AN ABSTRACT OF THE DISSERTATION OF

Nathaniel M. Brown for the degree of Doctor of Philosophy in Microbiology presented on March 1, 2016.

Title: Advances in Freshwater Cyanobacterial Genomics

Abstract approved: _____

Theo W. Dreher

Limnology is undergoing a transition to high-throughput -omic analysis of freshwater bacterial communities. An important first step in making the transition is to characterize several genomes that can be used as references to guide metagenome assembly and analysis. Here I characterize four new freshwater cyanobacterial genomes, a pair of lake community metagenomes, and a temperate phage.

- The *Anabaena* sp. WA102 genome is sequenced with long-read sequencing to finished status, unique structural features of the genome are analyzed, a comparative genomic analysis with other members of the *Nostocaceae* is carried out, and its capacity to produce anatoxin-a (and related toxin variants) is assessed

- Two metagenomes of the cyanobacterial bloom community in Anderson Lake, Jefferson County, Washington State, USA are analyzed, using the finished

Anabaena sp. WA102 genome to identify the dominant anatoxin-a-producing strain in the metagenomes and determine that the dominant cyanobacterial strain is nearly identical and likely clonal between blooms in 2012 and 2013

- Two new *Nostocaceae* genomes, *Anabaena* sp. AL93 and *Aphanizomenon* sp. WA102, are sequenced and compared with *Anabaena* sp. WA102
- A new freshwater *Cyanobium* species is isolated and its genome is sequenced, a temperate cyanobacterial phage that infects the strain is also isolated and sequenced and its integration into the host genome is characterized

Advances in Freshwater Cyanobacterial Genomics

by

Nathaniel M. Brown

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented March 1, 2016
Commencement June 2016

Doctor of Philosophy dissertation of Nathaniel M. Brown presented on March 1, 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.

Nathaniel M. Brown, Author

ACKNOWLEDGEMENTS

I thank my advisor, Dr. Theo Dreher, for his patience and quixotic positivity. I thank my parents, Tim and Jamie Brown, for their support. I thank my grandfather, Cyrus Austin, for explaining to me that the most difficult achievements are also the most worthwhile. I thank my friends, Nathan and Carly Hill, for keeping me grounded in reality. I thank my labmates, Tim Otten and Connor Driscoll, for listening to my endless talking.

TABLE OF CONTENTS

Page

1 Objective and Background 1

1.1 Objective 1

1.2 Background 2

2 Introduction 4

2.1 Freshwater cyanobacterial blooms 4

2.2 Environmental metagenomics 6

2.3 Current state of cyanobacterial genomics 8

3 Structural and Functional Analysis of the Closed Genome of the Recently Isolated Toxic *Anabaena* sp. WA102 12

3.1 Introduction 13

3.2 Results 15

3.2.1 The *Anabaena* sp. WA102 culture and genome 15

3.2.2 Comparison of *Anabaena* sp. WA102 long- and short-read genome assemblies 16

3.2.3 The *Anabaena* sp. AL93 culture and genome 18

3.2.4 Phylogenomic relationship between *Anabaena* sp. WA102, AL93, and other fully sequenced *Nostocacaea* 19

3.2.5 Comparing gene content and metabolic capabilities of *Anabaena* sp. WA102 and AL93 with other *Nostocaceae* genomes 19

3.2.6 Capacity for synthesis of anatoxin-a and other secondary metabolites 22

3.2.7 Lack of synteny with *Anabaena* sp. 90 24

3.2.8 The mobilome 26

3.2.9 Relationship between the *Anabaena* sp. WA102 genome and the Anderson Lake metagenome 29

3.2.10 A recent deletion event in the *Anabaena* sp. WA102 genome 29

3.2.11 Tandem repeat of the anatoxin-a *anaBCD* promoter region . 30

3.3 Discussion 31

3.3.1 The recently cultured toxic isolate, *Anabaena* sp. WA102, closely reflects the parent strain in Anderson Lake 31

3.3.2 Closing the genome reveals details about genome architecture 33

3.3.3 Predicted ecologic profile of *Anabaena* sp. WA102 34

TABLE OF CONTENTS (Continued)

Page

3.3.4 Evolution of the *Anabaena* sp. WA102 genome 35
3.4 Methods 36
3.4.1 Sample collection 36
3.4.2 Culture establishment and maintenance 37
3.4.3 LC-MS/MS 38
3.4.4 DNA extraction and amplification 39
3.4.5 DNA sequencing 40
3.4.6 Draft genome binning 40
3.4.7 Finished *Anabaena* sp. WA102 genome analysis 42
3.4.8 Comparative genomics among members of the *Nostocaceae* . 43
3.4.9 Accession numbers used in study 44
3.5 Tables and Figures 45

4 Identification of the major anatoxin-a producing cyanobacterium in Anderson Lake, its dynamics, and its distribution in the Puget Sound region 61
4.1 Introduction 62
4.2 Results 63
4.2.1 *Anabaena* sp. WA102 is the dominant cyanobacterial species in the 2012 Anderson Lake bloom metagenome sample . . . 63
4.2.2 *Aphanizomenon* sp. WA102, a novel non-toxic Nostocaceae species in Anderson Lake 66
4.2.3 Comparison of *Anabaena*/*Aphanizomenon* population genomes clustered within the July 2012 and May 2013 metagenomes. 67
4.2.4 *Anabaena* colony morphology correlates with anatoxin-a production 69
4.2.5 Distribution of *Anabaena* sp. WA102 across the Puget Sound region and *Nostocaceae* diversity 71
4.3 Discussion 75
4.3.1 *Anabaena* sp. WA102 was the major anatoxin-a producer in Anderson Lake in metagenome samples from July 2012 and May 2013 75
4.3.2 *Anabaena* sp. WA102 is not always the dominant nitrogen-fixing autotroph in Anderson Lake 77
4.3.3 *Anabaena* sp. WA102 is sparsely distributed throughout the Puget Sound Region 79

TABLE OF CONTENTS (Continued)

Page

4.4 Methods . . . 80
4.4.1 Sample collection . . . 80
4.4.2 Single-colony isolation, toxin extraction, and DNA sequencing 81
4.4.3 DNA extraction from lake samples for shotgun metagenomic and amplicon sequencing . . . 81
4.4.4 *cpcBA*-IGS amplicon primer design, amplification, and sequencing . . . 82
4.4.5 *cpcBA*-IGS amplicon analysis . . . 83
4.4.6 Metagenome analysis . . . 84
4.4.7 Genome comparisons . . . 86
4.5 Tables and Figures . . . 87

5 The genome of a novel freshwater picocyanobacterium, *Cyanobium* sp. LC18, and lysogenization by of one of its temperate phages, C-CRS01 96
5.1 Introduction . . . 97
5.2 Results . . . 98
5.2.1 *Cyanobium* sp. LC18 is a novel cyanobacterial species from the Klamath River System . . . 98
5.2.2 S-CRS01 infects *Cyanobium* sp. LC18 . . . 99
5.2.3 The C-CRS01 genome . . . 100
5.2.4 C-CRS01 lysogenizes *Cyanobium* sp. LC18 . . . 103
5.3 Discussion . . . 104
5.3.1 Relevance of the *Cyanobium* sp. LC18 genome to diversity-driven sequencing of the *Cyanobacteria* . . . 104
5.3.2 Isolation of a novel freshwater temperate *Siphovirus*, C-CRS01 105
5.3.3 Lysogenization of *Cyanobium* sp. LC18 by C-CRS01 . . . 106
5.4 Methods . . . 107
5.4.1 Isolating *Cyanobium* sp. LC18 and culture maintenance . . . 107
5.4.2 Isolating C-CRS01 on *Cyanobium* sp. LC18 . . . 107
5.4.3 Transmission Electron Microscopy of C-CRS01 . . . 109
5.4.4 Isolating *Cyanobium* sp. LC18 lysogens . . . 109
5.4.5 DNA preparation for C-CRS01 and *Cyanobium* sp. LC18 . . . 110
5.4.6 High-throughput sequencing of *Cyanobium* sp. LC18 lysogen and C-CRS01 genomes . . . 110

TABLE OF CONTENTS (Continued)

Page

5.4.7 Assembling and analyzing the *Cyanobium* sp. LC18 lysogen and C-CRS01 genomes 111
5.4.8 PCR amplification of prophage junctions 112
5.5 Tables and Figures 114

6 Conclusion 120

7 Contributions from authors 125
7.1 Chapter 3: Structural and Functional Analysis of the Closed Genome of the Recently Isolated Toxic *Anabaena* sp. WA102 125
7.2 Chapter 4: Identification of the major anatoxin-a producing cyanobacterium in Anderson Lake, its dynamics, and its distribution in the Puget Sound region 125
7.3 Chapter 5: The genome of a novel freshwater picocyanobacterium, *Cyanobium* sp. LC18, and lysogenization by of one of its temperate phages, C-CRS01 126

Bibliography 127

LIST OF FIGURES

Figure | Page

3.1 Location of Anderson Lake and picture of culture. A) A map of the Puget Sound region in Washington State, USA. *Anabaena* sp. WA102 was isolated from Anderson Lake at 48.0190 N, 237.1963 W on the Olympic Peninsula. B) A brightfield micrograph of *Anabaena* sp. WA102 at 200x magnification. Vegetative cells measure 7.1 by 6 μm on average. Colonies are heterocystous because the culture is maintained in nitrogen-free medium (BG-11_0). 47

3.2 HPLC-MS/MS survey of anatoxin-a and derivatives. A) HPLC elution of compounds extracted from the *Anabaena* sp. WA102 culture and two Anderson Lake samples (WA102 and WA103). Anatoxin-a elutes at approximately 2 minutes, as indicated by the anatoxin-a standard. Anatoxin-a peaks are surrounded by a gray dashed line. No variants of anatoxin-a were detected. B) Ion mass spectra for anatoxin-a are compared from lake sample WA102 (May 20th, 2013 with 12.5 μg/L anatoxin-a), lake sample WA103 (June 17th, 2013 with 35.8 μg/L anatoxin-a), and the culture. All spectra match the spectrum of the anatoxin-a standard closely. 48

3.3 PacBio read length distribution for the *Anabaena* sp. WA102 culture. PacBio read length average 8.5 kbp, allowing complete assembly of the *Anabaena* sp. WA102 across long repeat regions. 49

Figure Page

3.4 Plot of the *Anabaena* sp. WA102 genome. A) The genome is plotted as a black ring with demarcations every 100 kbp. Average GC content in 10 kbp non-overlapping windows is plotted outside of the genome ring. The first track within the genome ring includes the location of the *oriC* and RNA elements. The *oriC* was determined to lie downstream of *dnaA* among DnaA-binding motifs. The following two interior rings denote predicted protein-coding sequences, first on the positive strand (clockwise) and then on the negative strand (counter-clockwise). NRPS-PKS clusters identified by antiSMASH are shown as red tiles in the fourth interior track. Mobile elements - homing endonucleases and transposases - are plotted on the fifth interior track as orange and yellow tiles, respectively. Contigs from the binned Illumina genome of the culture (Figure 3.6) were aligned to the closed genome and 229 gaps in the Illumina assembly are represented as green tiles in the sixth interior track. Green arcs across the center connect repeated regions in the genome, determined by blastn alignment of the finished genome against itself. Note that repeat regions often coincide with gaps in the Illumina assembly. B) Genome-wide plot of cumulative GC skew. GC skew was averaged across 1 kbp non-overlapping windows of the genome and then cumulatively summed. Minimum and maximum points on the cumulative GC skew plot should indicate *oriC* and *terC*, respectively. However, the signal from the cumulative GC skew is weakened, preventing precise prediction of *oriC*, *terC*, and the replicon arms. . . . 50

3.5 *Nostocacaea* phylogenetic tree. A phylogenetic tree constructed from amino-acid alignments of single-copy orthologs present in all genomes of some of the fully sequenced members of the *Nostocacaea*. 51

3.6 KEGG orthologs (KO) differentially represented among the compared *Nostocaceae* genomes. All proteins from each *Nostocaceae* genome were mapped to the online KO database. Orthologs with significant differences among the genomes were highlighted in the above table for comparison. *Nostocaceae* genomes are arranged according to the phylogenetic tree for easy comparison. The *Anabaena* sp. WA102 genome encodes a sulfur metabolism cluster absent or incomplete in 6 out of 11 *Nostocaceae* genomes. 52

Figure Page

3.7 Nucleotide alignment of anatoxin-a clusters from *Cyanobacteria*. *anaA-G* and *anaI* are all conserved in *Anabaena* sp. WA102 and *Anabaena* sp. AL93, though *anaH* is missing from both. The 5' region of *anaB* and upstream promoter region is triplicated in *Anabaena* sp. WA102. The anatoxin-a cluster from *Anabaena* sp. WA102 is most similar to that from *Anabaena* sp. 37. The three *Anabaena* strains share a gene of unknown function downstream of *anaG* (colored pink). The *anaG* genes differ in size, correlated with different variants of anatoxin-a. Shorter variants of AnaG omit or truncate a putative methyl transferase domain. The *anaF* and *anaG* genes share a region of 86% nucleotide identity that is likely a homologous protein domain. *Anabaena* sp. WA102 and AL93 encode two of the shortest *anaG* genes and produce anatoxin-a, *Cylindrospermum* sp. PCC 7417 produces dihydroanatoxin-a (likely due to the unique gene Cylst 6226), and *Oscillatoria* sp. PCC 6506 primarily produces homoanatoxin-a. 53

3.8 Comparison of the AnaG protein domains among Cyanobacteria. The AnaG protein sequences from *Oscillatoria* sp. PCC 6506 and *Anabaena* sp. 37 have methyltransferase domains not present in any other AnaG protein sequences. The methyltransferase domains are divergent. The methyltransferase in *Oscillatoria* sp. PCC 6506 is proposed to contribute a methyl group that makes the homoanatoxin-a variant of anatoxin-a. AnaG lacking a methyltransferase domain (or containing a non-functional domain) likely prevents production of homoanatoxin-a. In support of that, no homoanatoxin-a was detected in the *Anabaena* sp. WA102 culture. 54

3.9 Nucleotide alignment between *Anabaena* sp. 90 and WA102. Although *Anabaena* sp. 90 and WA102 share 91.5% average nucleotide identity, they nearly entirely lack synteny. Additionally, the *Anabaena* sp. 90 genome is divided between two chromosomes, unlike the single chromosome of the *Anabaena* sp. WA102 genome. 55

3.10 Dotplots and average ortholog similarity for pairwise comparisons within three bacterial genera. Dotplots illustrate preservation or absence of long-range nucleotide similarity (synteny) between paired genomes from *Anabaena* in this study and *Pseudomonas* and *Streptococcus* (originally compared in Novichkov et al., 2009). 56

Figure Page

3.11 Probability density of the local colinear block (LCB) lengths for three bacterial genera. The same pairwise genomes comparisons from the dotplots in Figure 9 are aligned in Mauve. Mauve generates LCBs, which are syntenous regions defined by conserved termini, and that may contain large insertions. The lengths of these LCBs are plotted in a probability density plot for each pairwise genome comparison. The mean LCB length for each pairwise genome comparison is shown as a dotted line with the value printed above the graph. *Pseudomonas* genomes have a mean LCB length of 32.5 kbp, *Anabaena* 3.3 kbp, and *Streptococcus* 210 bp, quantifying what can be observed in the dotplots. 57

3.12 Comparing synteny within a local colinear block between *Anabaena* sp. WA102 and 90 (nucleotides 1,179,150-1,203,874 and 2,682,853-2,688,083, respectively). Within this local colinear block, there is evidence of interruption by transposases. Most of the six instances of broken synteny in this LCB are not clearly attributable to a particular mechanism. 57

3.13 Phylogenetic tree of transposase protein sequences encoded in the *Anabaena* sp. WA102 genome. The phylogenetic relationship between 130 annotated transposase protein sequences is sketched out in the tree. Two large clades of closely related transposases dominate the tree. The IS4Sa clade includes 25 transposases and the IS10 clade includes 20 transposases, which both belong to the larger IS4 transposase family. These transposases have a DDE-type active site that facilitates cut-and-paste transposition. The IS4Sa clade has an identical terminal direct repeat sequence: CCGCCTTGTCACCCGTTAAG. The IS10 clade has the terminal direct repeat sequence: ATTCAACAYTTCTG. 58

3.14 Nucleotide alignment between *Anabaena* sp. WA102 chromosome and plasmid. Nucleotide similarity between the chromosome and the plasmid indicates that the plasmid may be integrative and form genomic islands either by integrating into a site on the chromosome or by homologous recombination with the chromosome. The plasmid may be integrative because it encodes site-specific integrases, which can also be found at the homologous site on the chromosome. . . . 59

LIST OF FIGURES (Continued)

Figure Page

3.15 Deletion mutation detected in *Anabaena* sp. WA102 culture. A deletion mutation was detected in the PacBio long-read assembly of the *Anabaena* sp. WA102 culture. Mapping reads to the indel region showed that the deletion occurred between nucleotides 4,800,950 and 4,804,900. The deletion arose after December 2013 and expanded through the population to roughly two-thirds of the culture population by December 2014. 59

Figure Page

3.16 Tandem duplication of the putative *anaBCD* promoter region. A) Alignment of the *anaB* gene and upstream promoter region between different assemblies of the *Anabaena* sp. WA102 culture. Promoters were identified with the Virtual Footprint online server, and only promoters with PWM alignment scores greater than 12 were plotted. The 5' end of the *anaB* gene and upstream promoter region are triplicated in the PacBio assembly. None of the Illumina assemblies correctly assemble the tandem triplication. Assembly of 100 bp reads by IDBA v1.1.1 failed to correctly assemble the *anaB* gene and the promoter region. Assembly by PriceTI v1.0.1, using the IDBA contig to seed the assembly, produced two alternate versions of the *anaB* region. In the first version, the *anaB* gene and the upstream promoter region are both improperly assembled. In the second, the *anaB* gene and the most proximal portion of the promoter region are correctly assembled, but triplication is not assembled. B) Read coverage across the promoter region upstream of the *anaB* gene. Illumina metagenome reads from a toxic bloom in Anderson Lake (WA25, blue line), *Anabaena* sp. AL93 culture (green line), and *Anabaena* sp. WA102 culture are mapped across *anaB* and its upstream promoter region. Coverage is summed at each nucleotide and illustrates the absence of two junctions formed between the triplications where the green line drops to zero for the *Anabaena* sp. AL93 culture. In contrast, both the *Anabaena* sp. WA102 culture and the Anderson Lake metagenome contain the junctions formed by the triplication because read coverage does not fall to zero at those loci. Presence of the triplication in the Anderson Lake metagenome indicates that it formed in the *Anabaena* sp. WA102 genome nearly a year prior to establishing the culture. It has been under selection in the environment and continues to be selected for in culture. *Read coverage values for the July 2012 Anderson Lake metagenome have been divided by 10 to facilitate comparison along the ordinate. 60

LIST OF FIGURES (Continued)

Figure Page

4.1 Anatoxin-a occurrence in Anderson Lake. A) Lakes in Washington State, USA, with anatoxin-a levels measuring above the 1μg/L state guideline level for recreational exposure. The mean measured anatoxin-a level is shown for each lake, though it must be noted that these means are based on an unevenly sampled data and include extreme outlier values. Data is from https://www.nwtoxicalgae.org/. B) Anatoxin-a levels measured over the past 7 years in Anderson Lake, Jefferson County, WA. The summer months June-August are highlighted in yellow and the 1μg/L guideline level is shown as a dashed line. Points in black represent anatoxin-a measurements >1 μg/L and points in gray represent measurements <1 μg/L. **C)** Total (Kjeldahl) nitrogen:total phosphorus ratios (TN:TP) measured over the past 7 years. A TN:TP of 35, at which nitrogen-fixing cyanobacteria are thought to be uncompetitive with non-nitrogen-fixing cyanobacteria is shown as a dashed gray line. Measurements of TN:TP >35 are shown as gray dots, and measurements <35 are shown as black dots. 87

Figure Page

4.2 Metagenome analysis of Anderson Lake samples. A) Average coverage depth plot of contigs from the July 7th, 2012 Anderson Lake metagenome sample. Sequencing reads (30 Gbp of Illumina HiSeq 100-nt paired-end reads) were assembled into 230,285 contigs with total size of 255 Mbp and an N50 of 1,530 bp. Contigs belonging to population genomes are clustered on the plot according to coverage depth from the sequenced *Anabaena* sp. WA102 culture (y-axis) and coverage depth in the July 2012 Anderson Lake metagenome sample (x-axis). Two species of *Anabaena* were identified by PhylopythiaS+. The *Anabaena* population genome with average coverage depth of 1,200 is nearly identical to *Anabaena* sp. WA102 (Table 1). The *Anabaena* population genome with average coverage depth of 30 in the Anderson Lake metagenome is actually *Aphanizomenon* sp. WA102 (see text). The only anatoxin-a biosynthetic (ana) genes identified cluster with the *Anabaena* sp. WA102 population genome (dark red circles). B) Average coverage depth plot of contigs from the May 20th, 2013 Anderson Lake metagenome sample. Contigs were assembled from 4.6 Gbp of Illumina MiSeq 250-bp paired-end reads. The total assembly is 223 Mbp and has an N50 of 1,165. The *Aphanizomenon* sp. WA102 population genome has an average coverage depth of 100 in the Anderson Lake metagenome and the *Anabaena* sp. WA102 population genome has an average coverage depth of 20. The only anatoxin-a genes detected on contigs cluster with the *Anabaena* sp. WA102 population genome (dark red circles), which was confirmed by mapping reads to known anatoxin-a biosynthetic gene clusters. 88

LIST OF FIGURES (Continued)

Figure | Page

4.3 *Anabaena* morphotypes in Anderson Lake. A) Two *Anabaena* morphologies present in Anderson Lake on June 18th, 2013 (phase contrast, 200x magnification). A colony exhibiting the large-cell *Anabaena-crassa*-like morphology is shown on the left. Several intertwined filaments exhibiting the small-cell *Anabaena-flos-aquae*-like morphology are on the right. B) Filamentous colonies of the *Anabaena* sp. WA102 culture, resembling the *Anabaena flos-aquae*-like colonies in panel A. C) Anatoxin-a detection in individual colonies of *Anabaena* isolated from Anderson Lake on 18th and 25th June, 2013. Ten colonies with an *Anabaena-crassa*-like and 25 colonies with an *Anabaena-flos-aquae*-like morphology were tested for anatoxin-a by HPLC-MS/MS. 89

4.4 Unrooted phylogenetic tree showing relationship between the 19 most abundant *Nostocaeae cpcBA* OTUs detected in Puget Sound area lakes in 2012. Five monophyletic clades emerge, which we denote as OTU clades. The OTUs subsumed by each group are listed in parentheses underneath the OTU group name. OTU clade 1 represents *Anabaena* sp. WA102. These clades likely represent *Nostocaceae* strains detected in Puget Sound Region lakes, with intra-strain variation within each group represented by individual OTUs. 90

4.5 A heat-map of the OTU communities in *cpcBA*-IGS amplicon libraries by lake. The OTU clades as defined by Figure 6 are shown on the y-axis. The lake samples are arranged by similarity according to MDS of the weighted unifrac metric for each OTU community. . 91

4.6 The proportion of OTU clades in each lake and their distribution across lakes sampled in the Puget Sound region in 2012. 94

4.7 Comparison of *cpcBA*-IGS primer sequences and primer annealing sites in *Aphanizomenon* sp. WA102 and *Anabaena* sp. WA102. *Aphanizomenon* sp. WA102 has one more mismatch than *Anabaena* sp. WA102, presumably preventing its detection in DNA samples from Anderson Lake. 95

5.1 Micrograph of *Cyanobium* sp. LC18. A 200x magnification bright-field micrograph of *Cyanobium* sp. LC18 114

LIST OF FIGURES (Continued)

Figure Page

5.2 Phylogenomic tree of the *Cyanobium* sp. LC18 genome and 24 other *Cyanobium* and *Synechococcus* genomes. 514 amino acid sequences for orthologs shared between all 25 genomes in the analysis were concatenated and aligned to propose the evolutionary relationship between *Cyanobium* sp. LC18 and other picocyanobacteria. 115

5.3 Evidence of the novel cyanobacterial phage C-CRS01 infecting *Cyanobium* sp. LC18. A) Plaques less than 1mm in diameter forming on a lawn of *Cyanobium* sp. LC18 in a BG-11 top-agar plate. B) Pulsed-field gel electrophoresis of DNA extracted from two different plaques shows a phage genome size of approximately 60 kbp. C) 45,000x magnification transmission electron micrographs reveal that the phage is a member of the *Siphoviridae*, with a flexible, non-contractile tail and prolate icosahedral capsid. D) Comparison of a liquid lysate of *Cyanobium* sp. LC18 and an uninfected culture 3 days after infection. 116

5.4 The C-CRS01 genome. The 60,581 bp C-CRS01 genome is shown with annotated ORFs in the outermost track, all detected ORFs shown in the first inner track, average GC skew calculated in a 500-nt sliding window shown in the next innermost track, and average GC content calculated in a 500-nt sliding window in the innermost track. 117

Figure Page

5.5 Integration of C-CRS01 genome into the host *Cyanobium* sp. LC18 genome. A) Contig 28 from the assembly of a putative *Cyanobium* sp. LC18 lysogen indicates that the C-CRS01 genome does indeed integrate into the host genome (to scale). B) PCR verification of the prophage junctions in the *Cyanobium* sp. LC18 lysogen: lanes 1 and 2 are amplifications of the *attL* and *attR* junctions (respectively) from a non-lysogenic *Cyanobium* sp. LC18 strain, showing no detection of either junction, lanes 3 and 4 are amplifications of the *attL* and *attR* junctions (respectively) from a *Cyanobium* sp. LC18 lysogen, showing detection of both junctions as expected, lane 5 shows amplification of the *attP* junction in C-CRS01 genomic DNA (no bacterial DNA) C) Diagram (not to scale) showing the site of integration for C-CRS01 in *Cyanobium* sp. LC18. C-CRS01 integrates into a putative DNA binding protein with a helix-turn-helix domain. The amino acid and DNA sequence beginning from the *attB* region of the putative protein to the C-terminus of the protein is shown. D) After integration, C-CRS01 fuses the disrupted protein at the *attL* site with a homologous amino acid sequence encoded in its genome that includes an 8-amino-acid addition at the C terminus: AAAADDPA. The sequence of the fusion protein formed is shown at the bottom. 119

LIST OF TABLES

Table Page

3.1 Summary statistics of sequencing data and binned *Anabaena* genomes. PE indicates paired-end reads. *Three contigs include the chromosome, plasmid, and the contig representing the insertion variant with the *xseA* gene. 45

3.2 Summary of *Anabaena* sp. WA102 genome (Genbank:CP011456-7) annotation. Annotation according to the Prokka script and NCBI Prokaryotic Annotation Pipeline. 46

4.1 Differences between the *Anabaena* sp. WA102 population genome and the *Anabaena* sp. WA102 reference genome. The *Anabaena* sp. WA102 population genome clustered from the July 7th, 2012 metagenome of Anderson Lake surface water shows 57 differences that occur in all reads mapped to the *Anabaena* sp. WA102 reference genome. Average read depth coverage was 399x and 6,199 nucleotides had no read coverage. 93

5.1 Primers used for PCR amplification across phage integration junctions. 114

5.2 ORF annotations for C-CRS01. Annotations for ORFs identified in C-CRS01 were made by transferring annotations from homologs identified by structure-guided alignment in Phyre2 and primary amino-acid sequence alignment in BLASTP. 118

Chapter 1 Objective and Background

1.1 Objective

The purpose of this study is to expand the characterization of freshwater cyanobacteria by sequencing and characterizing new freshwater cyanobacterial genomes. Bacterial genome sequences contain a wealth of information that can inform us about a strain's ecological niche [47], its ability to produce useful or novel natural products [4], its ability to produce toxins or pathogenicity factors [148], and - by comparing it with other bacterial genomes - its evolutionary history [67]. However, acquiring genomic information for freshwater cyanobacteria poses many challenges. Sometimes a strain cannot be isolated in culture, so that it must be sequenced directly from environmental samples along with many other bacteria. The metagenomic sampling techniques and computational tools required for this have only just recently become available. Additionally, cyanobacterial genomes are often filled with repetitive DNA sequence that can fragment the genome assembly. The DNA sequencing technology to overcome this barrier currently is becoming cheap enough to be practical for small academic laboratories, but it is not always clear how to apply this technology to environmental samples. Solutions for these principal problems were sought in this body of work while attempting to identify the major anatoxin-a producing cyanobacterial strain in Anderson Lake, Jefferson

County, Washington, USA.

1.2 Background

The *Cyanobacteria* are relatively well studied as a phylum, but many genera within this phylum remain poorly characterized. The marine *Prochlorococcus* and *Synechococcus* genera are disproportionately well studied because they are responsible for a large fraction of the oxygenation and carbon fixation on Earth [110]. Marine filamentous cyanobacteria are often examined for natural products that can serve as pharmaceuticals, and so receive a large amount of research attention [153, 61]. In contrast, terrestrial and freshwater cyanobacteria, though representing much of the diversity within the *Cyanobacteria*, have received less attention. Freshwater cyanobacteria, the focus of this work, are important for several reasons. They form eutrophic blooms that deoxygenate water, cause fish kills, and shade out native aquatic plant species [106]. These blooms can also produce toxic compounds that may cause cancer or paralysis [100, 18]. At lower concentrations, freshwater cyanobacteria can produce taste and odor compounds that make drinking water unpalatable [64]. Freshwater cyanobacteria can also provide nutrition, and are used in some health-food products [45]. Like marine cyanobacteria, freshwater strains can also produce a variety of useful natural compounds [55]. Recognizing the importance of freshwater cyanobacteria and leveraging innovations in DNA sequencing technology, there have been several recent efforts to acquire a more balanced sampling of genomes from the *Cyanobacteria*, including freshwater cyanobacteria

[135, 23, 94].

