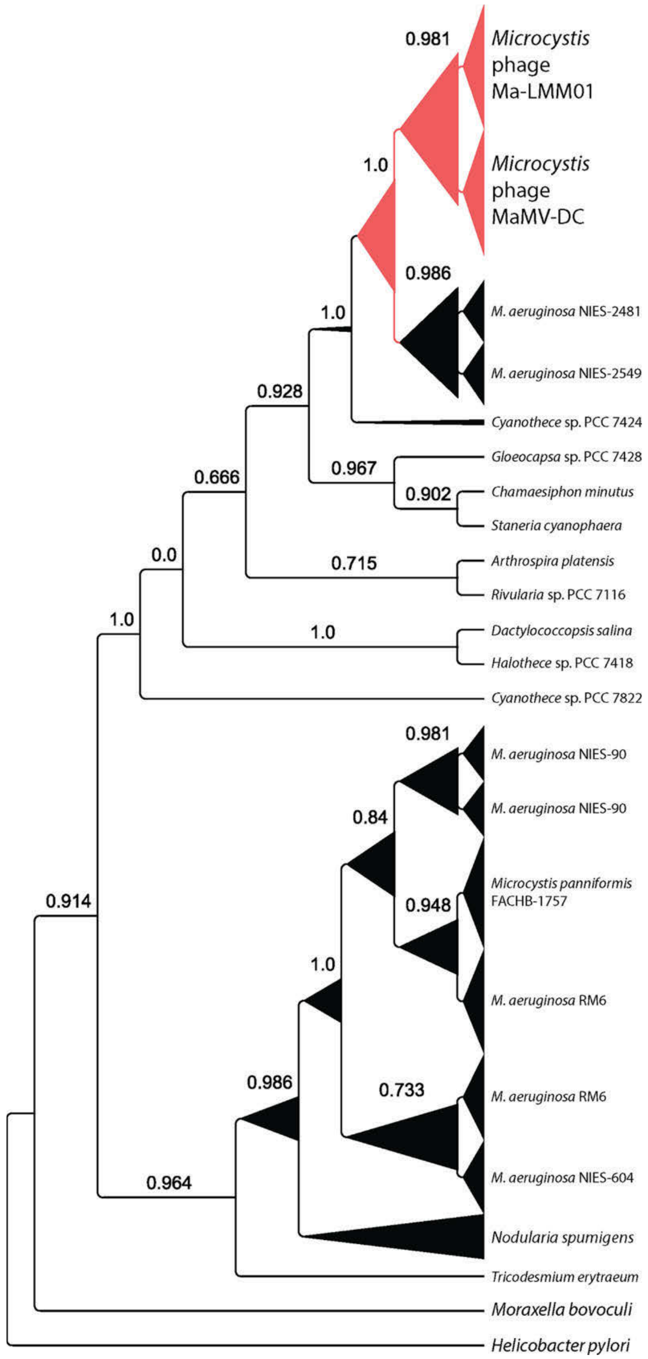

Figure 2.5. Phylogenetic distribution of IS607 reads. Bootstrapped phylogenetic tree of the mobile element IS607 where branch widths indicate abundance of Lake Tai dataset reads mapping to that branch. Branches belonging to the *Microcystis* phage are colored red.

of action from metatranscriptomic data is limited, and the lack of similarity to better characterized phage systems, such as Lambda phage, makes comparisons with *Microcystis* phage difficult to draw at this time. This is further complicated by the current unavailability of *Microcystis*-infecting phage for controlled studies. That said, it is clear from the consensus of the scientific community that we cannot discount the importance of this (and similar) environmental molecular studies (Simmonds *et al.*, 2017).

That *Microcystis* blooms can proliferate to massive densities (Steffen *et al.*, 2014) and yet somehow escape infection by the community of abundant phage (Long & Short, 2016) remains a perplexing ecological problem. This may be explained by the ability to resist infection by lytic viruses due to lysogen-induced resistance to superinfection. Indeed, while many observations lie in contrast, other studies that have suggested a "Piggyback-the-winner" model (Knowles *et al.*, 2016), which proposes that the spread of viral genomic material is best served by lysogenizing rapidly growing host cells that can persist at high densities. Clues to how this occurs mechanistically may lie in the uncharacterized genes coexpressed with the transposase and recombinase, namely *gp171* and *gp067*. While neither of these genes have close hits in the NCBI database, their implied relationship with the putative lysogenic genes suggests involvement in prophage maintenance. However, without culture work to identify their function, this remains speculation.

We observed that *Microcystis* phage gene expression could consistently be detected in *Microcystis* blooms and that a dramatic shift expression of lytic vs lysogenic gene groups was tied to environmental cues. Although the cause and effect of these cues needs further study, we hypothesize that *Microcystis*-infecting phage may actively integrate into the host genome – a state that can be distinguished from the lytic cycle *via* the relative transcription of *gp091* and *gp135*. While these new observations need continued validation and a better resolution of mechanistic controls, this study demonstrates that phage may have a strong influence population dynamics of this harmful bloom forming species.

Acknowledgments

We gratefully acknowledge Taihu Laboratory for Lake Ecosystem Research (TLLER), Jingchen Xue and Dr. Keqiang Shao for helping with sample collection and water chemical analysis. We also thank Drs. Gary LeCleir, Hans Paerl, Silvia Newell, Mark McCarthy, Guangwei Zhu, Boqiang Qin and Ferdi Hellweger for assistance and valuable discussions.

Chapter II Appendix

Table 2.1. Name, date, time, and location of each of the samples taken from Lake Tai and used for metranscriptomic sequencing. Also recorded is the environmental data collected for each sample, including water temperature during sampling (WaterTem °C), electric conductivity (EC), concentration of total dissolved solids (TDS), salinity (Sal), pH, nephelometric turbidity unit (NTU), YSI chlorphyll (YSI-CHL), phycocyanin (PC), dissolved oxygen (DO), Secchi depth (SD), depth of the lake at the sampling site (WaterDep), total nitrogen concentration (TN), total dissolved nitrogen concentration (TDN), ammonium concentration (NH4), total phosphorous concentration (TP), total dissolved phosphorous (TDP), phosphate concentration (PO4), and chlorophyll A concentration (CHLa).

Sample Name	MG-RAST Sample ID	Date	Time	Latitude	Longitude	WaterTem(ºC)	EC(µS/cm)	TDS(g/L)	Sal(‰)
T06_1	mgm4663025.3	6/7/2014	9:12	120.19067	31.51317	24.80	670	0.437	0.33
T06_2	mgm4663263.3	6/7/2014	17:26	120.22055	31.41747	27.87	712	0.439	0.33
T06_3	mgm4663272.3	6/7/2014	16:47	120.22945	31.39438	26.87	689	0.433	0.32
T06_4	mgm4663273.3	6/7/2014	8:34	120.18796	31.43609	23.72	637	0.424	0.32
T06_5	mgm4663274.3	6/7/2014	9:49	120.11638	31.44719	25.31	632	0.408	0.30
T06_7	mgm4663278.3	6/7/2014	12:11	120.18017	31.33833	25.32	608	0.392	0.29
T06_9	mgm4663280.3	6/7/2014	11:07	119.94500	31.3145	26.45	699	0.442	0.33
T07_1	mgm4664215.3	7/3/2014	13:17	120.19067	31.51317	26.03	620	0.359	0.29
T07_2	mgm4664214.3	7/3/2014	12:31	120.22055	31.41747	27.26	665	0.415	0.31
T07_3	mgm4664209.3	7/3/2014	17:53	120.22945	31.39438	27.96	691	0.425	0.32
T07_4	mgm4664210.3	7/3/2014	8:42	120.18796	31.43609	24.64	647	0.423	0.32
T07_5	mgm4664213.3	7/3/2014	13:46	120.11638	31.44719	25.24	590	0.382	0.28
T07_6	mgm4664212.3	7/3/2014	14:38	120.02817	31.45001	26.69	558	0.351	0.26
T07_8	mgm4664211.3	7/3/2014	14:54	120.03182	31.39761	27.85	570	0.352	0.26
T07_9	mgm4664208.3	7/3/2014	15:22	119.94500	31.3145	26.14	602	0.383	0.28
T08_1	mgm4664613.3	8/14/2014	8:00	120.19067	31.51317	26.70	543	0.342	0.25
T08_2	mgm4664610.3	8/14/2014	6:50	120.22055	31.41747	26.47	557	0.352	0.26
T08_4	mgm4664609.3	8/14/2014	7:30	120.18796	31.43609	26.70	557	0.351	0.26

Table 2.1. Continued.

Sample Name	MG-RAST Sample ID	Date	Time	Latitude	Longitude	WaterTem(ºC)	EC(μS/cm)	TDS(g/L)	Sal(‰)
T08_5	mgm4664612.3	8/14/2014	8:30	120.11638	31.44719	26.65	552	0.348	0.26
T08_8	mgm4664608.3	8/14/2014	9:30	120.03182	31.39761	26.61	467	0.294	0.22
T08_9	mgm4664611.3	8/14/2014	10:00	119.94500	31.3145	26.68	517	0.326	0.24
T09_1	mgm4664691.3	9/9/2014	14:15	120.19067	31.51317	27.37	542	0.337	0.25
T09_3	mgm4664695.3	9/9/2014	18:14	120.22945	31.39438	27.60	524	0.324	0.24
T09_4	mgm4664697.3	9/9/2014	13:52	120.18796	31.43609	27.41	532	0.331	0.24
T09_5	mgm4664692.3	9/9/2014	14:43	120.12684	31.44614	28.25	515	0.315	0.23
T09_6	mgm4664696.3	9/9/2014	15:12	120.02817	31.45001	27.38	470	0.292	0.21
T09_7	mgm4664694.3	9/9/2014	16:57	120.18017	31.33833	27.12	482	0.301	0.22
T09_8	mgm4664690.3	9/9/2014	15:33	120.03182	31.39761	27.55	486	0.301	0.22
T10_1	mgm4663615.3	10/8/2014	8:15	120.19067	31.51317	21.29	492	0.344	0.26
T10_2	mgm4663618.3	10/8/2014	13:30	120.22055	31.41747	21.91	473	0.327	0.24
T10_3	mgm4663619.3	10/8/2014	13:15	120.22945	31.39438	22.71	469	0.319	0.24
T10_4	mgm4663617.3	10/8/2014	7:45	120.18796	31.43609	20.69	475	0.337	0.25
T10_6	mgm4663620.3	10/8/2014	9:30	120.02817	31.45001	21.47	512	0.357	0.27
T10_7	mgm4663630.3	10/8/2014	11:45	120.18017	31.33833	22.20	467	0.321	0.24
T10_9	mgm4663634.3	10/8/2014	10:30	119.94500	31.3145	21.70	479	0.322	0.25

Table 2.2. Environmental data collected for each sample, including pH, nephelometric turbidity unit (NTU), YSI chlorphyll (YSI-CHL), phycocyanin (PC), dissolved oxygen (DO), Secchi depth (SD), depth of the lake at the sampling site (WaterDep), total nitrogen concentration (TN), total dissolved nitrogen concentration (TDN), ammonium concentration (NH4), total phosphorous concentration (TP).

Sample Name	pH	NTU	YSI-CHL(µg/L)	PC (cells/mL)	DO(%)	DO(mg/L)	SD(cm)	WaterDep(m)	TDN (mg/L)	NH4 (mg/L)	TP (mg/L)	TN (mg/L)
T06_1	8.19	13.3	15.2	2831	113.9	9.43	40	2.0	2.50	0.077	0.092	2.97
T06_2	8.77	26.4	45.1	11107	128.7	10.08	15	1.2	1.83	0.076	1.176	17.23
T06_3	9.08	10.0	22.4	5050	145.0	11.56	40	1.8	1.99	0.071	0.515	9.39
T06_4	8.35	19.5	6.7	7456	106.7	9.01	45	2.4	2.09	0.068	0.136	3.55
T06_5	7.94	18.1	11.1	1374	54.4	4.46	30	2.8	3.89	0.827	0.728	10.77
T06_7	8.31	21.8	4.8	2125	113.5	9.31	30	2.6	1.79	0.076	0.077	2.38
T06_9	8.18	23.2	17.0	2516	90.4	7.26	20	1.5	3.18	0.928	0.215	4.02
T07_1	8.75	20.8	5.0	2482	106.7	8.64	30	2.4	1.40	0.084	0.0602	1.78
T07_2	8.90	143.8	17.4	62212	98.3	7.79	0	1.5	1.17	0.119	0.2333	4.21
T07_3	9.73	24.6	7.6	10098	145.0	11.34	30	2.1	1.04	0.055	0.0940	1.94
T07_4	9.03	39.4	5.9	9534	99.9	8.29	30	2.7	1.06	0.138	0.2292	4.94
T07_5	8.99	25.3	13.0	2304	76.0	6.24	30	3.2	2.78	0.338	0.1216	3.30
T07_6	9.01	21.0	15.9	7035	136.0	10.91	40	2.0	4.07	1.313	0.1744	4.86
T07_8	9.11	18.5	4.0	6043	113.9	8.93	60	2.5	4.62	1.304	0.1739	5.24
T07_9	8.93	33.4	9.0	1425	47.5	3.84	30	1.8	4.20	1.672	0.2273	4.66
T08_1	9.57	57.9	12.0	22185	82.6	6.60	35	2.5	0.71	0.094	0.162	1.38
T08_2	8.98	51.5	8.0	18561	29.1	2.33	40	1.5	1.35	0.652	0.203	2.40
T08_4	9.38	37.7	10.8	21232	92.3	7.38	50	2.9	0.53	0.063	0.152	1.55
T08_5	9.38	71.3	18.6	36136	58.0	4.64	25	3.1	1.29	0.283	2.132	21.52
T08_8	9.32	41.3	12.7	21316	84.7	6.79	40	2.7	1.63	0.606	0.324	2.72
T08_9	9.20	45.0	12.0	2159	36.9	2.95	32	2.2	2.97	1.155	0.343	3.73
T09_1	10.03	28.2	21.6	10618	126.2	9.98	40	2.6	0.52	0.070	0.1704	1.33
T09_3	9.96	31.4	10.6	8714	121.6	9.57	30	2.2	0.58	0.109	0.4505	4.52
T09_4	9.60	33.2	12.5	9345	117.9	9.32	40	3.0	0.53	0.087	0.2040	1.39
T09_5	9.76	881.9	82.6	239256	67.7	5.27	0	2.7	1.31	0.363	11.6860	48.17
T09_6	9.88	27.3	15.3	32381	139.5	11.03	20	2.2	1.22	0.423	1.7083	17.79
T09_7	9.94	25.0	12.6	15009	159.0	12.63	30	3.1	0.67	0.162	4.2445	39.33
T09_8	9.98	57.6	17.7	29456	151.5	11.94	20	2.7	1.17	0.123	0.6253	7.01

Table 2.2. Continued.

Sample Name	pH	NTU	YSI-CHL(µg/L)	PC (cells/mL)	DO(%)	DO(mg/L)	SD(cm)	WaterDep(m)	TDN (mg/L)	NH4 (mg/L)	TP (mg/L)	TN (mg/L)
T10_1	9.38	34.2	34.2	4989	68.6	6.07	30	2.4	0.77	0.097	0.141	1.37
T10_2	10.05	48.5	15.3	4267	32.1	2.81	25	1.5	0.64	0.123	1.199	12.66
T10_3	9.94	79.1	23.2	43149	76.8	6.60	20	2.3	0.58	0.086	1.925	20.86
T10_4	9.52	43.1	13.0	8814	103.2	9.25	30	2.8	0.66	0.140	0.137	1.24
T10_6	9.17	23.1	22.9	3006	97.6	8.61	40	2.0	2.56	0.550	0.182	2.92
T10_7	9.77	47.0	20.5	13945	32.8	2.85	25	3.1	0.62	0.131	1.905	20.34
T10_9	9.21	85.9	31.3	15407	21.9	1.93	10	2.0	2.10	0.382	0.374	3.87

Table 2.3. Environmental data collected for each sample, including total dissolved phosphorous (TDP), phosphate concentration (PO4), and chlorophyll A concentration (CHLa).

Sample Name	TDP (mg/L)	PO4 (mg/L)	CHLa (µg/L)
T06_1	0.028	0.005	40.18
T06_2	0.029	0.006	959.76
T06_3	0.027	0.005	370.29
T06_4	0.027	0.009	54.01
T06_5	0.093	0.069	368.28
T06_7	0.010	0.012	27.34
T06_9	0.093	0.061	48.21
T07_1	0.0200	0.006	18.75
T07_2	0.0167	0.003	119.33
T07_3	0.0193	0.013	46.87
T07_4	0.0183	0.004	319.92
T07_5	0.0364	0.021	38.13
T07_6	0.0745	0.054	40.62
T07_8	0.0779	0.058	42.41
T07_9	0.1283	0.101	15.25
T08_1	0.080	0.050	74.58
T08_2	0.037	0.019	71.61
T08_4	0.037	0.015	89.22
T08_5	0.061	0.036	3257.60
T08_8	0.188	0.162	100.72
T08_9	0.138	0.112	21.11
T09_1	0.0746	0.060	57.14
T09_3	0.0667	0.053	249.59
T09_4	0.0686	0.058	50.89
T09_5	0.0343	0.059	16182.00
T09_6	0.0487	0.029	3414.96
T09_7	0.0452	0.027	1204.72
T09_8	0.0800	0.058	307.12
T10_1	0.064	0.054	32.6
T10_2	0.056	0.044	1037.9
T10_3	0.032	0.021	1622.1
T10_4	0.055	0.044	29.7
T10_6	0.098	0.082	37.5
T10_7	0.037	0.026	1507.6
T10_9	0.073	0.058	97.5

Table 2.4. Statistics from mapping metatranscriptomic reads to reference genomes from each of the Lake Tai samples. Cells show the total number of reads mapped to each of the given genomes during competitive read mapping with a 0.9 similarity and a 0.9 length fraction cutoff.

Organism	*M. a* NIES-843	*M. a* NIES-2549	*M. a* NIES-2481	*M. p* FACHB-1757	*Cyanothece* sp. PCC7424	Phage LMM01	Phage MaMV-DC	*S. elongatus* PCC6301
Accession	AP009552	CP011304	CP012375	CP011339	CP001287	AB231700	KF356199	AP008231
T06_1 Reads	2473749	1509966	1227663	2304363	37536	9660	11392	21183
T06_2 Reads	6838254	2350396	2407783	5756943	3043	19598	23946	1576
T06_3 Reads	6270643	2758914	2564920	5397864	3137	11938	14618	1251
T06_4 Reads	5501927	3028906	2569714	4839264	4755	32807	41931	4395
T06_5 Reads	6830252	3606800	3399024	7152728	2022	24223	30434	1101
T06_7 Reads	3317484	1908777	1608661	2949165	5585	5361	6759	979
T06_9 Reads	4178215	2405015	2050113	3823548	5202	21325	27033	3449
T07_1 Reads	2138281	1102837	938708	1774927	9008	3303	3510	10539
T07_2 Reads	5707816	2768949	1776582	5311607	1001	7305	8526	973
T07_3 Reads	4349279	1691464	1582329	3563322	2788	7434	8523	1711
T07_4 Reads	5635475	2780756	2481940	5481052	4052	4956	5704	2256
T07_5 Reads	5900773	3033124	2791331	5299654	1166	20886	24705	1070
T07_6 Reads	4995064	2396600	2091117	4201333	9762	14531	17239	6277
T07_8 Reads	6779750	2949575	2673246	6333455	3472	16875	19140	2201
T07_9 Reads	40590	31045	32837	55249	50741	49	34	60581
T08_1 Reads	5105630	2633705	3210020	5269680	6720	7276	6911	5947
T08_2 Reads	5150118	2394107	2969943	5942550	4190	4151	2591	3519
T08_4 Reads	3938926	1783820	2427106	3944607	6654	7162	7317	7191
T08_5 Reads	4741519	2174513	2646983	5801168	2490	3784	2424	2380
T08_8 Reads	4797307	1902221	2283664	4892917	2718	2699	1593	2694
T08_9 Reads	99338	76684	76165	99170	36441	68	42	30176
T09_1 Reads	1958698	950902	975932	1975702	71688	8140	9436	21297
T09_3 Reads	3628908	1662078	1813394	3352006	12997	2607	2653	6990
T09_4 Reads	3830668	2483296	2237795	3756592	24964	6512	7015	14943
T09_5 Reads	5866740	2354633	2914317	6822543	1933	7594	5634	1648
T09_6 Reads	5239495	2109641	2384288	6530059	3357	7539	4709	2704
T09_7 Reads	5194491	2722388	2597321	5562852	4549	3436	2255	1325
T09_8 Reads	6099883	2112075	2207356	5914912	2584	5601	4210	1752
T10_1 Reads	4781943	1230389	1201007	2542002	16099	3544	2866	20623
T10_2 Reads	7156145	2344360	2239095	4880539	4181	10471	8621	2500
T10_3 Reads	5109756	2561678	2413870	5769641	4453	6644	6558	1516
T10_4 Reads	5277495	1831066	1903650	4341048	6420	7541	5581	5104
T10_6 Reads	3487391	1393729	1435160	3427383	18629	3980	2356	21352
T10_7 Reads	9440537	4806299	4334415	8935964	22798	57073	66796	29297
T10_9 Reads	2352506	1092228	1020823	2163627	11213	11454	13012	15502

Table 2.5. Reference sequences used in Figure 2.5 and their accession numbers.

Organism	Accession Number
Microcystis aeruginosa NIES-90	AB543092
Microcystis aeruginosa NIES-90	AB543095
Nodularia spumigena CCY9414	CP007203
Trichodesmium erythraeum IMS101	CP000393
Moraxella bovoculi strain 33362	CP011379
Helicobacter pylori jhp1409	AF189015
Cyanothece sp. PCC 7822	CP002198
Dactylococcopsis salina PCC 8305	CP003944
Halothece sp. PCC 7418	CP003945
Arthrospira platensis YZ	CP013008
Rivularia sp. PCC 7116	CP003549
Cyanothece sp. PCC 7424	CP001291
Microcystis aeruginosa NIES-2481	CP012375
Microcystis aeruginosa NIES-2549	CP011304
Gloeocapsa sp. PCC 7428	CP003646
Chamaesiphon minutus PCC 6605	CP003600
Stanieria cyanosphaera PCC 7437 plasmid pSTA7437.01	CP003654
Microcystis aeruginosa NIES-112	AB543093
Microcystis aeruginosa NIES-604	AB543094
Microcystis panniformis FACHB-1757	CP011339
Microcystis aeruginosa RM6	AB543096
Microcystis Phage Ma-LMM01	AB231700
Microcystis Phage MaMV-DC	KF356199

Table 2.6. Branches collapsed in Figure 2.3. Each of the collapsed branches is identified by the letter from Figure 2.3, and beneath are listed the genes whose expression patterns were statistically indistinguishable by hierarchical clustering.

Collapsed Branch	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code	Gene Code
A	gp031	gp032											
B	gp146	gp147	gp15										
C	gp138	gp151											
D	gp027	gp043											
E	gp154	gp066	gp150	gp068	gp155								
F	gp047	gp158											
G	gp152	gp052	gp058										
H	gp061	gp054	gp055										
I	gp007	gp022											
J	gp040	gp044	gp179	gp163	gp041	gp030	gp164	gp162	gp161	gp122	gp124	gp123	gp160
K	gp059	gp075											
L	gp016	gp029											
M	gp023	gp024											
N	nblA	gp063											
O	gp159	gp026	gp034	gp172	gp175	gp184	gp013	gp028	gp131	gp174	gp177	gp004	gp021
O (cont.)	gp014	gp181	nrdA	gp170	gp176	gp180	gp178	gp003	gp173				
P	gp060	gp020	gp045	rIIA	gp025	gp072	gp009	gp084					
Q	gp017	gp183	gp019	gp039	gp165	gp051	gp168	gp169	gp166	gp167			
R	gp073	gp116											
S	nrdB	recA	gp088										
T	gp091	gp092	gp086	gp087	gp085	gp081	gp083						
U	gp137	gp141											
V	gp078	gp129											
W	gp065	gp069											
X	gp010	gp099	gp064	gp098	gp080	gp095							
Y	gp012	gp079											
Z	gp035	gp133											
AA	gp076	gp097	gp096	gp128	gp053	gp036	gp134	gp074	gp130				
BB	gp125	gp126											
CC	gp101	gp094	gp113	gp114	gp120	gp093	gp112						
DD	gp100	gp103	gp106	gp089	gp108	gp102	gp118	gp0109	gp115	gp111	gp117		
DD (cont.)	gp105	gp107	gp119										

CHAPTER III:

NOVEL VIRUSES WITHIN SPHAGNUM PEAT IDENTIFIED BY HIGH-THROUGHPUT TRANSCRIPT SEQUENCING DATA

Publication Note

This chapter is a version of an article in preparation for submission to *Applied and Environmental Microbiology* by Joshua M.A. Stough, Max Kolton, Joel Kostka, David Weston, Dale Pelletier, and Steven W. Wilhelm.

My contribution to this work was the analysis conceptualization and design, data management and analysis, and primary authorship and editing of the manuscript.

Abstract

Sphagnum peat bogs play an important role in global carbon storage, while also representing significant sources of economic and ecological value. Despite recent efforts to describe microbial diversity and metabolic potential of the *Sphagnum* microbiome, very little is known about the viral constituents of the community. Moreover, previous studies have focused on metagenomic sequencing to describe the resident *Sphagnum* microbes, restricting information to relative abundance and functional potential, rather than microbial activity. In this study, we used metatranscriptomics to describe the diversity and activity of viruses infecting hosts within the *Sphagnum* peat bog microbiome. Six *Sphagnum tissue* samples were obtained from peat bogs in Northern Minnesota and total RNA was extracted and sequenced. Metatranscriptome libraries were assembled and contigs screened for the presence of conserved viral marker genes. Using the phage major capsid protein, *gp23*, as a phylogenetic marker for phage, we identified 33 contigs representing phage strains active in the community. Similarly, RNA-dependent RNA polymerase and the NCLDV major capsid protein were used as markers for single-stranded RNA viruses and giant viruses, respectively. In total 114 contigs were identified as originating in currently undescribed ssRNA viruses, 22 of which represent near-complete RNA virus genomes encoding multiple genes. An additional 64 contigs were identified as being from novel giant viruses, many of which with significant similarity to the recently discovered Klosneuviruses, and 7 contigs were identified as putative virophage. Quantitative information from sequence read mapping was used to generate correlation co-occurrence networks with expression of host housekeeping gene *rpb1*. Clusters of co-expression were used to predict virus-host relationships, identifying 11 potential partnerships or groups. Together, our methods offer new tools for the identification of virus diversity in understudied clades, and suggest viruses play a considerable role in the ecology of the *Sphagnum* microbiome.

Introduction

Peat bogs represent one of the most significant biological carbon sinks on the planet, storing an estimated 25% of all terrestrial carbon in the form of partially decomposed organic matter. This accumulation of carbon is achieved through much slower rates of respiration and decomposition, due in large part to the low pH, nutrient-poor, and anaerobic environments created by the dominant moss population (van Breemen, 1995, Lamers *et al.*, 2000), of which the genus *Sphagnum* is most prevalent (Turetsky, 2003, Turetsky *et al.*, 2012). As these environmental conditions favor the growth of *Sphagnum* over vascular plants, primary production is dominated by the moss, which further retards decomposition due to production of antimicrobial compounds such as sphagnic acid (Verhoeven & Liefveld, 1997, Freeman *et al.*, 2001, Mellegard *et al.*, 2009) and sphagnan (Stalheim *et al.*, 2009, Hajek *et al.*, 2011). Despite this, *Sphagnum* and other peat mosses cultivate a diverse, symbiotic microbiome that appears to abate nutritional gaps for the moss and generally contribute to the unique biogeochemical characteristics of the peatland ecosystem (Lin *et al.*, 2014, Leppanen *et al.*, 2015, Kostka *et al.*, 2016). In addition to their value as reservoirs for microbial diversity, the partially decomposed organic matter, known as *Sphagnum* peat, serves as an important economic resource for use in horticulture. Many peat bogs have begun to experience stress due to anthropogenic disturbances such as climate change (Dudova *et al.*, 2013, Ireland *et al.*, 2014, Swindles *et al.*, 2015, Galka *et al.*, 2017). As such, the *Sphagnum* microbiome is of considerable interest in peatland conservation and the ecosystem's services to the surrounding environment.

While some work has been done characterizing the microbes that colonize *Sphagnum* biomass using rRNA sequencing, very little is currently known about the ecological factors that define community structure. Studies suggest that subtle differences in pH and available nutrients manipulated by different *Sphagnum* species and strains create distinct microbial consortia specific the moss host (Opelt *et al.*, 2007, Bragina *et al.*, 2013, Leppanen *et al.*, 2015), while others observe a more homogenous community (Bragina *et al.*, 2012), highlighting a need for further study. Culture-dependent experiments isolating endophytic bacteria indicates *Sphagnum* cultivates symbionts with antifungal activity (Opelt & Berg, 2004, Opelt *et al.*, 2007) and nitrogen fixation (Leppanen *et al.*, 2015), which may be passed on vertically to the moss progeny (Bragina *et al.*, 2013). Though environmental conditions and host-microbe symbiotic interactions are fundamental to the structure and function of microbial communities, the influence of virus populations on the *Sphagnum* microbiome remains almost entirely unexplored. Viruses are the most abundant biological entities on Earth, and play important roles in global ecosystems by driving the evolution of their hosts through predator-prey interactions and horizontal gene transfer (Brussaard *et al.*, 2008). In addition, viruses lyse single-celled primary producers and heterotrophs, releasing bioavailable nutrients tied up in the biomass of prokaryotes and eukaryotic protists (Jover *et al.*, 2014). Viruses also act as a top-down control on the composition and evenness of microbial communities by targeting hosts that reach higher cell densities, a phenomenon referred to as the "*kill-the-winner*" model (Thingstad & Lignell, 1997).

