AN ABSTRACT OF THE DISSERTATION OF

Connor B. Driscoll for the degree of Doctor of Philosophy in Microbiology presented on July 15, 2016.

Title: Comparative Genomics of Freshwater Bloom-Forming Cyanobacteria and Associated Organisms

Abstract approved: _____

Theo W. Dreher

The advent of improved DNA sequencing technologies has allowed the analysis of various microbial communities. Bloom-forming freshwater cyanobacteria can produce toxins and taste-and-odor compounds that can negatively affect drinking water supplies. Here, I have employed second- and third-generation sequencing technologies to characterize bloom-forming freshwater cyanobacterial genomes and their associated heterotrophic bacteria and viruses. These include nine novel freshwater *Nostocaceae* genomes, three genomes from heterotrophic bacteria associated with *Aphanizomenon* in a communal culture, and two novel *Microcystis* phage genomes.

- The genomes of three novel heterotrophic bacteria associated with *Aphanizomenon flos-aquae* in culture were sequenced and assembled to finished

quality with long-read sequencing. These genomes were sequenced together, highlighting the potential for using long-read sequencing towards metagenomics of low-diversity microbial communities. These genomes were analyzed to assess interactions between *Aphanizomenon flos-aquae* and these heterotrophs in culture. The presence of an ammonium-importer gene in two of these genomes suggests a putative dependency on fixed nitrogen from *Aphanizomenon flos-aquae*.

- The genomes of nine novel *Nostocaceae* genomes were sequenced and assembled to draft quality. Five of these genomes were assembled and extracted directly from three separate environmental short-read shotgun metagenomes. The remaining four strains were cultured, one of which was from this study (*Aphanizomenon* MDT14) and three that were provided by Gregory Dick's lab at the University of Michigan (*Anabaena* CPCC64, *Anabaena* LE011-02, and *Anabaena* AL09). All novel genomes were characterized relative to the rest of the *Nostocaceae* family to analyze evolutionary relationships and identify differences in gene content to evaluate potential phenotypic patterns/differences. Genes involved in toxin synthesis and sulfur metabolism are variably present in these genomes, with no patterns relative to phylogenomic relationships. Conversely, functionally diverse genes are present in genomes with close phylogenomic relationships.

- The genomes of two novel *Microcystis* phages were sequenced and assembled to finished quality from two separate environmental short-read shotgun

metagenomes. These novel genomes were similar to the previously sequenced *Microcystis* phages Ma-LMM01 and MaMV-DC, and all four genomes were characterized together to identify patterns of gene conservation in this geographically distributed phage group. Additionally, one of the completed phages was present in samples across a 6-week time series of environmental short-read shotgun metagenomes. Patterns of gene gain/loss and divergence were then analyzed in this *Microcystis* phage across the time series. Host-like genes involved in photosynthesis and phosphate starvation are present in all genomes, while presence of other host-like genes is less conserved. Genomes from the environmental time-series contain differences in presence/absence of several hypothetical genes, as well as sequence divergence in the tail collar gene, which may have implications for infection in the environment.

Comparative Genomics of Freshwater Bloom-Forming Cyanobacteria and Associated Organisms

by

Connor B. Driscoll

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented July 15, 2016
Commencement June 2017

Doctor of Philosophy dissertation of Connor B. Driscoll presented on July 15, 2016.

APPROVED:

Major Professor, representing Microbiology

Chair of the Department of Microbiology

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Connor B. Driscoll, Author

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Theo Dreher, for all of his assistance and advice. I would like to thank my committee members (Dr. Ryan Mueller, Dr. Rebecca Vega-Thurber, Dr. Valerian Dolja, and Dr. Jaga Giebultowicz) for their time and feedback. I would also like to thank my parents, Jolene and Michael Driscoll, for their immeasurable support. I would like to thank my grandfather, John Driscoll, for the years of sage wisdom and support. I would like to thank my uncle, Tom Driscoll, for the friendship and positivity through the difficult times. I thank my grandmother, Virginia Roberts, for always being there for me. I would also like to thank my uncle, Brad Kemper, for helping me to stay down-to-Earth. I also thank my labmates, Tim Otten and Nathan Brown, for the many productive and enjoyable conversations. I would like to thank my friends for keeping me well-grounded. I would like to thank my cat Polly for her much-needed companionship, and for changing my perspective during the short time we had together. I would also like to thank my new cats Shae and Elphaba, who have both already helped me through a difficult time. And finally, I would like to thank my stellar girlfriend, Riana Wernick, for all of her love and support.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

Chapter 1 Introduction

1.0.1 Freshwater bloom-forming cyanobacteria

When conditions allow, aquatic cyanobacteria can grow to high densities, and these populations are termed "blooms." Cyanobacterial blooms occur across the world in aquatic systems including marine, brackish, and freshwater. Most research to date has focused on the numerically abundant marine *Prochlorococcus* and *Synechococcus* genera, whose sheer numbers drive a major portion of global carbon fixation and oxygenation on Earth [Partensky et al. 1999]. Freshwater bloom-forming cyanobacteria, in contrast, have been studied much less. There is good reason to focus on freshwater cyanobacteria, especially since eutrophication has caused increased frequency of cyanobacterial blooms in freshwater systems over recent years [Oneil et al. 2012]. Bloom-forming cyanobacteria are also often capable of producing toxic compounds known to cause liver toxicity or paralysis [Nishiwaki-Matsushima et al. 1992; Carmichael et al. 1975; Cheung et al. 2013; Otten and Paerl 2015], threatening drinking water supplies as a result [Falconer 1999]. Additionally, some strains can produce taste and odor compounds that make drinking water supplies unpalatable [Jüttner and Watson 2007]. Cyanobacterial blooms also have major impacts on surrounding communities. For example, high-density blooms are responsible for water deoxygenation leading to "dead

zones" following bloom degradation, as well as having major food web impacts and preventing other organisms from photosynthesizing by shading out sunlight [Paerl et al. 2001; Huisman et al. 2004]. Freshwater cyanobacteria can also produce a wide range of natural compounds that carry potentially valuable activities [Harada 2004; Dittmann et al. 2015]. Taking these facts into consideration, it would be useful to apply current DNA sequencing technologies and analytical techniques to better understand environmental parameters that may be important for growth of freshwater cyanobacteria.

1.0.2 Bloom-associated bacteria

Heterotrophic bacteria associated with algal blooms have been shown to play important roles in nitrogen, sulfur, carbon and phosphorus cycling [Ask et al. 2009; González et al. 2000; Grossart et al. 2006; Grover 2000]. Some mutualistic interactions between algae and bacteria have been identified. For example, the vitamin B12 micronutrient is supplied to some eukaryotic algae by bacterial partners in exchange for fixed carbon [Croft et al. 2005; Amin et al. 2012]. Another example is the *Richelia intracellularis* cyanobacterium, which can fix nitrogen in a symbiotic relationship with diatoms [Foster et al. 2011]. Interactions between cyanobacteria and associated heterotrophic bacteria also occur. Cyanobacterial growth can be enhanced by the presence of heterotrophic bacteria in culture [Berg et al. 2009]. This may be a result of nutrient sharing. Alternatively, this may be caused by dependencies generated by reductive evolution via the Black Queen

Hypothesis, whereby gene loss in free-living organisms leaves them dependent on co-occurring microbes for lost metabolic or other functions [Morris et al. 2012]. These interactions could also be unidirectional. For example, others have shown that *Aphanizomenon flos-aquae* in the Baltic Sea fixes nitrogen, which is released from the cell as ammonium, and is taken up by heterotrophic and phototrophic bacteria with no identified benefit for *A. flos-aquae* [Ploug et al. 2010; Adam et al. 2016]. These interactions are important to understand as factors that potentially enhance or hinder growth of cyanobacterial strains in the environment, likely affect the overall microbial community composition, and may have an essential role in how cyanobacterial blooms initiate, persist, or collapse.

1.0.3 Cyanobacterial genomics

While cyanobacterial genomics have been well-studied over the past decade, a large majority of studied genomes belong to the marine genera *Prochlorococcus* and *Synechococcus* due to their importance in global biogeochemical cycles. In comparison, freshwater cyanobacterial genomes have been sequenced much less (166 *Prochlorococcus* genomes compared with 19 *Microcystis* genomes in NCBI's Genome database as of July 2016), leaving a number of clades with fewer sequenced representatives (22 genomes from entire *Nostocaceae* family in NCBI's Genome database as of July 2016). To better understand freshwater cyanobacterial evolution, physiology, and population dynamics, more genome sequences are needed from less-sequenced clades such as *Nostocaceae*. As more of these genomes are se-

quenced, we may begin to elucidate the distribution of toxin-synthesis genes, in addition to patterns of genes involved in niche differentiation. Freshwater cyanobacteria can also produce a number of secondary metabolites [Dittmann et al. 2015], and sequencing more genomes could reveal novel pathways for synthesizing undiscovered natural compounds.

A considerable amount of work has gone into studying the comparative genomics of cyanobacteria. This has included focus on primary metabolism [Beck et al. 2012], secondary metabolism and natural product synthesis [Baran et al. 2013; Calteau et al. 2014; Dittmann et al. 2015], the core and pan-genome [Shi and Falkowski 2008; Simm et al. 2014], phylogenomic characterization [Shih et al. 2013], and extracellular polysaccharide synthesis genes [Pereira et al. 2015]. However, most of these studies are phylum-wide analyses, which don't focus on patterns found in narrower groups such as genera. As a result, comparative studies of freshwater bloom-forming cyanobacteria would be useful in identifying genomic patterns that inform about their physiology and evolution.

In addition to characterizing functional gene content, comparative genomics helps reveal phyletic relationships [Daubin et al. 2002; Delsuc et al. 2005; Ciccarelli et al. 2006; Puigbo et al. 2010; Hug et al. 2016]. Single-gene phylogenies have been previously used to characterize relationships within cyanobacteria, which produces results inconsistent with current genus assignments of strains within the *Nostocaceae* family [Gugger et al. 2002; Rajaniemi et al. 2005]. These assignments are based on polyphasic classification, which takes into account morphological as well as genetic similarity [Komárek 2016]. However, morphological classifications

are subjective, and the genetic component relies solely on 16S rDNA phylogenies, which have not been capable of resolving the placement of *Anabaena*, *Aphanizomenon*, *Dolichospermum*, and *Nostoc* strains in phylogenetic trees. Further genome sequencing may allow for re-classification of these genomes, and proper classification of future isolates.

1.0.4 Viruses of freshwater cyanobacteria

As top-down predators, viruses infecting algae can control population density in the environment. For example, *Emiliana huxleyi* algal blooms in marine systems collapse in response to viral infection [Bratbak et al. 1993; Jacquet et al. 2002; Sorensen et al. 2009]. However, there are few instances of phage-induced freshwater cyanobacterial bloom collapse [Peduzzi et al. 2014]. Beyond controlling host population numbers, phages of marine cyanobacteria have been shown to drive host diversity and evolution as well [Rodriguez-Valera et al. 2009; Biller et al. 2015]. Since marine cyanophages play an important role in host ecology and evolution, studying freshwater cyanophages could inform about population dynamics and evolution of freshwater bloom-forming cyanobacteria.

However, as of June 2016, the genomes of only nine cyanophages that infect freshwater cyanobacteria have been sequenced. Eight of these viruses are tailed phages, while one (*Planktothrix* phage PaV-LD) is not tailed and is highly divergent from the others based on capsid gene phylogeny, and has thus remained unclassified [Gao et al. 2012]. The podoviruses, which consist of an *Anabaena* phage

(A-4L [Ou et al. 2015b]), a polar *Synechococcus*-infecting phage (SEIV-1 [Chénard et al. 2015]), and three *Phormidium* phages (PP-1 (unpublished), Pf-WMP3 [Liu et al. 2008], and Pf-WMP4 [Liu et al. 2007]) cluster together and separately from marine cyanopodoviruses based on a concatenated phylogeny of conserved genes [Ou et al. 2015b]. The genomes of only three freshwater cyanomyoviruses have been sequenced to date. One is a *Cyanobium*-infecting phage (S-CRM01) that is more closely related to marine cyanomyoviruses [Dreher et al. 2011] than the two other freshwater cyanomyoviruses which infect the potentially toxigenic *Microcystis aeruginosa* (Ma-LMM01 and MaMV-DC) [Yoshida et al. 2006; 2008; Ou et al. 2013; 2015a]. Both *Microcystis* phages have been characterized in culture and have fully sequenced genomes [Yoshida et al. 2006; 2008; Ou et al. 2013; 2015a]. About one-sixth of the MaMV-DC genome contains genes similar to host genes [Ou et al. 2015a]. These studies have revealed these lytic phages carry a host-like gene involved in regulating photosynthesis, *nblA*, which promotes phycobilisome degradation during infection [Ou et al. 2015a; Gao et al. 2012]. NblA may provide protection for the host photosystem II complex by preventing absorption of excess light energy (and therefore photoinhibition) through phycobilisome degradation [Yoshida-Takashima et al. 2012; Honda et al. 2014]. Alternatively, phycobilisome degradation may provide additional amino acids for phage structural synthesis [Yoshida-Takashima et al. 2012; Ou et al. 2015a], since phycobilisomes can constitute a large proportion of soluble cellular protein [Grossman et al. 1993]. Together with the *psb* genes found in other phages, this indicates that freshwater and marine cyanophages can employ different host-like genes in order to utilize resources

related to photosynthesis in their respective hosts. However, as fewer freshwater cyanophages have been sequenced than marine cyanophages, more genome sequences are needed to better understand the diversity and infection strategies available to freshwater cyanophages.

1.0.5 DNA sequencing and analysis

The amount of DNA sequencing has increased exponentially over the last decade [Buermans and Den Dunnen 2014]. Until recently, short-read DNA sequencing technologies dominated genomic and metagenomic studies [Morozova and Marra 2008; Bragg and Tyson 2014]. However, parsing these complex datasets has required the development of novel informatic tools for processes such as assembling genomes [Peng et al. 2012], calling and annotating genes [Delcher et al. 1999; Seemann 2014], taxonomic assignment [Gregor et al. 2014; Darling et al. 2014], and binning and evaluating whole genomes [Albertsen et al. 2013; Parks et al. 2015; Kang et al. 2015]. Long-read sequencing technologies are taking an increasing share of the market, and provide specific advantages not available to short-read sequencers [Koren and Phillippy 2015]. Primarily, long reads span repetitive genomic regions that short-reads cannot assemble to prevent assembly breaks [Lee et al. 2014]. As a result, finishing genome assemblies removes the need to bin draft genomes or manually close gaps with Sanger sequencing. Additionally, long-read sequences allow correct assembly of regions containing adjacent, short repetitive genes [Brown et al. 2016]. However, there are drawbacks to using long-read se-

quencers. They are much lower-throughput than short-read sequencers, which means that cost per base pair is higher. As a result, bacterial long-read sequencing has been primarily limited to cultured strains [Bashir et al. 2012; Zhang et al. 2016], with only a few exploratory steps into metagenomics [Frank et al. 2015; Mosher et al. 2014]. Also, high concentrations of DNA are necessary for long-read sequencing, which is sometimes difficult to obtain. These drawbacks are important to keep in mind for long-read sequencing projects.

1.0.6 Overview of chapters

The dissertation research presented here has five chapters, all focusing on comparative genomics of freshwater cyanobacteria, and the bacteria/viruses associated with them. In Chapter 2, we investigated the interactions of bacteria associated with *Aphanizomenon flos-aquae* from Upper Klamath Lake, OR by DNA sequencing a mixed-community culture. We employed long-read shotgun metagenome sequencing to completely assemble three bacterial genomes. Our results show that two genomes belong to the *Proteobacteria* phylum, and likely survive by importing fixed nitrogen released by *A. flos-aquae* grown in a nitrogen-free culture medium. We also discuss the possibility of sequencing microbial communities using long-read technology in order to fully assemble bacterial genomes.

In Chapter 3, we sequenced nine novel freshwater cyanobacterial genomes belonging to the *Nostocaceae* family. This family is of particular interest since some members have been shown to produce a variety of toxins (e.g. anatoxins, micro-

cystin, saxitoxin), as well as allelopathic compounds that can affect local organisms. As part of a collaborative study between four labs, we compared these nine novel genomes with previously-sequenced *Nostocaceae* family genomes to identify genes indicative of niche differentiation. Additionally, we sought to characterize the relationship of these novel genomes to the rest of the *Nostocaceae* family, where taxonomic inconsistencies abound regarding the placement of *Anabaena*, *Aphanizomenon*, and *Dolichospermum* strains. Five of the novel genomes were binned directly from environmental shotgun metagenomes (*Anabaena* CRKS33, *Anabaena* MDT14-2, *Anabaena* WA113, *Aphanizomenon* MDT14-1, and *Aphanizomenon* WA102), while the other four novel genomes were sequenced from cultures (*Anabaena* AL09, *Anabaena* CPCC64, *Anabaena* LE011-02, and *Aphanizomenon* MDT13), indicating the utility of available sequence analysis tools to sequence and extract draft-quality bacterial genomes without the need for culturing.

Chapter 4 focuses on phages that infect bloom-forming cyanobacteria. We assembled two novel strains of the *Microcystis* phage Ma-LMM01. Completing these genomes brings the total number of sequenced strains of this globally distributed phage type to four. We characterized these four genomes together to investigate patterns of conservation and variance across these genomes, in addition to searching for evidence of this virus in other freshwater metagenomes. This phage was present in a two-month metagenomic time-series from samples collected once every two weeks. We then compared fragmented genome assemblies of this phage over this time series to assess detectable genome variants within the same environment.

Together, this research provides novel insights into the lifestyles of freshwa-

ter bloom-forming cyanobacteria of the *Nostocaceae* family, as well as increasing our understanding of genome evolution of a *Microcystis* phage across and within environments.

Name	**Famly**	**Host**	**Genome size (bp)**	**No. of ORFS**	**% G+C**	**Genbank No.**
Ma-LMM01	*Myoviridae*	*Microcystis*	162,109	184	45.0	AB231700.1
S-CRM01	*Myoviridae*	*Synechococcus*	178,563	294	39.7	HQ615693.1
A-4L	*Podoviridae*	*Anabaena*	41,750	38	43.4	KF356198.1
Pf-WMP3	*Podoviridae*	*Phormidium*	43,249	41	46.5	EF537008.1
Pf-WMP4	*Podoviridae*	*Phormidium*	40,938	41	51.8	DQ875742.1
PP-1	*Podoviridae*	*Plectonema*	42,480	41	46.4	Unpublished
PaV-LD	unassigned	*Planktothrix*	95,299	142	41.5	HQ683709.1
SEIV-1	unassigned	*Synechococcus*	79,178	130	46.2	KJ410740

Table 1.1: Sequenced freshwater cyanophage genomes as of June 2016.

Chapter 2 Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture

Connor B. Driscoll[1], Timothy G. Otten[1], Nathan M. Brown[1], Theo W. Dreher[1,2]

[1]Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR, 97331, USA.

[2]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.

Formatted for *Standards in Genomic Sciences*

Biomed Central

Floor 6, 236 Gray's Inn Road

London

WC1X 8HB

United Kingdom

In review

2.1 Introduction

Metagenomic sequencing is the process of sampling DNA sequences from multiple genomes in a community of organisms, and has been applied to many environmental samples to assess both functional diversity and species richness of microbial communities [Gilbert and Dupont 2011; Escobar-Zepeda et al. 2015]. Recently, there has been a progression in metagenomic approaches associated with advances in sequencing technologies. Next-generation sequencing (NGS) methods [Mardis 2008] such as 454 and Illumina HiSeq/MiSeq greatly reduced sequencing costs per base relative to Sanger sequencing due to increased throughput, which facilitated high-throughput shotgun metagenomics (randomly sequencing all DNA in a sample). This provided several advantages over amplicon sequencing, where all variants of a single gene in a population are sequenced. For example, focus shifted from assigning taxa using single genes to using multiple genes and/or sequence composition instead [Escobar-Zepeda et al. 2015; Gregor et al. 2014]. It also permitted functional characterization of individual representatives or whole microbial communities [Sharon et al. 2013; Evans et al. 2015]. However, there are technical hurdles associated with short-read sequencing. Specifically, assembling short reads (50-300 bp) into contiguous sequences (contigs) rarely leads to complete genome assemblies due to repetitive genomic elements such as 16S rRNA genes [Rainey et al. 1996] and insertion sequence (IS) elements [Lawrence et al. 1992] that are 1 kb or greater in length. There are two consequences as a result. First, closing draft genomes by primer walking requires considerable manual effort and time. Second,

if closure is not possible, contigs must be clustered and binned using methods like differential coverage [Albertsen et al. 2013], co-abundance [Sharon et al. 2013; Imelfort et al. 2014; Kang et al. 2015], or gene/nucleotide composition [Cleary et al. 2015]. While useful, these methods are often not comprehensive and become even more difficult to implement when used in a metagenomic context, where multiple genomes (sometimes from closely related organisms) must be delineated [Hess et al. 2011]. Single-molecule real time sequencing (SMRT) technologies, such as PacBio and Oxford Nanopore, are part of the third-generation sequencing wave [Koren and Phillippy 2015]. These sequencers produce average read lengths in the 5-50 kb range, with 50% of reads longer than 14 kb [Lee et al. 2014], which exceed the size of repetitive elements in the average bacterial genome. Although more error-prone, these longer reads have proven advantageous for assembling closed genomes if sequencing depth is high enough to allow error correction [Koren et al. 2012]. To date, long-read sequencing has rarely been used for metagenomics for several reasons: 1) the amount of sequence data returned is a fraction of an Illumina run (up to 750 Gb/flow cell of Illumina HiSeq 3000 vs. up to 1 Gb/SMRT cell of PacBio Sequel based on company specifications), 2) the sequencing cost per base pair is higher, and 3) PacBio sequencing does not rely upon DNA amplification, so high concentrations of raw DNA are required. Due to these limitations, long-read metagenomics has so far been limited to whole-16S amplicon sequencing [Fichot and Norman 2013] and to improving binning from fragmented (short-read) assemblies [Frank et al. 2015]. Here, we have generated a PacBio shotgun metagenome from a non-axenic cyanobacterium culture established in summer 2013 originat-

ing from Upper Klamath Lake (UKL), OR. In this freshwater lake, the N_2-fixing filamentous cyanobacterium *Aphanizomenon flos-aquae* blooms annually. These blooms are harvested and sold as nutritional supplements. Little is known about the co-occurring microbial community in this lake, whose composition could be influenced by the presence of *A. flos-aquae* as the dominant primary producer [Bagatini et al. 2014; Louati et al. 2015]. By applying a selective growth medium lacking nitrogen, our goal was to sequence and assemble complete genomes from a relatively simple community, in turn assessing the possibility for using PacBio shotgun sequencing for environmental metagenomics. We closed three novel bacterial genomes, which provide insight into putative metabolic dependencies of these bacteria on *A. flos-aquae* in the co-culture. However, we were unable to close the *A. flos-aquae* genome, which is in draft quality and will be discussed elsewhere.

2.2 Organism information

2.2.1 Classification and features

The taxonomic placement of each genome was assessed three ways (Table 2.1). We used the SILVA SSU Ref NR database (accessed on March 9, 2016) to search for significant 16S rDNA matches in the Silva database [Quast et al. 2012]. Also, we generated 16S phylogenetic trees for each genome, using the SINA aligner [Pruesse et al. 2012] and FastTree [Price et al. 2010], with all classified *Alphaproteobacteria*, *Betaproteobacteria*, and *Bacteroidetes* representatives in SILVA, shown with

their nearest groups (Figures 2.1, 2.2, 2.3). For the second taxonomic placement method, we used PhyloPythiaS+ [Gregor et al. 2014], which searches for genomes with similar k-mer composition. The third method, Phylosift [Darling et al. 2014], is a pipeline that aligns 40 marker genes to generate a weighted probability score for specific taxonomic assignments. Based on consistency across these classification methods as well as confidence values from 16S trees, we named each genome *Hyphomonadaceae* UKL13-1, *Betaproteobacterium* UKL13-2, and *Bacteroidetes* UKL13-3, respectively. Minimum Information about the Genome Sequences is summarized in Table 2.2. Although we initiated and maintained this mixed-community culture for one year, the culture died and we did not obtain physiological information regarding these organisms. Sustaining long-term *A. flos-aquae* cultures is often difficult, and it is common for cultures to crash. Instead, we discuss insights from the genome annotations regarding these features below.

2.3 Genome sequencing information

2.3.1 Genome project history

Cultures were initiated from UKL, where annual *A. flos-aquae* blooms constitute a serious ecological disturbance but are also harvested and sold as nutritional supplements. The genome sequences were deposited to DDBJ/EMBL/GenBank under the accessions CP012156, CP012157, and CP012155 for the *Hyphomonadaceae* UKL13-1, *Betaproteobacterium* UKL13-2, and *Bacteroidetes* UKL13-3 genomes,

respectively. Project information is summarized in Table 2.3.

2.3.2 Growth conditions and genomic DNA preparation

One *Aphanizomenon flos-aquae* colony from a depth-integrated water sample from the UKL MDT site collected during August 2013 was transferred to Bold $3N_0$ medium (https://utex.org/products/bold-3n-medium) without $NaNO_3$. This medium consisted of 0.17 mM $CaCl_2$, 0.3 mM $MgSO_4$, 0.43 mM K_2HPO_4, 1.29 mM KH_2PO_4, 0.43 mM NaCl, P-IV trace metals, and 0.4 μmol vitamin B12 at pH 8.0. The culture was maintained under cool white fluorescent light (20 μE m^{-2} s^{-1}) with a light/dark cycle of 16 h/8 h at 24°C. Three separate DNA extractions were performed from this culture (Table 2.4). A sample taken in November 2013 was collected on a 1.2 μm GF/C filter (Whatman), and DNA extracted for Illumina sequencing using a DNA extraction kit (GeneRite DNA-EZ RWOC1). A similarly collected sample (Nov 2013) was extracted using phenol-chloroform [Sambrook and Russell 2006] and pooled with phenol-chloroform extracted DNA from an unfiltered sample of the culture collected during March 2015 (to balance proportion of sequencing associated with cyanobacteria and heterotrophic bacteria). This pooled sample was quantified with the Q32850 Quant-iT dsDNA BR Assay Kit. Approximately eight μg of DNA was submitted for PacBio sequencing.

2.3.3 Genome sequencing and assembly

The November 2013 sample was processed using a Nextera XT kit and sequenced using the Illumina HiSeq 2000 at the Oregon State University Center for Genome Research and Biocomputing (CGRB) to generate 17,617,259 paired-end reads (101 bp). The pooled (11/2013 & 3/2015) sample was processed for PacBio sequencing by the Molecular Biology and Genomics Core at Washington State University. Eight SMRT cells of PacBio RS sequencing generated 348,623 reads with an average length of 7,737 bp. PacBio sequences were assembled using HGAP [Chin et al. 2013] with three different parameter sets to optimize for assembly of different genomes (Table 2.5). Initially, only the *Bacteroidetes* genome assembled from 2 SMRT cells (167,289 PacBio reads), at a seed read length cutoff of 12.8 kb. The less abundant *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2 genomes required all 8 SMRT cells to close (348,623 reads). While the *Betaproteobacterium* genome closed with a seed read-length cutoff of 13.6 kb, the *Hyphomonadaceae* genome only assembled completely when this cutoff was lowered to 6 kb, likely since it had the lowest coverage of the three genomes. A lower cutoff directs more reads towards use in assembling, thereby improving chances of completing low-coverage assemblies [Forde et al. 2014]. However, this also reduces the number of reads used in error correction, which in turn increases the chances of assembly errors. These tradeoffs should be considered before performing assemblies, but it is notable that we would not have completed the *Hyphomonadaceae* UKL13-1 genome without lowering this cutoff. The *Hyphomonadaceae*, *Betapro-*

teobacterium, and *Bacteroidetes* genomes were of finished quality (Tables 2.4, 2.6), with each having average Phred scores (ASCII base 33) of 75.9, 76.0, and 81.9, respectively. We were unable to complete other genomes in the culture, including the draft-quality *A. flos-aquae* genome assembly. The Illumina-sequenced culture was assembled using the IDBA-Hybrid [Peng et al. 2012] software. We binned Illumina-assembled contigs from the three completed genomes by differential coverage of reads from both PacBio and Illumina samples. That is, Illumina and PacBio reads were separately mapped to each assembly using BWA-MEM [Li 2013] and BLASR [Chaisson and Tesler 2012], respectively. Contigs were then binned using the mmgenome R package [Albertsen et al. 2013] (Table 2.7).

2.3.4 Genome annotation

All genomes were annotated with the NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) [Angiuoli et al. 2008] and PROKKA [Seemann 2014] (included as additional files). Counts of features (Genes, CDS, pseudogenes,rRNAs, tRNAs, ncRNAs, and CRISPR arrays) come from PGAP annotations. Amino acid sequences were assigned to COG categories by searching against the COG protein database [Galperin et al. 2014] using RAPSEARCH [Zhao et al. 2012], taking only the top hits above an E-value of 1E-30. Amino acid sequences from each genome were also annotated using the KEGG database [Kanehisa et al. 2016] with the GhostKOALA [Kanehisa et al. 2015] pipeline and the "genus_prokaryotes" database on September 3, 2015.

2.4 Genome properties

Each genome assembled into one closed contig. The *Hyphomonadaceae* UKL13-1 genome consists of a single circular chromosome 3,501,508 bp long and a GC content of 56.12%. The genome contains a total of 3255 predicted genes, including 2934 predicted protein-coding sequences, 277 pseudogenes, and 44 RNA genes (40 tRNAs, one 16S-23S-5S rRNA operon, and 1 ncRNA) (Fig 2.4). The *Betaproteobacterium* UKL13-2 genome consists of a single circular chromosome 3,387,087 bp long and a GC content of 54.98%. The genome contains a total of 3087 predicted genes, including 2772 predicted protein-coding sequences, 265 pseudogenes, and 50 RNA genes (43 tRNAs, two 16S-23S-5S rRNA operons, and 1 ncRNA) (Fig. 2.5). The *Bacteroidetes* UKL13-3 genome consists of a single circular chromosome 3,236,529 bp long and a GC content of 37.33%. The genome contains a total of 2850 predicted genes, including 2598 protein-coding sequences, 211 pseudogenes, and 41 RNA genes (35 tRNAs and two 16S-23S-5S rRNA operons)(Fig. 2.6). Properties and statistics of each genome are shown in Table 2.8. The distribution of genes into COG functional categories is summarized in Table 2.9.

2.5 Insights from the genome sequence

2.5.1 PacBio metagenome and comparison to Illumina metagenome

The bacterial community associated with the *Aphanizomenon flos-aquae* culture was subjected to metagenomic analysis with 8 SMRT cells of PacBio reads, re-

sulting in three completed novel bacterial genomes: *Hyphomonadaceae* UKL13-1, *Betaproteobacterium* UKL13-2, and *Bacteroidetes* UKL13-3 (Table 2.6). There were insufficient reads to close the genome of *A. flos-aquae*, although 67 contigs could be clustered to represent an estimated 97% of the genome (Table 2.6). Contigs from partial genomes of two additional bacteria were also clustered: a novel *Flavobacterium* (63% estimated genome completeness) and a novel *Brevundimonas* (*Caulobacterales*) bacterium (17% estimated genome completeness) (Table 2.6), which were identified via PhylopythiaS+. The *Flavobacterium* genome contained 16S rDNA genes with 98% similarity to *Flavobacterium aquatile* DSM 1132, but no 16S gene was identified in the *Brevundimonas* contigs. Our results indicate the presence of at least six separate bacterial taxa in this non-axenic culture. A parallel Illumina HiSeq 2000 metagenome allowed comparison of PacBio-only and Illumina-only assemblies. When assembled with Illumina reads, the three predominant genomes separated into bins containing 100 or more contigs. The *Betaproteobacterium* genome bin contained more contigs than the *Hyphomonadaceae* and *Bacteroidetes* genomes, although it was sequenced at the highest Illumina depth of the three (63x coverage vs. 23x and 58x coverage, respectively). There was a 200 kb discrepancy between Illumina bin length and completed genome length for each of the three genomes. The total binned contig lengths for the *Bacteroidetes* and *Betaproteobacterium* were each shorter than the completed genomes, while the *Hyphomonadaceae* bin length was longer (Table 2.7). The additional sequences in the *Hyphomonadaceae* bin were primarily contigs shorter than 10 kb that were not part of the PacBio-assembled *Hyphomonadaceae* genome. The bin quality control

program CheckM [Parks et al. 2015] overestimated genome completeness or underestimated contamination when compared with the finished genome size. For example, CheckM estimated that the *Hyphomonadaceae* UKL13-1 bin contained 2% contamination, while comparing the bin length with the completed genome length suggests 6% contamination (Table 2.7). These discrepancies indicate that genome binning has a tendency to exclude important sequences or include extraneous sequences, and reveals the difficulty of assessing binned genome completeness and contamination without a reference. Incomplete binning is common for draft genomes, particularly from metagenomic assemblies [Hess et al. 2011].

We also assessed the extent to which genome repeats affected Illumina assemblies. Repeats in each genome were identified by using BLASTN to align each genome with itself, with a minimum E-value cutoff of 1E-30. Both intragenome BLASTN hits and missing Illumina coverage were then visualized with a circular genome plot (Figs. 2.7 - 2.9). Breaks in Illumina assemblies commonly co-localized with intragenomic repeats in each genome. In particular, the *Betaproteobacterium* UKL13-2 genome is enriched for repeat sequences relative to the other two genomes and contains larger regions unassembled by Illumina reads, factors that possibly contributed to the greater genome fragmentation (Table 2.7). We then analyzed gene functions in sequences missing from Illumina bins to assess the extent to which critical gene content was missing (Fig. 2.11). Most annotated genes in these regions were assigned to the mobilome category (esp. transposases), although genes from most other COG categories were also represented. Annotations within these regions included essential genes such as tRNAs, rRNA operons, translation-

associated genes (e.g. translation elongation factor Tu, ribosomal proteins L21, L27), and nucleotide metabolism genes (DNA polymerase III alpha subunit), in addition to a variety of enzymes and transporters (e.g., glycerol-3-phosphate dehydrogenase) (Tables 2.10-2.12). The presence of multiple rDNA sequences commonly produces breaks in short-read assemblies [Koren et al. 2013]. In such cases, rDNA sequences confined to small contigs lose their linkage to other genes. This makes assigning 16S sequences to draft genomes difficult when multiple organisms are present in the same sample, and can make it difficult to link 16S amplicon information to shotgun metagenomes. Also, the functional variety of non-mobilome-associated missing genes within these assembly breaks shows that they can hold informative sequences regarding physiology or lifestyle.

2.5.2 Novel Completed Genomes

To functionally characterize the three novel genomes, we searched all protein-coding sequences against the COG database using RAPSEARCH and a 1E-30 E-value cutoff. We then repeated this for all bacterial genomes in GenBank (collected on November 3, 2015) and compared these to our novel genomes to assess enrichment of protein-coding sequences associated with each COG functional category. These are shown as a percentage of all protein-coding sequences from each respective genome (Fig. 2.10). Our results indicate that the *Hyphomonadaceae* UKL13-1 genome contains more lipid metabolism (I) genes than most bacteria (at 5.01% vs. a mean of 2.96%), while the *Bacteroidetes* UKL13-3 genome contains more cell

wall/envelope/membrane biogenesis genes (M) (7.39%, vs. a mean of 4.61%) We then searched the KEGG database to identify complete and partial pathways in each genome. Identification of additional genes was aided by using Mauve whole- or partial-genome alignments [Darling et al. 2004] to reference genomes (*Cytophaga hutchinsonii*, *Roseobacter denitrificans*, *Rubrivivax gelatinosis*, and *Rhodobacter capsulatus*) and between *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2. The *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2 genomes contain anoxygenic photosynthesis and reaction center genes, as well as genes for bacteriochlorophyll and carotenoid synthesis. The 16S rDNA genes from these two genomes did not cluster near groups containing phototrophic bacteria (e.g. *Rhodobacter*, *Rhodoferax*) (Fig. 2.1, 2.2). Neither genome contains RuBisCO genes, consistent with these bacteria being aerobic anoxygenic phototrophs (AAP's). These are a class of heterotrophs that use phototrophy as a source of ATP production, but are unable to fix net carbon through photosynthesis [Moran and Miller 2007]. For *Betaproteobacterium* UKL13-2, the presence of genes for thiosulfate or sulfite oxidation (*soxABCDXYZ*), suggests that reduced sulfur compounds can serve as electron donors for ATP synthesis, perhaps in addition to organic compounds or during hypoxic conditions. Both *A. flos-aquae* and *Betaproteobacterium* UKL13-2 appear to be capable of assimilatory sulfate reduction of $MgSO_4$ (provided as the only S source in the growth medium), which is often used for amino acid synthesis. Photolithotrophic oxidation of reduced S compounds by the *Betaproteobacterium* would be energetically advantageous when using reduced S compounds derived from *A. flos-aquae*. Since neither genes for oxidation of re-

duced sulfur nor nitrogen compounds are evident in the *Hyphomonadaceae* genome, organic compounds likely serve as electron donors in this bacterium [Moran and Miller 2007].

In contrast with the proteobacterial genomes, *Bacteroidetes* UKL 13-3 contains no autotrophic genes, consistent with the typical lifestyle of these bacteria [Newton et al. 2011]. However, fewer genes were annotated from *Bacteroidetes* UKL13-3, and fewer completed KEGG pathway modules were identified than for the *Hyphomonadaceae* or *Betaproteobacterium* genomes (38 vs. 72 and 80, respectively). This could be due to protein-coding sequences carrying distant homology to those currently deposited in KEGG, limiting the ability to identify metabolic genes and pathways.