Chapter 2 Introduction

2.1 Freshwater cyanobacterial blooms

Freshwater cyanobacterial blooms occur when conditions allow freshwater cyanobacteria to grow and dominate an inland water body such as a lake or reservoir. They are bacterial dysbioses of freshwater ecosystems that can have outsized negative, non-linear effects on the entire ecosystem [29]. Blooms can produce toxins or anoxic conditions that kill fish [127] and other animals, in addition to killing macrophytes by blocking sunlight from reaching the benthos [63]. Some of the causes of freshwater cyanobacterial blooms are clear, but their relative importance is debated. Temperature, nutrient input, water column stratification, and light levels have all been implicated as (not necessarily independent) environmental drivers of freshwater cyanobacterial blooms [108, 83, 56, 10]. Blooms likely are becoming more frequent globally due to increased average global temperature [149, 107]. In addition to abiotic drivers, blooms can perpetuate themselves by increasing nutrient concentrations and cycling rates [22] and reduce the resilience of freshwater ecosystems to slight changes in environmental conditions, which in turn provokes bloom relapses [58]. As non-linear phenomena, blooms can also collapse suddenly. Hypothetical drivers of bloom collapse include both abiotic and biotic drivers: water column mixing, temperature change [167], grazing pres-

sure by predators, and viral-induced lysis [111, 49]. Freshwater cyanobacterial bloom collapse has received relatively less attention than bloom emergence, and many questions remain regarding the relative importance of hypothetical drivers of collapse. Bloom collapse often occurs in autumn, coincident with decreased day length and temperature and destratification of the water column, suggesting that abiotic environmental drivers are important. Although viral-induced algal bloom collapse has been thoroughly documented for *Emiliana huxleyi* in marine systems [9, 140, 59], data regarding viral-induced freshwater bloom collapses is scarce [111]. A barrier to determining if viruses cause freshwater cyanobacterial bloom collapse is the difficulty in measuring changes in the cyanobacterial bacteriophage population separately from changes in the entire lake's bacteriophage population and then correlate those changes to changes in the host population. For example, staining encapsidated viral DNA with epifluorescent dye and then counting virus-like particles (VLPs) does not distinguish between cyanobacteria-specific phages and non-cyanobacteria-specific viruses [101]. Higher resolution methods are needed. There has been a recent call to better understand the bacterial and viral drivers of bloom emergence and collapse, since abiotic drivers have thus far failed to explain many observed bloom dynamics [163]. DNA sequencing technology may provide the resolution needed to identify and quantify the dynamics of individual bacterial species and their corresponding bacteriophages during the lifecycle of a freshwater cyanobacterial bloom. The analytical technique that may make this possible is metagenomics, which is the sequencing, resolution, and characterization of a community of bacterial genomes present in a single environmental sample, such as

water taken from a lake. This study uses new sequencing techniques and metagenomic strategies to begin characterizing freshwater lake blooms and freshwater bloom-forming cyanobacteria in order to gain a better understanding of the biotic factors of a freshwater bloom lifecycle. In addition, a novel picocyanbacterium and its temperate phage are isolated and their genomes are sequenced. Progress is being made towards this goal, though observation of phage lysis terminating a freshwater cyanobacterial bloom remains elusive.

2.2 Environmental metagenomics

DNA sequencing platforms such as the Illumina HiSeq platform have reduced the price of DNA sequencing to the point where it has become practical to sequence the DNA from a bacterial community in an environmental sample [93]. This is usually done with Illumina short-read DNA sequencing technology, which can produce contiguous nucleotide sequences as long as 250bp. The short reads must be aligned with each other and concatenated with computational algorithms into larger nucleotide sequences that can be analyzed bioinformatically. These short-read assembly algorithms typically fail to concatenate the reads into contiguous nucleotide sequences (contigs) representing entire bacterial genomes. If more than one bacterial genome is present in the sample, such as is the case with environmental samples, then the multiple contigs that compose a single bacterial genome must be separated from contigs that compose other bacterial genomes. Clustering contigs from a metagenome into distinct bacterial genomes is a central challenge

of metagenomics and has important implications such as functional characterization of the genes and biochemical pathways encoded in each genome and assigning putative ecological roles for the bacteria represented by particular genomes. Additionally, clustering the contigs to determine the population dynamics of closely related bacterial species within a lake is important for answering many questions regarding freshwater cyanobacterial bloom emergence and collapse. Nucleotide contig clustering draws upon traditional statistical learning techniques to group contigs together by statistics that distinquish contigs from different genomes. It has recently been shown that two of the most important statistics by which to cluster contigs in a metagenome are tetranucleotide frequency and average contig coverage depth [1, 134]. Tetranucleotide or pentanucleotide frequency is a measure of the nucleotide composition of a contig determined by counting the number of four- or five- nucleotide combinations observed in the contig, and varies between genomes from different taxa [154]. Average contig coverage depth is a measure of how many times a contig was observed during DNA sequencing and indicates the relative abundance of a particular contig, with contigs from the same genome sharing the same relative abundance. Using these statistics, metagenome contigs often can be clustered into nearly complete representations of the original genomes, as measured by counts of bacterial universal unique marker genes. These clustered contigs are called population genomes, because they may contain more than one strain of a particular bacterial species [1]. This approach is new, and the quality and utility of these population genomes is still uncertain. However, they hold the promise of resolving the population dynamics of closely related bacterial species -

and their bacteriophages - in complex environmental samples such as lake water. It is important at this stage to assess the quality of these population genomes by comparing them with high-quality finished genomes derived from the environment.

2.3 Current state of cyanobacterial genomics

The *Cyanobacteria* are well studied as a phylum, however certain clades within the *Cyanobacteria* have been neglected, as measured by the small number of genomes sequenced from those clades [129, 34]. This has driven several groups to expand the number and diversity of sequenced genomes from the *Cyanobacteria*. The picocyanobacteria are the cyanobacterial clade with the largest number of sequenced genomes. The picocyanobacteria include the *Prochlorococcus*, *Synechococcus*, and *Cyanobium* genera and 194 total genome sequences deposited in NCBI Genbank as of February 2016. However, among those three genera, *Prochlorococcus* accounts for 151 genomes, *Synechococcus* for 39, and *Cyanobium* for 3. The outsized number of genomes sequenced among the *Synechococcus* and *Prochlorococcus* genera is due to intense study of these ubiquitous marine cyanobacteria for their importance in global biogeochemical cycles (though some of the *Synechococcus* genomes are from freshwater strains). While this is important, it does not contribute much to the study of freshwater systems and the cyanobacteria that inhabit them. The *Cyanobium* genus, on the other hand, has been shown to be globally important in freshwater ecosystems and able to produce useful natural products, but lacks representative genomes. In addition, bloom-forming cyanoabacteria (for

example, from the *Nostocaceae* family) are well known to produce similar secondary metabolites [32]. In order to put the study of freshwater cyanobacterial population dynamics (blooms in particular) on a firm foundation, more high-quality, finished freshwater cyanobacterial genomes need to be sequenced. These genomes will aid in vetting population genomes that are clustered from freshwater metagenomes, reveal prophages that may contribute to bloom collapse, and illuminate the population structure and evolution of closely related cyanobacterial strains that cause freshwater cyanobacterial blooms. New natural products with potentially useful or toxic properties are likely to be found in novel freshwater cyanobacterial genomes as well.

This work introduces four new freshwater bacterial genomes (for *Anabaena* sp. WA102, *Aphanizomenon* sp. WA102, *Anabaena* sp. AL93, and *Cyanobium* sp. LC18), an instructive freshwater lake metagenome analysis, and a novel freshwater temperate picocyanobacterial bacteriophage genome. Analysis from chapter 3 shows that long-read sequencing will be important for obtaining novel high-quality cyanobacterial genomes with relevance to lake environments, such as that of *Anabaena* sp. WA102, and reaffirms that many bloom-forming cyanobacterial genomes are prone to frequent rearrangement. Chapter 4, building on the high-quality genome sequence of *Anabaena* sp. WA102 discussed in chapter 3, investigates the clonality of a cyanobacterial bloom population, begins to probe the population dynamics of freshwater cyanobacteria as revealed by metagenomics, and tracks the distribution of a particular toxic strain across the region. The novel genome from a culture of the freshwater picocyanobacterium *Cyanobium* sp. LC18

and its temperate bacteriophage, C-CRS01, is discussed in chapter 5.

Chapter 3 Structural and Functional Analysis of the Closed Genome of the Recently Isolated Toxic *Anabaena* sp. WA102

Nathan M Brown, Ryan S Mueller, Jonathan W Shepardson, Zachary C Landry, Claudia S Maier, Jeffrey T Morré, F Joan Hardy, and Theo W Dreher

BMC Genomics

BioMed Central

Floor 6, 236 Gray's Inn Road

London

WC1X 8HB

United Kingdom

In revision

3.1 Introduction

Anabaena (some isolates are also named *Dolichospermum* [158]) are filamentous, nitrogen-fixing cyanobacteria that often form blooms in eutrophic water bodies. Traditionally, they have been studied as models of multicellular development in bacteria [48]. Their ability to fix both carbon and nitrogen makes them a key part of the biogeochemical cycle. Further, they can produce a range of bioactive secondary metabolites, which have been shown to threaten public health whenever toxic blooms occur in drinking or recreational water bodies [82, 15].

Anatoxin-a is one of the most toxic secondary metabolites produced by *Anabaena* species [18]. It acts as a nicotinic acetylcholine receptor agonist in animals, paralyzing muscles and causing death by asphyxiation [19]. The toxin is synthesized via a polyketide synthase (PKS) pathway encoded by a cluster of at least eight genes [92]. Anatoxin-a is known to be synthesized by five genera of *Cyanobacteria*: *Anabaena* (*Dolichospermum*), *Oscillatoria*, *Aphanizomenon*, *Cylindrospermum*, and *Phormidium* [121]. The entire PKS gene cluster has been sequenced and confirmed to produce anatoxin-a or a variant (homoanatoxin-a and dihydroanatoxin-a) in *Anabaena* sp. strain 37, *Oscillatoria* sp. strain PCC 6506, and *Cylindrospermum stagnale* PCC 7417 [121, 12, 15, 92]. We describe a novel species of anatoxin-a-producing *Anabaena* from Anderson Lake, Washington State, USA, *Anabaena* sp. WA102.

Many cyanobacterial genomes remain in draft form (51 according to [135]). *Cyanobacteria* genomes are often resistant to standard assembly approaches when

using Illumina short-insert DNA libraries, due to the fact that they have a large percentage of mobile elements (as much as 11% of the genome) that repeat throughout the genome [66]. These repeats, and other types of repetitive DNA, are nearly identical in sequence and longer than the insert size of typical DNA sequencing libraries. This causes ambiguous alignment and scaffolding of contigs on either side of the repeat and fragments the genome assembly [114]. While most of the gene content of these genomes properly assembles, reads from mobile element regions usually do not and are omitted from analysis. Structural variation in the genome, such as large deletions or tandem duplications, is also obscured in unfinished genome assemblies. Until recently, the only sequencing methods that have allowed assembly across repeat regions and produced finished *Nostocaceae* genomes have been Sanger sequencing and hybrid assembly of 454 and Illumina sequencing libraries that require laborious extra finishing steps. Increasing access to long-read sequencing platforms will circumvent these problems and help to close complex bacterial genomes in a single assembly step [70].

We describe a PacBio sequencing dataset of 8.5 kbp average read length that was used to finish and close the genome of *Anabaena* sp. WA102. We compare the long-read sequencing results to genome assembly from short-read sequences and describe structural features of potential physiological relevance that are missed in analysis of assemblies derived from short-read sequencing. We also compare the complete genome of the cultured isolate (Dec 2014) to the genome of a closely related population genome in Anderson Lake (Jul 2012).

3.2 Results

3.2.1 The *Anabaena* sp. WA102 culture and genome

Anabaena sp. WA102 was isolated from a water sample collected during a cyanobacterial bloom in Anderson Lake in Jefferson County, Washington, USA on May 20th, 2013 (Figure 3.1A). Anatoxin-a levels in the lake were 12.5μg/L. The non-axenic culture was first established in BG-11_0 medium, then a single contiguous colony - assumed to be clonal - was isolated from the established culture and serially propagated in BG-11_0. Colonies from the culture are heterocystous due to lack of nitrogen in the medium and have mean vegetative cell dimensions of 7.1 by 6 μm (Figure 3.1B). HPLC-MS/MS analysis showed that the culture produced anatoxin-a, with no detectable homoanatoxin-a nor dihydroanatoxin-a (Figure 3.2).

DNA extracted in December 2014 (19 months after culture establishment) was used to construct a library of size-selected fragments (over 8 kbp). Four PacBio SMRT cells were used to sequence a total of 1.13 Gbp with an average read length of 8.5 kbp (Table 3.1 and Figure 3.3). Two contigs representing the 5.7 Mbp chromosome (Figure 3.4A) and a 76.5 kbp plasmid that make up the complete *Anabaena* sp. WA102 genome were *de novo* assembled from the output of two PacBio SMRT cells (Genbank:CP011456-7). At an average nucleotide coverage of 49.8x, the average Phred quality score for the genome is 48.86 (a 1/76,913 probability of the assignment of an erroneous nucleotide).

The average GC content of the *Anabaena* sp. WA102 chromosome is 38.4%.

There are 5091 predicted genes on the chromosome, including 4667 protein-coding sequences (1824 of which encode hypothetical proteins), 365 pseudogenes, 5 ribosomal RNA operons, and 43 tRNA genes (Table 1). DnaA boxes and a surrounding AT-rich region identify a single putative origin of replication from nucleotides 1457-1702. The genome has an unusual GC skew pattern (Figure 3.4B) that does not allow for *terC* site prediction, as also seen with some other cyanobacteria [53, 161]. rRNA operons are scattered throughout the chromosome, not concentrated near the origin of replication, and in one case oriented against the presumed direction of replication. If *Anabaena* sp. WA102 is oligoploid like many cyanobacteria [52], then there may be less need to encode highly expressed genes such as the rRNA operons near the origin of replication to increase their copy number or orient them to optimize transcription during replication. The plasmid is 76.5 kbp long (1.3% of genome) and has an average GC content of 37.7%. There are 88 genes encoded on the plasmid, including 75 protein coding sequences, the majority of which are hypothetical proteins (57) or pseudogenes (13), and no rRNA or tRNA genes (Table 3.1).

3.2.2 Comparison of *Anabaena* sp. WA102 long- and short-read genome assemblies

DNA from the *Anabaena* sp. WA102 non-axenic culture was extracted in December 2014 (7 months after culture establishment) and used to construct an Illumina TruSeq metagenome. That library was sequenced as 100nt paired-end reads on

the HiSeq 2000 instrument, yielding 3.83 Gbp of total sequence, of which 738 Mbp (19%) mapped to the closed *Anabaena* sp. WA102 PacBio genome assembly. A draft *Anabaena* sp. WA102 genome was extracted from an assembly of this short-read Illumina sequencing data using the mmgenome package. The draft genome is not complete, but the sum length of contigs in the draft genome is within 1% of the length of the closed *Anabaena* sp. WA102 genome. When the draft genome is aligned against HMM profiles in an HMM profile database of essential bacterial genes from the mmgenome package, 105 essential genes found in other members of the *Nostocaceae* are also found in the new genomes (compared with 104 essential genes in the closed *Anabaena* sp. WA102 genome, see Table 3.2). This suggests that the draft genome is nearly complete and representative of actual gene content. Using blastn, 819 of 820 contigs in the *Anabaena* sp. WA102 draft genome align to the closed reference genome (e-value $\leq 10^{-30}$), further suggesting that the draft genome assembly has little contamination. Some of the contigs in the draft genome overlap when aligned to the closed genome, forming 230 regions of contiguous coverage with 229 gaps that are scattered around the circular genome (Figure 3.4A).

The gap regions sum to 34,166 bp (0.6% of the reference genome), containing 97 genes. Over half of these (56 genes) have more than one copy in the genome, including 26 genes from a single cluster of transposases. Many single-copy hypothetical genes that coincide with gaps have low complexity regions. Most gaps (green tiles on Figure 3.4A) coincide with long repeat regions in the genome, whose multiple copies are connected by green arcs (Figure 3.4A). The repeat regions include the

five rRNA operons, genes encoding transposons and homing endonucleases, and other repeat regions discussed in more detail below. In some cases gaps coincide with GC-rich regions. These results agree with previous observations of gaps in Illumina assemblies due to long repeat regions and regions of low nucleotide complexity [114]. The large number of contigs generated from the short-read Illumina sequences emphasizes the prevalence of long repetitive elements in the *Anabaena* sp. WA102 genome and the value of long-read sequencing technologies in producing finished genomes. This is further demonstrated by observations of tandem repeats in the long-read assembly, observation of structural variants in the population, analysis of genome synteny with another closely related *Anabaena* genome, and a full count of mobile elements within the genome (described below).

3.2.3 The *Anabaena* sp. AL93 culture and genome

Anabaena sp. AL93 is an anatoxin-a producing strain isolated in non-axenic culture from a toxic bloom in American Lake, Washington in 1993 (MA Crayton, personal communication). It provides local geographical context for *Anabaena* sp. WA102, since American Lake is only 100 km from Anderson Lake. It also provides some evolutionary context as a close relative of *Anabaena* sp. WA102 (see phylogeny below). The genome was sequenced with 1.36 Gbp of Illumina MiSeq 250-bp paired-end reads. Contigs representing 5.7 Mbp of the *Anabaena* sp. AL93 draft genome were binned using the mmgenome package to yield a nearly complete genome with 105 essential genes according to the database in the mmgenome

package.

3.2.4 Phylogenomic relationship between *Anabaena* sp. WA102, AL93, and other fully sequenced *Nostocacaea*

The closed genome from *Anabaena* sp. WA102 and the draft genome from *Anabaena* sp. AL93 can be placed phylogenetically among draft and full genomes from members of the *Nostocaceae*. The ancestral relationship of eleven genomes from the *Nostocaceae* constructed with a phylogenetic tree based on 1408 clusters of unique orthologs from each genome (Figure 3.5). Unanimity among 1000 tree constructions yielded 100% bootstrap support for every internal node. The tree was rooted at *Nostoc* sp. PCC 7107, according to [135]. *Anabaena* sp. WA102 and *Anabaena* sp. AL93 are most closely related to each other. They form a distinct clade with *Anabaena* sp. 90, a microcystin toxin-producing strain from Finland [161].

3.2.5 Comparing gene content and metabolic capabilities of *Anabaena* sp. WA102 and AL93 with other *Nostocaceae* genomes

The gene contents of *Anabaena* sp. WA102 and closely related *Nostocaceae* genomes were also assigned to metabolic pathways using the KEGG ortholog database. All genes necessary for nitrogen fixation (*nifDKH*) were found throughout these genomes. Figure 3.6 highlights metabolic pathways with differential

representation in *Anabaena* sp. WA102 and its relatives. Differences in sulfur metabolism are evident among the genomes. The *ssu* operon, which is involved in transport and metabolism of organic sulfur compounds [37], was intact in *Anabaena* sp. WA102. It was absent or incomplete in 6 of 11 *Nostocaceae*, including *Anabaena* sp. 90. *ssuABCDE* and *tauD* (taurine metabolism) are in the same gene cluster in *Anabaena* sp. WA102 and are likely co-regulated. *Anabaena* sp. WA102 also possesses the *fhuBC* genes, which encode two parts of the ferric hydroxamate ABC transporter. The presence of these genes suggest that *Anabaena* sp. WA102 is well equipped to import organic sulfur compounds and iron from the environment. This may provide a competitive advantage in providing the iron-sulfur clusters that are necessary for nitrogen fixation in niches with low sulfate availability.

Other genes present in *Anabaena* sp. WA102 but not in *Anabaena* sp. 90 or other *Nostocacaea* (Figure 3.6) may also provide competitive advantage under certain conditions. *btuB* is necessary for vitamin B12 uptake from the environment [72]. The *urtABCDE* cluster allows uptake and metabolism of nitrogen-rich urea [3]. *cydAB* encode the cytochrome *bd*-type oxidase, which has been shown to be necessary for *Nostoc* sp. PCC 7120 survival under nitrogen-limited conditions and is hypothesized to scavenge oxygen in heterocysts to prevent oxidation of nitrogenase [95]. The presence of *pixGHIJL* genes, which encode a phototactic system, suggests that *Anabaena* sp. WA102 is positively phototactic and likely motile [170]. In support of this hypothesis, *Anabaena* sp. WA102 encodes a twitching-motility pilus gene *pilT*, and a pilus assembly gene *pilC* (loci AA650_16975 and

16980). Gas vesicle genes present in two clusters (loci AA650_0781 to 07850 and AA650_07865 to 07870) support mobility through buoyancy control [159].

A number of metabolic genes are absent from *Anabaena* sp. WA102, but present in *Anabaena* sp. 90 or other *Nostocaceae*. *pecABCEF*, the genes responsible for phycoerythrocyanin synthesis [156]), are absent from *Anabaena* sp. WA102 but present in its close relative *Anabaena* sp. AL93 [152]. Phycoerythrocyanin is a photosynthetic pigment that absorbs light maximally at 575nm (green light) and confers a competitive advantage in coastal and freshwater environments where phytoplankton and turbid waters absorb much of the red light that is maximally absorbed by the ubiquitous phycocyanin pigment [156]. These two strains can be distinguished by their pigments, a critical element in niche adaptation. Both strains encode genes to synthesize the phycobilins phycocyanin and allophycocyanin, but only *Anabaena* sp. AL93 encodes the genes for phycoerythrocyanin synthesis. The absence of a phycoerythrocyanin operon suggests that *Anabaena* sp. WA102 would not compete well in shade from other photosynthetic organisms or deeper and murkier water because it cannot efficiently absorb green light. Rather, it may avoid shade or deeper water by positive phototaxis to the lake surface driven by gas vesicle buoyancy. The *psbJLM* components of the photosystem II apparatus are intermittently distributed throughout the *Nostocaceae* in this study but are completely absent from *Anabaena* sp. WA102. Different combinations of light harvesting genes in each genome, without a phylogenetic pattern, suggest that they are selected for under different light conditions and perhaps horizontally transferred.

3.2.6 Capacity for synthesis of anatoxin-a and other secondary metabolites

Cyanobacteria produce many secondary metabolites, including products of nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) genes. Much concern about freshwater cyanobacterial blooms stems from their production of toxic secondary metabolites. Fourteen gene clusters in the *Anabaena* sp. WA102 genome encode putative secondary metabolite synthesis proteins (Figure 3.4A). Anatoxin-a is made by proteins encoded in cluster eleven located between nucleotides 4,362,415 and 4,392,159, confirming that *Anabaena* sp. WA102 indeed is able to produce anatoxin-a, as detected by HPLC-MS/MS (Figure 3.2). The *anaA-G* genes in this 30 kbp cluster are syntenous with homologs in *Anabaena* sp. 37 and *Anabaena* sp. AL93 (Figure 3.7). However, genes *anaA*, *anaI*, and *anaJ* are rearranged between the *Anabaena* anatoxin-a clusters and the *Oscillatoria* and *Cylindrospermum* clusters [92].

Comparing *ana* clusters between *Anabaena* sp. WA102, *Anabaena* sp. AL93, *Anabaena* sp. 37 (Genbank:JF803645), *Cylindrospermum stagnale* sp. PCC 7417 (Genbank:NC019757), and *Oscillatoria* sp. PCC 6506 (Genbank:FJ477836) showed differences in the *anaG* gene (Figure 3.7 and 3.8). The AnaG protein plays a key role in determining the anatoxin variant produced [92]. AnaG adds an acetyl group and either one or two methyl groups to the bicyclic thioester precursor, forming either anatoxin-a or homoanatoxin-a, respectively. *Oscillatoria* sp. PCC 6506, which produces 99% homoanatoxin-a and 1% anatoxin-a, possesses

the largest methyltransferase domain in AnaG. The smaller AnaG methyltransferase domain in *Anabaena* sp. 37, a producer of anatoxin-a [121], is evidently not involved in homoanatoxin-a synthesis. The AnaG methyltransferase domain is missing entirely in *Anabaena* sp. WA102 and *Anabaena* sp. AL93, which are also producers of anatoxin-a (Figure 3.7 and 3.8). In *Cylindrospermum* sp. PCC 7417, which produces dihydroanatoxin-a, AnaG lacks the methyltransferase domain as well as the phosphopantetheine transferase domain on the extreme C-terminus (Figure 3.8). In the same strain, an oxidoreductase gene, *Cylst6226*, not present in the other *ana* clusters, is present (Figure 3.7) and implicated in dihydroanatoxin-a synthesis [92]. Note that annotation of genes *anaH-J* differs between [92] and [15]; we have chosen to follow [92].

The anatoxin-a synthetase gene cluster from *Anabaena* sp. AL93 revealed an organization most similar to that of *Anabaena* sp. WA102, although the AL93 AnaG gene is shorter in the C-terminal region. There are also differences in genes situated between *anaG* and *anaI*, which include genes not thought to be involved in anatoxin synthesis. Notably, all clusters (not shown for *Oscillatoria* sp. PCC 6506 in Figure 3.7, but referred to in [92]), share a MATE efflux pump homolog (*anaI*). MATE efflux pumps encoded within the saxitoxin gene cluster are known to export saxitoxin, another toxic secondary metabolite, from the producing cell [113]. They may play a similar role with anatoxin-a.

3.2.7 Lack of synteny with *Anabaena* sp. 90

Among the completely sequenced *Anabaena* genomes, *Anabaena* sp. WA102 is most closely related to *Anabaena* sp. 90, sharing an average nucleotide identity (ANI) of 91.5% and 2331 gene homologs. Despite this relatively close relationship, there are major differences in overall genome architecture. Whereas the *Anabaena* sp. 90 genome has two chromosomes of 4.33 and 0.82 Mbp, *Anabaena* sp. WA102 has a single chromosome. Local nucleotide alignment showed that there is little long-range synteny between the two *Anabaena* genomes (Figure 3.9).

Novichkov et al illustrated common paradigms of synteny between genomes within a genus using dotplots [102]. Aligning genome sequences between species of *Pseudomonas* yielded long stretches of synteny, but aligning genomes sequences between species of *Streptococcus* showed no synteny. Those dotplots are recreated and shown beside the dotplot for *Anabaena* sp. 90 and WA102 (Figure 3.10). Orthologs from each pair of aligned genomes were aligned by BLASTP, showing that average amino acid identity between orthologs of the *Anabaena* genomes was the highest (Figure 3.10). The dotplot of the *Anabaena* genomes is very fragmented, although these genomes are relatively closely related. The distinct X-shape to dotplots of *Pseudomonas* and *Streptococcus* genomes indicate chromosomal inversions around the origin of replication [38]. This pattern is missing in the dotplot of *Anabaena* genomes, indicating the infrequency or absence of these inversions. Figure 3.10 indicates that the *Anabaena* genomes have experienced a relatively faster rate of recombination versus point mutation. This is not uncommon among

bacterial genomes but varies among different taxa [141]. Length distributions of the local colinear blocks (LCB's) from alignments calculated by Mauve (Figure 3.11) support the general disruption of gene order between *Anabaena* sp. WA102 and 90. The largest local colinear blocks encompass biosynthetic gene clusters and a cryptic prophage discussed below. The LCBs are not clearly bounded by either repeat sequences or mobile elements, which does not lend a clear explanation for their rearrangement between the two bacteria.

In addition to long-range shuffling, we also detected local rearrangement of genes within clusters. For instance, an LCB at nucleotides 1,992,912-2,007,469 that includes thirteen genes in *Anabaena* sp. WA102 corresponds to the region between nucleotides 3,575,881-3,591,878 in *Anabaena* sp. 90 that includes fourteen genes (Figure 3.12). Genes in this syntenous region are putatively involved in complex carbohydrate biosynthesis and export (being mostly glycosyltransferases and including an ABC transporter). Of these, two glycosyltransferases, an acyltransferase, and a hypothetical protein are unique to *Anabaena* sp. WA102 and six glycosyltransferases are unique to *Anabaena* sp. 90. The remaining nine genes in *Anabaena* sp. WA102 and eight genes in *Anabaena* sp. 90 are homologous or share homologous domains. Two transposases are responsible for interrupting just one portion of synteny in this region, leaving 4 breaks in synteny unexplained. This suggests that recombination interrupts synteny even in otherwise conserved gene clusters, though the mechanism for recombination is not always clear.

3.2.8 The mobilome

108 transposases (79 intact and 29 pseudogenes) were automatically annotated by the NCBI pipeline, constituting 2% of the genome. Manual annotation with the aid of the IS Finder database [137] increased the number of intact and fragmented transposases to 130. In addition to transposases, 30 HNH homing endonuclease reverse transcriptases are encoded in the *Anabaena* sp. WA102 genome, bringing the total number of intact and degenerate mobile elements to 160. Phylogenetic relationships between insertion sequences show that two groups of closely related IS4-family insertion sequences predominate (20 in the IS10-like group and 25 in the IS4Sa-like group) among a wider representation of IS families (Figure 3.13). Aligning nucleotide sequences adjacent to each side of the coding sequence of these insertion sequences revealed the unique inverted repeat sequence for each group: ATTCAACAYTTCTG for the IS10-like group, and CCGCCTTGTCACCCGT-TAAG for the IS4Sa-like group. These two groups of transposases catalyze their transposition via three acidic residues in their active site: two aspartates and a glutamate, and transpose in a cut-and-paste fashion (non-replicative) [27].

Other common mobile elements found in bacterial genomes are prophages, cryptic prophages, and phage-like elements such as gene transfer agents (GTAs). No signature phage regions were detected with the PHAST phage-detection webserver. The IslandViewer 3 webserver, which detects genomic islands, highlighted an 18 kbp region between nucleotides 1,179,961 and 1,198,734. This region is also contained in the largest local colinear block (LCB) calculated by Mauve between

Anabaena sp. WA102 and *Anabaena* sp. 90. Within the LCB, there is a 19 kbp insertion in *Anabaena* sp. WA102 relative to *Anabaena* sp. 90. The LCB boundaries likely denote the exact boundaries of a cryptic prophage: nucleotides 1,179,211-1,198,554. This region contains a putative phage terminase large subunit that was automatically annotated by Prokka and confirmed with 100% confidence by Phyre2 structure-guided annotation. The terminase large subunit is a component of a DNA packaging protein unique to *Caudovirales*. Within this region also lie 21 hypothetical proteins, one IS-4 family transposase, one pseudogene, and one integrase. The large proportion of hypothetical proteins is consistent with a phage origin. The integrase lies 134 nucleotides downstream of a methionine tRNA, which may have served as an integration site (attB) of the prophage. The GC content in the region is 32.9%, lower than the genome average of 37.7% and consistent with a horizontally transferred region that has a distinct nucleotide composition. The small size of the region, lack of other identifiable phage proteins such as capsid or tail structure proteins, and the insertion of a transposon common to the bacterial genome suggest that this region is a partly degraded cryptic prophage. Several other phage integrases were automatically annotated, but these integrases are often functionally mislabeled. Alternatively, they may be site-specific integrases native to or co-opted by the bacterial genome for functions other than prophage integration and excision. These alternative functions are likely, considering the absence of other readily identifiable phage genes near these integrases.