As culturing viruses requires hosts that can be grown in a lab, environmental viruses are poorly understood and represented in reference databases. Previous efforts to describe environmental viromes have focused on DNA sequencing through metagenomics, usually by filtering virus particles based on size and density, followed by high-throughput sequencing. While this method has proven very powerful, rapidly expanding available reference material for bacteriophage (Roux *et al.*, 2016, Simmonds *et al.*, 2017), it leaves the considerable diversity of RNA viruses largely untapped (Steward *et al.*, 2013). Moreover, selecting for viruses based on size filters out giant viruses that have been shown to be both environmentally relevant and phylogenetically informative (Yutin *et al.*, 2009, Wilhelm *et al.*, 2017). In addition, metagenomic sequencing limits observations to relative abundance of virus particles, such that inferences on viral activity require tenuous assumptions. The advent of high-throughput RNA sequencing offers viral ecologists the opportunity to study active infections in the environment, as viruses only produce transcripts inside a host, while also capturing fragments of RNA virus genomes. Moreover, when sequencing is of sufficient depth and multiple samples are collected with spatial and temporal variability, these data present an opportunity to develop hypothetical relationships between virus and host markers (Moniruzzaman *et al.*, 2017).

In this study, we analyzed metatranscriptomes from the microbial community inhabiting the *Sphagnum* peat bogs of northern Minnesota, with the goal of describing the active viral constituents of the *Sphagnum* microbiome. Using marker genes conserved within several viral taxa, we identified an active and diverse bacteriophage population, largely undescribed in previous studies. We also identified a diverse consortium of “giant” viruses and potentially corresponding virophage, including several viruses closely related to the recently discovered Klosneuviruses (Schulz *et al.*, 2017), actively infecting hosts in the environment. Finally, a number of novel positive-sense ssRNA viruses, some of which have been assembled into near complete genomes, were observed. With this information in hand we developed statistical network analyses to relate co-expression of viral marker genes with housekeeping genes from potential hosts, proposing several virus-host groups that can be further tested in a laboratory setting. Together, these results demonstrate new potential model systems to study virus-host interactions in the peat bog ecosystem, and provide insight into the significant viral influence on the *Sphagnum* microbiome.

Materials and Methods

Sample collection and Survey of Environmental Conditions

Triplicate individual plants of *Sphagnum magellanicum and Sphagnum fallax* were collected on August 2015 from the SPRUCE experiment site at the S1 bog on the Marcell Experimental Forest (U.S. Forest Service http://mnspruce.ornl.gov/). The S1 Bog is an acidic and nutrient-deficient ombrotrophic *Sphagnum*-dominated peatland bog (surface pH≤4.0) located approximately 40 km north of Grand Rapids, Minnesota, USA (47°30.476′ N; 93°27.162′ W; 418 m above mean sea level) (Wilson *et al.*, 2016, Hanson *et al.*, 2017, Warren *et al.*, 2017). To characterize the *Sphagnum* virome, *Sphagnum* samples were collected as previously described. Briefly, randomly collected photosynthetically active *Sphagnum* stems (phyllosphere) were cleaned from unrelated plant debris, and frozen immediately on dry ice. Frozen samples were overnight shipped to the Georgia Institute of Technology for DNA and RNA extraction.

RNA Extraction and Sequencing

One gram of *Sphagnum* phyllosphere tissue was ground with a mortar and pestle under liquid nitrogen. The fine powder was transferred to 10 extraction tubes and total RNA was isolated using the PowerPlant RNA Isolation Kit with DNase according to the manufacturer's protocol (MoBio Laboratories, Carlsbad, CA, USA). DNA-depleted RNA was quantified using the Qubit RNA HS Assay Kit (Invitrogen, Carlsbad, CA, USA) and quality was assessed on the Agilent 2100 BioAnalyzer using the Agilent RNA 6000 Pico Kit (Agilent Technologies). Additionally, the absence of DNA contamination was confirmed by running a polymerase chain reaction using universal bacterial 16S rRNA primers 515F and 806R. Finally, RNA samples without detectible DNA contamination and exhibiting an RNA integrity number (RIN) > 6 were pooled. RNA samples were shipped to the Department of Energy Joint Genome Institute for rRNA depletion, cDNA library preparation, and sequencing on the Illumina™ HiSeq2000 platform using a single-end 250bp flow cell.

RNA-Seq Data Processing

Raw sequences were downloaded from the Department of Energy Joint Genome Institute server and processed using the CLC Genomics Workbench v. 10.0.1 (QIAGEN, Hilden, Germany). Reads below a 0.03 quality score cutoff were removed from subsequent analyses, and the remaining reads were trimmed of any ambiguous and low quality 5' bases. Samples were subjected to a subsequent *in silico* rRNA reduction using the SortmeRNA 2.0 software package (Kopylova *et al.*, 2012). Filtered paired reads were *de novo* assembled with cutoffs of 300 base minimum contig length and average coverage of 2, leaving a total of 705,526 contigs across all samples. Full RNA-seq libraries have been made publicly available on the JGI website under accession number Gp0146911.

Screening Assemblies for Marker Genes

To identify contigs specific to the NucleoCytoplasmic Large DNA Virus (NCLDV) clade, contig libraries were screened for the presence of 10 genes previously identified as core NCLDV genes (Table 3.1) as previously described (Moniruzzaman *et al.*, 2017, Stough & Wilhelm, 2017). Briefly, contig libraries were queried against NCVOG protein databases for each of the 10 marker genes in a Blastx search with a minimum e-value cutoff of 10^{-3}. Resulting hits were then queried against the refseq protein database and only contigs with top hits to virus genes were used in subsequent analyses. A similar method was used to identify virophage transcripts, where the virophage major capsid protein and packaging ATPase genes were used as markers.

Contigs derived from ssRNA viruses were identified by screening the contig library for RNA-dependent RNA Polymerase (RDRP). A BLAST database of RDRP sequences was downloaded from the pfam database (Finn *et al.*, 2016) under code pf00680. Contigs were aligned using blastx with a minimum evalue of 10^{-4}. Hits were queried against the refseq protein database and only hits to viral RDRP genes were retained for downstream analyses.

To identify RNA virus genome fragments, contig libraries were screened as described above using a core set of genes observed in RNA viruses (Table 3.2). BLAST databases for core RNA virus genes were constructed from reference sequences downloaded from pfam. Query sequences were then cross-referenced to identify contigs with hits to multiple RNA virus core genes. Only contigs > 1000 bases with at least one viral RDRP region were retained for further analysis. ORFs were predicted on these putative partial genomes using the CLC Genomics Workbench. Features on the partial genomes were predicted using the Pfam HMM domain and the NCBI Conserved Domain Database searches (Finn *et al.*, 2015, Marchler-Bauer *et al.*, 2015). Genome architecture was visualized using the Illustrator for Biological Sequences (IBS) software package (Liu *et al.*, 2015).

Phylogenetic Analysis

Reference sequences for viral marker genes were downloaded from the InterPro and RefSeq databases (Finn *et al.*, 2017). Reference sequences were aligned using the MUSCLE alignment algorithm (Edgar, 2004) in the MEGA v7.0.26 software package (Kumar *et al.*, 2016). Maximum likelihood phylogenetic trees were constructed in PhyML (Guindon *et al.*, 2010) with the LG substitution model and the aLRT SH-like likelihood method. Selected contigs assembled from the metatranscriptomes were translated into proteins according to the reading frame of the top BLAST hit. Translated proteins were placed on the reference trees in a maximum likelihood framework in pplacer (Matsen *et al.*, 2010). Trees with abundance data were visualized using the iToL web interface (Letunic & Bork, 2016).

Statistical Analysis

Quality filtered and trimmed reads were stringently mapped to the selected contigs (0.97 identity fraction, 0.7 length fraction) in CLC Genomics Workbench 10.0.1. Expression values were calculated as a modification of the TPM metric. Read counts were normalized by contig length in kb to determine the reads per kilobase (RPK) values for every contig within each library. These RPK values were then summed and divided by 1 million, to determine the sequencing depth scaling factor for each library. TPM for a contig was calculated by dividing its RPK value by the scaling factor for the library.

Expression values for contigs were imported into the PRIMER7 (Clark & Gorley, 2015) statistical software package and $\log_2$ transformed. Group average hierarchical clustering was performed using Pearson's correlation coefficient as the index of association. The SIMPROF test (Clarke *et al.*, 2008) was used to determine the statistical significance level of resulting clusters (alpha = 0.05, 1000 permutations). Statistically significant clusters with at least one viral contig, one *rpb*1 contig and less than 10 total members were visualized and annotated in Cytoscape 3.5.1 (Shannon *et al.*, 2003).

Results

Identification of Resident Phage Populations

To identify active bacteriophage in the peat bog metatranscriptomes, we screened contig libraries for the presence of four conserved phage genes: myovirus major capsid protein (*gp23*), phage portal protein (*gp20*), ribonucleotide reductase (*rnr*), and the lambda repressor (*recA*). Reads were mapped back onto conserved gene contigs, counted, and normalized for contig length and library size. Across all six *Sphagnum* tissue samples, 33 contigs were identified as transcripts encoding major capsid protein originating from bacteriophage, while only 6 contigs were identified from the three other marker genes. Concurrent with this, more reads were mapped to *gp23* contigs than the other marker genes combined, the most abundant of which were the three ribonucleotide reductase contigs.

To determine the phylogeny of the *gp23* contigs, we placed them on a maximum likelihood tree constructed using reference sequences downloaded from the InterPro database (Figure 3.1). Out of the 33 contigs, 18 are grouped in the *Eucampyvirinae* subfamily with *Campylobacter* viruses CP220 and PC18, while the rest are spread amongst the other Myovirus taxa, predominantly the *Tevenvirinae*. Gb0139905 contig 77559 was the most abundant, with consistently high expression across all samples, whereas other contigs dominate within one or two samples. Of the 6 contigs identified using the other 3 viral marker genes, one was identified as a potential *gp20* homologue, originating within *Myoviridae* with *Clostridium* virus phiCD119 as the closest relative (Figure 3.2). Two contigs were identified as *recA* contigs, likely originating in myovirus and siphovirus relatives (Figure 3.3), as were the remaining three contigs identified as ribonucleotide reductase transcripts (Figure 3.4).

Novel ssRNA virus diversity and abundance

To determine RNA virus diversity within the *Sphagnum* microbiome, we screened the metatranscriptome libraries for contigs with homology to positive-sense ssRNA virus RNA-dependent RNA polymerase. Contigs were placed on a reference RDRP tree to determine their phylogenetic identity (Figure 3.5). 114 contigs were identified as originating in RNA viruses, the majority of which belonged to the currently unassigned *Barnaviridae* and Astrovirus-like families. Additionally, a large number of Picornaviruses were observed, most of which were

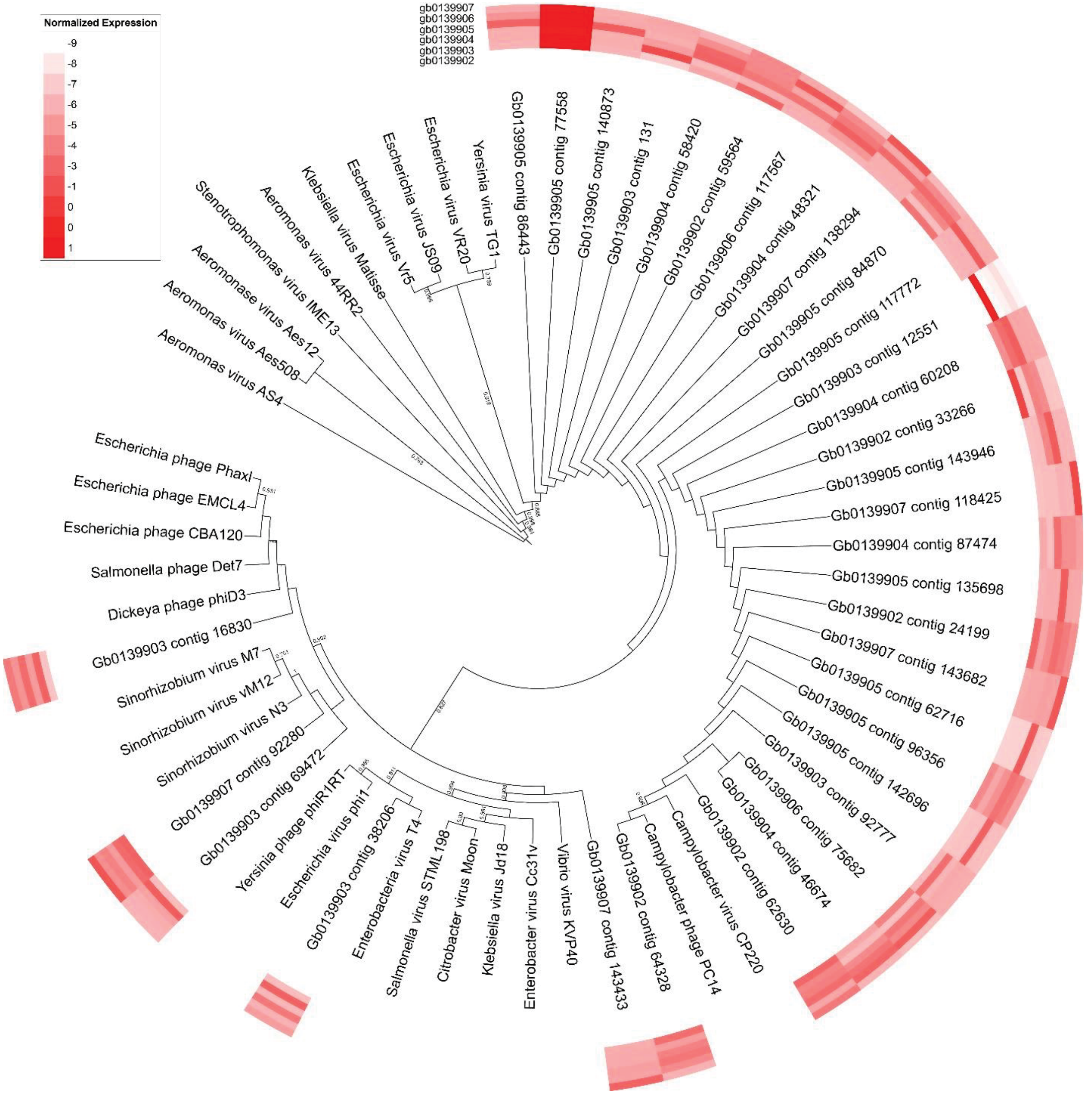

Figure 3.1. Phylogenetic placement of identified phage major capsid protein contigs on a Myovirus *gp23* maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown. Contig abundance (log$_2$ transformed TPM) within each of the six samples is shown in the heatmap surrounding the tree.

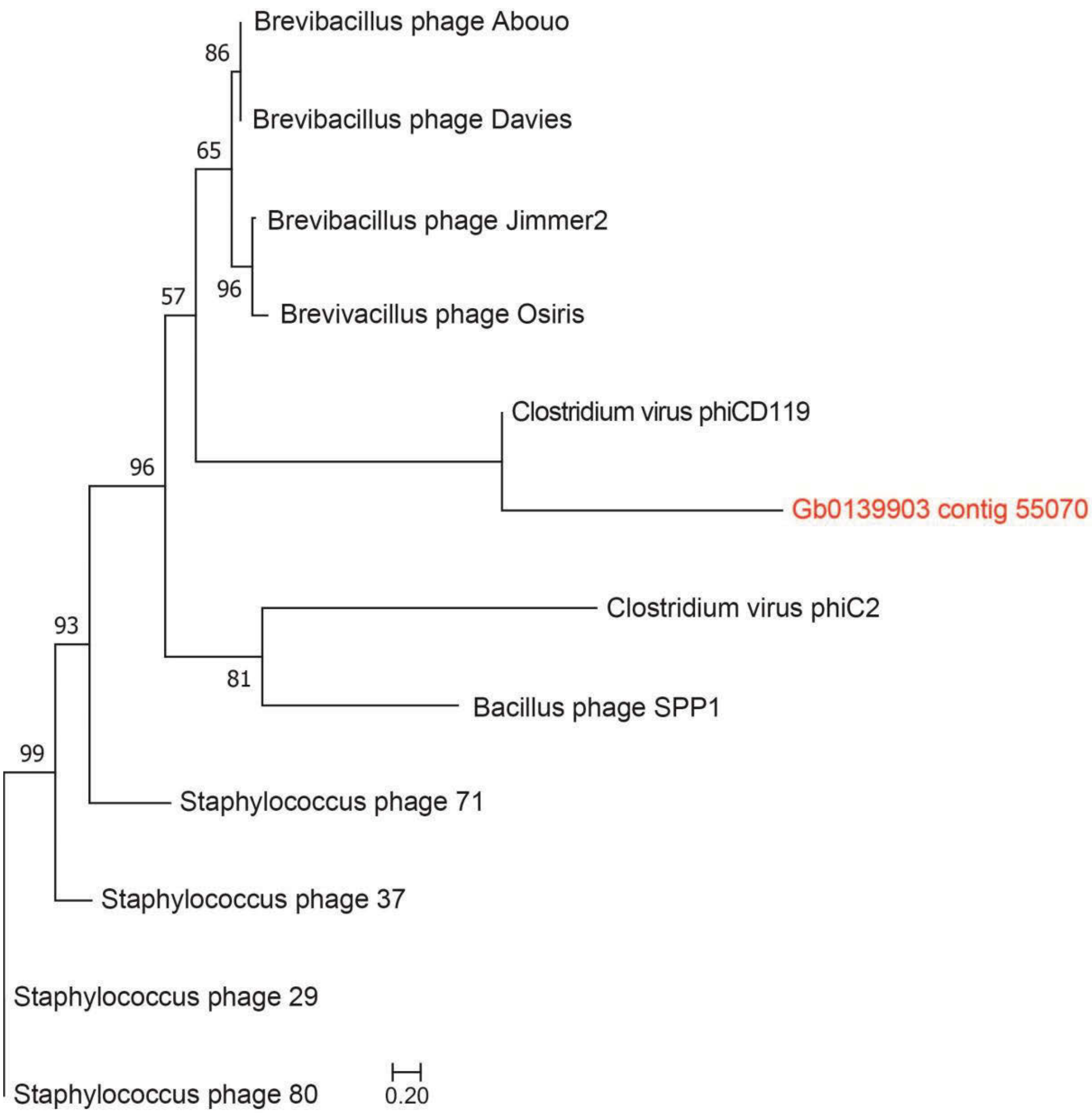

Figure 3.2. Phylogenetic placement of identified phage portal protein contigs on a *gp20* maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown.

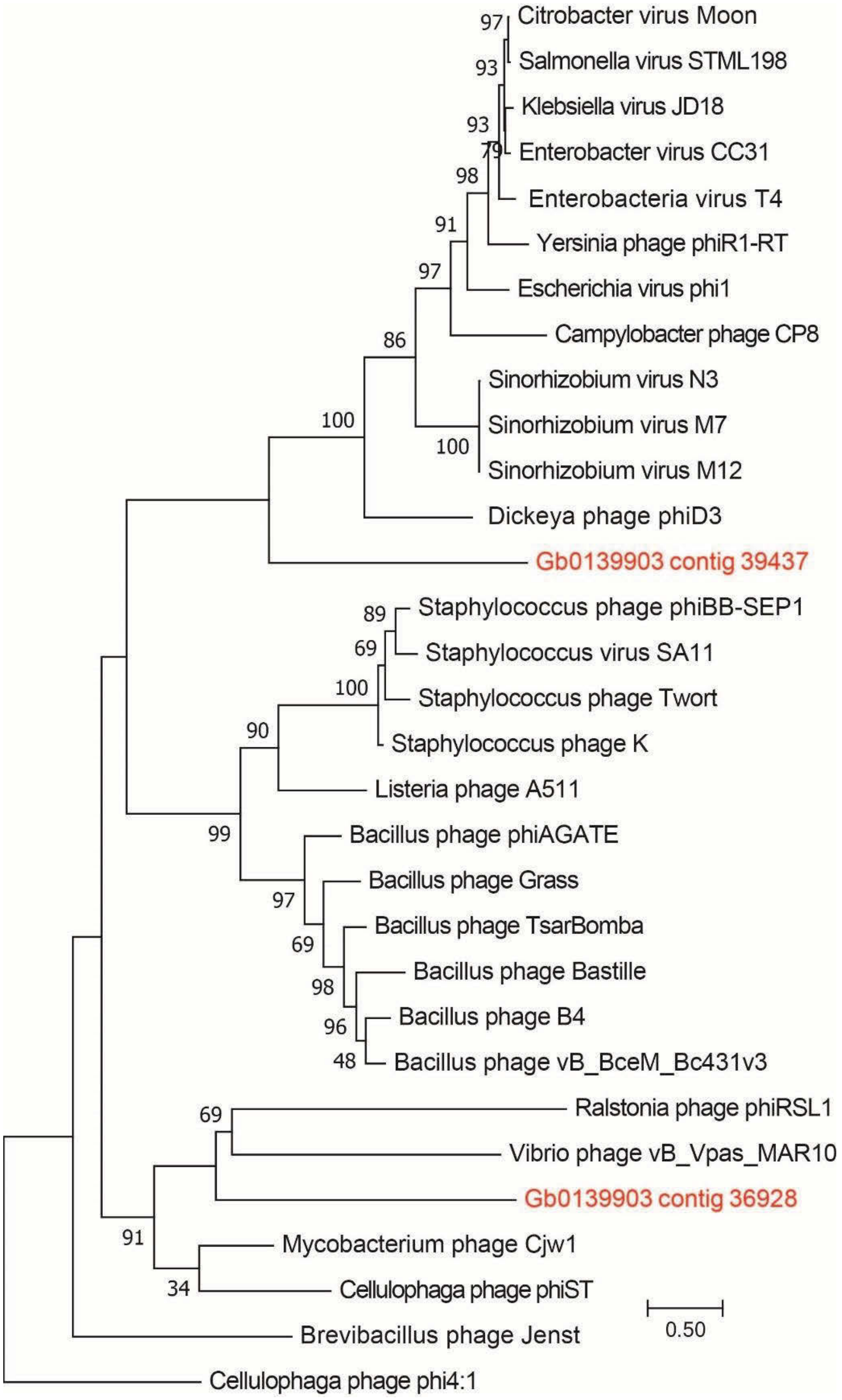

Figure 3.3. Phylogenetic placement of identified lambda repressor contigs on a RecA maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown.

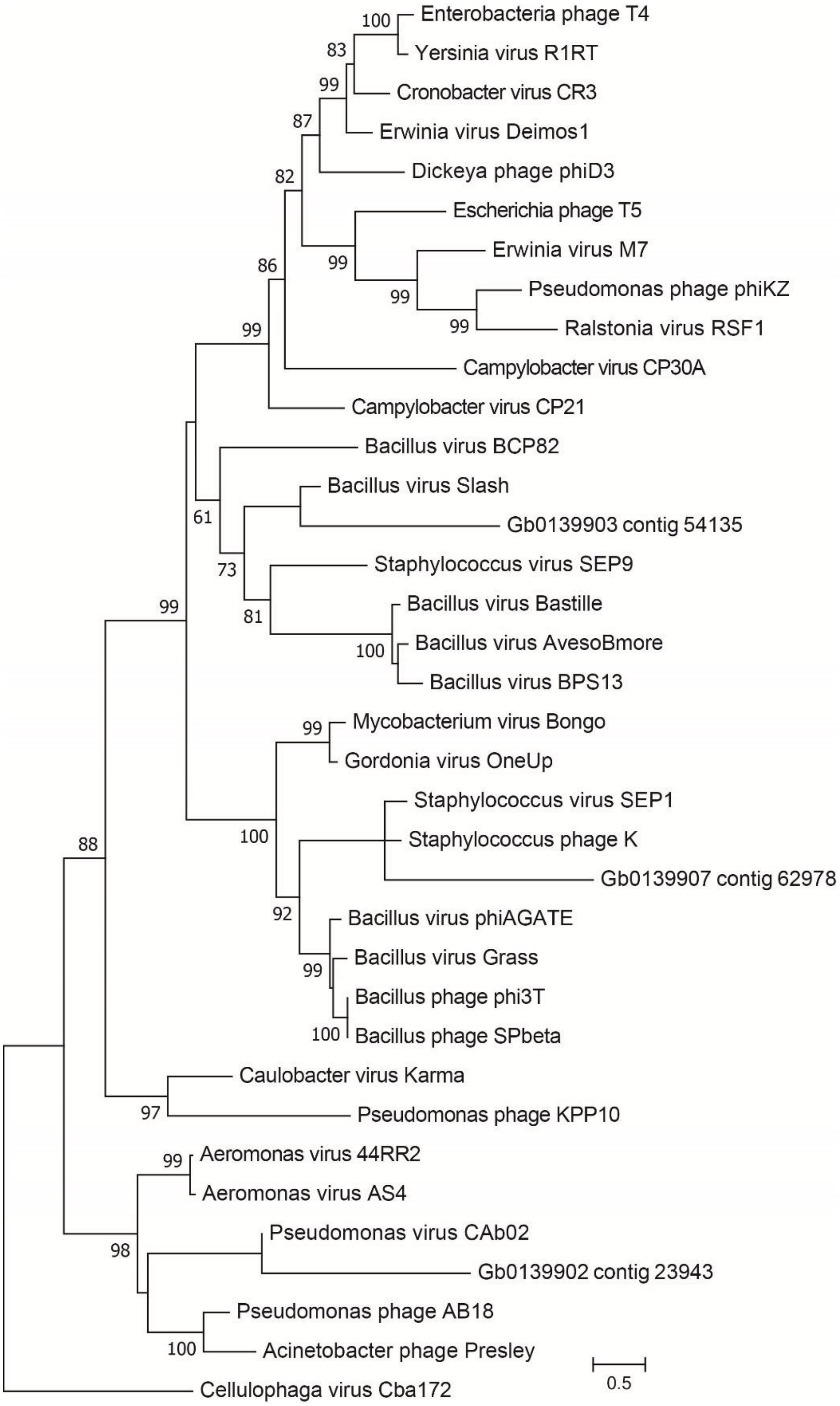

Figure 3.4. Phylogenetic placement of identified phage ribonucleotide reductase contigs on a maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown.

closely related to the unclassified marine *Aurantiochytrium* single-stranded RNA virus, and *Secoviridae* plant viruses. Lastly, several contigs were closely related to the *Nidovirales* clade, which generally infect animal species.

Among these, 22 contigs were found to be near complete ssRNA virus genomes (based on content and size), encoding multiple viral genes in addition to RDRP. Gene regions were identified and annotated using the NCBI conserved domain and PFam HMM search tools, and the full RDRP sequence was used to construct a maximum likelihood phylogenetic tree (Figure 3.6). Of the partial genomes observed, only 2 were missing conserved Rhv structural genes, and only one was missing the RNA virus Helicase. The majority of these contigs fall under the *Picornavirales* order, which also included some of the most complete viral genomes. As was observed with the shorter RDRP contigs above, most of the Picornavirus contigs were most closely related to either the unclassified marine species, or members of the *Secoviridae* clade, whose membership includes the Parsnip yellow fleck virus. A number of partial Picornavirus genomes were also identified as members of the *Dicistroviridae*. Outside the *Picornavirales*, most contigs clustered closely with the unassigned Astrovirus-like *Phytophthora infestans* RNA virus. In order to determine the relative abundance of RNA virus genomes in the peat bog samples, we mapped reads back to contigs and calculated TPM values to account for contig length and library size. The most abundant contig across all samples was Gb0139905_contig_3964, which was most closely related to the Rotifer birnavirus. All other contigs appear to be abundant prominently in one or two samples, and absent or in low abundance in the others, with no patterns of abundance apparent.