The *A. flos-aquae* genome was the only identified source of nitrogen fixing genes in the culture. Since the growth medium was nitrogen-deplete, all other bacteria in the community likely depend on reduced N provided by the cyanobacterium. Ploug et al. have shown that *A. flos-aquae* from the Baltic Sea fixes N_2 and releases it as NH_4^+, which is then taken up by surrounding heterotrophic or phototrophic bacteria [Ploug et al. 2010; Adam et al. 2016]. Both proteobacterial genomes contain the ammonium transporter gene *amtB*, which would allow uptake of NH_4^+ released by *A. flos-aquae*. No ammonia channel transport genes were annotated in the *Bacteroidetes* UKL13-3 genome. The proteobacterial genomes contained a number of chemotaxis and motility genes, which may be necessary for these organisms to stay associated and obtain benefits from *A. flos-aquae*, similar to other host-associated bacteria [Lertsethtakarn et al. 2011].

We searched the novel genomes for the presence of other transporters to inform of the needs for survival and growth. Both proteobacterial genomes contain transporters for alkanesulfonate, iron(III), phosphate, and phosphonate. The *Hyphomonadaceae* genome also contains a transporter for putrescine, while the *Betaproteobacterium* genome contains complete transporter modules for tungstate, molybdate, glutamate/aspartate, and branched-chain amino acids. Few, and only broadly functional transporter modules were identified in the *Bacteroidetes* genome. All three genomes appear to carry complete genetic pathways for nucleotide biosynthesis, as well as genes for synthesis of all 20 amino acids, indicating these organisms are self-sufficient in this regard. Because the *Flavobacterium* and *Brevundimonas* genomes were so incomplete, their gene content is not reported here.

We were unable to identify any plasmids in the assemblies. Shintani et al. classified the distribution of all plasmids in GenBank, and showed that the majority were found in *Proteobacteria* (47%), although most of these were associated with *Gammaproteobacteria* (63%), rather than *Alphaproteobacteria* (22%) or *Betaproteobacteria* (8.7%) [Shintani et al. 2015]. Plasmids from *Bacteroidetes* were much rarer at 1.6%. It may then be unsurprising that these bacteria lack plasmids.

2.5.3 Freshwater Bacteria Associated With Cyanobacterial Blooms

Bacteria from these three taxa are common in freshwater systems [Newton et al. 2011], are known to be commonly associated with cyanobacterial blooms, and can directly influence the growth of cyanobacteria in culture [Berg et al. 2009]. Some

Alphaproteobacteria have been identified in cyanobacterial-associated communities [Louati et al. 2015]. For example, Eiler et al. identified *Alphaproteobacteria* 16S rDNA sequences associated with another nitrogen-fixing cyanobacterium, *Gloeotrichia echinulata* [Eiler et al. 2006]. Interestingly, 16S rDNA from *Hyphomonadaceae* UKL13-1 shared significant identity (Table 2.6) with one of these sequences (A0904), suggesting that bacteria related to *Hyphomonadaceae* UKL13-1 are associated with various bloom-forming cyanobacteria. However, the extent to which such co-occurrences reflect physiological interdependencies remains to be explored.

Betaproteobacteria are often co-cultured with algae [Pernthaler et al. 2001], and have been seen physically associated with cyanobacteria [Louati et al. 2015; Eiler et al. 2006]. However, *Betaproteobacteria* are abundant in freshwater lakes [Hiorns et al. 1997], and their presence in co-culture may be due to their ability to survive off cell turnover. For example, many *Betaproteobacteria* are highly efficient at dissolved organic matter (DOM) degradation [Worm and Sondergaard 1998]. *Betaproteobacterium* UKL13-2 may thrive during increased *A. flos-aquae* cell turnover, which would provide DOM for survival. Based on 16S similarity searches, *Betaproteobacterium* UKL13-2 is not part of the widely distributed bet or Pnec clades found in freshwater lakes across the world (Table 2.1) [Newton et al. 2011]. With predicted chemotaxis and flagellar and twitching motility genes, both *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2 may actively seek out alive or dead *A. flos-aquae* cells as sources of nutrition. We have detected no genes by which these photoheterotrophic bacteria could obviously benefit *A.*

flos-aquae.

Bacteria from the *Bacteroidetes* phylum are commonly identified in, and sometimes dominate, freshwater lake systems [Pernthaler et al. 2004]. They are also frequently found in particle-associated communities and commonly degrade extracellular polysaccharide matrices that are grazed via bacteria that move through gliding motility [Lemarchand et al. 2006]. *Bacteroidetes* UKL13-3 possesses annotated gliding motility genes, which may indicate physical association with the originally isolated *A. flos-aquae* colony. Extracellular mucilage, as well as a range of nutrients (reduced C, N and S compounds) released by *A. flos-aquae*, may support the growth of *Bacteroidetes* UKL13-3, whose genome seems to lack many functionally annotated pathways. *Bacteroidetes* UKL13-3 has the only annotated extracellular peroxidase gene in the three genomes, which could protect against reactive oxygen species generated by photosynthesis in *A. flos-aquae*. Also, there are no annotated peroxidase genes in the *A. flos-aquae genome*. This may indicate a mutual benefit for both bacteria, and conform to the Black Queen Hypothesis defined for interactions between the unicellular cyanobacterium *Prochlorococcus* with other interacting bacteria [Morris et al. 2012]. On the other hand, large populations of *Bacteroidetes* bacteria are often observed following cyanobacterial bloom decline [Eiler and Bertilsson 2007] due to subsequently favorable conditions for copiotrophs [Zeder et al. 2009]. *A. flos-aquae* cell turnover may have provided dissolved organics for *Bacteroidetes* UKL13-3 growth in co-culture, as for the two *Proteobacteria*.

2.5.4 Metagenome Search

We also searched for the occurrence of these bacteria in 62 freshwater lake metagenomes from 8 sampling sites across the United States, including Oregon, Washington state, California, Texas, and Kansas (BioProject accessions: PRJNA312985, PRJNA282166, PRJNA312830, PRJNA312986, and PRJNA294203, respectively). To do so, we mapped reads from these metagenomes to the references with BWA-MEM with default parameters (0.067% error rate) and calculated average genome coverage. Matches were found in two samples. A metagenome from Copco Reservoir, CA on the Klamath River downstream of UKL on September 19, 2007 contained 86x read coverage of the *Hyphomonadaceae* UKL13-1 genome and 151x coverage of the *Bacteroidetes* UKL13-3 genome from 398,356,734 Illumina read pairs. Additionally, a metagenome from Cranberry Lake, WA on August 11, 2014 contained the *Betaproteobacterium* UKL13-2 genome at 99x coverage in from 13,955,857 Illumina read pairs. We also searched in 50 additional freshwater lake metagenomes in the IMG, MG-RAST, and SRA databases. The only detection found was the *Betaproteobacterium* UKL13-2 genome at 19x coverage in a metagenome consisting of 319,415,720 Illumina read pairs labeled "vibrio metagenome HEM-04" from a freshwater lake (BioProject accession: PRJNA64039). This initial analysis shows that the three novel bacteria are found elsewhere in freshwater habitats, although they do not appear to be ubiquitous or widely abundant.

2.6 Conclusions

Here, we have shown that completing multiple genome assemblies is possible from a simple microbial community using PacBio sequencing, a feat that is nearly impossible with short-read shotgun sequencing alone. There are several advantages to this approach. Completing genome assemblies from a shotgun metagenome avoids genome gaps and excludes contaminant sequences, which are significant issues with binned draft genomes. Absent sequences can contain functionally relevant information, such as gene clusters encoding secondary metabolites [Harrison and Studholme 2014] or antibiotic resistance genes near mobile elements [Zowawi et al. 2015]. Here we observed that key essential genes (Tables 2.10-2.12) were missing from each short-read assembly. Also, short-read assemblers can compress small repeats, potentially removing important functional information [Brown et al. 2016]. In addition to providing more complete genomic information, long-read sequencing of communities such as mixed cultures or environmental samples creates possibilities for new experimental designs. For example, complete genomes from novel organisms sequenced from the environment can be used as new references for culture-free resequencing efforts, such as to explore gene linkage patterns among alleles in a population. Further, long-read sequencers often detect DNA modifications, such as methylation, allowing capture of epigenetic information from environmental sequencing runs. Although PacBio sequencing is low-throughput compared with short-read sequencers, our results suggest that the current state of this technology allows genome sequencing from communities with relatively low di-

versity, such as those in extreme environments [Méndez-García et al. 2015] or when dominated by one or a few organisms [Lin et al. 2015]. Platform improvement, such as the recently released PacBio Sequel instrument, is expected to make long-read sequencing increasingly desirable for shotgun metagenomics in the future. Here, we have sequenced three novel genomes that may be associated with *A. flos-aquae* as part of the cyanobacterial phycosphere. Based on gene annotations and growth medium, both *Proteobacteria* are motile aerobic anoxygenic phototrophs that may utilize fixed nitrogen and carbon provided by *A. flos-aquae*. *Bacteroidetes* UKL13-3 is a heterotroph that likely has similar nutritional requirements, and may exist in a mutual relationship with *A. flos-aquae* through provision of an extracellular peroxidase. In future work, it will be interesting to explore the possible existence and nature of dependencies between these novel bacteria and *A. flos-aquae* colonies in blooms in Upper Klamath Lake and elsewhere.

Consensus Placement	**16S Silva**	**PhyloPythiaS+**	**Phylosift**
Hyphomonadaceae	Uncultured *Hyphomonadaceae* (99.79% identity)	*Alphaproteobacterium*	*Alphaproteobacterium*
Betaproteobacteria	Uncultured *Nitrosomonadaceae* (99.72%)	*Proteobacterium*	*Betaproteobacterium*
Bacteroidetes	*Sphingobacteriales*	*Flavobacterium*	*Bacteroidetes*

Table 2.1: Taxonomic placement of each novel genome by 16S similarity, composition (PhyloPythiaS+), and multiple marker gene similarities (Phylosift).

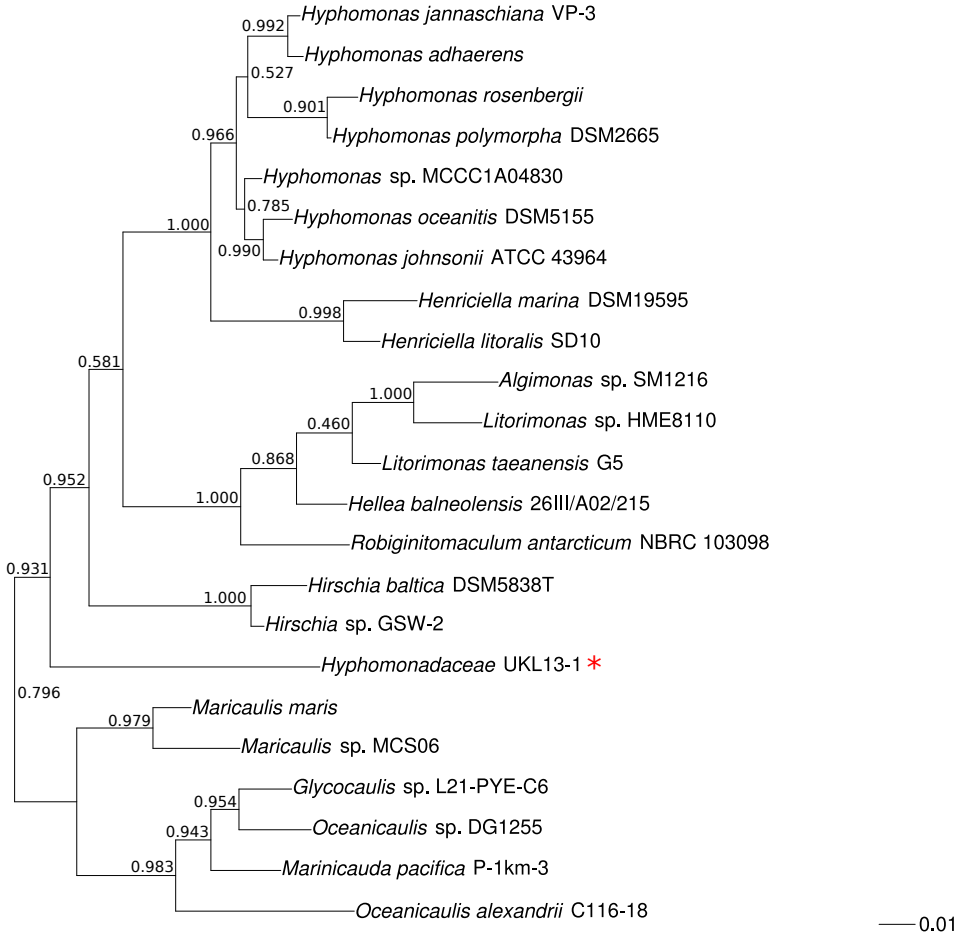

Figure 2.1: *Hyphomonadaceae* UKL13-1 1 16S phylogenetic tree.

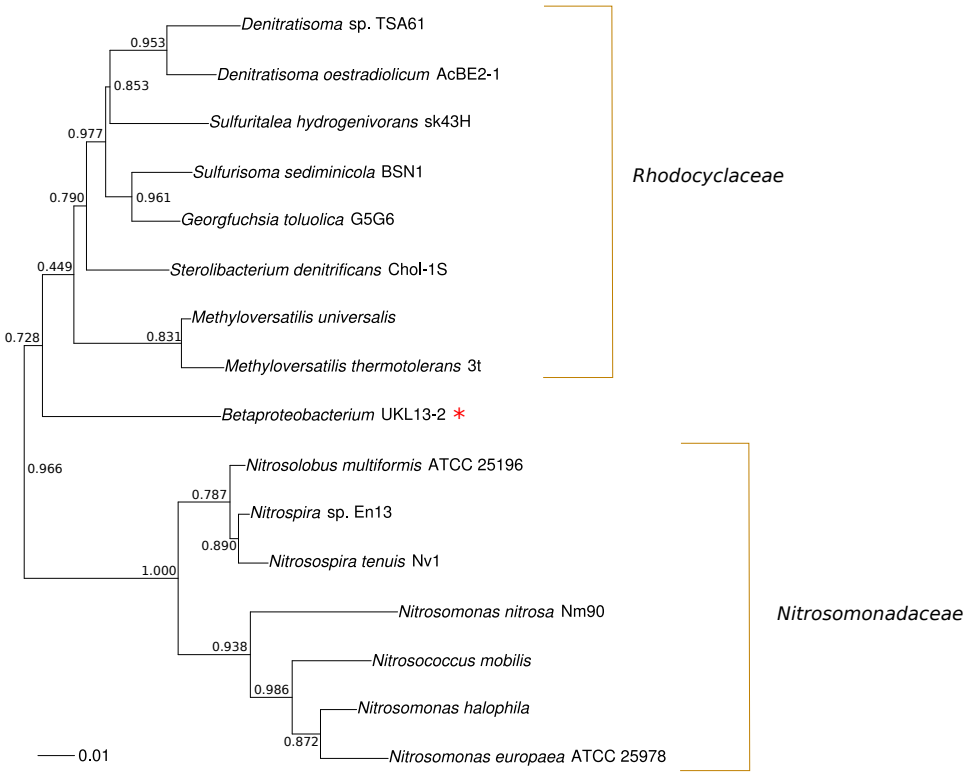

Figure 2.2: *Betaproteobacterium* UKL13-2 16S phylogenetic tree.

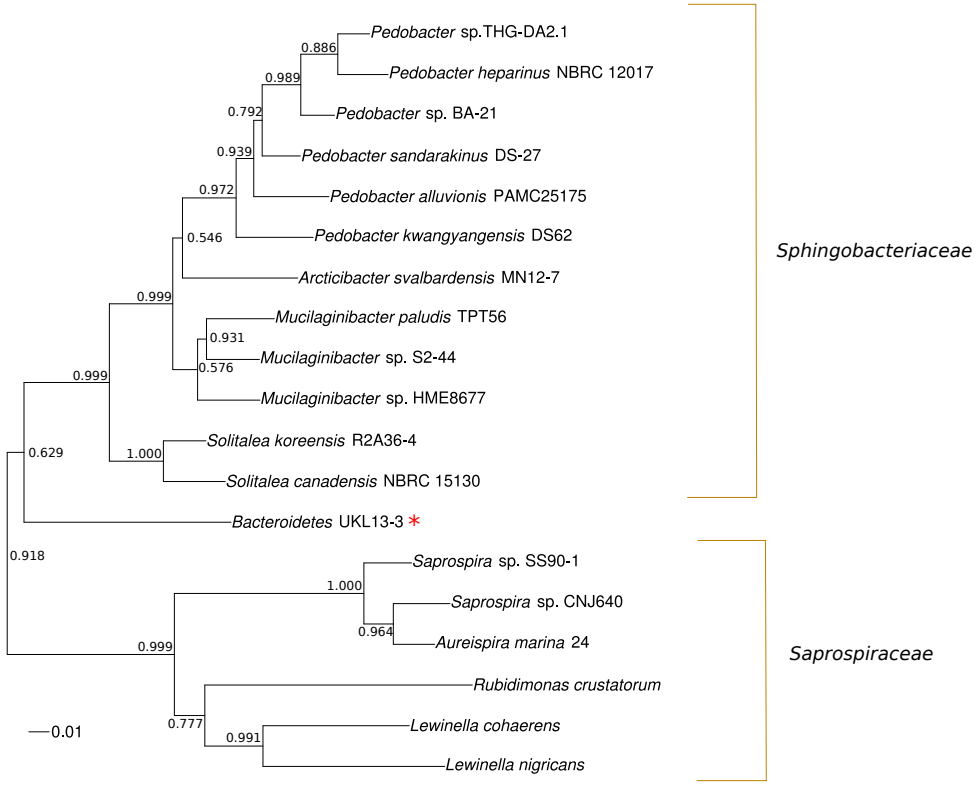

Figure 2.3: *Bacteroidetes* UKL13-3 16S phylogenetic tree.

		***Hyphomonadaceae* UKL13-1**		***Betaproteobacterium* UKL13-2**		***Bacteroidetes* bacterium UKL13-3**	
MIGS ID	**Property**	**Term**	**Evidence code**	**Term**	**Evidence code**	**Term**	**Evidence code**
	Classification	Domain Bacteria	TAS [63]	Domain Bacteria	TAS [Woese et al. 1990]	Domain Bacteria	TAS [Woese et al. 1990]
		Phylum *Proteobacteria*	TAS [Garrity et al. 2006]	Phylum *Proteobacteria*	TAS [Garrity et al. 2006]	Phylum *Bacteroidetes*	TAS [Krieg et al. 2010]
		Class *Alphaproteobacteria*	TAS [Garrity et al. 2005a]	Class *Betaproteobacteria*	TAS [Garrity et al. 2005b]		
		Order *Rhodobacterales*	TAS [Garrity et al. 2005c]				
		Family *Hyphomonadaceae*	TAS [Lee et al. 2005]				
	Gram stain	Unknown	NAS	Unknown	NAS	Unknown	NAS
	Cell shape	Unknown	NAS	Unknown	NAS	Unknown	NAS
	Motility	Unknown	NAS	Unknown	NAS	Unknown	NAS
	Sporulation	Unknown	NAS	Unknown	NAS	Unknown	NAS
	Temperature range	22-28°C	NAS	22-28°C	NAS	22-28°C	NAS
	Optimum temperature	Unknown	NAS	Unknown	NAS	Unknown	NAS
	pH range; Optimum	7.5-8.5; Unknown	NAS	7.5-8.5; Unknown	NAS	7.5-8.5; Unknown	NAS
	Carbon source	Unknown	NAS	Unknown	NAS	Unknown	NAS
	Terminal electron acceptor	Unknown	NAS	Unknown	NAS	Unknown	NAS
MIGS-6	Habitat	Freshwater lake	NAS	Freshwater lake	NAS	Freshwater lake	NAS
MIGS-6.3	Salinity	0.25%	NAS	0.25%	NAS	0.25%	NAS
MIGS-22	Oxygen requirement	Aerobic	NAS	Aerobic	NAS	Aerobic	NAS
MIGS-15	Biotic relationship	Syntrophic	TAS [Morris et al. 2013]	Syntrophic	TAS [Morris et al. 2013]	Syntrophic	TAS [Morris et al. 2013]
MIGS-14	Pathogenicity	Unknown	NAS	Unknown	NAS	Unknown	NAS
MIGS-4	Geographic location	Upper Klamath Lake, Oregon, USA	NAS	Upper Klamath Lake, Oregon, USA	NAS	Upper Klamath Lake, Oregon, USA	NAS
MIGS-5	Sample collection	Aug 6, 2013	NAS	Aug 6, 2013	NAS	Aug 6, 2013	NAS
MIGS-4.1	Latitude	42°22' N	NAS	42°22' N	NAS	42°22' N	NAS
MIGS-4.2	Longitude	-121°55' W	NAS	-121°55' W	NAS	-121°55' W	NAS
MIGS-4.4	Altitude	1,260 m	NAS	1,260 m	NAS	1,260 m	NAS

Table 2.2: Classification and general features of UKL genomes according to MIGS specifications [Field et al. 2008]

		***Hyphomonadaceae* UKL13-1**	***Betaproteobacterium* UKL13-2**	***Bacteroidetes* bacterium UKL13-3**
MIGS ID	**Property**	**Term**		
MIGS-31	Finishing quality	Complete	Complete	Complete
MIGS-28	Libraries used	SMRT library prep	SMRT library prep	SMRT library prep
MIGS-29	Sequencing platform	PacBio	PacBio	PacBio
MIGS-31.2	Fold coverage	94x	143x	112x
MIGS-30	Assemblers	HGAP	HGAP	HGAP
MIGS-32	Gene calling method	GeneMarkS+	GeneMarkS+	GeneMarkS+
	Locus tag	AEM38	AEM42	AEM57
	GenBank ID	CP012156	CP012156	CP012156
	GenBank date of release			
	GOLD ID	Gp0126808	Gp0126809	Gp0126810
	BIOPROJECT	PRJNA290648	PRJNA290650	PRJNA290651
MIGS-13	Source material identifier	UKL13	UKL13	UKL13
	Project relevance	Environmental	Environmental	Environmental

Table 2.3: Project information

Extraction	Handling	Extraction Procedure	Sample Date(s)	Sequencing
1	1.2 μm GF/C filtration	Kit	11/01/13	Illumina 100 bp paired-end HiSeq 2000
2	1.2 μm GF/C filtration and whole sample	Phenol-chloroform	Nov 2013 & March 2015	PacBio RS

Table 2.4: DNA extraction procedures and respective sequencing technologies.

Genome	**PacBio Reads (SMRT cells)**	**Minimum read length cutoff**
Hyphomonadaceae UKL13-1	348,623	6 kb
Betaproteobacterium UKL13-2	348,623	13.6 kb
Bacteroidetes bacterium UKL13-3	167,289	12.8 kb

Table 2.5: Assembly parameters for genome assemblies from PacBio reads. Minimum read length cutoff is lowest read-length used for assembly, with remaining reads used for error correction.

Genome	**Assembly Length (bp)**	**No. contigs**	**PB read coverage**	**Completeness estimate**	**Contamination estimate**
Hyphomonadaceae UKL13-1	3,501,508	1	94x	-	-
Betaproteobacterium UKL13-2	3,387,087	1	143x	-	-
Bacteroidetes bacterium UKL13-3	3,236,529	1	112x	-	-
Aphanizomenon flos-aquae	4,250,721	67	40x	96.67%	0.22%
Unknown *Flavobacterium*	2,347,065	96	22x	62.67%	0.25%
Unknown *Caulobacterales* bacterium	487,875	53	6x	17.15%	0.00%

Table 2.6: Genomes identified from PacBio assemblies. PacBio read coverage calculated by mapping with BLASR [Chaisson and Tesler 2012]. Completeness and contamination estimates for incomplete genomes are from CheckM [Parks et al. 2015].

	Illumina coverage	**# Illumina contigs**	**Bin Coverage Parameters**	**Bin assembly length (bp)**	**Bin assembly (% of genome)**	**Bin estimated completeness**	**Bin estimated contamination**
Hyphomonadaceae UKL13-1	23x	122	Illumina: 15-40x PacBio: >49x	3,716,244	106.13%	98.48%	2.19%
Betaproteobacterium UKL13-2	63x	162	Illumina: 37-87x PacBio: 71-211x	3,131,899	92.47%	96.15%	1.42%
Bacteroidetes bacterium UKL13-3	58x	96	Illumina: 44-103x PacBio: >228x	3,009,740	92.99%	97.81%	0.55%

Table 2.7: Illumina assembly statistics for each genome. Contig number and assembly length are from extracted bins. Illumina coverage calculated by mapping with BWA-MEM. Bin coverage parameters used to bin Illumina assemblies with mmgenome. Assembly as % of genome is comparison of contig bin length with actual genome length. Completeness and contamination estimated with CheckM.

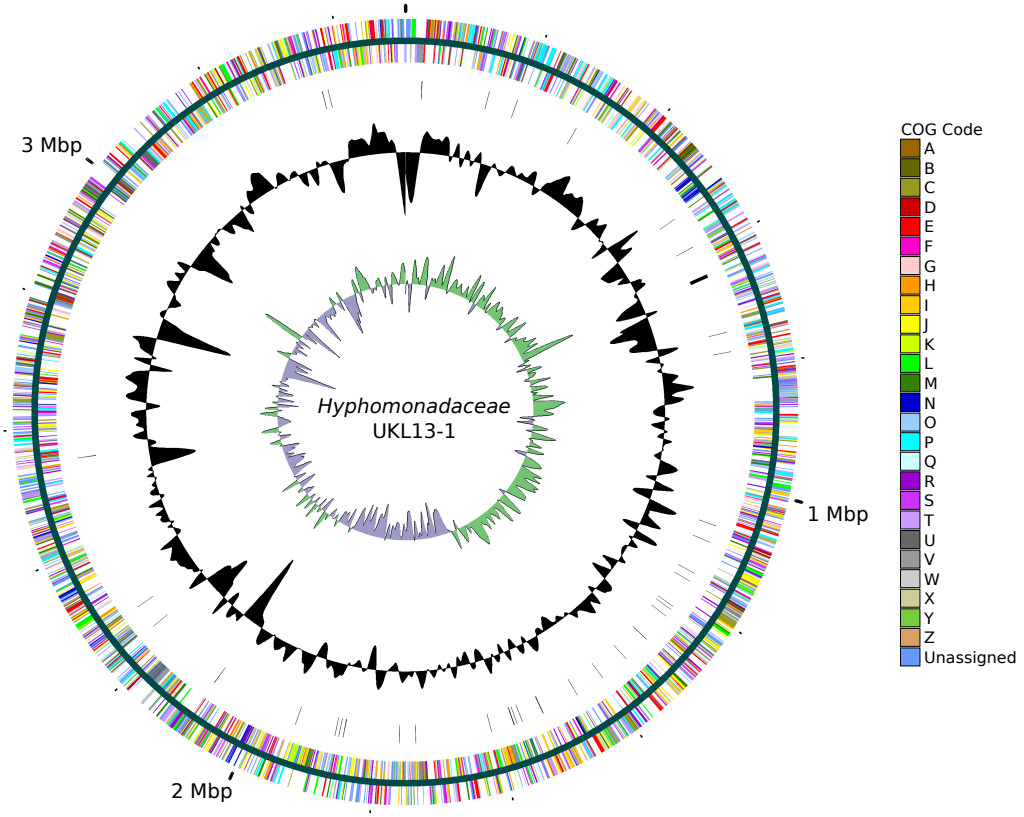

Figure 2.4: Circular map of the chromosome of *Hyphomonadaceae* UKL13-1. Circles from outermost radius to innermost: Predicted proteins encoded on the forward strand, colored by COG category; Predicted proteins encoded on the negative strand, colored by COG category; RNA genes; GC%, with peaks and troughs showing deviations from the average; GC skew, where green curves are positive skew values and purple curves represent negative skew values.

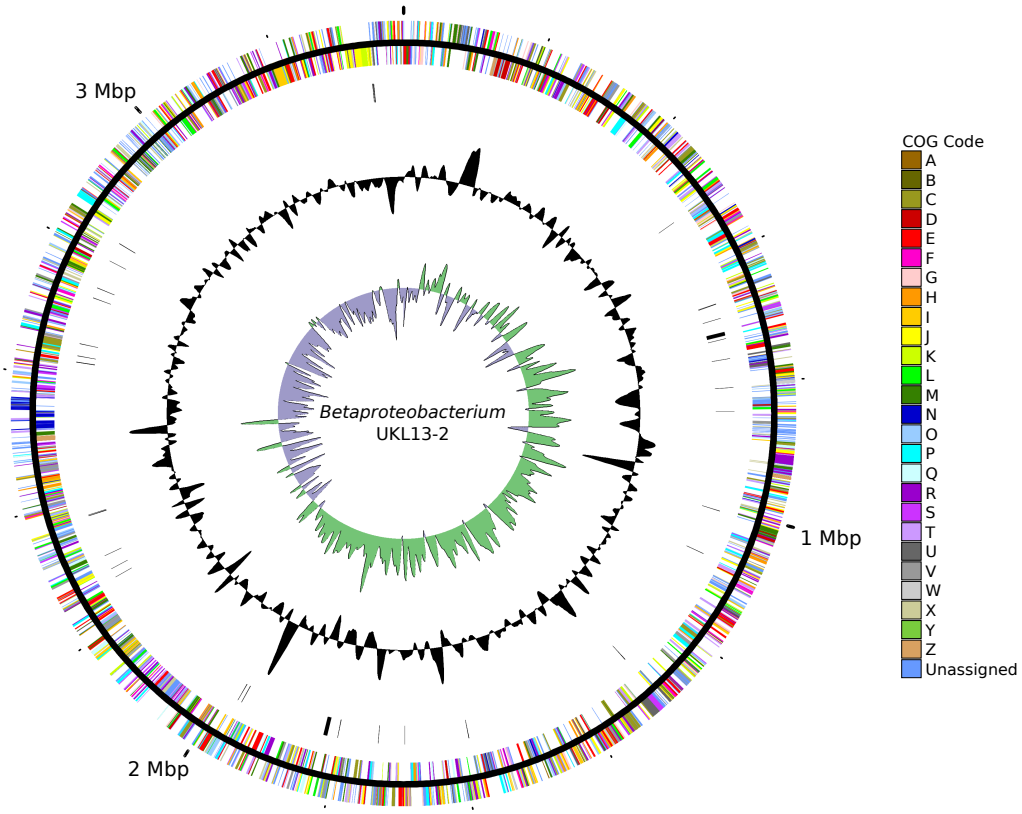

Figure 2.5: Circular map of the chromosome of *Betaproteobacterium* UKL13-2. See Fig. 2.4 caption for explanation.

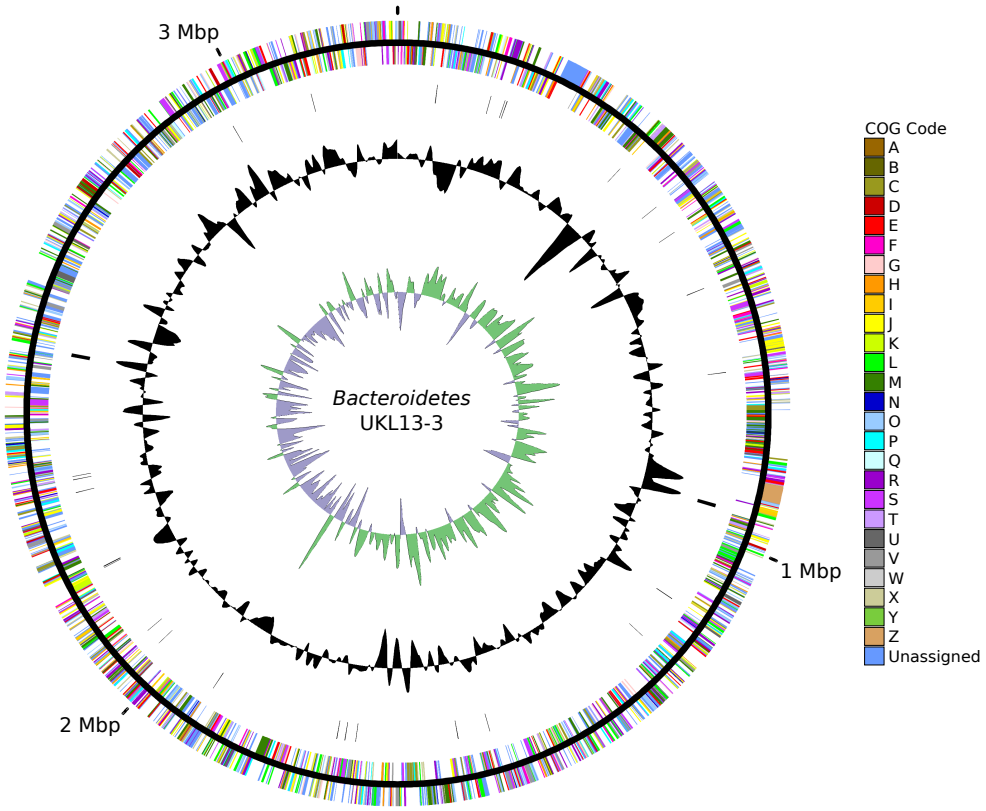

Figure 2.6: Circular map of the chromosome of *Bacteroidetes* bacterium UKL13-3. See Fig. 2.4 caption for explanation.

Attribute	***Hyphomonadaceae* UKL13-1**		***Betaproteobacterium* UKL13-2**		***Bacteroidetes* bacterium UKL13-3**	
	Value	% of total	Value	% of total	Value	% of total
Genome size (bp)	3,501,508	100	3,387,087	100	3,236,529	100
DNA coding (bp)	3,166,294	90.43	3,017,556	89.09	2,922,707	90.3
DNA G + C (bp)	1,964,937	56.12	1,862,116	54.98	1,208,228	37.33
DNA scaffolds	1		1		1	
Total genes	3255	100	3087	100	2850	100
Protein coding genes	2934	90.14	2772	89.8	2598	91.16
RNA genes	44	1.35	50	1.62	41	1.44
Pseudo genes	277	8.51	265	8.58	211	7.4
Genes in internal clusters	-	-	-	-	-	-
Genes with function prediction	2459	75.55	2300	74.51	1872	65.68
Genes assigned to COGs	2156	66.24	2078	67.31	1696	59.51
Genes with Pfam domains	2697	82.86	2489	80.63	2066	72.49
Genes with signal peptides	382	11.74	235	7.61	301	10.56
Genes with transmembrane helices (3)	310	9.52	271	8.78	255	8.95
CRISPR repeats	0		2		1	

Table 2.8: Properties and statistics for each genome.

	Hyphomonadaceae UKL13-1		*Betaproteobacterium* UKL13-2		*Bacteroidetes* bacterium UKL13-3		
Code	**Value**	**% of total**	**Value**	**% of total**	**Value**	**% of total**	**COG category**
J	184	4.91	187	5.07	175	5.34	Translation
A	1	0.03	1	0.03	1	0.03	RNA processing and modification
K	128	3.41	100	2.71	85	2.6	Transcription
L	109	2.91	100	2.71	126	3.85	Replication
B	2	0.05	2	0.05	1	0.03	Chromatin structure and dynamics
D	25	0.67	46	1.25	28	0.85	Cell cycle control
Y	0	0	0	0	0	0	Nuclear structure
V	69	1.84	77	2.09	74	2.26	Defense mechanisms
T	216	5.76	168	4.56	81	2.47	Signal transduction mechanisms
M	165	4.4	181	4.91	242	7.39	Cell wall/membrane/biogenesis
N	66	1.76	80	2.17	18	0.55	Cell motility
Z	0	0	18	0.49	1	0.03	Cytoskeleton
W	11	0.29	30	0.81	2	0.06	Extracellular structures
U	49	1.31	58	1.57	34	1.04	Intracellular trafficking
O	131	3.49	121	3.28	124	3.79	Posttranslational modification
C	135	3.6	187	5.07	114	3.48	Energy production and conversion
G	133	3.55	95	2.58	79	2.41	Carbohydrate transport and metabolism
E	197	5.25	224	6.08	127	3.88	Amino acid transport and metabolism
F	66	1.76	68	1.85	74	2.26	Nucleotide transport and metabolism
H	137	3.65	134	3.64	94	2.87	Coenzyme transport and metabolism
I	188	5.01	120	3.26	96	2.93	Lipid transport and metabolism
P	153	4.08	143	3.88	85	2.6	Inorganic ion transport and metabolism
Q	101	2.69	66	1.79	38	1.16	Secondary metabolites biosynthesis
R	223	5.95	213	5.78	211	6.44	General function prediction only
S	125	3.33	98	2.66	95	2.9	Function unknown
NA	1104	29.44	1083	29.39	1201	36.67	Not in COGs

Table 2.9: Number and proportion of genes associated with COG functional categories

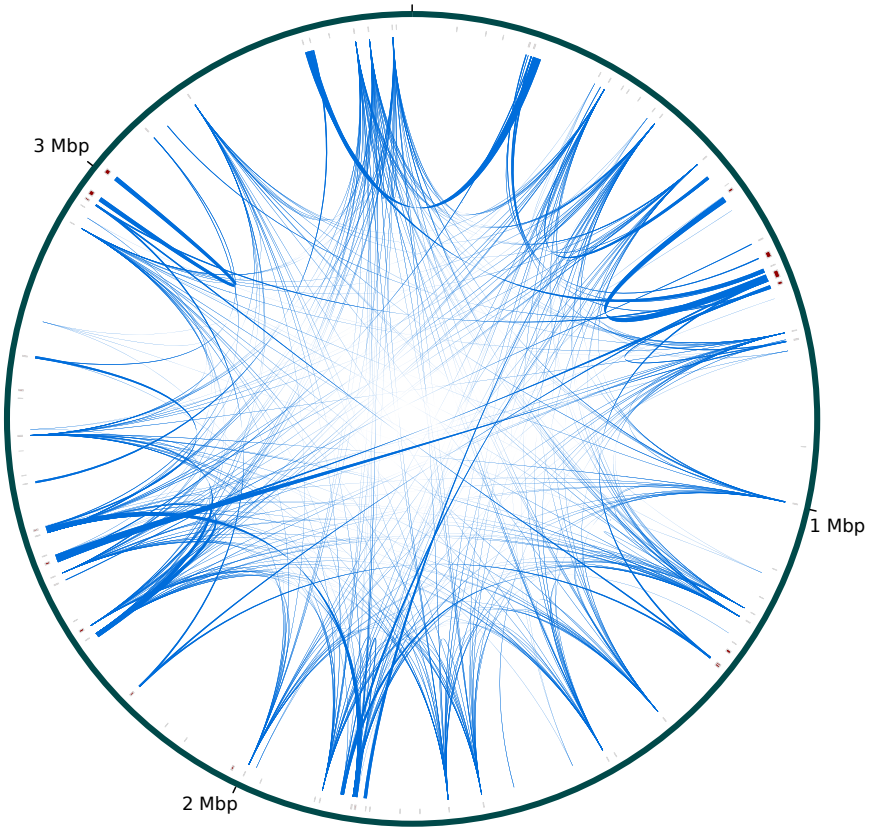

Figure 2.7: *Hyphomonadaceae* UKL13-1 genome repeats and Illumina breaks. Blue lines signify intragenomic repeats (based on BLASTN with a minimum E-value cutoff of 1E-30), and red bars mark sequences missing from Illumina assemblies.