Besides transposons and phage-like elements, a single plasmid was identified, rounding out the mobile element complement of *Anabaena* sp. WA102. The plas-

mid was identified as a 92 kbp contig assembled from PacBio reads. Fifteen kbp of nucleotide sequence from each end of the contig aligned with 99% similarity (overlapped with lower quality sequence at the extremities) and was trimmed from the final plasmid sequence. The trimmed plasmid sequence is 77 kbp long, with a 37.2% average GC content. The 88 genes on the plasmid include 75 intact and 13 pseudogenes. A *parAB* operon on the plasmid suggests that it is a low-copy plasmid (confirmed by an average read coverage less than that of the chromosome) with a well described partitioning mechanism [125]. The *parAB* operon and surrounding nucleotide sequence bears at least 86% similarity to the *parAB* operon and its surrounding sequence on the chromosome (Figure 3.14). Interestingly, the plasmid carries at least part of a non-ribosomal peptide synthase (NRPS) cluster. One protein within the cluster shows significant similarity to AdpD from the anabaenopeptilide cluster in *Anabaena* sp. 90 (BLASTP e-value = 6.9×10^{-121}). The other three biosynthetic proteins in the cluster show similarity to a malonyl CoA-acyl carrier protein transacylase, a β-ketoacyl synthase, and a short-chain dehydrogenase. Plasmid-borne NRPS clusters are not uncommon. A recent comprehensive survey of NRPS and polyketide synthase (PKS) clusters in all bacterial genomic data deposited at the National Center for Biotechnology Information (NCBI) revealed that 10% of NRPS/PKS clusters in *Cyanobacteria* are located on plasmids [160]. Importantly, the plasmid encodes four putative site-specific integrases, which may facilitate integration into a bacterial chromosome. Coupled with nucleotide similarity between the plasmid and the chromosome, where site-specific integrases can also be found, this indicates that the region of plasmid similarity

on the chromosome may be considered a genomic island.

3.2.9 Relationship between the *Anabaena* sp. WA102 genome and the Anderson Lake metagenome

To relate the *Anabaena* sp. WA102 genome to the bloom in Anderson Lake, the WA25 metagenome was sampled from Anderson Lake on July 7th, 2012. The sample was taken near the peak of a cyanobacterial bloom, when the anatoxin-a level was 187 μg/L (`https://www.nwtoxicalgae.org/Data.aspx`). The metagenome contains a genome from a strain of *Anabaena* sp. WA102 that is nearly identical to the culture and is likely an ancestor from 10 months before the culture strain was isolated and 2.5 years before it was sequenced. Reads from the July 2012 metagenome (short-read Illumina), the December 2013 culture (short-read Illumina), and the December 2014 culture (long-read PacBio) were mapped to the closed reference genome to observe differences in the genome.

3.2.10 A recent deletion event in the *Anabaena* sp. WA102 genome

The length of the PacBio reads not only allowed us to close the *Anabaena* sp. WA102 genome but also revealed structural variation in the population. The 21 kbp segment between nucleotides 4,790,517 and 4,812,024 was also present (99% similarity) on a 25 kbp contig in the PacBio assembly, reflecting the existence of a 4kbp indel variant within the genomes of the *Anabaena* sp. WA 102 culture

population (Figure 3.15). Mapping reads from the *Anabaena* sp. WA102 PacBio dataset showed that the contig had an average coverage of 25x, approximately one-third of the average coverage of the chromosome (73x), and that the deletion actually lies between nucleotides 4,800,950 and 4,804,900. This suggests that the deletion is present in two-thirds of the *Anabaena* sp. WA102 culture population. The indel appears to be a deletion that arose after December 2013, since the longer sequence is predominant in sequencing reads from both the July 2012 metagenome and the December 2013 culture (Figure 3.15). An XseA homolog (the large subunit of exonuclease VII) and two hypothetical gene products are deleted in the variant. In well characterized *Escherichia coli xseA* mutants, there is an increased recombination phenotype [20], suggesting the same may be true for two-thirds of the *Anabaena* sp. WA102 culture population.

3.2.11 Tandem repeat of the anatoxin-a *anaBCD* promoter region

Intriguingly, the anatoxin-a synthase region in the PacBio assembly of *Anabaena* sp. WA102 showed that the first 173 bp of the *anaB* gene and 398 bp upstream of the gene had been triplicated (Figures 3.7 and 3.16). This is in contrast with the genome of *Anabaena* sp. AL93, which does not have a triplication of the *anaB* promoter region. The 398 nucleotides upstream of *anaB* include four high-scoring putative promoters, identified *in silico* using Virtual Footprint and the PRODORIC database of position weight matrices for bacterial promoters [97]. Assembling Illumina reads from the *Anabaena* sp. WA102 culture with IDBA

v1.1.1 and PriceTI fails to correctly resolve the tandemly triplicated promoter region (Figure 3.16A). To determine when this triplication arose, reads from the July 2012, Dec 2013, and Dec 2014 sequencing runs were mapped to the triplicated region (Figure 3.16B). Illumina reads from the Anderson Lake metagenome and the *Anabaena* sp. WA102 culture mapped across the two unique junctions formed by the triple tandem repeats, confirming its presence as early as 2012 in Anderson Lake and also in the culture sequenced in December of 2013. In contrast, none of the reads from the *Anabaena* sp. AL93 culture mapped across the unique junctions formed by the tandem repeats (indicated by arrows in Figure 3.16B). This triplication is unique to *Anabaena* sp. WA102 among all known anatoxin-a cluster sequences and has been stable for at least 2.5 years, in both Anderson Lake and under culture conditions. Toxin production has been measured in the culture (Figure 3.2), so the tandem repeat is not interrupting transcription of the *anaBCD* operon. Instead, triplication of the putative promoter region may increase transcription of the operon.

3.3 Discussion

3.3.1 The recently cultured toxic isolate, *Anabaena* sp. WA102, closely reflects the parent strain in Anderson Lake

Anabaena sp. WA102 is a novel anatoxin-a-producing member of the *Nostocaceae* isolated from Anderson Lake on the Olympic Peninsula in Washington in 2013. It is

in stable non-axenic culture. The *Anabaena* sp. WA102 genome is unique among sequenced *Anabaena* genomes because it was sequenced within seven months of isolation. Other *Anabaena* strains have been in culture for several decades prior to whole genome sequencing and changes in a strain's genome can accumulate over such long periods. Sequencing a strain soon after isolation increases the relevance of the sequenced genome to the environment from which it was isolated and provides a reference point for later studies of the strains genome.

Anabaena sp. WA102 produces anatoxin-a in culture (Figure 3.2). The toxin is produced by NRPS and PKS enzymes encoded by the *anaA-J* gene cluster. A triple tandem repeat of the *anaB* putative promoter region in the cultured isolate (Figure 3.16A) is present in a nearly identical strain in the environment (July 2012 Anderson Lake sample, WA25 in Table 2), which suggests that it originates from and is relevant to the lake environment. Tandem repeats of genes and promoters commonly arise in bacterial genomes but are unstable and can collapse through homologous recombination or strand slippage at high frequency, unless the repeat is under selection [84, 123]. Thus, tandem repeats have been hypothesized to act as a crude selection-regulated response to environmental change [2, 122, 69]. Additionally, tandem repeats provide redundancy that drives the innovation, amplification, divergence (IAD) cycle that generates genetic novelty [99]. Tandemly repeated promoters, in particular, allow for promoter regions to generate or acquire new regulatory binding sites that can change the expression pattern of an operon [131]. Further study of this tandem repeat may be fruitful for several reasons. Most noteworthy is that these tandem repeats are 617nt long and identical, which makes

them highly susceptible to homologous recombination that can either expand or collapse the repeats [5]. Tandem repeats tend to be deleted rather than expanded unless deletion is selected against. This instability may be exacerbated by the deletion of the *xseA* gene in part of the population (Figure 3.15A), which causes a hyper recombination phenotype in *Escherichia coli*. That the tandem duplication can be detected in *Anabaena* sp. WA102 over a span of two years, including in Anderson Lake, suggests that a selective pressure in the lake and in the culture may be maintaining the triplication. Key questions are whether the tandem repeat increases expression of the *anaBCD* operon and production of anatoxin-a, and whether elevated expression is under selection. Determining the selective pressure preserving the tandem repeat in the *Anabaena* sp. WA102 culture may illuminate the function of anatoxin-a in the environment.

3.3.2 Closing the genome reveals details about genome architecture

Long-read sequencing technology will increasingly allow for bacterial genomes to be assembled in a single step [71]. Closing the *Anabaena* sp. WA102 genome with as few as two PacBio SMRT cells demonstrates that it is pragmatic to use non-axenic environmental enrichments of targeted bacterial species in order to obtain their finished genomes. The long-read library (PacBio C6-P4 technology) used in this study yielded an average read length of 8.5 kbp, which is long enough to span long-repeat regions in most bacterial genomes including refractory genomes such as those of the bloom-forming cyanobacteria *Anabaena* and *Microcystis* [70,

166]. Greater access to long-read sequencing raises expectations for the quality of bacterial genome assembly and will yield new insight into the mobilome and structural variation in bacterial populations. The mobilome in many bacteria may be under-represented because mobile elements that are repeated throughout bacterial genomes cannot be assembled correctly with short-insert DNA libraries. Observing structural variation such as erosion of synteny (Figure 3.9 and 3.12) and accumulation of local repeats (Figure 3.16) will enhance our understanding of bacterial evolution. In fact, short-insert libraries can be incorrectly assembled to suggest features that do not exist. An example of that is the missrepresentation of the *anaB* tandem repeat region in the *Anabaena* sp. WA102 genome (Figure 3.16A). *De novo* assembly of short-insert genomic libraries is not sufficient to determine the number of replicons in a genome or overall gene order. Further, this method is liable to miss structural variants within a population, such as the fractional presence of an *xseA*-bearing insertion (Figure 3.15A). While short-read sequencing possesses distinct shortcomings in describing structural features of a genome, nearly all single-copy genes that make up the majority of a bacterial genome can be assembled from short-read Illumina sequencing runs (Table 3.1 and Figure 3.4A).

3.3.3 Predicted ecologic profile of *Anabaena* sp. WA102

Mapping proteins from *Anabaena* sp. WA102 to the KEGG ortholog database indicates a metabolism acclimated to a nutrient-rich freshwater environment with

ample sunlight. The inability to produce phycoerythrocyanin, produced by some related *Anabaena*, coupled with positive phototaxis and gas vesicle operons suggest that it competes for light by outmaneuvering other photosynthetic organisms and rising to the surface of the water to avoid niches with less green light. Competition experiments between other nitrogen-fixing autotrophs and *Anabaena* sp. WA102 could test these hypotheses. Freshwater cyanobacteria are known to secrete hydroxamate-based siderophores to chelate iron in water [162]. These siderophores, including those encoded by the *fhu* genes in *Anabaena* sp. WA102, are then transported across the cell membrane by ferric-hydroxamate transporters [145]. Efficiently scavenging sulfur and iron would help maintain iron-sulfur clusters that are heavily used in nitrogen fixation and photosynthesis, so the predicted ability of *Anabaena* sp. WA102 to assimilate organic sulfur and oxidized iron from the lake environment may confer a growth advantage in some conditions over cyanobacteria lacking *ssu*, *tau* and *fhu* genes (Figure 3.6).

3.3.4 Evolution of the *Anabaena* sp. WA102 genome

A genomic island and a complementary plasmid carrying novel genetic cargo (Figure 3.14), tandem triplication of a promoter (Figure 3.16), observed deletion of a 4kb fragment of the genome (Figure 3.15), the ubiquity of mobile elements (Figure 3.13), and the nearly total absence of synteny with *Anabaena* sp. 90 (Figure 3.9) suggest that the genome is in rapid flux. The potential for the genome to radically rearrange may allow *Anabaena* sp. WA102 to respond to gradual changes

in the environment, such as climate change, if such changes offer the opportunity to adjust gene expression profiles. The increased availability of closed genomes as long-read sequencing becomes more widely used will allow us to quantify the rate of genome recombination in *Anabaena* and other bacteria. It will then be possible to test hypotheses for the most prevalent mechanisms and drivers of genome remodeling.

More genomes from closely related species need to be finished with long-read sequencing. These genomes can then be arranged in an alignable tight genome cluster and assayed for gene family growth and loss, and for rearrangements [118]. Alternatively, resequencing metagenomes of the original environment of *Anabaena* sp. WA102 - Anderson Lake - at regular intervals is currently feasible. This approach would generate a regular time series record of the evolution of the entire *Anabaena* sp. WA102 genome in its native environment with nucleotide resolution.

3.4 Methods

3.4.1 Sample collection

500 mL samples were collected from Anderson Lake, Washington State (48.0190 N, 237.1963 W) by the Jefferson County Public Health Department during the 2012 and 2013 cyanobacterial toxic bloom seasons. Samples were collected at a depth of 0-0.5m and may have included a dense windblown scum. Samples were shipped overnight on ice and several milliliters (depending on the sample

density) were filtered through 0.2μm Pall Supor 200 and 1.2μm-pore-size Whatman GF/C 24mm-diameter filters. Filters were stored at -80°C for later metagenomic sequencing. The culture was established upon sample arrival as described below.

3.4.2 Culture establishment and maintenance

A culture was established from a 0-0.5m deep bloom sample collected from Anderson Lake on May 20th, 2013. The lake sample was concentrated tenfold by low-speed centrifugation (5,000 RCF). No buoyant cells were observed. Approximately 20 μL of the concentrate was placed on a glass slide. *Anabaena* colonies were individually isolated by serially transferring the aliquot with an automatic pipette between at least five separate 50 μL MilliQ water droplets on the glass slide. Colonies were considered to be isolated when no other cells or cell debris were visible in the surrounding water droplet under 200x magnification on a Zeiss brightfield microscope. Isolated colonies were placed in 200 μL of BG-11_0 (i.e., BG-11 without nitrogen). BG-11_0 medium was prepared according to the Susan Golden Lab protocol (UC San Diego). One surviving colony was outgrown in BG-11_0 for several months, its identity was verified microscopically, and a single colony was again isolated into 200 μL of BG-11_0. The outgrown colony was then maintained long-term in non-axenic batch culture in BG-11_0 under white fluorescent illumination of approximately 20 $\mu Em^{-2}s^{-1}$ at 24°C with a light/dark cycle of 16hr/8hr. In addition to this culture, Dr. Mike Crayton from Pacific Lutheran University, Tacoma, Washington kindly shared a culture of *Anabaena* AL93 isolated in 1993

on BG-11 agar slants from American Lake, Pierce County, Washington State. It was maintained under the same conditions listed above but in BG-11 medium.

3.4.3 LC-MS/MS

Filters from lake samples were resuspended by dispersion in 500mL TNE buffer (50mM Tris-HCl (pH 7.5), 100mM NaCl, 0.1mM EDTA). Samples from resuspended filters or cultures were frozen and thawed for three cycles to release intracellular contents. Samples were centrifuged at 5,000 RCF for 5 min, and the supernatant was removed for LC-MS/MS analysis. LC-MS/MS analysis was conducted using a hybrid quadrupole-time of flight instrument (AB Sciex TripleTOF, Foster City, CA) coupled to a Shimadzu NexeraLC-30a UHPLC system (Shimadzu, Columbia, MD). The DuoSpray ion source (AB Sciex, Foster City, CA) was operated in the positive electrospray ionization mode and the following settings were used: ion source gas 1, 40 psi; ion source gas 2, 50 psi; curtain gas, 25 psi; gas temperature, 550C; and ion spray voltage, 5500 V. The declustering potential (DP) was 80 V and the collision energy (CE) was set to 27 V. The instrument was operated in positive ion polarity and high-resolution product ion mode. Precursor ion selection was performed in the quadrupole operated at unit resolution. Precursor ions screened included: m/z 166.1 (anatoxin-a, MH+, C10H16NO+), m/z 168.1 (dihydro-anatoxin-a, MH+, C10H18NO+), m/z 180.1 (homoanatoxin-a, MH+, C11H18NO+) and m/z 182.2 (dihydro-homoanatoxin-a, MH+, C11H20NO+). Product ion mass spectral data were acquired using a scan range of m/z 50650.

Auto calibrations were performed prior to each LC-MS/MS run. Chromatographic separations were carried out using an Agilent Zorbax RRHD SB-18 column (1.8μm particle size, 2.1x150mm) held at 40C. A binary solvent system was used consisting of water (solvent A, Fisher Optima LC/MS grade) and acetonitrile (solvent B, Fisher Optima LC/MS grade), both containing 0.1% formic acid (98% pure, Sigma Aldrich). The following gradient was applied: 5% B hold for 0.5 min then increase to 90% B within 5 min, reduce to 5% within 0.5 min and the hold for 5 min. Flow rate was 0.5 mL/min. Sample injection volume was 10 μL.

3.4.4 DNA extraction and amplification

DNA was extracted from cultures by concentrating the culture tenfold at 40,000 RCF and washing mucilage from the cell pellet with TNE buffer (50mM Tris-HCl (pH 7.5), 100mM NaCl, 0.1mM EDTA). The cell pellet was resuspended in TNE buffer and treated with a method from Neilan *et al.* [126] that had the following modifications. The protein fraction was removed with two 25:24:1 phenol/chloroform/isoamyl alcohol extractions followed by two chloroform extractions. Residual phenol was removed with a final diethyl-ether extraction. Total DNA from lake samples used for metagenome analysis was extracted from 1.2μm-pore-size filters by macerating the filters with a pestle and extracting DNA as described.

3.4.5 DNA sequencing

Samples are listed (Table 1). Each Illumina library was prepared and sequenced at the Oregon State University Center for Gene Research and Biotechnology, Corvallis, Oregon. The *Anabaena* sp. WA102 culture was also sequenced using the PacBio C6-P4 long-read sequencing platform at the Washington State University Molecular Biology and Genomics Core, Pullman, Washington. Prior to PacBio sequencing, DNA fragments were size-selected on the BluePippin system (Sage Science) to enrich for reads longer than 8 kbp. Raw reads were collected from four PacBio SMRT cells.

3.4.6 Draft genome binning

Illumina metagenomes were assembled using idba version 1.1.1 assembler software [112] on a 64-bit Linux server with 500GB of RAM. Prior to assembly, any reads containing ambiguous basecalls ("N") were culled. The large chromosome from the *Anabaena* sp. 90 genome (Genbank:NC019427) was used as a reference to guide assembly. Within idba, assemblies with kmer sizes ranging from 20nt to the sequence read length (100nt to 250nt) in 10nt increments were combined in the final assembly. Sequencing data from four PacBio SMRT cells for the *Anabaena* sp. WA102 culture was self-corrected, assembled, and polished using the Hierarchical Genome Assembly Process (HGAP) Pipeline at the Washington State University Molecular Biology and Genomics Core. Reads from original fastq files were mapped to the Illumina and PacBio assemblies using bwa ver-

sion 0.7.5a-r405 [80]. Average coverage depth for each contig was calculated using samtools version 0.1.18 (r982:295) and the calc.coverage.in.bam.depth.pl script from the mmgenome package (`https://github.com/MadsAlbertsen/mmgenome`) [1]. The mmgenome network.pl script generated a network of contigs based upon paired-end read data extracted from the bwa-generated SAM file. Bacterial and archaeal metagenome contigs were taxonomically classified using the PhylopythiaS+ support vector machine (SVM) classification software with only a contig fasta file and not a scaffold fasta file (`https://github.com/algbioi/ppsp`) [51]. 16S marker genes were detected in the contig file and used by PhylopythiaS+ to select an SVM training dataset automatically. Putative protein coding sequences were identified in each assembly fasta file using Prodigal version 2.6.2. To identify essential genes, putative protein sequences were aligned against a curated hmm database from the mmgenome package with the HMMER version 3.0 package (`http://hmmer.janelia.org/`) [35]. A custom data generation shell script based on the data.generation.2.1.0.sh script from mmgenome was used to combine the above processes (`https://github.com/russianconcussion/data.analysis.scripts/blob/master/mmgenome.datagen.sh`). Average coverage depth, network, taxonomic classification, and essential gene data for each assembly were imported into a data.frame structure in R. Finally, the mmgenome R package was used to generate a plot of genome clusters within the metagenomes, define and evaluate completeness of the clusters, and export well defined genome clusters as contigs in fasta format. Genome clusters in fasta format were annotated using Prokka version 1.11 [133].

3.4.7 Finished *Anabaena* sp. WA102 genome analysis

The finished *Anabaena* sp. WA102 genome was annotated using Prokka version 1.11 and the NCBI Prokaryotic Genome Annotation Pipeline after submission to Genbank

(Genbank:CP011456-7, for chromosome and plasmid, respectively). Non-ribosomal and polyketide synthesis gene clusters were annotated using the AntiSMASH webserver (`http://antismash.secondarymetabolites.org/`) [90]. The genome was scanned for prophages and genomic islands using the PHAST (`http://phast.wishartlab.com/`) and IslandViewer 3 (`http://www.pathogenomics.sfu.ca/islandviewer/`) webservers [173, 31]. Insertion sequences were manually annotated with the IS Finder database [136]. BLASTN and CIRCOS were used to detect local alignments between *Anabaena* sp. WA102 and *Anabaena* sp. 90 and plot the corresponding similarities (`http://circos.ca/`) [16, 74]. BLASTN, GenomicRanges, and CIRCOS were used to detect large repeat regions within the *Anabaena* sp. WA102 genome and map the *Anabaena* sp. WA102 Illumina assembly contigs to the finished genome [75]. Long and short repeat regions were also detected using RepeatScout to model repeat regions and RepeatMasker to annotate them (`http://www.repeatmasker.org`) [116]. Protein domains within the AnaG protein were identified with the SMART online protein domain database [78]. Whole genomes were aligned using Mauve 2.4.0 on default settings and Gepard 1.30.

3.4.8 Comparative genomics among members of the *Nostocaceae*

The putative protein-coding contents of *Anabaena* sp. WA102, *Anabaena* sp. AL93, *Dolichospermum* sp. AWQC131C, and *Dolichospermum* sp. AWQC310F was annotated using Prokka version 1.11. Protein content from *Anabaena variabilis* ATCC 29413, *Anabaena* sp. 90, *Anabaena* sp. PCC 7108, *Anabaena cylindrica* PCC 7122, *Nostoc* sp. PCC 7107, *Nostoc* sp. PCC 7120, and *Nostoc* sp. PCC 7524 were downloaded from Genbank. Protein-coding contents from each of the eleven genomes were used to build a genome-wide phylogenetic tree. The protein sequences were subjected to an all-versus-all BLASTP alignment to identify orthologs that occur once in each genome. These were clustered with the mcl algorithm and aligned with muscle [39, 36]. Protein alignments were masked with zorro to reduce noise from uninformative amino acid alignment positions and checked for a best fit among protein evolution models with ProtTest version 3.1 [165, 25]. The best-fit protein evolution model was used in RAxML to generate the final tree, which was rooted within the *Nostoc* genus outgroup at *Nostoc* sp. 7107, in accordance with Shih *et al.* [142, 135]. Proteins were also mapped to the free KEGG database from 2011 and compared across metabolic pathways [65]. A grid that correlates highlighted KEGG comparisons with the phylogenetic tree described above was generated using the adephylo package in R [60]. Proteins were also mapped to the COG database, which had been updated in 2014 to include four new functional categories [44].

3.4.9 Accession numbers used in study

Anabaena sp. WA102 [Genbank:CP011456-7], *Anabaena* sp. AL93 [Genbank:LJOU00000000], *Dolichospermum* sp. AWQC131C, *Dolichospermum* sp. AWQC310F, *Anabaena variabilis* ATCC 29413 [Genbank:NC007413], *Anabaena* sp. 90 [Genbank:NC_019427 and Genbank:CP003285], *Anabaena* sp. PCC 7108 [Genbank:KB235895], *Anabaena cylindrica* PCC 7122 [Genbank:NC_019771], *Nostoc* sp. PCC 7107 [Genbank:NC_019676], *Nostoc* sp. PCC 7120 [Genbank:NC_003272], *Nostoc* sp. PCC 7524 [Genbank:NC_019684], *Anabaena* sp. 37 anatoxin-a region [Genbank:JF803645], *Oscillatoria* sp. PCC 6506 anatoxin-a region [Genbank:FJ477836], *Cylindrospermum* sp. PCC 7417 [Genbank:NC_019757], and WA25 metagenome sample [SRA:SRP066506]

Sample	Sample date	Library prep/ Seq platform	Seq output (Gbp)	*Anabaena* sp. genome					
				N50	Mean cov.	No. contigs	Max contig (nt)	Total length (nt)	Unique core genes
Anabaena sp. WA102 culture est. from Anderson Lake May 2013	Dec 2014	Blue Pippin/ PacBio	1.13	5,715,573	50x	3*	5,705,437	5,807,452	104
	Dec 2013	TruSeq/ HiSeq2000 100bp PE	3.83	15,892	129x	819	66,878	5,698,213	105
WA25	July 2012/ Anderson Lake	TruSeq/ HiSeq2000 100bp PE	30.1	Shotgun metagenome of surface lake water	-	-	-	-	-
Anabaena sp. AL93 culture est. from American Lake 1993	Jan 2013	Nextera/ MiSeq 250bp PE	1.36	46,264	149x	314	133,848	5,757,055	105

Table 3.1: Summary statistics of sequencing data and binned *Anabaena* genomes. PE indicates paired-end reads. *Three contigs include the chromosome, plasmid, and the contig representing the insertion variant with the *xseA* gene.

Chromosome

Category	Element	NCBI	Prokka	Manual
Protein-coding genes:	Total	4667	5175	NA
	Hypothetical proteins	1824	2187	NA
	Transposases	79	82	130
	Homing endonucleases	7	30	30
	Histidine kinases	25	26	NA
RNA genes:	rRNA operons	5	5	NA
	tRNAs	43	44	NA
	Riboswitches	2	NA	NA
Pseudogenes:	Total	365	NA	NA
	Hypothetical proteins	186	NA	NA
	Transposases	29	NA	NA
	Homing endonucleases	6	NA	NA
	Histidine kinases	1	NA	NA

Plasmid

Category	Element	NCBI	Prokka	Manual
Protein-coding genes:	Total	75	96	NA
	Hypothetical proteins	57	66	NA
	Transposases	3	2	NA
	Homing endonucleases	0	0	NA
Pseudogenes:	Total	13	NA	NA
	Hypothetical proteins	10	NA	NA
	Transposases	0	NA	NA
	Homing endonucleases	1	NA	NA

Table 3.2: Summary of *Anabaena* sp. WA102 genome (Genbank:CP011456-7) annotation. Annotation according to the Prokka script and NCBI Prokaryotic Annotation Pipeline.

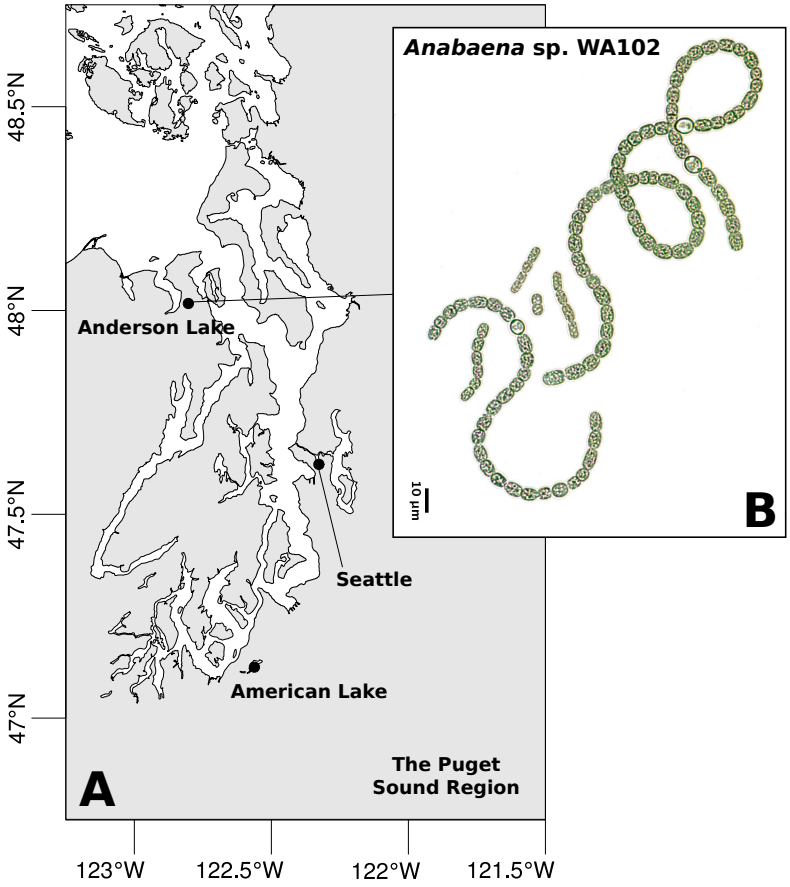

Figure 3.1: Location of Anderson Lake and picture of culture. A) A map of the Puget Sound region in Washington State, USA. *Anabaena* sp. WA102 was isolated from Anderson Lake at 48.0190 N, 237.1963 W on the Olympic Peninsula. B) A brightfield micrograph of *Anabaena* sp. WA102 at 200x magnification. Vegetative cells measure 7.1 by 6 μm on average. Colonies are heterocystous because the culture is maintained in nitrogen-free medium (BG-11_0).