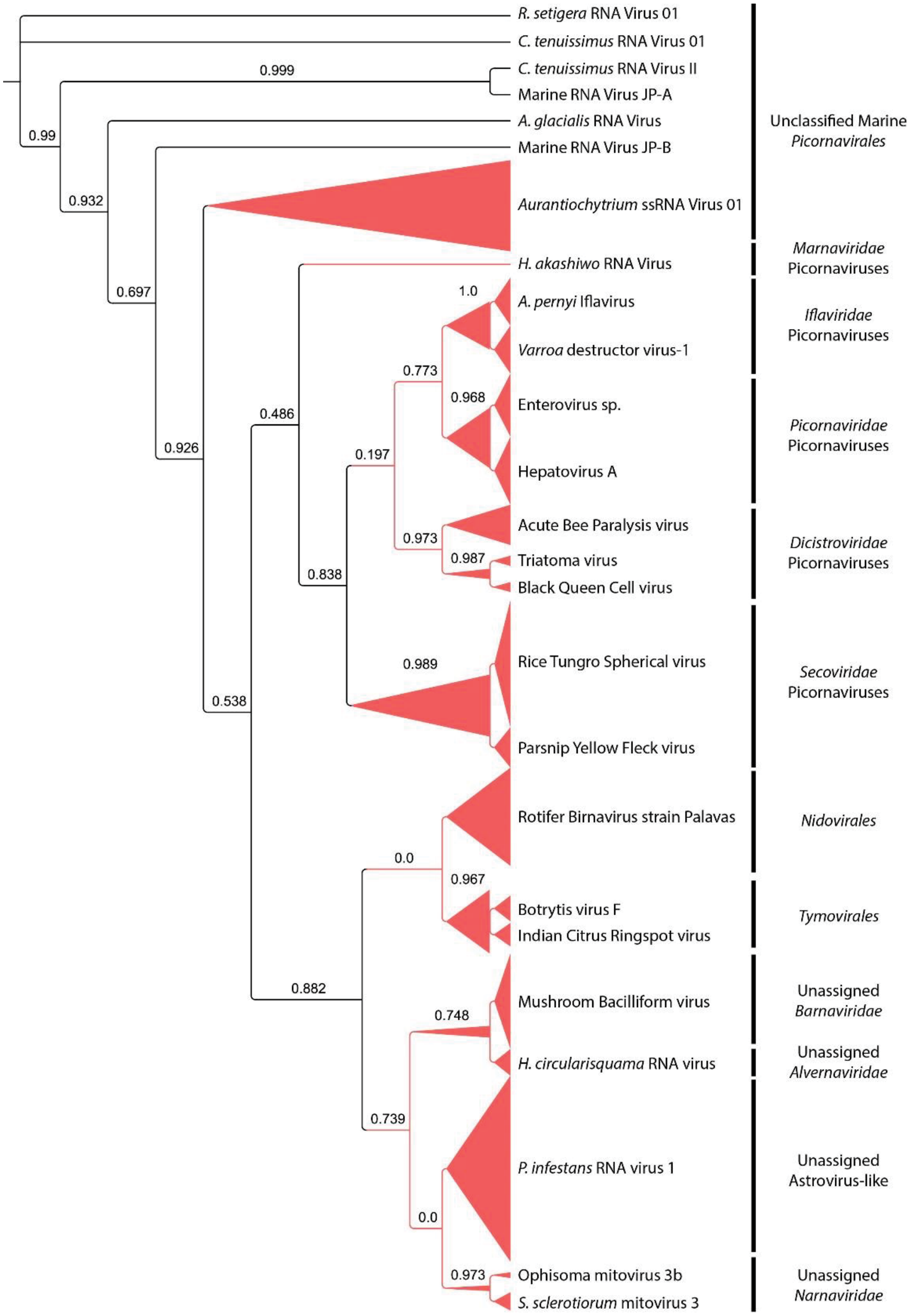
R. setigera RNA Virus 01
C. tenuissimus RNA Virus 01
0.999
C. tenuissimus RNA Virus II
Marine RNA Virus JP-A
A. glacialis RNA Virus
0.99
Marine RNA Virus JP-B
0.932
Aurantiochytrium ssRNA Virus 01
H. akashiwo RNA Virus
0.697
1.0
A. pernyi Iflavirus
Varroa destructor virus-1
0.773
0.968
Enterovirus sp.
0.486
0.926
0.197
Hepatovirus A
Acute Bee Paralysis virus
0.973
0.987
Triatoma virus
0.838
Black Queen Cell virus
0.989
Rice Tungro Spherical virus
0.538
Parsnip Yellow Fleck virus
Rotifer Birnavirus strain Palavas
0.0
0.967
Botrytis virus F
Indian Citrus Ringspot virus
0.882
Mushroom Bacilliform virus
0.748
H. circularisquama RNA virus
0.739
P. infestans RNA virus 1
0.0
0.973
Ophisoma mitovirus 3b
S. sclerotiorum mitovirus 3
Unclassified Marine Picornavirales
Marnaviridae Picornaviruses
Iflaviridae Picornaviruses
Picornaviridae Picornaviruses
Dicistroviridae Picornaviruses
Secoviridae Picornaviruses
Nidovirales
Tymovirales
Unassigned Barnaviridae
Unassigned Alvernaviridae
Unassigned Astrovirus-like
Unassigned Narnaviridae

Figure 3.5. Phylogenetic placement of identified ssRNA virus RNA-dependent RNA polymerase contigs on maximum likelihood reference tree. Branch width represents the number of contigs placed on the reference branch. Node support (aLRT-SH statistic) >50% are shown.

Activity of giant viruses in the *Sphagnum* microbiome

To identify active infections by giant viruses, we screened contigs libraries for the presence of 10 genes conserved amongst most NCVLDs. Of the 10 markers used, only the giant virus major capsid protein was detected in the metatranscriptome. 64 contigs were observed with homology to MCP, representing every known group of NCLDVs (Figure 3.7). Out of the 64 MCP contigs, 46 were placed within the *Mimiviridae* taxa. Most of these (25 contigs) are closely related to the recently discovered Klosneuviruses, with the Indivirus and Catovirus representing the most significant source of diversity in these samples. The next most abundant group were the "extended *Mimiviridae*" (7 contigs), species with known similarity to Mimiviruses that infect eukaryotic algae, and the *Asfarviridae* (6 contigs) which are primarily represented by the African swine fever Virus. Potential relatives of the giant virus outliers, Pandoravirus and Pithovirus, were not observed, and the *Iridoviriae* were poorly represented (1 contig). Using the virophage MCP and packaging ATPase as markers, we identified 7 contigs as transcripts originating in putative virophage, all of which were phylogenetically placed amongst freshwater isolates (Figure 3.8).

As described above, reads were mapped back onto contigs to determine the relative abundance of transcripts in the samples. As was observed with the other major viral taxa described, the majority of contigs were most abundantly expressed in one or two samples and present at very low levels in the rest. The most abundant MCP contig in the samples was Gb0139903 contig 73240, most closely related to *Megavirus chilensis*, which was the most highly expressed contig across all samples. Four other contigs (Gb0139907 contig 110585, Gb0139905 contigs 55722 and 141177, and Gb0139906 contig 119519) were highly expressed across all six samples.

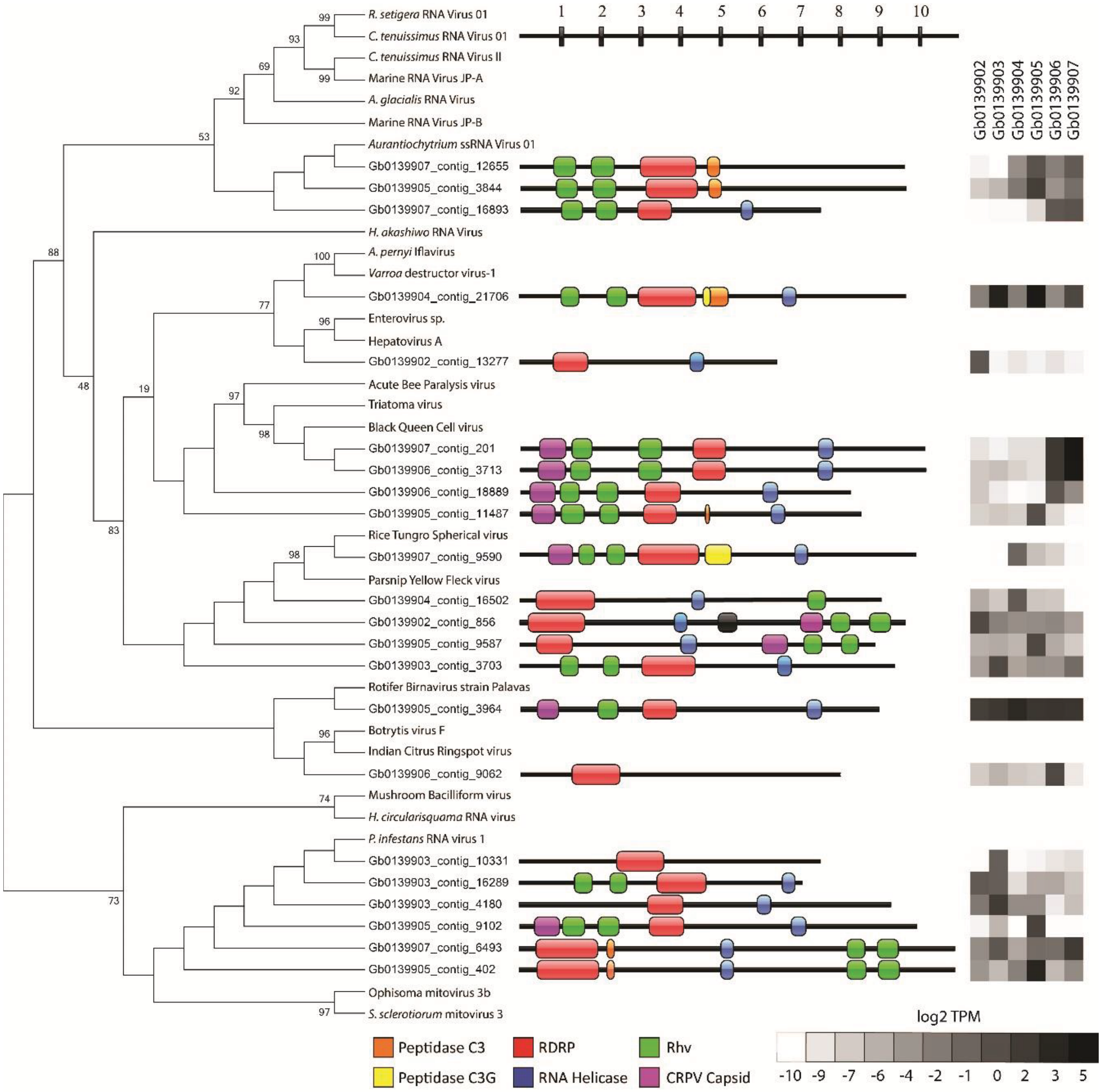

Figure 3.6. Phylogeny, genome architecture, and abundance of partial ssRNA virus genomes. Tree represents phylogenetic placement of RDRP gene regions from partial ssRNA virus genome contigs on a maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown. Center panel represents genome architecture determined by conserved domain search and ORF prediction. Length of contigs and gene regions is measured in kb. Heat map in right panel shows abundance of reads mapped to partial genome contigs in $\log_2$ TPM from each of the 6 metatranscriptome libraries.

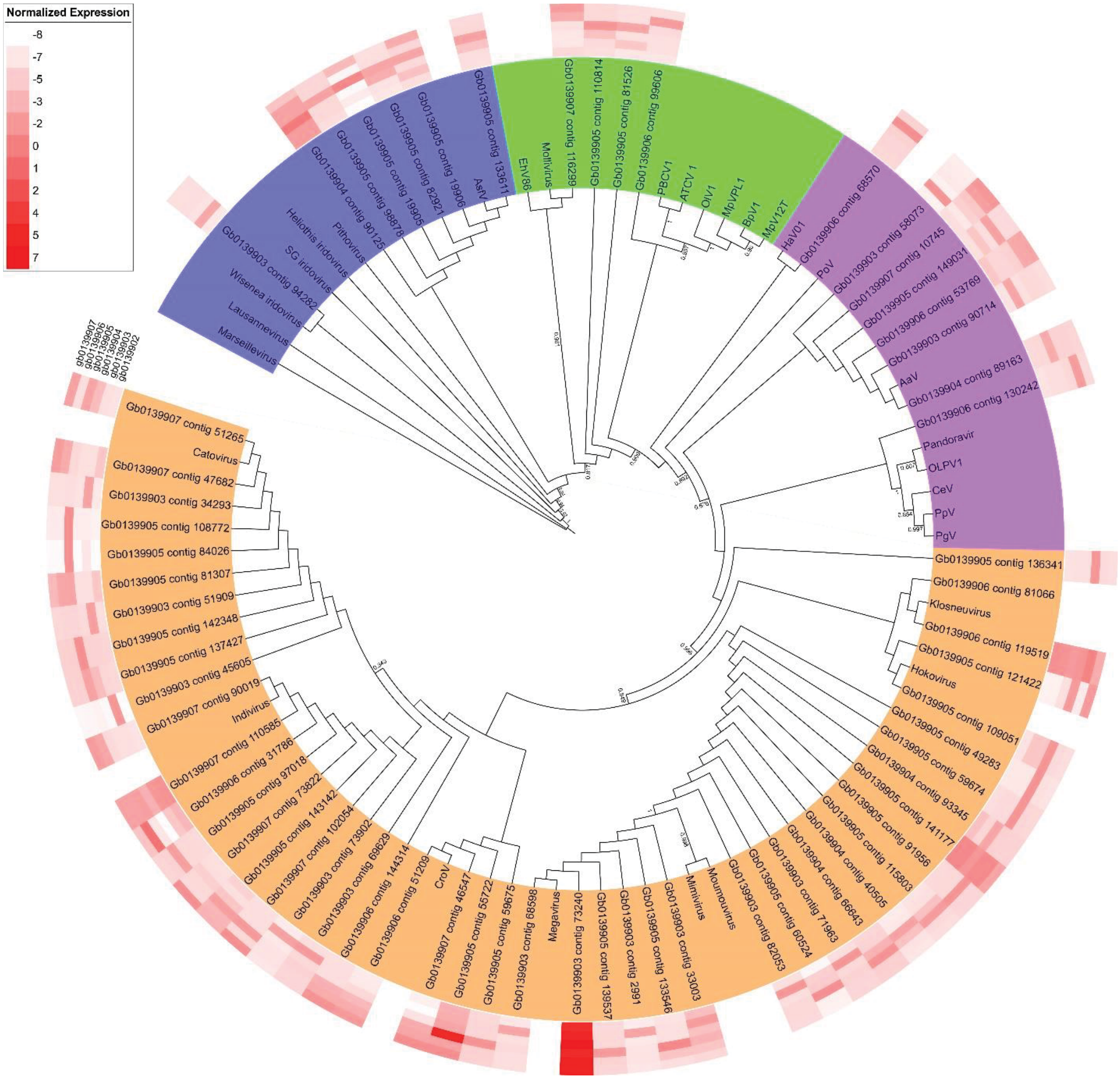

Figure 3.7. Phylogenetic placement of identified NCLDV major capsid protein contigs on a maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown. Contig abundance ($\log_2$ transformed TPM) within each of the six samples is shown in a heatmap surrounding the tree.

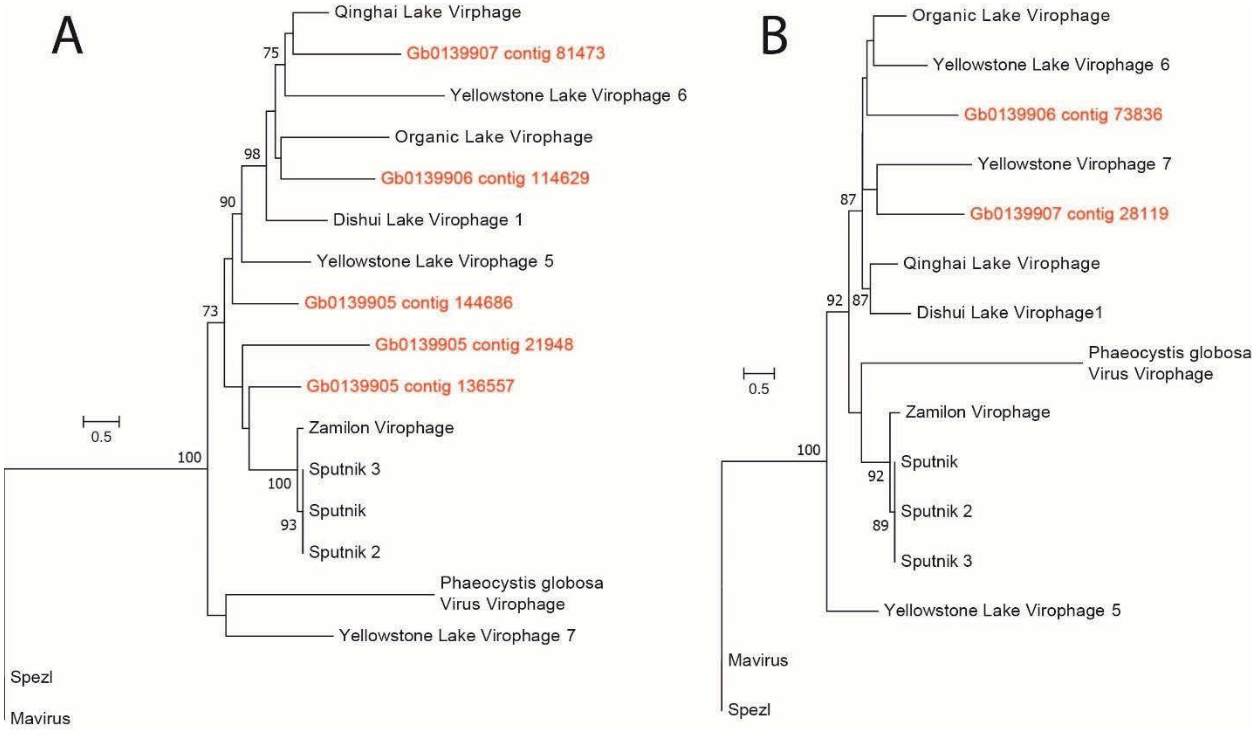

Figure 3.8. Phylogenetic placement of identified virophage A.) major capsid protein and B.) ATPase contigs on a maximum likelihood reference tree. Node support (aLRT-SH statistic) >50% are shown.

Prediction of virus-host pairs

By comparing expression of viral marker genes to *rpb1* expression by cellular organisms, we predicted potential virus-host groups in the *Sphagnum* peat bogs. Figure 3.9 shows networks containing at least one virus and one host, where co-occurrence and correlation were observed in more than one sample. A total of 11 virus-host groups were detected, spread across the major viral taxa detected in this dataset. Four relationships were predicted from bacteriophage *gp23* abundance, the simplest of which was a *Tevenvirinae* phage-*Proteobacteria* pair with a moderate correlation. The other 3 relationships are more complicated, containing multiple potential hosts and, for the largest predicted group, multiple viruses. Correlation coefficients for the phage-bacteria clusters were lower than was observed in the other major viral taxa, with low to moderate correlations between viruses and bacteria. Higher correlation values were observed, but they are restricted to bacterium-bacterium interactions.

We observed 3 predicted RNA virus-host clusters, all of which contained multiple hosts grouped with a single virus. Most of the predicted hosts are most closely related to plants, with multiple predicted animal hosts, primarily fish species. Correlation coefficients observed in these relationships are generally higher than was observed in the phage-host clusters, but with fewer clusters overall. The 4 predicted NCLDV-host clusters were the most highly correlated and the most complex. Predicted hosts are highly varied, ranging from diatoms to animals, though all virus members were placed either within *Mimiviridae* or the extended Mimivirus group. MCP contigs originating in close relatives of the recently discovered Klosneuviruses are present in both the 7- and 8-member clusters, in addition to a pair of contigs most closely related to *Aureococcus anophagefferens* Virus (AaV). An additional 15 statistically significant clusters

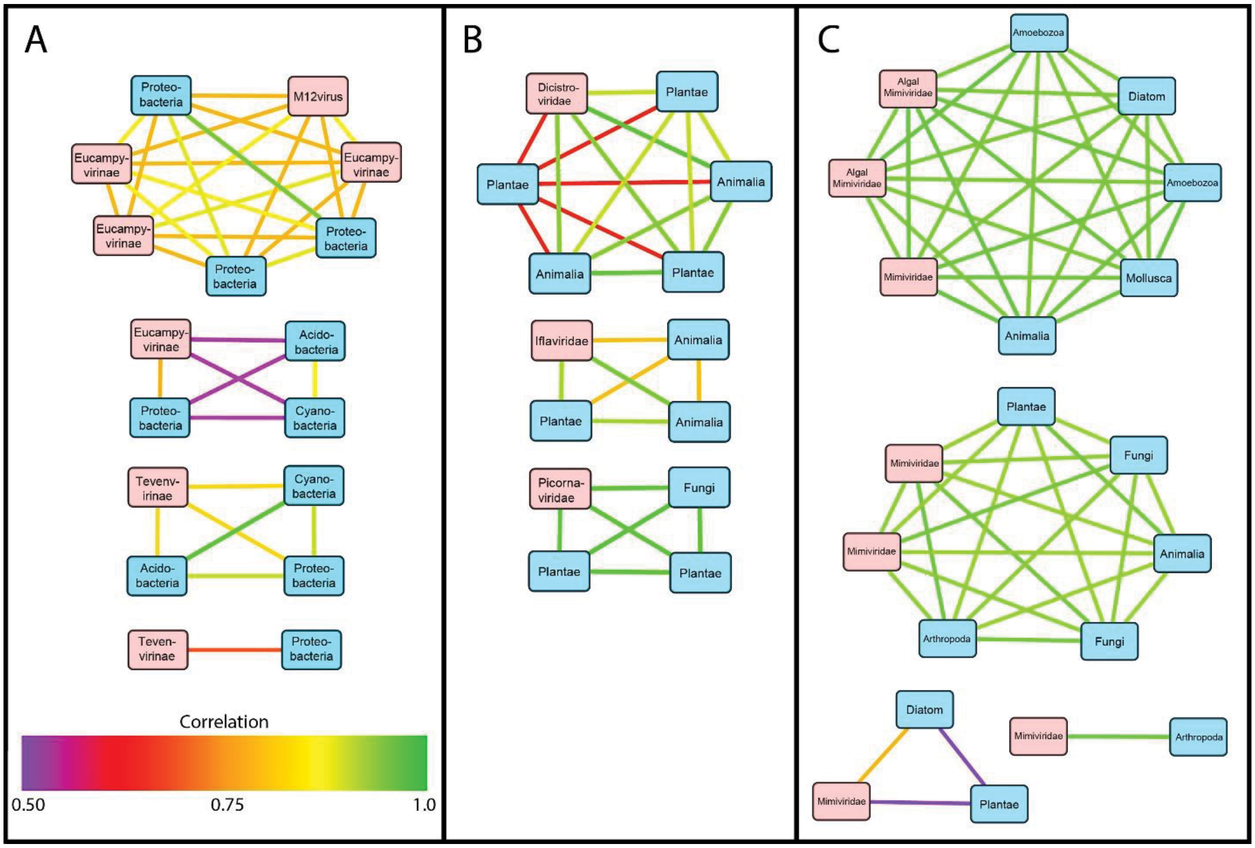

Figure 3.9. Correlation co-occurrence network analysis of conserved viral gene and host RNA polymerase (Rpb1) expression for A.) bacteriophage (Gp23), B.) ssRNA viruses (RDRP), and C.) NCLDVs (NCLDV MCP). Nodes in red represent virus contigs and blue nodes represent potential hosts. Host taxa were determined by best BLAST hit in a blastx search of the NCBI nr database. Nodes are connected by edges colored according to the Pearson correlation coefficient values between to contigs. Only relationships with contigs expressed in more than one sample are shown.

across all three viral taxa were observed where the virus and host were present in only one sample (not shown).

Discussion

In this study, we used community metatranscriptome data from *Sphagnum* tissue to describe the diversity and activity of the resident virus populations. We identified a considerable volume of novel viruses from multiple taxa, most of which are poorly represented in the literature and reference sequence databases. We used read mapping to quantify the relative abundance of viral transcripts and describe viral infections active when samples were taken. Lastly, we compared expression of viral transcripts to that of potential hosts, using correlation co-occurrence networks to predict potential hosts for the observed novel virus populations. Together, our results suggest that *Sphagnum* peat bogs represent a significant and largely untapped source of viral diversity. Viruses were highly active across all samples, some with individual viruses exhibiting abundant activity in single locations while others were more widespread. Our observations were based on RNA sequencing data, and thus most certainly do not represent a full accounting of the viruses present in the community. Metatranscriptomic data however allows us to distinguish active virus populations at the time of sampling. In addition, as viruses only transcribe their genes during infection, virus and host transcripts are expected to co-occur, and it is possible that the abundance of transcripts could be used to predict natural hosts of viruses observed in the ecosystem which can be tested in a laboratory or field setting. Ultimately, this study identifies from within a complex community a number of candidate virus-host model systems for future study.

Viral diversity and activity in *Sphagnum* peat

Out of the 10 genes used to screen the metatranscriptomes for giant virus sequences, only MCP transcripts were detected, suggesting that other genes were likely expressed at much lower levels. This is not surprising given the number of capsid proteins needed for viral replication, and indeed this transcriptional pattern was previously observed in marine systems by Moniruzzaman *et al.* (2017). Additionally, the RNA-seq dataset used in that study was poly-A selected, enriching for eukaryotic transcripts, and thus coverage of eukaryotic virus gene expression would be much higher than in the *Sphagnum* metatranscriptome. That we observed MCP expression in abundance suggests a significant number of infections occurred at the time of sampling. Another point to note is the diversity; 64 distinct MCP genotypes were observed, which is incredibly high when compared to one recent survey that identified 30 novel MCP transcripts from multiple environmental datasets (Wilhelm *et al.*, 2016), and another which observed 107 NCLDV sequences in 16 publicly available environmental metagenomes (Kerepesi & Grolmusz, 2017). Most of the MCP contigs identified here were placed in clusters around a small number of virus relatives, highlighting the under-sampled diversity of giant viruses in the literature, poor representation in reference databases, and the considerable diversity present in *Sphagnum* peat bogs. Similarly, a broad range of virophage transcripts were detected, indicating a significant response to infections by giant viruses in the system. Virophage are even more poorly represented in reference material, such that every available reference sequence is shown in Figure 3.9 and the novel sequences described here expand known representatives by more than half. All of the virophage observed in the *Sphagnum* microbiome are phylogenetically placed amongst freshwater isolates, and those that have been cultured infect the Mimiviruses and members of the extended *Mimiviridae*.

As was observed with the giant viruses, most RNA virus contigs were placed in clusters with a single represented species, suggesting a significant degree of uncharacterized diversity. This is not entirely surprising, as RNA viruses are expected to make up as much as half of the virus particles in the Earth's oceans, and yet they are almost as poorly understood and represented in sequencing databases as giant viruses (Steward *et al.*, 2013). This is in large part due to difficulty in detecting and quantifying the very small RNA virus particles (Tomaru & Nagasaki, 2007, Miranda *et al.*, 2016), and the inability to detect their genetic material in metagenomes (Steward *et al.*, 2013). As such, recent attempts to use metatranscriptomes to describe environmental RNA viruses have proven successful, not only identifying marker gene fragments in datasets, but assembling complete and near-complete genomes (Miranda *et al.*, 2016, Moniruzzaman *et al.*, 2017). Similarly, we were able to assemble and identify 22 near-complete RNA virus genomes, where completeness is determined primarily by size and the presence of the 6 core genes. As there are currently only 265 sequenced genomes within the *Picornavirales*, most of which grouped within the *Picornaviridae*, this represents a sizeable increase in the known diversity of ssRNA viruses, especially within the unassigned and unclassified taxa.

Given the importance of bacteria in the *Sphagnum* microbiome (Kostka *et al.*, 2016), the relatively low abundance of active bacteriophage in our samples was a surprise. Marker genes for identifying bacteriophage were chosen based on their conservation across phage taxa and their success in other environmental datasets. Gp20 (phage portal protein) and Gp23 (major capsid protein) have been shown previously to be highly conserved and effective for phylogenetic assignment of members of the *Myoviridae* (Dorigo *et al.*, 2004, Comeau & Krisch, 2008, Roux *et al.*, 2012). RecA (lambda repressor) is conserved across all three bacteriophage taxa and could illuminate lysogeny, and ribonucleotide reductase (RNR) has been used as an effective marker for screening novel viruses from marine sequencing datasets (Sakowski *et al.*, 2014). As such, we only identified 39 bacteriophage contigs using these markers, 33 of which were from Gp23. This may represent a similar pattern to the giant viruses above, where transcripts encoding structural proteins are much more abundant than other genes and sequencing lacked the depth to detect them. For the purpose of discovering novel phage species, DNA sequencing through metagenomics may prove more successful.

Most virus transcripts exhibited high abundance in one or two *Sphagnum* bog samples while low (or absent) in others, suggesting potential "boom-bust" infection dynamics previously observed in algal giant viruses (Short, 2012). In contrast, 5 MCP contigs and 2 RNA virus contigs were similarly expressed across all six samples at similar levels, consistent with a slow and persistent rate of infection observed in marine systems (Moniruzzaman *et al.*, 2017), but not yet described in freshwater or terrestrial ecosystems.

Virus-host predictions

As viruses produce transcripts only when actively infecting a host, positive correlation and co-occurrence between virus and host transcripts is expected, and might be used to predict host-virus relationships, provided an appropriate transcriptional proxy for growth and activity is available (Moniruzzaman *et al.*, 2017). In this study, we used the eukaryotic RNA-polymerase gene *rpb1* as a marker for abundance and activity in potential hosts, as it has been previously described as one of the more consistently expressed eukaryotic genes in marine systems, scaling well with the activity of the organism (Alexander *et al.*, 2015). We used NCLDV MCP abundance as a proxy for giant virus production, Gp23 for phage production, and RdRP for RNA virus production, as transcription is necessary for the assembly of new virus particles and transcript abundance in some appears to be closely linked to viral replication (Moniruzzaman *et al.*, 2017).

Correlation and co-occurrence matrices, clustered into groups by similarity and tested with the SIMPROF permutation test yielded 11 predicted groups of viruses and hosts. In giant viruses, several of the initial networks produced in the analysis included multiple bacterial species picked up in the RNA polymerase screen. As we have no reason to believe bacterial species are infected by NCLDVs, they were removed from the final virus-host groups. It is likely these predictions represent a confounding relationship between prokaryotes and potential eukaryotic hosts, observed in network analyses for all three viral taxa described here, where a beneficial interaction results in an indirect correlation with viral infection. Indeed, previous use of this method in marine systems showed a similar phenomenon, where an algal Mimivirus and a known host were grouped with a fungal species and another virus (Moniruzzaman *et al.*, 2017). Even after the removal of bacterial species from the predicted groups, some remain complicated with multiple viruses and potential hosts, which may be explained by a broader host range amongst giant viruses enabled by the expansion of genetic material and increased independence from host machinery. Similar complicated relationships were observed amongst RNA viruses, though these are more tenuous, as we are unable to distinguish whether sequencing reads originated transcripts or genomic material.

All together, we have identified a considerable amount of viral diversity from several major viral taxa active within a poorly understood microbial ecosystem. As they were identified from transcript sequencing data, the viruses described here likely only represent a fraction of the whole virus community, which may be elucidated through further culture-independent work. We have also used transcript abundance within a statistical framework to predict several host-virus relationships which can be sought out and tested in culture. These results establish an important and much needed foundation for future research into the microbial ecology in *Sphagnum* peat bogs.

Acknowledgments

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. Support for the SPRUCE experimental site is from the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. Oak Ridge National Laboratory is managed by UT- Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05- 00OR22725.

Chapter III Appendix

Table 3.1. NCLDV most highly conserved clusters of orthologous genes used to screen transcript libraries.

NCVOG	Product
NCVOG0249	A32 virion packagaing ATPase
NCVOG0262	VLFT-like transcription factor
NCVOG0024	Superfamily II Helicase II
NCVOG1117	mRNA capping enzyme
NCVOG0023	D5 helicase-primase
NCVOG0276	ribonucleotide reductase small subunit
NCVOG0271	RNA polymerase large subunit
NCVOG0274	RNA polymerase small subunit
NCVOG0038	B-family DNA polymerase
NCVOG0022	Major Capsid Protein

Table 3.2. Conserved RNA virus genes and their respective Pfam designations.