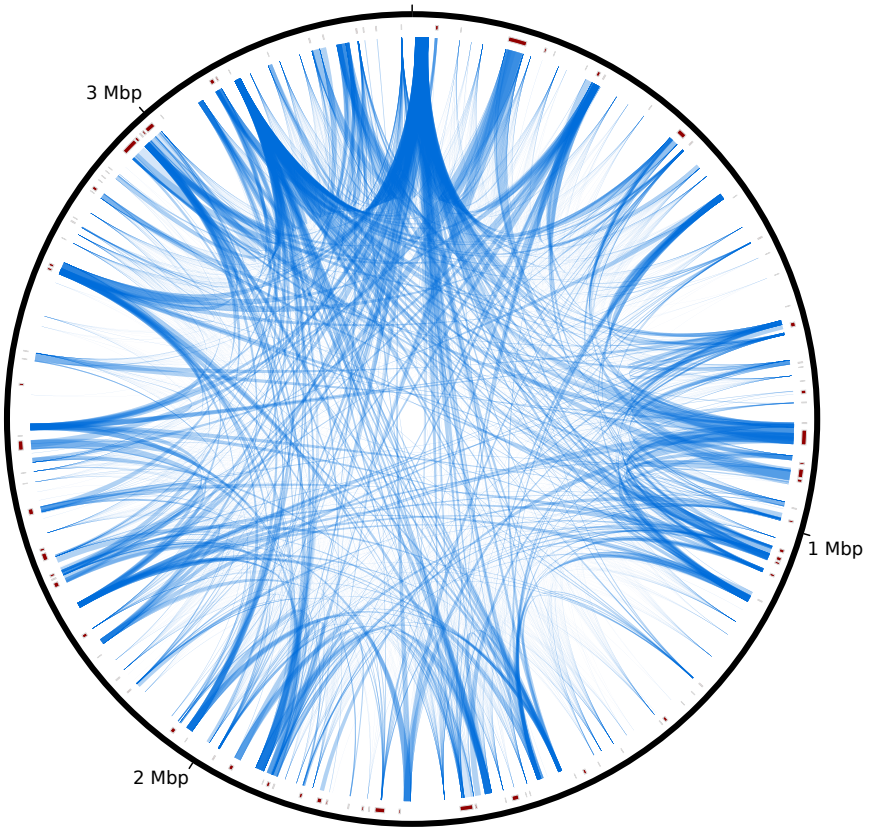

Figure 2.8: *Betaproteobacterium* UKL13-2 genome repeats and Illumina breaks. See Fig. 2.7 caption for explanation.

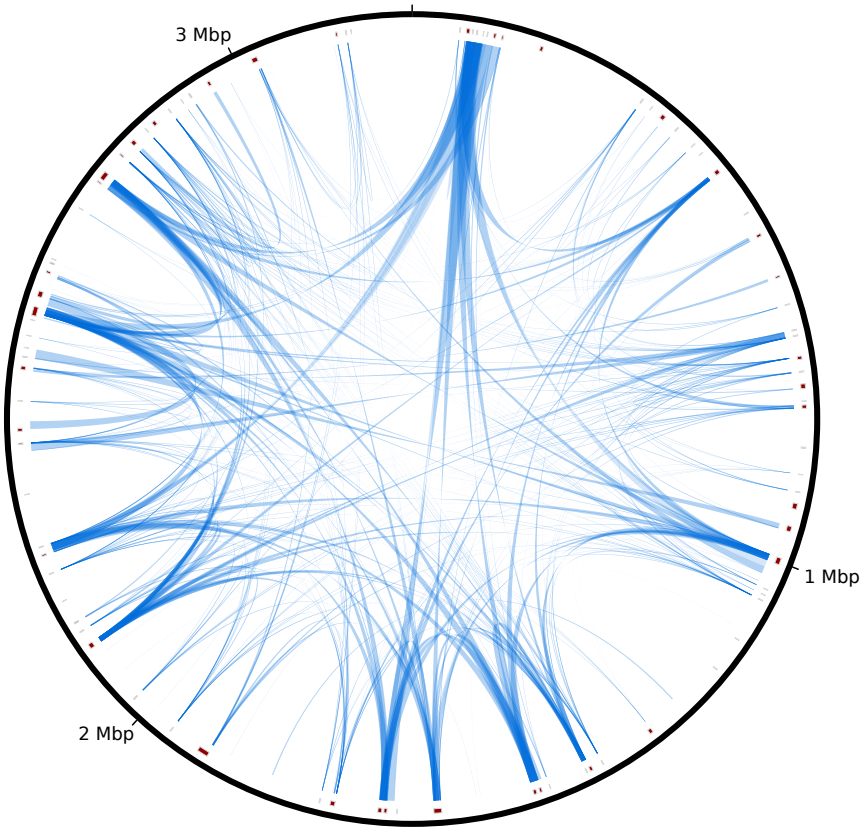

Figure 2.9: *Bacteroidetes* UKL13-3 genome repeats and Illumina breaks. See Fig. 2.7 caption for explanation.

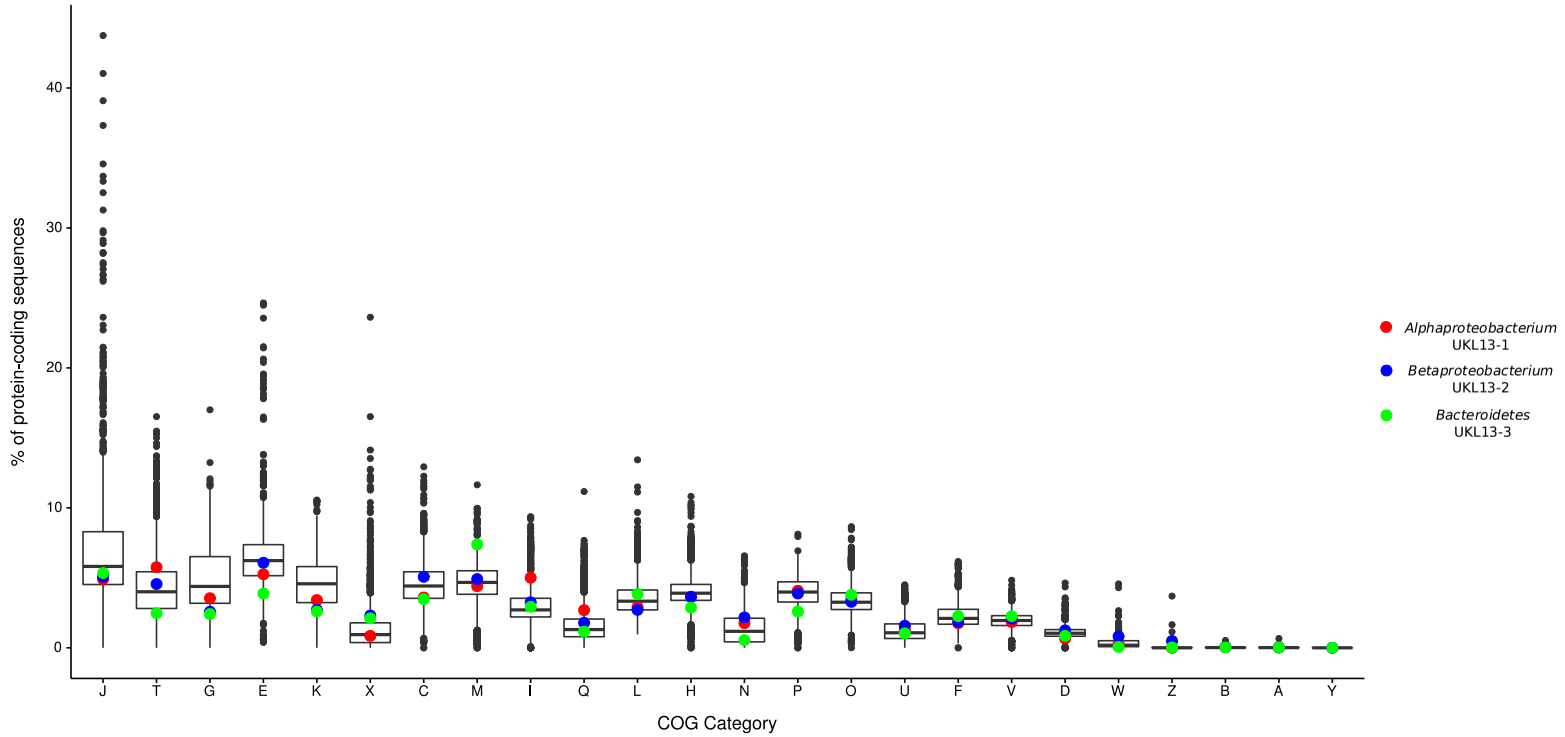

Figure 2.10: Percentage of protein-coding sequences from all bacterial genomes assigned to COG categories. Novel genomes are highlighted.

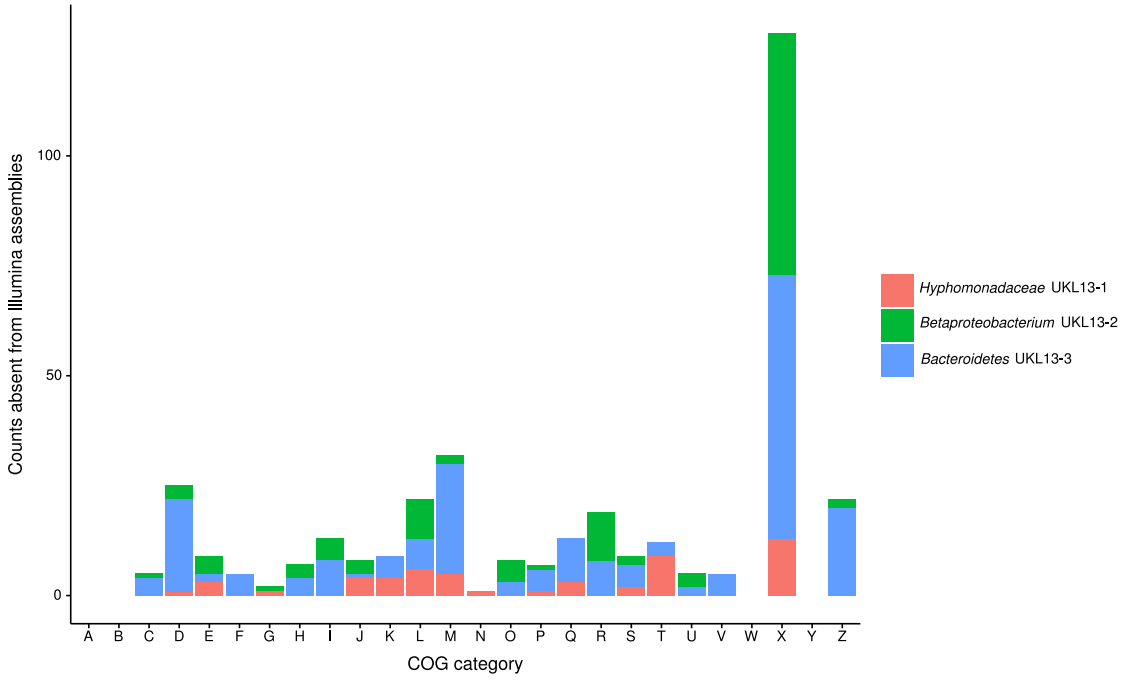

Figure 2.11: COG categories missing from Illumina assemblies determined by comparison to the closed genomes. Categories assigned with Rapsearch2. X is the mobilome COG category, while the rest of the category labels are annotated in Table 2.9

Genome start position	Genome end position	Annotation
105087	105752	Gram-negative bacterial tonB protein
314755	316659	magnesium chelatase subunit D
525899	526879	Cold shock-like protein 7.0
528897	528973	tRNA-Arg(acg)
635884	636315	Inosine-5'-monophosphate dehydrogenase
636962	638425	16S ribosomal RNA
659206	659775	cytochrome b561
659804	660520	Sensory transduction protein regX3
660517	661668	putative sensor histidine kinase TcrY
1093221	1094357	Anhydro-N-acetylmuramic acid kinase
1138645	1139184	Ribosomal large subunit pseudouridine synthase E
1163972	1165297	Multidrug export protein MepA
1210401	1213946	DNA polymerase III subunit alpha
1831751	1832941	Elongation factor Tu
2667538	2667849	50S ribosomal protein L21
2667873	2668142	50S ribosomal protein L27
3382686	3383465	Sulfite exporter TauE/SafE

Table 2.10: Notable annotated genes in *Hyphomonadaceae* UKL13-1 Illumina breaks (i.e., missing from Illumina assemblies). Genes called and annotated with PROKKA.

Genome start position	Genome end position	Annotation
409081	409818	Cytochrome c4
409906	411195	Glutamate-1-semialdehyde 2%2C1-aminomutase
411222	411950	Thiamine-phosphate synthase
411934	412848	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
413135	413317	Rubredoxin
712503	714033	16S ribosomal RNA
714162	714238	tRNA-Ile(gat)
714282	714357	tRNA-Ala(tgc)
714679	717940	23S ribosomal RNA
718207	718315	5S ribosomal RNA
1247149	1248405	Cobalt-zinc-cadmium resistance protein CzcB
1259988	1263734	rpt_family=CRISPR
1549239	1550837	All-trans-zeta-carotene desaturase
1551350	1554268	Vitamin B12 transporter BtuB
1734848	1734923	tRNA-Asn(gtt)
1809827	1810864	DNA ligase
1812260	1812910	DNA ligase
1819452	1820982	16S ribosomal RNA
1821111	1821187	tRNA-Ile(gat)
1821231	1821306	tRNA-Ala(tgc)
1821628	1824889	23S ribosomal RNA
1825156	1825264	5S ribosomal RNA
2345784	2346506	Lipoprotein-releasing system ATP-binding protein LolD
2346487	2347197	Lipoprotein-releasing system transmembrane protein LolE
2347402	2347869	Lipoprotein-releasing system ATP-binding protein LolD
2348328	2349311	Lipoprotein-releasing system transmembrane protein LolC
2349316	2350401	cofactor-independent phosphoglycerate mutase
2350471	2352180	Single-stranded-DNA-specific exonuclease RecJ
2356035	2357276	Anaerobic sulfatase-maturating enzyme
2587089	2588792	DNA repair protein RecN
2588806	2589681	putative inorganic polyphosphate/ATP-NAD kinase
2632025	2634364	Vitamin B12 transporter BtuB
2634720	2634809	tRNA-Ser(tga)
2873478	2877125	DNA polymerase III subunit alpha
2891997	2893424	GMP synthase [glutamine-hydrolyzing]
2898939	2899481	aldehyde dehydrogenase
3319589	3320779	Elongation factor Tu
3337015	3338205	Elongation factor Tu

Table 2.11: Notable annotated genes in *Betaproteobacterium* UKL13-2 Illumina breaks. Genes called and annotated with PROKKA.

Genome start position	Genome end position	Annotation
174030	174104	tRNA-Asn(gtt)
763067	763948	putative chromosome-partitioning protein ParB
764032	764742	Response regulator UvrY
846021	846407	S23 ribosomal protein
951944	953464	16S ribosomal RNA
953672	953746	tRNA-Ile(gat)
953760	953836	tRNA-Ala(tgc)
953954	956812	23S ribosomal RNA
956909	957014	5S ribosomal RNA
1372915	1373316	30S ribosomal protein S6
1373319	1373597	30S ribosomal protein S18
1373619	1374059	50S ribosomal protein L9
1374155	1374874	Riboflavin synthase
2109603	2109788	50S ribosomal protein L32
2109812	2110771	Phosphate acyltransferase
2110771	2111766	3-oxoacyl-[acyl-carrier-protein] synthase 3
2111864	2112343	Biotin carboxyl carrier protein of acetyl-CoA carboxylase
2112436	2113782	Biotin carboxylase
2411292	2412356	RNA polymerase-binding transcription factor DksA
3022790	3023815	Glycerol-3-phosphate dehydrogenase [NAD(P)+]

Table 2.12: Notable annotated genes in *Bacteroidetes* UKL13-3 Illumina breaks. Genes called and annotated with PROKKA.

Chapter 3 Nine novel *Anabaena* and *Aphanizomenon* genome sequences reveal the existence of a closely-related clade of globally distributed, bloom-forming cyanobacteria within the *Nostocaceae* family

Connor B. Driscoll[1], Timothy G. Otten[1], Nathan M. Brown[1], Kevin G. Meyer[3], Gregory J. Dick[3,4,5], Yanbin Yin[6], Zachary C. Landry[1], Theo W. Dreher[1,2]

[1]Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR, 97331, USA.

[2]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.

[3]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, Michigan, USA.

[4]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA.

[5]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA.

[6]Department of Biological Sciences, Northern Illinois University, DeKalb, Illinois, USA.

In preparation

3.1 Introduction

Cyanobacteria are a diverse set of primary producers that are important for ecosystems and global biogeochemical cycles. They have played an important role in atmospheric oxygen accumulation through oxygenic photosynthesis, while providing fixed carbon and occasionally nitrogen depending on the species [Karl et al. 1997; Canfield 2005]. Their diversity allows them to grow in a range of environments, including saltwater, freshwater, soil, and even deserts [Biller et al. 2014; Cheung et al. 2013; Lyra et al. 2001; Garcia-Pichel et al. 2001]. The *Nostocaceae* family primarily includes nitrogen-fixing filamentous cyanobacteria such as *Anabaena*, *Aphanizomenon*, and *Dolichospermum*, which commonly bloom in freshwater or brackish ecosystems around the world [Ikawa et al. 1982; Wang et al. 2012; D'Agostino et al. 2016b]. Some members of the *Nostocaceae* family threaten drinking-water supplies and recreational ecosystem use through production of harmful secondary metabolites [Cheung et al. 2013]. This issue is compounded by global climate change which facilitates increased frequency and duration of blooms (Paerl, 2009). Some members of the *Nostocaceae* family produce microcystin, cylindrospermopsin, or the potent saxitoxin or anatoxin neurotoxins [MacKintosh et al. 1990; Cheung et al. 2013]. A phylum-wide analysis of cyanobacterial genomes revealed widespread presence of non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) pathways, although most are associated with unknown end products [Calteau et al. 2014].

Previous studies attempting to connect morphological and genetic characteri-

zations of organisms belonging to this family have revealed taxonomic anomalies. Most prominently, *Anabaena* and *Aphanizomenon* strains are intermixed in phylogenies although their colony morphologies are visibly different [Gugger et al. 2002; Rajaniemi et al. 2005], while at least one *Anabaena* isolate has been reclassified and renamed as *Nostoc* [Shih et al. 2013]. Taxonomic revision proposals have affected this group over recent years, collecting many of the planktonic members into the genus *Dolichospermum* (leaving benthic species in the genus *Anabaena*) [Wacklin et al. 2009], but also creating two additional genera *Sphaerospermum* [Zapomělová et al. 2009] and *Chrysosporum* [Zapomělová et al. 2012]. These proposals have been based on polyphasic classification, which combines morphological and genetic information to create a taxonomy [Komárek 2016]. This approach has shortcomings, since morphological classification is subjective and colony morphologies are not always clearly distinguishable. Further, the genetic component underlying these proposed revisions has been narrow, relying solely on 16S rDNA phylogenies. These revisions have led to continual expansion of the *Nostocaceae* family through regular additions of putative novel genera. However, these classifications should be considered with caution pending genomic-level sequence information from a larger number and diversity of members of the clade.

The members of this family originate from diverse environments and exhibit varying lifestyles. For example, while almost all members fix nitrogen, *Raphidiopsis brookii* D9 does not [Stucken et al. 2010]. Also included in the family are a number of symbionts (*Richelia* [Gómez et al. 2005], *Nostoc punctiforme* PCC 73102 [Ran et al. 2007], and *Nostoc azollae* 0708) [Ran et al. 2010], the soil microbe *Cylin-*

drospermum stagnale PCC 7417, and saltwater-tolerant *Aphanizomenon flos-aquae* 2012/KM1/D3 [Šulčius et al. 2015] and *Nodularia spumigena* CCY9414 from the Baltic Sea. However, most members of this family were originally isolated from various freshwater systems.

Here, we have sequenced nine novel genomes consisting entirely of *Anabaena* and *Aphanizomenon* strains. Five of these genomes were computationally extracted from three separate environmental metagenomes, while the remaining six derive from cultures established from natural blooms. We assessed the phylogenomic relationships within these thirty-one genomes and assessed the distribution of secondary metabolite gene clusters. We also compared functional gene content to better understand cellular capabilities. In the process, we have identified a well-populated clade containing several subgroups that may represent a previously undersampled, but geographically widespread cyanobacterial lineage.

3.2 Methods

3.2.1 Genome selection and isolation

Novel genomes included in our analyses originated from a number of sites in the US, with each assembled from either environmental metagenomes or sequenced cultures (Table 3.1). Genomes obtained from environmentally sampled metagenomes (*Aphanizomenon* MDT14, *Anabaena* CRKS33, *Anabaena* MDT14, *Aphanizomenon* WA102-2, and *Anabaena* WA113) and the cultured *Aphanizomenon* MDT13 were

binned by differential coverage using the mmgenome R package [Albertsen et al. 2013]. Other genomes were binned with ESOM [Dick et al. 2009](*Anabaena* CPCC64, *Anabaena* AL09, and *Anabaena* LE011-02). The number of contigs, bin N50, and essential gene counts from mmgenome for each bin are listed in Table 3.2. We used CheckM to assess genome completeness and contamination [Parks et al. 2015](Table 3.3). Binned genomes were taxonomically classified using PhylopythiaS+ [Gregor et al. 2014].

3.2.2 Phylogenomic tree and group assignments

We generated a phylogenomic tree of the Nostocaceae family using the Hal pipeline [Robbertse et al. 2011]. In brief, this identifies orthologous protein clusters with all-vs-all BLASTP followed by MCL (Markov Cluster algorithm) clustering. Orthologous clusters are then aligned with MUSCLE, and the alignments are edited to remove segments that are poorly aligned. Each individual alignment is then concatenated into a single, super-alignment. An alignment model is then assigned with ProTest, and phylogenetic reconstructions performed with RAxML. The result is a phylogenomic tree built from alignments of all single-copy orthologues shared between all genomes.

Highly similar genomes were grouped based on whole-genome average nucleotide identity (gANI) and the fraction of each genome pair that is alignable (AF) [Varghese et al. 2015]. Varghese et al. suggest a cutoff for species assignments of 96.5% gANI and 0.6 AF. Here, we used a 95% gANI and 0.6 AF cutoff to

group genomes since *Anabaena* CRKS33 falls just outside the Varghese-suggested parameters, although we are not designating our groups as shared-species.

3.2.3 Core and pan-genome analysis

The core genome of the 31 *Nostocaceae* genomes were analyzed using the GET_HOMOLOGUES software package [Contreras-Moreira and Vinuesa 2013]. Homologous gene families were identified using the OrthoMCL clustering algorithm (OMCL) with sequence cluster reporting of t=0 and no Pfam-domain composition requirements [Vinuesa and Contreras-Moreira 2015; Contreras-Moreira and Vinuesa 2013; Fischer et al. 2011]. Core genome size was calculated using the exponential decay models of Tettelin and Willenbrock [Tettelin et al. 2005; Willenbrock et al. 2007] and the pan-genome size was estimated with the exponential model of Tettelin.

Additionally, a binomial mixture model [Snipen et al. 2009] classified genes based on distribution within all 31 analyzed genomes into four categories [Koonin and Wolf 2008]; core (occurring in all genomes), soft core (occurring in 95% of genomes and including core genes; [Kaas et al. 2012], shell (genes found in 3-18 genomes), and cloud (genes present in 1-2 genomes). A phylogenetic tree was produced by the PARS program of the PHYLIP suite [Felsenstein 2005] which used presence/absence data of the OMCL pan-genomic matrix [Contreras-Moreira and Vinuesa 2013].

The gene contents of individual taxa were compared using the parse_pangenome_matrix.pl script in GET_HOMOLOGUES.

3.2.4 Genome annotations

All genomes were annotated with GenBank's Prokaryotic Genome Annotation Pipeline (PGAP) [Angiuoli et al. 2008]. This pipeline includes rRNA and tRNA annotations by BLAST, and tRNAscan, respectively. In addition, all gene clusters from the pan-genome analysis were annotated with KEGG's BLASTKOALA using the genus_prokaryotes database (March 23, 2016) [Kanehisa et al. 2015]. Differences in gene content were assessed by the distribution of KO annotations, while specific gene categories (e.g., sulfur metabolism and photosynthesis) were also analyzed. Carotenoid-, vitamin-, and glutathione-synthesis pathways were assessed through KEGG annotations as well. All protein-coding sequences were also assigned to COG categories using Rapsearch 2.16 [Zhao et al. 2012] with the COG database and a 1E-30 E-value cutoff.

We searched through novel genomes for toxin synthesis gene clusters by BLASTN using a custom database containing secondary metabolite synthesis gene clusters as identified in Dittmann et al. [Dittmann et al. 2015]. This BLASTN search used an E-value cutoff of 1E-30, and clusters were identified where the total proportion of genes in a cluster were similar to greater than 50% of the reference cluster. In addition, we identified and counted gene clusters by using antiSMASH 3.04 without the inclusive option for all genomes [Weber et al. 2015].

All buoyancy genes were identified from PGAP annotations, including the previously characterized *gvpA* and *gvpC* genes. Peroxiredoxin, catalase, and superoxide dismutase genes were also identified from PGAP annotations. Insertion

sequences (IS) were identified using HMMSEARCH [Finn et al. 2011] with the TnPred IS Hidden Markov Model database (http://www.mobilomics.cl/downloads.html) and a 1E-30 E-value cutoff. This database contains 47 HMMs for 19 IS families. Extracellular polymeric synthesis (EPS) genes previously characterized by Pereira et al. [Pereira et al. 2009; 2015] were identified by BLASTP alignment against EPS genes in GenBank found in the *Nostocaceae* family with an E-value cutoff of 1E-30. The components of restriction-modification (R-M) systems within the genomes were identified by performing protein sequence searches with TBLASTN (e-value of 1E-100) against known R-M system protein sequences obtained from REBASE database (accessed on May 8, 2016) [Roberts et al. 2009].

3.3 Results

3.3.1 Evaluating binned genomes

There are benefits to analyzing binned genomes directly from environmental samples. Studies have shown that mutations occur in culture that affect the fitness of bacteria, often through reductive processes [Koskiniemi et al. 2012; Cooper et al. 2001; Wang et al. 2012]. By removing the bias of changes from cultivating these bacteria in a lab environment, we are able to observe these genomes in their natural state. This also reduces the effort necessary to establish difficult-to-culture organisms. Here we have included five novel genome bins extracted from three metagenomes, in addition to 4 novel cultured genomes.

While some of these genomes are completed, five were binned from metagenomes, and most others are draft quality (Table 3.1). We used CheckM on all genomes and the mmgenome R package to obtain universal gene counts and copy numbers for binned genomes (see Methods) (Tables 3.2, 3.3). In addition, contigs identified as contaminants by NCBI's WGS submission pipeline were removed. These results indicate low levels of contamination (0-4.22%), and each bin contains on average >97% of universal genes, with the exception of *Aphanizomenon* 2012/KM1/D3 and the three *Richelia* genomes (Table 3.3). Upon closer inspection we identified multiple, unique rDNA genes in some of these bins, which we subsequently removed from the respective bins. Our binning process could cluster contigs containing similar sequences from other bacteria, but our mmgenome, NCBI, and CheckM contamination measurements suggest this is minimal. Regardless, it is important to keep in mind there may be some small error in gene copy number counts within these bins. In addition, previous work has shown that draft genomes can exclude functionally relevant gene content, although at the level of single genes and not entire pathways (See Chapter 2).

3.3.2 *Nostocaceae* family phylogenomic characterization

In total, we have provided nine novel sequenced genomes belonging to the *Nostocaceae* family, bringing the total number up to 31 (as of September 2015). We assessed the evolutionary relationships within this family by generating a phylogenomic tree based on alignments of all single-copy shared orthologues from these

genomes [Robbertse et al. 2011] (Figure 3.1). In addition, we used pairwise genomic ANI and alignment fraction (AF) calculations to assign genomes to potential species groups [Varghese et al. 2015](Figure 3.1, Table 3.4). All but one of the new genomes (*Anabaena* CPCC64) formed a clade comprised of 4 or 5 species-level groups consisting of *Anabaena*, *Aphanizomenon*, and *Dolichospermum* strains. We refer to this clade as Clade AAD. We then grouped seventeen genomes into five separate groups (Figure 3.1), two of which (groups 3 and 4) contained representatives from both *Anabaena* and *Aphanizomenon* genera that were previously characterized by morphology [Brown et al. 2016; Šulčius et al. 2015; Cao et al. 2014]. In addition, gANI/AF grouping cutoffs were consistent with clusters in the phylogenomic tree. *Nostoc* and *Anabaena* genomes also don't clearly separate. For example, both *Anabaena* CPCC64 and *Anabaena* variabilis ATCC 29413 separate out with *Nostoc* PCC 7120 (also known as *Anabaena* PCC 7120).

Some of the other genomes also cluster together. For example, the nitrogen-fixing *Cylindrospermopsis raciborskii* CS-505 and the non-nitrogen-fixing *Raphidiopsis brookii* D9, which carry some of the smallest genomes for free-living filamentous cyanobacteria (3.9 and 3.2 Mb, respectively), form a monophyletic group, consistent with previous reports [Stucken et al. 2010; Shih et al. 2013]. The *Richelia* genomes also form a monophyletic cluster, although *Richelia intracellularis* RC01 is well separated from HH01 and HM01. Also, the large difference in size between these genomes indicates they are considerably diverged (5.4 Mb for RC01 compared with 3.2 and 2.2 Mb for HH01 and HM01, respectively). Alternatively, several genomes do not cluster closely with other genomes. These include *Nos-*

toc azollae 0708, *Cylindrospermum stagnale* PCC 7417, *Nostoc punctiforme* PCC 73102, and *Nodularia spumigena* CCY9414. The placement of these genomes indicate these groups are currently underrepresented and that there is likely more room for sequencing new members closely related to these strains.

Fifteen of the thirty-one genomes prominently belong to a single clade, including all but one of the novel genomes presented here (*Anabaena* CPCC64) (Figure 3.1). Within this clade, there are four subclades with high similarity over large portions of their genomes based on gANI calculations (Table 3.4). Each contained members from diverse geographic origins (Table 3.1). For example, group three included *Anabaena* WA93 and *Anabaena* WA102 from Washington State Lakes in the USA, while *Aphanizomenon flos-aquae* NIES-81 and *Aphanizomenon flos-aquae* 2012/KM1/D3 were isolated from Lake Kasumigaura in Japan and in the Baltic Sea, respectively. The variability of geographic origin and water-body indicates that these groups consist of strains that have in the past carried (or have obtained over time) the capability to survive in different environmental conditions. While it is likely the less-populated clades are underrepresented compared to this larger group, it still seems the strains of the 15-member clade are part of a closely-related, globally widespread group of genome-types that are separate from the remainder of the *Nosocaceae* family.

3.3.3 Core and pan-genome

The core genome for all *Nostocaceae* members was estimated by orthologous gene clustering. We identified 576 and 463.6 core genes with residual standard errors of 442.18 and 392.37 for Tettelin and Willenbrock fits, respectively (Figure 3.2). The pan-genome, estimated by Tettelin fit, was 16,298.3 genes with a residual standard error of 572.45 (Figure 3.3). Additionally, the binomial mixture model estimates 349 core (1.30%), 1372 soft core (5.13%), 6803 shell (25.41%), and 18,596 cloud genes (69.46%). These pan-genome numbers are likely underestimates, since the pan-genome curve is not asymptotic (Figure 3.3), which corresponds to the large size of the flexible genome and the shared taxonomic level of genomes included in this analysis.

Of the 349 core gene clusters, which are found in all thirty-one genomes, 322 (92.2%) were assigned to KEGG functional groups. The most prevalent core gene function is associated with the ribosome, of which there are 38 unique gene clusters in total (Figure 3.4). Genes associated with amino acid biosynthesis, photosynthesis, carbon metabolism, porphyrin/chlorophyll metabolism, and nucleotide metabolism are also common. Of the 1372 soft core genes, 585 (42.6%) were assigned to KEGG functional categories. These genes are found in all but one of the thirty-one genomes, and include the core genome set. More soft core genes are associated with amino acid biosynthesis than any other functional category. Carbon metabolism, ribosomal, ABC transporter, photosynthesis, and porphyrin/chlorophyll metabolism gene counts are abundant for this set.

Out of the 6803 shell genes (present in 3-18 genomes), 1510 (22.2%) were assigned to KEGG functional categories. ABC transporter genes are the most prevalent identified functional category in the shell genome, while amino acid biosynthesis and two-component systems are also found often. Only 1896 (10.2%) of the 18,596 genes in the cloud genome (present in 1-2 genomes) were assigned to KEGG functional categories. The distribution of genes were similar to the shell genome, where ABC transporter and two-component system genes are abundant, although carbon metabolism genes are more abundant here.

Overall, ribosomal genes, which are expected to be conserved, are found most often in the core gene sets [Shi and Falkowski 2008]. Genes whose presence are expected to be more variable (ABC transporters, two-component systems) are much more abundant in the shell and cloud genomes. Multiple functional categories are present in the shell and cloud, suggesting either variation in shared pathways, or similar function from paralogous genes which are assigned to different gene clusters. In particular, nitrogen metabolism genes are found more often in the shell and cloud genomes. Closer inspection suggests this is due to the lack of nitrogen fixation genes in *Raphidiopsis brookii* D9, which does not fix nitrogen, as well as the lack of transporters for nitrogenous compound uptake in the *Richelia* genomes as well as *Nostoc azollae* 0708. *Anabaena* MDT14 does not seem to contain nitrogen fixation genes, although it's possible these genes were not assembled.

From gene clustering, a total of 16,387 genes were identified as unique to a single genome. Of these genes, only 1477 genes were assigned to KEGG orthologues, leaving a large majority without functional annotation. The majority (5.8%) of

annotated unique sequences were assigned to ABC transporters, 23% of which were annotated as amino acid transporters. There is sometimes overlap in annotated functions across genomes (some of the *livGKM* branched-chain amino acid transporter gene clusters, for example), although sequence identity is low between these separate clusters. This suggests these genes either diverged while retaining function, or they are paralogs which have obtained new functions. This could explain why *Nostoc* PCC 7120 contains six copies of a putative iron complex transport system and *Anabaena* MDT14 carries four copies of the sulfonate transport gene *ssuA*. In addition, there are nitrogen, sulfur, iron, molybdate, and cobalt/nickel-related transport enzymes found uniquely throughout these genomes. The next largest group of annotated unique genes were annotated as carbon metabolism genes. Previous studies have shown that cyanobacterial central carbon metabolism is highly fragmented, which may be due to overlap in carbon metabolism pathways [Beck et al. 2012].

3.3.4 Toxin synthesis and secondary metabolite genes

Toxin synthesis in *Nostocaceae* members is of particular concern, since many of these strains grow in globally distributed freshwater systems and therefore may pose a threat to public health [Beltran and Neilan 2000; Bolch et al. 1999]. Additionally, secondary metabolites produced by cyanobacteria have allelopathic effects that can impact other organisms [Leão et al. 2009; Rzymski et al. 2014]. Previously sequenced *Nostocaceae* members have been characterized for their ability

to produce toxic compounds, so we identified toxin synthesis gene clusters in the novel genomes here by using BLASTN alignments against previously characterized nucleotide sequences. Of the nine novel genomes, none contained putative toxin biosynthesis gene clusters (Figure 3.5). However, we identified a number of other secondary metabolite synthesis clusters in all genomes. Geosmin synthesis genes, which encode the enzymes for synthesis of a taste-and-odor compound affecting drinking-water supplies [Jüttner and Watson 2007], were identified in six genomes, one of which was a novel genome (*Anabaena* CRKS33). Both *Anabaena* AL09, *Anabaena* LE011-02, and *Anabaena* 90 contain genes for synthesizing the protease inhibitor anabaenopeptolide [Rouhiainen et al. 2000], while these genomes and *Aphanizomenon flos-aquae* NIES-81 carry genes for anabaenopeptin synthesis [Itou et al. 1999; Murakami et al. 2000]. Genes for synthesizing the cyanobactin anacyclamide are found in *Anabaena* AL09, *Anabaena* LE011-02, *Anabaena* 90, *Anabaena* AL93, *Anabaena* WA102, and *Aphanizomenon flos-aquae* 2012/KM1/D3 [Leikoski et al. 2010]. *Cylindrospermum stagnale* PCC 7417 carries cylindrocyclophane synthesis genes, which encode for a proteasome inhibitor with measured cytotoxic effects [Chlipala et al. 2010]. *Anabaena* 90 contains a putative gene cluster for synthesis of hassallidin , which has been demonstrated to have antifungal properties [Vestola et al. 2014].

We also identified other secondary metabolite synthesis clusters. Polyketide synthase and terpene synthesis clusters were found in all genomes (Table 3.5). In addition, almost all genomes contained non-ribosomal peptide synthesis genes with the exception of the *Richelia* symbionts. Bacteriocins are found in sixteen

of the thirty-one genomes, although they are not exclusive or ubiquitous to any groups. Bacteriocins are toxic proteins that inhibit growth of other, sometimes closely related, bacteria that are often encoded in cyanobacterial genomes [Wang et al. 2011].