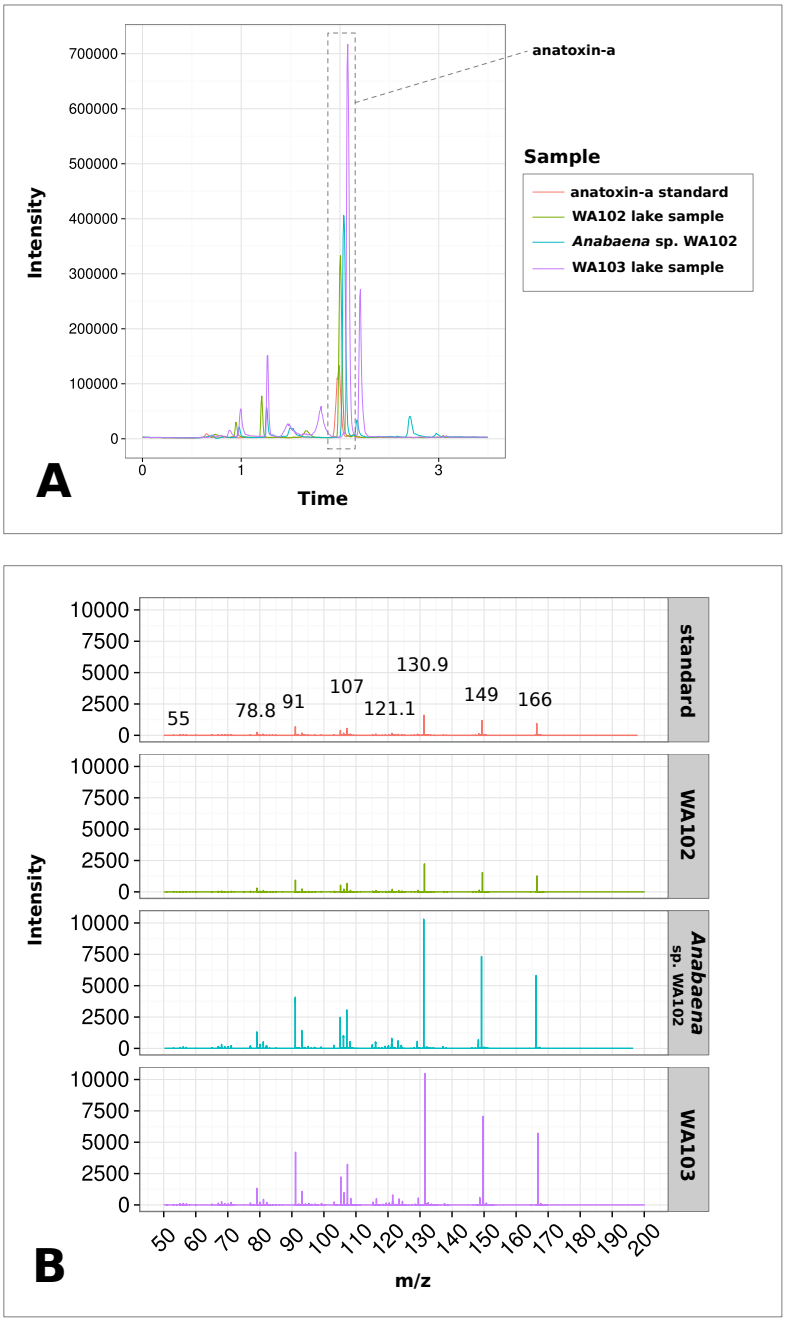

Figure 3.2: HPLC-MS/MS survey of anatoxin-a and derivatives. A) HPLC elution of compounds extracted from the *Anabaena* sp. WA102 culture and two Anderson Lake samples (WA102 and WA103). Anatoxin-a elutes at approximately 2 minutes, as indicated by the anatoxin-a standard. Anatoxin-a peaks are surrounded by a gray dashed line. No variants of anatoxin-a were detected. B) Ion mass spectra for anatoxin-a are compared from lake sample WA102 (May 20th, 2013 with 12.5 μg/L anatoxin-a), lake sample WA103 (June 17th, 2013 with 35.8 μg/L anatoxin-a), and the culture. All spectra match the spectrum of the anatoxin-a standard closely.

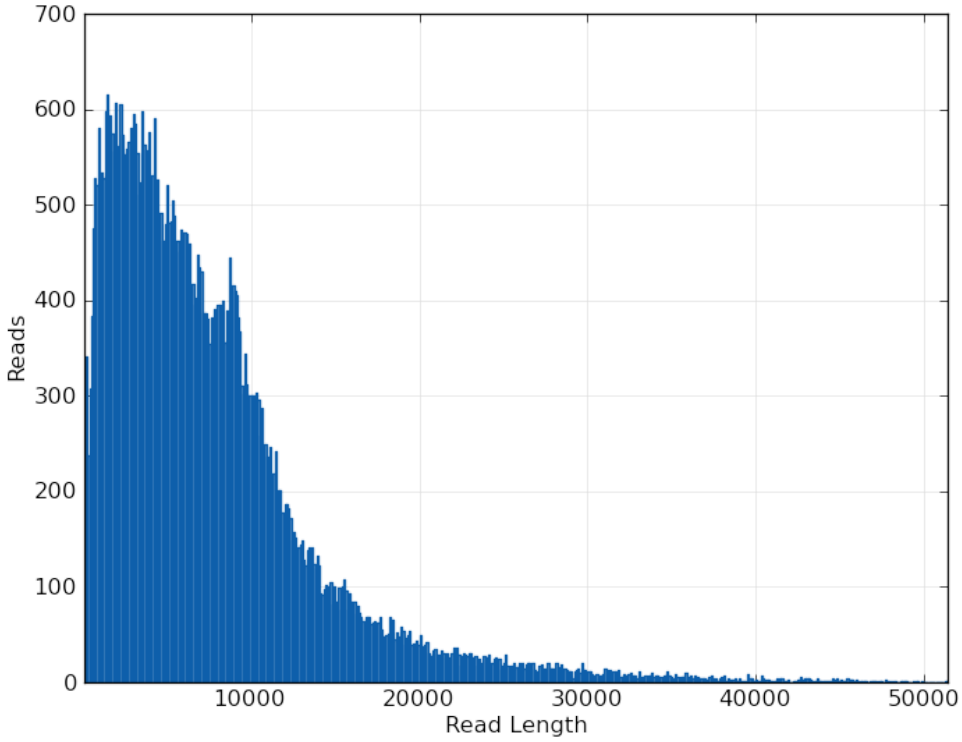

Figure 3.3: PacBio read length distribution for the *Anabaena* sp. WA102 culture. PacBio read length average 8.5 kbp, allowing complete assembly of the *Anabaena* sp. WA102 across long repeat regions.

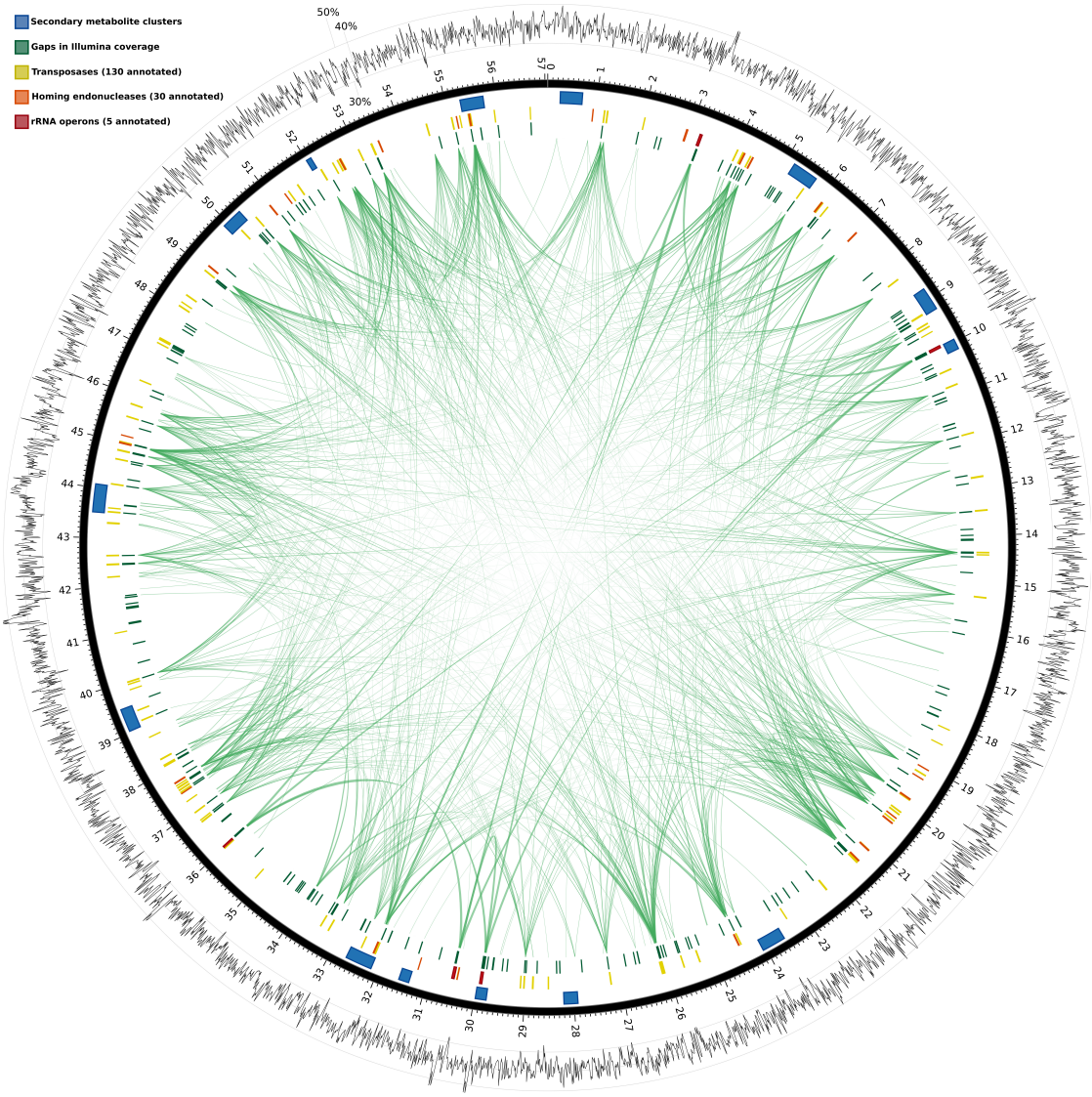

Figure 3.4: Plot of the *Anabaena* sp. WA102 genome. A) The genome is plotted as a black ring with demarcations every 100 kbp. Average GC content in 10 kbp non-overlapping windows is plotted outside of the genome ring. The first track within the genome ring includes the location of the *oriC* and RNA elements. The *oriC* was determined to lie downstream of *dnaA* among DnaA-binding motifs. The following two interior rings denote predicted protein-coding sequences, first on the positive strand (clockwise) and then on the negative strand (counter-clockwise). NRPS-PKS clusters identified by antiSMASH are shown as red tiles in the fourth interior track. Mobile elements - homing endonucleases and transposases - are plotted on the fifth interior track as orange and yellow tiles, respectively. Contigs from the binned Illumina genome of the culture (Figure 3.6) were aligned to the closed genome and 229 gaps in the Illumina assembly are represented as green tiles in the sixth interior track. Green arcs across the center connect repeated regions in the genome, determined by blastn alignment of the finished genome against itself. Note that repeat regions often coincide with gaps in the Illumina assembly. B) Genome-wide plot of cumulative GC skew. GC skew was averaged across 1 kbp non-overlapping windows of the genome and then cumulatively summed. Minimum and maximum points on the cumulative GC skew plot should indicate *oriC* and *terC*, respectively. However, the signal from the cumulative GC skew is weakened, preventing precise prediction of *oriC*, *terC*, and the replicon arms.

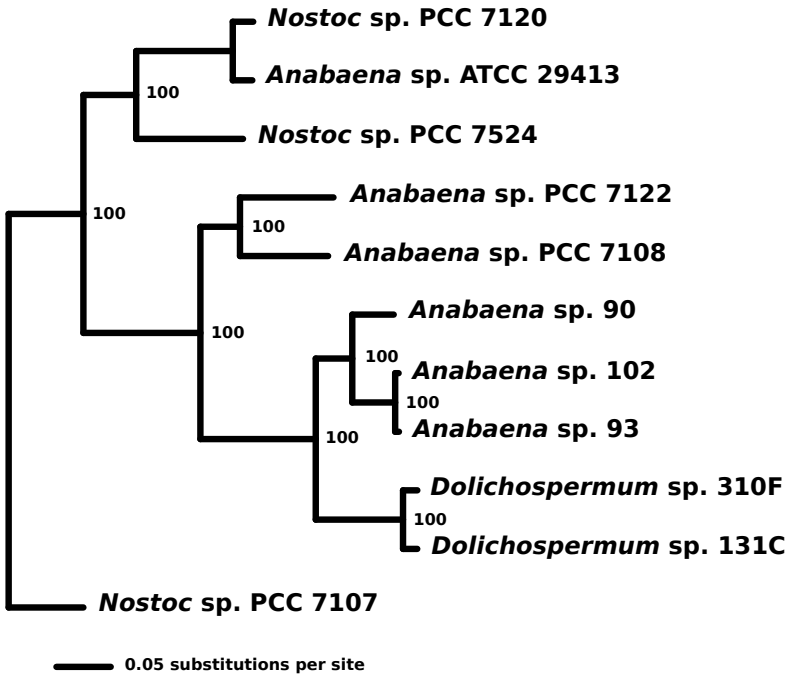

Figure 3.5: *Nostocacaea* phylogenetic tree. A phylogenetic tree constructed from amino-acid alignments of single-copy orthologs present in all genomes of some of the fully sequenced members of the *Nostocacaea*.

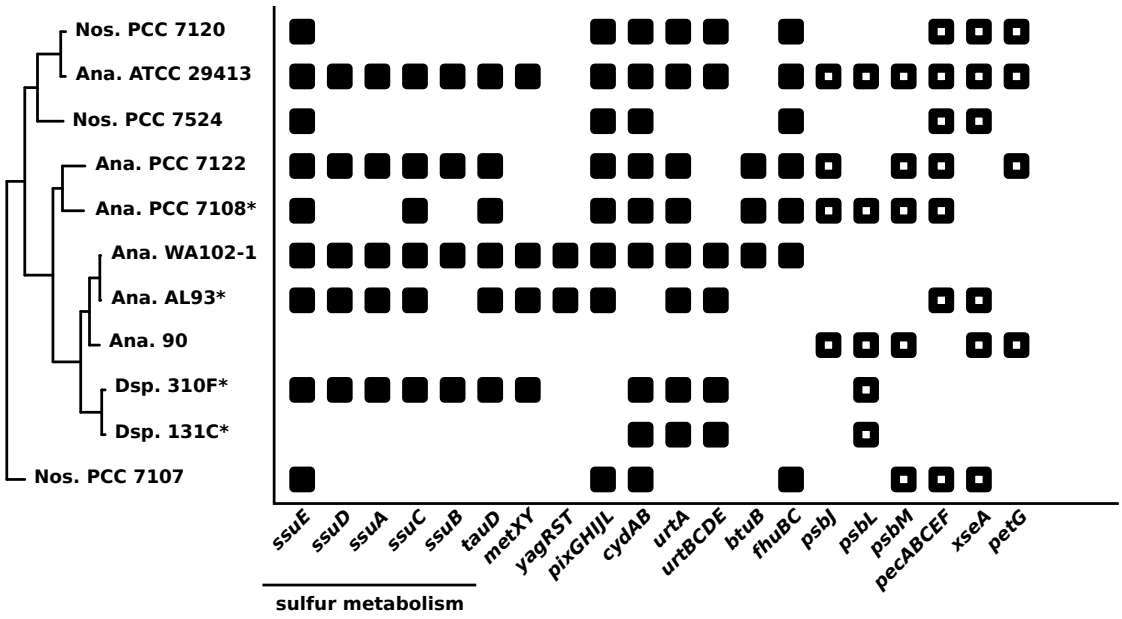

Figure 3.6: KEGG orthologs (KO) differentially represented among the compared *Nostocaceae* genomes. All proteins from each *Nostocaceae* genome were mapped to the online KO database. Orthologs with significant differences among the genomes were highlighted in the above table for comparison. *Nostocaceae* genomes are arranged according to the phylogenetic tree for easy comparison. The *Anabaena* sp. WA102 genome encodes a sulfur metabolism cluster absent or incomplete in 6 out of 11 *Nostocaceae* genomes.

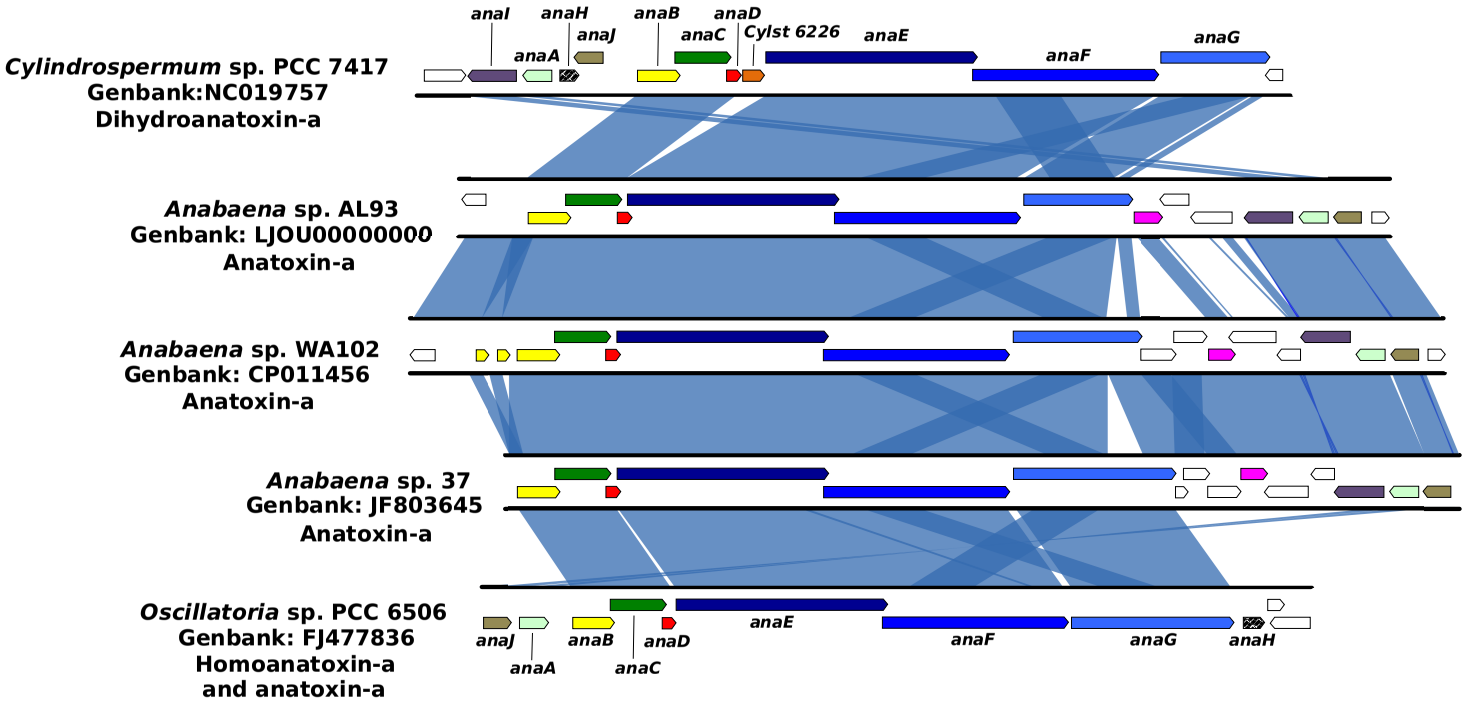

Figure 3.7: Nucleotide alignment of anatoxin-a clusters from *Cyanobacteria*. *anaA-G* and *anaI* are all conserved in *Anabaena* sp. WA102 and *Anabaena* sp. AL93, though *anaH* is missing from both. The 5' region of *anaB* and upstream promoter region is triplicated in *Anabaena* sp. WA102. The anatoxin-a cluster from *Anabaena* sp. WA102 is most similar to that from *Anabaena* sp. 37. The three *Anabaena* strains share a gene of unknown function downstream of *anaG* (colored pink). The *anaG* genes differ in size, correlated with different variants of anatoxin-a. Shorter variants of AnaG omit or truncate a putative methyl transferase domain. The *anaF* and *anaG* genes share a region of 86% nucleotide identity that is likely a homologous protein domain. *Anabaena* sp. WA102 and AL93 encode two of the shortest *anaG* genes and produce anatoxin-a, *Cylindrospermum* sp. PCC 7417 produces dihydroanatoxin-a (likely due to the unique gene Cylst 6226), and *Oscillatoria* sp. PCC 6506 primarily produces homoanatoxin-a.

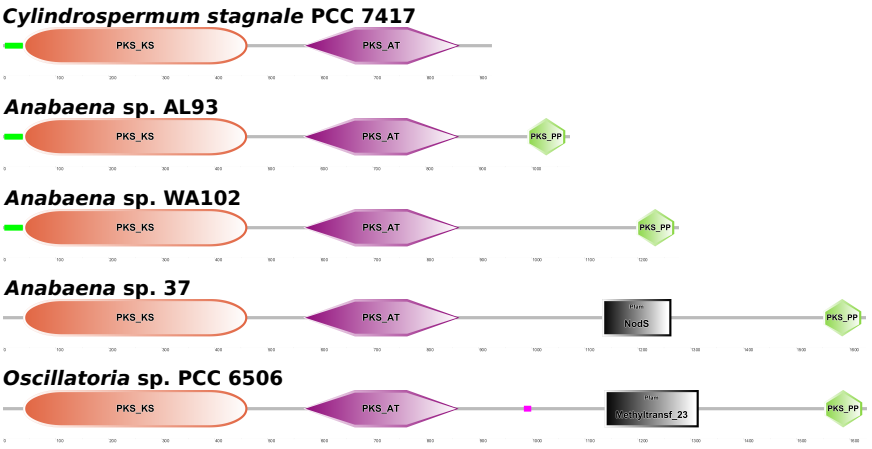

Figure 3.8: Comparison of the AnaG protein domains among Cyanobacteria. The AnaG protein sequences from *Oscillatoria* sp. PCC 6506 and *Anabaena* sp. 37 have methyltransferase domains not present in any other AnaG protein sequences. The methyltransferase domains are divergent. The methyltransferase in *Oscillatoria* sp. PCC 6506 is proposed to contribute a methyl group that makes the homoanatoxin-a variant of anatoxin-a. AnaG lacking a methyltransferase domain (or containing a non-functional domain) likely prevents production of homoanatoxin-a. In support of that, no homoanatoxin-a was detected in the *Anabaena* sp. WA102 culture.

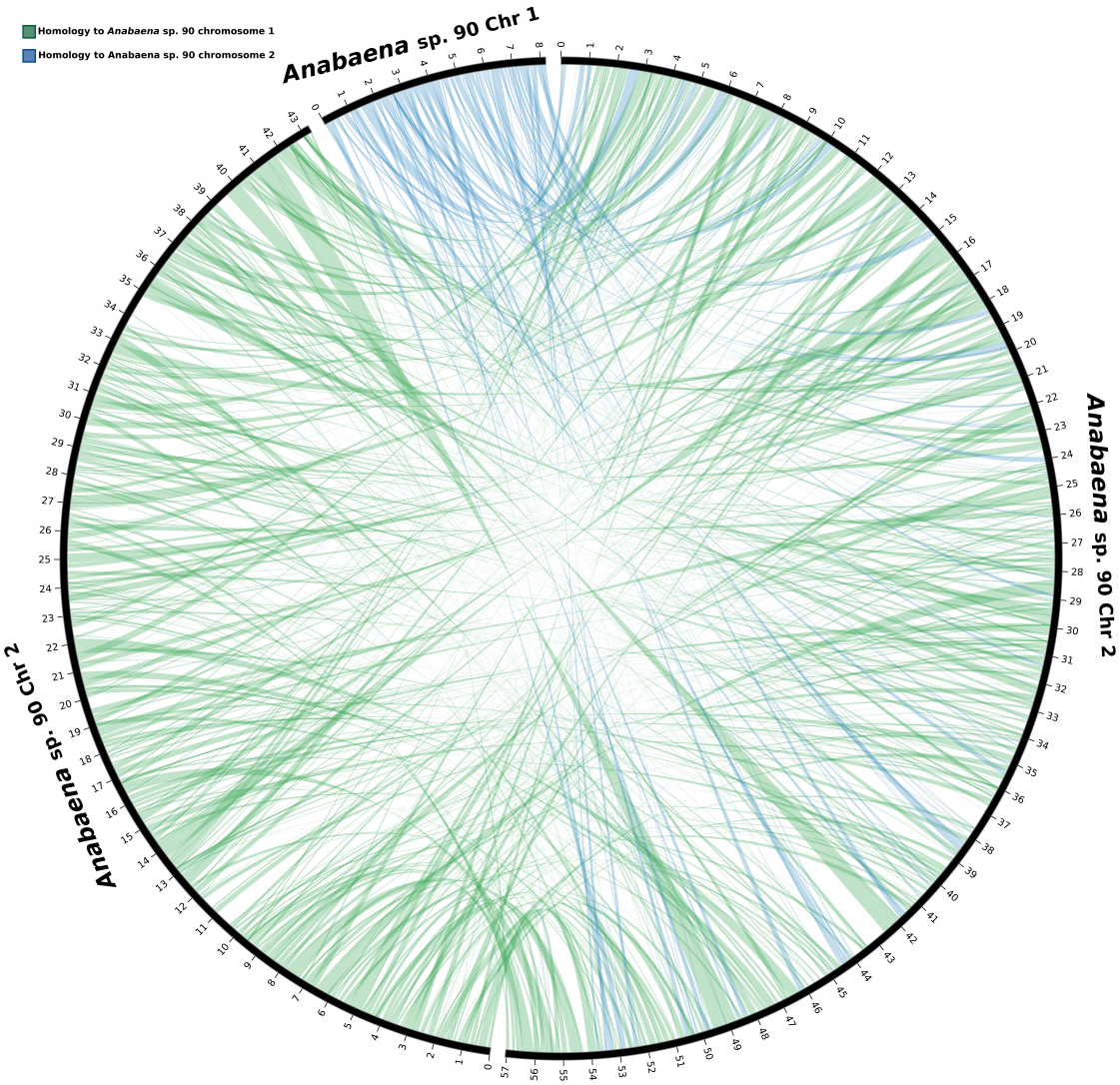

Figure 3.9: Nucleotide alignment between *Anabaena* sp. 90 and WA102. Although *Anabaena* sp. 90 and WA102 share 91.5% average nucleotide identity, they nearly entirely lack synteny. Additionally, the *Anabaena* sp. 90 genome is divided between two chromosomes, unlike the single chromosome of the *Anabaena* sp. WA102 genome.

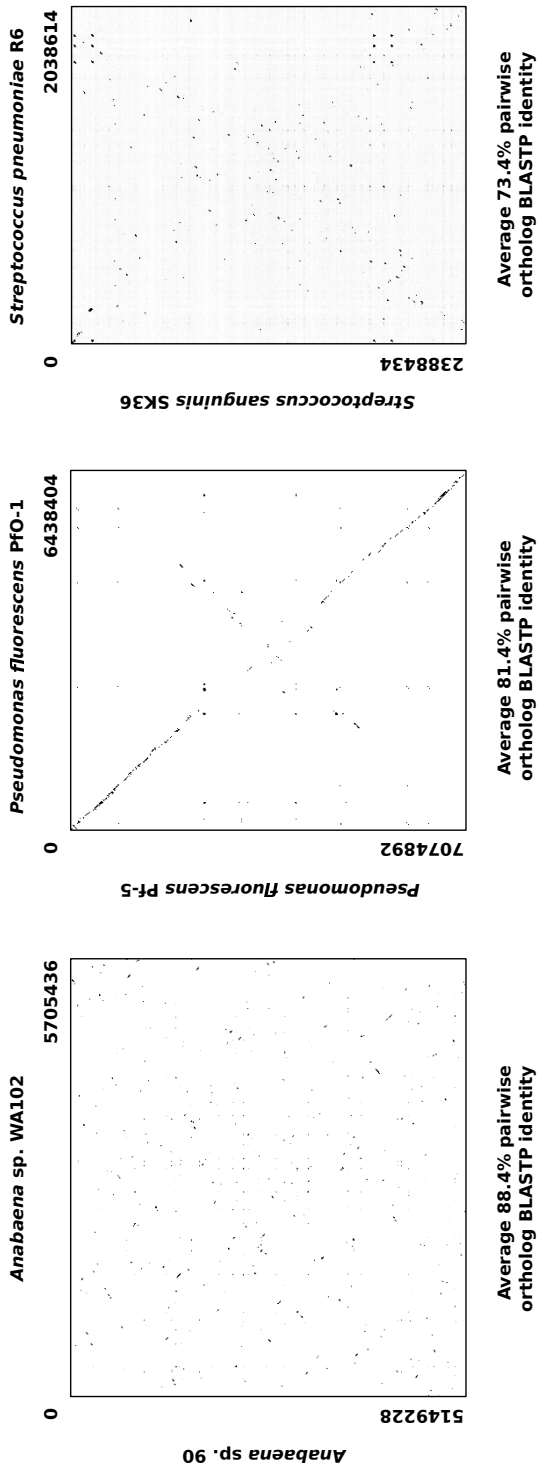

Figure 3.10: Dotplots and average ortholog similarity for pairwise comparisons within three bacterial genera. Dotplots illustrate preservation or absence of long-range nucleotide similarity (synteny) between paired genomes from *Anabaena* in this study and *Pseudomonas* and *Streptococcus* (originally compared in Novichkov et al., 2009).