Gene	Pfam
CRPV capsid	08762
VP4	11492
RNA-Dependent RNA Polymerase	00680
Peptidase C3	00548
Rhv	00073
RNA Helicase	00910
Peptidase C3G	12381

CHAPTER IV:

GENOME OF *CHRYSOCHROMULINA PARVA* VIRUS AND ITS CONSTITUENT VIROPHAGE

Publication Note

This chapter is a version of an article in preparation for submission to *Frontiers in Microbiology* by Joshua M.A. Stough, Yuri Chaban, Morgan M. Steffen, Mohammed Moniruzzaman, Eric R. Gann, Steven M. Short, and Steven W. Wilhelm.

My contribution to this work was data management and analysis, statistical analysis, and primary authorship and editing of the manuscript.

Abstract

Giant viruses represent an increasingly important viral clade that is frequently involved in the top-down control of single-celled eukaryotic algae populations in marine ecosystems across the globe. Despite increased interest in giant viruses with the discovery of Mimivirus, little is known about their physiology and ecology. In this study, we characterize the genome and functional potential of a virus capable of infecting the freshwater haptophyte *Chrysochromulina parva*, originally isolated from Lake Ontario. CpV is a member of the nucleocytoplasmic large DNA virus (NCLDV) group, and possesses a 437 kb genome encoding 503 ORFs with a GC content of 25%. Phylogenetic analysis of core functional genes places CpV amongst an emerging group of algae-infecting Mimiviruses informally referred to as the "extended *Mimiviridae*", making it the first to be isolated from freshwater ecosystems. During sequencing, we also captured and described the 22.7 kb genome of a virophage that appears to "infect" and exploit the activity of CpV to replicate. The virophage genome encodes 19 predicted ORFs, including all of the currently described core genes necessary for function, as well as several genes implied in genetic modification. Lastly, we used the obtained CpV and virophage reference sequence to recruit reads from available environmental metatranscriptomic data in order to estimate their activity in freshwater ecosystems. We observed moderate levels of virus and virophage transcript abundance in samples obtained during the *Microcystis aeruginosa* bloom in Lake Tai, China in 2013, with a spike in activity in one sample. In all, these results highlight the importance of giant viruses in the environment, and establish a foundation for future research on the physiology and ecology CpV as a model system for algal mimivirus dynamics in freshwater ecosystems.

Introduction

Viruses are the most abundant biological particles on Earth, and play important roles in global ecosystems by driving the evolution of their hosts and biogeochemical cycling (Brussaard *et al.*, 2008). The vast majority of currently described viruses are smaller than 200 nm, with genomes encoding the minimal functions necessary for replication and evasion of host defenses. As such, it came as a great surprise when the Mimiviruses were discovered infecting *Acanthamoeba* species, whose size and complexity rival that of many bacteria (Raoult *et al.*, 2004). Radically different from the conventional model of viruses, the members of this novel lineage possess hundreds of genes, many of which are responsible for functions previously only found in cellular life (Filee *et al.*, 2008), including translational machinery and auxiliary metabolic functions. Together, these unusual features allow giant viruses to replicate largely independent of host machinery, blurring traditional boundaries between cellular life and viruses (Yutin *et al.*, 2009, Claverie & Abergel, 2010). Indeed, while thought to originally be bacteria, Mimiviruses were shown to be physically larger members of a more diverse viral group, generally referred to as the Nucleocytoplasmic Large DNA Viruses (NCLDVs). Despite dramatic variety in genome content and preferred hosts, giant viruses appear to share a common evolutionary ancestor, sparking heated debate over whether this family constitutes a new domain of life (Boyer *et al.*, 2010, Yutin *et al.*, 2014). The study of giant viruses has also led to the discovery of virophage, small viruses that rely on NCLDV machinery to replicate, usually at the expense of the “host” giant virus (La Scola *et al.*, 2008, Gaia *et al.*, 2014).

While the excitement over giant viruses has led to a considerable amount of speculation and debate over their origin and the definition of life, relatively few have been isolated and described. Only a handful of giant viruses have sequenced genomes, making proper analysis of evolutionary history and impact on global ecosystems challenging. These difficulties are compounded by the lack of comparable genetic references in public databases, leaving as much as 93% of the genes unidentifiable (Philippe *et al.*, 2013). As giant viruses have been implicated both in human disease (Popgeorgiev *et al.*, 2013, Yolken *et al.*, 2014) and the collapse of harmful algal blooms (Gastrich *et al.*, 2004), there is a desperate need for further expansion of known physiology and diversity. Indeed, some researchers have attempted to address this using the *Acanthamoeba* host as bait to isolate novel strains from multiple environmental systems with some success (Boyer *et al.*, 2009, Legendre *et al.*, 2014, Legendre *et al.*, 2015). However, it has been hypothesized that the high degree of independence from host cell machinery, combined with phagocytosis as the mechanism of entry, may permit giants to infect a much broader range of hosts than other viruses (Koonin & Yutin, 2010). Concurrent with this, a sub-family of Mimiviruses is emerging, often called the extended *Mimiviridae*, whose members infect single-celled eukaryotic algae but are phylogenetically distinct from the *Phycodnaviridae* (Moniruzzaman *et al.*, 2014). As such, this novel taxon may represent a source of viral diversity that is not only evolutionarily informative for the study of giants, but also environmentally relevant. Recently, a giant virus was isolated infecting the freshwater algae *Chrysochromulina parva* in Lake Ontario. Initial sequence analysis of conserved NCLDV genes amplified from culture indicated a close phylogenetic relationship with the extended *Mimiviridae*, primarily *Phaeocystis globosa* Virus (PgV) and *Chrysochromulina ericina* Virus (CeV), making it the first algal Mimivirus described in freshwater ecosystems (Mirza *et al.*, 2015).

In this study, we sequenced and characterized the genome of the *Chrysochromulina parva* Virus (CpV). During sequencing, we also captured and characterized the genome of a putative virophage that we predict exploits the infection cycle of CpV to replicate. We used genomic sequences obtained here to screen publicly available metatranscriptomic datasets isolated from freshwater ecosystems in multiple locations and time points for the presence and activity of CpV, observing significant activity in Lake Tai, China (*Taihu* in Chinese). Given the virus was originally isolated in Lake Ontario, this suggests that close relatives within the algae-infecting *Mimiviridae* may be globally distributed and active players in freshwater ecosystems. This study further establishes a foundation for future research with CpV, which may serve as a useful model system for freshwater algal Mimiviruses.

Materials and Methods

CpV propagation and purification

Viruses infecting the Prymnesiophyte algae *Chrysochromulina parva* CCMP 291 were originally isolated in 2011 (Mirza et al, 2015) and have been maintained in the laboratory. To produce virus genomic material for sequencing, CpV lysates were generated from a series of 150 and 500 mL mid-log phase *C. parva* batch cultures grown at a constant temperature of 15°C, with a12 h light-dark cycle at approximately 23 µE m^{-2} s^{-1} in DY-V medium (Anderson, 2005). The resulting lysates were filtered through 47 mm diameter, 0.50 µm pore-size borosilicate glass microfiber Advantec® filters (Life Science Products, Inc.) followed by filtration through 47 mm diameter, 0.22 µm pore-size PVDF Durapore® membranes (EMD Millipore). The filtered lysates were then concentrated approximately 200-fold via ultracentrifugation using a SW32Ti rotor (Beckman Coulter) as previously described (Short *et al.*, 2011). After ultracentrifugation, the pelleted material was resuspended in 10 mM Tris-Cl (pH 8.5), pooled, and stored at 4°C.

Filtered and concentrated lysates were further purified using Optiprep™ (Iodixanol, MilliporeSigma Canada Co.) step gradients. Four-step gradients were created using Optiprep™ solutions diluted in ultrapure H_2O to final concentrations of 40%, 35%, 30%, and 25% v/v, whereby 2.50 mL of each step was bottom loaded in 13.2 mL Ultra-Clear™ ultracentrifuge tubes (Beckman Coulter Canada, LP) starting with the 25% solution and ending with the 40 % solution following Moniruzzaman et al. (2014). Three mL of concentrated lysate was then loaded on the top of the gradient which were ultracentrifuged in a SW40Ti rotor (Beckman Coulter) for 14 h 45 min at 39,000 rpm. Following ultracentrifugation, visible bands formed approximately one-third of the distance from the top of the tube, and 1.50 mL of this band and immediately surrounding gradient medium was collected by aspiration and stored at 4°C.

CpV DNA extraction and purification

Nucleic acids were extracted from gradient-purified bands using a QIAamp® MinElute® Virus Spin Kit (Qiagen) following the manufacturer's recommendations with the following modifications: each MinElute column was loaded with lysed material twice, and 50 µL of Buffer AVE (RNase-free water with 0.04 % Sodium azide) was used during each elution step. To further concentrate purified genomic DNA, ethanol precipitation was conducted by mixing pooled, extracted DNA with 0.1 x volume of 3M NaOAc and 3 x volume absolute ethanol followed by incubation at -20 °C overnight. Precipitated DNA was then collected by centrifugation for 1 h at 14,000 x g at 4°C, the supernatant was decanted, and the DNA pellet was washed twice with ice-cold 70% ethanol. After being left to dry at room temperature, the DNA pellet was resuspended with MilliQ H_2O, and was stored at -20 °C. DNA concentration was quantified using an Invitrogen® Qubit® 3.0 Fluorometer and dsDNA HS Assay kit (Thermo Fisher Scientific). In total 100 µL of DNA at a concentration of approximately 5 ng μL^{-1} was submitted to HudsonAlpha Institute for Biotechnology for sequencing.

Quality control and assembly

Raw sequences were imported into the CLC Genomics Workbench v. 10.0.1 (QIAGEN, Hilden, Germany) and processed for quality control. Reads below 0.03 quality score cutoff were removed from subsequent analyses, and the remaining reads were trimmed of any ambiguous and low quality 5' bases and only reads at the full length were retained for assembly. Quality controlled reads were then assembled using the SPAdes 3.10.1 assembler with 9 iterative kmer assemblies (kmers 21, 33, 55, 65, 77, 85, 99, 113, 127) and the careful option turned on for contig correction. Scaffolds with length >5000 bp were then imported into CLC Genomics workbench for contig quality assessment and analysis. Quality controlled reads were mapped onto scaffolds with high stringency (> 0.7 length fraction, > 0.97 similarity fraction) to determine coverage. In order to reduce scaffolds to only those viral in origin, scaffold libraries were aligned to the NCBI 16S rRNA database and hits were removed from future analyses. The remaining scaffolds were blasted against a protein database containing sequences from all currently sequenced giant virus genomes. Open reading frames were predicted using CLC Genomics workbench and coding sequences were imported into BLAST2GO for functional identification.

Phylogenetic Analysis

Reference amino acid sequences for viral marker genes were downloaded from the NCBI refseq database. These reference sequences were aligned with CpV coding sequence translations using the MUSCLE alignment algorithm (Edgar, 2004) in the MEGA v7.0.26 software package (Kumar *et al.*, 2016). Maximum likelihood phylogenetic trees were constructed in PhyML (Guindon *et al.*, 2010) with the LG substitution model and the aLRT SH-like likelihood method. Trees were edited in MEGA. Whole genome alignments were performed using the BLAST Ring Image Generator (BRIG) (Alikhan *et al.*, 2011).

Environmental Quantification and Statistical Analysis

Quality filtered and trimmed reads were mapped to the CpV (0.8 identity fraction, 0.5 length fraction) and CpVV (0.7 identity fraction, 0.5 length fraction) genomes in CLC Genomics Workbench 10.0.1. Expression values were normalized per million reads within each library. Expression values were imported into the R statistical software package (R Core Team, 2015) and correlations were calculated using the hmisc package (Harrell Jr., 2016). Data were visualized in SigmaPlot v.12.5 (Systat Software, Inc.).

Results

Assembly and annotation of CpV

Sequencing yielded 26,745,770 reads for genome assembly, which was reduced to 26,729,526 after quality control. The genome of *Chrysochromulina parva* Virus was stringently assembled using SPAdes with read correction and post-assembly scaffold checking. The result was 1099 scaffolds over 5000 bp in length, which were screened for the presence of NCLDV core genes. The best result was a 437,255bp scaffold with an average coverage of 127.44 and a GC content of 25%, encoding 503 predicted open reading frames. In functional and taxonomic assignments of predicted ORFs (Figure 4.1), more than half had top BLAST hits to NCLDV genes, the vast majority of which were from *Phaeocystis globosa* Virus (PgV), although a few showed similarity to members of the *Phycodnaviridae* and the recently discovered Hokovirus (Schulz *et al.*, 2017). The remaining genes with taxonomic assignments were split primarily amongst eukaryotes and bacteria, with a handful originating in viruses and virophage, and 165 genes with no BLAST hits. Phylogenetic analysis of 3 core NCLDV genes (Figure 4.2) yielded similar results, with PgV identified as the closest relative for all three. The remaining 7 of the ten predicted "core" NCLDV genes (Supplemental Material) also exhibited similar phylogeny with only two exceptions. The CpV RNA polymerase small subunit was most closely related to

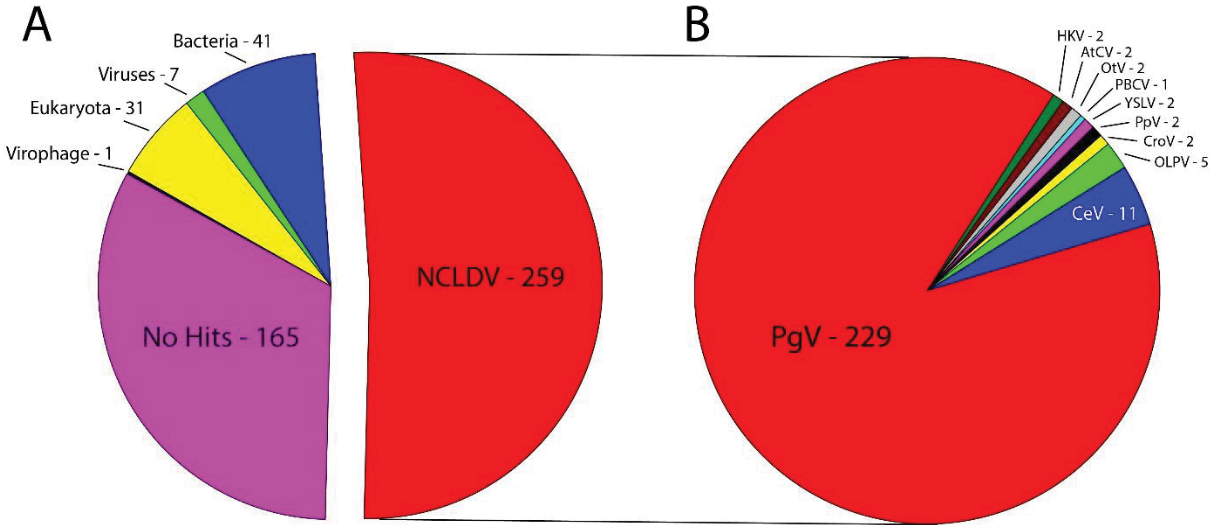

Figure 4.1. Best BLAST hits of CpV predicted open reading frames against A.) NCLDVs, other viruses, virophage, and the three domains of life. B.) Specific NCLDV representatives. PgV – *Phaeocystis globosa* Virus, CeV – *Chrysochromulina ericina* Virus, OLPV – Organic Lake Phycodnavirus, CroV – *Cafeteria roenbergensis* Virus, PpV – *Pyramimonas pouchettii* Virus, YSLV – Yellowstone Lake Phycodnavirus, PBCV – *Paramecium bursaria Chlorella* Virus, OtV – *Ostreococcus tauri* Virus, AtCV – *Acanthamoeba turfacea Chlorella* Virus, HKV – Hokovirus.

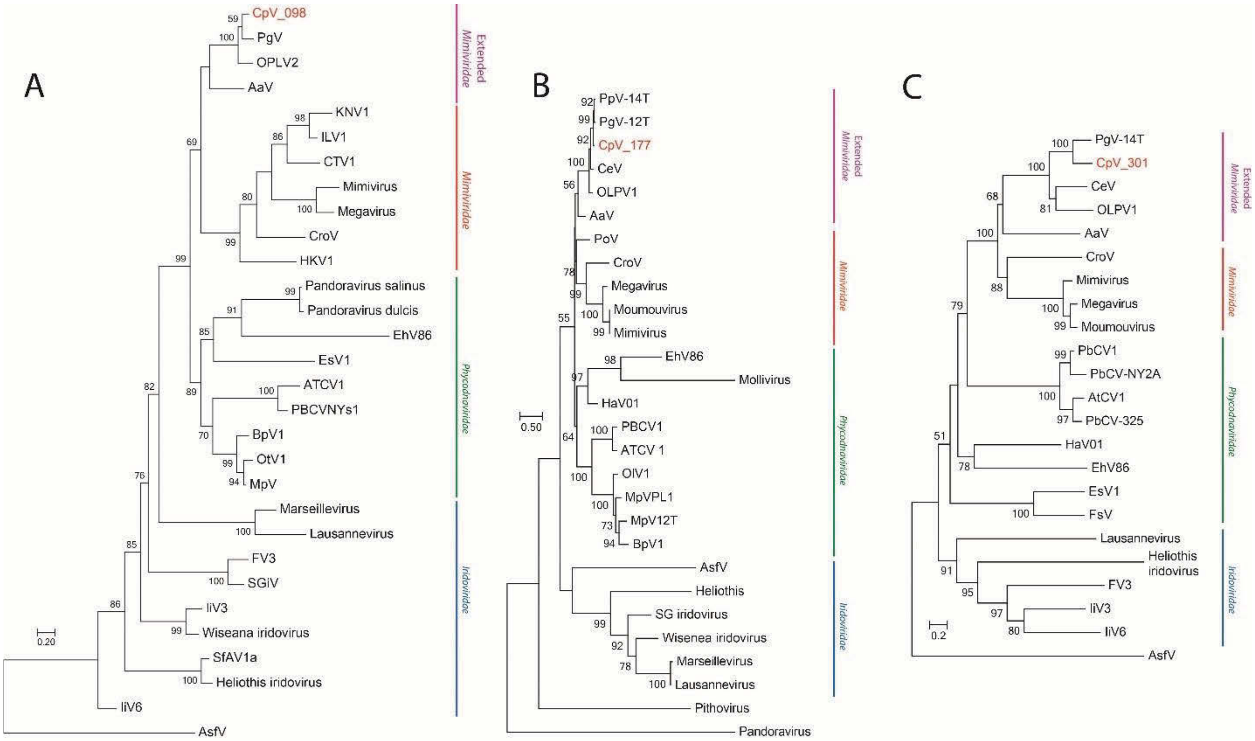

Figure 4.2. Maximum-likelihood phylogenetic trees of A.) A32-like virion packaging ATPase, B. Major capsid protein, and C.) B-family DNA polymerase. Node support (aLRT-SH statistic) >50% are shown.

Organic Lake Phycodnavirus 1, but with PgV as the next closest relative, and Superfamily II Helicase II protein was placed close to a poxvirus, but with extended *Mimiviridae* member *Aureococcus anophagefferens* Virus (AaV) as the next closest relative.

Information on functional annotation is limited, as 320 of the 503 ORFs have no predicted function (Figure 4.3). The remaining ORFs include the expected virus structural components, such as Major Capsid Protein (CpV_105, CpV_177), virion construction (A32-like packaging protein; CpV_098), viral transcriptional regulation (late transcription factor VLTF3; CpV_176), and DNA replication genes were observed, including replication factor C, DNA polymerase, DNA Helicase, Mismatch repair factor (MutS7 and MutS8), and ribonucleotide reductase. CpV also encodes a large contingent of genes with predicted DNA modification activity, including 14 DNA methyltransferases and a histone acetyltransferase. The genome features a group of 5 eukaryotic E3 ubiquitin ligases, 3 of which are clustered together in one location on the genome (CpV_167, CpV_169, CpV_171).

Funtional potential and phylogeny of CpVV

The *Chrysochromulina parva* Virus Virophage was assembled along with the "host" virus genome using SPAdes. The CpVV contig is 22,761 bases long with a GC content of 37.8% and ~42,000x average coverage. The coverage is the highest in the assembly, at 6 fold higher than the next highest contig, and 1000 fold higher than the vast majority of contigs. Open reading frames were predicted in CLC Genomic Workbench and features were predicted in BLAST2GO. CpVV encodes 19 ORFs (Figure 4.4), of which 8 have predicted functions and 1 has a conserved domain of unknown function. 2 of the predicted open reading frames encode predicted packaging ATPases, each of which is phylogenetically distinct. CpVV_08 is most closely related to the PgV virophage, as is observed with many of the other predicted ORFs, whereas CpVV_05 is more

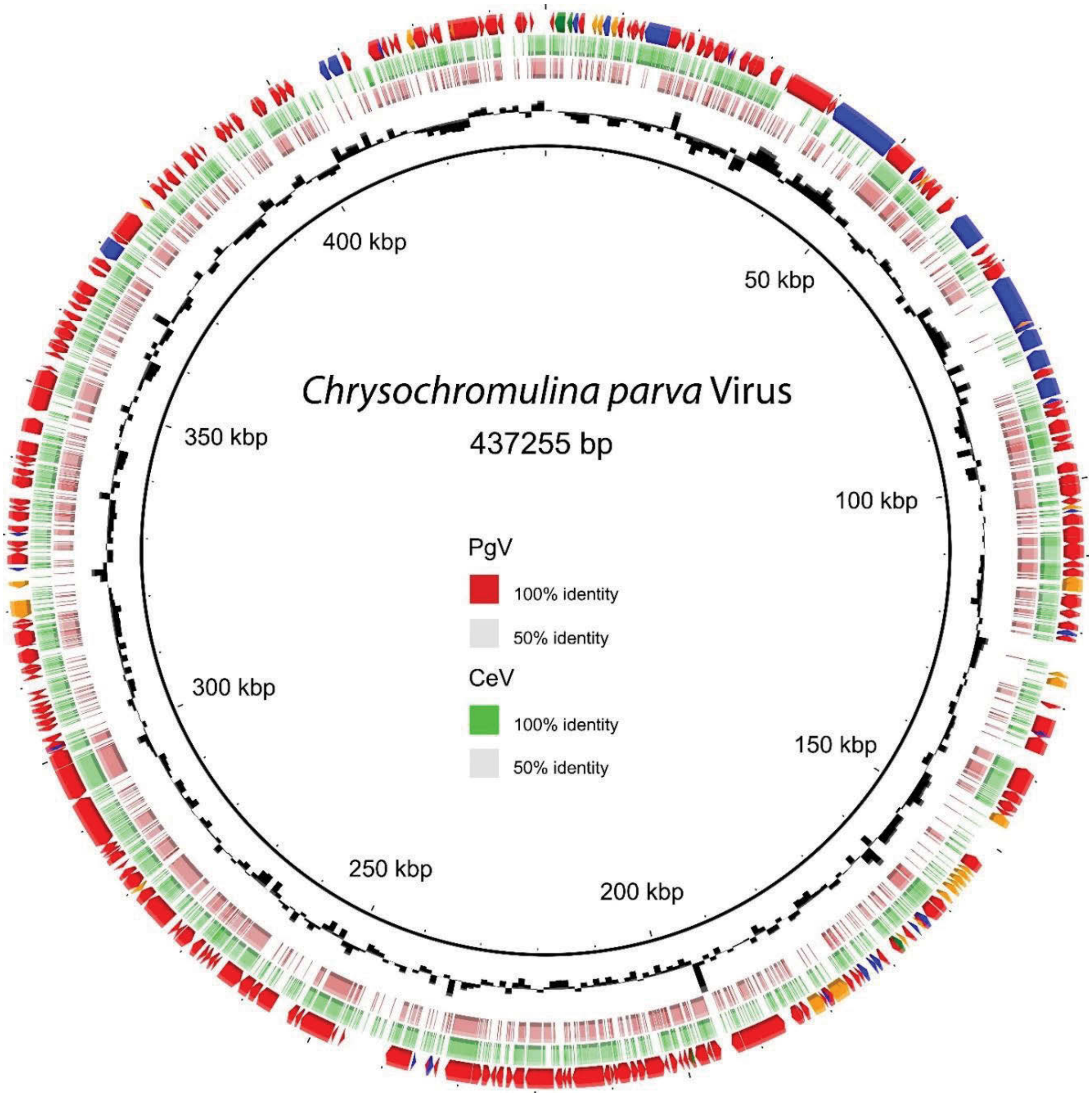

Figure 4.3. CpV ORF prediction and whole genome alignment. Outer ring – ORF prediction and top BLAST hit (Red arrows – NCLDV, Blue arrows – Bacteria, Orange arrows – Eukaryotes, Green arrows – Viruses). Second outermost ring in green – whole genome alignment with CeV where color gradient represents sites similar to CeV. Third outermost ring in red – whole genome alignment with PgV where color gradient represents sites similar to PgV. Innermost ring in black – GC content. Genome is presented as circular for convenience.

closely related to the environmental assembly known as Dishui Lake Virophage. However, BLAST results of CpVV_05 show top hits to ATPase genes encoded in members of the *Phycodnaviridae*, namely *Paramecium bursaria Chlorella* Virus. The genome also encodes a minor capsid protein (CpVV_11), and a major capsid (CpVV_0012) which appears to be related to the environmental assembly known as the Yellowstone Lake Virophage. Despite the similarity to YSLV7, PgVV is the top BLAST hit to both capsid proteins and CpVV_03. It should also be noted that the MCP, mCP, ATPase and CpVV_03 are all in similar genomic locations to their corresponding ORFs in PgVV (Figure 4.5). CpVV_0017 encodes a hypothetical protein with similarity to the Qinghai Lake Virophage gene QLV_03, and CpVV_18 shows similarity to mobile elements present in *Guillardia theta*, *Muricauda* sp., and *Tetrahymena thermophile* genomes. Outside of expected virophage genes, CpVV also encodes a predicted HNH Homing Endonuclease and DNA-methyltransferase, the top hits for both of which are bacterial. One additional open reading frame encodes a protein with predicted E3 ubiquitin ligase activity.

Environmental abundance of CpV and CpVV

To determine whether CpV and CpVV genes are expressed in currently available environmental metatransciptomes, reads from these datasets were mapped to the CpV and CpVV genomes with a minimum length cutoff of 0.7 and similarity fractions of 0.8 and 0.5, respectively. In every metatranscriptome dataset, a considerable number of reads were mapped to the short non-coding regions on the ends of the CpVV genome, and as such were not included in the abundance estimates shown here. Of the datasets mapped, metatranscriptomes sequenced from *Microcystis aeruginosa* blooms in Lake Tai, China during 2013 showed a moderate level of activity with good coverage of both CpV and CpVV genomes across all of the samples with a spike in activity in the SL48086 sample (Figure 4.6), however transcript abundance was too low

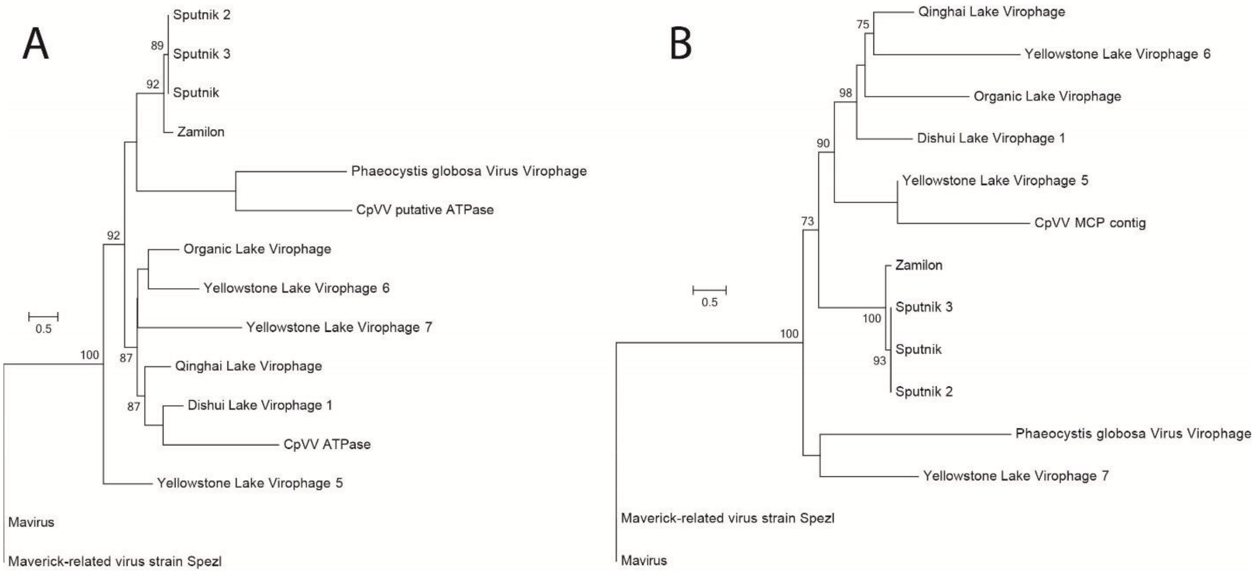
A
Sputnik 2
89
Sputnik 3
92
Sputnik
Zamilon
Phaeocystis globosa Virus Virophage
CpVV putative ATPase
92
Organic Lake Virophage
Yellowstone Lake Virophage 6
0.5
Yellowstone Lake Virophage 7
100
87
Qinghai Lake Virophage
87
Dishui Lake Virophage 1
CpVV ATPase
Yellowstone Lake Virophage 5
Mavirus
Maverick-related virus strain Spezl
B
Qinghai Lake Virophage
75
Yellowstone Lake Virophage 6
98
Organic Lake Virophage
Dishui Lake Virophage 1
90
Yellowstone Lake Virophage 5
CpVV MCP contig
73
Zamilon
0.5
Sputnik 3
100
Sputnik
100
93
Sputnik 2
Phaeocystis globosa Virus Virophage
Yellowstone Lake Virophage 7
Maverick-related virus strain Spezl
Mavirus

Figure 4.4. Maximum-likelihood phylogenetic trees of A.) virion packaging ATPase, B.) Major capsid protein. Node support (aLRT-SH statistic) >50% are shown.