In addition, several genomes contain gene clusters for the synthesis of cyanobactins, which are bioactive cyclic peptides that are potential leads for novel antitumor, antimalarial, or other compounds [Donia et al. 2008]. Eleven of the thirty-one genomes analyzed here contained putative cyanobactin synthesis gene clusters, and in total fifteen were identified across all genomes. No more than two cyanobactin synthesis clusters were found in each genome (Table 3.5). Also, none were associated with predicted chemical structures by antiSMASH. Of the eleven putative clusters, nine were identified in the AAD clade, including all genomes in groups 1 and 2. Only the *Anabaena* genomes in group 3 carry cyanobactin synthesis genes, while no genomes from group 4 contain these genes. This suggests there may be group-specific patterns within the AAD clade in their ability to produce cyanobactins.

Other secondary metabolite gene clusters were identified, although they were not as prevalent. Lantipeptide synthesis genes were identified in six of the thirty-one genomes, with four found in genomes related with and part of group 5. Lantipeptides are another group of potentially valuable bioactive peptides that include the lantibiotic antimicrobials [Knerr and van der Donk 2012]. Their increased presence in *Nostoc* and related genomes indicates the potential for identifying novel lantipeptides produced by *Nostoc* strains. Microviridin synthesis genes were iden-

tified in three genomes (*A. flos-aquae* NIES-81, *N. spumigena* CCY 9414, and *Anabaena/Nostoc* PCC 7120). Microviridins are a group of serine-protease inhibitors, some of which can kill grazers [Rohrlack et al. 2004; Ziemert et al. 2010]. Ladderane synthesis genes were identified in seven genomes, although the distribution was inconsistent with the phylogenomic tree. Ladderane lipids may provide denser membranes than conventional cell membrane lipids, and are used by annamox bacteria to enclose the annamoxasome [Rattray et al. 2010]. Other clusters putatively synthesize proteusin, resorcinol, arylpolyene, lassopeptides, and thiopeptides, although these were identified in four or fewer genomes each.

3.3.5 Functional gene comparisons

To assign function to protein-coding sequences in each genome, we annotated clusters generated from the pan-genome analysis with the KEGG database and compared differences across the family (Figure 3.5). Some exceptions for more specific searches were used for some groups, as detailed in Methods. We then highlighted differences in annotated gene content between genomes.

3.3.5.1 Photosynthesis-associated genes

The distribution of photosynthetic genes associated with photosystem complex II (PSII) assembly is either dispersed or sparse depending on the genes in question, with no phylogenomic pattern. Twenty to thirty-one genomes contain the *psbOP*-

TUVXYZ, *psb27*, and *psb28* genes, suggesting presence of these genes is generally conserved in this family. In contrast, fewer genomes contain the *psbJKLM* and *psb28-2* genes, which are non-essential photosystem genes whose presence can affect photoautotrophic growth rates in cyanobacteria [Lind et al. 1993; Ikeuchi et al. 1991; Sakata et al. 2013; Bentley et al. 2008].

Twenty-seven of these genomes contain complete genes for synthesis of phycocyanin (*cpcABCDEFG*), a light-harvesting pigment ubiquitous in cyanobacteria which absorbs primarily orange/red light at 620 nm [Myers and Kratz 1955]. *Anabaena* MDT14 contained no *cpc* genes, although we hypothesize these were lost during the assembly/binning process. Phycoerythrin synthesis genes (*cpeABCRSTUYZ*) were identified in only four of the genomes, all of which were symbionts (the three *Richelia* strains and the plant symbiont *Nostoc punctiforme* PCC 73102) [Meeks et al. 2001]. Genes encoding the green-light harvesting pigment phycoerythrocyanin (*pecABCEF*) are dispersed among eleven of the genomes, and are found in group 5 as well as other *Nostoc* and *Anabaena* strains, while only found in two genomes from the AAD clade (*Anabaena* LE011-02 and *Anabaena* WA93). These genes are likely carried in strains that are in highly competitive environments for red-light absorption, or perhaps in deeper or more opaque aquatic systems [Ting et al. 2002]. The differential distribution of light-harvesting and photosynthesis genes suggests they are under differential selection depending on their respective environments, which likely vary in light availability.

3.3.5.2 Sulfur metabolism genes

Of the thirty-one genomes, sixteen carry all or most of the *ssuABCDE* operon, which is involved in organic sulfur uptake [van der Ploeg et al. 1999]. The *tauD* gene, involved in metabolizing taurine to sulfite for sulfur metabolism, was found in the same sixteen genomes in addition to *Anabaena* PCC 7108. Notably, these genes are entirely absent from group 4 genomes.This indicates there may be differential dependencies on sulfur in certain strains, or variation in sulfur availability in some environments. Some or all of the genes for assimilatory sulfate reduction (*cysCH*, *sat*, *sir*) in addition to the sulfate transporter *cysP* are found in all genomes but *Richelia intracellularis* HM01.

3.3.5.3 Nitrogen metabolism genes

The *nifV* gene was found in all genomes except *C. stagnale* CS-505, *R. brookii* D9, and *Anabaena* MDT14. This gene encodes for a homocitrate synthase which, when present, increases nitrogen fixation efficiency in *Nostoc* PCC 7120 [Stricker et al. 1997]. Additionally, twenty-one of the thirty-one genomes contain the *cydAB* genes, which encode for an oxidase essential for *Nostoc* PCC 7120 growth under nitrogen-limiting conditions [Mikulic 2013]. Previous work has also raised the possibility this oxidase scavenges oxygen in heterocysts to prevent nitrogenase oxidation. However, *cydAB* genes are present in *Raphidiopsis brookii* D9, which neither forms heterocysts nor encodes for nitrogenases [Stucken et al. 2010].

3.3.5.4 Phototaxis genes

We also identified the *pixJ* gene in fourteen of the genomes, although it is notably absent from most of the AAD clade with the exception of three of the group 3 genomes. The *pixJ* gene is essential for type IV pili-directed positive phototaxis in *Synechocystis* PCC 6803, and in the *Nostocaceae* genomes, is commonly found near annotated putative chemotaxis homologs *cheWAY* genes, indicating their importance for phototaxis in these genomes [Schuergers et al. 2016; Campbell et al. 2015]. However, most strains in the AAD clade lack *pixJ*, suggesting they are either non-motile or are using a currently unannotated protein(s) for phototaxis.

3.3.5.5 Transporters

Several transporters are found throughout many of these genomes. The neutral amino acid complex genes *natCDE* and the manganese transporter genes *manRS* are nearly ubiquitous, while the vitamin B12-importer gene, *btuB*, is commonly found throughout these genomes as well [Picossi et al. 2005; Yamaguchi et al. 2002; Köster 2001]. Genes encoding the urea transporter complex (*urtABCDE*) [Beckers et al. 2004] are found in nearly all of the fifteen-member clade with the exception of *Anabaena* 90. Iron transport genes (ABC.FEV.AP) [Katoh et al. 2001] are less frequently found, although they are spread throughout the family. The presence of these transporters indicates that *Nostocaceae* strains can use external sources of amino acids, manganese, and vitamin B12, while fewer strains utilize iron uptake. This may indicate a form of mixotrophy in this group, similar to

how some marine picocyanobacteria are capable of taking up organic compounds to fuel growth [Zubkov et al. 2003].

3.3.5.6 Group-specific functional genes

Several genes associated with amino acid transport and retention were identified in group 1 and 5 genomes, as well as two *Nostoc* genomes and *Cylindrospermum* PCC 7417. One, found in 14 genomes including group 1 strains, is a gamma-glutamyltransferase (*ggt*) which increases non-polar amino acid solubility and may prevent the loss of non-polar amino acids via gamma-glutamylation [Suzuki et al. 2007; Baran et al. 2011]. Also, nine genomes, including group 1 strains, contain all or most components of the high-affinity branched-chain amino acid transport system (*livGHKM*). These genomes already contain the genes necessary to synthesize these amino acids, so the presence of uptake/retention genes suggests their requirements for these amino acids may be greater than their capacity for synthesis. Alternatively, this may be a way to shunt cellular resources towards other growth-related processes or could provide a mechanism of competitive exclusion against co-occurring bacteria.

Thirteen genomes encode the *tynA* gene, including all four genomes in group 2 (*Anabaena* MDT14, *Anabaena* 90, *Anabaena* AL09, and *Anabaena* LE011-02), as the only members of the AAD clade. The *tynA* gene encodes a primary-amine oxidase, which catalyzes oxidative deamination of aromatic amines to aldehydes [Elovaara et al. 2015]. Previous work in *E. coli* revealed *tynA* confers the ability to

grow in the presence of phenylethylamine [Elovaara et al. 2015]. Further analysis of environmental bacterial genomes suggests this gene is found more often when nutrients are less abundant, indicating *tynA* may encode an alternative metabolic enzyme when carbon or nitrogen availability is low [Elovaara et al. 2015]. A trade-off of this growth is that H_2O_2 is produced, and there is a net release outside of the cell [Kumar and Imlay 2013]. Cyanobacteria already undergo increased oxidative stress due to photosynthesis, and therefore require various strategies to mitigate reactive oxygen species [Paerl and Otten 2013]. As a result, this may increase oxidative load and require devoting more cellular resources to addressing this problem. Alternatively, increasing extracellular H_2O_2 could increase lethality for surrounding organisms, thereby reducing resource competition or predation and possibly promoting initiation or sustenance of high-density blooms [Jansen et al. 2002; Selva et al. 2009].

Nine of the genomes, contain *mocA*, a molybdenum cytidylyltransferase that is necessary for creating the molybdopterin cytosine dinucleotide cofactor (MOC) [Neumann et al. 2009]. Another set of genes found in eight of these nine genomes is *yagTRS*. These encode a xanthine dehydrogenase which requires MOC to function; this is consistent with the presence of *mocA* in many of the same genomes [Neumann et al. 2009]. Of the 15-member clade, only group 3 genomes (*Anabaena* WA93, *Anabaena* WA102, *Aphanizomenon* 2012/KM1/D3, and *Aphanizomenon* NIES-81) contain these genes, which may provide a distinguishing trait in comparison with the rest of the clade. These genes convert xanthine into urate to create NADH, possibly for reducing agent [Self 2002]. Previous studies indicate

that purines can act as the sole source of nitrogen or carbon in *Klebsiella pneumoniae* and the unicellular algae *Chlorella* [Ammann and Lynch 1964; Tyler 1978]. Alternatively, a similar pathway in *E. coli* was revealed to salvage purines rather than use them as a nitrogen source [Xi et al. 2000]. Group 3 genomes may then use purines as either another form of NADH synthesis, as a nitrogen source, or as recycled organic matter.

Group 4 genomes contain a type I-C CRISPR-Cas system encoding Cas5d, Csd1, and Csd2. A similar *Bacillus halodurans* Cas5d nuclease has been characterized, which revealed specific RNA nuclease activity [Punetha et al. 2014]. However, it also carries out a metal-dependent, non-specific DNase activity, hinting at a more generalized defense strategy. For example, promiscuous restriction mechanisms can increase bacterial fitness in the presence of phage or plasmid DNA through degradation without sequence specificity [Vasu et al. 2012]. Polyamines or proteins attached to cellular DNA may protect the host's genome, allowing for nucleation of newly introduced DNA. This generalized defense mechanism may then provide this group with a selective advantage for protection against parasitic plasmids or phages.

Eight of the thirty-one genomes contain *fruB*, a gene that is part of a fructose phosphotransferase system, which uses phosphoenolpyruvate to power fructose import [Geerse et al. 1989]. Notably, six of these eight genomes are *Anabaena* and *Nostoc* strains including the symbiont *Nostoc azollae* 0708, and none of the eight genomes belong to the AAD clade. Since some cyanobacteria have shown the ability to utilize external carbon sources [Anderson and McIntosh 1991], it's pos-

sible the presence of these fructose importers indicate an optional heterotrophic phenotype not seen in the AAD clade.

We also looked for unique genes found in all members of the AAD clade relative to all 31 genomes to identify signature genes for this group. However, no genes were ubiquitously and uniquely found in the AAD clade. Also, there were no annotated genes unique within genomes isolated from the same sampling sites, nor were there genes unique to *Aphanizomenon* genomes in comparison with *Anabaena* genomes that might explain morphological differences.

3.3.5.7 Buoyancy genes

Excluding genomes from symbiotic bacteria (*Nostoc azollae* 0708, *Nostoc punctiforme* PCC 73102, and all *Richelia*), the number of *gvpA* copies is highly variable (mean = 3.0, SD = 3.5) (Table 3.6). This may be due to artifacts of assembly, where short arrays of highly similar genes can lead to assembly errors [Brown et al. 2016]. The same is true for counts of all gas vesicle-related genes (mean = 6.4, SD = 5.2). Previous work in *Anabaena* and *Microcystis* cultures revealed that gene loss or rearrangements within the gas vesicle operon led to observable losses in buoyancy in culture [Wang et al. 2012; Mlouka et al. 2004]. To see if this happened with the genomes analyzed here, we extracted and aligned GvpG translations to compare with the truncated sequence from *Anabaena* sp. 90. While variable in length, there is no evidence suggesting any have lost their function. To investigate this further, we compared gas vesicle gene counts between genomes

from cultured (18 genomes) and uncultured (8 genomes) strains. The uncultured group on average contained 5.0 more gas vesicle-related genes than the cultured group (independent sample t-test; mean diff. = 5.0, p-value = 0.038). Additionally, we found that GvpC protein predicted molecular weight was variable across these genomes (15-27 kDa). Previous work has suggested that GvpC size is negatively correlated with ability for gas vesicles to withstand greater pressures, which may explain the variation seen here [Bright and Walsby 1999]. As a result, strains such as *Anabaena* CPCC64 may have more durable gas vesicles than strains such as *Anabaena* WA113.

3.3.5.8 Genes for ROS defense

Photosynthetic electron transport in cyanobacteria generates harmful reactive oxygen species (ROS) that phototrophic cyanobacteria must defend against [Latifi et al. 2009]. To assess patterns in the strategies used by the *Nostocaceae* family (including novel strains), we searched for genes associated with oxidative stress responses. We identified superoxide dismutase (SOD) genes in all genomes, and most also carry peroxiredoxin (Prx-s) genes (with the exception of the complete *Anabaena* sp. 90 and *Anabaena* WA102 genomes, which indicate true absences)(Table 3.7). SOD is essential for countering superoxide activity, while Prx-s reduces H_2O_2 and other ROS [Latifi et al. 2007]. Rubrerythrin-encoding genes are also commonly found in these genomes. This enzyme uses an electron from NADPH or NADH to convert H_2O_2 to water, and experiments in *Anabaena* PCC 7120 have demon-

strated its role as a peroxidase and protector of nitrogenase in heterocysts [Kurtz 2006; Zhao et al. 2007]. Other genes encoding peroxidase and catalase are also found in some of the genomes, although they are found less than previously described ROS-mitigating genes. Catalases only dismutate H_2O_2, while peroxidases can target different types of peroxides [Chelikani et al. 2004].

In addition to enzymatic oxidative stress relievers, nonenzymatic antioxidants can also reduce the burden of ROS in cyanobacteria [Latifi et al. 2009]. We identified genes necessary for synthesis of the carotenoids zeaxanthin and myxoxanthophyll in all of these genomes. Myxoxanthophyll protects against peroxidation, while mutants of *Synechococcus* PCC 7942 lacking zeaxanthin have been shown to be more susceptible to high light and oxidative stress [Schäfer et al. 2005]. Vitamin E synthesis genes are also present in most of these genomes. Vitamin E protects membrane lipids from peroxidation in plants, and previous work in *Synechocystis* PCC 6803 suggests a similar role in cyanobacteria as well [Havaux et al. 2005; Maeda et al. 2005]. We also identified the glutathione synthesis pathway in all genomes except *Richelia* HM01; glutathione is a nonribosomal peptide that contributes to oxidative stress resistance, as observed in *Synechocystis* PCC 6803 [Cameron and Pakrasi 2010].

3.3.5.9 Extracellular polymeric substance synthesis and export genes

Cyanobacteria can produce extracellular polymeric substances (EPS) which often associate with the outside of cells in the form of sheaths or capsules. They provide

a number of benefits, including protection against desiccation or UV damage, and can assist in maintaining an anoxic environment in heterocysts [Pereira et al. 2009; Kehr and Dittmann 2015]. Most of the genomes characterized here contain genes associated with the three characterized EPS pathways in cyanobacteria, which include the *wzy* and *bcsA* genes (hallmarks of the Wzy- and Synthase-dependent pathways, respectively) (Table 3.8) [Pereira et al. 2015]. Genes encoding extracellular polysaccharide biosynthesis proteins are nearly ubiquitously found in these genomes, while seven strains also contain putative capsular synthesis genes (Table 3.8). These putative genes may be involved in sheath/capsule/mucilage biosynthesis, and if so may indicate strains that contain this extracellular feature.

3.3.5.10 RNA genes

The *Nostocaceae* genomes contain from 1 to 12 rRNA operons, although some of the draft genomes have no copies of some of the rRNA genes (Table 3.9). Binning fragmented assemblies can exclude rRNA genes entirely, which may explain why some genomes contained no identified rRNA genes (See chapter 2). As a result, we are unable to determine true number of rRNA operons in these genomes. In addition, the *Nostocaeae* genomes contain to from 26 to 76 tRNA genes, which is highly variable, with larger genomes carrying more tRNAs (Table 3.10). The least variable tRNA genes were tRNA-His, tRNA-Cys, and tRNA-Trp (Std. Dev. = 0.18, 0.34, and 0.37, respectively), likely since they are commonly found in only 1-2 copies per genome. On the other hand, the most variable tRNA genes were

tRNA-Ala, tRNA-Leu, and tRNA-Ile (Std. Dev. = 1.70, 1.60, 1.33, respectively).

3.3.5.11 IS elements

We identified the number of insertion sequence (IS) elements found throughout the genomes using the TnPred database. The distribution of total insertion sequences is variable across all thirty-one genomes, with a range of 0-251 per genome (Table 3.11). Genomes lacking IS elements are the symbiotic *Richelia* HH01 and HM01 genomes, which are highly reduced. In contrast, the largest number of IS sequences reside in *Richelia* RC01, which has a much larger genome than strains HH01 and HM01, and which presumably has not undergone as much reductive evolution. The distribution of characterized IS sequences is varied across these genomes, including those isolated from the same sites (e.g., *Aphanizomenon* MDT13 and *Aphanizomenon* MDT 14). This, in addition to the majority of IS sequences being uncharacterized, suggests there is a large diversity of these mobile genes within and between environments. However, it is important to keep in mind that these sequences may be underestimated in draft genomes due to assembly breaks at repetitive elements.

3.4 Discussion

Here, we have characterized nine novel *Anabaena* and *Aphanizomenon* genomes relative to the rest of the *Nostocaceae* family. We focused on identifying novel

toxin or taste-and-odor compound synthesizing genes, assessing consistency between phylogenomic signal and morphological characterization, and searching for differences in functional gene content.

3.4.1 Phylogenomics reveals morphology-phylogeny inconsistencies

Previous reports based on single-copy gene phylogenies of filamentous cyanobacteria have revealed inconsistencies in the placements of the *Anabaena*, *Aphanizomenon*, and *Nostoc* genera [Gugger et al. 2002; Rajaniemi et al. 2005; Shih et al. 2013]. Supporting these reports, these genomes again do not cleanly separate into distinct clusters by genus in our phylogenomic tree, and placement of the novel genomes presented here introduces new inconsistencies in phylogenomic and morphological classification (Figure 3.1). All of these results conflict with taxonomic assignment of filamentous cyanobacteria using polyphasic approaches that often weigh heavily towards subjective colony or cell morphology characterizations [Zapomělová et al. 2009; 2012; Komárek 2016]. This suggests that current genus-level assignments may need to be reconsidered, especially those characterized primarily by morphology. While polyphasic taxonomic assignments may include relevant phenotypic information, these descriptors do not adequately reflect evolutionary relationships [Stanier and Niel 1962; Sapp 2005]. As a result, morphological classifications should be considered uncertain at best and should routinely require genetic information to verify. We believe our phylogenomic tree better indicates evolutionary relationships within this family because it evaluates the similarity

of 279 single-copy orthologues found across all genomes presented here. In-depth assessment of the trends in prokaryotic gene evolution have revealed patterns of dominant vertical inheritance with low amounts of HGT for widely-shared genes [Puigbo et al. 2010], and since our tree is built from family-wide shared genes, we believe this supports the branch assignments. Some patterns in phylogenomic relationships emerge from these analyses. Notably, *Nostoc* genomes are never found in the AAD clade, suggesting a clear distinction between *Nostoc* and many *Anabaena/Aphanizomenon/Dolichospermum*. Additionally, the *Anabaena* genomes that cluster with *Nostoc* strains may be more *Nostoc*-like, suggesting that reclassification of these strains to distinguish between *Nostoc*-like *Anabaena* and potentially toxigenic AAD-like *Anabaena* may be prudent. It is important to note that of our genome set, none are part of the benthic *Anabaena*'s [Surakka et al. 2005], revealing the potential for future work to expand the phylogenomics presented here.

3.4.2 Distribution of toxic/secondary metabolite synthesis genes

Toxicity of strains in this family is a relevant phenotype that must be addressed. From our analysis, none of the novel genomes contain putative toxin synthesis genes, although we identified geosmin synthesis genes in *Anabaena* CRKS33. Overall, six of the thirty-one strains produce toxic compounds (identified by toxin measurements previously), and there is a broad range of synthesized toxin types for those that do. In addition, closely related strains sometimes contain toxic and non-

toxic members, which has also been previously reported [D'Agostino et al. 2016a]. This supports the possibility that toxins are not consistently vertically inherited and are often lost in certain lineages, similar to discussions by others that toxic gene clusters are horizontally transferred, although specific instances have yet to be identified [Stucken et al. 2010; Jiang et al. 2012]. The inconsistent distribution of toxin synthesis genes, especially within the AAD clade, supports this possibility. Toxic strains may gain these genes through lateral transfer events, and retain them due to some selective advantage, but analysis of more closely-related genome pairs with and without toxin synthesis genes is necessary to address these possibilities.

Although common, the scattered presence of bacteriocins in these genomes indicates they are not conserved. Aharonovich et al. recently identified bacteriocin genes that were upregulated in *Prochlorococcus* when co-cultured with a marine heterotroph, indicating putative utility for controlling co-occurring bacterial population growth [Aharonovich and Sher 2016]. Carrying bacteriocin genes could provide an advantage to bloom-forming cyanobacteria by inhibiting competitors for nutrient acquisition. However, their lack of conservation across these genomes from similar geographic origin may indicate any advantage from retaining these genes is not environment-specific.

Of the eleven genomes which carry cyanobactin synthesis clusters, nine are part of the AAD clade. Previous work has shown that *Anabaena* strains produce a range of diverse cyanobactins, indicating the potential for biomining these genomes to look for valuable new compounds [Leikoski et al. 2010]. Further analysis indicates that six of these clusters are anacyclamide synthesis genes, a diverse group of

cyanobactins [Sivonen et al. 2010]. However, the bioactivity of anacyclamides are still unknown.

3.4.3 Functional gene content comparisons

Freshwater systems are often distinct from each other due to physical separation with no direct linkage. As a result, the environmental parameters from different systems can be drastically different. Most of the genomes analyzed here come from geographically disparate freshwater systems (Table 3.1). By comparing functional gene content of these genomes, we revealed variation in presence of genes associated with multiple pathways. These included genes encoding auxiliary photosystem components, pigments, sulfur metabolism enzymes, transporter, and phototaxis proteins. Differential environmental parameters from each respective isolation site may mediate selective pressures that drive either retention or loss of these genes, and could indicate their persistence in some genomes after acquisition through horizontal gene transfer.

In addition, several of the groups within the AAD clade carry genes unique from the rest of the clade, suggesting physiological differences that may provide specific advantages in a range of environments. These unique genes may then indicate strategies by which certain groups retain an advantage relative to co-occurring bacteria in their respective environments. For example, group 1 genomes contain genes associated with amino acid uptake and retention, which may benefit these strains by allowing resources to shift from amino acid synthesis to other

growth-related processes. Additionally, other group-specific genes are associated with alternative metabolic pathways for utilizing purines or phenylethylamine, and a non-specific targeting CRISPR system. Overall, this suggests that these groups may use functionally diverse strategies to obtain niche-specific competitive advantages across environments.

3.5 Conclusions

Here, we have sequenced nine novel *Anabaena* and *Aphanizomenon* genomes, and compared them with all sequenced genomes from the *Nostocaceae* family. Phylogenomic analyses of these strains indicates that eight of the nine novel genomes belong to a single, newly expanded clade, adding to the availability of sequenced genomes from this group. Consequently, fifteen of the thirty-one *Nostocaceae* genomes belong to this clade, which consists entirely of planktonic bloom-forming strains of *Anabaena*, *Aphanizomenon*, and *Dolichospermum*. Within this fifteen-member clade are four distinct subclades consisting of highly similar genomes (>95% nucleotide identity over >60% of their genome). These genomes consist of mixed genera previously classified by morphology, and indicates the utility in acquiring genomic information as a cautious step towards validating taxonomic assignments.

There are no clear patterns of toxin gene presence throughout the *Nostocaceae* family, indicating the possibility that these genes are transferred horizontally. Additionally, we identified genes unique to genomes from each group relative to the AAD clade. These genes varied widely in function and included amino acid

transport/retention, utilization of alternative nitrogen sources, and a DNA/RNA-targeting CRISPR system. Genes for organic sulfur uptake are variable across these genomes. As a result, variation in these pathways suggests some *Nostocaceae* strains utilize multiple strategies for acquisition of sulfur and nitrogen. These genomic comparisons can serve as a guideline for future classifications of bloom-forming, filamentous cyanobacteria, in addition to informing about the important and unique genomic characters that may help to better understand these potentially toxigenic cyanobacteria.

Genome	Isolation source	Cultured?	Genome Size (Mbp)	Accession Number	Contig number (including plasmids)	GC%	No. CDS	No. Pseudogenes	CRISPR arrays	Ref.
Anabaena 90	Lake Vesijrvi, Finland (1986)	Y	5.3	NC_019427	5	38.1	4444	163	2	[Wang et al. 2012]
Anabaena CRKS33*	Cheney Reservoir, KS, USA (August 30, 2013)	N	4.95	LJOT00000000	1109	37.6	4638	224	3	This work
Anabaena cylindrica PCC 7122	Pond in Cambridge, UK (1939)	Y	7.06	NC_019771	7	38.79	5775	212	13	None
Anabaena flos-aquae CPCC64*	Lake Ontario, Western Basin (June 7, 2009)	Y	6.91	LJOR00000000	80	41.33	5474	146	8	This work
Anabaena lemmermannii AL09*	Lake Ontario, Western Basin (August 1, 2005)	Y	4.65	LJOQ00000000	109	38.1	3988	307	0	This work
Anabaena lemmermannii LE011-02*	Lake Erie (July 12, 2011)	Y	4.74	LJOP00000000	122	38.06	4072	257	2	This work
Anabaena MDT14*	MDT site, Upper Klamath Lake, USA (June 4, 2014)	N	4.95	LJOV00000000	1227	38.9	4546	329	2	This work
Anabaena PCC 7108	Moss Beach, CA, USA (1970)	Y	5.88	NZ_KB235895	3	38.77	4875	122	8	[Shih et al. 2013]
Anabaena variabilis ATCC 29413	Mississippi, USA (1964)	Y	7.1	NC_007413	5	41.41	5721	49	8	[Thiel et al. 2014]
Anabaena WA102	Anderson Lake, WA, USA (May 20, 2013)	Y	5.78	NZ_CP011456	2	38.38	4880	223	4	[Brown et al. 2016]
Anabaena WA113*	Cranberry Lake, WA, USA (August 11, 2014)	N	4.69	LJOS00000000	279	37.22	4002	231	3	This work
Anabaena WA93	American Lake, WA, USA (1993)	Y	5.66	LJOU00000000	217	38.36	4693	355	5	This work
Aphanizomenon flos-aquae 2012/KM1/D3	Curonian Lagoon, Baltic Sea (2012)	Y	5.74	NZ_JSDP01000254	325	38.22	4601	836	7	[Šulčius et al. 2015]
Aphanizomenon flos-aquae MDT13 culture*	MDT site, Upper Klamath Lake, USA (August, 2013)	Y	4.43	LJOY00000000	307	37.05	3787	191	3	This work
Aphanizomenon flos-aquae MDT14*	MDT site, Upper Klamath Lake, USA (June 4, 2014)	N	4.63	LJOX00000000	193	37.11	3936	211	4	This work
Aphanizomenon flos-aquae NIES-81	Lake Kasumigaura, JP (1978)	Y	5.85	NZ_KI928192	103	37.37	4744	325	14	[Cao et al. 2014]
Aphanizomenon flos-aquae WA102-2*	Anderson Lake, WA, USA (May 20, 2013)	N	5.94	LJOW00000000	1160	39.12	5296	405	3	This work
Cylindrospermopsis raciborskii CS-505	Solomon Dam, Australia (1996)	Y	3.87	NZ_ACYA01000093	93	40.23	3176	173	13	[Stucken et al. 2010]
Cylindrospermum stagnale PCC 7417	Soil from greenhouse in Stockholm, Sweden (1972)	Y	7.61	NC_019757	4	42.2	6118	183	10	None
Dolichospermum circinale AWQC131C	Lake Cargelligo, NSW, Australia (Date unknown)	Y	4.44	NZ_KE384588	121	37.01	3750	268	1	[D'Agostino et al. 2016a]
Dolichospermum circinale AWQC310F	Farm Dam, Millawa, VIC, Australia (1995)	Y	4.4	NZ_KE384663	82	37.33	3676	269	3	[D'Agostino et al. 2016a]
Nodularia spumigena CCY 9414	Baltic Sea, near Bornholm island (Date unknown)	Y	5.46	NZ_CP007203	1	41.19	4510	148	2	None
Nostoc azollae 0708	Water fern, unknown location	Y	5.48	NC_014248	3	38.37	3985	1158	0	[Ran et al. 2010]
Nostoc PCC 7107	Pond in Point Reyes Peninsula, CA, USA (1970)	Y	6.32	NC_019676	1	40.36	5200	117	13	None
Nostoc PCC 7120	Unknown	Y	7.21	NC_003272	7	41.27	5823	143	11	[Ohmori et al. 2001]
Nostoc PCC 7524	Hot spring from Amparai District, Maha Oya, Sri Lanka (1973)	Y	6.71	NC_019684	3	41.53	5326	105	6	None
Nostoc punctiforme PCC 73102	Macrozamia sp. root section, Australia (1973)	Y	9.05	NC_010628	6	41.35	6966	388	6	[Meeks et al. 2001]
Raphidiopsis brookii D9	Billings Reservoir, Brazil (1996)	Y	3.18	NZ_ACYB01000047	47	40.06	2602	276	5	[Stucken et al. 2010]
Richelia intracellularis HH01	Western Gulf of Mexico (Date unknown)	Y	3.24	NZ_CAIY01000090	90	33.71	1872	97	0	[Hilton et al. 2013]
Richelia intracellularis HM01	Western Gulf of Mexico (Date unknown)	Y	2.21	NZ_CAIS01000941	941	33.76	1119	592	0	[Hilton et al. 2013]
Richelia intracellularis RC01	Unknown	Y	5.48	NZ_CBZS010000001	857	39.16	4330	1254	0	[Hilton 2014] (Thesis)

Table 3.1: Genome information. *'s denote novel genomes

Genome	Number of contigs	N50	Total No. Universal Genes	No. Unique Universal Genes
Anabaena AL09	109	64,976	109	101
Anabaena CPCC64	80	134,379	118	105
Anabaena CRKS33	1109	13,210	113	104
Anabaena LE011-02	122	64,747	112	103
Anabaena MDT14	1227	7,760	113	101
Anabaena WA113	279	71,945	115	105
Aphanizomenon flos-aquae MDT13 culture	307	55,427	115	105
Aphanizomenon flos-aquae MDT14	193	60,437	117	106
Aphanizomenon WA102-2	1160	15,892	115	105

Table 3.2: Novel genomes information

Genome	Estimated Completeness (%)	Estimated Contamination (%)	Taxon
Anabaena 90*	99.67	0	Cyanobacteria
***Anabaena* AL09**	98.11	0	Cyanobacteria
***Anabaena* CPCC64**	99.33	0	Cyanobacteria
***Anabaena* CRKS33**	99.44	1.78	Cyanobacteria
Anabaena cylindrica PCC 7122*	99.44	0	Cyanobacteria
***Anabaena* LE011-02**	99.22	0.11	Cyanobacteria
***Anabaena* MDT14**	97.17	4.22	Cyanobacteria
Anabaena PCC 7108	99.63	0.3	Cyanobacteria
Anabaena variabilis ATCC 29413*	99.33	0	Cyanobacteria
Anabaena WA102*	99.78	0.22	Cyanobacteria
***Anabaena* WA113**	99.89	0.44	Cyanobacteria
Anabaena WA93	99.67	0.52	Cyanobacteria
Aphanizomenon flos-aquae 2012/KM1/D3	87.52	7.22	Cyanobacteria
***Aphanizomenon flos-aquae* MDT13 culture**	99.67	0.37	Cyanobacteria
***Aphanizomenon flos-aquae* MDT14**	99	1	Cyanobacteria
Aphanizomenon flos-aquae NIES-81	99.67	0.56	Cyanobacteria
***Aphanizomenon flos-aquae* WA102**	99.89	3.6	Cyanobacteria
Cylindrospermopsis raciborskii CS-505	99.85	0	Cyanobacteria
Cylindrospermum stagnale PCC 7417*	99.78	0.68	Cyanobacteria
Dolichospermum circinale AWQC131C	99.56	0	Cyanobacteria
Dolichospermum circinale AWQC310F	99.56	0	Cyanobacteria
Nodularia spumigena CCY 9414	99.78	0.67	Cyanobacteria
Nostoc azollae 0708*	98.89	0	Cyanobacteria
Nostoc PCC 7107*	99.28	0.36	Cyanobacteria
Nostoc PCC 7120*	99.19	0	Cyanobacteria
Nostoc PCC 7524*	99.28	0	Cyanobacteria
Nostoc punctiforme PCC 73102*	99.56	0.22	Cyanobacteria
Raphidiopsis brookii D9	99.37	0	Cyanobacteria
Richelia intracellularis RC01	94.34	1.85	Cyanobacteria
Richelia intracellularis HH01	93.44	0.11	Cyanobacteria
Richelia intracellularis HM01	64.75	0.56	Cyanobacteria

Table 3.3: CheckM results on binned genomes. Bolded genomes are novel genomes presented in this study. *'s denote genomes that are finished-quality, while the remainder are draft-quality.