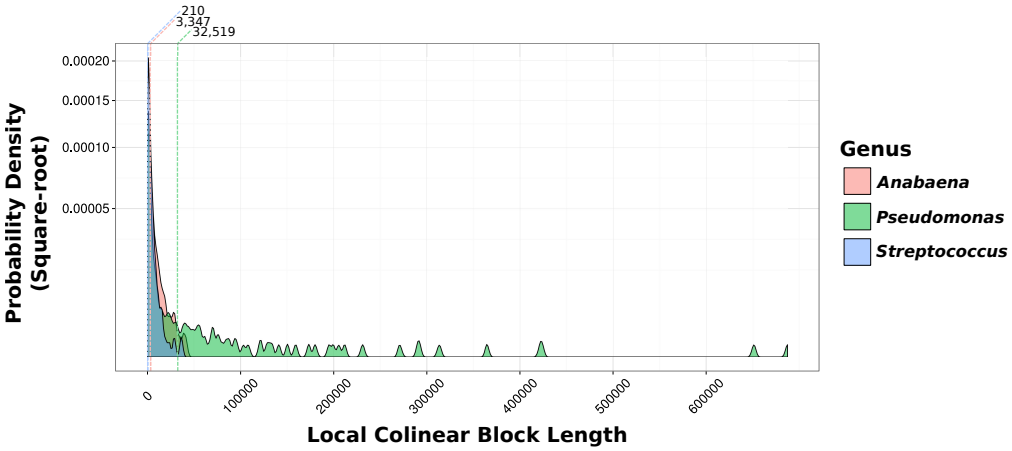

Figure 3.11: Probability density of the local colinear block (LCB) lengths for three bacterial genera. The same pairwise genomes comparisons from the dotplots in Figure 9 are aligned in Mauve. Mauve generates LCBs, which are syntenous regions defined by conserved termini, and that may contain large insertions. The lengths of these LCBs are plotted in a probability density plot for each pairwise genome comparison. The mean LCB length for each pairwise genome comparison is shown as a dotted line with the value printed above the graph. *Pseudomonas* genomes have a mean LCB length of 32.5 kbp, *Anabaena* 3.3 kbp, and *Streptococcus* 210 bp, quantifying what can be observed in the dotplots.

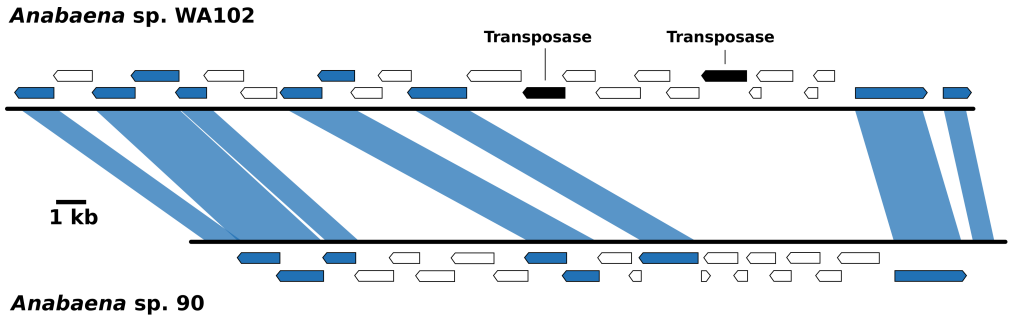

Figure 3.12: Comparing synteny within a local colinear block between *Anabaena* sp. WA102 and 90 (nucleotides 1,179,150-1,203,874 and 2,682,853-2,688,083, respectively). Within this local colinear block, there is evidence of interruption by transposases. Most of the six instances of broken synteny in this LCB are not clearly attributable to a particular mechanism.

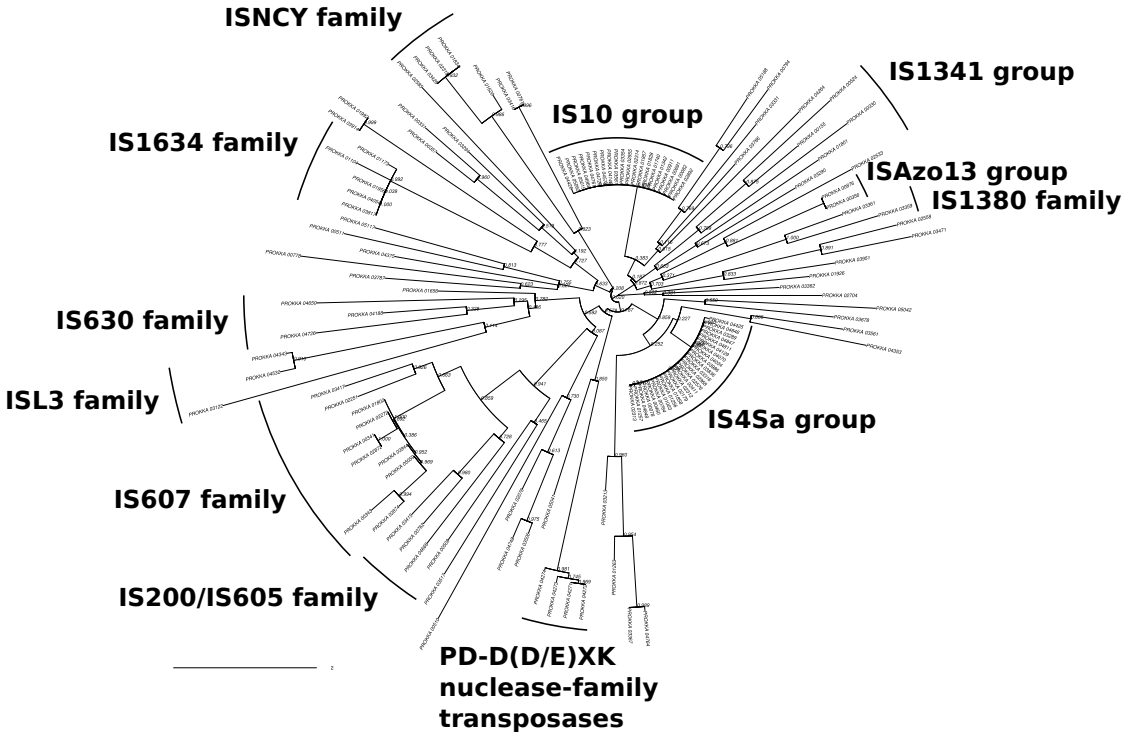

Figure 3.13: Phylogenetic tree of transposase protein sequences encoded in the *Anabaena* sp. WA102 genome. The phylogenetic relationship between 130 annotated transposase protein sequences is sketched out in the tree. Two large clades of closely related transposases dominate the tree. The IS4Sa clade includes 25 transposases and the IS10 clade includes 20 transposases, which both belong to the larger IS4 transposase family. These transposases have a DDE-type active site that facilitates cut-and-paste transposition. The IS4Sa clade has an identical terminal direct repeat sequence: CCGCCTTGTCACCCGTTAAG. The IS10 clade has the terminal direct repeat sequence: ATTCAACAYTTCTG.

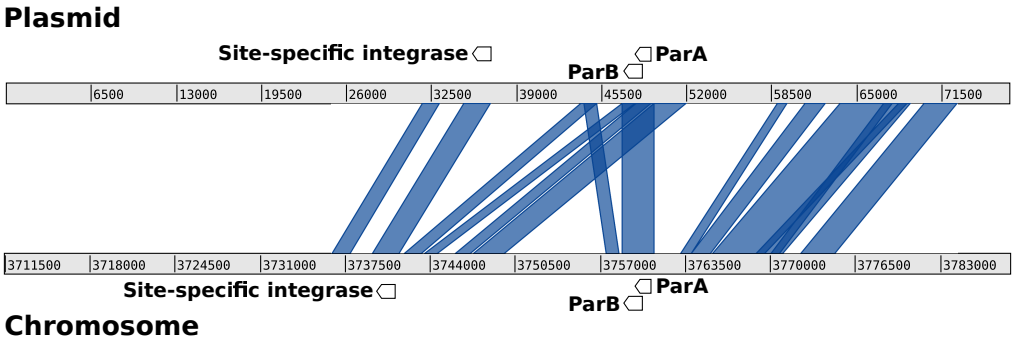

Figure 3.14: Nucleotide alignment between *Anabaena* sp. WA102 chromosome and plasmid. Nucleotide similarity between the chromosome and the plasmid indicates that the plasmid may be integrative and form genomic islands either by integrating into a site on the chromosome or by homologous recombination with the chromosome. The plasmid may be integrative because it encodes site-specific integrases, which can also be found at the homologous site on the chromosome.

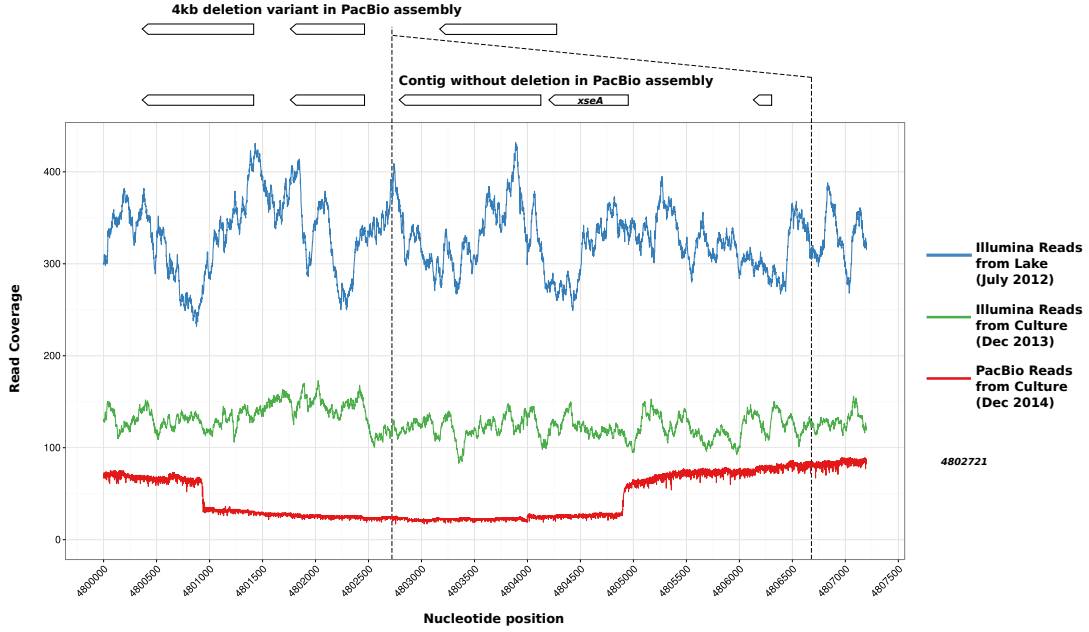

Figure 3.15: Deletion mutation detected in *Anabaena* sp. WA102 culture. A deletion mutation was detected in the PacBio long-read assembly of the *Anabaena* sp. WA102 culture. Mapping reads to the indel region showed that the deletion occurred between nucleotides 4,800,950 and 4,804,900. The deletion arose after December 2013 and expanded through the population to roughly two-thirds of the culture population by December 2014.

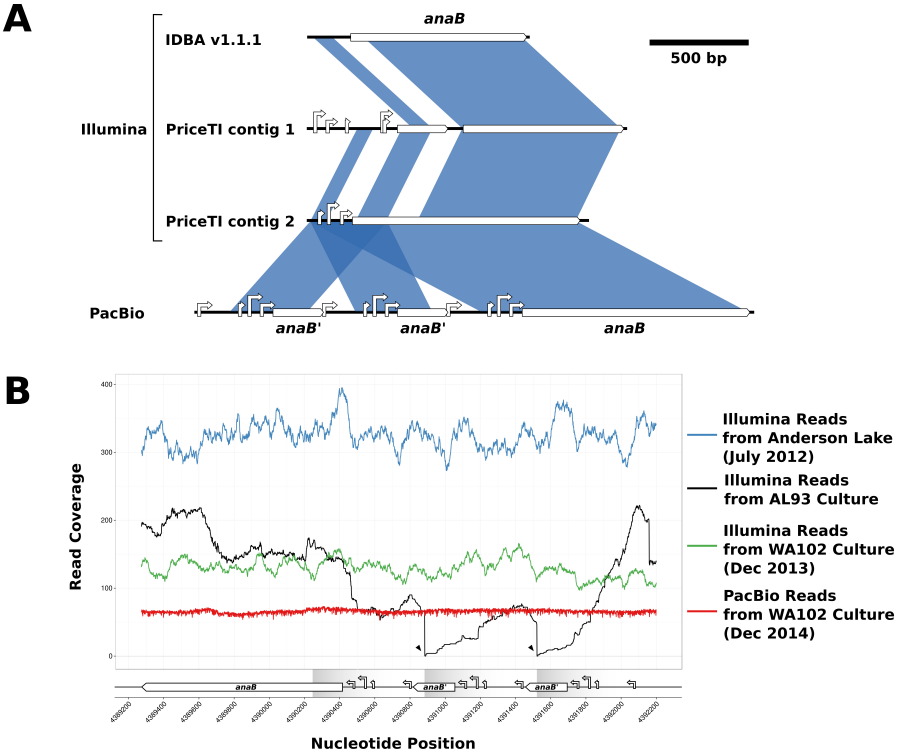

Figure 3.16: Tandem duplication of the putative *anaBCD* promoter region. A) Alignment of the *anaB* gene and upstream promoter region between different assemblies of the *Anabaena* sp. WA102 culture. Promoters were identified with the Virtual Footprint online server, and only promoters with PWM alignment scores greater than 12 were plotted. The 5' end of the *anaB* gene and upstream promoter region are triplicated in the PacBio assembly. None of the Illumina assemblies correctly assemble the tandem triplication. Assembly of 100 bp reads by IDBA v1.1.1 failed to correctly assemble the *anaB* gene and the promoter region. Assembly by PriceTI v1.0.1, using the IDBA contig to seed the assembly, produced two alternate versions of the *anaB* region. In the first version, the *anaB* gene and the upstream promoter region are both improperly assembled. In the second, the *anaB* gene and the most proximal portion of the promoter region are correctly assembled, but triplication is not assembled. B) Read coverage across the promoter region upstream of the *anaB* gene. Illumina metagenome reads from a toxic bloom in Anderson Lake (WA25, blue line), *Anabaena* sp. AL93 culture (green line), and *Anabaena* sp. WA102 culture are mapped across *anaB* and its upstream promoter region. Coverage is summed at each nucleotide and illustrates the absence of two junctions formed between the triplications where the green line drops to zero for the *Anabaena* sp. AL93 culture. In contrast, both the *Anabaena* sp. WA102 culture and the Anderson Lake metagenome contain the junctions formed by the triplication because read coverage does not fall to zero at those loci. Presence of the triplication in the Anderson Lake metagenome indicates that it formed in the *Anabaena* sp. WA102 genome nearly a year prior to establishing the culture. It has been under selection in the environment and continues to be selected for in culture. *Read coverage values for the July 2012 Anderson Lake metagenome have been divided by 10 to facilitate comparison along the ordinate.

Chapter 4 Identification of the major anatoxin-a producing cyanobacterium in Anderson Lake, its dynamics, and its distribution in the Puget Sound region

Nathan M Brown, Ryan S Mueller, Jonathan W Shepardson, Zachary C Landry, Claudia S Maier, Souyun Ahn, F Joan Hardy, and Theo W Dreher

In preparation

4.1 Introduction

Toxic cyanobacterial blooms are increasing in frequency and severity [107, 149]. Rising global temperatures and increasing eutrophication are among the factors enabling freshwater cyanobacteria to expand their range and intensify bloom events. These freshwater cyanobacteria often produce noxious or toxic secondary metabolites that can foul drinking water or prevent recreational use [17]. The genomic characteristics, growth dynamics, and repertoire of secondary metabolites of many of these toxic cyanobacteria are still unknown.

Anatoxin-a is a cyanobacterial neurotoxin that can paralyze animals, causing death by asphyxiation. It is produced by at least six cyanobacteria genera, including *Anabaena*, *Dolichospermum*, *Oscillatoria*, *Aphanizomenon*, *Cylindrospermum*, and *Phormidium* [121]. It was recently implicated in the death of a herd of approximately 100 elk that drank anatoxin-a-contaminated water from a water trough in New Mexico [88]. Public health officials often close freshwater lakes to public access for recreation or drinking water when anatoxin-a levels rise, as in the case of Anderson Lake [33].

Anderson Lake has some of the highest anatoxin-a levels among Washington State lakes (Figure 4.1A). It has had toxic cyanobacterial blooms annually for at least the past 7 years (Figure 4.1B). The blooms usually occur between May and September and produce anatoxin-a levels well in excess of the Washington State guidelines for safe recreational exposure (1μg anatoxin-a per liter). These blooms are often composed of nitrogen-fixing species such as *Anabaena* and *Apha-*

nizomenon, which is likely due to the low total nitrogen:total phosphorus ratio in the lake (Figure 4.1C) [147]. Several of the cyanobacterial species that could potentially produce anatoxin-a have been observed in Anderson Lake [33]. We took multiple approaches to identify the cyanobacteria responsible for producing excessive levels of anatoxin-a during summer blooms in Anderson Lake during 2012 and 2013. We had previously isolated the putative major toxin producer, *Anabaena* sp. WA102, in culture and completed its genome [11]. In this work, we took a broader approach by surveying the lake bloom bacterial community with deep metagenomic sequencing, single-colony isolation and genomic sequencing, and HPLC-MS/MS in the spring and summer of 2013. We also conducted a cyanobacteria-specific phylogenetic marker survey to identify the distribution of the anatoxin-a producer in surrounding freshwater lakes during the summer and fall of 2012. These studies confirmed *Anabaena* sp. WA102 as the major producer of anatoxin-a in Anderson Lake during sampling periods in 2012 and 2013.

4.2 Results

4.2.1 *Anabaena* sp. WA102 is the dominant cyanobacterial species in the 2012 Anderson Lake bloom metagenome sample

A sample of surface water was taken from Anderson Lake on July 7th, 2012 during a toxic cyanobacterial bloom event (Figure 4.1B), with anatoxin-a levels at 187 μg/L on July 9th (`https://www.nwtoxicalgae.org`). The sample was collected as

a retentate on a 1.2 μm-pore-size filter, enriching for large cyanobacterial colonies. Shotgun metagenomic sequencing of total DNA extracted from the sample yielded a metagenome of 100-nt paired-end reads totaling 30.1 Gb. The metagenome was assembled into 230,285 contigs with total size of 255 Mbp and an N50 of 1,530 bp. Nonpareil [128] estimates that the metagenome has covered 92.0% of the bacterial community in the sample. Clustering the contigs by read coverage depth from this metagenome and a metagenome of the *Anabaena* sp. WA102 non-axenic culture yielded separable bacterial population genomes (Figure 4.2). Genome clusters for five cyanobacterial genera were identified in the metagenome sample. The dominant bacterial genome cluster in the July 7th lake metagenome was classified within the *Anabaena* genus using Phylophythia S+. This matched the most abundant cluster of contigs in the *Anabaena* sp. WA102 culture metagenome (see the y-axis of Figure 4.2), suggesting that the genome from the lake is closely related to the *Anabaena* sp. WA102 species.

To quantify the relationship between the lake population genome and the *Anabaena* sp. WA102 culture reference genome, the metagenome was randomly downsampled 10% and mapped to the reference genome (to facilitate mutation identification by Breseq). Of the sequence reads from the lake sample metagenome that mapped to the contigs clustered and labeled as *Anabaena* sp. WA102 in Figure 4.2A, 99.4% also mapped to the *Anabaena* sp. WA102 reference genome [Genbank: CP011456-7]. That is, only 0.6% of reads were recruited from incorrectly clustered contigs in the metagenome (Figure 4.2A). Reads failed to map to a total of 6,199 nucleotides (0.11% of the 5,782,034 bp reference genome). Average coverage depth

was 399. Fifty-seven nucleotide differences that occurred with 100% frequency in mapped reads were detected between the reference genome and the population genome from the lake (Table 4.1). Adding 36 polymorphisms found among the mapped reads (including a 72-nt indel) brings the detected differences between the bloom population genome and the reference genome to 164 nucleotides, giving a 99.997% nucleotide identity between the two genomes. Therefore the bloom and reference genomes can for practical purposes be considered to represent the same strain.

Importantly, the only genes in the July lake metagenome sample that encode for anatoxin-a production occur within the dominant *Anabaena* sp. WA102 population genome (contigs with anatoxin-a genes highlighted in dark red in Figure 4.2A). Mapping all metagenome reads to known anatoxin-a synthetase cluster nucleotide sequences (from *Cylindrospermum stagnale* PCC 7417 [Genbank:NC_019757.1], *Oscillatoria* sp. PCC 6506 [Genbank:FJ477836.1], and *Anabaena* sp. WA102 [Genbank:CP011456]) confirms that the synthetase cluster from *Anabaena* sp. WA102 is the only one detected. These results indicate that the dominant cyanobacterium in the July lake metagenome sample is the previously described *Anabaena* sp. WA102, the only bacterium detected in that sample able to produce anatoxin-a. The dominant *Anabaena* sp. WA102 strain shows minimal nucleotide diversity if polymorphisms among the mapped reads are considered. There are 36 polymorphisms among the reads from the *Anabaena* sp. WA102 bloom strain (only one of these differences spans more than a single nucleotide: a 72-nt indel that maps to position 1,479,670 in the reference genome), indicating that the bloom is likely

clonal.

4.2.2 *Aphanizomenon* sp. WA102, a novel non-toxic Nostocaceae species in Anderson Lake

A second lake surface water sample was taken on May 20th, 2013, prior to the bloom peak, when the anatoxin-a level in Anderson Lake was 12.5 μg/L. The highest level observed in 2013 was 38.7 μg/L on 28 May (`https://www.nwtoxicalgae.org`). Shotgun metagenomic sequencing of total DNA extracted from the sample collected onto a 1.2 μm filter yielded a 4.6 Gb metagenome with a total size of 223 Mbp and an N50 of 1,165 bp. Nonpareil estimates that the metagenome sample has covered 67.0% of the bacterial community in the sample. The metagenomes were assembled and contigs were clustered into population genomes by average read coverage depth as described above. The dominant cyanobacterial population genome (100 average coverage depth, see x-axis in Figure 4.2A) corresponded to a cluster of contigs classified by Phylopythia S+ as *Anabaena* with sequences that were distinct from *Anabaena* sp. WA102 (Figure 4.2B). This cluster of contigs contained 4.5 Mbp total, including 105 unique single-copy essential marker genes and 112 total single-copy essential marker genes according to the mmgenome R package. Closed genomes from the *Nostocaceae* have between 104 and 106 unique essential genes according to mmgenome, indicating that the binned genome is nearly complete with little contamination. Analysis with CheckM [109] reached the same conclusion (99.89% complete, 3.6% contamination). Visual inspection of

the May 20th 2013 lake water sample indicated the predominance of cyanobacterial colonies with a morphology consistent with *Aphanizomenon flos-aquae*. The *Anabaena/Aphanizomenon* clade is known to be intermixed [120] and it is not surprising for a species with *Aphanizomenon* morphology to be identified as *Anabaena* by sequence-dependent means (PhylopythiaS+). We refer to this species as *Aphanizomenon* sp. WA102. Its population genome assembly contains no anatoxin-a biosynthetic genes, nor biosynthetic genes for other known toxins.

A second cyanobacterial population genome, classified as *Anabaena* by Phylopythia S+ and containing anatoxin-a biosynthesis genes, was present in the metagenome at a lower average coverage depth (20 average coverage depth in Figure 4.2B). This population genome matched the sequence of *Anabaena* sp. WA102. With a total of 4.1 Mbp, this cluster of contigs represented 71% of the 5.7 Mbp *Anabaena* sp. WA102 reference genome.

4.2.3 Comparison of *Anabaena/Aphanizomenon* population genomes clustered within the July 2012 and May 2013 metagenomes.

In each metagenome, two *Nostocaceae* (*Anabaena/Aphanizomenon*) genome bins were identified (Figure 4.2). Pairwise comparison of the genome-wide average-nucleotide identities (gANI) between these four population genomes supported the identifications made in Figure 4.2. The gANI between the 2012 and 2013 *Anabaena* sp. WA102 population genomes was 99.85%, and the gANI between the 2012 and

2013 *Aphanizomenon* sp. WA102 population genomes was 99.94%. This indicates that the same two cyanobacteria were present in the 2012 and 2013 samples. These two cyanobacteria are markedly distinct, however, with a gANI of 88.7% that is well below the 96.5% level recommended as the boundary between species [157].

Predicted genes from the *Anabaena* sp. 102 and *Aphanizomenon* sp. WA102 genomes were mapped to the online KEGG ortholog database and compared across KEGG metabolic pathway maps. Of 1,313 non-redundant KEGG orthologs found in the two genomes, 1,213 (92.4%) are shared, while *Anabaena* sp. WA102 encodes 85 unique orthologs, and *Aphanizomenon* sp. WA102 encodes 15 unique orthologs. Since the *Aphanizomenon* sp. WA102 population genome is a nearly complete population genome, some genes may be missing. However, if several genes from a KEGG ortholog pathway are absent from *Aphanizomenon* sp. WA102 and present in *Anabaena* sp. WA102, it is unlikely that those several genes failed to assemble or cluster with the genome, and it is likely that *Aphanizomenon* sp. WA102 actually lacks those orthologs. The organic sulfur uptake and metabolism genes *ssuABCDE* and *tauD*, positive phototaxis genes *pixGHIJL*, cytochrome *cydBA*, red light response genes *cph1/rcp1*, and iron import genes *fhuBC* are absent in *Aphanizomenon* sp. WA102 but present in *Anabaena* sp. WA102. The large number of sulfur metabolism genes absent from *Aphanizomenon* sp. WA102 suggests that this pattern is not due to poor assembly of the genes or improper contig clustering. Rather, it suggests a meaningful difference in the ability to uptake organic sulfur and metabolize it, which may contribute to swapped dominance at different times between *Anabaena* sp. WA102 and *Aphanizomenon* sp. WA102 in Anderson

Lake.

4.2.4 *Anabaena* colony morphology correlates with anatoxin-a production

At times, two morphologies of *Anabaena* are present in Anderson Lake, a smaller-celled *Anabaena flos-aquae*-like morphotype and a larger *Anabaena crassa*-like type (Figure 4.3). Each of these *Anabaena* morphologies is distinct from the *Aphanizomenon flos-aquae* morphology. A third *Nostocaceae* genome that may correspond to the *crassa*-like morphotype was not resolved in the metagenomes discussed above. Morphology, particularly size, can be determined both genetically and environmentally [171, 138], and the environmental determinant can be strong in the case of *Anabaena* [172]. If the cell size and colony morphology of *Anabaena* spp. in Anderson Lake is plastic, then both morphologies of *Anabaena* may be associated with anatoxin-a production. To find if the morphology of the *Anabaena* colonies can be used to guide prediction of anatoxin-a production in Anderson Lake, we isolated 25 *Anabaena flos-aquae*-like colonies and 10 *Anabaena crassa*-like colonies directly from lake surface water samples on June 18th and 25th, 2013. The anatoxin-a concentration in Anderson Lake on June 17th, 2013 was 35.8 μg/L, indicating that the anatoxin-a producer was present in the lake. Anatoxin-a could be measured in 19/25 *flos-aquae*-like colonies, but in only 2/10 *crassa*-like colonies (Figure 4.3C). This suggests that *Anabaena flos-aquae*-like colonies are more likely to be associated with anatoxin-a production. Consistent with this, *Anabaena* sp.

WA102, an anatoxin-a producer from Anderson Lake that is in culture, has a morphology similar to *Anabaena flos-aquae*.

To determine if the genomes from the *flos-aquae*-like and *crassa*-like *Anabaena* colonies are related to the genomes identified from the metagenomes, DNA was extracted and amplified from a single colony each from the larger- and smaller-celled *Anabaena* morphotypes with multiple-displacement amplification (MDA) and sequenced with 250-nt paired-end reads on the Illumina MiSeq platform. Sequencing reads from the MDA were mapped to the *Anabaena* sp. WA102 and *Aphanizomenon* sp. WA102 genomes. Reads from MDA of the *flos-aquae*-like colony covered 93.1% of the *Anabaena* sp. WA102 genome, including the anatoxin-a gene cluster, and 47.65% of the *Aphanizomenon* sp. WA102 genome. Reads from the MDA of the *crassa*-like colony mapped to 16.8% of the *Anabaena* sp. WA102 genome, excluding the anatoxin-a gene cluster, and 4.8% of the *Aphanizomenon* sp. WA102 genome. These results suggest that the *flos-aquae*-like colony morphology, which resembles the morphology of *Anabaena* sp. WA102 in culture, is indicative of the presence of an anatoxin-a producing species, whereas the *crassa*-like colony morphology is not. Although this relationship is not causal, it can guide lake management decisions. Additionally, the *crassa*-like colony morphology does not belong to either *Anabaena* sp. WA102 nor *Aphanizomenon* sp. WA102, nor could the anatoxin-a synthetase region be detected in its amplified DNA.

4.2.5 Distribution of *Anabaena* sp. WA102 across the Puget Sound region and *Nostocaceae* diversity

To assess the distribution of *Anabaena* sp. WA102 across the region, surface water from eleven freshwater lakes surrounding the Puget Sound, including Anderson Lake, were sampled ten times each (every two weeks) from May to October 2012. Three additional samples taken in 2012 from Anderson Lake and a sample taken on February 21st, 2013 from Clear Lake were also included. Samples were filtered onto 1.2 μm-pore-size filters to enrich for cyanobacterial colonies. Total DNA was extracted from the filters and the *cpcBA*-IGS phylogenetic marker region was amplified from each sample with the polymerase chain reaction (PCR). We included a positive control composed of DNA from the *Anabaena* sp. WA102 culture and a negative control composed of DNA from a culture of *Synechococcus* sp. PCC 7942. The forward primer from [8] was used in combination with a newly designed reverse primer to create an amplicon approximately 430 bp in size that can be fully sequenced on the Illumina MiSeq platform. The marker includes parts of the phycobilin phycocyanin genes *cpcA*, *cpcB*, and the intergenic region between them. This region is unique to cyanobacteria and contains highly variable sequence, enabling cyanobacterial species alone to be identified in lake microbial communities and to be distinguished from each other. The primers were designed to anneal to conserved DNA motifs in members of the family *Nostocaceae*, but not other members of the *Cyanobacteria* (such as *Synechococcus* sp. PCC 7942).