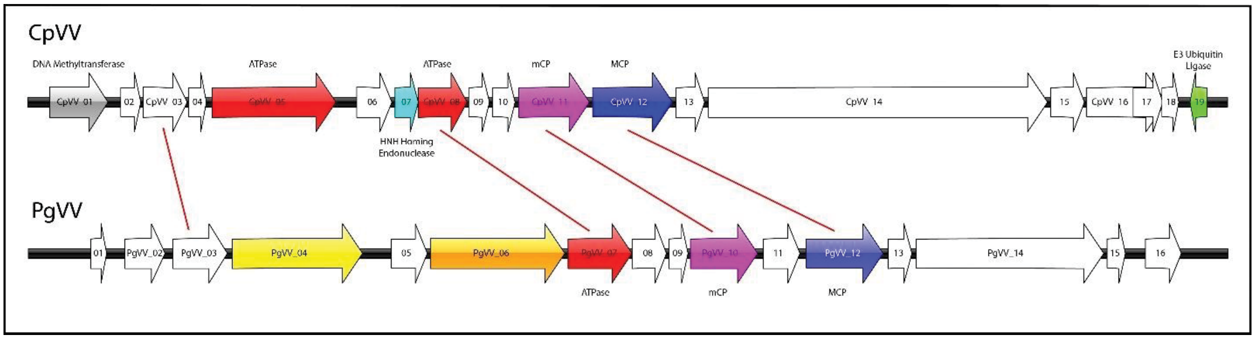

Figure 4.5. Genome architecture of CpVV (top) and PgVV (bottom). Synteny between virophage shown by connecting red lines.

to discern distinct transcriptional patterns from more than one sample. Comparison of the number of reads mapped to the genome of CpV to CpVV showed transcript abundance was highly correlated (Figure 4.6b; r = 0.82, p = 0.0005).

Discussion

In this study, we determined the genome sequence and functional potential of the *Chrysochromulina parva* Virus and its corresponding virophage. We also used currently available environmental metatranscriptomic data to estimate the activity of CpV and its virophage in freshwater ecosystems. The genome content suggests CpV is a versatile giant virus with a close evolutionary relationship with the marine algal mimiviruses PgV and CeV. As with many giant viruses, CpV possesses a mosaic arrangement of genes originating in viruses and all three domains of life, as well as a contingent of genes of unknown origin or function. However, the similarity of CpV to closely related marine algal mimiviruses suggests efforts to sample the diversity of NCLDVs may be filling many of the gaps in sequenced representatives. CpV shares a similar genome size, GC%, and content with PgV and CeV, who make up top Blast hits for more than half of the predicted open reading frames. In support of this close evolutionary relationship, CpVV also appears to exhibit similar gene content and synteny with PgVV. Beyond what is shared with its marine relatives, both CpV and CpVV appear to have incorporated a number of genes predicted to have originated in the cellular host, which may serve roles in genetic regulation and protection from host defenses.

The first group of genes unique to CpV is a large collection of genes involved in DNA modification, including 13 DNA methyltransferases, a histone demethylase, a histone acetyl transferase, and at least 8 restriction modification systems. In addition, the virophage also appears to encode a DNA methyltransferase, likely horizontally transferred either from CpV or

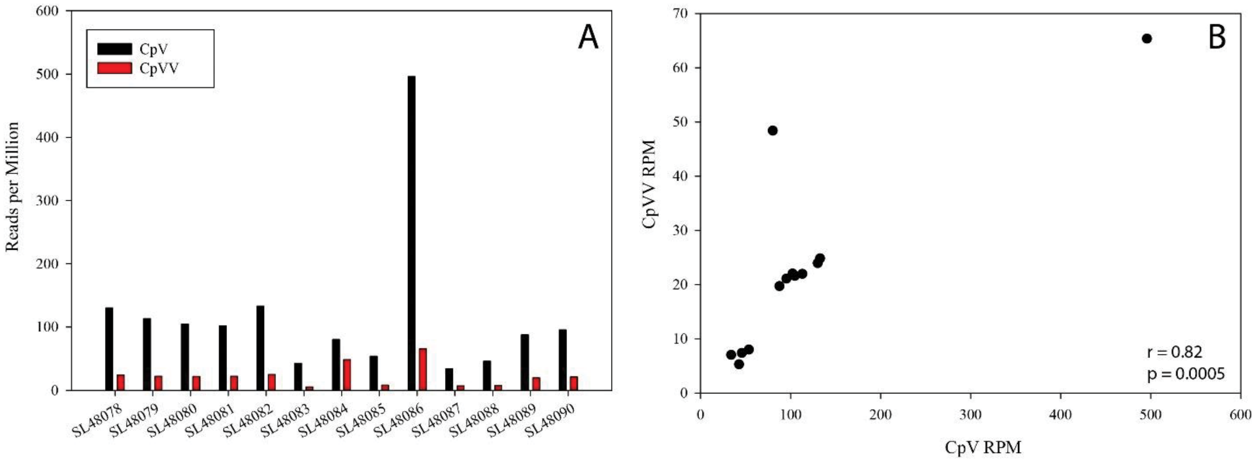

Figure 4.6. Relative abundance of CpV and CpVV in Lake Tai during the *M. aeruginosa* bloom during 2013. A.) Reads per million mapped to CpV and CpVV genomes within each sample. B.) Reads per million mapped to CpV plotted against reads per million mapped to CpVV, with Pearson correlation coefficient and p-value.

the cellular host. Methyltransferases have been indicated in the physiology of giant viruses for some time, as *Chlorella* viruses have long been shown to exhibit heavily methylated genomes encoding large numbers of methylation genes accompanied by an unexpected contingent of corresponding restriction endonucleases (Nelson *et al.*, 1998, Etten & Meints, 1999). DNA methyltransferases have been observed in the genomes of other giants as well, though not nearly as many and generally not accompanied by R-E systems. In both cases, as with CpV and CpVV, the purpose of these functional genes is unknown, though the similarity to epigenetic regulation of transcription in cellular organisms is apparent (Bird, 2002). Of similar interest is the presence of putative histone modification enzymes, histone demethylase (CpV_120) and histone acetyltransferase (CpV_503). As it is extremely unlikely that the viral genome is packaged bound to histones, the most obvious purpose of these two genes, assuming they were properly identified, is to manipulate the host's transcriptional regulation (Kouzarides, 2007).

While DNA modification machinery is frequently observed amongst NCLDVs (Santini *et al.*, 2013, Moniruzzaman *et al.*, 2014), though not always as densely, CpV also possesses a cluster of 6 E3-ubiquitin ligases, all eukaryotic in origin and most grouped within 1kb of one another. Similar RING-finger E3-ubiquitin ligases have been observed in several Mimiviruses (Iyer *et al.*, 2006), they are generally fewer in number and spread throughout the genome. Only *Aureococcus anophagefferens* Virus (AaV) appears to have a similar group of coding sequences (Moniruzzaman *et al.*, 2014). While function of these proteins in NCLDVs has yet to be validated in culture, they are hypothesized to inhibit host cell defenses (Iyer *et al.*, 2006, Chaurushiya *et al.*, 2012). The presence of E3-ubiquitin ligase coding sequence in CpVV appears to be entirely new. Activity by virophage generally occurs at the expense of growth by the "host" virus (Fischer & Suttle, 2011, Fischer, 2012), but contribution of an active, functional Ub-ligase might serve to further protect the corresponding giant virus from host defenses, potentially challenging the antagonistic nature of the virophage. Indeed, as the virophage particle was likely packaged within the CpV virion before DNA extraction and sequencing as was also observed in PgVV (Santini *et al.*, 2013), CpVV may also rely on successful infection by its "host" virus to spread and replicate.

To further explore the ecological role CpV and its virophage may play in freshwater ecosystems, we mapped currently available metatranscriptome reads isolated and sequenced from *M. aeruginosa* blooms in Lake Erie and Lake Tai, China during the years 2013 and 2014. While most of the datasets showed a complete absence of activity by CpV, expression was detected in Lake Tai during the 2013 bloom across all samples. Reads mapping to CpVV cooccurred with its "host" in all samples, exhibiting a high correlation. Considering that the metatranscriptomes used here were isolated from microbial communities dominated freshwater cyanobacteria, and were not poly-A selected, activity by CpV must have been very considerable to be detected. Additionally, as infection cycles in the environmental datasets are not synchronized, the relationship between CpV and its corresponding virophage must be very close to be observed through a noisy dataset. As CpV was originally isolated from Lake Ontario, Canada (Mirza *et al.*, 2015), it is likely that the virus and virophage observed here in Lake Tai are close relatives. Despite potential differences in physiology however, these results suggest that closely related freshwater algal Mimiviruses are globally distributed and environmentally relevant.

Altogether, the results here suggest that CpV represents an abundant and active member of the extended *Mimiviridae* with a unique functional potential. As the first freshwater representative of the algal mimiviruses to be isolated and maintained in culture, CpV stands as an important model virus for the future study of mimivirus ecology and physiology in freshwater ecosystems. Its virophage, though similar to PgVV, offers a unique opportunity to study the emerging taxon of giant virus-infecting-viruses in culture, as well as a significant expansion to the currently underrepresented virophage diversity.

Acknowledgements

This work was support by a NSERC Discovery award to SMS and a GBMF Grant to SWW. The authors are grateful for the support of the *Kenneth and Blaire Mossman Endowment* at the University of Tennessee.

CHAPTER V:

FUNCTIONAL CHARACTERISTICS OF THE GUT MICROBIOME IN C57BL/6 MICE DIFFERENTIALLY SUSCEPTIBLE TO *PLASMODIUM YOELII*

Publication Note

This chapter is a version of a peer-reviewed article published in *Frontiers in Microbiology* 7:11 by Joshua M.A. Stough, Stephen Dearth, Joshua Denny, Gary LeCleir, Nathan Schmidt, Shawn Campagna, and Steven Wilhelm.

My contribution to this work was the experimental design, sample collection and processing, data management and analysis, and primary authorship and editing of the manuscript.

Abstract

C57BL/6 mice are widely used for in vivo studies of immune function and metabolism in mammals. In a previous study, it was reported that when C57BL/6 mice purchased from different vendors were infected with Plasmodium yoelii, a causative agent of murine malaria, they exhibited both differential immune responses and significantly different parasite burden. These patterns were reproducible when gut contents were transplanted into gnotobiotic mice. To gain insight into the mechanism of resistance, we removed whole ceca from mice purchased from two vendors, Taconic Biosciences (low parasitemia) and Charles River Laboratories (high parasitemia), to determine the combined host and microflora metabolome and metatranscriptome. With the exception of two Charles River samples, we observed ≥90% similarity in overall bacterial gene expression within vendors and ≤80% similarity between vendors. In total 33 bacterial genes were differentially expressed in Charles River mice (p-value < 0.05) relative to the mice purchased from Taconic. Included among these, fliC, ureABC, and six members of the nuo gene family were overrepresented in microbiomes susceptible to more severe malaria. Moreover, 38 mouse genes were differentially expressed in these purported genetically identical mice. Differentially expressed genes included basigin, a cell surface receptor required for P. falciparum invasion of red blood cells. Differences in metabolite pools were detected, though their relevance to malaria infection, microbial community activity, or host response is not yet understood. Our data have provided new targets that may connect gut microbial activity to malaria resistance and susceptibility phenotypes in the C57BL/6 model organism.

Introduction

Since its development in the 1940's, the C57BL/6 inbred mouse strain has become one of the most widely used murine genetic backgrounds for diverse biomedical research. The strength of these inbred mice as model organisms is their reproducibility, allowing independent researchers to carry out experiments on genetically identical mice (Silver, 1995). Use of this inbred strain became so widespread it was selected as the first murine genome to be sequenced (Waterston *et al.*, 2002). However in recent years, attention has been drawn to the split in the strain's ancestral line during the 1950's when mice were separately bred and maintained by the National Institutes of Health (NIH) and Jackson Laboratory, now known as C57BL/6N and C57BL/6J, respectively (Bailey, 1978, Altman & Kats, 1979). Concern has arisen over use of these divergent substrains interchangeably as model organisms following multiple reports of changes in behavior (Crawley *et al.*, 1997), differential tolerance to ethanol (Khisti *et al.*, 2006, Green *et al.*, 2007), deletion of the gene encoding nicotinamide nucleotide transhydrogenase (*nnt*) in the C57BL/6J lineage (Freeman *et al.*, 2006), and discovery of multiple SNPs between derived mouse genomes (Mekada *et al.*, 2009). The importance of these strains to the scientific community has led to major efforts to describe the genomic (Simon *et al.*, 2013) and regulatory (Keane *et al.*, 2011) differences between the various lineages, and catalogue them for proper selection of model organisms (Grubb *et al.*, 2014).

While the genetic differences and the resulting phenotypic alterations between the major C57BL/6 lineages may be increasingly considered by researchers during experimental design, only recently can this be said for their "second genome": the microbiome. The importance of tissue-associated microbial symbionts to mammalian metabolism and immunity has become well established. Gut microbial communities in particular make up the majority of the microbial consortia and diversity in the body (Savage, 2002), and play an important role in early post-natal development of the immune system, protection from gut pathogens, and host metabolism. Members of the taxa Firmicutes and Bacteroidetes dominate intestinal communities, largely responsible for the catabolism of hundreds of different glycans indigestible by mammalian enzymes, giving the host access to otherwise recalcitrant nutrients (Backhed *et al.*, 2005). The resulting pool of monosaccharides are fermented to short-chain fatty acids, which not only provide energy for the host, but have been shown to influence immune function. Acetate and butyrate, influenced by dietary fiber content, can signal through G-protein-coupled receptors expressed on CD4+ T helper cells resulting in the regulation of cytokine expression and resolution of intestinal inflammation (Kau *et al.*, 2011). Indeed, just as immune cells use receptors to detect infection and tissue damage signals, it is apparent that the same receptors are used in different combinations to detect beneficial microbial activity and prevent harmful response (Swiatczak & Cohen, 2015). Despite the profound influence that even subtle changes in gut community composition and activity can have on host physiology, the impact of the gut microbiome on mice used as model organisms remains poorly understood.

It was recently shown that when C57BL/6 mice purchased from different vendors were infected with malaria parasite *Plasmodium yoelii*, they exhibited significantly different parasite burdens and immune responses. This was confirmed to be the result of microbial interaction with the mouse host when both resistant and susceptible phenotypes were reproduced via fecal transplant to gnotobiotic mice (Villarino *et al.*, 2016). Subsequent sequencing of 16S rRNA gene libraries obtained from the transplanted gut microbiomes showed conservation in gut microbial community composition within, but major differences between, samples obtained from mice from different vendors. To elucidate the mechanisms underlying microbiome-mediated resistance to malaria, the cecum microbial and host metatranscriptome was sequenced. Significant differences were observed in both host and bacterial transcription patterns. Additionally, the metabolic profiles of cecum whole tissue samples were determined and analyzed. Overall differences in individual metabolite concentrations call into question the interchangeable use of mice from different sources. These data begin to elucidate factors that may influence susceptibility to *P. yoelli* infection, and these results also provide further evidence that caution is needed when comparing results from experiments using mice from separate C57BL/6 sublineages and/or vendors.

Materials and Methods

Mice and Infections

Female C57BL/6 mice were purchased from Taconic Biosciences (Hudson, NY) and Charles River Laboratories (Wilmington, MA). Mice were housed and maintained at University of Tennessee animal care facility under biosafety level 2 conditions. Mice were fed NIH-31 Modified Open Formula Mouse/Rat Irradiated Diet (Envigo 7913; Envigo, Indianapolis, IN) and provided autoclaved municipal tap water to drink. To verify vendor-dependent malaria disease severity, mice were infected with 10^5 *Plasmodium yoelii* parasitized red blood cells (pRBCs) *via* tail vein injection after a two-week acclimation period upon arrival at the animal care facility. Parasite burden was determined from thin blood smears. Blood samples were obtained by performing tail snips. Slides were fixed with methanol, followed by Giemsa stain (Thermo Fisher Scientific) diluted 1:20 in ddH_20 for 30 min. Percent parasitemia was calculated as the percent of total RBCs that contain a blood stage parasite averaged from the counts RBCs within a 10x10 grid from five microscope fields (1000x) per sample. All studies were performed in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and approved by the University of Tennessee Institutional Animal Care and Use Committee.

Gut Microbiome Sampling

As sampling directly from gut tissue is destructive, mice used for microbiome sampling were not used to track parasite burden. To limit potential variation in gut microbial communities and to ensure that the disease severity phenotype was consistent, mice used for microbiome sampling were purchased in the same batch as those used for tracking disease progression. Six mice from each vendor (Taconic Biosciences and Charles River Laboratories) were acclimated for two weeks upon arriving at the animal facility. After acclimation the mice were sacrificed and a necropsy performed. Whole ceca were removed, weighed, and immediately flash frozen in liquid nitrogen and stored at -80 °C. Cecum samples were divided in half for metabolomics analysis and metatranscriptome sequencing.

RNA extraction and Sequencing

Total RNA was isolated from whole ceca using the MOBIO Power Microbiome™ RNA extraction kit. RNA concentration and purity was determined using a NanoDrop ND-1000 spectrophotometer. Measurements were taken three times to account for variability in the readings. Extracted RNA was tested for DNA contamination by running a polymerase chain reaction using universal bacterial 16S rRNA primers 27F and 1492R. DNA contamination was removed with the MOBIO RTS DNase kit. 12 purified RNA samples were shipped to the Hudson Alpha Institute Genomic Services Laboratory (Huntsville, AL) for rRNA reduction and sequencing on the Illumina™ HiSeq platform using a paired-end 100bp flow cell.

Metabolite extraction and analysis

Contents were removed from ceca placed in 1.5 mL centrifuge tubes and suspended in 1.3 mL of extraction solvent (40:40:20 HPLC grade methanol, acetonitrile, water with 0.1% formic acid) kept at 4°C. Extraction proceeded for 20 min at -20°C before samples were centrifuged for 5 min (16.1 rcf) at 4˚C and supernatants were transferred to new vials. The remaining cecal contents were resuspended in 200 µL of cold (4˚C) extraction solvent. The extraction was again allowed to proceed for 20 min at -20°C before being centrifuged for 5 min (16.1 rcf) at 4°C. These supernatants were also transferred to the vials and another 200 µL of extraction solvent was added to the pelleted cell for a final wash by repeating the previous extraction once more. The vials containing all of the combined extraction supernatants were placed in a nitrogen drying apparatus until all the extraction solvent had been evaporated. The residual solid was resuspended in 300 µL of sterile water and transferred to 300 µL autosampler vials. Samples were immediately placed in a 4˚C autosampler for mass spectrometric analysis.

A 10 µL injection of each sample was separated through a Synergi 2.5 micron Hydro-RP 100 Å, 100 x 2.00 mm LC column (Phenomenex, Torrance, CA) maintained at 25°C. The mass spectrometer and chromatographic separation were performed similar to a reported method (Lu *et al.*, 2010). The eluent was introduced into the mass spectrometer *via* an electrospray ionization source in negative mode before entering an Exactive Plus orbitrap mass spectrometer (Thermo Scientific, Waltham, MA) through a 0.1-mm internal diameter fused silica capillary tube. The samples were run with a spray voltage of 3 kV, a nitrogen sheath gas flow rate of 10 units, a capillary temperature set at 320°C, and an AGC target set to 3e6. The samples were analyzed in full scan mode with a resolution of 140,000 and a scan window of 85 to 800 *m/z* for from 0 to 9 min and 110 to 1000 *m/z* from 9 to 25 min. Solvent A consisted of 97:3 HPLC grade water:methanol, 10 mM tributylamine, and 15 mM acetic acid. Solvent B was HPLC grade methanol. The mobile phase gradient from 0 to 5 mins was 0% B, from 5 to 13 min was 20% B, from 13 to 15.5 min was 55% B, from 15.5 to 19 min is 95% B, and from 19 to 25 min was 0% B while maintaining a constant flow rate of of 200 µL/min.

Data Processing

Raw sequences were downloaded from the HudsonAlpha Institute server and checked for quality using FastQC application (Babraham Institute, Cambridge, England). Unless noted, all bioinformatics and statistical software were used at default settings. Samples were subjected to a subsequent *in silico* rRNA reduction using the SortmeRNA 2.0 software package (Kopylova *et al.*, 2012). Since RNA was extracted from whole cecum tissue and would contain mRNA of murine origin, processed reads were paired and mapped to the *Mus musculus* reference genome using the CLC Genomics Workbench v8.5 (Waltham, MA). Mouse reads were annotated and further analyzed in CLC. Unmapped reads were assumed to originate from the gut microbiome and were uploaded to the Metagenomics RAST server (MG-RAST) (Meyer *et al.*, 2008) for alignment and identification. All sequencing data were submitted to the Short Reads Archive (SRA) under accession code SRP075802.

For metabolome data, .raw files generated by Xcalibur were converted to the open-source mzML format (Martens *et al.*, 2011) *via* the ProteoWizard package (Chambers *et al.*, 2012). MAVEN software (Clasquin *et al.*, 2002) (Princeton University) was used to automatically perform non-linear retention time correction for each sample. Metabolites were manually identified by *m*/*z* (± 5 ppm) and retention time for each sample using MAVEN to calculate associated peak areas. Relative concentrations (*i.e.,* in the absence of internal standards for all metabolites) were normalized by mass of the processed tissue sample. Fold changes were calculated and the data were transformed and clustered using Cluster software (de Hoon *et al.*, 2004). Heat maps were generated from clustered data using Microsoft Excel software.

Statistical Analysis

Microbial transcript abundances annotated from the SEED Subsystem database (Overbeek *et al.*, 2005) (evaluated as raw read counts) were exported from the MG-RAST server and normalized by library size. Normalized gene expression data and relative metabolite concentration were log transformed, and used to generate a Bray-Curtis dissimilarity matrix and non-metric multidimensional scaling plots in the PRIMER7 software suite (Clark & Gorley, 2015). PRIMER7 was also used to perform ANOSIM tests comparing overall expression and metabolite profiles. Differences in individual gene expression, between gut microbial communities from the two vendors, were determined using the edgeR Bioconductor package in R Statistics software (Robinson & Smyth, 2007, Robinson & Smyth, 2008, Robinson *et al.*, 2010, McCarthy *et al.*, 2012, Zhou *et al.*, 2014). Differential expression of individual mouse genes between vendors was determined using the edgeR test implemented in CLC Genomics Workbench. Figures were generated using SigmaPlot (Systat Software, Inc.). As p-values from statistical tests were false discovery rate adjusted for multiple comparisons, a p-value cutoff of 0.1 was used to provide thorough detailing of differences between mouse substrains that may be useful to researchers. Additionally, Cohen's *d* effect size (Cohen, 1988) was calculated for each gene from relative transcript abundances. All significantly different genes and metabolites are presented with their p-values, fold changes, and effect sizes in Table 5.1.

Results

Differential susceptibility to *P. yoelii*

C57BL/6N mice from Taconic and Charles River were infected with *P. yoelii* pRBCs. Parasitemia in Taconic mice peaked 13 days post-infection at ~15% and was cleared by 23 days post-infection (Figure 5.1). Charles River mice exhibited higher parasite burden, peaking at ~60% parasitemia 19 days post-infection and delayed clearance (day 29 post-infection) compared to Taconic mice. These data are consistent with previous observations that showed *P. yoelii* infection of C57BL/6 mice from Taconic and Jackson Laboratories had lower parasitemia than C57BL/6 mice from Charles River, National Cancer Institute, and Envigo (formally Harlan) (Villarino *et al.*, 2016).

Transcriptome results

Ribosomal RNA reduction, cDNA synthesis, and sequencing on the Illumina™ HiSeq yielded a total of 294 million paired-end 100bp reads across 12 samples. An average of 43.8% of reads were removed during *in silico* rRNA reduction using SortMeRNA. One of the Taconic samples exhibited much higher attrition, with 73.2% of its reads removed. As a result, the number of reads annotated from this sample were a full order of magnitude lower than the other samples, so it was removed from further analyses because of dissimilarity. Reads passing quality control were mapped to the mouse genome and subsequently used to determine murine transcriptional patterns. The remaining reads were uploaded to MG-RAST for characterization of microbial transcriptional patterns. The quality control pipeline removed an average of 14.6% of reads due to read quality, artificial duplication, and estimated sequencing error. An average of 2.1 million reads per sample were annotated as microbial transcripts and divided into functional categories.

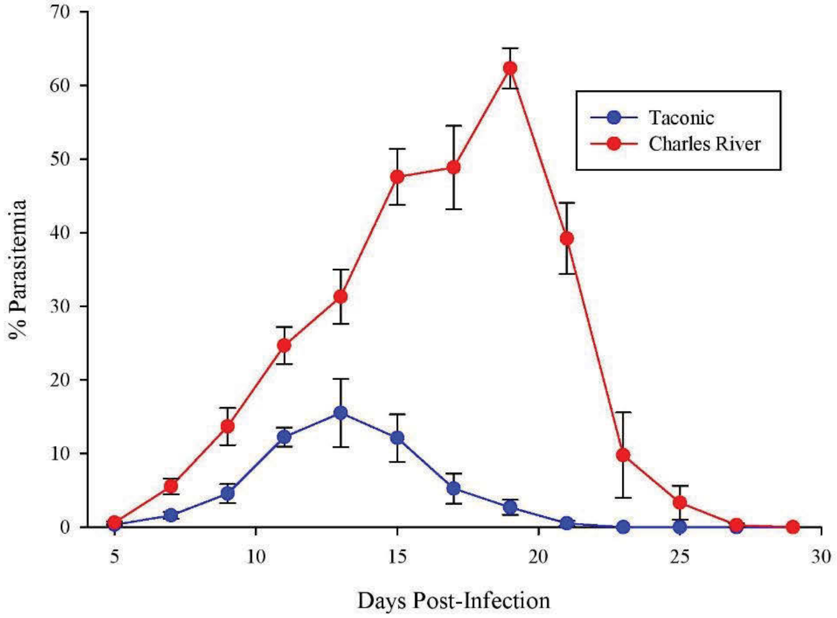

Figure 5.1. C57BL/6 mice from Taconic exhibit reduced parasitemia compared to mice from Charles River. Mice were infected with 10^5 *P. yoelii* pRBCs. Percent parasitemia was determined on the indicated days. Data (mean±s.d.) are cumulative results (n=7-8 mice per group) from two independent experiments.

Community Structure and Function

The phylogenetic composition of the cecal microbial community as determined by metatranscriptomic analysis is represented in Figure 5.2. The microbial community transcriptional profile is dominated by the bacterial phyla Firmicutes and Bacteroidetes, the reads from which make up an average of 90.1% ± 6.3 of each sample. The next most abundant source of transcripts originate in Proteobacteria at 3.2% ± 0.20 of reads, followed by Actinobacteria, 1.7% ± 0.16, and Fusobacteria, 0.53% ± 0.03. Within the phylum Bacteroidetes, families Bacteroidaceae and Porphyromonadaceae are most prevalent, 46.7% and 51.1% of the phylum respectively. The Firmicutes portion of the community is split predominantly between orders Lactobacillales and Clostridiales, 8.9% and 82.5% of the phylum respectively. The MG-RAST pipeline identified 0.04% ± 0.007 of the reads as being of viral origin, all of which were bacteriophage. Archaea made up 0.24% ± 0.009 of the transcripts, with the Euryarchaeota dominating at 92.3% of the Archaeal reads.

Non-metric multidimensional scaling plot of Bray-Curtis dissimilarity analysis is represented in Figure 5.3. Sample Taconic 6 was left out of this analysis due to significant dissimilarity caused by methodology that skews the plot. Overall bacterial transcript abundances in the 5 Taconic and 6 Charles River samples are at least 80% similar. With the exception of two Charles River samples (designated by asterisks in Figure 5.3), mouse groups cluster with at least 85% similarity and as high as 98%. These two samples more closely resemble expression profiles of the Taconic gut communities. As mice from these two substrains are so closely related, some overlap within the internal variation of the mouse groups was to be expected. However, ANOSIM analysis comparing overall expression of bacterially-derived transcripts determined that community

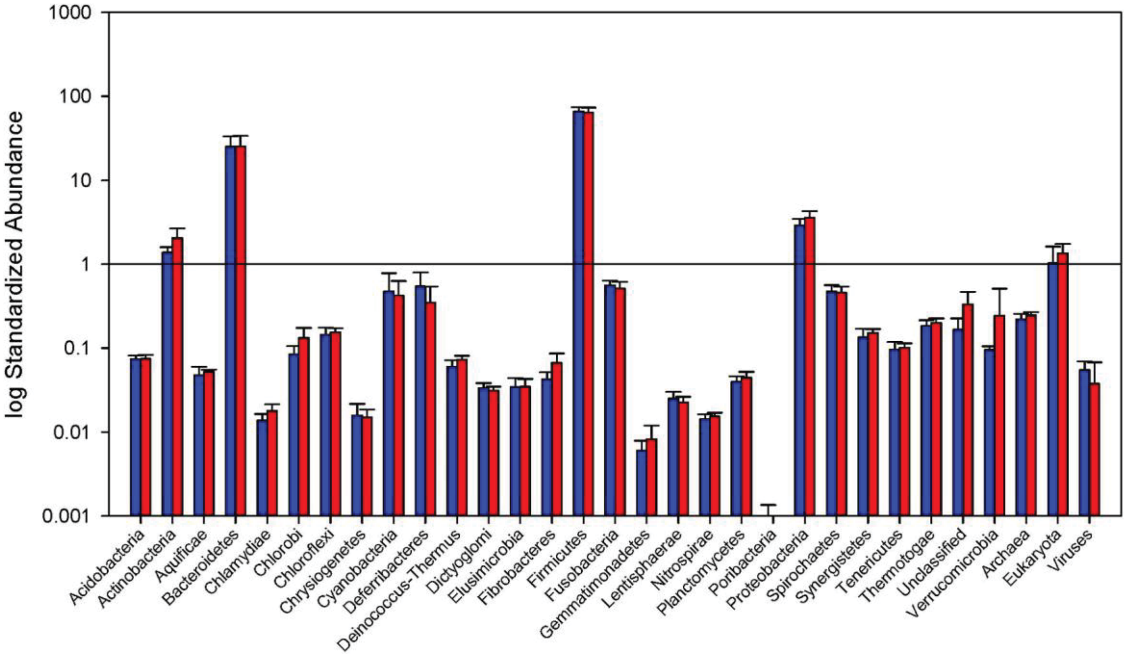

Figure 5.2. Relative abundance of Bacterial phyla and total Archaea, Eukarya, and virus reads. Read counts normalized by library size from the samples in each group. Blue bars represent abundance in mice purchased from Taconic Biosciences. Red bars represent abundance in mice purchased from Charles River Laboratories. Error bars represent standard deviation. Data (mean ± s.d.) are from n=5 Tac and n=6 CR mice.