Genome 1	*Anabaena* 90	*Anabaena* AL09	*Anabaena* CPCC64	*Anabaena* CRKS33	*Anabaena* LE011-02	*Anabaena* MDT14	*Anabaena variabilis* ATCC 29413	*Anabaena* WA102	*Anabaena* WA113	*Anabaena* WA93	*Aphanizomenon flos-aquae* 2012/KM1/D3	*Aphanizomenon flos-aquae* MDT13	*Aphanizomenon flos-aquae* MDT14	*Aphanizomenon flos-aquae* NIES-81	*Aphanizomenon flos-aquae* WA102-2	*Dolichospermum circinale* AWQC131C	*Dolichospermum circinale* AWQC310F
***Anabaena* 90**	100/1.0	97.17/0.69	75.65/0.52	87.19/0.61	97.15/0.71	97.28/0.65	75.7/0.52	91.99/0.7	88.83/0.63	91.96/0.69	91.6/0.62	88.93/0.62	88.95/0.62	91.11/0.68	88.85/0.63	87/0.62	87.06/0.61
***Anabaena* AL09**	97.18/0.79	100/1.0	75.72/0.58	87.24/0.68	98.47/0.85	97.46/0.73	75.73/0.58	91.7/0.76	88.99/0.7	91.74/0.75	91.33/0.67	88.99/0.7	88.98/0.69	91.11/0.74	89.01/0.69	87.05/0.67	87.16/0.68
***Anabaena* CPCC64**	75.64/0.4	75.75/0.39	100/1.0	75.42/0.37	75.69/0.4	75.77/0.37	100/0.99	75.61/0.41	75.27/0.37	75.58/0.41	75.7/0.37	75.29/0.37	75.28/0.37	75.53/0.42	75.3/0.37	75.48/0.36	75.5/0.37
***Anabaena* CRKS33**	87.19/0.67	87.24/0.65	75.42/0.53	100/1.0	87.26/0.67	87.17/0.62	75.43/0.53	86.87/0.7	88.17/0.67	86.81/0.69	86.84/0.62	88.19/0.67	88.18/0.67	86.68/0.69	88.1/0.66	95.98/0.75	95.9/0.73
***Anabaena* LE011-02**	97.37/0.72	97.44/0.7	75.78/0.52	87.17/0.62	100/1.0	97.54/0.72	75.78/0.53	91.85/0.7	88.67/0.62	91.72/0.68	91.4/0.62	88.68/0.62	88.62/0.62	91.06/0.67	88.72/0.62	86.88/0.61	86.96/0.62
***Anabaena* MDT14**	81.37/0.54	81.5/0.53	76.52/0.56	80.72/0.49	97.54/0.71	100/1.0	76.6/0.56	81.19/0.55	80.7/0.5	81.15/0.54	81.21/0.49	80.67/0.5	80.61/0.5	80.99/0.54	80.6/0.5	80.6/0.49	80.6/0.49
***Anabaena variabilis* ATCC 29413**	75.8/0.39	75.78/0.38	100/0.97	75.43/0.36	75.71/0.38	75.76/0.36	100/1.0	75.75/0.4	75.3/0.37	75.6/0.4	75.86/0.36	75.47/0.36	75.44/0.36	75.68/0.41	75.35/0.36	75.63/0.35	75.64/0.36
***Anabaena* WA102**	91.97/0.66	91.71/0.63	75.62/0.51	86.87/0.6	91.72/0.64	91.85/0.60	75.66/0.51	100/1.0	88.39/0.62	98.91/0.86	97.33/0.65	88.49/0.62	88.51/0.61	96.74/0.73	88.59/0.61	86.56/0.59	86.56/0.61
***Anabaena* WA113**	88.85/0.71	89.01/0.69	75.32/0.55	88.19/0.69	89.04/0.7	88.68/0.64	75.33/0.55	88.43/0.73	100/1.0	88.42/0.73	88.25/0.66	98.42/0.83	98.45/0.82	88.15/0.72	99.26/0.89	87.9/0.69	87.98/0.69
***Anabaena* WA93**	91.93/0.68	91.74/0.65	75.61/0.52	86.81/0.62	91.63/0.66	91.7/0.61	75.62/0.52	98.9/0.89	88.42/0.64	100/1.0	97.39/0.68	88.47/0.64	88.49/0.63	96.86/0.74	88.41/0.63	86.59/0.61	86.65/0.62
***Aphanizomenon flos-aquae* 2012/KM1/D3**	91.58/0.66	91.33/0.62	75.72/0.51	86.89/0.61	91.4/0.63	91.37/0.56	75.76/0.51	97.32/0.73	88.32/0.61	97.37/0.73	100/1.0	88.38/0.61	88.4/0.61	97.03/0.74	88.26/0.61	86.58/0.59	86.56/0.6
***Aphanizomenon flos-aquae* MDT13 culture**	88.94/0.74	88.99/0.72	75.32/0.58	88.18/0.74	89.03/0.73	88.64/0.67	75.38/0.58	88.49/0.77	98.42/0.88	88.48/0.77	88.34/0.69	100/1.0	99.29/0.92	88.15/0.75	98.45/0.85	87.95/0.72	88.02/0.73
***Aphanizomenon flos-aquae* MDT14**	88.96/0.71	88.97/0.7	75.31/0.56	88.18/0.7	88.99/0.71	88.59/0.64	75.37/0.56	88.49/0.74	98.44/0.84	88.5/0.74	88.37/0.66	99.29/0.89	100/1.0	88.13/0.72	98.44/0.85	87.94/0.69	87.98/0.7
***Aphanizomenon flos-aquae* NIES-81**	91.08/0.67	91.11/0.64	75.54/0.53	86.65/0.62	91.07/0.64	91.06/0.6	75.59/0.53	96.76/0.75	88.1/0.62	96.86/0.74	97.05/0.68	88.16/0.61	88.17/0.61	100/1.0	88.12/0.62	86.36/0.61	86.4/0.61
***Aphanizomenon flos-aquae* WA102**	88.91/0.59	89/0.57	75.34/0.45	88.11/0.57	89.06/0.58	88.72/0.53	75.36/0.45	88.51/0.6	99.26/0.74	88.41/0.6	88.2/0.54	98.44/0.67	98.46/0.69	88.16/0.59	100/1.0	87.89/0.57	87.87/0.57
***Dolichospermum circinale* AWQC131C**	86.98/0.74	87.06/0.7	75.5/0.56	95.97/0.82	87.05/0.71	86.89/0.67	75.54/0.56	86.58/0.75	87.9/0.73	86.6/0.74	86.55/0.67	87.95/0.73	87.95/0.72	86.37/0.74	87.89/0.72	100/1.0	97.27/0.82
***Dolichospermum circinale* AWQC310F**	87.06/0.74	87.17/0.72	75.51/0.58	95.9/0.82	87.09/0.74	86.96/0.68	75.55/0.58	86.64/0.77	87.97/0.74	86.67/0.77	86.53/0.68	88.02/0.74	87.99/0.73	86.42/0.76	87.87/0.73	97.27/0.84	100/1.0

Table 3.4: Genomic average nucleotide identity (gANI) and alignment fraction (AF) values for each pairwise genome comparison. Values above the gray divider are calculated by aligning Genome 1 (row name) to Genome 2 (column name). Values below the gray divider are from aligning Genome 2 to Genome 1.

Genome	PKS	NRPS	Terpene	Bacteriocin	Microviridin	Lantipeptide	Proteusin	Ladderane	Siderophore	Cyanobactin	Resorcinol	Lassopeptide	Thiopeptide	Arylpolyene	Oligosaccharide	PUFA	Other
Richelia intracellularis RC01	2		1														
Richelia intracellularis HH01	1		1														
Richelia intracellularis HM01	1		1													1	
Raphidiopsis brookii D9	1	1	3											1	1		1
Cylindrospermopsis raciborskii CS-505	3	1	3											1			
Nostoc azollae 0708	2	1	1				1										
Anabaena cylindrica PCC 7122	4	5	4	4							1						
Anabaena PCC 7108	5	4	3	1													1
Anabaena CRKS33	2	1	4	1						1							1
Dolichospermum circinale AWQC131C	3	1	4				1			1							3
Dolichospermum circinale AWQC310F	2	2	4				1			2							3
Anabaena AL09	2	3	3							2							2
Anabaena LE011-02	2	4	3	1						1							
Anabaena MDT14	4	4	4	1						2		1					1
Anabaena 90	3	5	3	1						1							
Anabaena WA93	4	2	3							1							1
Anabaena WA102	3	5	3	1						1							2
Aphanizomenon flos-aquae 2012/KM1/D3	2	3	3														
Aphanizomenon flos-aquae NIES-81	3	2	4		1												2
Aphanizomenon flos-aquae MDT14	3	3	3	1													1
Aphanizomenon flos-aquae MDT13 culture	2	2	3														
Anabaena WA113	3	2	3														
Aphanizomenon WA102	2	2	3														
Cylindrospermum stagnale PCC 7417	8	9	6	1		2		1		1		1	1				1
Nostoc punctiforme PCC 73102	10	7	4	8		5											2
Nodularia spumigena CCY 9414	5	2	1	1	1					2							
Anabaena CPCC64	4	6	3	2		1	1	1	1								1
Anabaena variabilis ATCC 29413	6	6	3	2		1	1	1	1								1
Nostoc PCC 7120	4	3	4	2	1	1		1	1								
Nostoc PCC 7524	4	1	5	2		2						1					1
Nostoc PCC 7107	2	1	4	3								1					1

Table 3.5: Number of secondary metabolite gene clusters identified in each genome by category

Genome	No. GvpA	No. GvpC	GvpC length (AA)	GvpC estimated MW (Da)	Isolation site max depth (meters)	Vesicle Genes
Richelia intracellularis RC01	0	0			Unknown	0
Richelia intracellularis HH01	0	0			Unknown	0
Richelia intracellularis HM01	0	0			5267	1
Raphidiopsis brookii D9	1	1	220	24844.51	19	3
Cylindrospermopsis raciborskii CS-505	2	1	220	24985.54	13.4	3
Nostoc azollae 0708	0	0			Unknown	1
Anabaena cylindrica PCC 7122	0	0			Unknown	0
Anabaena PCC 7108	0	0			Unknown	1
Anabaena CRKS33	1	1	193	22084.74	12.8	6
Dolichospermum circinale AWQC131C	0	1	226	25847.76	6	4
Dolichospermum circinale AWQC310F	1	1	226	25955.71	Unknown	5
Anabaena AL09	0	0			244.1	3
Anabaena LE011-02	1	1	211	24087.75	64	7
Anabaena MDT14	2	1	193	22028.54	15.2	7
Anabaena 90	7	1	193	21984.52	6	7
Anabaena WA93	7	1	193	22042.56	27	12
Anabaena WA102	7	1	194	22042.56	7.6	17
Aphanizomenon flos-aquae 2012/KM1/D3	3	1	127	14616.73	5	4
Aphanizomenon flos-aquae NIES-81	4	1	197	22484.95	7	5
Aphanizomenon flos-aquae MDT14	12	1	193	22187.8	15.2	17
Aphanizomenon flos-aquae MDT13 culture	4	1	193	22187.8	15.2	10
Anabaena WA113	11	1	193	22217.77	11.6	17
Aphanizomenon WA102	8	1	193	22217.77	7.6	14
Cylindrospermum stagnale PCC 7417	0	0			None (soil, greenhouse)	1
Nostoc punctiforme PCC 73102	1	1	235	27339.55	None (symbiotic with cycad)	6
Nodularia spumigena CCY 9414	0	0			458.7	4
Anabaena CPCC64	2	1	129	15282.4	244.1	8
Anabaena variabilis ATCC 29413	2	1	129	15282.4	Unknown	5
Nostoc PCC 7120	0	0			Unknown	0
Nostoc PCC 7524	0	0			Unknown	1
Nostoc PCC 7107	2	1	129	15218.12	Unknown	4

Table 3.6: Buoyancy genes

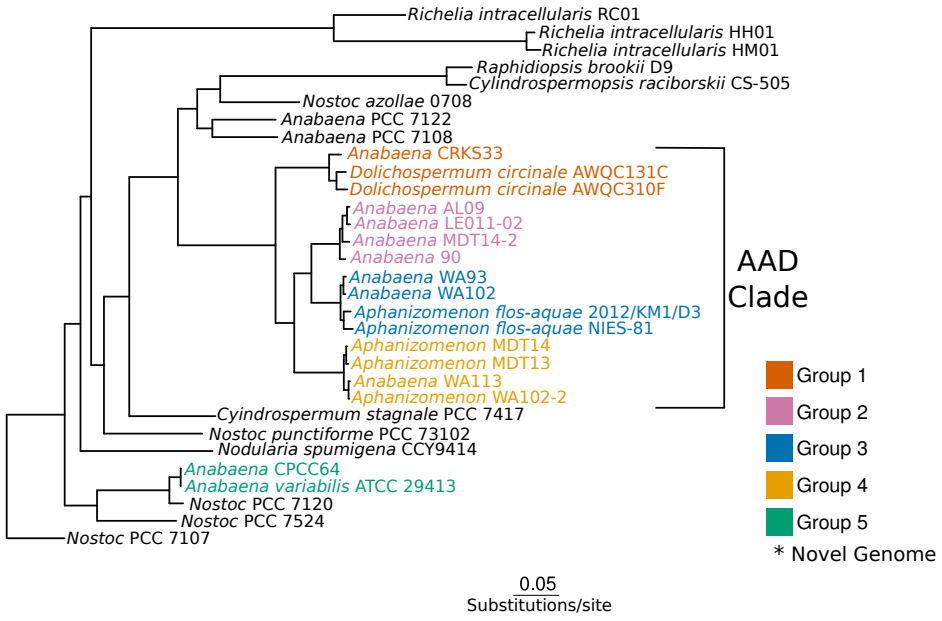

Figure 3.1: Phylogenomic tree of Nostocaceae clade. The tree was built using the HAL pipeline, which uses a concatenated alignment of all single-copy orthologues that are found in all genomes. Genome names are colored based on groupings, which are specified by genomic ANI (gANI) >95% and the aligned genome fraction (AF) with a 0.6 minimum cutoff. Genomes new to this study are highlighted with an asterisk.

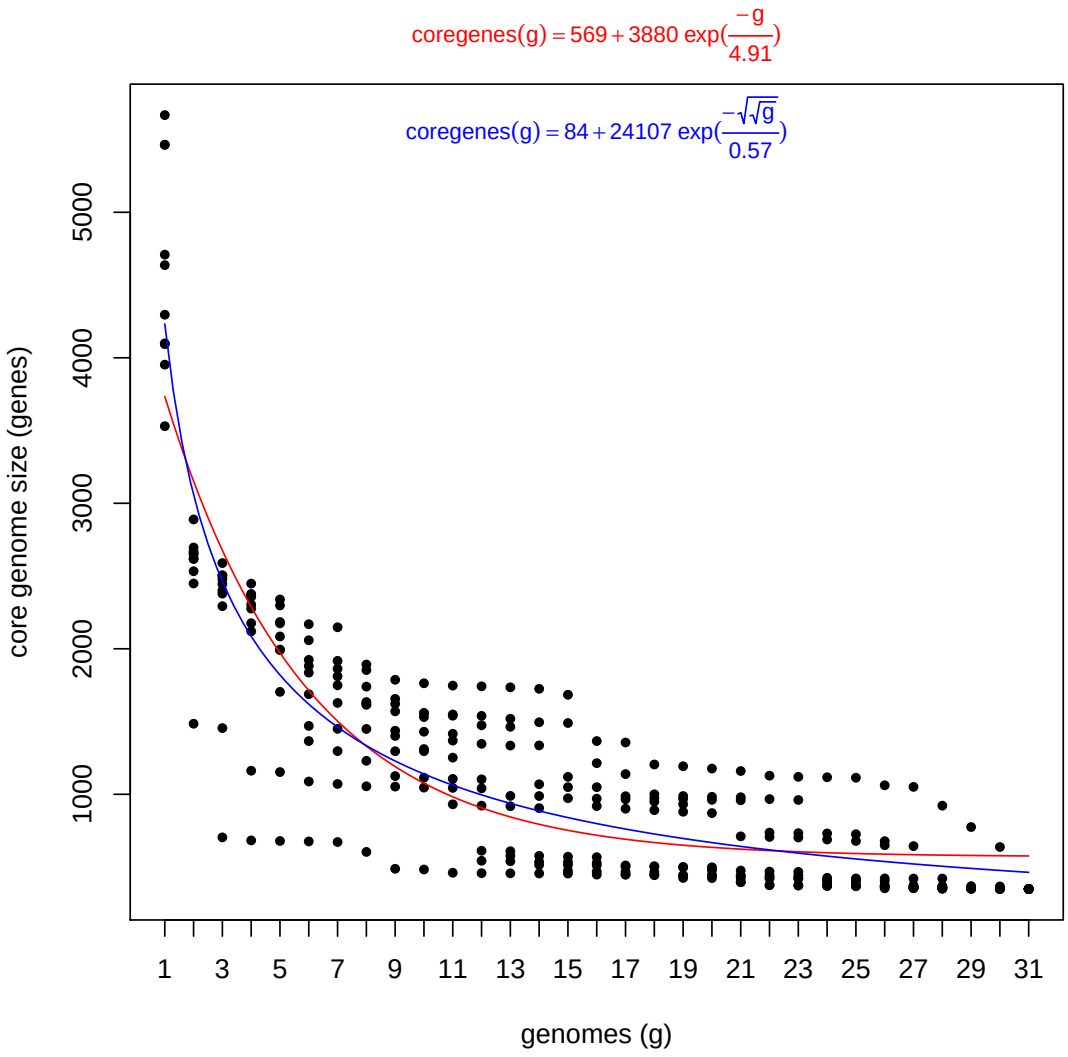

Figure 3.2: The core genome curve from the thirty-one *Nostocaceae* genomes determined by the OrthoMCL algorithm. The red line is the Tettelin exponential decay model estimate, while the blue line is the Willenbrock exponential decay model estimate. Number of genomes sampled are on the x-axis, while the number of genes included in the core genome are on the y-axis. Dots represent single iterations of core genome calculation.

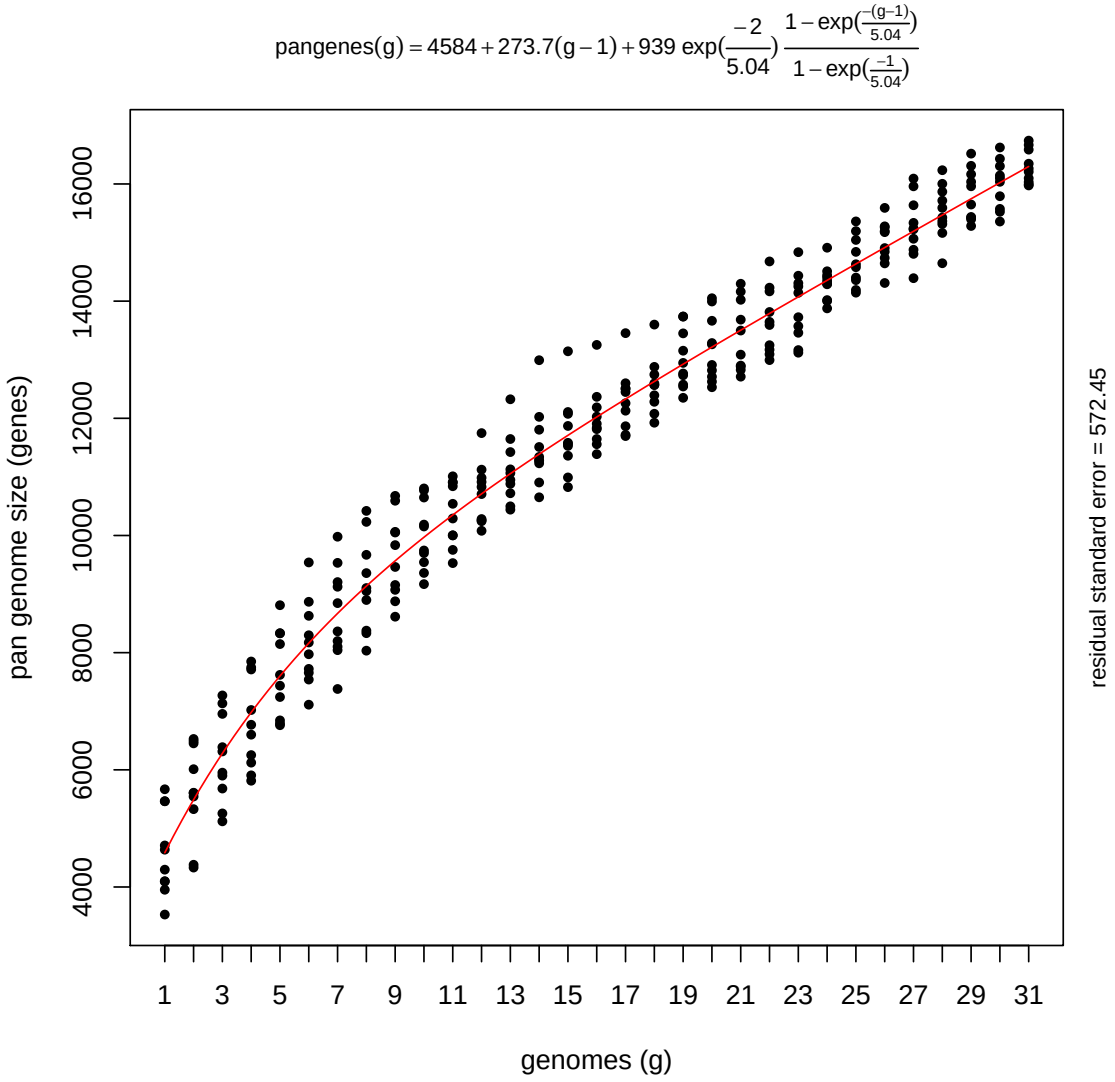

Figure 3.3: The flexible genome curve from the thirty-one *Nostocaceae* genomes determined by the OrthoMCL algorithm. Number of genomes sampled are on the x-axis, while the number of genes included in the core genome are on the y-axis. Dots represent single iterations of core genome calculation.

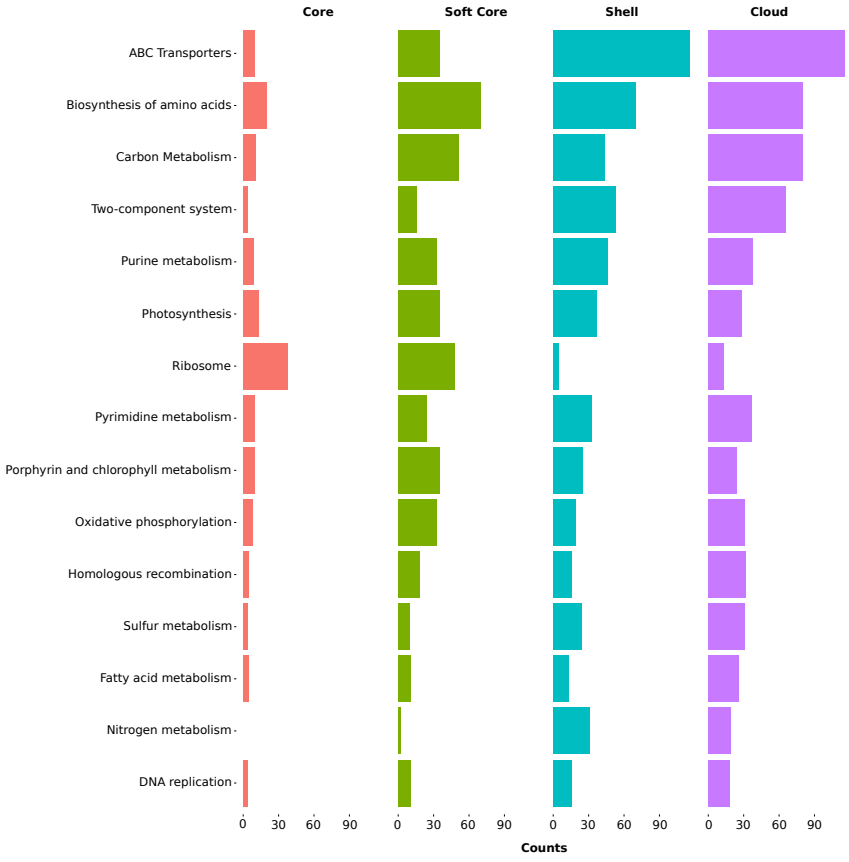

Figure 3.4: Counts of gene clusters associated with KEGG categories in the core (present in all genomes), soft core (core genes + genes absent in one genome), shell (genes in 3-18 genomes), and cloud (genes in 1-2 genomes) genomes.

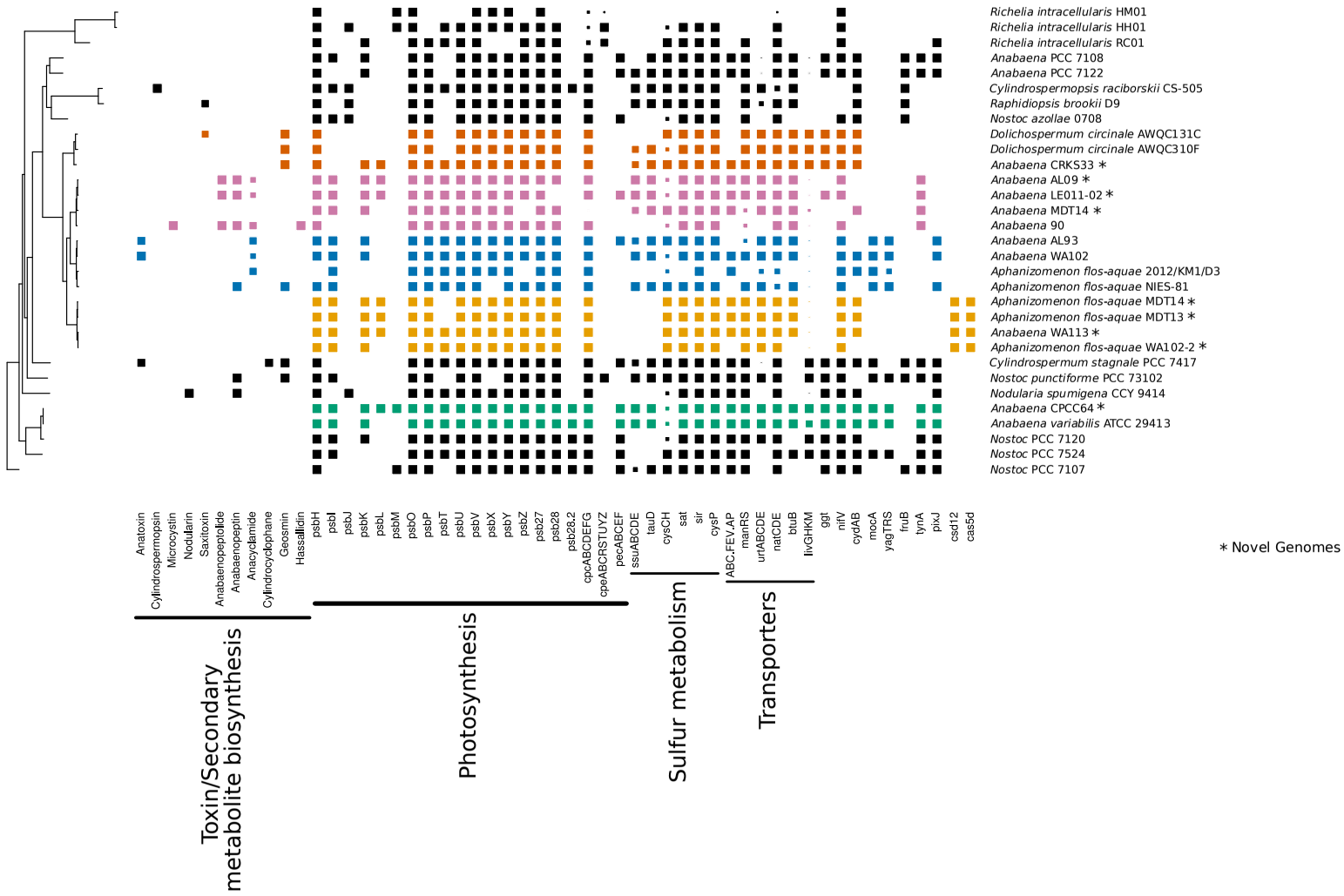

Figure 3.5: Functional gene content of *Nostocaceae* genomes. The tree on the left is the phylogenomic tree from Figure 3.1. Boxes inform regarding the presence/absence of functional genes. Additionally, the size of the boxes correlates with number of genes from that pathway that are present (smaller the box, less complete the gene set). Boxes are colored by the groups designated in Figure 3.1. Gene categories are grouped and labeled on the x-axis.

Genome	SOD	Peroxiredoxin	Rubrerythrin	Glutathione peroxidase	Cytochrome c peroxidase	Catalase	Glutathione?	Vitamin E (ubiquinone biosynthesis)?	Vitamin A (retinol)?	Vitamin C (ascorbic acid)?	Carotenoids (Zeaxanthin, Myxols)
Richelia intracellularis RC01	2	1					Y	N	N	N	Y
Richelia intracellularis HH01	2	1					N	N	N	N	Y
Richelia intracellularis HM01	1	3					Y	Y	N	N	Y
Raphidiopsis brookii D9	1	2					Y	Y	N	N	Y
Cylindrospermopsis raciborskii CS-505	1	3	1				Y	Y	N	N	Y
Nostoc azollae 0708	3	7				1	Y	Y	N	N	Y
Anabaena cylindrica PCC 7122	2	1			1		Y	Y	N	N	Y
Anabaena PCC 7108	2	2	1			1	Y	Y	N	N	Y
Anabaena CRKS33	1	3					Y	Y	N	N	Y
Dolichospermum circinale AWQC131C	1	3	1				Y	Y	N	N	Y
Dolichospermum circinale AWQC310F	1	3	1				Y	Y	N	N	Y
Anabaena AL09	1	3	1		1	1	Y	Y	N	N	Y
Anabaena LE011-02	1	3	1			1	Y	Y	N	N	Y
Anabaena MDT14	1	3	1				Y	Y	N	N	Y
Anabaena 90	1		1				Y	Y	N	N	Y
Anabaena WA93	1	3	1				Y	Y	N	N	Y
Anabaena WA102	1		1				Y	Y	N	N	Y
Aphanizomenon flos-aquae 2012/KM1/D3	1	2	1		1		Y	Y	N	N	Y
Aphanizomenon flos-aquae NIES-81	1	3	1				Y	Y	N	N	Y
Aphanizomenon flos-aquae MDT14	1	3	1				Y	Y	N	N	Y
Aphanizomenon flos-aquae MDT13 culture	1	3	1				Y	Y	N	N	Y
Anabaena WA113	1	3	1				Y	Y	N	N	Y
Aphanizomenon WA102	1	3	1				Y	Y	N	N	Y
Cylindrospermum stagnale PCC 7417	2	10	1	1	1	4	Y	Y	N	N	Y
Nostoc punctiforme PCC 73102	3	2	1	2	1	4	Y	Y	N	Y	Y
Nodularia spumigena CCY 9414	2	2					Y	Y	N	N	Y
Anabaena CPCC64	2	3	1	1	1	1	Y	Y	N	N	Y
Anabaena variabilis ATCC 29413	2	2	1	1	1	2	Y	Y	N	N	Y
Nostoc PCC 7120	2	2	1			1	Y	Y	N	N	Y
Nostoc PCC 7524	2	10	1		1	2	Y	Y	N	N	Y
Nostoc PCC 7107	2	2	1	1		1	Y	Y	N	N	Y

Table 3.7: Genes associated with oxidative stress

Genome	No. wzy	No. kpsT	No. BcsA (alg8/alg44)	No. Exopolysaccharide biosynthesis genes	No. Capsular Exopolysaccharide genes
Richelia intracellularis RC01	1	0	1	0	0
Richelia intracellularis HH01	1	0	1	0	0
Richelia intracellularis HM01	5	1	2	1	0
Raphidiopsis brookii D9	1	0	2	1	0
Cylindrospermopsis raciborskii CS-505	2	1	6	1	0
Nostoc azollae 0708	5	3	5	7	5
Anabaena cylindrica PCC 7122	2	5	3	4	3
Anabaena PCC 7108	3	2	3	1	0
Anabaena CRKS33	4	2	6	1	0
Dolichospermum circinale AWQC131C	3	3	4	1	0
Dolichospermum circinale AWQC310F	2	2	4	1	0
Anabaena AL09	2	2	4	2	0
Anabaena LE011-02	2	2	3	2	0
Anabaena MDT14	2	2	5	2	0
Anabaena 90	2	2	6	6	2
Anabaena WA93	3	1	4	2	0
Anabaena WA102	3	1	5	5	2
Aphanizomenon flos-aquae 2012/KM1/D3	2		3	2	0
Aphanizomenon flos-aquae NIES-81	3	2	3	2	0
Aphanizomenon flos-aquae MDT14	2	1	5	2	0
Aphanizomenon flos-aquae MDT13 culture	2	1	4	2	0
Anabaena WA113	2	2	4	2	0
Aphanizomenon WA102	2	2	4	2	0
Cylindrospermum stagnale PCC 7417	5	2	5	12	5
Nostoc punctiforme PCC 73102	4	1	6	1	0
Nodularia spumigena CCY 9414	3	3	3	0	0
Anabaena CPCC64	5	1	7	2	0
Anabaena variabilis ATCC 29413	5	1	7	1	0
Nostoc PCC 7120	2	2	3	2	2
Nostoc PCC 7524	2	3	5	9	4
Nostoc PCC 7107	4	3	6	1	0

Table 3.8: EPS genes

Genome	5S	16S	23S
Richelia intracellularis RC01	1	1	1
Richelia intracellularis HH01	1	1	2
Richelia intracellularis HM01	1	1	1
Raphidiopsis brookii D9	3	2	3
Cylindrospermopsis raciborskii CS-505	3	3	3
Nostoc azollae 0708	4	4	4
Anabaena cylindrica PCC 7122	4	4	4
Anabaena PCC 7108	4	3	3
Anabaena CRKS33		2	1
Dolichospermum circinale AWQC131C		2	1
Dolichospermum circinale AWQC310F	2	1	2
Anabaena AL09			
Anabaena LE011-02			
Anabaena MDT14	5	8	4
Anabaena 90	5	5	5
Anabaena WA93	5	5	3
Anabaena WA102	5	5	5
Aphanizomenon flos-aquae 2012/KM1/D3	5	5	4
Aphanizomenon flos-aquae NIES-81	5	6	3
Aphanizomenon flos-aquae MDT14	5	12	9
Aphanizomenon flos-aquae MDT13 culture		3	3
Anabaena WA113	8	15	11
Aphanizomenon WA102	10	11	14
Cylindrospermum stagnale PCC 7417	4	4	4
Nostoc punctiforme PCC 73102	4	4	4
Nodularia spumigena CCY 9414	4	5	2
Anabaena CPCC64		1	
Anabaena variabilis ATCC 29413	12	12	12
Nostoc PCC 7120	4	4	4
Nostoc PCC 7524	5	5	5
Nostoc PCC 7107	4	4	4

Table 3.9: rRNA genes

Genome	tRNA -Ala	tRNA -Arg	tRNA -Asn	tRNA -Asp	tRNA -Cys	tRNA -Gln	tRNA -Glu	tRNA -Gly	tRNA -His	tRNA -Ile	tRNA -Leu	tRNA -Lys	tRNA -Met	tRNA -Phe	tRNA -Pro	tRNA -Ser	tRNA -Thr	tRNA -Trp	tRNA -Tyr	tRNA -Val	tRNA -Other	Sum
Richelia intracellularis RC01	3	4	1	1	1	1	1	2	1	1	4	1	3	1	3	4	3	1	1	2	0	39
Richelia intracellularis HH01	3	3	1	0	1	1	0	1	1	1	3	0	0	1	3	3	1	1	1	1	0	26
Richelia intracellularis HM01	3	4	1	1	1	1	0	2	1	1	3	1	4	1	3	4	3	1	1	2	0	38
Raphidiopsis brookii D9	4	4	1	1	1	1	1	3	1	2	4	2	2	1	3	4	3	1	1	2	0	42
Cylindrospermopsis raciborskii CS-505	4	4	1	1	1	1	1	3	1	2	4	2	2	1	3	4	3	1	1	2	0	42
Nostoc azollae 0708	7	6	2	2	2	3	3	4	2	3	9	3	5	2	4	6	4	2	3	3	1	76
Anabaena cylindrica PCC 7122	4	4	2	2	2	2	1	4	1	3	7	4	3	2	4	5	3	2	2	2	2	61
Anabaena PCC 7108	4	4	1	1	1	1	1	3	1	2	4	3	2	1	3	4	3	1	1	2	0	43
Anabaena CRKS33	3	4	1	1	1	1	1	3	1	0	4	2	3	1	4	4	3	1	1	2	0	41
Dolichospermum circinale AWQC131C	3	4	1	1	1	1	1	3	1	0	4	2	3	1	3	4	3	1	1	2	0	40
Dolichospermum circinale AWQC310F	3	4	1	1	1	1	1	3	1	0	4	2	3	1	3	4	3	1	1	2	0	40
Anabaena AL09	3	4	1	1	1	1	1	2	1	0	5	2	3	1	3	4	3	1	1	2	0	40
Anabaena LE011-02	3	4	1	1	1	1	1	2	1	0	5	2	3	1	3	4	3	1	1	2	0	40
Anabaena MDT14	5	4	2	2	1	1	1	2	1	2	6	2	3	1	3	3	3	1	1	3	0	47
Anabaena 90	5	4	1	1	1	1	1	2	1	2	5	2	3	1	3	4	3	1	1	2	0	44
Anabaena WA93	3	4	1	1	1	1	1	2	1	0	4	2	3	1	2	4	3	1	1	2	0	38
Anabaena WA102	5	4	1	1	1	1	1	2	1	2	4	2	3	1	3	4	3	1	1	2	0	43
Aphanizomenon flos-aquae 2012/KM1/D3	3	4	1	1	1	1	1	2	1	0	5	2	3	0	3	3	4	1	1	1	0	38
Aphanizomenon flos-aquae NIES-81	4	4	1	1	1	1	1	2	1	1	4	2	3	1	3	4	3	1	1	2	0	41
Aphanizomenon flos-aquae MDT14	5	4	1	1	1	1	1	2	1	2	4	2	4	1	3	4	3	1	1	2	0	44
Aphanizomenon flos-aquae MDT13 culture	4	4	1	1	1	1	1	2	1	1	4	2	3	1	3	4	3	1	1	2	0	41
Anabaena WA113	6	4	1	1	1	1	1	2	1	3	4	2	3	1	3	4	3	1	1	2	0	45
Aphanizomenon WA102	8	7	4	4	1	2	3	3	1	4	6	3	6	1	4	5	4	2	1	3	0	72
Cylindrospermum stagnale PCC 7417	7	5	2	2	2	2	3	5	1	2	7	3	4	2	4	5	5	1	1	3	1	67
Nostoc punctiforme PCC 73102	8	5	2	2	2	3	2	3	1	4	9	2	6	2	5	5	6	2	1	2	2	74
Nodularia spumigena CCY 9414	7	4	1	1	1	1	1	3	1	4	4	2	3	1	3	4	3	1	1	2	0	48
Anabaena CPCC64	3	4	2	1	1	1	1	3	1	0	4	2	3	1	3	4	3	1	1	2	0	41
Anabaena variabilis ATCC 29413	6	4	2	1	1	1	1	3	1	3	4	2	3	1	3	4	3	1	1	2	0	47
Nostoc PCC 7120	4	4	1	1	1	1	1	3	1	1	4	2	3	1	3	4	3	1	1	2	0	42
Nostoc PCC 7524	7	5	2	2	1	3	1	3	1	3	6	4	3	2	4	4	4	2	1	2	0	60
Nostoc PCC 7107	7	6	3	1	1	3	1	4	1	3	8	3	4	2	4	6	4	1	2	3	1	68

Table 3.10: tRNA genes

Genome	IS200-IS605	IS4	IS630	IS982	ISAs1	Tn3	IS3-IS150	IS5-IS1031	IS982	IS110	ISL3	IS1	IS256	IS30	IS5-IS5	IS3-IS407	IS66	IS1380	IS3-IS3	IS91	IS5-ISL2	Uncharacterized	Total
Richelia intracellularis RC01								1														0	1
Richelia intracellularis HH01																						0	0
Richelia intracellularis HM01					23			39				8			5							176	251
Raphidiopsis brookii D9												4										12	16
Cylindrospermopsis raciborskii CS-505		4									3	1										34	42
Nostoc azollae 0708								3														22	25
Anabaena cylindrica PCC 7122		5			1	2		2	2		4	1										14	31
Anabaena PCC 7108		2	1		1	2																14	20
Anabaena CRKS33			1		1						1				1	1	4					38	47
Dolichospermum circinale AWQC131C			1	2	1							1										13	18
Dolichospermum circinale AWQC310F				1	1																	11	13
Anabaena AL09	2					2					1											26	31
Anabaena LE011-02						1		1	1			2										24	29
Anabaena MDT14		1	1	1			1	1		1		2			1			1	1	1		68	80
Anabaena 90	2				1			1														34	38
Anabaena WA93	1			2		1				1												49	54
Anabaena WA102	3	4		3	1	1				2	2							2				68	86
Aphanizomenon flos-aquae 2012/KM1/D3	5	2	2	1	2	1																32	45
Aphanizomenon flos-aquae NIES-81	1	1	1	1	1			1		1	1											24	32
Aphanizomenon flos-aquae MDT14	1				1			1														44	47
Aphanizomenon flos-aquae MDT13 culture	1						1	1	1													31	35
Anabaena WA113					3		1	1						2								67	74
Aphanizomenon WA102		1		1	4	1		2			2	3	1	2	1							87	105
Cylindrospermum stagnale PCC 7417			1	1	2	1		1														58	64
Nostoc punctiforme PCC 73102			2	1	1	6		2		1	2	1									1	48	65
Nodularia spumigena CCY 9414	1							1														47	49
Anabaena CPCC64		1						3														34	38
Anabaena variabilis ATCC 29413		1						6														41	48
Nostoc PCC 7120	1							34													1	68	104
Nostoc PCC 7524		3	1		1					1		1										39	46
Nostoc PCC 7107	2			3				6		1	5											50	67

Table 3.11: IS sequences

Chapter 4 Genome sequencing of two novel Ma-LMM01-like strains reveals patterns of conservation and divergence in a globally distributed *Microcystis* phage type

Connor B. Driscoll[1], Timothy G. Otten[1], Theo W. Dreher[1,2]

[1]Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR, 97331, USA.