A total of 823,202 amplicons were sequenced from 96 samples (including con-

trols). After removing low-quality sequences, 3,506 operational taxonomic units (OTUs) were clustered at 3% nucleotide similarity according to [104]. Removing OTUs with fewer than 20 sequences in order to simplify analysis, 597,665 amplicons from 19 OTUs remained for analysis. Despite attempts to equalize the mass of DNA contributed from each sample prior to pooling and sequencing, 80% of processed amplicons could be attributed to only two samples, the positive control (328,879 amplicons) and the June 18th, 2012 sample from Lake Ketchum (142,623 amplicons). This is likely due to the near absence of *Nostocaceae*, or any cyanobacterial bloom, in most samples. The negative control contained 52 amplicons that matched amplicon sequences from the positive control. The negative control and positive control were processed in adjacent wells, and aerosol from the positive control (328,879 amplicons) likely contaminated the negative control prior to PCR amplification. This suggests that low counts may be erroneous. To protect against this source of error, 62 samples with fewer than 500 amplicons (ten times the number of amplicons observed in the negative control, which would avoid contamination of the same magnitude seen in the negative control) were removed before analysis. 584,651 amplicons (71% of all sequenced amplicons) from the remaining 23 samples, including the positive control, were analyzed. Representative nucleotide sequences from the 19 remaining OTUs were realigned against each other and placed in a phylogenetic tree, with poorly supported nodes (bootstrap support $< 90\%$) omitted (Figure 4.4). The tree shows five well supported clades that differ between each other in the alignment largely due to indels. The tree shows that OTUs 1, 5, 6, 9, 12, and 19 form a well supported clade of closely re-

lated sequences. These OTUs were detected in significant numbers in the positive control, which is known to be a single isolate. Inspection of the alignment indicated that each of these OTUs shared the same indels and differed by one nucleotide in the approximately 430 nt amplicon. There are two other similar clades formed by OTUs 2, 7, 8, 10, 16, and 17 and OTUs 3, 11, 13, 14, and 18, as well as two well resolved OTUs (4 and 15), resulting in five distinct OTU clades (Figure 4.4).

The *Nostocaceae* community from each sample is displayed in a heat-map (Figure 4.5), in which samples are arranged by similarity according to multi-dimensional scaling of the weighted unifrac metric for each community. The positive control is most similar to samples from Anderson Lake, with large counts of OTU clade 1. This is expected, considering that the positive control contains DNA from *Anabaena* sp. WA102, which was in turn isolated from Anderson Lake. It also indicates that OTU clade 1 represents the toxic *Anabaena* sp. WA102 species, confirmed by the fact that the *cpcBA*-IGS sequence from the *Anabaena* sp. WA102 genome and the OTU 1 representative sequence are identical. The October 2012 Anderson Lake sample is not clustered with the May and July 2012 samples, showing changes in the *Nostocaceae* community in Anderson Lake over the bloom season. The February 2013 sample from Clear Lake and October 2012 sample from Echo Lake cluster with the May 2012 sample from Anderson Lake (during a bloom with anatoxin-a levels exceeding state guidelines in Anderson Lake) because of high counts of OTU group 3. However, counts of OTU clade 1 in the February 2013 Clear Lake and October 2012 Echo Lake samples are similar to counts found in the negative control (124 and 37, respectively, whereas the negative

control had 49) and may not indicate the presence of OTU 1. OTU clades 2 and 3 are common among many lakes, and occur in combination with OTU groups 1, 4, and 5. The geographic distribution of the OTU clades can be seen in the map of the Puget Sound Region (Figure 4.6), showing the prevalence of OTU clades 2 and 3 and the near absence of OTU clade 1 except in Anderson Lake. *Anabaena* sp. WA102 (OTU 1) was thus not widespread throughout lakes in the Puget Sound region in 2012, though it may be present at low levels in lakes throughout the region.

Intriguingly, the *cpcBA*-IGS from *Aphanizomenon* sp. WA102 could not be detected in the July 7th, 2012 sample from Anderson Lake, despite its abundant presence in the same sample according to the shotgun metagenome data from the same sample (Figure 4.2). There are three mismatches on the reverse *cpcBA*-IGS primer compared to the *Aphanizomenon* sp. WA102 *cpcBA*-IGS region. In contrast, there are two mismatches on the forward primer and a single mismatch on the reverse primer compared to the *Anabaena* sp. WA102 *cpcBA*-IGS, which was successfully detected by PCR (Figure 4.7). Presumably the additional mismatch on the forward primer destabilized annealing to *Aphanizomenon* sp. WA102 *cpcBA*-IGS DNA sufficiently to prevent amplification.

4.3 Discussion

4.3.1 *Anabaena* sp. WA102 was the major anatoxin-a producer in Anderson Lake in metagenome samples from July 2012 and May 2013

Deep shotgun metagenome sequencing of surface water collected from the Anderson Lake toxic bloom on July 7th, 2012 revealed *Anabaena* sp. WA102 to be the dominant cyanobacterial component. Its genome coverage depth was approximately 40-fold greater than that of the next-most populous cyanobacterium, *Aphanizomenon* sp. WA102. With only 57 nucleotide differences occurring in all reads mapped between the *Anabaena* sp. WA102 population genome clustered from the metagenome and the closed *Anabaena* sp. WA102 reference genome, the identity of the dominant cyanobacterium in this lake sample is clear. The only anatoxin-a genes identified within the metagenome assembly were found to cluster in the *Anabaena* sp. WA102 population genome, identifying *Anabaena* sp. WA102 as the sole anatoxin-a producer detected in the metagenome. The metagenome is estimated to cover 92% of the genomes in the environmental sample, leaving the possibility that a less abundant anatoxin-a-producing cyanobacterium remains undetected by the metagenome. However, it is clear that the most abundant bacterial component, *Anabaena* sp. WA102, is also the major producer of anatoxin-a in the July 2012 Anderson Lake bloom. There may be other sources of anatoxin-a in the lake, such as cyanobacterial anatoxin-a producers in the benthos, that were

not captured in the surface-water bloom samples. Further metagenomic analysis of Anderson Lake should include samples from several depths of the water column, as well as the benthos.

Isolating single colonies of *Anabaena* from Anderson Lake confirmed that *Anabaena* sp. WA102 was the dominant anatoxin-a producer in Anderson Lake. Of the two morphologies of *Anabaena* colonies that coexisted in Anderson Lake on June 18th and June 25th, 2013, only the *Anabaena-flos-aquae*-like colonies were strongly associated with anatoxin-a production (Figure 4.3C). This is the same morphology as the *Anabaena* sp. WA102 culture (Figure 4.3). Further, the DNA sequence from one of these colonies mapped to 93.1% of the *Anabaena* sp. WA102 genome, including the anatoxin-a sythetase gene cluster, confirming its identity. The correlation between *Anabaena* sp. WA102 morphology and anatoxin-a production suggests that it may be reasonable to predict whether or not a bloom is toxic by visual inspection, though more samples should be considered.

PCR amplifying the *cpcBA*-IGS region from DNA isolated from three Anderson Lake samples throughout the 2012 bloom season verified the presence of *Anabaena* sp. WA102 when anatoxin-a levels are high in the lake (Figures 4.5, 4.6). Samples from May 2013 and July 2013, when anatoxin-a levels were above Washington state guidelines, showed large counts of the OTU 1 amplicon, which represents the *cpcBA*-IGS from *Anabaena* sp. WA102. These combined results from 2012 and 2013 suggest that *Anabaena* sp. WA102 may be a perennial cause of high anatoxin-a levels in Anderson Lake.

4.3.2 *Anabaena* sp. WA102 is not always the dominant nitrogen-fixing autotroph in Anderson Lake

Although *Anabaena* sp. WA102 was dominant in the July 7th, 2012 bloom sample (Figure 4.2A), it is not always the dominant nitrogen-fixing *Nostocaceae* in Anderson Lake. The May 20th, 2013 metagenome sample showed that *Anabaena* sp. WA102 was less abundant than *Aphanizomenon* sp. WA102 (Figure 4.2B). In addition, the *Nostocaceae* community from the May and October 2012 *cpcBA*-IGS samples showed higher abundances of OTU 3 or OTU 2, respectively, than OTU 1 (*Anabaena* sp. WA102) (Figure 4.5). The identities of the *Nostocaceaea* represented by OTUs 2 and 3 are presently unknown, but they are distinct from *Aphanizomenon* sp. WA102, which was not detected in the amplicon study because of primer mismatches (Figure 4.7) that were unknown at the outset of these experiments. Independence from PCR primers is a major advantage of shotgun metagenomics, especially when studying organisms with uncharacterized genomes.

Comparing the *Anabaena* sp. WA102 and *Aphanizomenon* sp. WA102 genomes shows that they share all but 100 of 1,313 KEGG orthologs that mapped to the two genomes. As far as KEGG orthologs represent core metabolic functions in bacteria, this suggests that *Anabaena* sp. WA102 and *Aphanizomenon* sp. WA102 share a sizable number (92.4%) of core metabolic functions. The few differences in core metabolism between these two *Nostocaceae*, which share a photoautotrophic, nitrogen-fixing niche in the same lake, may indicate the reasons that they exchange dominance in Anderson Lake throughout the seasons. Two biochemical pathways

with the greatest differences between the two species are light-response and sulfur metabolism pathways.

Conspecific cyanobacteria such as different ecotypes of *Prochlorococcus* spp. specialize in collecting different light spectra, thus occupying non-competing niches [156]. However, both of the Anderson Lake *Nostocaceae* encode phycocyanin and allophycocyanin, but none of the other known phycobilins, so that they are in competition for the same light spectrum. The ability of *Aphanizomenon* sp. WA102 to respond to changes in light conditions seems limited compared to *Anabaena* sp. WA102, since it lacks the *pixGHIJL* positive phototaxis genes [170] and *cph1/rcp1* genes for the phytochrome red-light-response two-component system [168]. The ability to move towards light conferred by the *pix* operon or to modify the photosystems in response to a shift in the red-light spectrum conferred by phytochrome suggests that *Anabaena* sp. WA102 would outcompete *Aphanizomenon* sp. WA102 for light in a dynamic light environment. This may become important during the well-lit summer months, when the cyanobacterial bloom forms patches of shade in contrast to well lit open areas of the lake. *Anabaena* sp. WA102 would be able to modify its photosystems in response to being shaded by a bloom, as well as be able to migrate towards open sunlit areas of the lake.

In addition, *Aphanizomenon* sp. WA102 lacks the *ssuABCDE* operon and *tauD* gene involved in sulfur metabolism [37]. These genes would give *Anabaena* sp. WA102 a competitive edge if sulfate levels drop in the water column, perhaps during a bloom, allowing it to assimilate sulfonates. The role of sulfur in freshwater cyanobacterial ecology is not well understood, largely because sulfur is not well

measured in freshwater cyanobacterial ecology studies. However, some freshwater cyanobacteria have evolved dramatic responses to sulfur deprivation, indicating that sulfur depletion may limit growth on some occasions, and may be relevant to freshwater cyanobacterial ecology. For example, *Fremyella diplosiphon*, a filamentous freshwater cyanobacterium, uses phycocyanin, which can compose half of a cyanobacterium's protein biomass, to store sulfur [54]. During sulfur deprivation, it will stop production of sulfur-rich phycocyanin, consume the protein, and initiate production of a sulfur-poor phycocyanin homolog. As shotgun metagenomics becomes more common for characterizing freshwater lake systems at the genetic and genomic levels, complementary measurements of elements such as sulfur and iron may illuminate new aspects of cyanobacterial metabolism and competition in freshwater lakes.

4.3.3 *Anabaena* sp. WA102 is sparsely distributed throughout the Puget Sound Region

Tracking the *Nostocaceae* communities across the Puget Sound Region with the *cpcBA*-IGS phylogenetic marker showed that among the 11 sampled lakes, *Anabaena* sp. WA102 is only abundant in Anderson Lake. Low counts of *cpcBA*-IGS OTU 1, which represents the *cpcBA*-IGS sequence from *Anabaena* sp. WA102, were detected in all but Spanaway, Harts, and Cassidy Lakes, but such low counts are below our threshold of confidence since similar counts of OTU 1 were found in the negative control.

Invasive members of the *Nostocales* are spreading across the world, likely in response to increasing global temperatures. Their shared ability to fix nitrogen, grow efficiently under different light conditions, persist in cold temperatures between seasons by forming akinetes, and produce toxic and allelopathic secondary metabolites such as anatoxin-a enables them to invade freshwater lakes. The scarcity of *Anabaena* sp. WA102 throughout the Puget Sound Region may give public health officials the chance to prevent its spread to the region by treating Anderson Lake with herbicides. It will be important in the future to monitor lakes in the region for the spread of this potent anatoxin-a-producing species.

4.4 Methods

4.4.1 Sample collection

500 mL samples were collected from Anderson Lake, Jefferson County, Washington State (48.0190°N, 237.1963°W) by the Jefferson County Public Health Department during the 2012 and 2013 cyanobacterial toxic bloom seasons. Samples were collected at a depth of 0-0.5 m and may have included a dense windblown scum. Samples were shipped overnight on ice and 10-25 mL (depending on the sample density) were filtered through 1.2 μm-pore-size Whatman GF/C 24 mm-diameter filters. Filters were stored at -80°C for later DNA analysis.

4.4.2 Single-colony isolation, toxin extraction, and DNA sequencing

Freshwater samples were collected on June 17th and 24th, 2013 from Anderson Lake. Samples were stored at 4°C for up to a week for colony isolation (no chlorosis was observed). *Anabaena* colonies were individually isolated by serial transfer between MilliQ water droplets on a glass slide. Colonies were considered to be isolated when no other cell debris was visible in the surrounding water droplet under 200x magnification on a Zeiss brightfield microscope. Isolated colonies were suspended in 50μL sterile MilliQ water in a micro-centrifuge tube before being frozen by immersion in liquid nitrogen and thawed in a water bath at 25°C to gently release cytoplasmic contents (freeze/thaw cyle was repeated three times). Samples then were centrifuged at 23,000 RCF for 5 min to pellet crude cell mass and supernatants were removed for HPLC-MS/MS toxin analysis. DNA from the pellets of these samples was extracted with three -80°C/20°C freeze-thaw cycles followed by a cold alkaline lysis [119]. DNA was then amplified from the colony with multiple displacement amplification (Qiagen REPLI-g kit, cat no. 150023) and sequenced with 250-nt paired-end reads on an Illumina MiSeq instrument.

4.4.3 DNA extraction from lake samples for shotgun metagenomic and amplicon sequencing

Total DNA from lake samples was extracted from 1.2 μm-pore-size filters by macerating the filters in 500 μL TNE (50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 0.1 mM

EDTA) with a pestle. Cell material was pelleted with low-speed centrifugation, resuspended in TNE buffer, and DNA was extracted by a method from Neilan et al. [126] that had the following modifications. The protein fraction was removed with two 25:24:1 phenol/chloroform/isoamyl alcohol extractions followed by two chloroform extractions. Residual phenol was removed with a final diethyl-ether extraction. DNA was stored at -20°C.

4.4.4 *cpcBA*-IGS amplicon primer design, amplification, and sequencing

Nucleotide sequences for complete *cpcB* and *cpcA* genes and the *cpcBA* intergenic sequence from the *Nostocaceae* family were downloaded from NCBI (November 2015). The nucleotide sequences from each gene (*cpcB* and *cpcA* separately) were translated and codon-aligned using the pal2nal.pl script with default settings [151]. These codon alignments were used to locate conserved primer binding sites in the *cpcB* and *cpcA* coding sequence that would allow polymerase chain reaction (PCR) amplification across the most variable regions of *cpcB*, *cpcA*, and the *cpcBA* intergenic sequence. Primers were also chosen to produce a PCR product less than 450 bp, so that an overlap between paired-end 250-nt Illumina MiSeq reads is created when sequenced. The chosen primers were:

forward - NB78 5' GGCTGCTTGTTTACGCGACA,

reverse - NB81 5' GTCCTTGGGTATCAGCAGATGC.

The forward primer was reported by [8]. Each primer has a T_m of 57°C and to-

gether they yield a PCR product of approximately 430 bp (the size of the *cpcBA* intergenic spacer varies). The following PCR program was used with the Kapa HotStart ReadyMix HiFi Kit (cat no. KK2601, Kapa Biosystems, Wilmington, MA) to amplify from extracted environmental DNA: initial denaturation at 95°C for 3:00 followed by denaturation at 98°C for 0:20, primer annealing at 60°C for 0:15, and extension at 72°C for 0:30 for 30 cycles, with storage at 4°C. The sequencing library was prepared using the Nextera XT Index Kit. Prior to pooling the indexed amplicons for sequencing, their molarities were equalized using the SequalPrep Normalization Plate Kit (cat no. A10510-01, ThermoFisher Scientific, Waltham, MA). The library was sequenced on the Illumina MiSeq platform with 300-nt paired-end reads.

4.4.5 *cpcBA*-IGS amplicon analysis

Fastq files from each sample were processed in mothur v1.36.1 [132] using a batch file that can be found at `https://github.com/russianconcussion/data.analysis.scripts/blob/master/cpcba.batch`. This workflow was based on the Mothur MiSeq SOP [73]. Amplicons from each sample were deduplicated, aligned, and clustered at a 3% nucleotide similarity threshold [104] in mothur to form 3,506 operational taxonomic units (OTUs). The "*.shared" file from mothur was then imported into R with the phyloseq package, incorporated into a complex phyloseq object with metadata for each sample and a phylogenetic tree of OTUs. The data was visualized for exploratory data analysis in phyloseq

[89] using the R script found at `https://github.com/russianconcussion/data.analysis.scripts/blob/master/cpcba.R`. To simplify the dataset for interpretation while keeping the majority of the data, OTUs with fewer than 20 sequences were removed from further analysis, leaving 19 OTUs for analysis. Representative sequences from these OTUs were realigned against each other with Muscle v3.8.31 [36] run on default settings to obtain a more efficient alignment and placed in a phylogenetic tree using the general time-reversible nucleotide substitution model in FastTree 2.1.8 [117]. Internal nodes with less than 90% bootstrap support were removed using Newick Utilities v. 1.6 [62]. OTUs were plotted on a map of the Puget Sound region using the maptools package [6] in R.

4.4.6 Metagenome analysis

DNA extracted from the July 7th, 2012 Anderson Lake sample was shotgun sequenced using 100-nt paired-end reads on the Illumina HiSeq 2000 platform. Reads from the 30.1 Gbp metagenome were uploaded to the NCBI Short-Read Archive [SRA:SRS1169983]. The May 20th, 2013 Anderson Lake sample was shotgun sequenced using 300-nt paired-end reads on the Illumina MiSeq platform, and reads from the 4.6 Gbp metagenome were uploaded to the NCBI Short-Read Archive [SRA:SRR2939624]. Metagenomes were assembled using idba version 1.1.1 assembler software [112] on a 64-bit Linux server with 500GB of RAM. Prior to assembly, any reads containing ambiguous basecalls ("N") were culled. The large chromosome from the *Anabaena* sp. 90 genome [Genbank:NC019427]

was used as a reference to guide assembly. Within idba, assemblies with kmer sizes ranging from 20 nt to the sequence read length (100 nt to 250 nt) in 10nt increments were combined in the final assembly. Reads from original fastq files were mapped to the assemblies using bwa version 0.7.5a-r405 [80]. Average coverage depth for each contig was calculated using samtools version 0.1.18 (r982:295) [81] and the calc.coverage.in.bam.depth.pl script from the mmgenome package (`https://github.com/MadsAlbertsen/mmgenome`) [1]. The mmgenome network.pl script generated a network of contigs based upon paired-end read data extracted from the bwa-generated SAM file. Bacterial and archaeal metagenome contigs were taxonomically classified using the PhylopythiaS+ support vector machine (SVM) classification software with only a contig fasta file and not a scaffold fasta file (`https://github.com/algbioi/ppsp`) [51]. 16S marker genes were detected in the contig file and used by PhylopythiaS+ to select an SVM training dataset automatically. Putative protein coding sequences were identified in each assembly fasta file using Prodigal version 2.6.2 [57]. To identify single-copy essential marker genes, putative protein sequences were aligned against a curated hmm database from the mmgenome package with the HMMER version 3.0 package (`http://hmmer.janelia.org/`) [35]. A custom data generation shell script based on the data.generation.2.1.0.sh script from mmgenome was used to combine the above processes (`https://github.com/russianconcussion/data.analysis.scripts/blob/master/mmgenome.datagen.sh`). Average coverage depth, network, taxonomic classification, and essential gene data for each assembly were imported into a data.frame structure in R. Finally, the mmgenome

R package was used to generate a plot of genome clusters within the metagenomes, define and evaluate completeness of the clusters, and export well defined genome clusters as contigs in fasta format. Genome clusters in fasta format were annotated using Prokka version 1.11 [133]. Contigs from the binned genome of *Anabaena* sp. WA102 and the bam file of reads mapped to the contigs were used to extract raw reads from the original fastq files that map to the contigs. Using breseq v0.27.1, the extracted reads were mapped to the *Anabaena* sp. WA102 reference genome [Genbank:CP011456-7] and mutations were called [28]. Mutations were displayed using Circos v0.67-7 [74]. Metagenome completeness was estimated using Nonpareil v2.4 on default settings [128].

4.4.7 Genome comparisons

Population genomes were compared using the webserver for the average nucleotide identity (ANI) tool from the Konstantinidis Lab (`http://enve-omics.ce.gatech.edu/ani/`) [50]. Gene content was mapped to the KEGG ortholog database with the KEGG Mapper webserver (`http://www.genome.jp/kegg/tool/map_pathway1.html`) [65]. Comparisons between KEGG orthologs found in each genome were made using the iPATH webserver (`http://pathways.embl.de/`) [79] and the pathview package in R [85].

4.5 Tables and Figures

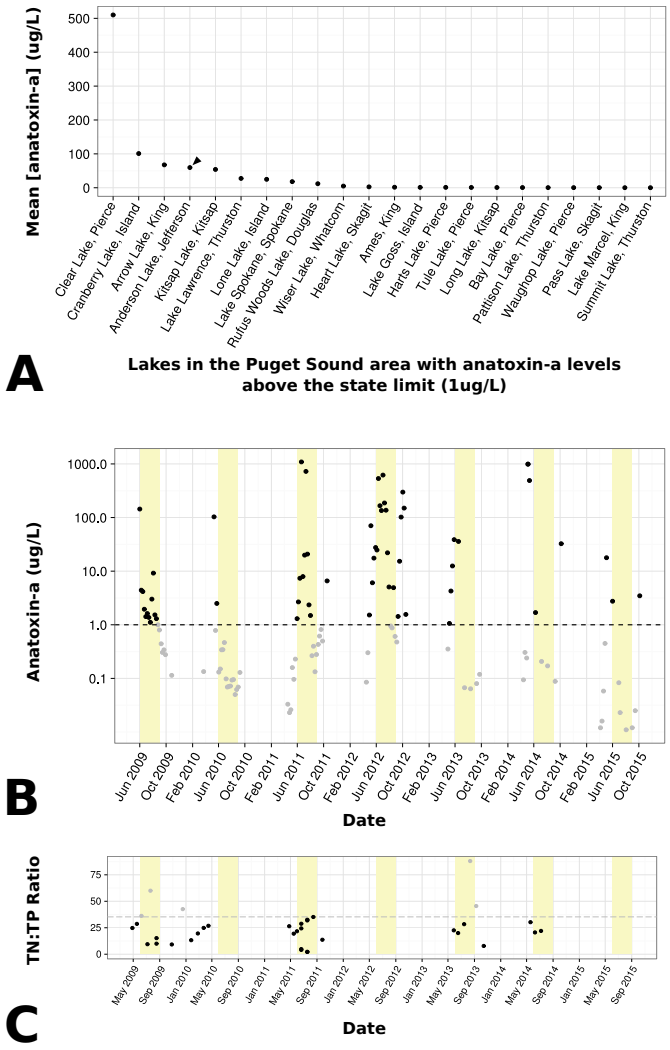

Figure 4.1: Anatoxin-a occurrence in Anderson Lake. A) Lakes in Washington State, USA, with anatoxin-a levels measuring above the 1μg/L state guideline level for recreational exposure. The mean measured anatoxin-a level is shown for each lake, though it must be noted that these means are based on an unevenly sampled data and include extreme outlier values. Data is from https://www.nwtoxicalgae.org/. B) Anatoxin-a levels measured over the past 7 years in Anderson Lake, Jefferson County, WA. The summer months June-August are highlighted in yellow and the 1μg/L guideline level is shown as a dashed line. Points in black represent anatoxin-a measurements >1 μg/L and points in gray represent measurements <1 μg/L. **C)** Total (Kjeldahl) nitrogen:total phosphorus ratios (TN:TP) measured over the past 7 years. A TN:TP of 35, at which nitrogen-fixing cyanobacteria are thought to be uncompetitive with non-nitrogen-fixing cyanobacteria is shown as a dashed gray line. Measurements of TN:TP >35 are shown as gray dots, and measurements <35 are shown as black dots.

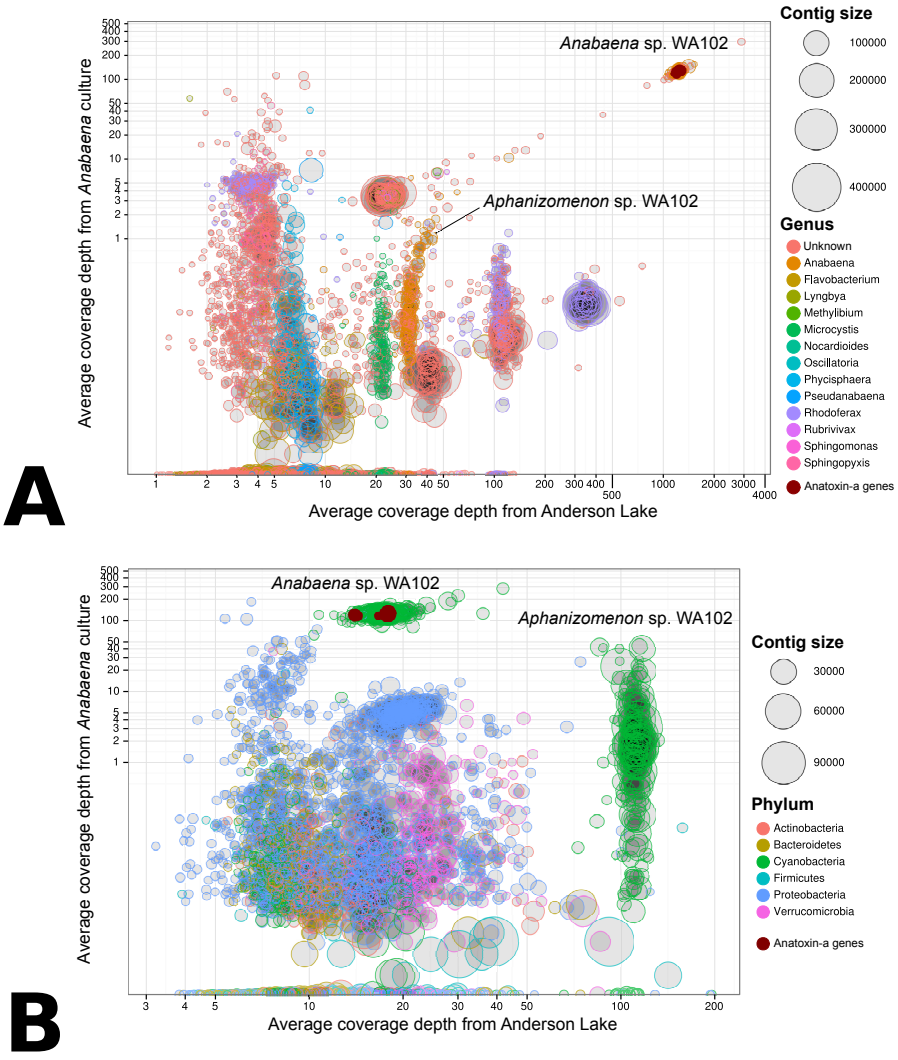

Figure 4.2: Metagenome analysis of Anderson Lake samples. A) Average coverage depth plot of contigs from the July 7th, 2012 Anderson Lake metagenome sample. Sequencing reads (30 Gbp of Illumina HiSeq 100-nt paired-end reads) were assembled into 230,285 contigs with total size of 255 Mbp and an N50 of 1,530 bp. Contigs belonging to population genomes are clustered on the plot according to coverage depth from the sequenced *Anabaena* sp. WA102 culture (y-axis) and coverage depth in the July 2012 Anderson Lake metagenome sample (x-axis). Two species of *Anabaena* were identified by PhylopythiaS+. The *Anabaena* population genome with average coverage depth of 1,200 is nearly identical to *Anabaena* sp. WA102 (Table 1). The *Anabaena* population genome with average coverage depth of 30 in the Anderson Lake metagenome is actually *Aphanizomenon* sp. WA102 (see text). The only anatoxin-a biosynthetic (ana) genes identified cluster with the *Anabaena* sp. WA102 population genome (dark red circles). B) Average coverage depth plot of contigs from the May 20th, 2013 Anderson Lake metagenome sample. Contigs were assembled from 4.6 Gbp of Illumina MiSeq 250-bp paired-end reads. The total assembly is 223 Mbp and has an N50 of 1,165. The *Aphanizomenon* sp. WA102 population genome has an average coverage depth of 100 in the Anderson Lake metagenome and the *Anabaena* sp. WA102 population genome has an average coverage depth of 20. The only anatoxin-a genes detected on contigs cluster with the *Anabaena* sp. WA102 population genome (dark red circles), which was confirmed by mapping reads to known anatoxin-a biosynthetic gene clusters.