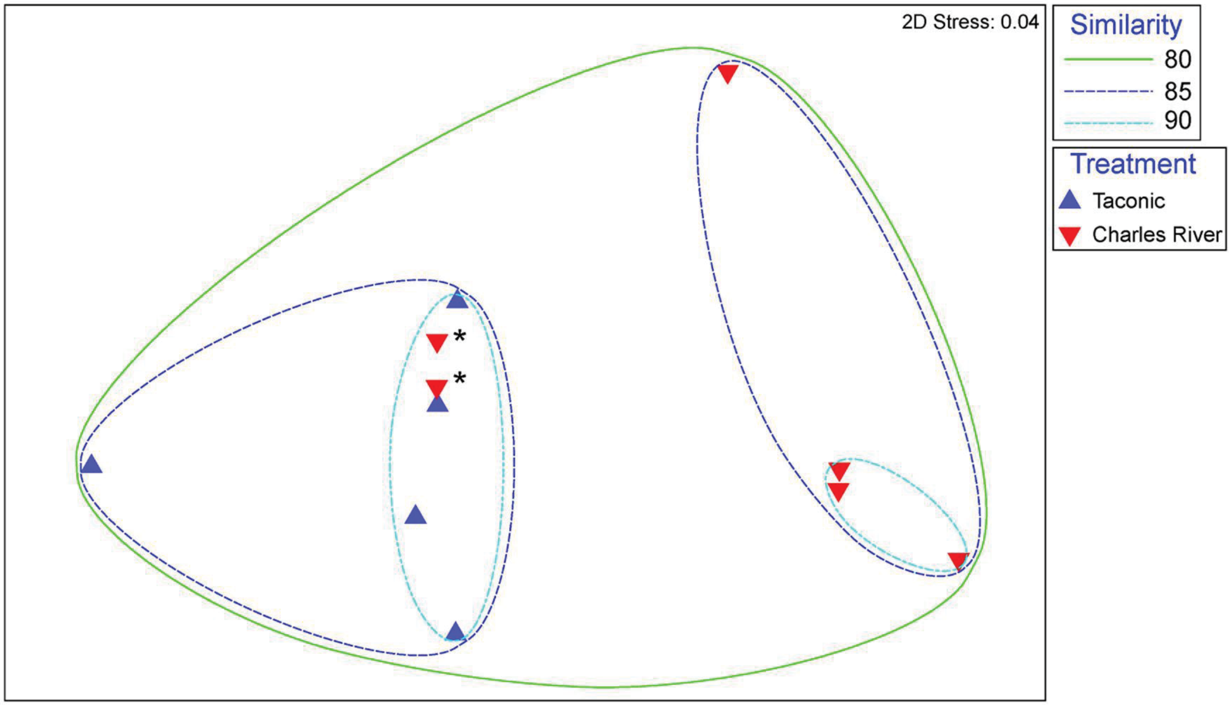

Figure 5.3. Non-metric multidimensional scaling of Bray-Curtis similarity matrix comparing overall abundances of bacterially derived transcripts. Blue points represent samples isolated from Taconic Biosciences mice. Red points represent samples isolated from Charles River Laboratories mice. Ellipses represent lines of 80, 85, and 90% similarity between samples. Asterisks designate two Charles River samples addressed in text.

expression between mouse groups was statistically different ($p = 0.048$).

In general, the distribution of sequences within SEED Subsystem categories were consistent between the two mouse groups (Figure 5.4). Combining 11 metatranscriptomes, the most abundant functional groups are Carbohydrate Metabolism (19.5%), Protein Metabolism (14.0%), and Amino Acid Metabolism (7.7%). A significant portion (13.3%) of the sequences are categorized as clustering-based subsystems, whose functions are bioinformatically identified, but not yet experimentally validated. An unpaired t-test comparing normalized expression of individual Level 1 SEED Subsystem categories between the two treatment groups yielded significant ($p < 0.05$), or trending towards significant ($p < 0.08$), differences in Protein Metabolism ($p = 0.029$), Cell Wall and Capsule synthesis ($p = 0.053$), Motility and Chemotaxis ($p = 0.047$), Sulfur Metabolism ($p = 0.038$), Iron Acquisition and Metabolism ($p = 0.077$), Secondary Metabolism ($p = 0.059$), and Potassium Metabolism ($p = 0.014$).

Differentially Expressed Bacterial Genes

To determine whether specific transcripts significantly differed in expression between the resistant and susceptible phenotypes, statistical analysis of differential gene expression of bacterially derived transcripts was performed using the edgeR Bioconductor package. A total of 60 bacterial genes were differentially expressed ($p \leq 0.1$), 33 of which with false discovery rate (FDR) adjusted p-values less than 0.05 and 11 with p-values less than 0.001 (Figure 5.5). Of these, 51 of 60 genes were overrepresented in Charles River mice compared to Taconic. The majority of differentially expressed genes are involved in energy, amino acid, and carbon metabolisms. Overexpressed in Charles River mice were transcripts encoding FliC, the flagellar body protein, which is heavily proinflammatory. Only three genes were determined to be significantly overrepresented in resistant mice purchased from Taconic Biosciences. All

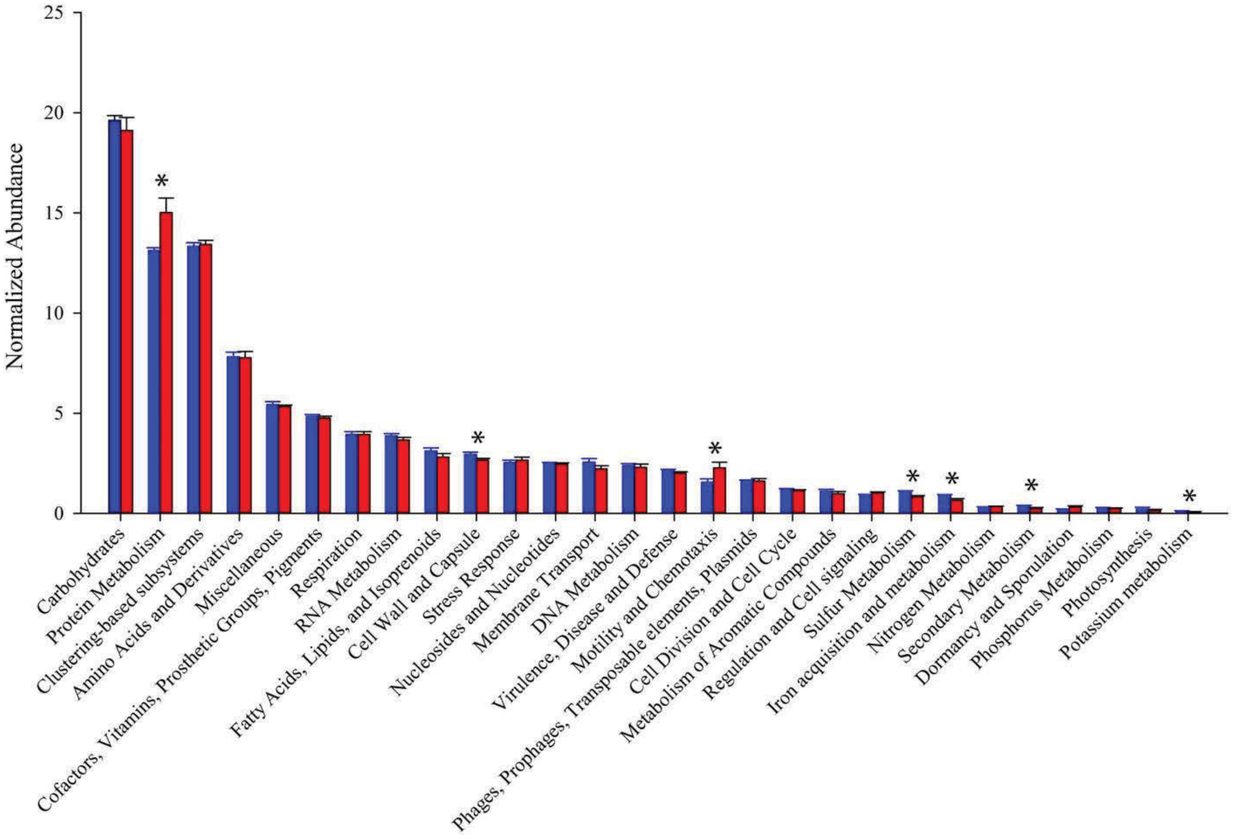

Figure 5.4. Relative abundance of SEED subsystems functional categories. Read counts normalized by library size from samples within each group. Blue bars represent abundance in mice purchased from Taconic Biosciences. Red bars represent abundance in mice purchased from Charles River Laboratories. Data (mean ± s.d.) are from n=5 Tac and n=6 CR mice. Asterisks indicate functional categories significantly different ($p < 0.05$) or trending towards significant ($p < 0.8$) in a comparison via unpaired Student's t-test.

statistically significant bacterial genes, with the exception of three, exhibited an effect size greater than 0.8, the value typically used as the cutoff for a strong effect.

Differentially Expressed Mouse Genes

Since sequencing also yielded mouse transcripts within the samples, differential gene expression amongst the murine transcripts was also analyzed. Fold change in gene expression and FDR adjusted p-values from the exact test are presented in the volcano plot in Figure 5.6. Twenty genes were differentially expressed with a p-value less than 0.1, 12 of which had p-values less than 0.05. Of these, 11 genes were significantly overrepresented in Charles River mice and one in Taconic mice. The overrepresented transcripts in Charles River mice include Galectin-9 (*LGALS9*), which is an important immune signaling molecule (Merani *et al.*, 2015), and Basigin (*bsg*), a cell surface receptor whose expression is required for infection of RBCs by the human malaria parasite *Plasmodium falciparum* (Crosnier *et al.*, 2011). All statistically significant mouse genes exhibited an effect size greater than 1.0, with the lowest being 1.1

Metabolite Pools

Relative metabolite concentrations were normalized by mass of the processed tissue sample, and these data were used to calculate fold change and cluster analyses. Comparison of normalized metabolite abundances determined that differences in the metabolome of Charles River and Taconic mice were present ($p = 0.082$). Normalized abundance of significantly different metabolites are presented in Figure 5.7. Of the 129 metabolites detected in the samples, 36 were found in significantly higher relative concentrations in Charles River mice, and two (NADH and N-acetyl-L-alanine) were found in higher concentrations in Taconic mice ($p < 0.1$). All statistically significant metabolites exhibited an effect size greater than 1.0, with the lowest being 1.17. The majority of significant metabolites were nucleotides, amino acids, or the

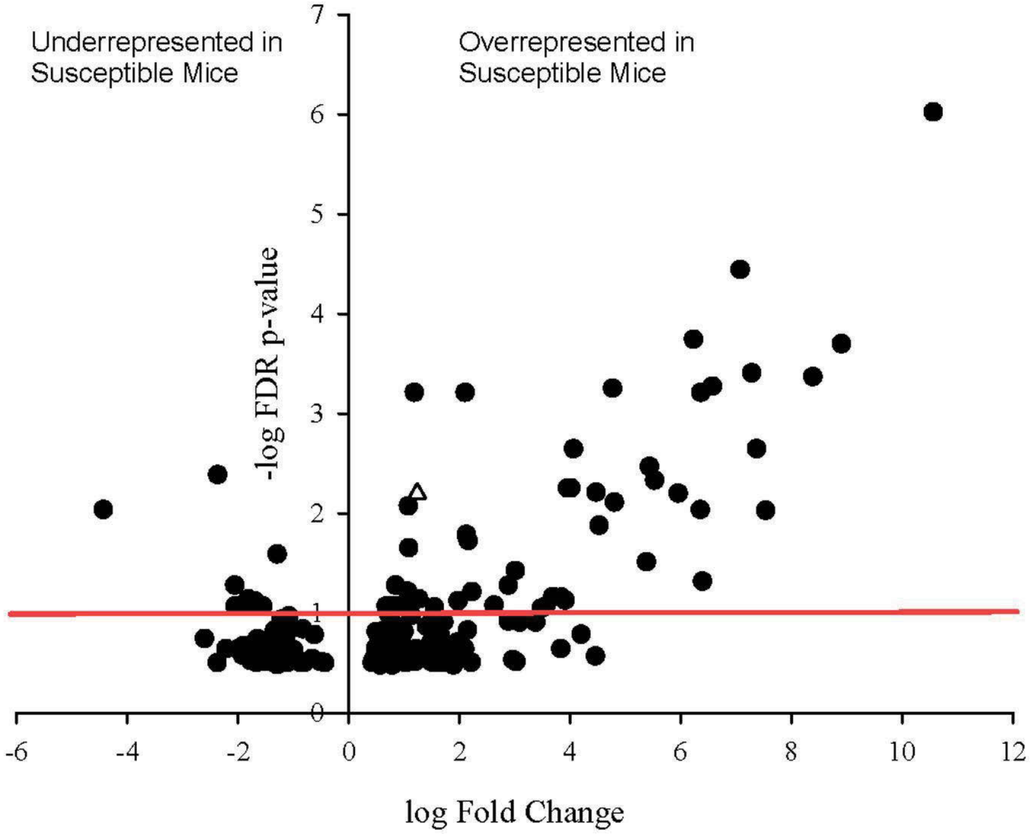

Figure 5.5. Volcano plot showing degree of differential expression of bacterially-derived genes in Charles River Laboratories mice compared to Taconic Biosciences. Log-transformed fold change in expression is plotted on the x-axis and log-transformed false discovery rate-adjusted p-values plotted on the y-axis. The red horizontal line represents the 0.1 p-value cutoff. Empty triangle: *fliC* (Flagellin).

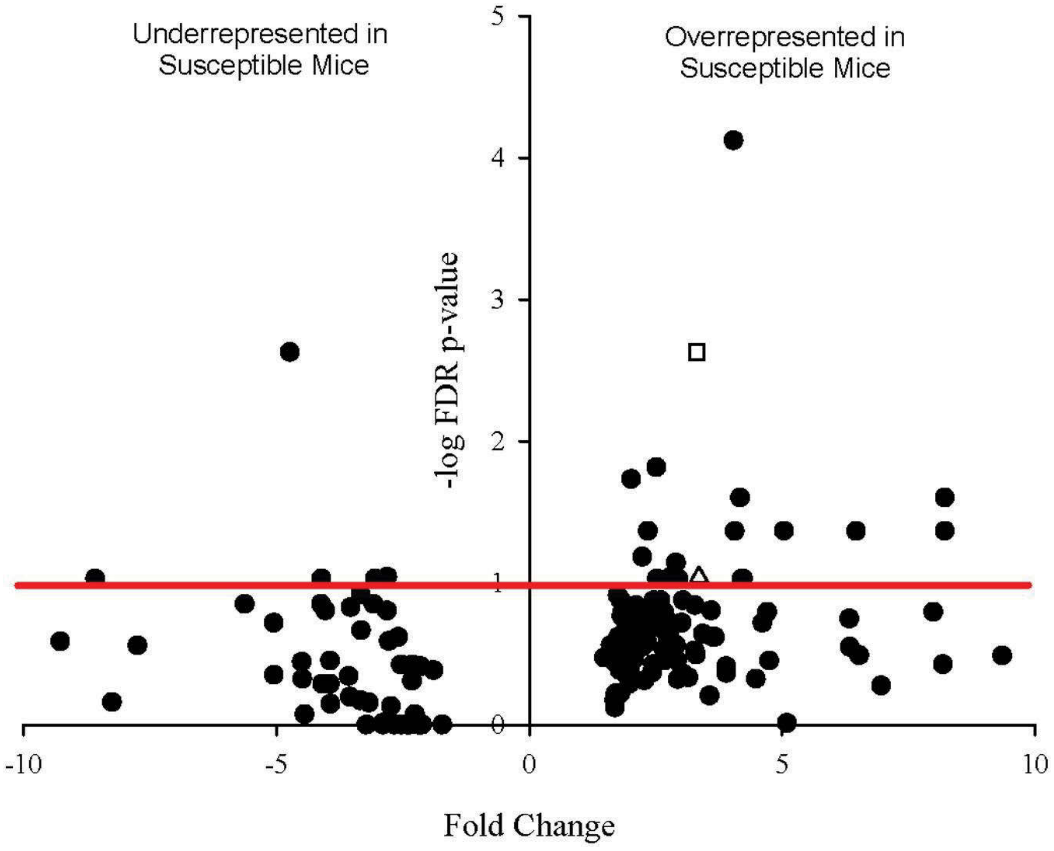

Figure 5.6. Volcano plot showing degree of differential expression of mouse-derived genes in Charles River Laboratories mice compared to Taconic Biosciences. Log-transformed fold change in expression is plotted on the x-axis and log-transformed false discovery rate-adjusted p-values plotted on the y-axis. The red horizontal line represents the 0.1 p-value cutoff. Empty square: *bsg* (Basigin). Empty triangle: *lgals9* (Galectin-9).

substrates involved in the biosynthesis of these compounds. While a number of additional transcripts and metabolites were differentially abundant between mouse substrains, we have restricted our discussion to only those where a mechanism influential in gut microbial symbiosis, immune regulation, and malaria infection are clear.

Discussion

Previous studies have demonstrated that the microbiome of C57BL/6 mice can modulate the severity of *Plasmodium* infections in mice (Villarino *et al.*, 2016). The resistant and susceptible phenotypes were not only reproducible across cohorts, but transmissible as part of cecal transplants to germ-free mice. Differences in parasite burden and bacterial community composition of Taconic and Charles River mice in the current study were consistent with previous research. Taconic mice exhibited significantly lower peak parasite burden and recovered from infection more quickly than Charles River mice. These findings strongly suggest that, as with our previous study, differences in parasite burden are the result of some currently unidentified interaction between the host and the gut microbiota, rather than the effects of epigenetic regulation, genetic or environmental effects. However, differential expression of mouse genes and differential abundance of metabolite pools are purely associative until further gut transplant studies are carried out.

Phylogenetically, the vast majority of transcripts were produced by bacteria, with Bacteroidetes and Firmicutes the most abundant among them. And while reliance on transcript abundance as an indicator of community composition is tenuous, the data are consistent with 16S rRNA and metagenomic studies of both mice and humans (Backhed *et al.*, 2005, Ley *et al.*, 2006, Sekirov *et al.*, 2010). Overall, community composition inferred from transcript abundance did not differ at the phylum level between mice from the two vendors sampled. However,

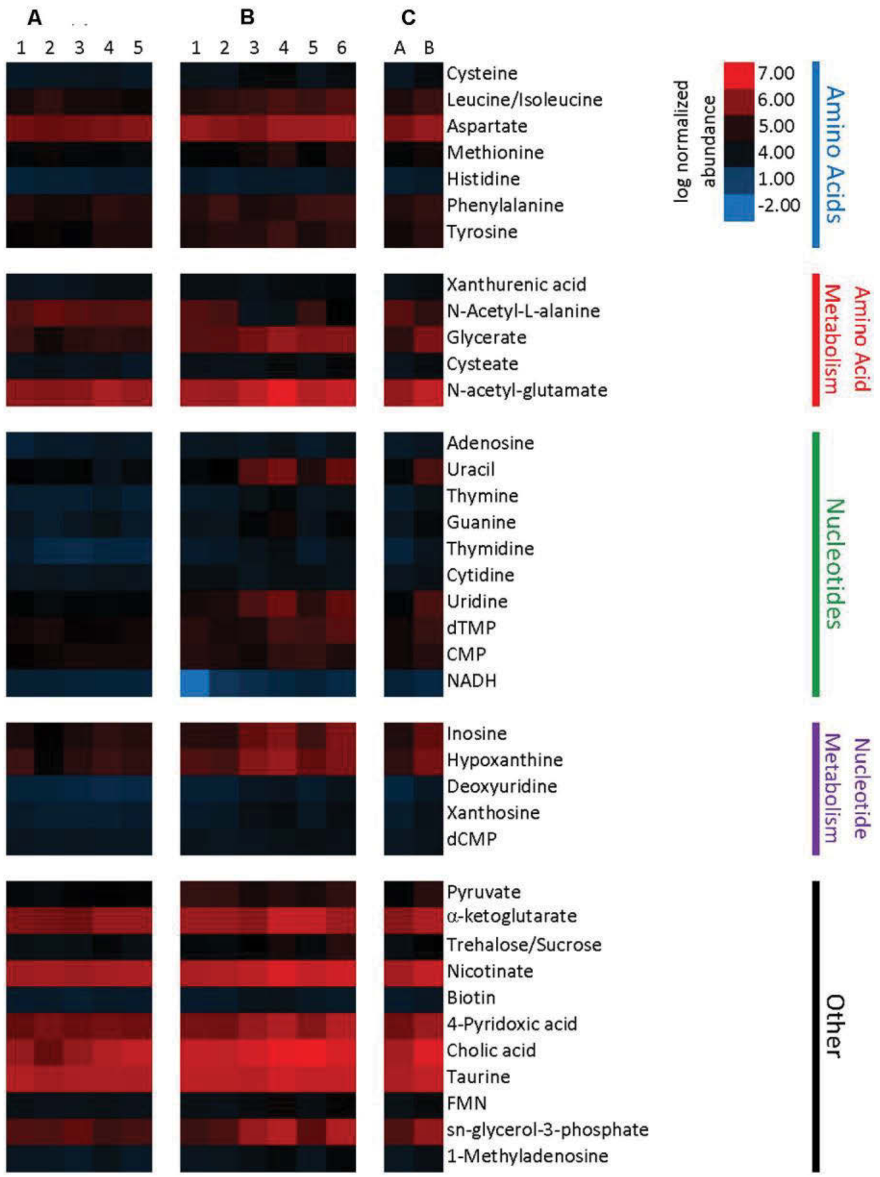

Figure 5.7. Heatmap representing metabolite abundances normalized to sample tissue mass and log transformed. Metabolites displayed are significantly different with a p-value cutoff of 0.1. A: five columns represent metabolite abundances for each of five Taconic Biosciences mice.B: six columns represent each of the six Charles River mice. C. Columns represent the mean abundances for Taconic (A) and Charles River (B).

relevant differences in community functional profiles from overall expression patterns suggest that the factors involved in affecting host phenotype may exist at a finer scale.

Within the context of our study neither Archaea nor viruses make up a significant portion of transcriptional activity, although their contribution cannot be discounted. Previous studies have also shown their abundance is lower than their bacterial counterparts (Hoffmann *et al.*, 2013); however, it is likely that this community was not sequenced deeply enough to detail their role. Viruses in particular may require targeted approaches to better resolve their influence on community dynamics and host phenotype. The role of phage populations may be limited to top-down control of the bacterial community with no direct influence over host cells (Ogilvie & Jones, 2015).

Differential bacterial gene expression in the cecum, in part, reflects differences in microbial community composition between mouse strains that are often used interchangeably in research and provides important targets to unveil the mechanism underlying resistance to malaria. Overrepresentation of transcripts encoding flagellin in Charles River mice suggests a mechanism that may involve indirect modulation of the immune system by the gut microflora. Flagellin is the principal protein component of the bacterial flagellum, encoded by the gene *fliC*. While the majority of the gut microbial diversity is capable of producing flagella, flagellin levels are generally low in the healthy gastrointestinal tract (Verberkmoes *et al.*, 2009). Increased flagellin expression can be associated with mucosal barrier breakdown and inflammation (Sanders, 2005, Gewirtz, 2006). It has been hypothesized that anti-flagellin antibodies down-regulate *fliC* expression in resident non-pathogenic microbes (Cullender *et al.*, 2013) and this prevents colonization by potential pathogens (Ghose *et al.*, 2016). However, it is currently unclear whether local stimulation of innate and adaptive immune response in the gut *via* Toll-like receptor 5 (TLR5) (Gewirtz *et al.*, 2006) is relevant to the immune response to *Plasmodium* infection.

Differential regulation of murine gene expression between groups of mice purchased from different vendors is compelling evidence of non-genomic C57BL/6N strain divergence. Of particular interest is the overrepresentation of basigin (BSG) in Charles River mice and its possible involvement in malaria resistance. Also referred to as CD147 or EMMPRIN (extracellular matrix metalloprotease inducer), basigin is a cell surface receptor in the immunoglobulin superfamily. It is commonly expressed on many tissue types and is involved in a wide variety of biological functions, such developmental processes, nutrient transport, and inflammation (Xiong *et al.*, 2014, Hahn *et al.*, 2015). The basigin gene, *bsg*, can encode four different variants through alternative splicing, each of which is expressed in different tissues (Liao *et al.*, 2011). Subsequent assembly and analysis of Basigin transcripts from our dataset identified that the vast majority of reads encoded isoform Bsg-2, the most abundant and best characterized isoform in human and mouse tissue. While basigin is involved in many processes, it became relevant to human health when it was found to induce expression of matrix metalloproteases, which can promote tumor cell development, invasion, and metastasis (Hahn *et al.*, 2015). Perhaps more relevant to the current work, a recent study identified Bsg-2 as a key receptor for reticulocyte-binding protein homologue 5 (PfRh5), the parasite ligand required for erythrocyte invasion by *Plasmodium falciparum* (Crosnier *et al.*, 2011). In total these observations results in the new hypothesis that decreased expression of basigin isoform Bsg-2 in Taconic mice may contribute to their malaria resistance.

Another overrepresented transcript in Charles River mice encodes the β-galactoside-binding protein galectin-9. Galectins bind specifically to glycosylated proteins and are typically involved in cell signaling and regulation. As a result, dysfunction of galectin activity and expression is closely linked to cancer development (Thijssen *et al.*, 2015) and autoimmune disorders (Blidner *et al.*, 2015). As a ligand for the type-I glycoprotein Tim-3, galectin-9 modulates the innate immune response to viral infection by inducing apoptosis in infected T cells (Merani *et al.*, 2015). Dysfunctional expression and activation of the Tim-3 signaling molecule has been linked to $CD4^+$ and $CD8^+$ T cell "exhaustion" in chronic HIV (Jones *et al.*, 2008) and hepatitis C (Golden-Mason *et al.*, 2009) infection. It is possible that underrepresentation of galectin-9 in Taconic mice may improve T cell response to *Plasmodium* infection. However, interest in galectin proteins as important immune signaling molecules has emerged only recently. As the regulation of these proteins is poorly understood, the mechanism by which the gut microbiota may influence galectin expression is unclear.

As part of our analysis we mapped both transcripts and metabolite data ($p \leq 0.1$) onto microbial metabolic pathways to identify biological processes that may link the two. However, we were unable to find connections beyond two or three features within any pathway. This may be due to the relatively low transcript coverage of the vast metabolic capabilities of the microbiome, but is likely also related to the transient nature of gut contents and the constant flux of new material combined with the temporal disconnect between transcriptional and metabolic responses. Additionally, it can be difficult to determine whether differential relative concentrations of specific molecules are the cause or result of physiological change. However, the presence of significant differences in specific gut metabolites, as well as relevant difference in overall metabolite pools, between C57BL/6N mice is of serious concern to those that rely on them for reproducibility. Previous work has also shown that the murine microbiome can alter the concentration of circulating metabolite in the host (Villarino *et al.*, 2016), further complicating the comparison of results between vendors and substrains.

This study identified key differences in the gene expression of both the microbial and murine components of the gastrointestinal tract, including the cell surface receptor basigin, as a potential link between the gut microbiome and the previously observed malaria resistance. Differential expression of the immune signaling protein galectin-9 was also noted, and this alteration may play a role in regulation of the differential immune response observed in the prior study. Additionally, a relevant difference in the overall metabolome and significant differences in multiple individual metabolites were observed. While the differences in gene expression and metabolism we observed provide evidence against the interchangeability of mice obtained from different vendors, they shed new light on potential avenues for investigation into the effects of the microbiome on the severity of malaria.

Acknowledgements

This work was supported by NIH Grant 1R21AI113386 and funds from the Kenneth & Blaire Mossman endowment to the University of Tennessee.

Chapter V Appendix

Table 5.1. Significantly differentially expressed mouse and bacterial genes (p ≤ 0.1), log Fold Change, and False Discovery Rate Adjusted p-values.

Source	Gene Name	Gene Product	logFC	FDR	Effect Size
Bacteria	GATM	glycine amidinotransferase	10.56203129	9.48E-07	1.3052741
	SpeA	arginine decarboxylase	7.074700747	3.60E-05	1.5428224
	NuoM	NADH-quinone oxidoreductase subunit M	6.231465975	0.000180273	1.4258402
	TusE/DsrC	tRNA 2-thiouridine synthesizing protein E	8.902038577	0.000200091	1.6800906
	PurT	phosphoribosylglycinamide formyltransferase 2	7.283113669	0.00039202	1.4756676
	DsrB	sulfite reductase beta subunit	8.385858709	0.000427852	1.6403626
	NuoN	NADH-quinone oxidoreductase subunit N	6.573771144	0.000530932	1.624988
	UreC	urease subunit alpha	4.7696686	0.000557846	1.440464
	Buk	butyrate kinase	2.102034749	0.000614343	1.6461527
	FTCD	glutamate formiminotransferase / formiminotetrahydrofolate cyclodeaminase	6.359338898	0.000614343	1.2068993
	GdhA	glutamate dehydrogenase (NADP+)	1.18441685	0.000614343	2.1202833
	NuoK	NADH-quinone oxidoreductase subunit K	7.370181564	0.002263382	1.6583615
	GlpQ/UgpQ	glycerophosphoryl diester phosphodiesterase	4.062422123	0.002263382	1.0715194
	UreA	urease subunit gamma	5.435311192	0.003406626	1.3049208
	GlcD	glycolate oxidase	-2.368672855	0.00412958	1.2855642
	RegX3	two-component system, OmpR family, response regulator RegX3	5.529495891	0.00467813	1.0360145
	NuoB	NADH-quinone oxidoreductase subunit B	3.953272974	0.005621745	1.5099251
	NuoL	NADH-quinone oxidoreductase subunit L	4.01725117	0.005621745	1.275322
	UreB	urease subunit beta	4.46892511	0.006184066	1.6861034
	IolB	5-deoxy-glucuronate isomerase	5.954879914	0.006332655	1.4514659
	FliC	flagellin	1.233262625	0.006332655	2.108153
	PTS-Aga-EIIC/AgaW	PTS system, N-acetylgalactosamine-specific IIC component	4.802058041	0.007779713	1.6118904
	SerC/PSAT1	phosphoserine aminotransferase	1.074156786	0.008473378	1.8291656
	GlpC	glycerol-3-phosphate dehydrogenase subunit C	6.354592197	0.009210915	1.5117792

Table 5.1. Continued.