[2]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.

In preparation

4.1 Introduction

Microcystis aeruginosa is a toxic, bloom-forming cyanobacterium present in globally distributed eutrophic freshwater systems [Wu et al. 2007; Marmen et al. 2016]. It can produce microcystins, a group of potent hepatotoxins that have been implicated in the deaths of livestock and humans, and may also cause hepatocellular carcinoma [Nishiwaki-Matsushima et al. 1992; Yoshizawa et al. 1990]. As a result, freshwater systems at risk for *Microcystis* blooms must be monitored for water management purposes.

Cyanophages are a diverse set of viruses that infect cyanobacteria in both marine and freshwater systems across the world [Mann et al. 2005; Brussaard and Martinez 2008; Dreher et al. 2011]. The marine cyanophages have been shown to play an important role in biogeochemical cycles, as well as regulating cyanobacterial populations and mediating horizontal gene transfer events between hosts [Clokie and Mann 2006; Mann 2003; Mühling et al. 2005; Mann and Clokie 2012]. Core and pan-genome analysis of marine cyanophages has revealed a set of core shared genes, which are frequently host-associated genes such as *phoH*, *mazG*, and *psbA* in addition to structural and replication genes [Sullivan et al. 2010]. Similar analyses of closely related T4 strains identified the presence of interspersed hyperplastic genome regions [Comeau et al. 2007]. These genomic segments often contain unique genes (ORFans) that are found in novel phage genomes [Yin and Fischer 2008]. However, their origin, role in natural phage population dynamics, and their gain or loss over time in natural systems has not been characterized.

Previously studied cyanophages primarily have been isolated from the marine picocyanobacteria (*Prochlorococcus* and *Synechococcus*), while fewer freshwater cyanophage genomes have been sequenced [Chen and Lu 2002; Millard et al. 2009; Sullivan et al. 2005; 2010]. Thus far, three freshwater cyanomyoviruses have been isolated and sequenced. One was isolated from a *Synechococcus* strain from Copco Reservoir in the USA [Dreher et al. 2011]. Also, two strains of a phage that infects *M. aeruginosa* have been isolated and sequenced. The first, Ma-LMM01, is a *Microcystis*-specific phage isolated from Lake Mikata, Japan which only infected *M. aeruginosa* NIES-298 out of nine tested *Microcystis* strains [Yoshida et al. 2006]. The second, MaMV-DC, was isolated from Lake Dianchi, China and only infected *M. aeruginosa* FACHB-524 out of nine tested *Microcystis* strains [Ou et al. 2013]. Both have an icosahedral head and a contractile tail, and have been characterized as myoviruses based on these morphological features [Yoshida et al. 2006; Ou et al. 2013]. While both are lytic, they each carry putative prophage antirepressor genes, which may suggest a possible lysogenic lifestyle [Lemire et al. 2011]. Their stringent host specificity indicates a significant hurdle complicating freshwater cyanophage isolation, and suggests the necessity of culturing both host and phage from the same environment to increase the likelihood of successful isolation.

Here, we used a culture-independent approach to sequence and assemble two novel phage genomes sharing high similarity with phages Ma-LMM01 and MaMV-DC from shotgun metagenomes of geographically disparate *Microcystis* blooms in North America. We compared these genomes to better understand gene conservation and host-phage evolution in this widespread phage-type. In addition, one

of the novel phages was present in four samples collected two weeks apart from Cheney Reservoir, KS, USA. We compared these genomes to assess population variability or gene gain/loss in the environment.

4.2 Methods

4.2.1 Sequenced samples and assembly

The MaCRKS23 genome originated from a depth-integrated, 0.2 μm filtered sample collected from Cheney Reservoir, Kansas on July 8, 2013 at 37.7597° latitude, -97.835° longitude. MaSF12 originated from a depth-integrated, 0.2 μm filtered sample collected from near Mildred Island (38.9860° latitude, -121.5204° longitude) in the San Francisco Delta on August 27, 2012. Total DNA was extracted from filters with Gene-Rite DNA-EZ RW01 extraction kits. Libraries were prepared with Nextera XT library kits, and samples were sequenced with Illumina HiSeq 2000. Reads were assembled with IDBA-UD with default parameters [Peng et al. 2012], and contigs with significant similarity (¡1e-30 BLASTN E-value) to phage Ma-LMM01 were extracted. These fragmented assemblies were completed using PriceTI [Ruby et al. 2013] with the following parameters: PriceTI -icf inputcontigs.fasta 1 1 2 -fpp Fwdreads.fastq Revreads.fastq 500 90 -nc 81 -nco 5 -rqf 95 0.998 0 14 -rqf 95 0.99 14 6 -rqf 95 0.9 20 10 -rqf 90 0.9 30 10 -rqf 80 0.6 40 20 -trim 25 2 -trim 35 2 -trim 45 2 -trim 55 2 -trim 65 3 -trim 70 2 -lenf 60 1 -lenf 70 5 -lenf 80 20 -reset 5 10 14 18 20 25 30 35 40 45 50 55 59 60 63 65 70 75 -target 90

3 2 2.

All assemblies were then validated by manual assessment of paired read-mapping from each original metagenome read-set using BWA-MEM with default parameters [Li 2013].

4.2.2 Genome annotation and gene clustering

Genomes were annotated with Prokka [Seemann 2014], and the resulting GenBank files were used for input to the BYU implementation of Phamerator ([Cresawn et al. 2011], https://github.com/byuphamerator/phamerator-dev/). This process uses ClustalOmega [Sievers and Higgins 2014] and BLASTP to align protein sequences within and between genomes. Then, sequences are clustered into "phamilies" using specified lower minimum cutoff parameters (we used 32.5% identity and an E-value of 1e-50). Conserved domains in each protein were identified using the cddSearch.py script that is part of the BYU implementation of Phamerator to compare proteins against the Conserved Domain Database (CDD). Protein sequences were then searched against the non-redundant protein (nr) database using BLASTP with a 1e-10 maximum E-value cutoff.

4.2.3 Phylogenetic tree

The protein-coding sequences for the large terminase subunit were extracted from each genome, including a set of previously-sequenced freshwater and marine cyanophages.

These were subsequently used for multiple sequence alignment with PROMALS3D, which uses a combination of sequence-similarity alignments with predicted secondary structures [Pei et al. 2008]. A maximum-likelihood tree was then generated using FastTree with default parameters [Price et al. 2010].

4.2.4 Metagenome search

We searched for evidence of these genomes in 62 freshwater metagenomes we previously collected from eight sampling sites in different states across the USA, including Oregon, Washington state, California, Texas, and Kansas (BioProject accessions: PRJNA312985, PRJNA282166, PRJNA312830, PRJNA312986, and PRJNA294203, respectively). Additionally, we searched through 50 additional freshwater metagenomes from the IMG [Markowitz et al. 2012], MG-RAST [Glass et al. 2010], and SRA [Leinonen et al. 2010] databases. All metagenome searches were performed using BWA-MEM using default parameters [Li 2013].

4.2.5 Cheney metagenome comparisons

Metagenomes from Cheney Reservoir samples consisting of 100bp paired-end Illumina HiSeq 2000 reads which contained >10x read coverage over the MaCRKS23 genome were assembled with IDBA-UD [Peng et al. 2012]. Sequences associated with MaCRKS23 were extracted from each assembly, and assemblies improved using PriceTI with the same parameters as mentioned earlier [Ruby et al. 2013]. As-

semblies were annotated with Prokka [Seemann 2014]. Additionally, MaCRKS23-like phage reads from each metagenome were mapped and extracted with BWA-MEM with default parameters. These reads were subsequently mapped back to each genome to identify genomic variants/missing genomic sequences with Breseq [Deatherage and Barrick 2014], and these variants were manually verified by comparing assemblies between each time point using progressiveMauve alignments [Darling et al. 2010]. Genes between these assemblies were compared by creating codon alignments with Pal2Nal [Suyama et al. 2006], using both Clustal Omega amino acid alignments [Sievers and Higgins 2014] and DNA sequences as input. Then, PAML was used to calculate dN/dS, and non-synonymous and synonymous substitutions using these codon alignments [Yang 2007].

4.3 Results

4.3.1 Isolating assembled sequences from metagenomes

The two novel genomes here were assembled from cellular fraction metagenomes, suggesting the likelihood that these phage sequences have been extracted from cells undergoing an active phage infection cycle. Previously, fosmid clones from environmental DNA have contained phage DNA sequences [DeLong et al. 2006; Ghai et al. 2010; Zhao et al. 2013], while analysis of cellular shotgun metagenomes has revealed an abundance of phage-derived sequences [Mizuno et al. 2013]. Since we have also identified phages from cellular metagenomes, our results indicate

the identification of actively infecting phages in the population may be possible through cellular metagenomics.

4.3.2 General characteristics

Both genomes were assembled into circular contigs, which is consistent with the linear, circularly permuted genomes reported from the previously sequenced strains [Yoshida et al. 2008; Ou et al. 2015a]. The genome sizes of both MaCRKS23 and MaSF12 (173,787 and 176,940 bp, respectively) are larger than those of Ma-LMM01 and MaMV-DC (162,109 and 169,223, respectively). The number of protein-coding genes is variable (Table 4.1), likely as a result of the genome size differences and variation in certain genes, some of which are found in hyperplastic genomic regions (Figure 4.1). The Ma-LMM01 genome contains 21 small ORFs in this region, while the MaMV-DC genome contains 10 small ORFs, the MaCRKS23 genome contains 15 small ORFs, and the MaSF12 genome contains 25 small ORFs in their respective hyperplastic regions. This suggests this hyperplastic region is undergoing expansion and contraction events likely through indels of these small ORFs. Analyses of hyperplastic regions of marine cyanophages indicates the small ORFs are often host-derived [Millard et al. 2009]. However, the genes within these regions do not share significant similarity to known *Microcystis* genes.

The average GC-content of these genomes is stable at near 46% (Figures 4.2 and 4.3). Also, the number of tRNA-encoding genes in these genomes are mostly consistent, with each carrying tRNA's for methionine and tyrosine, and MaCRKS23

carrying an extra methionine tRNA.

Pairwise ANI comparison of the Ma-LMM01-like phages revealed a range of 92.60%-97.04%, with Ma-LMM01 and MaMV-DC sharing the highest similarity, and the American strains sharing the least similarity (Table 4.2). This suggests the Japanese and Chinese strains are more closely related than any other pair, while the North American strains are the most divergent two genomes.

4.3.3 Phylogenetic characterization

We characterized these genomes by phylogenetic analysis of the conserved TerL protein-coding sequences from all currently-sequenced freshwater cyanophages and several representative marine cyanophages (Figure 4.4). All *Microcystis* phages clustered closely together in a single clade, reflecting their close relationship and separation from other known myophages. They separate into a larger clade with the marine and freshwater *Synechococcus*-infecting myoviruses and the uncharacterized *Planktothrix* phage PaV-LD, while the *Anabaena* phage A-4L and the *Phormidium* podoviruses separate into a diverse clade, with the T7-like marine viruses clustering together. As a result, a similar diversity of freshwater cyanomyoviruses may arise as more genomes are sequenced.

4.3.4 Gene content

Using Phamerator, we compared shared phams, or clustered protein sequences from each completed genome (Table 4.3). Of the 238 total gene clusters identified here, 124 (52%) are in all four genomes, 24 (10%) are in three genomes, 24 (10%) are in two genomes, and 66 (28%) are unique to a single genome. Similar to the previously-annotated Ma-LMM01 and MaMV-DC genomes, MaCRKS23 and MaSF12 consist primarily of hypothetical genes (156/201 = 78% of MaCRKS23 genes; 157/210 = 75% of MaSF12 genes) (Figures 4.2 and 4.3). The genes in the hyperplastic regions are generally clustered together.

4.3.4.1 Conserved genes

All genes identified as being associated with replication and virion structure are conserved across these four genomes (Table 4.3). A putative prophage antirepressor was also conserved across all genomes (pham 9), while Ma-LMM01 and MaCRKS23 each carried an additional putative antirepressor gene (pham 161) that is non-homologous with the first. Although conserved across these genomes, the gene encoding the ribonucleotide reductase alpha subunit (*nrdA*, pham 6) is interrupted by an in-frame intein sequence in MaSF12. Previously-sequenced phage *nrd* genes contain in-frame introns and inteins, suggesting these sequences are particularly susceptible to interruption by these mobile elements [Dwivedi et al. 2013].

Several host-like sequences are found in all four genomes. Each genome encodes multiple host-like serine/threonine protein kinase genes (pham 39) with the excep-

tion of a single copy found in MaCRKS23. Additionally, each genome encodes a single serine/threonine protein phosphatase gene (pham 24), although there was not significant similarity to any host genes. This suggests these phages are capable of modulating the phosphorylation state of host or other protein(s). For example, phage T7 encodes a serine/threonine kinase which phosphorylates multiple host proteins, while the lambdoid phage 933W expresses a kinase gene in response to co-infection by phage HK97 [Gone and Nicholson 2012; Robertson 2011]. However, the advantages to carrying these genes is not known.

The *nblA* gene is found in all genomes, although these sequences can diverge as shown previously [Ou et al. 2015a; Nakamura et al. 2014]. Aligning them reveals that the Ma-LMM01 and MaCRKS23 copies are most similar, since they both contain similar fourteen-residue N-terminal extensions relative to the copies in MaMV-DC and MaSF12 (Figure 4.5). Ou et al. showed that gene expression of *nblA* in MaMV-DC is associated with reduced phycocyanin levels during phage maturation and release *in vivo* [Ou et al. 2015a]. They suggest this is to recycle the abundant host phycobilisome proteins to create amino acid supplies necessary for phage growth. Others have suggested the *nblA* gene increases rates of photosynthesis by preventing absorption of excess light energy (and therefore photoinhibition) through phycobilisome degradation [Yoshida-Takashima et al. 2012; Honda et al. 2014]. The presence of this gene in all genomes suggests it is an important component driving successful infection of *Microcystis*.

Both a putative chitinase and chitin-binding protein are found in each of these genomes. Previous characterization of chitinase genes suggests there is structural

similarity with chitinase sequences in plants and lysozymes in phages [Holm and Sander 1994]. Furthermore, overexpression in *E. coli* of a putative chitinase from a *Ralstonia* myovirus revealed lytic-like activity in which rod-shaped cells became round and aggregated [Yamada et al. 2010]. This may suggest the putative chitinase sequences in these genomes provide lytic activities for these phages.

Additionally, all genomes carry a Cas4-like nuclease-encoding gene. Cas4-like genes have been identified in *Campylobacter* phages, and previous work by Hooton and Connerton identified that infection with these phages led to increased host-derived spacer acquisition [Hooton and Connerton 2015]. This may act as a phage-driven form of autoimmune activation, whereby host CRISPR-Cas activity is diverted towards host DNA degradation, and not phage [Hooton et al. 2016]. Other conserved genes include a putative L-lysine 6-monooxygenase gene and the phosphate-starvation gene *phoH* that is commonly found in marine cyanophages.

4.3.4.2 Variable genes

Several gene clusters with annotated functions are present in some, but not all of these genomes. A putative host-like pentapeptide repeat protein (pham 31) is present in MaCRKS23, MaSF12, and is present in two copies in MaMV-DC. Pentapeptide-repeat proteins in cyanobacteria have a variety of functions, including heterocyst maturation and differentiation [Black et al. 1995; Liu et al. 2002], and manganese uptake [Chandler et al. 2003].

There are six putative transposase gene clusters found throughout these four

genomes. Three are unique to single genomes (phams 165, 194, and 213), while others are in two or three genomes (phams 123, 164, and 195). These genomes carry between one and five putative transposase genes, with MaSF12 carrying the greatest number. All transposase sequences are similar to putative transposases in *Microcystis* genomes, indicating that phage-encoded transposases are often shuttled between cells by these phages.

Three genomes contain putative antitoxin genes. MaCRKS23 uniquely carries a putative *higA* antitoxin (pham 235), while MaMV-DC and MaSF12 both carry an XRE family antitoxin (pham 33). Some instances have been shown where toxin-antitoxin (TA) systems in bacteria can protect against phage infection. For example, TA systems can function as abortive infection (Abi) systems which increase time to phage maturation and diminish burst size [Pecota and Wood 1996], probably due to skewed toxin:antitoxin ratios following phage-altered translational levels [Fineran et al. 2009; Koga et al. 2011]. To counteract this, phage can carry antitoxin-mimicking genes that protect against Abi systems. For example, phage T4 encodes a broadly effective antitoxin which protects against multiple toxins [Otsuka and Yonesaki 2012].

Both MaMV-DC and MaCRKS23 contain putative selenoprotein O homologs (pham 13), which share 97% amino acid identity with the identically-annotated protein sequence from *Microcystis panniformis* FACHB-1757. Selenoproteins can provide antioxidative functions, which may be beneficial for infecting photosynthetic organisms. Orthologues of this protein are found in many different bacterial and eukaryotic genomes, and recent work identified these proteins engaging in re-

dox reactions with unknown proteins in mammalian cell mitochondria [Han et al. 2014]. Additionally, they may also be protein kinases [Lenart and Pawłowski 2013]. If true, it's possible these proteins play a role in intracellular signaling associated with photosynthesis-induced ROS build-up. Furthermore, others have suggested that phages may promote host resistance to oxidative stress by increasing production of host-encoded selenoproteins [Szemes et al. 2012].

4.3.5 Environmental metagenome search and time-series comparisons

To further assess the geographic distribution of this virus, we searched through 50 metagenomes from freshwater environments in the MG-RAST, SRA, and IMG databases by mapping reads, but were unable to identify samples with reads mapped to Ma-LMM01-like genomes. We also searched through 62 freshwater metagenomes collected by our laboratory. We only identified consistent, $\geq$20x coverage in three metagenomes from Cheney Reservoir, KS, USA. These samples (CRKS24, CRKS25, CRKS27) are derived from the same environment and sampling site as the completed MaCRKS23 genome. Additionally, these samples, starting with CRKS23, are part of a time-series with two week intervals in-between (July 8, 2013 to August 19, 2013). While phage CRKS23 was present throughout the 6-week period in summer 2013, we did not identify reads associated with these phage genomes in samples from the following year (February 19,2014 - December 16, 2014; 26 samples), suggesting these phages are not consistently abundant from

year-to-year. As a result, we have identified a 6-week-long period where this phage is present in this system, indicating the persistence of infection in this environment during this time.

Further, we assembled portions of these genomes from each time point and compared to each other to better understand genomic variants and population dynamics in the environment. We employed read-mapping to identify regions of low coverage, indicating which segments of the genome may have significant changes, and then compared across genome assemblies.

In comparing read-mapping and assemblies in the time series, we identified gene insertion/deletions and sequence divergence in certain genes (Table 4.4). These genes were primarily annotated as hypothetical proteins. Of the eight gene insertions identified, three of the best hits are to genes found in MaMV-DC. The others include transposases with best hits to *Microcystis* and *Oscillatoria* genes, and a lysine-tRNA gene, while the remainder had no significantly similar sequences in the *nr* database.

In particular, one annotated gene that is different in sequence composition between the time-series assemblies is a putative tail collar domain protein (Fig. 4.6). We calculated pairwise dN/dS ratios of the tail collar gene annotated from each time point (Table 4.5). For comparison, we also calculated dN/dS for the major capsid gene, which is expected to be highly conserved. The average pairwise dN/dS ratio is much higher for the tail collar gene compared with the major capsid gene pairwise dN/dS ratio (1.0912 vs. 0.04945, respectively). However, the range of dN/dS values for each gene comparison is notably different (0.5213-1.6975 for

the tail collar gene, 0.001-0.1131 for the major capsid gene). Additionally, there is a significant difference in the ratio of non-synonymous to synonymous mutations between these genes (Fisher's exact test, p-value = 2.2e-16). These results indicate the capsid gene is under strong purifying selection. On the other hand, selective pressures on the tail collar gene may vary over time, as indicated by the range in dN/dS values from pairwise comparisons, but overall are neutral or slightly positive.

4.4 Discussion

4.4.1 Novel genomes add to undersampled freshwater cyanophage genomes

Including these novel genomes, only eleven freshwater cyanophage genomes have been sequenced to date, indicating the potential for future research. These four are currently the only sequenced *Microcystis*-infecting phage genomes available. In comparison to mycobacteriophages, which are the most well-sampled group of sequenced phage genomes available [Hatfull 2010], these four *Microcystis* phages are in the upper end (94%) of genome similarity based on ANI values in comparison with clustered mycobacteriophage genomes from geographically distant isolation sites [Pope et al. 2011]. Notably, these genomes also come from geographically distributed environments (East Asia and North America). The similarity in these genomes suggests this is a successful phage group that is broadly capable

of infecting *Microcystis* strains across the world. Comparisons of new freshwater cyanophages with the similarities within the Ma-LMM01-like group may contribute to our understanding of the forces affecting the evolution of freshwater cyanophages, and may help identify new, globally distributed phage archetypes infecting bloom-forming freshwater cyanobacteria.

A search through environmental metagenomes yielded positive hits for Ma-LMM01-like viruses in the Cheney Reservoir in several samples following the July 8, 2013 sample. Our search also included metagenomes from the 2014 year during a *Microcystis* bloom. In these metagenomes, we did not detect read-coverage for the Ma-LMM01-like viruses. This suggests that the virus is either not present in the population at this time, or is present at such a low level so as not to be detected by shotgun sequencing. Previous work by Kimura et al. employed QPCR to track abundance of Ma-LMM01 and *Microcystis* cells in Hirosawanoike Pond, Japan [Kimura et al. 2012]. Their results suggest that Ma-LMM01 abundance is variable, but it persists across a seven-month time span in this environment. Other studies have also shown the presence of Ma-LMM01-like DNA sequences in freshwater samples in Lake Ontario, Canada and Sulejow Reservoir, Poland [Mankiewicz-Boczek et al. 2016; Rozon and Short 2013]. The same may be true in Cheney Reservoir across these two sampling seasons, where the 2013 season harbored more abundant, detectable phage numbers than the 2014 season.

4.4.2 Conserved and variable genes inform about consistency and differences in lifestyle

The novel genomes presented here are similar to previously-sequenced *Microcystis* phages (Ma-LMM01 and MaMV-DC) (Figure 4.4, Table 4.2). In total, there are 124 gene clusters in these four genomes, about 52% of which were identified in all four strains. Clusters found in all four strains, on average, make up 65% of the number of predicted ORFs in each genome. As a result, as much as a third of the ORFs in any given Ma-LMM01-like genome are part of the "shell" genome for these strains. The percent of conserved phams across the four strains is relatively small in comparison with mycobacteriophage clusters with a similar number of sequenced genomes [Hatfull et al. 2010], some of which are clustered at much lower ANI (as low as 54%) compared with the *Microcystis* phages analyzed here. These 124 conserved gene phamilies may then represent the core or essential gene content for this globally distributed group of phages, and inform about processes necessary for this group to infect *Microcystis*. The remaining 48% of non-conserved gene phamilies may represent genes in a state of flux, that are gained or lost within or between populations. Many of these genes are small ORFans, which are part of hyperplastic regions of the genome [Comeau et al. 2007].

Putatively essential genes consist of structural- and recombination-associated genes. They also include host-like genes such as *nblA* and *phoH*, suggesting that regulating cellular phosophate uptake and phycobilisome degradation are important, if not essential, for successful infection. Marine cyanophage replication is

strongly affected by diurnal cycles [Lindell et al. 2005], and Kimura et al. showed that Ma-LMM01 gp91 copy numbers in the environment increased between 6 to 9 hours after dawn [Kimura et al. 2012], which is similar to the latent period in culture (6 to 12 hours) [Yoshida et al. 2006]. Also, transcript analysis of Ma-LMM01 infection in culture by Honda et al. revealed the upregulation of stress-induced genes involved in protecting the photosynthetic apparatus [Honda et al. 2014]. Alternatively, *nblA* may provide recycled amino acids for phage growth [Ou et al. 2015a]. It's possible these genes could provide both functions during infection, although further work is necessary to determine if this is the case.

Additionally, a Cas4-encoding gene is conserved in these four strains, similar to *Campylobacter* phages [Hooton and Connerton 2015]. *Microcystis* genomes consistently harbor CRISPR-Cas gene arrays, and many carry the Type I-D system which requires the Cas4 protein [Yang et al. 2015]. While CRISPR protospacer mutations have been shown in Ma-LMM01-like viruses [Kimura et al. 2013], alternative mechanisms may be necessary for phages to infect *Microcystis* strains which encode diverse CRISPR systems [Kuno et al. 2012]. Infection of *Campylobacter* with a phage encoding a *cas4*-like gene led to increased acquisition of host-like spacer sequences in the host CRISPR [Hooton and Connerton 2015]. The role of this during phage infection is unknown, but it may act as an alternative CRISPR escape mechanism [Hooton et al. 2016]. Although host-derived CRISPR spacers may control gene expression, this is unlikely since Ma-LMM01 has experimentally been shown to alter the expression of few host genes during infection [Honda et al. 2014].

Additionally, all genomes analyzed contain putative prophage antirepressor genes, and closer inspection of these genes using HHPRED further supported these annotations. Antirepressor proteins play a role in prophage induction, which may indicate that these phages exhibit a temperate lifestyle [Lemire et al. 2011]. However, this is the only lysogeny-related gene found in these genomes (the previously-annotated site-specific integrase in Ma-LMM01 is actually more similar to a transposase). Also, these phages have not shown any evidence of lysogeny *in vivo*. Further experiments are needed to verify the function of these genes.

Phams found in three or fewer genomes consist of the shell genome of this phage group. Genes in this group include host-like toxin antidote genes, a pentapeptide repeat gene, and a selenoprotein. These genes may provide benefits for infecting certain *Microcystis* strains, perhaps for escaping strain-specific TA systems or protecting against excess oxidative damage from *Microcystis* strains found in high-light environments. As a result, these genes may be under variable selection based on environment-specific host strain differences.

4.4.3 Ma-LMM01-like phages in the environment show evidence of gene gain, loss, and divergence

The Ma-LMM01-like phages identified in the Cheney Reservoir time-series carry differences in their respective genomes at each time point. These differences may be due to existing variation in the population. Alternatively, these differences may be due to mutations occurring in the time between each sampling. Phage genomes are

known to be mosaic, and are susceptible to gene gain, loss, and swapping through HGT events [Sullivan et al. 2006; Hatfull et al. 2010]. Here, we identified variable presence or absence of some genes over the time-series that were most closely-related to genes found in Ma-LMM01 or MaMV-DC. In particular, genes newly present in later time points may indicate their persistence in a common genetic pool that spans different environments, and is made available to phage genomes through frequent HGT events. In turn, this could support the idea of a large common genetic pool available to all dsDNA phage genomes proposed by Hendrix et al. [Hendrix et al. 1999]. Additionally, genes that appear and later disappear in these genomes (Table 4.4, CRKS24_00035-00036, CRKS24_00043, CRKS27_00014-00015) may indicate these genes persist at some level in the phage population over this time span.

There are also patterns of divergence in some genes between these samples. Notably, the tail collar-encoding gene seems to be under positive selection in stark contrast to the major capsid gene (Fig. 4.6). Phage tail collar proteins act as environmental sensors that bind to tail fibers to sequester them from binding host receptors until certain conditions (pH, ionic strength) are reached [Conley and Wood 1975]. An alignment of tail collar genes from the Cheney time-series indicates that most variation occurs in the C-terminal sequence. Previous structural studies of the T4 phage neck indicate the N-terminus of tail collar genes associate with the phage head, while residues closer to the C-terminus interact with the tail fibers [Fokine et al. 2013]. The variation in this gene may then be due to neutral or positive selection at segments of the gene encoding the protein C-terminus.

Tail fiber proteins allow tailed phages to bind to host cell receptors [Duplessis and Moineau 2001; Heller 1984; Rakhuba et al. 2010], and these genes tend be variable relative to host range [Tetart et al. 1998]. If so, positive selection in tail collar genes could allow co-evolution of the tail fiber and tail collar genes so phages retain adequate control over infection. This could protect untimely activation of the phage injection machinery from particulate matter in the environment. These variants may co-exist in the population, or arise during the observed sampling period. Regardless, genotype dominance may change over time depending on selective forces, similar to previous reports of the Ma-LMM01 tail sheath gene [Kimura et al. 2013; Mankiewicz-Boczek et al. 2016]. As a result, the tail collar gene in these Ma-LMM01-like phages may be integral to the co-evolutionary "arms race" between these viruses and their hosts [Hall et al. 2011].

4.5 Conclusions

Here we present two novel genomes that are very similar to previously characterized and sequenced *Microcystis* phage genomes. Together, these *Microcystis* phages comprise a globally distributed group of viruses with a similar genomic archetype. These genomes encode a variety of genes to escape host defenses. While some are conserved, others are variable, suggesting that certain genes may provide an advantage for infecting particular strains of *Microcystis*. In a single environment over a short time-period, some gene content varies relative to the MaCRKS23 genome. In addition, selective pressures on different structural genes are variable,

indicating genes encoding components of the virion structure are more susceptible to mutation than others.

Genome	**Genome size (bp)**	**GC%**	**Protein-coding genes**	**No. tRNAs**
Ma-LMM01	162,109	46.0	184	2
MaMV-DC	169,223	46.0	170	2
MaSF12	176,940	45.5	210	2
MaCRKS23	173,787	45.4	201	3

Table 4.1: General characteristics of *Microcystis* phage strains.

Genome	**Ma-LMM01**	**MaMV-DC**	**MaCRKS23**	**MaSF12**
Ma-LMM01	100	97.02	94.12	94.06
MaMV-DC	97.04	100	93.38	94.24
MaCRKS23	94.18	93.92	100	92.61
MaSF12	94.23	94.3	92.6	100

Table 4.2: Pairwise ANI calculations for Ma-LMM01-like phages.

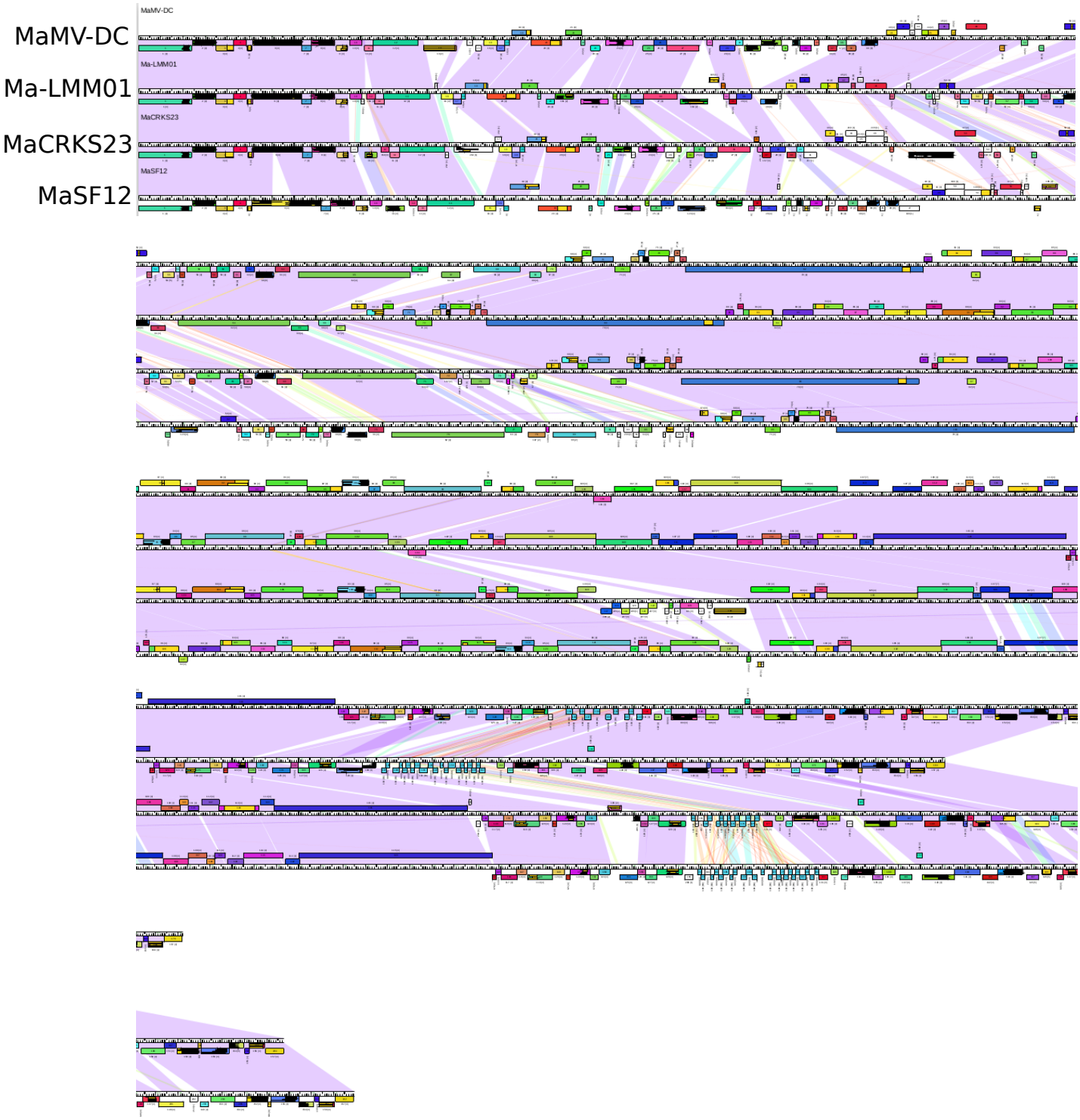

Figure 4.1: Phamerator-generated genome maps of Ma-LMM01 phage strains

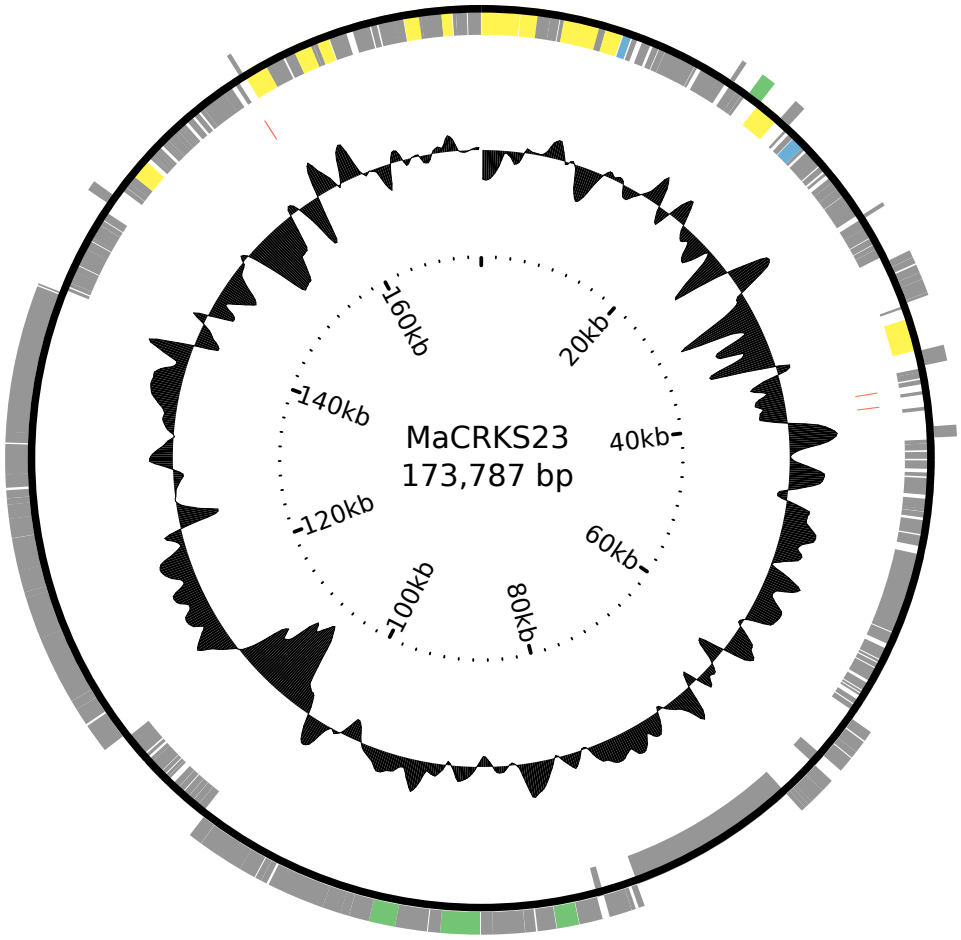

Figure 4.2: Circular genome plot of MaCRKS23. Outermost, black circle represents the genome, with outside marks showing forward orientation ORFs, and inside marks showing reverse orientation ORFs. Grey marks are coding sequences with no known function, while yellow marks show sequences with replication function, green marks sequences encoding virion structural components, and blue marks sequences indicative of viral lifestyle. Further towards the center, red marks show tRNA-encoding sequences. The next circle shows GC% of genome regions relative to the average GC%.