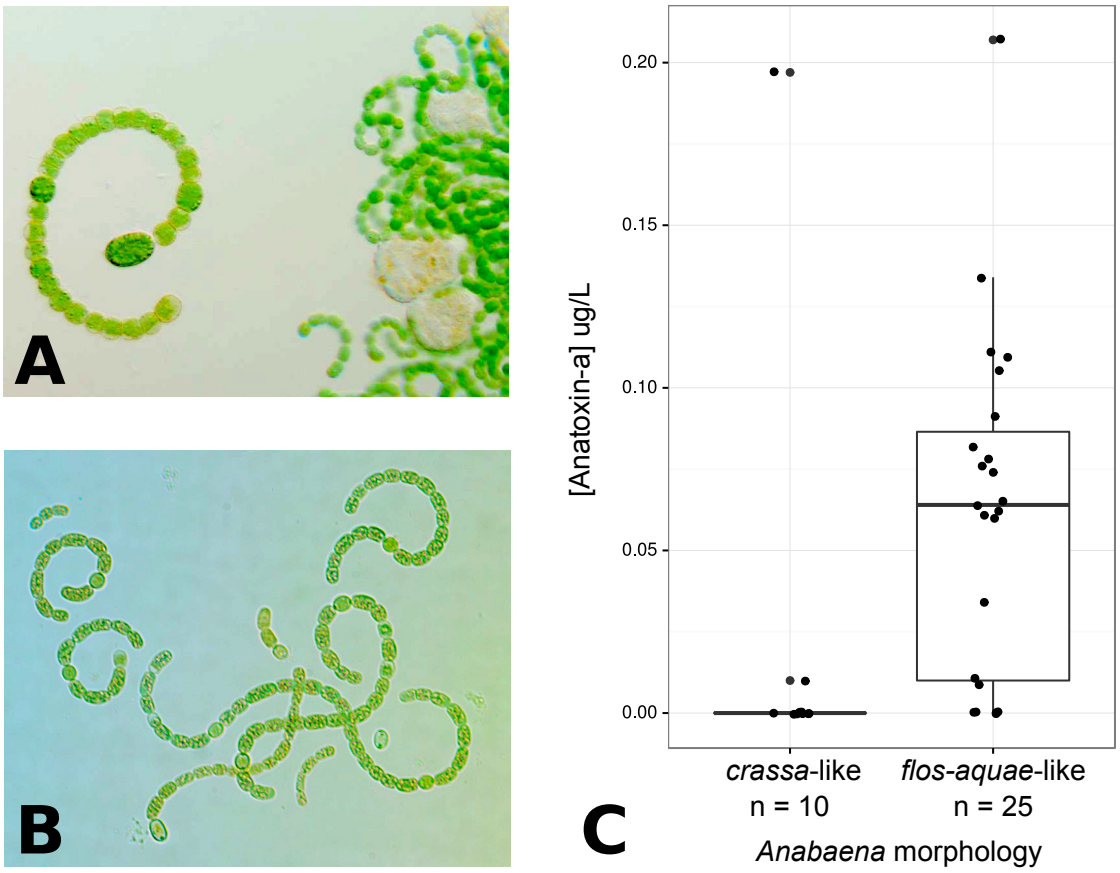

Figure 4.3: *Anabaena* morphotypes in Anderson Lake. A) Two *Anabaena* morphologies present in Anderson Lake on June 18th, 2013 (phase contrast, 200x magnification). A colony exhibiting the large-cell *Anabaena-crassa*-like morphology is shown on the left. Several intertwined filaments exhibiting the small-cell *Anabaena-flos-aquae*-like morphology are on the right. B) Filamentous colonies of the *Anabaena* sp. WA102 culture, resembling the *Anabaena flos-aquae*-like colonies in panel A. C) Anatoxin-a detection in individual colonies of *Anabaena* isolated from Anderson Lake on 18th and 25th June, 2013. Ten colonies with an *Anabaena-crassa*-like and 25 colonies with an *Anabaena-flos-aquae*-like morphology were tested for anatoxin-a by HPLC-MS/MS.

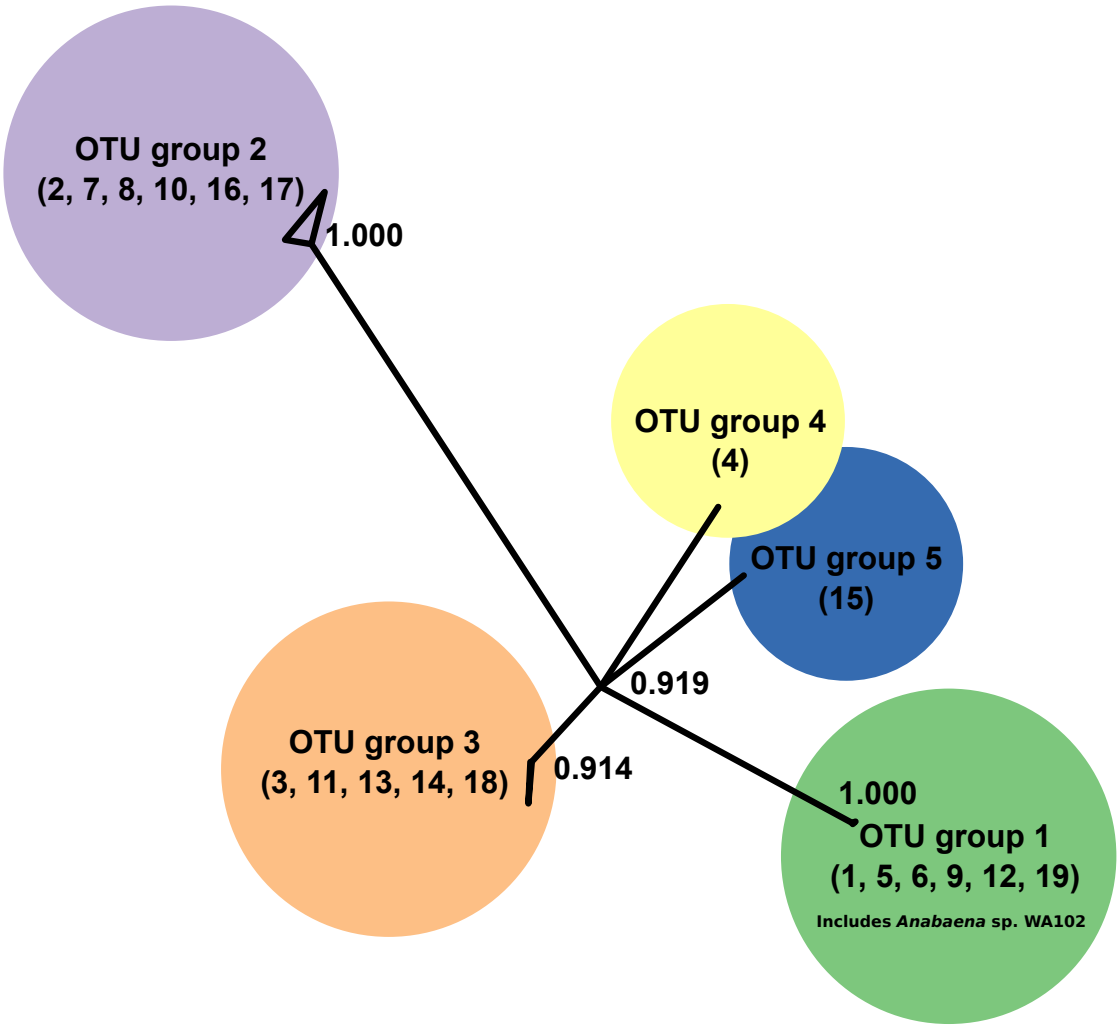

Figure 4.4: Unrooted phylogenetic tree showing relationship between the 19 most abundant *Nostocaeae cpcBA* OTUs detected in Puget Sound area lakes in 2012. Five monophyletic clades emerge, which we denote as OTU clades. The OTUs subsumed by each group are listed in parentheses underneath the OTU group name. OTU clade 1 represents *Anabaena* sp. WA102. These clades likely represent *Nostocaceae* strains detected in Puget Sound Region lakes, with intra-strain variation within each group represented by individual OTUs.

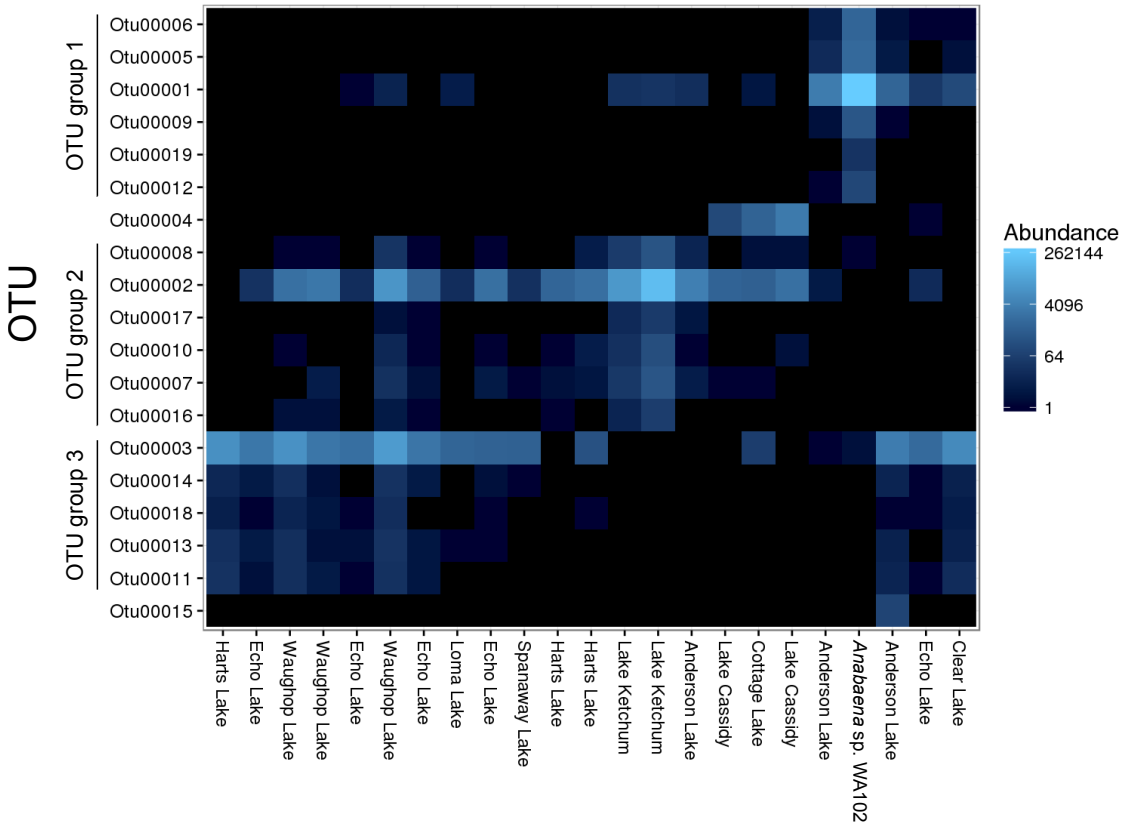

Figure 4.5: A heat-map of the OTU communities in *cpcBA*-IGS amplicon libraries by lake. The OTU clades as defined by Figure 6 are shown on the y-axis. The lake samples are arranged by similarity according to MDS of the weighted unifrac metric for each OTU community.

Position	Mutation	Frequency	Description
1152 (plasmid)	+A	1	intergenic
1473 (plasmid)	+A	1	intergenic
79606	+A	1	intergenic
239095	+G	1	intergenic
385092	+G	1	coding
469686	+T	1	intergenic
707437	+A	1	intergenic
954599	+G	1	intergenic
993032	+A	1	intergenic
1165656	+G	1	intergenic
1242731	+A	1	intergenic
1289386	+T	1	intergenic
1444344	+T	1	intergenic
1445624	+G	1	coding
1500146	Δ1 bp	1	intergenic
1882597	+G	1	intergenic
2354763	+T	1	intergenic
2366065	+G	1	intergenic
2609407	+A	1	intergenic
2644835	Δ1 bp	1	intergenic
2700751	+T	1	coding
2805367	+G	1	intergenic
2842547	+T	1	intergenic
2994651	+T	1	intergenic
3016229	+G	1	intergenic
3046399	+T	1	intergenic
3068191	+T	1	intergenic
3123663	+G	1	intergenic
3210975	Δ1 bp	1	intergenic
3546967	+A	1	intergenic
3564151	Δ1 bp	1	intergenic
3692513	+G	1	coding
3736799	+A	1	intergenic
3736816	+A	1	intergenic
3736983	+T	1	intergenic
3747047	+C	1	intergenic
3769563	+T	1	intergenic
3774081	+G	1	intergenic
3774218	+G	1	intergenic
3788976	+T	1	coding
3964228	Δ1 bp	1	intergenic
3973756	+T	1	intergenic
3995848	+A	1	intergenic
4071859	+A	1	intergenic
4075444	+C	1	intergenic
4175477	+C	1	intergenic
4331943	+T	1	intergenic
4572417	+C	1	intergenic
4720867	+T	1	intergenic
4742240	+T	1	intergenic
4798252	+A	1	intergenic
4840966	Δ1 bp	1	intergenic
4895400	+C	1	intergenic
5011934	+G	1	intergenic
5076417	+T	1	intergenic
5145703	+C	1	intergenic
5652574	+A	1	coding
2462571	A→G	0.547	intergenic
3075204	Δ25 bp	0.533	intergenic
452792	A→C	0.481	L24V
932525	G→A	0.378	intergenic
3229658	A→G	0.333	intergenic
2462570	T→G	0.316	intergenic
3229649	A→C	0.292	intergenic
932534	T→A	0.285	intergenic
931010	A→G	0.277	intergenic
1479670	Δ72 bp	0.27	
2464837	C→T	0.256	G54G
694185	C→T	0.245	intergenic
3229657	Δ1 bp	0.235	intergenic
2531697	A→G	0.23	intergenic
930986	T→G	0.218	intergenic
2531702	G→A	0.207	intergenic
5187836	G→A	0.202	P13S
930993	A→G	0.198	intergenic
3541890	T→G	0.198	intergenic
2464846	C→T	0.171	V51V
2531705	T→G	0.149	intergenic
618885	T→C	0.142	*71Q
618886	A→G	0.135	*71W
372989	A→G	0.132	intergenic
2531706	A→G	0.125	intergenic
932553	T→C	0.124	intergenic
618890	G→A	0.118	intergenic
5187827	G→A	0.113	P16S
3541911	G→A	0.104	intergenic
1869592	+T	0.1	intergenic
658806	T→C	0.095	intergenic
4156524	+G	0.095	intergenic
2663634	Δ1 bp	0.079	intergenic
3280532	C→T	0.07	intergenic
790525	A→G	0.069	intergenic
932542	A→G	0.056	intergenic

Table 4.1: Differences between the *Anabaena* sp. WA102 population genome and the *Anabaena* sp. WA102 reference genome. The *Anabaena* sp. WA102 population genome clustered from the July 7th, 2012 metagenome of Anderson Lake surface water shows 57 differences that occur in all reads mapped to the *Anabaena* sp. WA102 reference genome. Average read depth coverage was 399x and 6,199 nucleotides had no read coverage.

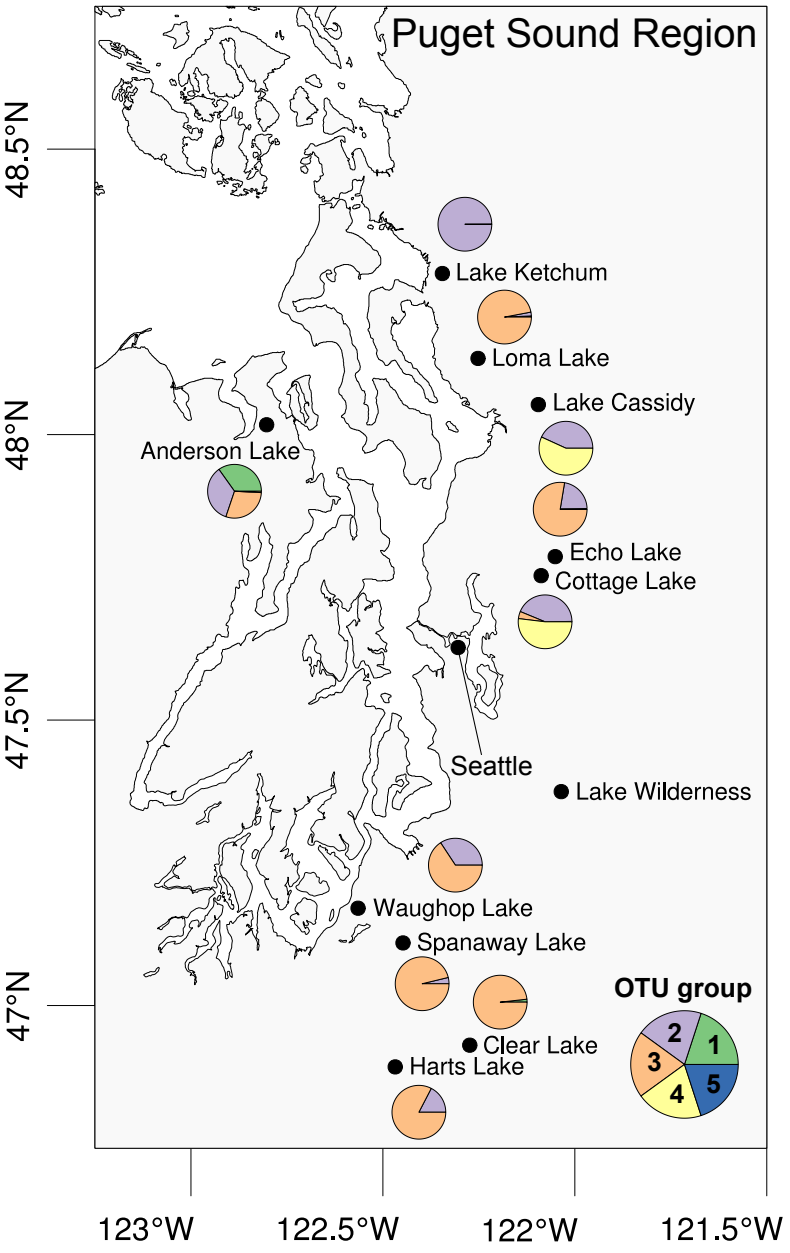

Figure 4.6: The proportion of OTU clades in each lake and their distribution across lakes sampled in the Puget Sound region in 2012.

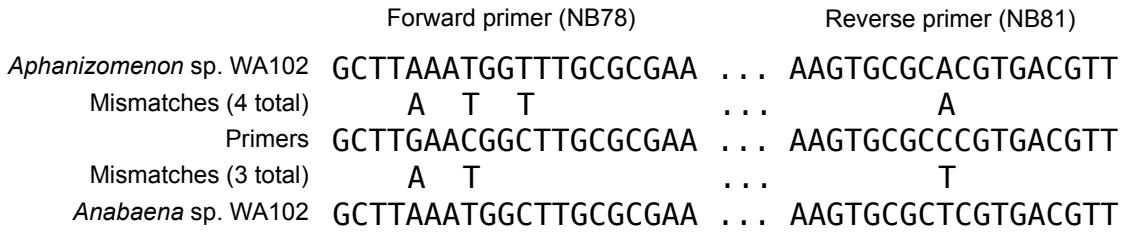

Figure 4.7: Comparison of *cpcBA*-IGS primer sequences and primer annealing sites in *Aphanizomenon* sp. WA102 and *Anabaena* sp. WA102. *Aphanizomenon* sp. WA102 has one more mismatch than *Anabaena* sp. WA102, presumably preventing its detection in DNA samples from Anderson Lake.

Chapter 5 The genome of a novel freshwater picocyanobacterium, *Cyanobium* sp. LC18, and lysogenization by of one of its temperate phages, C-CRS01

Nathan M Brown, Ryan S Mueller, Jeff H Chang, Connie Bozarth, and Theo W Dreher

In preparation

5.1 Introduction

The picocyanobacteria include three genera of cyanobacteria, *Prochlorococcus*, *Synechococcus*, and *Cyanobium*, that form a clade and are characterized by cell diameters less than 3 μm [13]. These genera are thought to be some of the earliest members of the *Cyanobacteria*, and resolving their phylogeny is important for understanding major biogeochemical events in Earth's ancient history [7]. *Prochlorococcus* and marine *Synechococcus* are well studied in marine systems, however relatively poorly studied in freshwater ecosystems. There are both freshwater and marine strains among the *Synechococcus* and *Cyanobium* genera, but of the 43 genome sequences from these two genera deposited in NCBI Genbank, only 11 (approximately 25%) are from freshwater strains as of February 2016. Considering the greater diversity among freshwater strains than among marine strains identified in marker gene studies of the *Synechococcus* and *Cyanobium* genera [130], much of the genomic diversity of this ancient clade remains unsampled. In addition, the *Cyanobium* genus is the most common of the picocyanobacterial genera found in freshwater lakes [14], where picocyanobacteria can contribute as much as 80% of photosynthetic activity in a freshwater lake [146]. While picocyanobacteria are considered non-bloom-forming cyanobacteria, they can grow densely enough to out-compete larger photoautotrophic species in oligotrophic lakes [146].

Cyanobacteria-bacteriophage (phage) interactions are not well studied in lakes, but evidence is accumulating that suggests cyanobacterial phages may sometimes play a role in freshwater cyanobacterial strain turnover and bloom collapse

[111, 49]. Studies of phage that infect marine picocyanobacteria have yielded insights into cyanobacterial physiology and ecology, showing how bacteriophages manipulate their hosts' metabolism during infection [155, 87], and unveiling a viral carbon shunt in the global carbon cycle [24].

We isolated a novel picocyanobacterial species, *Cyanobium* sp. LC18, from Upper Klamath Lake in Klamath County, Oregon, USA. We sequenced the genome of *Cyanobium* sp. LC18 and determined its phylogentic relationship to other member of the *Synechococcus* and *Cyanobium* genera. We also isolated a novel temperate cyanophage, S-CRS01, on *Cyanobium* sp. LC18 and described the integration of the S-CRS01 genome into the genome of *Cyanobium* sp. LC18.

5.2 Results

5.2.1 *Cyanobium* sp. LC18 is a novel cyanobacterial species from the Klamath River System

Cyanobium sp. LC18 was isolated from the Klamath River System in 2007 (Figure 5.1). The non-axenic culture was sequenced and assembled into 27,025 contigs with an N50 of 3,485 bp and total size of 47.0 Mbp. A draft genome of 133 contigs was clustered from the assembly using the mmgenome package to separate them out by average contig coverage depth and average GC content. The draft genome is 3,290,107 nucleotides in size, and contains 108 single-copy essential marker genes from the mmgenome HMM marker gene database and 106 unique single-copy es-

sential marker genes. That there are two single-copy essential marker genes (108 - 106) that are not unique indicates that there is some level of contamination. Both the number of unique single-copy essential marker genes and genome size are in concordance with the reference *Cyanobium* sp. PCC 6307 genome, which contains the same number of marker genes from the same database, and a genome size of 3,342,364 nucleotides. This suggests that the *Cyanobium* sp. LC18 draft genome bin is nearly complete and relatively uncontaminated. A phylogenetic tree showing its relationship with other picocyanobacterial genomes was built using 514 amino acid sequences from putative proteins shared between 25 *Synechococcus* and *Cyanobium* genomes (Figure 5.2). *Cyanobium* sp. LC 18 formed a clade with other freshwater *Cyanobium* species among species of marine *Synechococcus*. Its closest relative is *Cyanobium* sp. PCC 6307, and with a genome-wide average nucleotide identity (gANI) of 93.11%, it is a distinct species (proposed gANI threshold for species is 96.5%) [157].

5.2.2 S-CRS01 infects *Cyanobium* sp. LC18

Surface water samples from Copco Reservoir were collected on September 30th, 2008 and viruses were concentrated with PEG precipitation. The virus concentrate was added to exponential-phase *Cyanobium* sp. LC18 in BG-11 top-agar. Plaques less than 1 mm in diameter were observed on BG-11 top-agar plates seeded with culture (Figure 5.3A). Agar plugs taken from several of these plaques were resuspended in SM buffer, and sterile-filtered. DNA from 8 resuspended plaques was

prepared for pulsed-field gel electrophoresis. Each plaque showed a single band with the same migration pattern, indicating a putative phage genome of between 50-60kb (Figure 5.3B), within the typical size range for phage genomes (which are most often from the order *Caudovirales*). Transmission electron microscopy showed that the plaques contained *Siphoviridae*, members of the *Caudovirales* with prolate icosahedral heads (59 nm wide by 84 nm long) and flexible, non-contractile tails (Figure 5.3C). The sterile-filtered plaque suspension was used to inoculate a second top-agar plate seeded with exponential-phase *Cyanobium* sp. LC18. These steps were repeated once more on a third top-agar plate, confirming that a passageable, filterable agent was causing infection of *Cyanobium* sp. LC18. An isolated plaque resuspended from the third top-agar plate was used to infect a culture of *Cyanobium* sp. LC18 in liquid BG-11 media and prepare a phage stock (Figure 5.3D). The density of C-CRS01 virions was 1.459 g/mL when measured by the refraction index of the phage fraction from an equilibrium CsCl gradient, indicating that the virion is approximately 40% DNA and 60% protein (common for *Caudovirales*).

5.2.3 The C-CRS01 genome

DNA isolated from the phage stock was sequenced on the Illumina HiSeq platform with 100 bp paired-end reads and assembled into a single 60,518 nt contig. The GC content of the genome was 66.3%, slightly lower than the host genome GC content of 69.6%. 85 ORFs were identified with Prodigal and 31 were annotated. 28 ORFs

were annotated with high confidence (greater than 80%) using Phyre2 tertiary structure-guided alignment and 3 ORFs were annotated by BLASTP alignment, leaving 54 ORFs annotated as hypothetical proteins (Table 5.2 and Figure 5.4). Among annotated proteins were the expected structural proteins such as the major capsid protein, portal protein, major tail protein, and tail tape-measure protein. In addition, a phage-like integrase and Xis-like directionality factor suggested a temperate lifestyle for C-CRS01, explored further below.

Like many other members of the *Siphoviridae*, C-CRS01 encodes a pair of RecTE-like proteins (C-CRS01_00018 and C-CRS01_00020). Although the role of these proteins in phage replication is not clearly understood, they have been co-opted as genetic engineering proteins for recombineering, a high-efficiency *in vivo* bacterial genome engineering method [115]. Importantly, efficient RecTE-driven recombineering is limited to bacteria that are closely related to the bacterial host to which the RecTE phage proteins are native. The potential of these proteins for genome engineering in picocyanobacteria is currently being explored.

Three proteins in the C-CRS01 genome, the NblA-like protein, Rubisco-fold-like protein, and spore photoproduct lyase (C-CRS01_00050, C-CRS01_00057, and C-CRS01_00067) may be a signature of the photosynthetic lifestyle of its host, *Cyanobium* sp. LC18. Phages that infect photosynthetic hosts and encode genes involved in photosynthesis have been dubbed photosynthetic phages, because they modulate photosythesis during infection [86]. NblA is a polypeptide in many cyanobacteria necessary for photobleaching in response to sulfur, nitrogen, and phosphate starvation and has been found in other cyanophages that infect fresh-

water cyanobacteria. Cyanobacteria photobleach under nutrient starvation, hydrolyzing their phycobilisome proteins, which can comprise half of the protein mass of a cell, to harvest the significant stores of sulfur, nitrogen, and phosphate for cellular metabolism. The C-CRS01 NblA-like amino acid sequence is similar to protein L107_02539 in *Cyanobium* sp. LC18 (BLASTP e-value 1x10-14, 46.15% identity). L107_02539 is also an NblA-like protein-coding gene (Phyre2 confidence score 92.8%). The NblA protein has been found in many cyanophages, is relatively highly conserved in amino acid sequence, and is considered a signature gene of cyanophages [169, 46, 98]. Rubisco (d-ribulose-1,5-bisphosphate carboxylase/oxygenase) is present in all photosynthetic organisms and is responsible for fixing carbon from atmospheric carbon dioxide. The amino acid sequence from the C-CRS01_00057 gene found in C-CRS01 bears similarity to the C-terminal 40 amino acids of the Rubisco small subunit from spinach (amino acids 80-121 in Uniprot Q43832). Although the role of the Rubisco small subunit is unknown, directed point mutations in amino acids 88-104 from the C terminus region of *Synechococcus* sp. PCC6301 increased the Km of the Rubisco holoenzyme for CO_2 by 0.5 to 2-fold [41]. It is possible that the polypeptide encoded in C-CRS01 may alter the activity of the host Rubisco holoenzyme during infection or latency. Spore photoproduct (5-thyminyl-5,6-dihydrothymine) is the most common pyrimidine photodimer that forms in dsDNA under UV irradiation. The spore photoproduct lyase, originally found in *Bacillus* spores, monomerizes thymine nucleobases from the spore photoproduct using SAM radical chemistry (rather than the photochemical energy used by many photolyases) after the spores germinate [139]. Its activity

contributes to the extreme longevity of *Bacillus* spores. Surprisingly, this is the first time that a spore photoproduct lysase or photolyase has been observed in a cyanophage. A spore photoproduct photolyase would provide effective protection for phage DNA in an environment with high levels of UV radiation, such as those inhabited by cyanophages and their photosynthetic hosts.