Source	Gene Name	Gene Product	logFC	FDR	Effect Size
Bacteria	SDHA/SDH1	succinate dehydrogenase (ubiquinone) flavoprotein subunit	-4.432885653	0.009210915	0.4563344
	GctA	glutaconate CoA-transferase, subunit A	7.53547712	0.009404321	1.0240919
	NuoH	NADH-quinone oxidoreductase subunit H	4.523208595	0.013254562	1.3783323
	Ptb	phosphate butyryltransferase	2.124189103	0.016398458	1.4304567
	YgeU/XdhC	xanthine dehydrogenase iron-sulfur-binding subunit	2.154890651	0.018858598	3.3008939
	Eda	2-dehydro-3-deoxyphosphogluconate aldolase / (4S)-4-hydroxy-2-oxoglutarate aldolase	1.082844645	0.022255509	2.1391447
	E3.2.1.24	alpha-mannosidase	-1.291538009	0.025748223	2.0848625
	ALAS	5-aminolevulinate synthase	5.379671095	0.03089919	0.5861588
	DsrA	sulfite reductase alpha subunit	6.39127666	0.048137885	0.9054653
	Cgn	Cingulin	2.881220804	0.05284059	1.6751635
	MmsA/IolA/ALDH6A1	malonate-semialdehyde dehydrogenase (acetylating) / methylmalonate-semialdehyde dehydrogenase	-2.060221714	0.05284059	1.2174738
	ATPF1A/AtpA	F-type H+-transporting ATPase subunit alpha	0.844771592	0.05284059	1.3782354
	RP-S15/MRPS15/RpsO	Small Subunit Ribosomal Protein	1.062973115	0.061089357	1.6536516
	PatA	putrescine aminotransferase	2.222866162	0.061196835	1.5502705
	ITPK1	inositol-1,3,4-trisphosphate 5/6-kinase / inositol-tetrakisphosphate 1-kinase	3.690990259	0.068577927	0.6381211
	NuoA	NADH-quinone oxidoreductase subunit A	3.840319059	0.069622603	1.2586251
	LYS1	saccharopine dehydrogenase (NAD+, L-lysine forming)	1.148515411	0.072628937	1.1809167
	CheV	two-component system, chemotaxis family, response regulator CheV	1.137598186	0.072628937	1.7032707
	ABC-2.A	ABC-2 type transport system ATP-binding protein	1.268058818	0.072628937	1.1852107
	Enr	2-enoate reductase	-1.816291338	0.072628937	1.9947845
	PsaA	photosystem I P700 chlorophyll a apoprotein A1	3.908347932	0.074550244	0.8354202
	DapD	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase	1.972015719	0.076022433	1.4479565
	ALDO	fructose-bisphosphate aldolase, class I	-1.702917526	0.076022433	1.4723451
	MDH1	malate dehydrogenase	2.622145266	0.083780158	1.2696492
	RP-S6/MRPS6/RpsF	Small subunit ribosomal protein S6	0.849315223	0.08574694	2.3462309

Table 5.1. Continued.

Source	Gene Name	Gene Product	logFC	FDR	Effect Size
Bacteria	ATPF1B/AtpB	F-type H+-transporting ATPase subunit beta	0.738843482	0.08574694	1.3055021
	NadX	aspartate dehydrogenase	-2.056355974	0.08574694	1.9236227
	PanE/ApbA	2-dehydropantoate 2-reductase	-1.862885298	0.08574694	1.5857412
	Eno	enolase	0.677777702	0.08574694	1.0884583
	ArgG/ASS1	argininosuccinate synthase	0.921274891	0.08574694	1.163186
	FabB	3-oxoacyl-[acyl-carrier-protein] synthase I	3.561095503	0.08574694	1.1025828
	TatC	sec-independent protein translocase protein TatC	-1.55099591	0.08574694	2.0887403
	KdsA	2-dehydro-3-deoxyphosphooctonate aldolase (KDO 8-P synthase)	1.547237925	0.087896354	1.3833127
	FucK	L-fuculokinase	3.481884866	0.089880237	1.5672104
	ABC.PE.S	peptide/nickel transport system substrate-binding protein	0.835541407	0.089880237	2.2754899
	Fhs	formate--tetrahydrofolate ligase	0.703741962	0.094662643	1.4771
Mouse	Ahcyl2	adenosylhomocysteinase	6.11069988	7.54438E-05	3.0206194
	Bsg	Basigin	5.021955403	0.002347438	1.2340425
	Rsrp1	Arginine/serine-rich protein 1	-3.492567218	0.002347438	1.9022358
	Ndufa7	NADH dehydrogenase 1 alpha subcomplex subunit 7	3.823154723	0.015297121	3.1831989
	Cyp2c55	Cytochrome P450 2C55	2.962176928	0.018513833	3.4631036
	Gsdmc4	Gasdermin-C4	6.062150002	0.025095072	2.7076045
	Ndufb8	NADH dehydrogenase 1 beta subcomplex subunit 8	13.81400494	0.025095072	1.1636012
	Hmgcs2	Hydroxymethylglutaryl-CoA synthase	3.469436143	0.042944276	2.7083165
	Rpl38	60S Ribosomal protein L38	7.647852391	0.042944276	1.5492865
	Azin1	Antizyme Inhibitor 1	6.290758297	0.043120905	2.1475965
	Hadhb	Trifunctional enzyme subunit beta	10.51963033	0.043120905	2.2005213
	Psmb4	Proteasome Subunit beta type-4	3.343521152	0.065559612	2.54816
	H3f3a	Histone H3.3	4.133121776	0.090810428	1.4220731
	Lgals9	Galectin-9	5.108298738	0.090810428	2.1805035
	Arpc1a	Actin-related protein 2/3 complex subunit 1A	4.444238214	0.093437914	1.5775638
	Hist1h4h	Histone H4	3.780447723	0.093437914	2.4061417
	Ndufs4	NADH dehydrogenase iron-sulfur protein 4	3.810497174	0.093437914	2.9942464
	Pcca	Propionyl-CoA carboxylase alpha chain	6.612551547	0.093437914	1.9207781
	Pnkd	Probable Hydrolase PNKD	6.633368035	0.093437914	2.6403529
	Slc25a20	Mitochondrial carnitine/acylcarnitine carrier protein	-2.961285001	0.093437914	1.3733935

Table 5.2. Fold Change and Cluster Analysis p-values for individual metabolites.

Compound	Pathway	Fold Change	p value	Effect Size
cysteine	Amino Acid	4.264451517	0.048392497	1.482565924
leucine/isoleucine	Amino Acid	2.436147564	0.022166626	1.743857484
aspartate	Amino Acid	2.007974347	0.009882875	2.078749142
methionine	Amino Acid	1.891779813	0.075978816	1.224121478
histidine	Amino Acid	2.261762804	0.063734379	1.305951153
phenylalanine	Amino Acid	1.64334625	0.07620887	1.22177407
tyrosine	Amino Acid	1.814098773	0.047735051	1.400186304
Xanthurenic acid	Amino Acid Metabolism	2.110636707	0.030706401	1.542142043
N-Acetyl-L-alanine	Amino Acid Metabolism	0.384640226	0.011383081	1.917705446
glycerate	Amino Acid Metabolism	6.255104016	0.011354755	2.228397124
Cysteate	Amino Acid Metabolism	2.585023342	0.079345053	1.23520779
N-acetyl-glutamate	Amino Acid Metabolism	2.81404649	0.083747638	1.2273061
uracil	Nucleotide	10.66350413	0.070807952	1.319535042
thymine	Nucleotide	7.113874218	0.085233507	1.232174327
guanine	Nucleotide	5.704544345	0.089711477	1.200954906
thymidine	Nucleotide	12.31455573	0.043687853	1.535083428
cytidine	Nucleotide	2.073567308	0.010745541	2.103744183
uridine	Nucleotide	8.48703799	0.044962788	1.533516235
dTMP	Nucleotide	2.916461634	0.050213852	1.476807495
CMP	Nucleotide	1.710297432	0.026978638	1.716741428
NADH	Nucleotide	0.341028811	0.006417009	2.121411393
adenosine	Nucleotide	2.302531269	0.061374337	1.298884329
inosine	Nucleotide Metabolism	5.002214937	0.059828075	1.392609086
hypoxanthine	Nucleotide Metabolism	5.085169781	0.027413684	1.749107264
deoxyuridine	Nucleotide Metabolism	10.38925436	0.051010272	1.472521792
xanthosine	Nucleotide Metabolism	7.301096662	0.052336967	1.459539016
dCMP	Nucleotide Metabolism	2.092157232	0.014158991	2.06543616
pyruvate	Glycolysis/TCA cycle	3.501114712	0.0031527	2.829660522
a-ketoglutarate	Glycolysis/TCA cycle	2.272752652	0.064501908	1.312544356
trehalose/sucrose	Carbon Metabolism	3.461951999	0.065543262	1.322667299
nicotinate	Vitamin	2.054801917	0.029668976	1.7110114
biotin	Vitamin	2.415017868	0.088746356	1.20510263
4-Pyridoxic acid	Vitamin Metabolism	2.554636577	0.064361789	1.356401173
Cholic acid	Bile Acid	3.786874682	0.026447811	1.694773436
Taurine	Bile Acid Biosynthesis	1.655596046	0.001840094	2.926294892
FMN	Oxydative Phosphorylation	2.718121434	0.096874767	1.17243847
sn-glycerol-3-phosphate	Glycerolipid Biosynthesis	4.255672265	0.098616854	1.166615844

Table 5.2. Continued.

Compound	Pathway	Fold Change	p value	Effect Size
1-Methyladenosine	Other	3.51953833	0.083512912	1.226670205
2-Hydroxy-2-methylbutanedioic acid	Other	3.169547335	0.068111382	1.284917215

CHAPTER VI:

CONCLUSION

While jumping between disparate microbial communities and environments, the work described here is unified behind the themes of 1.) developing targeted approaches that allow researchers to reduce the size and complexity of the notoriously large high-throughput sequencing datasets in order to propose hypotheses that can be tested in a field or laboratory setting and 2.) discovering or predicting the relationships between microbes and their hosts, with a particular focus on viruses. As a whole, this work has established a firm foundation for future refinement of these methods and application to as yet unexamined microbial communities.

Building off of previous studies suggesting the strong influence of viral activity on *Microcystis aeruginosa* blooms (Steffen *et al.*, 2015, Steffen *et al.*, 2017), we used currently available metatranscriptome sequencing data isolated from the hypereutrophic Lake Tai, China to search for *Microcystis*-specific viruses and characterize their activity during cyanobacterial blooms. During this study, we discovered transcriptional patterns, consistent across temporal and geographic scales, that suggested rampant lysogeny occurs during bloom season. These results further suggest a series of viral expression markers that could be used to further predict lytic and lysogenic activity in *Microcystis* phage. As lysogeny may protects the host from subsequent infection by other lytic phage, these observations may provide an explanation for *Microcystis* success as a bloom former and its ability to defy Hutchinson's paradox of the plankton and the kill-the-winner-hypothesis (Hutchinson, 1961, Thingstad & Lignell, 1997).

While the microbial diversity in cyanobacterial blooms is relatively well described, study of the microbiome of *Sphagnum* peat bog environments is in its infancy (Kostka *et al.*, 2016). As the viral constituents of these ecosystems are almost entirely unknown, we applied a pipeline in development to characterize viruses from multiple taxa using currently available metatranscriptome sequencing isolated from *Sphagnum* tissue in Northern Minnesota. In this study, we were able to identify a broad diversity of both DNA and RNA viruses. As viral mRNA molecules are only produced during active infections, we were also able to use the relative abundances of viral transcripts to build correlation co-occurrence networks in order to predict the hosts of many of the identified viruses. This viral diversity and these predicted relationships now stand as proposed hypotheses that can be tested in future, substantially simplifying a complex microbial community into a group of potential model systems for the study of virus host interactions in *Sphagnum*.

One such example of a potential model system in freshwater ecosystems is the haptophyte *Chrysochromulina parva*, its virus CpV and virophage CpVV. To lay the foundation for future culture-based studies, we sequenced, assembled, and annotated the genomes of CpV and CpVV. Therein we observed a versatile giant virus clearly originating in an emerging clade within the NCLDV group. CpV possessed an expanded potential for the genetic and regulatory modification of its host, with corresponding capabilities in its virophage. Subsequent examination of CpV activity using currently available metatranscriptome sequencing data from freshwater ecosystems revealed significant activity in Lake Tai, China during the cyanobacterial bloom in 2013. These results suggest that CpV is globally distributed and an important contributor to freshwater ecosystems.

Lastly, in order to determine the contribution of the gut microbiome to malaria resistance in mice, we isolated and sequenced the bacterial community and mouse metatranscriptomes. During our analysis we were able to identify multiple genes potentially involved in the interface between gut microbes and their host that may contribute to resistance to infection by *Plasmodium*. These results also have significant implications in the selection of mouse strains for scientific research, as even genetically identical mice can have radically different phenotypes depending on the activity of their gut microbes. Altogether this body of work establishes a collection of powerful methods for targeting specific organisms and activities in diverse microbial ecosystems, and proposes hypotheses that advance the understanding of the environments studied herein.

LIST OF REFERENCES

Abedon ST (1992) Lysis of lysis-inhibited bacteriophage T4-infected cells. *J Bacteriol* **174**: 8073-8080.

Abedon ST (1999) Bacteriophage T4 resistance to lysis-inhibition collapse. *Genetics Research* **74**: 1-11.

Adams MD, Kelley JM, Gocayne JD*, et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651-1656.

Alexander H, Jenkins BD, Rynearson TA & Dyhrman ST (2015) Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences* **112**: E2182-E2190.

Alikhan NF, Petty NK, Ben Zakour NL & Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**: 10.

Altman PL & Kats DD (1979) *Inbred and Genetically Defined Strains of Laboratory Animals, Part 1 Mouse and Rat*. Federation of American Societies for Experimental Biology, Bethesda, Maryland.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Anderson NL & Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**: 1853-1861.

Anderson NL & Anderson NG (2002) The human plasma proteome - History, character, and diagnostic prospects. *Molecular & Cellular Proteomics* **1**: 845-867.

Anderson RT, Vrionis HA, Ortiz-Bernad I, Resch CT, Long PE, Dayvault R, Karp K, Marutzky S, Metzler DR & Peacock A (2003) Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Applied and environmental microbiology* **69**: 5884-5891.

Avery OT, Macleod CM & McCarty M (1944) STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med* **79**: 137-158.

Aziz RK, Bartels D, Best AA*, et al.* (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75-75.

Backhed F, Ley RE, Sonnenburg JL, Peterson DA & Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* **307**: 1915-1920.

Bailey DW (1978) Sources of Subline Divergence and their Relative Importance for Sublines of Six Major Inbred Strains of Mice. . *Origins of Inbred Mice,*(Morse III HC, ed.) p.^pp. 197-215. Academic Press, New York.

Baltimore D (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**: 1209-1211.

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

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP & Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. *Genome research* **12**: 177-189.

Beres SB & Musser JM (2007) Contribution of Exogenous Genetic Elements to the Group A Streptococcus Metagenome. *PloS one* **2**: 14.

Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S & Cohen SN (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99**: 9697-9702.

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6-21.

Bishop CT, Anet EFLJ & Gorham PR (1959) Isolation and identification of the Fast-Death Factor in *Microcystis aeruginosa* NRC-1. *Can J Biochem Physiol* **37**: 453-471.

Blackstock WP & Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol* **17**: 121-127.

Blidner AG, Mendez-Huergo SP, Cagnoni AJ & Rabinovich GA (2015) Re-wiring regulatory cell networks in immunity by galectin-glycan interactions. *Febs Letters* **589**: 3407-3418.

Bobay LM, Rocha EPC & Touchon M (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* **30**: 737-751.

Boocock MR & Rice PA (2013) A proposed mechanism for IS607-family serine transposases. *Mob DNA* **4**: 9.

Boschker H, Nold S, Wellsbury P & Bos D (1998) Direct linking of microbial populations to specific biogeochemical processes by 13C-labelling of biomarkers. *Nature* **392**: 801.

Boyer GL (2007) The occurrence of cyanobacterial toxins in new York lakes: Lessons from the MERHAB-Lower great lakes. *Lake Reserv Manag* **23**: 153-160.

Boyer M, Madoui MA, Gimenez G, La Scola B & Raoult D (2010) Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4(th) Domain of Life Including Giant Viruses. *PloS one* **5**: 8.

Boyer M, Yutin N, Pagnier I*, et al.* (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A* **106**: 21848-21853.

Bragina A, Cardinale M, Berg C & Berg G (2013) Vertical transmission explains the specific Burkholderia pattern in Sphagnum mosses at multi-geographic scale. *Frontiers in microbiology* **4**: 10.

Bragina A, Maier S, Berg C, Muller H, Chobot V, Hadacek F & Berg G (2012) Similar diversity of Alphaproteobacteria and nitrogenase gene amplicons on two related Sphagnum mosses. *Frontiers in microbiology* **3**: 10.

Brittain SM, Wang J, Babcock-Jackson L, Carmichael WW, Rinehart KL & Culver DA (2000) Isolation and characterization of Microcystins, cyclic heptapeptide hepatotoxins from a Lake Erie strain of *Microcystis aeruginosa*. *Journal of Great Lakes Research* **26**: 241-249.

Brulc JM, Antonopoulos DA, Miller MEB*, et al.* (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A* **106**: 1948-1953.

Brum JR, Hurwitz BL, Schofield O, Ducklow HW & Sullivan MB (2016) Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *The ISME journal* **10**: 437-449.

Brunberg AK (1999) Contribution of bacteria in the mucilage of Microcystis spp. to benthic and pelagic production in a hypereutrophic lake. *FEMS microbiology ecology* **29**: 13-22.

Brussaard CPD, Wilhelm SW, Thingstad F*, et al.* (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *The ISME journal* **2**: 575-578.

Canard B & Sarfati RS (1994) DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**: 1-6.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N & Knight R (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* **108**: 4516-4522.

Caporaso JG, Kuczynski J, Stombaugh J*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**: 335-336.

Carmichael W (1996) Toxic *Microcystis* and the environment. *Toxic microcystis* 1-11.

Chambers MC, Maclean B, Burke R*, et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**: 918-920.

Chaurushiya MS, Lilley CE, Aslanian A*, et al.* (2012) Viral E3 Ubiquitin Ligase-Mediated Degradation of a Cellular E3: Viral Mimicry of a Cellular Phosphorylation Mark Targets the RNF8 FHA Domain. *Mol Cell* **46**: 79-90.

Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB & Welschmeyer NA (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**: 340-343.

Clark KR & Gorley RN (2015) *PRIMER v7: User manual/tutorial*. PRIMER-E, Plymouth.

Clarke KR, Somerfield PJ & Gorley RN (2008) Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. *J Exp Mar Biol Ecol* **366**: 56-69.

Clasquin MF, Melamud E & Rabinowitz JD (2002) LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine. *Current Protocols in Bioinformatics,*p.^pp. John Wiley & Sons, Inc.

Claverie JM & Abergel C (2010) Mimivirus: the emerging paradox of quasi-autonomous viruses. *Trends Genet* **26**: 431-437.

Clokie MRJ, Shan JY, Bailey S, Jia Y, Krisch HM, West S & Mann NH (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental microbiology* **8**: 827-835.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J.

Comeau AM & Krisch HM (2008) The capsid of the T4 phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol Biol Evol* **25**: 1321-1332.

Crawley JN, Belknap JK, Collins A*, et al.* (1997) Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology* **132**: 107-124.

Crick FH, Barnett L, Brenner S & Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* **192**: 1227-1232.

Crosnier C, Bustamante LY, Bartholdson SJ*, et al.* (2011) Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature* **480**: 534-U158.

Cullender TC, Chassaing B, Janzon A*, et al.* (2013) Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**: 571-581.

D Ainsworth T, Krause L, Bridge T*, et al.* (2015) The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *The ISME journal* **9**: 2261-2274.

de Hoon MJL, Imoto S, Nolan J & Miyano S (2004) Open source clustering software. *Bioinformatics* **20**: 1453-1454.

de Saussure N (1804) Recherches chimiques sur la vegetation. 327.

Doolittle WF & Papke RT (2006) Genomics and the bacterial species problem. *Genome biology* **7**: 116.

Dorigo U, Jacquet S & Humbert JF (2004) Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. *Applied and environmental microbiology* **70**: 1017-1022.

Driks A (1999) Bacillus subtilis spore coat. *Microbiology and Molecular Biology Reviews* **63**: 1-20.

Dudova L, Hajkova P, Buchtova H & Opravilova V (2013) Formation, succession and landscape history of Central-European summit raised bogs: A multiproxy study from the Hruby Jesenik Mountains. *Holocene* **23**: 230-242.

Eden PA, Schmidt TM, Blakemore RP & Pace NR (1991) Phylogenetic analysis of Aquaspirillum magnetotacticum using polymerase chain reaction-amplified 16S rRNA-specific DNA. *International journal of systematic bacteriology* **41**: 324-325.

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

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* **5**: 1-19.

Erez Z, Steinberger-Levy I, Shamir M*, et al.* (2017) Communication between viruses guides lysis-lysogeny decisions. *Nature* **541**: 18.

Etten JLV & Meints RH (1999) Giant Viruses Infecting Algae. *Annual Review of Microbiology* **53**: 447-494.

Feldmann H, Aigle M, Aljinovic G*, et al.* (1994) Complete DNA sequence of yeast chromosome II. *The EMBO journal* **13**: 5795-5809.

Fiehn O (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and functional genomics* **2**: 155-168.

Fiehn O, Kloska S & Altmann T (2001) Integrated studies on plant biology using multiparallel techniques. *Current opinion in biotechnology* **12**: 82-86.

Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA & Knight R (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *Isme J* **6**: 1007-1017.

Filee J, Pouget N & Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol Biol* **8**: 13.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A & Eddy SR (2015) HMMER web server: 2015 update. *Nucleic Acids Research* **43**: W30-W38.

Finn RD, Coggill P, Eberhardt RY*, et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**: D279-D285.

Finn RD, Attwood TK, Babbitt PC*, et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* **45**: D190-D199.

Fischer MG (2012) Sputnik and Mavirus: more than just satellite viruses. *Nat Rev Micro* **10**: 78-78.

Fischer MG & Suttle CA (2011) A virophage at the origin of large DNA transposons. *Science* **332**: 231-234.

Fischer SG & Lerman LS (1979) Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. *Cell* **16**: 191-200.

Fischer SG & Lerman LS (1980) Separation of random fragments of DNA according to properties of their sequences. *Proceedings of the National Academy of Sciences* **77**: 4420-4424.

Freeman C, Ostle N & Kang H (2001) An enzymic 'latch' on a global carbon store - A shortage of oxygen locks up carbon in peatlands by restraining a single enzyme. *Nature* **409**: 149-149.

Freeman HC, Hugill A, Dear NT, Ashcroft FM & Cox RD (2006) Deletion of nicotinamide nucleotide transhydrogenase - A new quantitive trait locus accounting for glucose intolerance in C57BL/6J mice. *Diabetes* **55**: 2153-2156.

Freeman WM, Walker SJ & Vrana KE (1999) Quantitative RT-PCR: pitfalls and potential. *BioTechniques* **26**: 112-122, 124-115.

Fuhrman JA & Steele JA (2008) Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquatic Microbial Ecology* **53**: 69-81.

Fulwyler MJ (1965) Electronic separation of biological cells by volume. *Science* **150**: 910-911.

Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D & La Scola B (2014) Zamilon, a Novel Virophage with Mimiviridae Host Specificity. *PloS one* **9**: 8.

Galka M, Tobolski K, Gorska A & Lamentowicz M (2017) Resilience of plant and testate amoeba communities after climatic and anthropogenic disturbances in a Baltic bog in Northern Poland: Implications for ecological restoration. *Holocene* **27**: 130-141.

Gardner RS, Wahba AJ, Basilio C, Miller RS, Lengyel P & Speyer JF (1962) Synthetic polynucleotides and the amino acid code. VII. *Proc Natl Acad Sci U S A* **48**: 2087-2094.

Gastrich MD, Leigh-Bell JA, Gobler CJ, Anderson OR, Wilhelm SW & Bryan M (2004) Viruses as potential regulators of regional brown tide blooms caused by the alga, Aureococcus anophagefferens. *Estuaries* **27**: 112-119.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P & Thompson FL (2005) Opinion: Re-evaluating prokaryotic species. *Nature reviews Microbiology* **3**: 733.

Gewirtz AT (2006) Flag in the crossroads: flagellin modulates innate and adaptive immunity. *Current Opinion in Gastroenterology* **22**: 8-12.

Gewirtz AT, Vijay-Kumar M, Brant SR, Duerr RH, Nicolae DL & Cho JH (2006) Dominant-negative TLR5 polymorphism reduces adaptive immune response to flagellin and negatively associates with Crohn's disease. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **290**: G1157-G1163.

Ghose C, Eugenis I, Sun XM, Edwards AN, McBride SM, Pride DT, Kelly CP & Ho DD (2016) Immunogenicity and protective efficacy of recombinant Clostridium difficile flagellar protein FliC. *Emerging Microbes & Infections* **5**: 10.

Gifford SM, Sharma S, Rinta-Kanto JM & Moran MA (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *Isme J* **5**: 461-472.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM & Handelsman J (2002) Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Applied and environmental microbiology* **68**: 4301-4306.

Golden-Mason L, Palmer BE, Kassam N, Townshend-Bulson L, Livingston S, McMahon BJ, Castelblanco N, Kuchroo V, Gretch DR & Rosen HR (2009) Negative Immune Regulator Tim-3 Is Overexpressed on T Cells in Hepatitis C Virus Infection and Its Blockade Rescues Dysfunctional CD4(+) and CD8(+) T Cells. *Journal of Virology* **83**: 9122-9130.

Good IJ (1946) Normal Recurring Decimals. *Journal of the London Mathematical Society* **s1-21**: 167-169.

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG & Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol* **22**: 245-252.

Green ML, Singh AV, Zhang YZ, Nemeth KA, Sulik KK & Knudsen TB (2007) Reprogramming of genetic networks during initiation of the fetal alcohol syndrome. *Developmental Dynamics* **236**: 613-631.

Grubb SC, Bult CJ & Bogue MA (2014) Mouse Phenome Database. *Nucleic Acids Research* **42**: D825-D834.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W & Gascuel O (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**: 307-321.

Hahn JN, Kaushik DK & Yong VW (2015) The role of EMMPRIN in T cell biology and immunological diseases. *Journal of Leukocyte Biology* **98**: 33-48.

Hajek T, Ballance S, Limpens J, Zijlstra M & Verhoeven JTA (2011) Cell-wall polysaccharides play an important role in decay resistance of Sphagnum and actively depressed decomposition in vitro. *Biogeochemistry* **103**: 45-57.

Handelsman J, Rondon MR, Brady SF, Clardy J & Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology* **5**: R245-249.

Hanson PJ, Riggs JS, Nettles WR*, et al.* (2017) Attaining whole-ecosystem warming using air and deep-soil heating methods with an elevated CO2 atmosphere. *Biogeosciences* **14**: 861-883.

Hargreaves KR, Anderson NJ & Clokie MRJ (2013) Recovery of viable cyanophages from the sediments of a eutrophic lake at decadal timescales. *FEMS microbiology ecology* **83**: 450-456.

Harke MJ, Jankowiak JG, Morrell BK & Gobler CJ (2017) Transcriptomic responses in the bloom-forming cyanobacterium *Microcystis* induced during exposure to zooplankton. *Applied and environmental microbiology* **83**.

Harke MJ, Steffen MM, Gobler CJ, Otten TG, Wilhelm SW, Wood SA & Paerl HW (2016) A review of the global ecology, genomics, and biogeography of the commonly toxic cyanobacterium, *Microcystis*. *Harmful Algae* **54**: 4-20.

Harrell Jr. FE (2016) Hmisc: Harrell miscellaneous. p.^pp.

Hatzenpichler R (2012) Diversity, Physiology, and Niche Differentiation of Ammonia-Oxidizing Archaea. *Applied and environmental microbiology* **78**: 7501-7510.

Hershey AD & Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology* **36**: 39-56.

Hewson I, Poretsky RS, Beinart RA*, et al.* (2009) In situ transcriptomic analysis of the globally important keystone N-2-fixing taxon Crocosphaera watsonii. *Isme J* **3**: 618-631.

HilleRisLambers J, Adler PB, Harpole WS, Levine JM & Mayfield MM (2012) Rethinking Community Assembly through the Lens of Coexistence Theory. *Annual Review of Ecology, Evolution, and Systematics, Vol 43,* Vol. 43 (Futuyma DJ, ed.) p.^pp. 227-248. Annual Reviews, Palo Alto.

Hoffmann C, Dollive S, Grunberg S, Chen J, Li HZ, Wu GD, Lewis JD & Bushman FD (2013) Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PloS one* **8**: 12.

Hsiao CL & Carbon J (1979) HIGH-FREQUENCY TRANSFORMATION OF YEAST BY PLASMIDS CONTAINING THE CLONED YEAST ARG4 GENE. *Proc Natl Acad Sci U S A* **76**: 3829-3833.

Hutchinson GE (1961) The paradox of the plankton. *The American Naturalist* **95**: 137-145.

Hutchinson GE (1961) THE PARADOX OF THE PLANKTON. *Am Nat* **95**: 137-145.