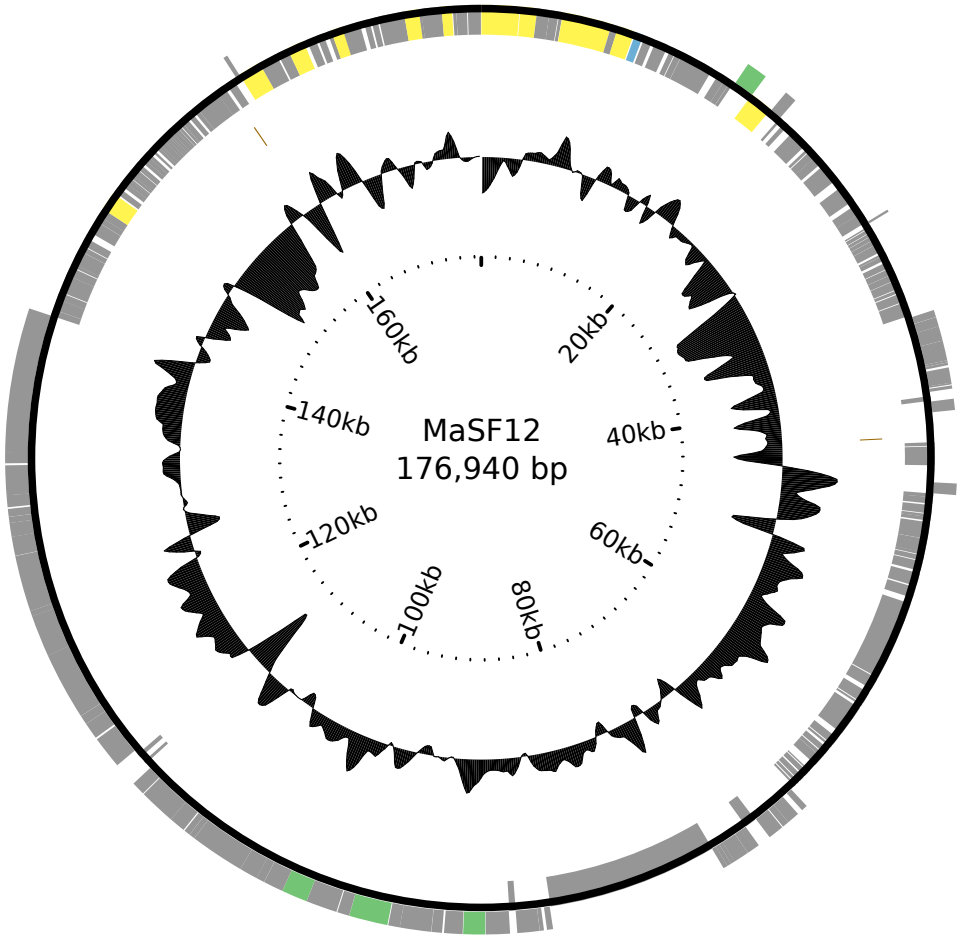

Figure 4.3: Circular genome plot of MaSF12. Each circle is as described in the Figure 4.2 caption.

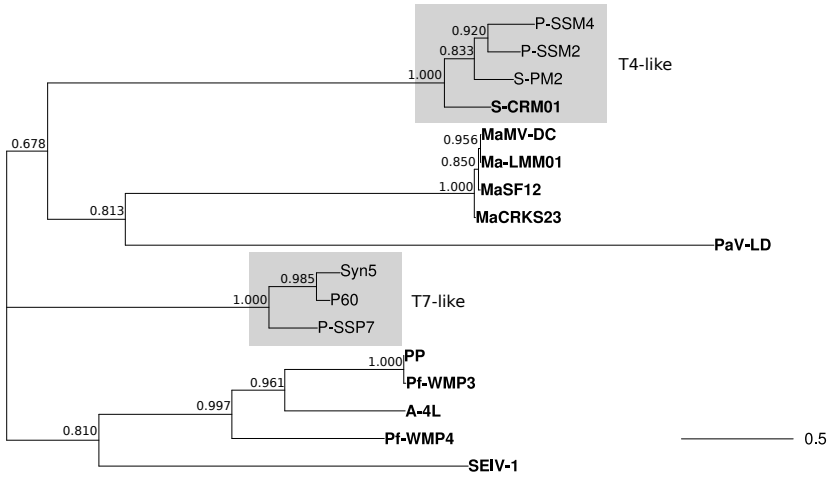

Figure 4.4: TerL phylogeny of freshwater and some marine cyanophages relative to the newly sequenced MaSF12 and MaCRKS23 phages. Bolded genome labels are freshwater cyanophages. Grey boxes indicate phages classified as either T4-like or T7-like.

Figure 4.5: Multiple sequence alignment of NblA protein sequences encoded by the four *Microcystis* phages.

Putative function	Pham	Number of members	Mean translation length	Ma-LMM01	MaCRKS23	MaMV-DC	MaSF12	BLAST hit Accession no.
Ribonucleoside-diphosphate reductase subunit alpha	6	4	811.5	Ma-LMM01_gp006	MaCRKS23_00006	MaMV-DC_gp006	MaSF12_00006	WP_045361327.1
Prophage antirepressor	9	4	164.25	Ma-LMM01_gp009	MaCRKS23_00009	MaMV-DC_gp009	MaSF12_00009	
Selenoprotein O	13	2	481.5		MaCRKS23_00129	MaMV-DC_gp013		CCI09371.1
Serine/threonine protein phosphatase	24	4	394.75	Ma-LMM01_gp025	MaCRKS23_00028	MaMV-DC_gp024	MaSF12_00023	
Pentapeptide repeat protein	31	4	248.25		MaCRKS23_00011	MaMV-DC_gp031, MaMV-DC_gp032	MaSF12_00011	CCI08969.1
Putative antitoxin protein	33	2	95			MaMV-DC_gp034	MaSF12_00036	CCI32511.1
Serine/threonine protein kinase	39	9	240.6667	Ma-LMM01_gp062, Ma-LMM01_gp037	MaCRKS23_00068	MaMV-DC_gp040, MaMV-DC_gp063	MaSF12_00044, MaSF12_00046, MaSF12_00073, MaSF12_00074	WP_045360364.1
Putative transposase	123	4	351	Ma-LMM01_gp135		MaMV-DC_gp127	MaSF12_00027, MaSF12_00061	WP_016515614.1, WP_008206903.1
Prophage antirepressor	161	2	270	Ma-LMM01_gp024	MaCRKS23_00026			
Putative transposase	164	2	393	Ma-LMM01_gp031			MaSF12_00028	WP_016516871.1
Putative transposase	165	1	139	Ma-LMM01_gp032				WP_002769560.1
Putative transposase	194	1	98				MaSF12_00057	WP_061430971.1
Putative transposase	195	2	213.5		MaCRKS23_00157		MaSF12_00058	BAF99932.1
Putative transposase	213	1	575		MaCRKS23_00016			WP_061433345.1
Putative antitoxin protein	235	1	99		MaCRKS23_00158			WP_045358675.1

Table 4.3: Pham clusters of interest.

Gene	Annotation	Type of difference	BLASTP hit	Accession No.	Similarity	CRKS23	CRKS24	CRKS25	CRKS27
CRKS23_00063	hypothetical protein	ORF lost	MaMV-DC	YP_851070.1	73% (E=5e-31)	Y	Y	Y	N
CRKS23_00075	hypothetical protein	Split into two genes (CRKS27_00014-15) and large indel	MaMV-DC	YP_009217751.1	85% (E=2e-161)	Y	N	N	Split
CRKS23_00084	hypothetical protein	Divergence	MaMV-DC	YP_009217757.1	77% (E=1e-175)	Y	Y	Y	Y
CRKS24_00004	Hypothetical protein	Truncation	Ma-LMM01	YP_851112.1	88% (E=0.0)	Y	Truncated	Y	Y
CRKS24_00031	Hypothetical protein	Divergence	MaMV-DC	YP_009217757.1	77% (1e-174)	Y	Y	Y	Y
CRKS24_00035	hypothetical protein	Insertion	MaMV-DC	YP_009217752.1	82% (E=0.0)	N	Y	N	Y
CRKS24_00036	hypothetical protein	Insertion		None		N	Y	N	Y
CRKS24_00043	tRNA-Lys	Insertion		None		N	Y	N	N
CRKS24_00104	hypothetical protein	Split into two genes (CRKS27_00002-3)	MaMV-DC	YP_009217688.1	86% (E=1e-86)	Y	Y	Y	Split
CRKS25_00006	Transposase	Insertion	Microcystis	WP_008206903.1	96% (E = 0.0)	N	Y	Y	Unknown
CRKS25_00033	hypothetical protein	Multiple indels	MaMV-DC	YP_009217757.1	51% (E=2e-91)	Y	Y	Y	Y
CRKS25_00106	hypothetical protein	Insertion	MaMV-DC	YP_009217834.1	88% (E = 6e-109)	N	Unknown	Y	Y
CRKS25_00107	hypothetical protein	Insertion		None		N	Unknown	Y	Y
CRKS25_00108	hypothetical protein	Insertion		None		N	Y	Y	Y
CRKS25_00176	Tail collar protein	Divergent sequence				Y	Y	Y	Y
CRKS27_00014	hypothetical protein	Insertion				Y	N	N	Y
CRKS27_00015	hypothetical protein	Insertion	MaMV-DC	YP_009217751.1	79% (E=2e-53)	Y	N	N	Y
CRKS27_00078	Transposase	Divergence and split gene	Oscillatoria	WP_044196576.1	66% (2e-180)	Y	Y	Y	Split
CRKS27_00105	hypothetical protein	Divergence				Y	Unknown	Y	Truncated

Table 4.4: Differences in MaCRKS23 over 2013 time series.

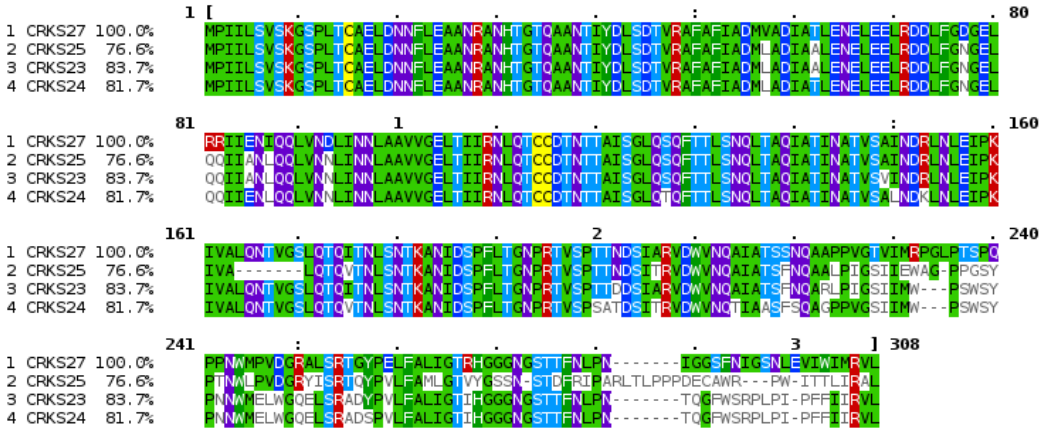

Figure 4.6: MUSCLE alignment of tail collar protein sequences assembled from Cheney time series. Colors indicate similarities based on amino acid sequence and properties.

Tail Collar Proteins

	Synonymous sites	Non-synonymous sites	dN	dS	dN/dS	Non-synonymous mutations	Synonymous mutations
CRKS23 vs. CRKS24	266.8	624.2	0.0326	0.0625	0.5213	20	17
CRKS23 vs. CRKS25	228.7	629.3	0.1089	0.0641	1.6975	69	15
CRKS23 vs. CRKS27	240.3	650.7	0.1094	0.0923	1.1854	71	22
CRKS24 vs. CRKS25	233.2	624.8	0.1269	0.1243	1.0209	79	29
CRKS24 vs. CRKS27	243.3	647.7	0.117	0.1318	0.8879	76	32
CRKS25 vs. CRKS27	235.5	631.5	0.1371	0.1111	1.2342	87	26
Mean					1.0912		
Sum						402	141

Capsid

	Synonymous sites	Non-synonymous sites	dN	dS	dN/dS	Non-synonymous mutations	Synonymous mutations
CRKS23 vs. CRKS24	405.5	905.5	0.0034	0.0371	0.0903	3	15
CRKS23 vs. CRKS25	241.6	1069.4	0	0	0.001	0	0
CRKS23 vs. CRKS27	298.6	1012.4	0	0.0072	0.001	0	2
CRKS24 vs. CRKS25	405.5	905.5	0.0034	0.0371	0.0903	3	15
CRKS24 vs. CRKS27	422.1	888.9	0.0034	0.0303	0.1131	3	13
CRKS25 vs. CRKS27	298.6	1012.4	0	0.0072	0.001	0	2
Mean					0.04945		
Sum						9	47

Table 4.5: dN/dS calculations for tail collar and capsid genes compared across the time series.

Chapter 5 Conclusion

The work presented in this dissertation analyzes the genomics of freshwater bloom-forming cyanobacteria, as well as associated heterotrophic bacteria and viruses. Second- and third-generation sequencing technologies were employed to generate novel genomic sequences that were subsequently compared to increase understanding of microbes in freshwater bloom habitats.

Three novel heterotrophic bacterial genomes, *Hyphomonadaceae* UKL13-1, *Betaproteobacterium* UKL13-2, and *Bacteroidetes* UKL13-3 were assembled from a long-read shotgun metagenome derived from a non-axenic *Aphanizomenon flos-aquae* culture grown in medium without nitrogen. The presence of an ammonium transporter gene, *amtB*, in *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2 suggests these bacteria are obtaining fixed nitrogen from *Aphanizomenon flos-aquae*, which likely releases fixed nitrogen in the form of ammonium, similar to previous reports [Ploug et al. 2010]. Based on gene content, *Hyphomonadaceae* UKL13-1 and *Betaproteobacterium* UKL13-2 both contain the genes necessary for aerobic anoxygenic photosynthesis, but not RuBisCO, which indicates their mixotrophic lifestyle.

Nine novel genomes from strains in the *Nostocaceae* family were sequenced and assembled to draft quality by our lab and Gregory Dick's lab at the University of Michigan. The relationships of these novel strains to all other sequenced *Nos-*

tocaceae genomes indicates that eight of these nine strains belong to one large clade. This clade is named the AAD clade because it consists entirely of globally distributed, bloom-forming *Anabaena*, *Aphanizomenon*, and *Dolichospermum* strains. Also, this clade separates into four separate groups of closely-related strains, and these groups have unique gene signatures relative to the remainder of the AAD clade involved in amino acid transport and retention, alternative nitrogen metabolism, and CRISPR-mediated defense. The novel genomes do not carry toxin synthesis genes, although *Anabaena* CRKS33 does contain genes for synthesis of the taste-and-odor compound geosmin. The distribution of toxin synthesis genes throughout the *Nostocaceae* family is scattered, and five separate toxin synthesis gene clusters are found in eight of the genomes. This indicates the lack of any pattern of descent for these toxin synthesis clusters as seen by others [Stucken et al. 2010; Jiang et al. 2012], and raises the questions of how these genes are retained or obtained by these strains, as well as what advantages are conferred by the production of each toxin.

Finally, two novel phage strains similar to the *Microcystis* phages Ma-LMM01 and MaMV-DC were assembled directly from environmental short-read shotgun metagenomes. These strains are part of a globally distributed *Microcystis* phage genome archetype, perhaps indicating their success infecting *Microcystis* strains worldwide. Comparison of these genomes indicates that host-like *nblA* and *phoH* genes are conserved, while genes putatively involved in escaping host defenses can be more variable (a Cas4-encoding gene is conserved across all genomes, while antitoxin genes are not). Comparison of fragmented genomes from an environmen-

tal metagenomic time-series revealed the presence of certain genes were variable. Also, the sequence encoding the tail collar gene, which encodes a structural component that sequesters the receptor-binding protein to control initiation of infection [Conley and Wood 1975; Fokine et al. 2013], was variable across this time-series. Together, these results suggest the existence of variation in the phage population, and raises the possibility of succession events where particular genomic variants may become fixed or dominant in the population over short time-spans.

The techniques employed throughout this work include long-read metagenomic sequencing as well as assembling and binning genomes from short-read environmental shotgun metagenomes. Several of the genomes analyzed were binned or completely assembled from environmental short-read shotgun metagenomes, including five draft-quality cyanobacterial genomes and two complete cyanophage genomes. This was possible due to novel techniques for parsing DNA sequencing data from complex microbial communities which have become available over the last decade. As a result, culturing these strains was not necessary. Additionally, since culturing bacteria can lead to genomic evolution, genomes extracted directly from environmental sequencing data are in their natural state.

If possible, completely assembling genomes obviates the need for binning fragmented genomes and assessing contamination. The three novel heterotrophic bacteria associated with *Aphanizomenon flos-aquae* were sequenced and assembled from a single mixed community culture with long reads alone. Since long-read sequencing is much lower-throughput than short-read sequencers, assembling these genomes was possible due to the low diversity of the culture. As a result, long-

read sequencing of relatively non-diverse communities towards assembling complete genomes is possible, and may become more feasible as the comparative depth of long-read sequencers increases.

This body of work includes advances in understanding the genomics of bloom-forming cyanobacteria and their associated organisms through applying current analytical techniques. Altogether, this lays the groundwork for genomics-based methods by which cyanobacterial blooms may be studied to better understand factors driving bloom formation and collapse.

Chapter 6 Contributions from authors

6.1 Chapter 2: Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture

Connor B. Driscoll and Theo W. Dreher conceived and designed the experimental plan, and wrote the manuscript with input from other authors. Connor B. Driscoll conducted most of the experiments and bioinformatic analyses. Timothy G. Otten initiated sequenced cultures, and provided extracted DNA for Illumina sequencing. Nathan M. Brown provided some data analysis scripts.

6.2 Chapter 3: Nine novel *Anabaena* and *Aphanizomenon* genome sequences reveals the existence of a closely-related clade of globally distributed, bloom-forming cyanobacteria within the *Nostocaceae* family

Connor B. Driscoll and Theo W. Dreher conceived and designed the experimental plan, with input from the remaining authors. Connor B. Driscoll conducted most of the experiments. Connor B. Driscoll and Theo W. Dreher wrote the

manuscript. Connor B. Driscoll, Nathan M. Brown, and Gregory J. Dick provided genome sequences for analysis. Timothy G. Otten performed DNA extractions and assembled metagenomes. Kevin Meyer performed core- and pan-genome analyses, and provided gene clusters. Yanbin Yin conducted secondary metabolite analysis. Zachary C. Landry performed phylogenomic analysis.

6.3 Chapter 4: Genome sequencing of two novel Ma-LMM01-like strains reveals patterns of conservation and divergence in a globally distributed *Microcystis* phage type

Connor B. Driscoll and Theo W. Dreher conceived and designed the experimental plan. Connor B. Driscoll conducted most of the experiments. Timothy G. Otten performed DNA extractions and arranged DNA metagenome datasets. Connor B. Driscoll performed bioinformatic analyses. Connor B. Driscoll and Theo W. Dreher wrote the manuscript.

Bibliography

Adam, B., Klawonn, I., Svedén, J. B., Bergkvist, J., Nahar, N., Walve, J., Littmann, S., Whitehouse, M. J., Lavik, G., Kuypers, M. M., et al. (2016). N_2-fixation, ammonium release and N-transfer to the microbial and classical food web within a plankton community. *The ISME journal*, 10(2):450–459.

Aharonovich, D. and Sher, D. (2016). Transcriptional response of *Prochlorococcus* to co-culture with a marine *Alteromonas*: differences between strains and the involvement of putative infochemicals. *The ISME journal*.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538.

Amin, S. A., Parker, M. S., and Armbrust, E. V. (2012). Interactions between diatoms and bacteria. *Microbiology and Molecular Biology Reviews*, 76(3):667–684.

Ammann, E. C. and Lynch, V. H. (1964). Purine metabolism by unicellular algae ii. adenine, hypoxanthine, and xanthine degradation by chlorella pyrenoidosa. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Nucleic Acids and Related Subjects*, 87(3):370–379.

Anderson, S. L. and McIntosh, L. (1991). Light-activated heterotrophic growth of the cyanobacterium *Synechocystis* sp. strain PCC 6803: a blue-light-requiring process. *Journal of Bacteriology*, 173(9):2761–2767.

Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G. M., Kodira, C. D., Kyrpides, N., Madupu, R., Markowitz, V., et al. (2008). Toward an online repository of Standard Operating Procedures (SOPs) for (meta) genomic annotation. *OMICS A Journal of Integrative Biology*, 12(2):137–141.

Ask, J., Karlsson, J., Persson, L., Ask, P., Byström, P., and Jansson, M. (2009). Whole-lake estimates of carbon flux through algae and bacteria in benthic and pelagic habitats of clear-water lakes. *Ecology*, 90(7):1923–1932.

Bagatini, I. L., Eiler, A., Bertilsson, S., Klaveness, D., Tessarolli, L. P., and Vieira, A. A. H. (2014). Host-specificity and dynamics in bacterial communities associated with bloom-forming freshwater phytoplankton. *PloS One*, 9(1):e85950.

Baran, R., Bowen, B. P., and Northen, T. R. (2011). Untargeted metabolic footprinting reveals a surprising breadth of metabolite uptake and release by *Synechococcus* sp. PCC 7002. *Molecular BioSystems*, 7(12):3200–3206.

Baran, R., Ivanova, N. N., Jose, N., Garcia-Pichel, F., Kyrpides, N. C., Gugger, M., and Northen, T. R. (2013). Functional genomics of novel secondary metabolites from diverse cyanobacteria using untargeted metabolomics. *Marine Drugs*, 11(10):3617–3631.

Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., et al. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, 30(7):701–707.

Beck, C., Knoop, H., Axmann, I. M., and Steuer, R. (2012). The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13(1):1.

Beckers, G., Bendt, A. K., Krämer, R., and Burkovski, A. (2004). Molecular identification of the urea uptake system and transcriptional analysis of urea transporter-and urease-encoding genes in *Corynebacterium glutamicum*. *Journal of Bacteriology*, 186(22):7645–7652.

Beltran, E. C. and Neilan, B. A. (2000). Geographical segregation of the neurotoxin-producing cyanobacterium *Anabaena circinalis*. *Applied and Environmental Microbiology*, 66(10):4468–4474.

Bentley, F. K., Luo, H., Dilbeck, P., Burnap, R. L., and Eaton-Rye, J. J. (2008). Effects of inactivating psbM and psbT on photodamage and assembly of photosystem II in *Synechocystis* sp. PCC 6803. *Biochemistry*, 47(44):11637–11646.

Berg, K. A., Lyra, C., Sivonen, K., Paulin, L., Suomalainen, S., Tuomi, P., and Rapala, J. (2009). High diversity of cultivable heterotrophic bacteria in association with cyanobacterial water blooms. *The ISME journal*, 3(3):314–325.

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., Roache-Johnson, K. H., Ding, H., Giovannoni, S. J., Rocap, G.,

Moore, L. R., and Chisholm, S. W. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data*, 1:140034.

Biller, S. J., Berube, P. M., Lindell, D., and Chisholm, S. W. (2015). *Prochlorococcus*: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13(1):13–27.

Black, K., Buikema, W. J., and Haselkorn, R. (1995). The *hglK* gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium *Anabaena* sp. strain PCC 7120. *Journal of Bacteriology*, 177(22):6440–6448.

Bolch, C. J., Orr, P. T., Jones, G. J., and Blackburn, S. I. (1999). Genetic, morphological, and toxicological variation among globally distributed strains of *Nodularia* (*Cyanobacteria*). *Journal of Phycology*, 35(2):339–355.

Bragg, L. and Tyson, G. W. (2014). Metagenomics using next-generation sequencing. *Environmental Microbiology: Methods and Protocols*, pages 183–201.

Bratbak, G., Egge, J. K., and Heldal, M. (1993). Viral mortality of the marine alga *Emiliania huxleyi* (*Haptophyceae*) and termination of algal blooms. *Marine Ecology Progress Series*.

Bright, D. and Walsby, A. (1999). The relationship between critical pressure and width of gas vesicles in isolates of *Planktothrix rubescens* from Lake Zürich. *Microbiology*, 145(10):2769–2775.

Brown, N. M., Mueller, R. S., Shepardson, J. W., Landry, Z. C., Morré, J. T., Maier, C. S., Hardy, F. J., and Dreher, T. W. (2016). Structural and functional analysis of the finished genome of the recently isolated toxic *Anabaena* sp. WA102. *BMC Genomics*, 17(1):1.

Brussaard, C. P. and Martinez, J. M. (2008). Algal bloom viruses. *Plant Viruses*, 2(1):1–13.

Buermans, H. and Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941.

Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., Kerfeld, C. A., Sivonen, K., Piel, J., and Gugger, M. (2014). Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics*, 15:977.

Cameron, J. C. and Pakrasi, H. B. (2010). Essential role of glutathione in acclimation to environmental and redox perturbations in the cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Physiology*, 154(4):1672–1685.

Campbell, E. L., Hagen, K. D., Chen, R., Risser, D. D., Ferreira, D. P., and Meeks, J. C. (2015). Genetic analysis reveals the identity of the photoreceptor for phototaxis in hormogonium filaments of *Nostoc punctiforme*. *Journal of Bacteriology*, 197(4):782–791.

Canfield, D. E. (2005). THE EARLY HISTORY OF ATMOSPHERIC OXYGEN: Homage to Robert M. Garrels. *Annual Review of Earth and Planetary Sciences*, 33:1–36.

Cao, H., Shimura, Y., Masanobu, K., and Yin, Y. (2014). Draft genome sequence of the toxic bloom-forming cyanobacterium *Aphanizomenon flos-aquae* NIES-81. *Genome Announcements*, 2(1):e00044–14.

Carmichael, W. W., Biggs, D. F., and Gorham, P. R. (1975). Toxicology and pharmacological action of *Anabaena flos-aquae* toxin. *Science*, 187(4176):542–544.

Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1):238.

Chandler, L. E., Bartsevich, V. V., and Pakrasi, H. B. (2003). Regulation of manganese uptake in *Synechocystis* 6803 by RfrA, a member of a novel family of proteins containing a repeated five-residues domain. *Biochemistry*, 42(18):5508–5514.

Chelikani, P., Fita, I., and Loewen, P. C. (2004). Diversity of structures and properties among catalases. *Cellular and Molecular Life Sciences CMLS*, 61(2):192–208.

Chen, F. and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Applied and Environmental Microbiology*, 68(5):2589–2594.

Chénard, C., Chan, A., Vincent, W., and Suttle, C. (2015). Polar freshwater cyanophage S-EIV1 represents a new widespread evolutionary lineage of phages. *The ISME journal*.

Cheung, M. Y., Liang, S., and Lee, J. (2013). Toxin-producing cyanobacteria in freshwater: a review of the problems, impact on drinking water safety, and efforts for protecting public health. *The Journal of Microbiology*, 51(1):1–10.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569.

Chlipala, G. E., Sturdy, M., Krunic, A., Lantvit, D. D., Shen, Q., Porter, K., Swanson, S. M., and Orjala, J. (2010). Cylindrocyclophanes with proteasome inhibitory activity from the Cyanobacterium *Nostoc* sp. *Journal of Natural Products*, 73(9):1529–1537.

Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287.

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E. J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*, 33(10):1053–1060.

Clokie, M. R. and Mann, N. H. (2006). Marine cyanophages and light. *Environmental Microbiology*, 8(12):2074–2082.

Comeau, A. M., Bertrand, C., Letarov, A., Tétart, F., and Krisch, H. (2007). Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology*, 362(2):384–396.

Conley, M. P. and Wood, W. B. (1975). Bacteriophage T4 whiskers: a rudimentary environment-sensing device. *Proceedings of the National Academy of Sciences*, 72(9):3701–3705.

Contreras-Moreira, B. and Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, 79(24):7696–7701.

Cooper, V. S., Schneider, D., Blot, M., and Lenski, R. E. (2001). Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *Journal of Bacteriology*, 183(9):2834–2841.

Cresawn, S. G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R. W., and Hatfull, G. F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics*, 12(1):395.

Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J., and Smith, A. G. (2005). Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature*, 438(7064):90–93.

D'Agostino, P. M., Song, X., Neilan, B. A., and Moffitt, M. C. (2016a). Proteogenomics of a saxitoxin-producing and non-toxic strain of *Anabaena circinalis* (cyanobacteria) in response to extracellular NaCl and phosphate depletion. *Environmental Microbiology*.

D'Agostino, P. M., Woodhouse, J. N., Makower, A. K., Yeung, A. C., Ongley, S. E., Micallef, M. L., Moffitt, M. C., and Neilan, B. A. (2016b). Advances in genomics, transcriptomics and proteomics of toxin-producing cyanobacteria. *Environmental Microbiology Reports*, 8(1):3–13.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403.

Darling, A. E., Jospin, G., Lowe, E., Matsen IV, F. A., Bik, H. M., and Eisen, J. A. (2014). Phylosift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS One*, 5(6):e11147.

Daubin, V., Gouy, M., and Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research*, 12(7):1080–1090.

Deatherage, D. E. and Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Engineering and Analyzing Multicellular Systems: Methods and Protocols*, pages 165–188.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496–503.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., and Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10(8):1.

Dittmann, E., Gugger, M., Sivonen, K., and Fewer, D. P. (2015). Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. *Trends in Microbiology*, 23(10):642–652.

Donia, M. S., Ravel, J., and Schmidt, E. W. (2008). A global assembly line to cyanobactins. *Nature Chemical Biology*, 4(6):341.

Dreher, T. W., Brown, N., Bozarth, C. S., Schwartz, A. D., Riscoe, E., Thrash, C., Bennett, S. E., Tzeng, S.-C., and Maier, C. S. (2011). A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. *Environmental Microbiology*, 13(7):1858–1874.

Duplessis, M. and Moineau, S. (2001). Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages. *Molecular Microbiology*, 41(2):325–336.

Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A., and Breitbart, M. (2013). A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology*, 13(1):1.

Eiler, A. and Bertilsson, S. (2007). Flavobacteria blooms in four eutrophic lakes: linking population dynamics of freshwater bacterioplankton to resource availability. *Applied and Environmental Microbiology*, 73(11):3511–3518.

Eiler, A., Olsson, J. A., and Bertilsson, S. (2006). Diurnal variations in the auto- and heterotrophic activity of cyanobacterial phycospheres (*Gloeotrichia echinulata*) and the identity of attached bacteria. *Freshwater Biology*, 51(2):298–311.

Elovaara, H., Huusko, T., Maksimow, M., Elima, K., Yegutkin, G. G., Skurnik, M., Dobrindt, U., Siitonen, A., McPherson, M. J., Salmi, M., et al. (2015). Primary amine oxidase of *Escherichia coli* is a metabolic enzyme that can use a human leukocyte molecule as a substrate. *PloS One*, 10(11):e0142367.

Escobar-Zepeda, A., de León, A. V.-P., and Sanchez-Flores, A. (2015). The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, 6.

Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., and Tyson, G. W. (2015). Methane metabolism in the archaeal phylum *Bathyarchaeota* revealed by genome-centric metagenomics. *Science*, 350(6259):434–438.

Falconer, I. R. (1999). An overview of problems caused by toxic blue–green algae (cyanobacteria) in drinking and recreational water. *Environmental Toxicology*, 14(1):5–12.

Felsenstein, J. (2005). PHYLIP: Phylogenetic inference program, version 3.6. *University of Washington, Seattle*.

Fichot, E. B. and Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*, 1(1):10.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glockner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S. A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T.,

San Gil, I., Wilson, G., and Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–547.

Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., and Salmond, G. P. (2009). The phage abortive infection system, ToxIN, functions as a protein–RNA toxin–antitoxin pair. *Proceedings of the National Academy of Sciences*, 106(3):894–899.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, page gkr367.

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S., and Stoeckert, C. J. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new Ortholog groups. *Current Protocols in Bioinformatics*, pages 6–12.

Fokine, A., Zhang, Z., Kanamaru, S., Bowman, V. D., Aksyuk, A. A., Arisaka, F., Rao, V. B., and Rossmann, M. G. (2013). The molecular architecture of the bacteriophage T4 neck. *Journal of Molecular Biology*, 425(10):1731–1744.

Forde, B. M., Zakour, N. L. B., Stanton-Cook, M., Phan, M.-D., Totsika, M., Peters, K. M., Chan, K. G., Schembri, M. A., Upton, M., and Beatson, S. A. (2014). The complete genome sequence of *Escherichia coli* ec958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* o25b: H4-st131 clone.

Foster, R. A., Kuypers, M. M., Vagner, T., Paerl, R. W., Musat, N., and Zehr, J. P. (2011). Nitrogen fixation and transfer in open ocean diatom–cyanobacterial symbioses. *The ISME journal*, 5(9):1484–1493.

Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G., McHardy, A. C., Nederbragt, A. J., and Pope, P. B. (2015). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *bioRxiv*, page 026922.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2014). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, page gku1223.

Gao, E.-B., Gui, J.-F., and Zhang, Q.-Y. (2012). A novel cyanophage with a cyanobacterial nonbleaching protein A gene in the genome. *Journal of Virology*, 86(1):236–245.

Garcia-Pichel, F., Lopez-Cortes, A., and Nubel, U. (2001). Phylogenetic and morphological diversity of cyanobacteria in soil desert crusts from the Colorado plateau. *Applied and Environmental Microbiology*, 67(4):1902–1910.

Garrity, G., Staley, J. T., Boone, D. R., De Vos, P., Goodfellow, M., Rainey, F. A., Schleifer, K.-H., Brenner, D. J., and Krieg, N. R. (2006). *Bergey's Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria*. Springer Science & Business Media.

Garrity, G. M., Bell, J. A., and Lilburn, T. (2005a). *Bergey's Manual of Systematic Bacteriology: Volume Two The Proteobacteria Part C The Alpha-, Beta-, Delta-, and Epsilonproteobacteria*, chapter Class I. Alphaproteobacteria class. nov., pages 1–574. Springer US, Boston, MA.

Garrity, G. M., Bell, J. A., and Lilburn, T. (2005b). *Bergey's Manual of Systematic Bacteriology: Volume Two The Proteobacteria Part C The Alpha-, Beta-, Delta-, and Epsilonproteobacteria*, chapter Class II. Betaproteobacteria class. nov., page 575. Springer US, Boston, MA.

Garrity, G. M., Bell, J. A., and Lilburn, T. (2005c). *Bergey's Manual of Systematic Bacteriology: Volume Two The Proteobacteria Part C The Alpha-, Beta-, Delta-, and Epsilonproteobacteria*, chapter Order III. Rhodobacterales class. nov., page 161. Springer US, Boston, MA.

Geerse, R., Izzo, F., and Postma, P. (1989). The PEP: fructose phosphotransferase system in *Salmonella typhimurium*: FPr combines enzyme IIIFru and pseudo-HPr activities. *Molecular and General Genetics MGG*, 216(2-3):517–525.

Ghai, R., Martin-Cuadrado, A.-B., Molto, A. G., Heredia, I. G., Cabrera, R., Martin, J., Verdu, M., Deschamps, P., Moreira, D., López-García, P., et al. (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *The ISME journal*, 4(9):1154–1166.

Gilbert, J. A. and Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Annual Review of Marine Science*, 3:347–371.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010(1):pdb–prot5368.

Gómez, F., Furuya, K., and Takeda, S. (2005). Distribution of the cyanobacterium *Richelia intracellularis* as an epiphyte of the diatom *Chaetoceros compressus* in the western Pacific Ocean. *Journal of Plankton Research*, 27(4):323–330.

Gone, S. and Nicholson, A. W. (2012). Bacteriophage T7 protein kinase: site of inhibitory autophosphorylation, and use of dephosphorylated enzyme for efficient modification of protein *in vitro*. *Protein Expression and Purification*, 85(2):218–223.

González, J. M., Simó, R., Massana, R., Covert, J. S., Casamayor, E. O., Pedrós-Alió, C., and Moran, M. A. (2000). Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Applied and Environmental Microbiology*, 66(10):4237–4246.

Gregor, I., Dröge, J., Schirmer, M., Quince, C., and McHardy, A. (2014). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *arXiv preprint arXiv:1406.7123*.

Grossart, H.-P., Czub, G., and Simon, M. (2006). Algae–bacteria interactions and their effects on aggregation and organic matter flux in the sea. *Environmental Microbiology*, 8(6):1074–1084.

Grossman, A. R., Schaefer, M. R., Chiang, G. G., and Collier, J. L. (1993). The phycobilisome, a light-harvesting complex responsive to environmental conditions. *Microbiological Reviews*, 57(3):725–749.

Grover, J. P. (2000). Resource competition and community structure in aquatic micro-organisms: experimental studies of algae and bacteria along a gradient of organic carbon to inorganic phosphorus supply. *Journal of Plankton Research*, 22(8):1591–1610.

Gugger, M., Lyra, C., Henriksen, P., Coute, A., Humbert, J. F., and Sivonen, K. (2002). Phylogenetic comparison of the cyanobacterial genera *Anabaena* and *Aphanizomenon*. *International Journal of Systematic and Evolutionary Microbiology*, 52(Pt 5):1867–1880.

Hall, A. R., Scanlan, P. D., Morgan, A. D., and Buckling, A. (2011). Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters*, 14(7):635–642.

Han, S.-J., Lee, B. C., Yim, S. H., Gladyshev, V. N., and Lee, S.-R. (2014). Characterization of mammalian selenoprotein o: a redox-active mitochondrial protein. *PloS One*, 9(4):e95518.

Harada, K.-i. (2004). Production of secondary metabolites by freshwater cyanobacteria. *Chemical and Pharmaceutical Bulletin*, 52(8):889–899.

Harrison, J. and Studholme, D. J. (2014). Recently published *Streptomyces* genome sequences. *Microbial Biotechnology*, 7(5):373–380.

Hatfull, G. F. (2010). Mycobacteriophages: genes and genomes. *Annual Review of Microbiology*, 64:331–356.

Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C.-C., Weber, R. J., Patel, M. C., Germane, K. L., Edgar, R. H., et al. (2010). Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *Journal of Molecular Biology*, 397(1):119–143.

Havaux, M., Eymery, F., Porfirova, S., Rey, P., and Dörmann, P. (2005). Vitamin E protects against photoinhibition and photooxidative stress in *Arabidopsis thaliana*. *The Plant Cell*, 17(12):3451–3469.

Heller, K. J. (1984). Identification of the phage gene for host receptor specificity by analyzing hybrid phages of T5 and BF23. *Virology*, 139(1):11–21.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the worlds a phage. *Proceedings of the National Academy of Sciences*, 96(5):2192–2197.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–467.

Hilton, J. A. (2014). Ecology and evolution of diatom-associated cyanobacteria through genetic analyses.

Hilton, J. A., Foster, R. A., Tripp, H. J., Carter, B. J., Zehr, J. P., and Villareal, T. A. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nature Communications*, 4:1767.

Hiorns, W. D., Methe, B. A., Nierzwicki-Bauer, S. A., and Zehr, J. P. (1997). Bacterial diversity in Adirondack mountain lakes as revealed by 16S rRNA gene sequences. *Applied and Environmental Microbiology*, 63(7):2957–2960.

Holm, L. and Sander, C. (1994). Structural similarity of plant chitinase and lysozymes from animals and phage. *FEBS Letters*, 340(1-2):129–132.

Honda, T., Takahashi, H., Sako, Y., and Yoshida, T. (2014). Gene expression of *Microcystis aeruginosa* during infection of cyanomyovirus Ma-LMM01. *Fisheries Science*, 80(1):83–91.

Hooton, S. and Connerton, I. F. (2015). *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Frontiers in Microbiology*, 5:744.

Hooton, S. P., Brathwaite, K. J., and Connerton, I. F. (2016). The bacteriophage carrier state of *Campylobacter jejuni* features changes in host non-coding RNAs and the acquisition of new host-derived CRISPR spacer sequences. *Frontiers in Microbiology*, 7.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. *Nature Microbiology*, 1:16048.

Huisman, J., Sharples, J., Stroom, J. M., Visser, P. M., Kardinaal, W. E. A., Verspagen, J. M., and Sommeijer, B. (2004). Changes in turbulent mixing shift competition for light between phytoplankton species. *Ecology*, 85(11):2960–2970.

Ikawa, M., Wegener, K., Foxall, T. L., and Sasner, J. J. (1982). Comparison of the toxins of the blue-green alga *Aphanizomenon flos-aquae* with the *Gonyaulax* toxins. *Toxicon*, 20(4):747–752.

Ikeuchi, M., Eggers, B., Shen, G., Webber, A., Yu, J., Hirano, A., Inoue, Y., and Vermaas, W. (1991). Cloning of the psbK gene from *Synechocystis* sp. PCC

6803 and characterization of photosystem II in mutants lacking PSII-K. *Journal of Biological Chemistry*, 266(17):11111–11115.

Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., and Tyson, G. W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603.

Itou, Y., Suzuki, S., Ishida, K., and Murakami, M. (1999). Anabaenopeptins G and H, potent carboxypeptidase A inhibitors from the cyanobacterium *Oscillatoria agardhii* (NIES-595). *Bioorganic & Medicinal Chemistry Letters*, 9(9):1243–1246.

Jacquet, S., Heldal, M., Iglesias-Rodriguez, D., Larsen, A., Wilson, W., and Bratbak, G. (2002). Flow cytometric analysis of an *Emiliana huxleyi* bloom terminated by viral infection. *Aquatic Microbial Ecology*, 27(2):111–124.

Jansen, W., Bolm, M., Balling, R., Chhatwal, G., and Schnabel, R. (2002). Hydrogen peroxide-mediated killing of *Caenorhabditis elegans* by *Streptococcus pyogenes*. *Infection and Immunity*, 70(9):5202–5207.

Jiang, Y., Xiao, P., Yu, G., Sano, T., Pan, Q., and Li, R. (2012). Molecular basis and phylogenetic implications of deoxycylindrospermopsin biosynthesis in the cyanobacterium *Raphidiopsis curvata*. *Applied and Environmental Microbiology*, 78(7):2256–2263.

Jüttner, F. and Watson, S. B. (2007). Biochemical and ecological control of geosmin and 2-methylisoborneol in source waters. *Applied and Environmental Microbiology*, 73(14):4395–4406.

Kaas, R. S., Friis, C., Ussery, D. W., and Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, 13(1):577.

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

Kanehisa, M., Sato, Y., and Morishima, K. (2015). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*.

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

Karl, D., Letelier, R., Tupas, L., Dore, J., Christian, J., and Hebel, D. (1997). The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature*, 388(6642):533–538.

Katoh, H., Hagino, N., Grossman, A. R., and Ogawa, T. (2001). Genes essential to iron transport in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *Journal of Bacteriology*, 183(9):2779–2784.

Kehr, J.-C. and Dittmann, E. (2015). Biosynthesis and function of extracellular glycans in cyanobacteria. *Life*, 5(1):164–180.

Kimura, S., Sako, Y., and Yoshida, T. (2013). Rapid gene diversification of *Microcystis* cyanophages revealed by long-and short-term genetic analysis of the tail sheath gene in a natural pond. *Applied and Environmental Microbiology*, pages AEM–03751.

Kimura, S., Yoshida, T., Hosoda, N., Honda, T., Kuno, S., Kamiji, R., Hashimoto, R., and Sako, Y. (2012). Diurnal infection patterns and impact of *Microcystis* cyanophages in a Japanese pond. *Applied and Environmental Microbiology*, 78(16):5805–5811.

Knerr, P. J. and van der Donk, W. A. (2012). Chemical synthesis and biological activity of analogues of the lantibiotic epilancin 15X. *Journal of the American Chemical Society*, 134(18):7648–7651.

Koga, M., Otsuka, Y., Lemire, S., and Yonesaki, T. (2011). *Escherichia coli rnlA* and *rnlB* compose a novel toxin–antitoxin system. *Genetics*, 187(1):123–130.

Komárek, J. (2016). A polyphasic approach for the taxonomy of cyanobacteria: principles and applications. *European Journal of Phycology*, pages 1–8.

Koonin, E. V. and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719.

Koren, S., Harhay, G. P., Smith, T., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., and Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, 14(9):R101.

Koren, S. and Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., et al. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700.

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLoS Genetics*, 8(6):e1002787.

Köster, W. (2001). ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B 12. *Research in Microbiology*, 152(3):291–301.

Krieg, N. R., Ludwig, W., Euzéby, J., and Whitman, W. B. (2010). *Bergey's Manual of Systematic Bacteriology: Volume Four The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, chapter Phylum XIV. Bacteroidetes phyl. nov., pages 25–469. Springer New York, New York, NY.

Kumar, S. R. and Imlay, J. A. (2013). How *Escherichia coli* tolerates profuse hydrogen peroxide formation by a catabolic pathway. *Journal of Bacteriology*, 195(20):4569–4579.

Kuno, S., Yoshida, T., Kaneko, T., and Sako, Y. (2012). Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. *Applied and Environmental Microbiology*, 78(15):5353–5360.

Kurtz, D. M. (2006). Avoiding high-valent iron intermediates: superoxide reductase and rubrerythrin. *Journal of Inorganic Biochemistry*, 100(4):679–693.

Latifi, A., Ruiz, M., Jeanjean, R., and Zhang, C.-C. (2007). PrxQ-A, a member of the peroxiredoxin Q family, plays a major role in defense against oxidative stress in the cyanobacterium *Anabaena* sp. strain PCC7120. *Free Radical Biology and Medicine*, 42(3):424–431.

Latifi, A., Ruiz, M., and Zhang, C.-C. (2009). Oxidative stress in cyanobacteria. *FEMS Microbiology Reviews*, 33(2):258–278.

Lawrence, J. G., Ochman, H., and Hartl, D. (1992). The evolution of insertion sequences within enteric bacteria. *Genetics*, 131(1):9–20.

Leão, P. N., Vasconcelos, M. T. S., and Vasconcelos, V. M. (2009). Allelopathy in freshwater cyanobacteria. *Critical Reviews in Microbiology*, 35(4):271–282.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., and Schatz, M. (2014). Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*, page 006395.

Lee, K. B., Liu, C. T., Anzai, Y., Kim, H., Aono, T., and Oyaizu, H. (2005). The hierarchical system of the '*Alphaproteobacteria*': description of *Hyphomonadaceae* fam. nov., *Xanthobacteraceae* fam. nov. and *Erythrobacteraceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology*, 55(Pt 5):1907–1919.

Leikoski, N., Fewer, D. P., Jokela, J., Wahlsten, M., Rouhiainen, L., and Sivonen, K. (2010). Highly diverse cyanobactins in strains of the genus *Anabaena*. *Applied and Environmental Microbiology*, 76(3):701–709.

Leinonen, R., Sugawara, H., and Shumway, M. (2010). The sequence read archive. *Nucleic Acids Research*, page gkq1019.

Lemarchand, C., Jardillier, L., Carrias, J.-F., Richardot, M., Debroas, D., Sime-Ngando, T., and Amblard, C. (2006). Community composition and activity of prokaryotes associated to detrital particles in two contrasting lake ecosystems. *FEMS Microbiology Ecology*, 57(3):442–451.

Lemire, S., Figueroa-Bossi, N., and Bossi, L. (2011). Bacteriophage crosstalk: coordination of prophage induction by trans-acting antirepressors. *PLoS Genetics*, 7(6):e1002149.

Lenart, A. and Pawłowski, K. (2013). Intersection of selenoproteins and kinase signalling. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1834(7):1279–1284.

Lertsethtakarn, P., Ottemann, K. M., and Hendrixson, D. R. (2011). Motility and chemotaxis in *Campylobacter* and *Helicobacter*. *Annual Review of Microbiology*, 65:389–410.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Lin, K.-H., Liao, B.-Y., Chang, H.-W., Huang, S.-W., Chang, T.-Y., Yang, C.-Y., Wang, Y.-B., Lin, Y.-T. K., Wu, Y.-W., Tang, S.-L., et al. (2015). Metabolic characteristics of dominant microbes and key rare species from an acidic hot spring in Taiwan revealed by metagenomics. *BMC Genomics*, 16(1):1.

Lind, L., Shukla, V. K., Nyhus, K., and Pakrasi, H. (1993). Genetic and immunological analyses of the cyanobacterium *Synechocystis* sp. PCC 6803 show that the protein encoded by the psbJ gene regulates the number of photosystem II centers in thylakoid membranes. *Journal of Biological Chemistry*, 268(3):1575–1579.

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

Liu, P., Tian, G.-l., Lee, K.-S., Wong, M.-S., and Ye, Y.-h. (2002). Full enzymatic synthesis of a precursor of bioactive pentapeptide OGP (10-14) in organic solvents. *Tetrahedron Letters*, 43(13):2423–2425.

Liu, X., Kong, S., Shi, M., Fu, L., Gao, Y., and An, C. (2008). Genomic analysis of freshwater cyanophage Pf-WMP3 infecting cyanobacterium *Phormidium foveolarum*: the conserved elements for a phage. *Microbial Ecology*, 56(4):671–680.

Liu, X., Shi, M., Kong, S., Gao, Y., and An, C. (2007). Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium *Phormidium foveolarum*: complete genome sequence and DNA translocation. *Virology*, 366(1):28–39.

Louati, I., Pascault, N., Debroas, D., Bernard, C., Humbert, J.-F., and Leloup, J. (2015). Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PloS One*, 10(11).

Lyra, C., Suomalainen, S., Gugger, M., Vezie, C., Sundman, P., Paulin, L., and Sivonen, K. (2001). Molecular characterization of planktic cyanobacteria of *Anabaena*, *Aphanizomenon*, *Microcystis* and *Planktothrix* genera. *International Journal of Systematic and Evolutionary Microbiology*, 51(Pt 2):513–526.

MacKintosh, C., Beattie, K. A., Klumpp, S., Cohen, P., and Codd, G. A. (1990). Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. *FEBS Letters*, 264(2):187–192.

Maeda, H., Sakuragi, Y., Bryant, D. A., and DellaPenna, D. (2005). Tocopherols protect *Synechocystis* sp. strain PCC 6803 from lipid peroxidation. *Plant Physiology*, 138(3):1422–1435.

Mankiewicz-Boczek, J., Jaskulska, A., Pawelczyk, J., Gagala, I., Serwecinska, L., and Dziadek, J. (2016). Cyanophages Infection of Microcystis Bloom in Lowland Dam Reservoir of Sulejw, Poland. *Microb. Ecol.*, 71(2):315–325.

Mann, N. H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiology Reviews*, 27(1):17–34.

Mann, N. H. and Clokie, M. R. (2012). Cyanophages. In *Ecology of Cyanobacteria II*, pages 535–557. Springer.

Mann, N. H., Clokie, M. R., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A., and Krisch, H. (2005). The genome of S-PM2, a photosynthetic T4-type bacteriophage that infects marine *Synechococcus* strains. *Journal of Bacteriology*, 187(9):3188–3200.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Reviews of Genomics and Human Genetics*, 9:387–402.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, 40(D1):D115–D122.

Marmen, S., Aharonovich, D., Grossowicz, M., Blank, L., Yacobi, Y. Z., and Sher, D. J. (2016). Distribution and habitat specificity of potentially-toxic *Microcystis* across climate, land, and water use gradients. *Frontiers in Microbiology*, 7.

Meeks, J. C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., and Atlas, R. (2001). An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynthesis Research*, 70(1):85–106.

Méndez-García, C., Peláez, A. I., Mesa, V., Sánchez, J., Golyshina, O. V., and Ferrer, M. (2015). Microbial diversity and metabolic networks in acid mine drainage habitats. *Frontiers in Microbiology*, 6:475.

Mikulic, M. (2013). *Knock-out mutants of respiratory terminal oxidases in the cyanobacterium Anabaena sp. strain PCC 7120*. PhD thesis, uniwien.

Millard, A. D., Zwirglmaier, K., Downey, M. J., Mann, N. H., and Scanlan, D. J. (2009). Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. *Environmental Microbiology*, 11(9):2370–2387.

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLoS Genetics*, 9(12):e1003987.

Mlouka, A., Comte, K., Castets, A.-M., Bouchier, C., and de Marsac, N. T. (2004). The gas vesicle gene cluster from *Microcystis aeruginosa* and DNA rearrangements that lead to loss of cell buoyancy. *Journal of Bacteriology*, 186(8):2355–2365.

Moran, M. A. and Miller, W. L. (2007). Resourceful heterotrophs make the most of light in the coastal ocean. *Nature Reviews Microbiology*, 5(10):792–800.

Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264.

Morris, B. E., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial syntrophy: interaction for the common good. *FEMS Microbiology Reviews*, 37(3):384–406.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, 3(2):e00036–12.

Mosher, J. J., Bowman, B., Bernberg, E. L., Shevchenko, O., Kan, J., Korlach, J., and Kaplan, L. A. (2014). Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *Journal of Microbiological Methods*, 104:59–60.

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

Murakami, M., Suzuki, S., Itou, Y., Kodani, S., and Ishida, K. (2000). New anabaenopeptins, potent carboxypeptidase-A inhibitors from the cyanobacterium *Aphanizomenon flos-aquae*. *Journal of Natural Products*, 63(9):1280–1282.

Myers, J. and Kratz, W. (1955). Relations between pigment content and photosynthetic characteristics in a blue-green alga. *The Journal of General Physiology*, 39(1):11–22.

Nakamura, G., Kimura, S., Sako, Y., and Yoshida, T. (2014). Genetic diversity of *Microcystis* cyanophages in two different freshwater environments. *Archives of Microbiology*, 196(6):401–409.

Neumann, M., Mittelstädt, G., Seduk, F., Iobbi-Nivol, C., and Leimkühler, S. (2009). Moca is a specific cytidylyltransferase involved in molybdopterin cytosine dinucleotide biosynthesis in *Escherichia coli*. *Journal of Biological Chemistry*, 284(33):21891–21898.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews*, 75(1):14–49.

Nishiwaki-Matsushima, R., Ohta, T., Nishiwaki, S., Suganuma, M., Kohyama, K., Ishikawa, T., Carmichael, W. W., and Fujiki, H. (1992). Liver tumor promotion by the cyanobacterial cyclic peptide toxin microcystin-LR. *Journal of Cancer Research and Clinical Oncology*, 118(6):420–424.

Ohmori, M., Ikeuchi, M., Sato, N., Wolk, P., Kaneko, T., Ogawa, T., Kanehisa, M., Goto, S., Kawashima, S., Okamoto, S., et al. (2001). Characterization of genes encoding multi-domain proteins in the genome of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Research*, 8(6):271–284.

Otsuka, Y. and Yonesaki, T. (2012). Dmd of bacteriophage T4 functions as an antitoxin against *Escherichia coli* LsoA and RnlA toxins. *Molecular Microbiology*, 83(4):669–681.

Otten, T. G. and Paerl, H. W. (2015). Health effects of toxic cyanobacteria in us drinking and recreational waters: our current understanding and proposed direction. *Current Environmental Health Reports*, 2(1):75–84.

Ou, T., Gao, X.-C., Li, S.-H., and Zhang, Q.-Y. (2015a). Genome analysis and gene *nblA* identification of *Microcystis aeruginosa* myovirus (MaMV-DC) reveal the evidence for horizontal gene transfer events between cyanomyovirus and host. *Journal of General Virology*, 96(12):3681–3697.

Ou, T., Li, S., Liao, X., and Zhang, Q. (2013). Cultivation and characterization of the MaMV-DC cyanophage that infects bloom-forming cyanobacterium *Microcystis aeruginosa*. *Virologica Sinica*, 28(5):266–271.

Ou, T., Liao, X.-Y., Gao, X.-C., Xu, X.-D., and Zhang, Q.-Y. (2015b). Unraveling the genome structure of cyanobacterial podovirus A-4L with long direct terminal repeats. *Virus Research*, 203:4–9.

Oneil, J., Davis, T. W., Burford, M. A., and Gobler, C. (2012). The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. *Harmful Algae*, 14:313–334.

Paerl, H. W., Fulton, R. S., Moisander, P. H., and Dyble, J. (2001). Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *The Scientific World Journal*, 1:76–113.

Paerl, H. W. and Otten, T. G. (2013). Blooms bite the hand that feeds them. *Science*, 342(6157):433–434.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055.

Partensky, F., Blanchot, J., and Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. *Bulletin-Institut Oceanographique Monaco-Numero Special*, pages 457–476.

Pecota, D. C. and Wood, T. K. (1996). Exclusion of T4 phage by the *hok/sok* killer locus from plasmid R1. *Journal of Bacteriology*, 178(7):2044–2050.

Peduzzi, P., Gruber, M., Gruber, M., and Schagerl, M. (2014). The virus's tooth: cyanophages affect an African flamingo population in a bottom-up cascade. *The ISME journal*, 8(6):1346.

Pei, J., Kim, B.-H., and Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, 36(7):2295–2300.

Peng, Y., Leung, H. C., Yiu, S.-M., and Chin, F. Y. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Pereira, S., Zille, A., Micheletti, E., Moradas-Ferreira, P., De Philippis, R., and Tamagnini, P. (2009). Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. *FEMS Microbiology Reviews*, 33(5):917–941.

Pereira, S. B., Mota, R., Vieira, C. P., Vieira, J., and Tamagnini, P. (2015). Phylum-wide analysis of genes/proteins related to the last steps of assembly and export of extracellular polymeric substances (EPS) in cyanobacteria. *Scientific Reports*, 5.

Pernthaler, J., Posch, T., Simek, K., Vrba, J., Pernthaler, A., Glöckner, F. O., Nübel, U., Psenner, R., and Amann, R. (2001). Predator-specific enrichment of actinobacteria from a cosmopolitan freshwater clade in mixed continuous culture. *Applied and Environmental Microbiology*, 67(5):2145–2155.

Pernthaler, J., Zöllner, E., Warnecke, F., and Jürgens, K. (2004). Bloom of filamentous bacteria in a mesotrophic lake: identity and potential controlling mechanism. *Applied and Environmental Microbiology*, 70(10):6272–6281.

Picossi, S., Montesinos, M. L., Pernil, R., Lichtlé, C., Herrero, A., and Flores, E. (2005). ABC-type neutral amino acid permease N-I is required for optimal diazotrophic growth and is repressed in the heterocysts of *Anabaena* sp. strain PCC 7120. *Molecular Microbiology*, 57(6):1582–1592.

Ploug, H., Musat, N., Adam, B., Moraru, C. L., Lavik, G., Vagner, T., Bergman, B., and Kuypers, M. M. (2010). Carbon and nitrogen fluxes associated with the cyanobacterium *Aphanizomenon* sp. in the Baltic Sea. *The ISME journal*, 4(9):1215–1223.

Pope, W. H., Ferreira, C. M., Jacobs-Sera, D., Benjamin, R. C., Davis, A. J., DeJong, R. J., Elgin, S. C., Guilfoile, F. R., Forsyth, M. H., Harris, A. D., et al. (2011). Cluster K mycobacteriophages: insights into the evolutionary origins of mycobacteriophage TM4. *PLoS One*, 6(10):e26750.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3):e9490.

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829.

Puigbo, P., Wolf, Y. I., and Koonin, E. V. (2010). The tree and net components of prokaryote evolution. *Genome Biology and Evolution*, 2:745–756.

Punetha, A., Sivathanu, R., and Anand, B. (2014). Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type IC system. *Nucleic Acids Research*, 42(6):3846–3856.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, page gks1219.

Rainey, F. A., Ward-Rainey, N. L., Janssen, P. H., Hippe, H., and Stackebrandt, E. (1996). *Clostridium paradoxum* DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences. *Microbiology*, 142 (Pt 8):2087–2095.

Rajaniemi, P., Hrouzek, P., Kastovska, K., Willame, R., Rantala, A., Hoffmann, L., Komarek, J., and Sivonen, K. (2005). Phylogenetic and morphological evaluation of the genera *Anabaena*, *Aphanizomenon*, *Trichormus* and *Nostoc (Nostocales*, Cyanobacteria). *International Journal of Systematic and Evolutionary Microbiology*, 55(Pt 1):11–26.

Rakhuba, D., Kolomiets, E., Dey, E. S., and Novik, G. (2010). Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. *Polish Journal of Microbiology*, 59(3):145–155.

Ran, L., Huang, F., Ekman, M., Klint, J., and Bergman, B. (2007). Proteomic analyses of the photoauto-and diazotrophically grown cyanobacterium *Nostoc* sp. PCC 73102. *Microbiology*, 153(2):608–618.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J. A., Ininbergs, K., Zheng, W.-W., Lapidus, A., Lowry, S., Haselkorn, R., and Bergman, B. (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS One*, 5(7):e11486.

Rattray, J. E., van de Vossenberg, J., Jaeschke, A., Hopmans, E. C., Wakeham, S. G., Lavik, G., Kuypers, M. M., Strous, M., Jetten, M. S., Schouten, S., et al. (2010). Impact of temperature on ladderane lipid distribution in anammox bacteria. *Applied and Environmental Microbiology*, 76(5):1596–1603.

Robbertse, B., Yoder, R. J., Boyd, A., Reeves, J., and Spatafora, J. W. (2011). Hal: an automated pipeline for phylogenetic analyses of genomic data. *PLoS Currents Tree of Life*.

Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2009). REBASEa database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, page gkp874.

Robertson, E. S. (2011). Survival of the fittest: a role for phage-encoded eukaryotic-like kinases. *Molecular Microbiology*, 82(3):539–541.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T. F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836.

Rohrlack, T., Christoffersen, K., Kaebernick, M., and Neilan, B. A. (2004). Cyanobacterial protease inhibitor microviridin J causes a lethal molting disruption in *Daphnia pulicaria*. *Applied and Environmental Microbiology*, 70(8):5047–5050.

Rouhiainen, L., Paulin, L., Suomalainen, S., Hyytiainen, H., Buikema, W., Haselkorn, R., and Sivonen, K. (2000). Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in *Anabaena* strain 90. *Molecular Microbiology*, 37(1):156–167.

Rozon, R. and Short, S. (2013). Complex seasonality observed amongst diverse phytoplankton viruses in the bay of quinte, an embayment of lake ontario. *Freshwater Biology*, 58(12):2648–2663.

Ruby, J. G., Bellare, P., and DeRisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3: Genes—Genomes— Genetics*, 3(5):865–880.

Rzymski, P., Poniedziałek, B., Kokociński, M., Jurczak, T., Lipski, D., and Wiktorowicz, K. (2014). Interspecific allelopathy in cyanobacteria: Cylindrospermopsin and *Cylindrospermopsis raciborskii* effect on the growth and metabolism of *Microcystis aeruginosa*. *Harmful Algae*, 35:1–8.

Sakata, S., Mizusawa, N., Kubota-Kawai, H., Sakurai, I., and Wada, H. (2013). Psb28 is involved in recovery of photosystem II at high temperature in *Synechocystis* sp. PCC 6803. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1827(1):50–59.

Sambrook, J. and Russell, D. W. (2006). Purification of nucleic acids by extraction with phenol: chloroform. *Cold Spring Harbor Protocols*, 2006(1):pdb–prot4455.

Sapp, J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiology and Molecular Biology Reviews*, 69(2):292–305.

Schäfer, L., Vioque, A., and Sandmann, G. (2005). Functional *in situ* evaluation of photosynthesis-protecting carotenoids in mutants of the cyanobacterium *Synechocystis* PCC6803. *Journal of Photochemistry and Photobiology B: Biology*, 78(3):195–201.

Schuergers, N., Lenn, T., Kampmann, R., Meissner, M. V., Esteves, T., Temerinac-Ott, M., Korvink, J. G., Lowe, A. R., Mullineaux, C. W., and Wilde, A. (2016). Cyanobacteria use micro-optics to sense light direction. *Elife*, 5:e12620.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, page btu153.

Self, W. T. (2002). Regulation of purine hydroxylase and xanthine dehydrogenase from *Clostridium purinolyticum* in response to purines, selenium, and molybdenum. *Journal of Bacteriology*, 184(7):2039–2044.

Selva, L., Viana, D., Regev-Yochay, G., Trzcinski, K., Corpa, J. M., Novick, R. P., Penadés, J. R., et al. (2009). Killing niche competitors by remote-control bacteriophage induction. *Proceedings of the National Academy of Sciences*, 106(4):1234–1238.

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120.

Shi, T. and Falkowski, P. G. (2008). Genome evolution in cyanobacteria: the stable core and the variable shell. *Proceedings of the National Academy of Sciences*, 105(7):2510–2515.

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K. W., Han, C. S., Rubin, E. M., Eisen, J. A., Woyke, T., Gugger, M., and Kerfeld, C. A. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences*, 110(3):1053–1058.

Shintani, M., Sanchez, Z. K., and Kimbara, K. (2015). Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology*, 6.

Sievers, F. and Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple Sequence Alignment Methods*, pages 105–116.

Simm, S., Keller, M., Selymesi, M., and Schleiff, E. (2014). The composition of the global and feature specific cyanobacterial core-genomes. *Frontiers in Microbiology*, 6:219–219.

Sivonen, K., Leikoski, N., Fewer, D. P., and Jokela, J. (2010). Cyanobactinsribosomal cyclic peptides produced by cyanobacteria. *Applied Microbiology and Biotechnology*, 86(5):1213–1225.

Snipen, L., Almøy, T., and Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*, 10(1):385.

Sorensen, G., Baker, A. C., Hall, M. J., Munn, C. B., and Schroeder, D. C. (2009). Novel virus dynamics in an *Emiliania huxleyi* bloom. *Journal of Plankton Research*, 31(7):787–791.

Stanier, R. Y. and Niel, C. v. (1962). The concept of a bacterium. *Archives of Microbiology*, 42(1):17–35.

Stricker, O., Masepohl, B., Klipp, W., and Böhme, H. (1997). Identification and characterization of the *nifV-nifZ-nifT* gene region from the filamentous

cyanobacterium *Anabaena* sp. strain PCC 7120. *Journal of Bacteriology*, 179(9):2930–2937.

Stucken, K., John, U., Cembella, A., Murillo, A. A., Soto-Liebe, K., Fuentes-Valdés, J. J., Friedel, M., Plominsky, A. M., Vásquez, M., and Glöckner, G. (2010). The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS One*, 5(2):e9235.

Šulčius, S., Alzbutas, G., Kvederavičiūtė, K., Koreivienė, J., Zakrys, L., Lubys, A., and Paškauskas, R. (2015). Draft genome sequence of the cyanobacterium *Aphanizomenon flos-aquae* strain 2012/KM1/D3, isolated from the Curonian Lagoon (Baltic Sea). *Genome Announcements*, 3(1):e01392–14.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biology*, 3(5):e144.

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A. M., Kelly, L., Weigele, P. R., DeFrancesco, A. S., Kern, S. E., Thompson, L. R., Young, S., et al. (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environmental Microbiology*, 12(11):3035–3056.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., and Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology*, 4(8):e234.

Surakka, A., Sihvonen, L. M., Lehtimaeki, J. M., Wahlsten, M., Vuorela, P., and Sivonen, K. (2005). Benthic cyanobacteria from the Baltic Sea contain cytotoxic *Anabaena*, *Nodularia*, and *Nostoc* strains and an apoptosis-inducing *Phormidium* strain. *Environmental Toxicology*, 20(3):285–292.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl 2):W609–W612.

Suzuki, H., Yamada, C., and Kato, K. (2007). γ-Glutamyl compounds and their enzymatic production using bacterial γ-glutamyltranspeptidase. *Amino Acids*, 32(3):333–340.

Szemes, T., Vlková, B., Minárik, G., Drahovská, H., Turňa, J., and Celec, P. (2012). Does phage P22 contribute to resistance of *Salmonella* to oxidative stress? *Medical Hypotheses*, 79(4):484–486.

Tetart, F., Desplats, C., and Krisch, H. (1998). Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity. *Journal of Molecular Biology*, 282(3):543–556.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences*, 102(39):13950–13955.

Thiel, T., Pratte, B. S., Zhong, J., Goodwin, L., Copeland, A., Lucas, S., Han, C., Pitluck, S., Land, M. L., Kyrpides, N. C., et al. (2014). Complete genome sequence of *Anabaena variabilis* atcc 29413. *Standards in Genomic Sciences*, 9(3):562.

Ting, C. S., Rocap, G., King, J., and Chisholm, S. W. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends in Microbiology*, 10(3):134–142.

Tyler, B. (1978). Regulation of the assimilation of nitrogen compounds. *Annual Review of Biochemistry*, 47(1):1127–1162.

van der Ploeg, J. R., Iwanicka-Nowicka, R., Bykowski, T., Hryniewicz, M. M., and Leisinger, T. (1999). The *Escherichia coli ssuEADCB* gene cluster is required for the utilization of sulfur from aliphatic sulfonates and is regulated by the transcriptional activator Cbl. *Journal of Biological Chemistry*, 274(41):29358–29365.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., and Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, page gkv657.

Vasu, K., Nagamalleswari, E., and Nagaraja, V. (2012). Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. *Proceedings of the National Academy of Sciences*, 109(20):E1287–E1293.

Vestola, J., Shishido, T. K., Jokela, J., Fewer, D. P., Aitio, O., Permi, P., Wahlsten, M., Wang, H., Rouhiainen, L., and Sivonen, K. (2014). Hassallidins, antifungal glycolipopeptides, are widespread among cyanobacteria and are the end-product of a nonribosomal pathway. *Proceedings of the National Academy of Sciences*, 111(18):E1909–E1917.

Vinuesa, P. and Contreras-Moreira, B. (2015). Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. *Bacterial Pangenomics: Methods and Protocols*, pages 203–232.

Wacklin, P., Hoffmann, L., Komárek, J., et al. (2009). Nomenclatural validation of the genetically revised cyanobacterial genus *Dolichospermum* (ralfs ex bornet et flahault) comb. nova. *Fottea*, 9(1):59–64.

Wang, H., Fewer, D. P., and Sivonen, K. (2011). Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS One*, 6(7):e22384.

Wang, H., Sivonen, K., Rouhiainen, L., Fewer, D. P., Lyra, C., Rantala-Ylinen, A., Vestola, J., Jokela, J., Rantasarkka, K., Li, Z., and Liu, B. (2012). Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. *BMC Genomics*, 13:613.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R., Wohlleben, W., et al. (2015). antiSMASH 3.0a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, 43(W1):W237–W243.

Willenbrock, H., Hallin, P. F., Wassenaar, T. M., and Ussery, D. W. (2007). Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biology*, 8(12):1.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

Worm, J. and Sondergaard, M. (1998). Dynamics of heterotrophic bacteria attached to *Microcystis* spp.(*Cyanobacteria*). *Aquatic Microbial Ecology*, 14(1):19–28.

Wu, Z.-X., Gan, N.-Q., and Song, L.-R. (2007). Genetic diversity: Geographical distribution and toxin profiles of *Microcystis* strains (*Cyanobacteria*) in China. *Journal of Integrative Plant Biology*, 49(3):262–269.

Xi, H., Schneider, B. L., and Reitzer, L. (2000). Purine catabolism in *Escherichia coli* and function of xanthine dehydrogenase in purine salvage. *Journal of Bacteriology*, 182(19):5332–5341.

Yamada, T., Satoh, S., Ishikawa, H., Fujiwara, A., Kawasaki, T., Fujie, M., and Ogata, H. (2010). A jumbo phage infecting the phytopathogen *Ralstonia solanacearum* defines a new lineage of the *Myoviridae* family. *Virology*, 398(1):135–147.

Yamaguchi, K., Suzuki, I., Yamamoto, H., Lyukevich, A., Bodrova, I., Los, D. A., Piven, I., Zinchenko, V., Kanehisa, M., and Murata, N. (2002). A two-component Mn_2+-sensing system negatively regulates expression of the *mntCAB* operon in *Synechocystis*. *The Plant Cell*, 14(11):2901–2913.

Yang, C., Lin, F., Li, Q., Li, T., and Zhao, J. (2015). Comparative genomics reveals diversified CRISPR-Cas systems of globally distributed *Microcystis aeruginosa*, a freshwater bloom-forming cyanobacterium. *Frontiers in Microbiology*, 6:394.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yin, Y. and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics*, 9(1):1.

Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y., Sakamoto, S., Hiroishi, S., and Ogata, H. (2008). Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *Journal of Bacteriology*, 190(5):1762–1772.

Yoshida, T., Takashima, Y., Tomaru, Y., Shirai, Y., Takao, Y., Hiroishi, S., and Nagasaki, K. (2006). Isolation and characterization of a cyanophage infecting the toxic cyanobacterium *Microcystis aeruginosa*. *Applied and Environmental Microbiology*, 72(2):1239–1247.

Yoshida-Takashima, Y., Yoshida, M., Ogata, H., Nagasaki, K., Hiroishi, S., and Yoshida, T. (2012). Cyanophage infection in the bloom-forming cyanobacteria *Microcystis aeruginosa* in surface freshwater. *Microbes and Environments*, 27(4):350–355.

Yoshizawa, S., Matsushima, R., Watanabe, M. F., Harada, K.-i., Ichihara, A., Carmichael, W. W., and Fujiki, H. (1990). Inhibition of protein phosphatases by microcystins and nodularin associated with hepatotoxicity. *Journal of Cancer Research and Clinical Oncology*, 116(6):609–614.

Zapomělová, E., Jezberová, J., Hrouzek, P., Hisem, D., Řeháková, K., and Komárková, J. (2009). Polyphasic characterization of three strains of *Anabaena reniformis* and *Aphanizomenon aphanizomenoides* (cyanobacteria) and their reclassification to *Sphaerospermum* gen. nov.(incl. *Anabaena kisseleviana*) 1. *Journal of Phycology*, 45(6):1363–1373.

Zapomělová, E., Skácelová, O., Pumann, P., Kopp, R., and Janeček, E. (2012). Biogeographically interesting planktonic *Nostocales* (Cyanobacteria) in the Czech Republic and their polyphasic evaluation resulting in taxonomic revisions of *Anabaena bergii* Ostenfeld 1908 (*Chrysosporum* gen. nov.) and *A. tenericaulis* Nygaard 1949 (*Dolichospermum tenericaule* comb. nova). *Hydrobiologia*, 698(1):353–365.

Zeder, M., Peter, S., Shabarova, T., and Pernthaler, J. (2009). A small population of planktonic *Flavobacteria* with disproportionally high growth during the spring phytoplankton bloom in a prealpine lake. *Environmental Microbiology*, 11(10):2676–2686.

Zhang, J.-Y., Guan, R., Zhang, H.-J., Li, H., Xiao, P., Yu, G.-L., Du, L., Cao, D.-M., Zhu, B.-C., Li, R.-H., et al. (2016). Complete genome sequence and genomic characterization of *Microcystis panniformis* FACHB 1757 by third-generation sequencing. *Standards in Genomic Sciences*, 11(1):1.

Zhao, W., Ye, Z., and Zhao, J. (2007). RbrA, a cyanobacterial rubrerythrin, functions as a FNR-dependent peroxidase in heterocysts in protection of nitrogenase from damage by hydrogen peroxide in *Anabaena* sp. PCC 7120. *Molecular Microbiology*, 66(5):1219–1230.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126.

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., Ellisman, M., Deerinck, T., Sullivan, M. B., and Giovannoni, S. J. (2013). Abundant SAR11 viruses in the ocean. *Nature*, 494(7437):357–360.

Ziemert, N., Ishida, K., Weiz, A., Hertweck, C., and Dittmann, E. (2010). Exploiting the natural diversity of microviridin gene clusters for discovery of novel tricyclic depsipeptides. *Applied and Environmental Microbiology*, 76(11):3568–3574.

Zowawi, H. M., Forde, B. M., Alfaresi, M., Alzarouni, A., Farahat, Y., Chong, T.-M., Yin, W.-F., Chan, K.-G., Li, J., Schembri, M. A., et al. (2015). Stepwise evolution of pandrug-resistance in *Klebsiella pneumoniae*. *Scientific Reports*, 5.

Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H., and Amann, R. (2003). High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. *Applied and Environmental Microbiology*, 69(2):1299–1304.