Ireland AW, Clifford MJ & Booth RK (2014) Widespread dust deposition on North American peatlands coincident with European land-clearance. *Veg Hist Archaeobot* **23**: 693-700.

Iyer LM, Balaji S, Koonin EV & Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research* **117**: 156-184.

Jiang H, Dong H, Zhang G, Yu B, Chapman LR & Fields MW (2006) Microbial Diversity in Water and Sediment of Lake Chaka, an Athalassohaline Lake in Northwestern China. *Applied and environmental microbiology* **72**: 3832-3845.

Jin X & Tu Q (1990) The standard methods for observation and analysis in lake eutrophication. *Chinese Environmental Science Press, Beijing* **240**.

Jones RB, Ndhlovu LC, Barbour JD*, et al.* (2008) Tim-3 expression defines a novel population of dysfunctional T cells with highly elevated frequencies in progressive HIV-1 infection. *J Exp Med* **205**: 2763-2779.

Jones SE & Lennon JT (2010) Dormancy contributes to the maintenance of microbial diversity. *Proc Natl Acad Sci U S A* **107**: 5881-5886.

Jousset A, Bienhold C, Chatzinotas A*, et al.* (2017) Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME journal* **11**: 853-862.

Jover LF, Effler TC, Buchan A, Wilhelm SW & Weitz JS (2014) The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat Rev Micro* **12**: 519-528.

Kaneko T, Nakajima N, Okamoto S*, et al.* (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research* **14**: 247-256.

Kau AL, Ahern PP, Griffin NW, Goodman AL & Gordon JI (2011) Human nutrition, the gut microbiome and the immune system. *Nature* **474**: 327-336.

Keane TM, Goodstadt L, Danecek P*, et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289-294.

Kerepesi C & Grolmusz V (2017) The "Giant Virus Finder" discovers an abundance of giant viruses in the Antarctic dry valleys. *Arch Virol* **162**: 1671-1676.

Kerridge PT (1925) The Use of the Glass Electrode in Biochemistry. *The Biochemical journal* **19**: 611-617.

Kersulyte D, Mukhopadhyay AK, Shirai M, Nakazawa T & Berg DE (2000) Functional organization and insertion specificity of IS607, a chimeric element of *Helicobacter pylori*. *J Bacteriol* **182**: 5300-5308.

Khisti RT, Wolstenholme J, Shelton KL & Miles MF (2006) Characterization of the ethanol-deprivation effect in substrains of C57BL/6 mice. *Alcohol* **40**: 119-126.

Kittelmann S, Seedorf H, Walters WA, Clemente JC, Knight R, Gordon JI & Janssen PH (2013) Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PloS one* **8**: e47879.

Knietsch A, Waschkowitz T, Bowien S, Henne A & Daniel R (2003) Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on Escherichia coli. *Applied and environmental microbiology* **69**: 1408-1416.

Knowles B, Silveira CB, Bailey BA*, et al.* (2016) Lytic to temperate switching of viral communities. *Nature* **531**: 466-470.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N & Ratcliff RM (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences* **105**: 2504-2509.

Koonin EV & Yutin N (2010) Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. *Intervirology* **53**: 284-292.

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

Koropatkin NM, Cameron EA & Martens EC (2012) How glycan metabolism shapes the human gut microbiota. *Nature reviews Microbiology* **10**: 323.

Kostka JE, Weston DJ, Glass JB, Lilleskov EA, Shaw AJ & Turetsky MR (2016) The Sphagnum microbiome: new insights from an ancient plant lineage. *New Phytol* **211**: 57-64.

Kourilsky P (1975) Lysogenization by bacteriophage lambda: II.-identification of genes involved in the multiplicity dependent processes. *Biochimie* **56**: 1511-1516.

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128**: 693-705.

Kowalchuk GA, Stephen JR, DeBoer W, Prosser JI, Embley TM & Woldendorp JW (1997) Analysis of ammonia-oxidizing bacteria of the beta subdivision of the class Proteobacteria in coastal sand dunes by denaturing gradient gel electrophoresis and sequencing of PCR-amplified 16S ribosomal DNA fragments. *Applied and environmental microbiology* **63**: 1489-1497.

Krausfeldt LE, Tang X, van de Kamp J, Gao G, Bodrossy L, Boyer GL & Wilhelm SW (2017) Spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu. *FEMS Microbiol Ecol* **93**: fix024.

Kromkamp J, Konopka A & Mur LR (1988) Buoyancy in Light-Limited Continuous Cultures of Microcystis aeruginosa. *Journal of Plankton Research* **10**: 171-183.

Kruger MC, Bertin PN, Heipieper HJ & Arsene-Ploetze F (2013) Bacterial metabolism of environmental arsenic-mechanisms and biotechnological applications. *Appl Microbiol Biotechnol* **97**: 3827-3841.

Kumar S, Stecher G & Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**: 1870-1874.

Kuno S, Yoshida T, Kamikawa R, Hosoda N & Sako Y (2010) The distribution of a phage-related insertion sequence element in the cyanobacterium, *Microcystis aeruginosa*. *Microbes Environ* **25**: 295-301.

La Scola B, Desnues C, Pagnier I*, et al.* (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* **455**: 100-U165.

Lamers LPM, Bobbink R & Roelofs JGM (2000) Natural nitrogen filter fails in polluted raised bogs. *Glob Change Biol* **6**: 583-586.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M & FitzHugh W (2001) Initial sequencing and analysis of the human genome.

Legendre M, Bartoli J, Shmakova L*, et al.* (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci U S A* **111**: 4274-4279.

Legendre M, Lartigue A, Bertaux L*, et al.* (2015) In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc Natl Acad Sci U S A* **112**: E5327-E5335.

Leppanen S, Rissanen A & Tiirola M (2015) Nitrogen fixation in Sphagnum mosses is affected by moss species and water table level. *Plant and Soil* **389**: 185-196.

Letunic I & Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**: W242-W245.

Ley RE, Peterson DA & Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.

Liao CG, Kong LM, Song F*, et al.* (2011) Characterization of Basigin Isoforms and the Inhibitory Function of Basigin-3 in Human Hepatocellular Carcinoma Proliferation and Invasion. *Mol Cell Biol* **31**: 2591-2604.

Lin X, Tfaily MM, Green SJ, Steinweg JM, Chanton P, Imvittaya A, Chanton JP, Cooper W, Schadt C & Kostka JE (2014) Microbial Metabolic Potential for Carbon Degradation and

Nutrient (Nitrogen and Phosphorus) Acquisition in an Ombrotrophic Peatland. *Applied and environmental microbiology* **80**: 3531-3540.

Liu WZ, Xie YB, Ma JY*, et al.* (2015) IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**: 3359-3361.

Long AM & Short SM (2016) Seasonal determinations of algal virus decay rates reveal overwintering in a temperate freshwater pond. *ISME J*.

Longhurst A, Sathyendranath S, Platt T & Caverhill C (1995) An estimate of global primary production in the ocean from satellite radiometer data. *Journal of plankton Research* **17**: 1245-1271.

Lovley DR (1993) Anaerobes into heavy metal: dissimilatory metal reduction in anoxic environments. *Trends in ecology & evolution* **8**: 213-217.

Lu WY, Clasquin MF, Melamud E, Amador-Noguez D, Caudy AA & Rabinowitz JD (2010) Metabolomic Analysis via Reversed-Phase Ion-Pairing Liquid Chromatography Coupled to a Stand Alone Orbitrap Mass Spectrometer. *Analytical Chemistry* **82**: 3212-3221.

Manichanh C, Rigottier-Gois L, Bonnaud E*, et al.* (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**: 205-211.

Mankiewicz-Boczek J, Jaskulska A, Paweczyk J, Gagala I, Serwecinska L & Dziadek J (2016) Cyanophages infection of *Microcystis* bloom in lowland dam reservoir of Sulejw, Poland. *Microb Ecol* **71**: 315-325.

Marchler-Bauer A, Derbyshire MK, Gonzales NR*, et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Research* **43**: D222-D226.

Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Martens L, Chambers M, Sturm M*, et al.* (2011) mzML-a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics* **10**: 7.

Matsen FA & Evans SN (2013) Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PloS one* **8**: 15.

Matsen FA, Kodner RB & Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* **11**: 16.

Mayr E (1982) *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

McCarthy DJ, Chen YS & Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**: 4288-4297.

McDaniel L & Paul JH (2005) Effect of nutrient addition and environmental factors on prophage induction in natural populations of marine *Synechococcus* species. *Applied and environmental microbiology* **71**: 842-850.

McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F & Paul JH (2008) Metagenomic Analysis of Lysogeny in Tampa Bay: Implications for Prophage Gene Expression. *PloS one* **3**: 9.

Mekada K, Abe K, Murakami A, Nakamura S, Nakata H, Moriwaki K, Obata Y & Yoshiki A (2009) Genetic Differences among C57BL/6 Substrains. *Experimental Animals* **58**: 141-149.

Mellegard H, Stalheim T, Hormazabal V, Granum PE & Hardy SP (2009) Antibacterial activity of sphagnum acid and other phenolic compounds found in Sphagnum papillosum against food-borne bacteria. *Lett Appl Microbiol* **49**: 85-90.

Merani S, Chen W & Elahi S (2015) The bitter side of sweet: the role of Galectin-9 in immunopathogenesis of viral infections. *Rev Med Virol* **25**: 175-186.

Meyer F, Paarmann D, D'Souza M*, et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**: 386.

Meyer F, Paarmann D, D'Souza M*, et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**: 8.

Miller R & Day M (2008) Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology. *Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses,*(Abedon S, ed.) p.^pp. 114-144. Cambridge University Press, Cambridge.

Miranda JA, Culley AI, Schvarcz CR & Steward GF (2016) RNA viruses as major contributors to Antarctic virioplankton. *Environmental microbiology* **18**: 3714-3727.

Mirza SF, Staniewski MA, Short CM, Long AM, Chaban YV & Short SM (2015) Isolation and characterization of a virus infecting the freshwater algae Chrysochromulina parva. *Virology* **486**: 105-115.

Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ & Wilhelm SW (2017) Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat Commun* **8**: 10.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH & Wilhelm SW (2014) Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. *Virology* **466**: 60-70.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH & Wilhelm SW (2014) Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology* **466-467**: 60-70.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA & Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806-810.

Muyzer G, De Waal EC & Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology* **59**: 695-700.

Myers EW, Sutton GG, Delcher AL*, et al.* (2000) A Whole-Genome Assembly of Drosophila. *Science* **287**: 2196-2204.

Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Nelson M, Burbank DE & Van Etten JL (1998) Chlorella viruses encode multiple DNA methyltransferases. *Biological Chemistry* **379**: 423-428.

Nyren P, Pettersson B & Uhlen M (1993) Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry* **208**: 171-175.

O'Connor M, Peifer M & Bender W (1989) Construction of large DNA segments in Escherichia coli. *Science* **244**: 1307-1312.

Ogilvie LA & Jones BV (2015) The human gut virome: a multifaceted majority. *Frontiers in microbiology* **6**: 12.

Oliver SG, Winson MK, Kell DB & Baganz F (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol* **16**: 373-378.

Opelt K & Berg G (2004) Diversity and antagonistic potential of bacteria associated with bryophytes from nutrient-poor habitats of the Baltic Sea coast. *Applied and environmental microbiology* **70**: 6569-6579.

Opelt K, Chobot V, Hadacek F, Schonmann S, Eberl L & Berg G (2007) Investigations of the structure and function of bacterial communities associated with Sphagnum mosses. *Environmental microbiology* **9**: 2795-2809.

Oppenheim AB, Kobiler O, Stavans J, Court DL & Adhya S (2005) Switches in bacteriophage lambda development. *Annual Review of Genetics,* Vol. 39 p.^pp. 409-429. Annual Reviews, Palo Alto.

Ou T, Li S, Liao X & Zhang Q (2013) Cultivation and characterization of the MaMV-DC cyanophage that infects bloom-forming cyanobacterium *Microcystis aeruginosa*. *Virologica Sinica* **28**: 266-271.

Ou T, Gao XC, Li SH & Zhang QY (2015) Genome analysis and gene *nblA* identification of *Microcystis aeruginosa* myovirus (MaMV-DC) reveal the evidence for horizontal gene transfer events between cyanomyovirus and host. *J Gen Virol* **96**: 3681-3697.

Overbeek R, Begley T, Butler RM*, et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33**: 5691-5702.

Parada AE & Fuhrman JA (2017) Marine archaeal dynamics and interactions with the microbial community over 5 years from surface to seafloor. *The ISME journal*.

Parada AE, Needham DM & Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental microbiology* **18**: 1403-1414.

Paul JH (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *Isme J* **2**: 579-589.

Payet JP & Suttle CA (2013) To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol Oceanogr* **58**: 465-474.

Petrov D, Margreitter C, Grandits M, Oostenbrink C & Zagrovic B (2013) A systematic framework for molecular dynamics simulations of protein post-translational modifications. *PLoS Comput Biol* **9**: e1003154.

Philippe N, Legendre M, Doutre G*, et al.* (2013) Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341**: 281-286.

Popgeorgiev N, Boyer M, Fancello L*, et al.* (2013) Marseillevirus-Like Virus Recovered From Blood Donated by Asymptomatic Humans. *J Infect Dis* **208**: 1042-1050.

Pozio E & La Rosa G (2003) PCR-derived methods for the identification of Trichinella parasites from animal and human samples. *PCR detection of microbial pathogens* 299-309.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA & Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**: 336-341.

Pufahl PK & Hiatt EE (2012) Oxygenation of the Earth's atmosphere–ocean system: A review of physical and chemical sedimentologic responses. *Marine and Petroleum Geology* **32**: 1-20.

Qin BQ, Zhu GW, Gao G, Zhang YL, Li W, Paerl HW & Carmichael WW (2010) A drinking water crisis in Lake Taihu, China: linkage to climatic variability and lake management. *Environ Manag* **45**: 105-112.

R Core Team (2015) R: A language and environment for statistical computing. p.^pp. R Foundation for Statistical Computing, Vienna, Austria.

Radajewski S, Ineson P, Parekh NR & Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* **403**: 646.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M & Claverie JM (2004) The 1.2-megabase genome sequence of mimivirus. *Science* **306**: 1344-1350.

Roach JC, Boysen C, Wang K & Hood L (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**: 345-353.

Robinson MD & Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**: 2881-2887.

Robinson MD & Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**: 321-332.

Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Ronaghi M, Uhlén M & Nyrén P (1998) A Sequencing Method Based on Real-Time Pyrophosphate. *Science* **281**: 363-365.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA & Minor C (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology* **66**: 2541-2547.

Rothschild LJ & Mancinelli RL (2001) Life in extreme environments. *Nature* **409**: 1092-1101.

Roux S, Tournayre J, Mahul A, Debroas D & Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC bioinformatics* **15**: 76.

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T & Debroas D (2012) Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PloS one* **7**: 12.

Roux S, Brum JR, Dutilh BE*, et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689-+.

Rozon RM & Short SM (2013) Complex seasonality observed amongst diverse phytoplankton viruses in the Bay of Quinte, an embayment of Lake Ontario. *Freshw Biol* **588**: 2648-2663.

Rual JF, Venkatesan K, Hao T*, et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**: 1173-1178.

Sachse K & Hotzel H (2003) Detection and differentiation of Chlamydiae by nested PCR. *PCR detection of microbial pathogens* 123-136.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA & Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**: 1350-1354.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB & Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.

Sakaguchi Y, Hayashi T, Kurokawa K, Nakayama K, Oshima K, Fujinaga Y, Ohnishi M, Ohtsubo E, Hattori M & Oguma K (2005) The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc Natl Acad Sci U S A* **102**: 17472-17477.

Sakowski EG, Munsell EV, Hyatt M, Kress W, Williamson SJ, Nasko DJ, Polson SW & Wommack KE (2014) Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proc Natl Acad Sci U S A* **111**: 15786-15791.

Sanders DSA (2005) Mucosal integrity and barrier function in the pathogenesis of early lesions in Crohn's disease. *Journal of Clinical Pathology* **58**: 568-572.

Sanger F & Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441-448.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM & Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687-695.

Santini S, Jeudy S, Bartoli J*, et al.* (2013) Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci U S A* **110**: 10800-10805.

Savage DC (2002) Intestinal Microbiology for the 21st Century. *Bioscience Microflora* **20**: 107-114.

Schloesing J & Muntz A (1877) Sur la Nitrification par les Ferments Organises. *C R Acad Sci Paris* 301-303.

Schloss PD & Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology* **6**: 229.

Schloss PD, Gevers D & Westcott SL (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PloS one* **6**: 14.

Schloss PD, Westcott SL, Ryabin T*, et al.* (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology* **75**: 7537-7541.

Schmidt TSB, Matias Rodrigues JF & von Mering C (2014) Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale. *PLOS Computational Biology* **10**: e1003594.

Schroeder DC, Oke J, Hall M, Malin G & Wilson WH (2003) Virus Succession Observed during an Emiliania huxleyi Bloom. *Applied and environmental microbiology* **69**: 2484-2490.

Schulz F, Yutin N, Ivanova NN*, et al.* (2017) Giant viruses with an expanded complement of translation system components. *Science* **356**: 82-+.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS & Huttenhower C (2011) Metagenomic biomarker discovery and explanation. *Genome biology* **12**: 18.

Sekirov I, Russell SL, Antunes LCM & Finlay BB (2010) Gut Microbiota in Health and Disease. *Physiological Reviews* **90**: 859-904.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B & Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome research* **13**: 2498-2504.

Shen H, Niu Y, Xie P, Tao MIN & Yang XI (2011) Morphological and physiological changes in *Microcystis aeruginosa* as a result of interactions with heterotrophic bacteria. *Freshwater Biology* **56**: 1065-1080.

Shizuya H & Kouros-Mehr H (2001) The development and applications of the bacterial artificial chromosome cloning system. *The Keio journal of medicine* **50**: 26-30.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y & Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89**: 8794-8797.

Short SM (2012) The ecology of viruses that infect eukaryotic algae. *Environmental microbiology* **14**: 2253-2271.

Silver LM (1995) *Laboratory Mice*. Oxford University Press, New York.

Simmonds P, Adams MJ, Benko M*, et al.* (2017) Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15**: 161-168.

Simmonds P, Adams MJ, Benko M*, et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Micro* **15**: 161-168.

Simon MM, Greenaway S, White JK*, et al.* (2013) A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome biology* **14**: 22.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB & Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674-679.

Smith TF & Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* **147**: 195-197.

Sneath PH & Sokal RR (1973) *Numerical taxonomy. The principles and practice of numerical classification*.

Stalheim T, Ballance S, Christensen BE & Granum PE (2009) Sphagnan - a pectin-like polymer isolated from Sphagnum moss can inhibit the growth of some typical food spoilage and food poisoning bacteria by lowering the pH. *J Appl Microbiol* **106**: 967-976.

Steffen MM, Belisle BS, Watson SB, Boyer GL & Wilhelm SW (2014) Status, causes and controls of cyanobacterial blooms in Lake Erie. *Journal of Great Lakes Research* **40**: 215-225.

Steffen MM, Li Z, Effler TC, Hauser LJ, Boyer GL & Wilhelm SW (2012) Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. *PloS one* **7**: e44002.

Steffen MM, Belisle BS, Watson SB, Boyer GL, Bourbonniere RA & Wilhelm SW (2015) Metatranscriptomic Evidence for Co-Occurring Top-Down and Bottom-Up Controls on Toxic Cyanobacterial Communities. *Applied and environmental microbiology* **81**: 3268-3276.

Steffen MM, Belisle BS, Watson SB, Boyer GL, Bourbonniere RA & Wilhelm SW (2015) Metatranscriptomic evidence for co-occurring top-down and bottom-up controls on toxic cyanobacterial communities. *Applied and environmental microbiology* **81**: 3268-3276.

Steffen MM, Dearth SP, Dill BD, Li Z, Larsen KM, Campagna SR & Wilhelm SW (2014) Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in *Microcystis*. *Isme J* **8**: 2080-2092.

Steffen MM, Davis TW, Stough JMA*, et al.* (2017) Lesson from 2014 Lake Erie bloom: business as usual for the western basin? *Environ Sci Technol* **in review**.

Steffen MM, Davis TW, McKay RM*, et al.* (2017) Ecophysiological examination of the Lake Erie Microcystis bloom in 2014: linkages between biology and the water supply shutdown of Toledo, Ohio. *Environmental Science & Technology*.

Stevenson A, Burkhardt J, Cockell CS*, et al.* (2015) Multiplication of microbes below 0.690 water activity: implications for terrestrial and extraterrestrial life. *Environmental microbiology* **17**: 257-277.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M & Poisson G (2013) Are we missing half of the viruses in the ocean? *The ISME journal* **7**: 672-679.

Stewart EJ (2012) Growing unculturable bacteria. *J Bacteriol* **194**: 4151-4160.

Stone L & Roberts A (1990) THE CHECKERBOARD SCORE AND SPECIES DISTRIBUTIONS. *Oecologia* **85**: 74-79.

Stone NE, Fan JB, Willour V, Pennacchio LA, Warrington JA, Hu A, de la Chapelle A, Lehesjoki AE, Cox DR & Myers RM (1996) Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. *Genome research* **6**: 218-225.

Stough JMA & Wilhelm SW (2017) Screening Sequencing Datasets for Marker Genes in CLC Genomics. p.^pp.

Swiatczak B & Cohen IR (2015) Gut feelings of safety: tolerance to the microbiota mediated by innate immune receptors. *Microbiol Immunol* **59**: 573-585.

Swindles GT, Turner TE, Roe HM, Hall VA & Rea HA (2015) Testing the cause of the Sphagnum austinii (Sull. ex Aust.) decline: Multiproxy evidence from a raised bog in Northern Ireland. *Rev Palaeobot Palynology* **213**: 17-26.

Takashima Y, Yoshida T, Yoshida M, Shirai Y, Tomaru Y, Takao Y, Hiroishi S & Nagasaki K (2007) Development and application of quantitative detection of cyanophages phyogenetically related to cyanophage Ma-LMM01 infecting *Microcystis aeruginosa* in fresh water. *Microbes and Environments* **22**: 207-213.

Temin HM & Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**: 1211-1213.

Thijssen VL, Heusschen R, Caers J & Griffioen AW (2015) Galectin expression in cancer diagnosis and prognosis: A systematic review. *Biochim Biophys Acta* **1855**: 235-247.

Thingstad TF & Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13**: 19-27.

Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J & Chisholm SW (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci U S A* **108**: E757-E764.

Tomaru Y & Nagasaki K (2007) Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *J Oceanogr* **63**: 215-221.

Tucker S & Pollard P (2005) Identification of cyanophage Ma-LBP and infection of the cyanobacterium *Microcystis aeruginosa* from an Australian subtropical lake by the virus. *Applied and environmental microbiology* **71**: 629-635.

Tuomela M, Vikman M, Hatakka A & Itävaara M (2000) Biodegradation of lignin in a compost environment: a review. *Bioresource Technology* **72**: 169-183.

Turetsky MR (2003) The role of bryophytes in carbon and nitrogen cycling. *Bryologist* **106**: 395-409.

Turetsky MR, Bond-Lamberty B, Euskirchen E, Talbot J, Frolking S, McGuire AD & Tuittila ES (2012) The resilience and functional role of moss in boreal and arctic ecosystems. *New Phytol* **196**: 49-67.

Tweeddale H, Notley-McRobb L & Ferenci T (1998) Effect of Slow Growth on Metabolism of Escherichia coli, as Revealed by Global Metabolite Pool ("Metabolome") Analysis. *J Bacteriol* **180**: 5109-5116.

Valls M & De Lorenzo V (2002) Exploiting the genetic and biochemical capacities of bacteria for the remediation of heavy metal pollution. *FEMS Microbiology Reviews* **26**: 327-338.

van Breemen N (1995) How Sphagnum bogs down other plants. *Trends in ecology & evolution* **10**: 270-275.

Venter JC & Adams MD & Myers EW*, et al.* (2001) The sequence of the human genome. *Science* **291**: 1304-1351.

Verberkmoes NC, Russell AL, Shah M*, et al.* (2009) Shotgun metaproteomics of the human distal gut microbiota. *Isme J* **3**: 179-189.

Verhoeven JTA & Liefveld WM (1997) The ecological significance of organochemical compounds in Sphagnum. *Acta Bot Neerl* **46**: 117-130.

Villarino NF, LeCleir GR, Denny JE, Dearth SP, Harding CL, Sloan SS, Gribble JL, Campagna SR, Wilhelm SW & Schmidt NW (2016) Composition of the gut microbiota modulates the severity of malaria. *Proc Natl Acad Sci U S A* **113**: 2235-2240.

Wahba AJ, Gardner RS, Basilio C, Miller RS, Speyer JF & Lengyel P (1963) Synthetic polynucleotides and the amino acid code. VIII. *Proc Natl Acad Sci U S A* **49**: 116-122.

Waller AS, Hug LA, Mo K, Radford DR, Maxwell KL & Edwards EA (2012) Transcriptional Analysis of a Dehalococcoides-Containing Microbial Consortium Reveals Prophage Activation. *Applied and environmental microbiology* **78**: 1178-1186.

Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV & Bork P (2014) Classification and quantification of bacteriophage taxa in human gut metagenomes. *Isme J* **8**: 1391-1402.

Wang D-Z, Kong L-F, Li Y-Y & Xie Z-X (2016) Environmental Microbial Community Proteomics: Status, Challenges and Perspectives. *International Journal of Molecular Sciences* **17**: 1275.

Wang Z, Gerstein M & Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* **10**: 57-63.

Warren MJ, Lin XJ, Gaby JC*, et al.* (2017) Molybdenum-Based Diazotrophy in a Sphagnum Peatland in Northern Minnesota. *Applied and environmental microbiology* **83**: 14.

Washburn MP, Wolters D & Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **19**: 242-247.

Waterston RH & Lindblad-Toh K & Birney E*, et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Watkins SC, Smith JR, Hayes PK & Watts JEM (2014) Characterisation of host growth after infection with a broad-range freshwater cyanopodophage. *PloS one* **9**: 8.

Watson JD & Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737-738.

Weitz JS & Wilhelm SW (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol Rep* **4**: 17.

Whitman WB, Coleman DC & Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci U S A* **95**: 6578-6583.

Wilhelm S, Bird J, Bonifer K*, et al.* (2017) A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. *Viruses* **9**: 46.

Wilhelm SW & Suttle CA (1999) Viruses and nutrient cycles in the sea. *Bioscience* **49**: 781-788.

Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M & Stough JMA (2016) Standing on the shoulders of giant viruses: five lessons learned about large viruses infecting small eukaryotes and the opportunities they create. *PLoS Pathog* **12**: 5.

Wilhelm SW, Bird JT, Bonifer KS*, et al.* (2017) A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. *Viruses-Basel* **9**: 18.

Williams RJ, Howe A & Hofmockel KS (2014) Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in microbiology* **5**: 358.

Wilson HR, Yu DG, Peters HK, Zhou JG & Court DL (2002) The global regulator RNase III modulates translation repression by the transcription elongation factor N. *Embo J* **21**: 4154-4161.

Wilson RM, Hopple AM, Tfaily MM*, et al.* (2016) Stability of peatland carbon to rising temperatures. *Nat Commun* **7**: 10.

Winogradsky S (1887) Concerning sulfur bacteria. *Bot Ztg*.

Winogradsky S (1890) On the nitrifying organisms. *Sciences* **110**: 1013-1016.

Winogradsky S (1928) The direct method in soil microbiology and its application to the study of nitrogen fixation. *Soil Sci* **25**: 37-43.

Winogradsky S & Brock T (1889) Physiological studies on the sulfur bacteria. *Milestones in Microbiology*.

Woese CR & Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences* **74**: 5088-5090.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S & Nasko DJ (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6**: 427-439.

Xie Y, Wu G, Tang J*, et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**: 1660-1666.

Xiong LJ, Edwards CK & Zhou LJ (2014) The Biological Function and Clinical Utilization of CD147 in Human Diseases: A Review of the Current Scientific Literature. *International Journal of Molecular Sciences* **15**: 17411-17441.

Yamaguchi H, Suzuki S, Tanabe Y, Osana Y, Shimura Y, Ishida K-i & Kawachi M (2015) Complete genome sequence of *Microcystis aeruginosa* NIES-2549, a bloom-forming cyanobacterium from Lake Kasumigaura, Japan. *Genome Announcements* **3**.

Yolken RH, Jones-Brando L, Dunigan DD*, et al.* (2014) Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci U S A* **111**: 16106-16111.

Yoshida-Takashima Y, Yoshida M, Ogata H, Nagasaki K, Hiroishi S & Yoshida T (2012) Cyanophage infection in the bloom-forming cyanobacteria *Microcystis aeruginosa* in surface freshwater. *Microbes Environ* **27**: 350-355.

Yoshida M, Yoshida T, Kashima A, Takashima Y, Hosoda N, Nagasaki K & Hiroishi S (2008) Ecological dynamics of the toxic bloom-forming cyanobacterium *Microcystis aeruginosa* and its cyanophages in freshwater. *Applied and environmental microbiology* **74**: 3269-3273.

Yoshida T, Takashima Y, Tomaru Y, Shirai Y, Takao Y, Hiroishi S & Nagasaki K (2006) Isolation and characterization of a cyanophage infecting the toxic cyanobacterium *Microcystis aeruginosa*. *Applied and environmental microbiology* **72**: 1239-1247.

Yoshida T, Nagasaki K, Takashima Y, Shirai Y, Tomaru Y, Takao Y, Sakamoto S, Hiroishi S & Ogata H (2008) Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *J Bacteriol* **190**: 1762-1772.

Yutin N, Wolf YI & Koonin EV (2014) Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466**: 38-52.

Yutin N, Wolf YI, Raoult D & Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology Journal* **6**: 13.

Zhou XB, Lindsay H & Robinson MD (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research* **42**: 10.

VITA

Joshua Stough was born in Merced, CA on November 17th, 1988. He attended Nashua High School South in Nashua, NH and graduated in 2007. He received his Bachelor of Science degree in Biological Sciences from Auburn University in 2011, his Master of Science in Biological Sciences at Auburn University in 2013, and began the PhD program at the University of Tennessee in August 2013.