5.2.4 C-CRS01 lysogenizes *Cyanobium* sp. LC18

That C-CRS01 encodes an integrase and Xis-like directionality factor suggests that it can integrate into the *Cyanobium* sp. LC18 genome. To test this hypothesis, putative *Cyanobium* sp. LC18 [C-CRS01] lysogens were grown by isolating and outgrowing a colony that grew in the middle of a lacuna formed by C-CRS01 on a lawn of *Cyanobium* sp. LC18 in BG-11 top agar. The isolated colony was outgrown in liquid BG-11 and DNA was extracted from the culture. The genome was sequenced and analyzed as described above. A single 161.1 kbp contig was assembled that contained the full C-CRS01 prophage with 61.5 kbp of bacterial genome sequence adjacent to the *attL* site and 38.9 kbp of bacterial genome sequence adjacent to the *attR* site (Figure 5.5A). The *att* core sequence shared by *attP*, *attB*, *attL*, and *attR* was determined to be TACGACATCCGTGA. The integration junctions were verified by PCR, using primers that anneal on either side of the *attB*, *attP*, *attR*, and *attL* sites and extend across the junctions (Figure 5.5B). Interestingly, C-CRS01 integrated into a hypothetical-protein-coding gene in *Cyanobium* sp. LC18. The hypothetical protein encodes a helix-turn-helix do-

main and may bind DNA. Often temperate phages will integrate into the middle of tRNA sequences on the host chromosome and complement the disrupted tRNA by supplying the missing portion of the tRNA in its own genome [124]. In this case, C-CRS01 complements the disrupted hypothetical protein with a homologous sequence encoded in its own genome (Figure 5.5C and D). The C-terminus of the phage-encoded protein sequence includes an additional eight amino acids, AAAADDPA. The alanine-rich nature of this C-terminal sequence suggests that it may be a proteolytic tag [42].

5.3 Discussion

5.3.1 Relevance of the *Cyanobium* sp. LC18 genome to diversity-driven sequencing of the *Cyanobacteria*

Cyanobium sp. are the most common cyanobacterial species in freshwater ecosystems [14]. In addition, *Cyanobium* species can encode useful natural products [76]. However, only 3 genomes representing this genus are available in NCBI Genbank (as of February 2016). Considering their ubiquity in freshwater ecosystems and their potential utility, more genome sequences from this genus should be contributed to the public databases. Future freshwater ecological studies and natural product biomining efforts will benefit from greater genome sequencing coverage of this genus. Towards this end, we have isolated a novel species of picocyanobacteria, *Cyanobium* sp. LC18, from the freshwater Klamath River System and deposited

its draft genome sequence in Genbank.

5.3.2 Isolation of a novel freshwater temperate *Siphovirus*, C-CRS01

The novel temperate cyanobacterial phage C-CRS01 was shown to infect and lyse *Cyanobium* sp. LC18. This specific interaction between a bacteriophage and a member of the the most abundant genus in freshwater lakes may shed light on bacteriophage/cyanobacterial dynamics in freshwater ecosystems in general. Some studies have observed a proliferation of virus-like particles (VLPs) during a collapse in freshwater cyanobacterial blooms, though these observations have not carried convincing statistical support for the correlation between the two events [111, 49]. However, the hypothesis of viral-induced freshwater cyanobacterial bloom collapses is now beginning to attract serious attention [111, 163]. A key challenge of testing the hypothesis that bacteriophages cause freshwater cyanobacterial bloom collapses is identifying and quantifying the bacteriophages responsible for infecting cyanobacterial bloom strains. This is difficult when measuring the quantity of all bacteriophages in a lake ecosystem, since they are so abundant for all the diverse bacterial species that may be present. Effective methods for identifying and quantifying the dynamics of a particular bacterium or phage include quantitative or digital PCR. In order to track a specific organism with PCR methods, an appropriate PCR amplicon marker must be selected. Sequencing phage and host cyanobacterial genomes will expand the list of possible targets to track via highly specific and quantitative methods such as quantitative or digital PCR.

Several putative proteins in the C-CRS01 genome are of special interest. The spore photoproduct lyase is the first to be seen in a cyanobacterial phage genome. Most phages rely on bacterial host photolyases to repair pyrimidine dimers that accumulate in their DNA while packaged in the capsid [21]. Host photolyases usually require activation by blue light in order to repair dimers. The spore photoproduct lyase is interesting in that it is independent of light, suggesting that there is some utility to the phage in resolving pyrimidine dimers without a need for blue light. The spore photoproduct lyase may also confer similar longevity to C-CRS01 phage that are inert for long periods as it does to *Bacillus* spores [30]. The RecT-like recombinase and RecE-like exonuclease are also especially interesting. These homologs of the RecTE proteins from the Rac prophage in *Escherichia coli* are likely capable of facilitating recombineering. Recombineering is an *in vivo* genomic engineering technique that efficiently incorporates single-stranded donor DNA molecules (such as synthetic oligonucleotides) into a host chromosome at the replication fork [96]. Importantly, the RecTE enzymes are most efficient in bacteria closely related to the host bacterium of the phage in which the RecTE enzymes are found [26]. These RecTE enzymes may facilitate recombineering in *Cyanobium* species and related species among the picocyanobacteria.

5.3.3 Lysogenization of *Cyanobium* sp. LC18 by C-CRS01

Many temperate cyanobacterial bacteriophages have been isolated [105, 77, 43, 150] and lysogeny is a well established phenomenon in freshwater cyanobacteria

[144]. However, C-CRS01 is the first temperate cyanobacterial phage for which the integration site has been described and experimentally verified. Much recent attention has been focused on the benefits that temperate phages confer on their hosts [103, 40]. Lysogenic conversion is the process by which a prophage alters the phenotype of its host, which can have a positive impact on the fitness of a lysogen. C-CRS01 and *Cyanobium* sp. LC18 constitute an experimentally tractable system to study the effects of lysogenic conversion in cyanobacteria.

5.4 Methods

5.4.1 Isolating *Cyanobium* sp. LC18 and culture maintenance

Cyanobium sp. LC18 was isolated on May 31st, 2008 from Upper Klamath Lake, Klamath County, Oregon (42.4160174°N , 237.9062507°W). The non-axenic culture was maintained under 10 μEm^{-2}s^{-1} cool white fluorescent light with a light/dark cycle of 16 hr/8 hr at 24°C in BG-11 medium.

5.4.2 Isolating C-CRS01 on *Cyanobium* sp. LC18

Samples from Upper Klamath Lake were collected from surface water at multiple sites in the reservoir on September 30th, 2008 and stored at 4°C. Samples were combined and approximately 2L was filtered through a 0.2 μM-pore-size Supor membrane filter (Pall Corporation, Port Washington, NY, USA) to remove most cellular organisms. NaCl and PEG 8000 were added to the filtrate (final con-

centrations: 1M NaCl and 2.7% PEG 8000) and stirred at 4°C in the dark for 5 days. The solution was centrifuged at 9,000 RCF for 30 min, supernatant was decanted, and the pellet was resuspended in 5 mL of SM buffer (100 nM NaCl, 8 mM $MgSO_4$·7H2O, 50 mM Tris·HCl (1M, pH 7.5)). Chloroform was added to the resuspension, gently mixed, and centrifuged at low speed for 2 min to remove remaining PEG 8000. The aqueous phase was extracted and stored at 4°C in the dark. 20 μL of the phage resuspension was mixed with 0.5 mL of *Cyanobium* sp. LC18 exponential phase culture in 4.5 mL of molten BG-11 top agar (0.75% w/v agar) at 45°C. The mixture was kept at 45°C for 15 min to encourage phage adsorption, then plated on BG-11 agar plates and incubated under 10 $\mu Em^{-2}s^{-1}$ cool white fluorescent light with a light/dark cycle of 16 hr/8 hr at 24°C. Plaques were observed after 1 month of bacterial growth. Plaques were picked with a sterile pipette tip, resuspended in 100 μL SM buffer, and stored at 4°C in the dark. A resuspended plaque was then diluted and serially replated twice on fresh *Cyanobium* sp. LC18 top-agar plates, as described above, to purify the phage. Phage stock was prepared by ultracentrifuging 60 mL of sterile-filtered C-CRS01 lysate (10 mL per each of 6 ultraclear thinwall tubes (Beckman-Coulter, Brea, CA), layered on top of a 1 mL 5% sucrose cushion) in a Beckman-Coulter L8-70M ultracentrifuge using a SW41 rotor at 150k RCF for 170 min. Each pellet was resuspended in 100 μL SM buffer, combined with the other resuspensions, and stored at 4°C in the dark. 1 mL of a second phage stock was purified on a CsCl equilibrium gradient by isopycnic ultracentrifugation (layered on top of 8mL of 1.5 g/mL CsCl solution in a tube in an SW41 swinging rotor and centrifuged at 150k RPM for 20 hr). The

characteristic opalescent phage band was extracted from the centrifuge tube and the refractive index of the suspension was measured with an Abbé refractometer.

5.4.3 Transmission Electron Microscopy of C-CRS01

CsCl gradient-purified phage was applied to a glow-discharged carbon-type B, 300-mesh copper grid (Ted Pella, Redding, CA, USA) and stained with 1% phosphotungstic acid, pH 6.5. Samples were observed on a Philips CM-12 transmission electron microscope at 60 kEV.

5.4.4 Isolating *Cyanobium* sp. LC18 lysogens

Cyanobium sp. LC18 lysogens containing the C-CRS01 prophage were obtained by spotting 10 μL of a 10^9 PFU/mL C-CRS01 phage stock onto a BG-11 top-agar lawn of *Cyanobium* sp. LC18. After two weeks of growth, a lacuna approximately 6 mm in diameter developed where the phage had been placed, with well separated cyanobacterial colonies growing within the clearing. One of these isolated cyanobacterial colonies - presumed to be a lysogen growing due to superinfection immunity - was picked with a sterile plastic pipette tip, used to inoculate 10 mL of liquid BG-11 media, and outgrown. 1 mL of the outgrown culture was added to 1 mL sterile 50% glycerol and stored in a 2 mL cryovial at -80°C. DNA was isolated from the remaining culture for genome sequencing.

5.4.5 DNA preparation for C-CRS01 and *Cyanobium* sp. LC18

Genomic DNA was extracted from C-CRS01 phage stock by adding RNAse A, 1 M CaCl, and DNAse I (to final concentration of 250 ng/mL, 50 mM, and 0.1 U/mL, respectively) and incubating for 1 hr at room temperature to remove exogenous nucleic acids. Capsids were digested by adding 20% w/v SDS and 20 mg/mL Proteinase K (to a final concentration of 0.2% w/v and 0.2 mg/mL, respectively) and incubating for 1 hr at 37°C. Excess protein was separated in a 1:1 phenol:chloroform phase (pH 8.0) and DNA was extracted and then precipitated with 0.5 vol of 7.5 M ammonium acetate and 2 vol isopropanol at room temperature for 10 min, centrifuged for 10 min at 13k RCF, the supernatant was decanted, and the pellet was air-dried and resuspended in TE buffer (10 mM Tris-HCl, pH 8.0 and 1 mM EDTA). *Cyanobium* sp. LC18 genomic DNA was extracted by centrifuging 1 mL of culture at 5k RCF, decanting the supernatant, and resuspending the pellet in TNE buffer. The resuspended pellet was treated using a method from Neilan *et al.* [126].

5.4.6 High-throughput sequencing of *Cyanobium* sp. LC18 lysogen and C-CRS01 genomes

Cyanobium sp. LC18 lysogen and C-CRS01 genomic DNA were sequenced as individual libraries prepared with the Nextera XT library preparation kit on the Illumina HiSeq 2000 platform using 100-nt paired-end reads.

5.4.7 Assembling and analyzing the *Cyanobium* sp. LC18 lysogen and C-CRS01 genomes

The *Cyanobium* sp. LC18 genome was assembled using idba_hybrid v1.1.1 on default settings, except with a kmer step size of 10. The putative protein-coding contents of *Cyanobium* sp. LC18 was annotated using Prokka version 1.11. Protein contents from all other strains used in the phylogenomic tree was downloaded from NCBI Genbank (NCBI Genbank Accession Numbers: NZ_AATZ00000000, NZ_ADXL00000000, NZ_ADXM00000000, NC_008319, NC_007516, NZ_AZXL00000000, NC_007513, NZ_LFEK00000000, NC_007776, NC_007775, NZ_CP006269, NZ_CP006270, NZ_CP006271, NZ_BAWS00000000, NZ_BAUB00000000, NC_019681, NC_019680, NC_010480, NC_010479, NC_010478, NC_010477, NC_010476, NC_010475, NC_010474, CP014003, CP014002, CP014001, CP014000, CP013999, CP013998, NZ_ABRV00000000, NZ_CM001776, NZ_ALWC00000000, NC_019692, NC_019691, NC_019702, CP000100, NC_009482, NZ_AAUA00000000, NZ_AANP00000000, NZ_CP006473, NZ_CP006472, NZ_CP006471, NZ_AANO00000000, NC_009481, NZ_AAOK00000000, NZ_AGIK00000000, NZ_CP011941, NC_005070, NZ_LN847356, NZ_CP006882, NC_019675, NZ_DS990557, NZ_DS990556, JMRP01000001-JMRP01000071). Protein-coding contents from each of the 25 genomes were used to build a genome-wide phylogenetic tree. The protein sequences were subjected to an all-versus-all BLASTP alignment to identify orthologs that occur once in each genome. These were clustered with the mcl algorithm and aligned with muscle [39, 36]. Protein

alignments were masked with zorro to reduce noise from uninformative amino acid alignment positions and checked for a best fit among protein evolution models with ProtTest version 3.1 [165, 25]. The best-fit protein evolution model was used in RAxML to generate the final tree, which was rooted within the *Nostoc* genus outgroup at *Nostoc* sp. 7107, in accordance with Shih *et al.* [142, 135]. Phyre2 and BLASTP were used to annotate putative proteins in the C-CRS01 genome [68, 16]. The C-CRS01 genome was plotted using Circos [74].

5.4.8 PCR amplification of prophage junctions

The *attL*, *attR*, *attB*, and *attP* sites of the C-CRS01 phage, the *Cyanobium* sp. LC18 lysogen, and the non-lysogenic *Cyanobium* sp. LC18 were PCR amplified to detect their presence in each sample. The *attL* and *attR* amplicons from *Cyanobium* sp. LC18 lysogen were sequenced to verify the prophage junction sequences that were assembled from high-throughput sequencing of the lysogen. Primers NB37 and 40 were used to amplify the *attB* site, NB37 and 38 to amplify the *attL* junction, NB39 and 40 to amplify the *attR* junction, and NB38 and 39 to amplify the *attP* site (Table 5.1). A master mix of 0.5 μL Phusion HiFi polymerase (ThermoScientific, Pittsburgh, PA), 1 μL dNTPs (10mM), 10 μL 5x Phusion GC buffer, 2.5 μL forward primer (10μM), 2.5 μL reverse primer (10 μM), 1 μL template DNA, 3 μL 1,2-propanediol, and 29.5 μL H2O was used. PCR cycling conditions were: initial denaturation at 95°C for 3:00, then 35 cycles of denaturation at 98°C for 0:20, primer annealing at 71°C for 0:15, and extension at 72°C for 0:30, followed

by storing the samples at 4°C prior to agarose gel electrophoresis.

5.5 Tables and Figures

Primer	Sequence	Position	T_m
NB0037	AGCCCAAGCCGTTCTTCTG	attL 5' end (bacterial chromosome)	67
NB0038	AGGGTCTGCAACAGCTTGG	attL 3' end (prophage chromosome)	65.9
NB0039	ATCCGCTGCTCGGGTTGATG	attR 5' end (prophage chromosome)	71.9
NB0040	AACGCCTGGGACCGTTTCTC	attR 3' end (bacterial chromosome)	69.8

Table 5.1: Primers used for PCR amplification across phage integration junctions.

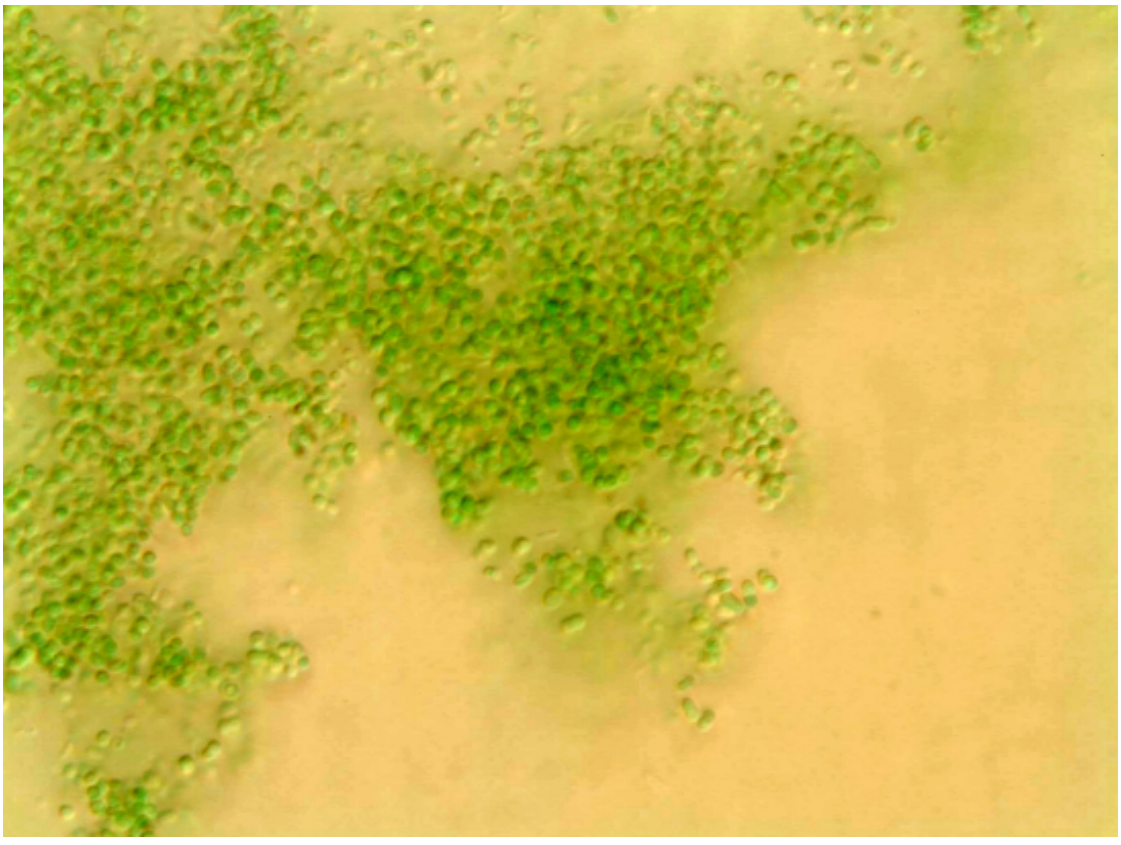

Figure 5.1: Micrograph of *Cyanobium* sp. LC18. A 200x magnification brightfield micrograph of *Cyanobium* sp. LC18

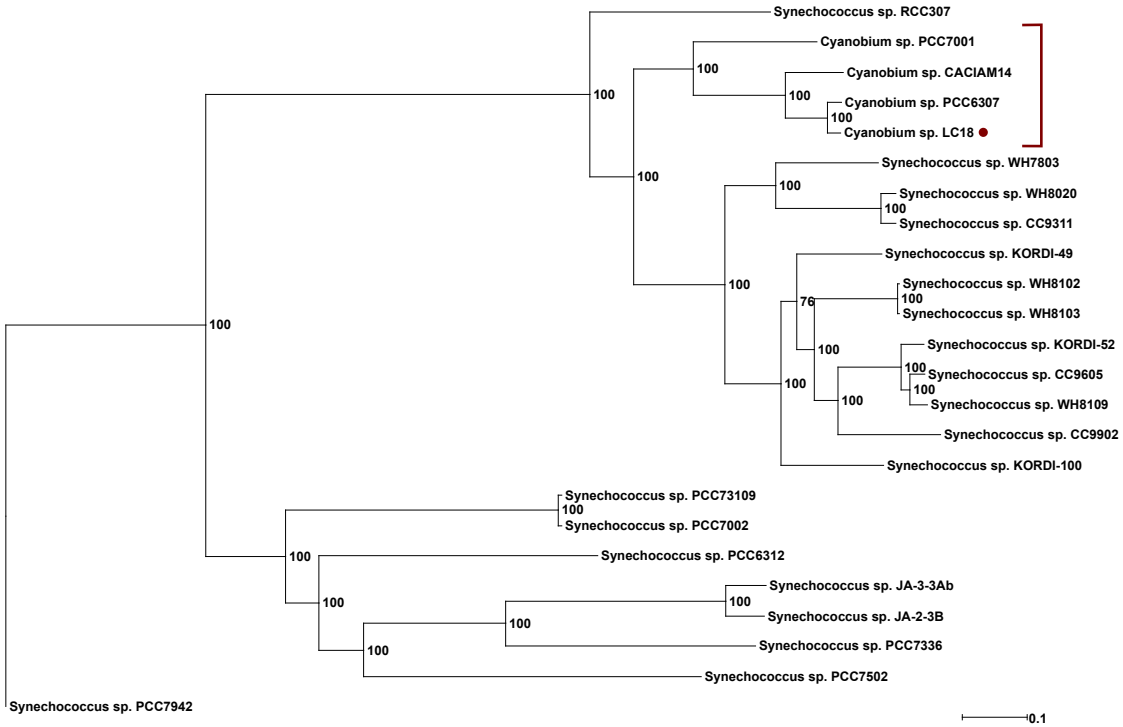

Figure 5.2: Phylogenomic tree of the *Cyanobium* sp. LC18 genome and 24 other *Cyanobium* and *Synechococcus* genomes. 514 amino acid sequences for orthologs shared between all 25 genomes in the analysis were concatenated and aligned to propose the evolutionary relationship between *Cyanobium* sp. LC18 and other picocyanobacteria.

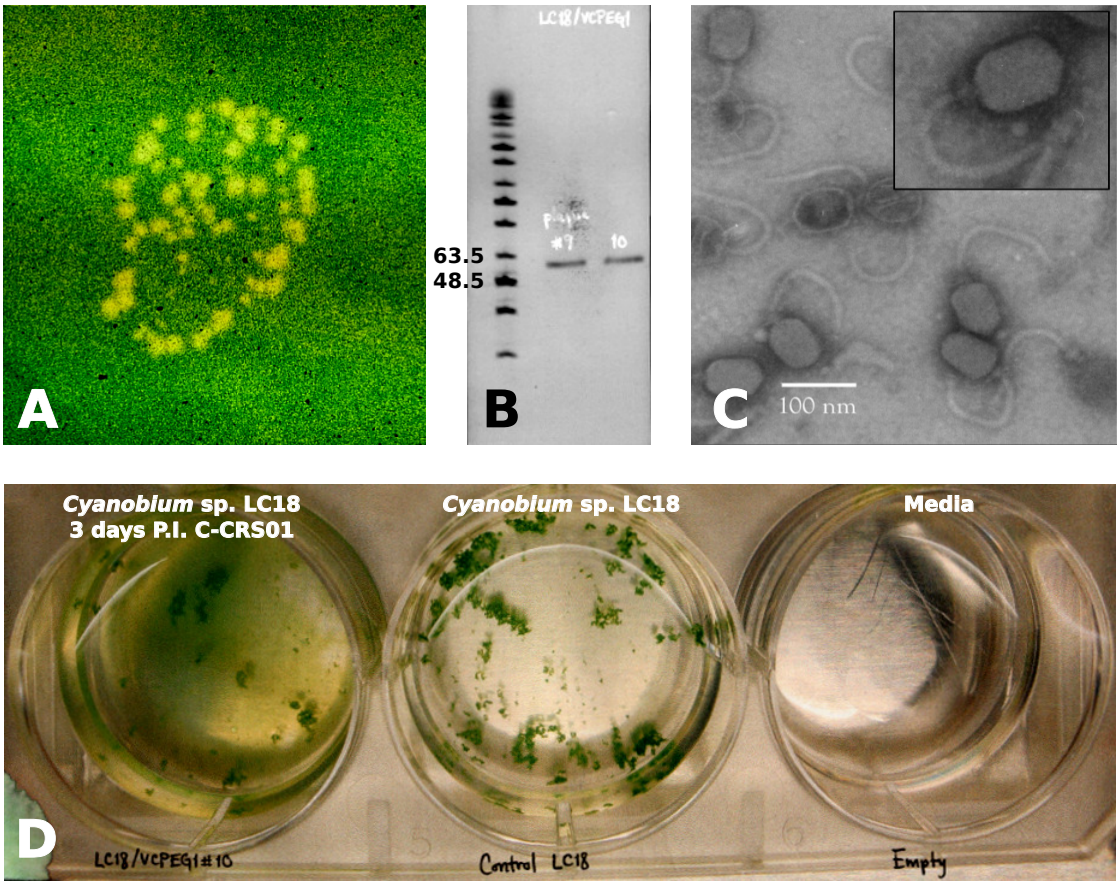

Figure 5.3: Evidence of the novel cyanobacterial phage C-CRS01 infecting *Cyanobium* sp. LC18. A) Plaques less than 1mm in diameter forming on a lawn of *Cyanobium* sp. LC18 in a BG-11 top-agar plate. B) Pulsed-field gel electrophoresis of DNA extracted from two different plaques shows a phage genome size of approximately 60 kbp. C) 45,000x magnification transmission electron micrographs reveal that the phage is a member of the *Siphoviridae*, with a flexible, non-contractile tail and prolate icosahedral capsid. D) Comparison of a liquid lysate of *Cyanobium* sp. LC18 and an uninfected culture 3 days after infection.

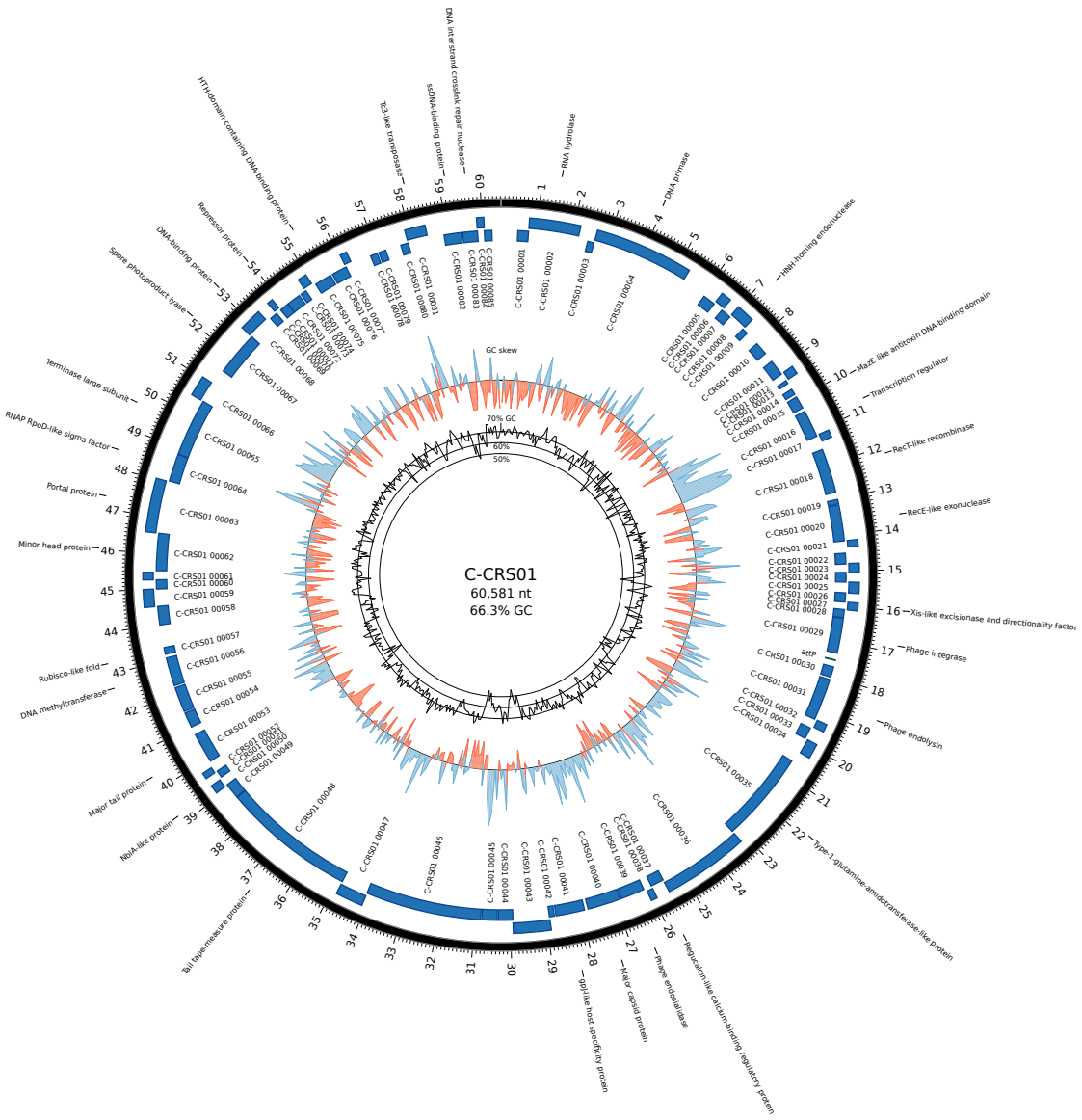

Figure 5.4: The C-CRS01 genome. The 60,581 bp C-CRS01 genome is shown with annotated ORFs in the outermost track, all detected ORFs shown in the first inner track, average GC skew calculated in a 500-nt sliding window shown in the next innermost track, and average GC content calculated in a 500-nt sliding window in the innermost track.

Annotation	Locus tag	Phyre2 % confidence/ BLASTP E-value	Phyre2 % identity/ BLASTP % identity	Top PDB hit/ Top NCBI NR hit
RNA hydrolase	C-CRS01_00002	100 / -	19 / -	2xgj / -
DNA primase	C-CRS01_00004	100 / -	15 / -	2au3 / -
HNH homing endonuclease (I-HmuI-like)	C-CRS01_00008	99.8 / -	28 / -	1u3e / -
MazE-like antitoxin DNA-binding domain	C-CRS01_00015	94.5 / -	31 / -	2mrn / -
Transcription regulator	C-CRS01_00016	97 / -	20 / -	3pxp / -
RecT-like recombinase	C-CRS01_00018	- / 3×10^{-105}	- / 58	- / WP_007082136.1
RecE-like exonuclease	C-CRS01_00020	100 / -	27 / -	3h4r / -
Xis-like excisionase and directionality factor	C-CRS01_00027	90.6 / -	28 / -	4j2n / -
Phage integrase	C-CRS01_00029	100 / -	16 / -	1z1b / -
Phage endolysin	C-CRS01_00031	100 / -	24 / -	1xjt / -
Type 1 glutamine amidotransferase-like protein	C-CRS01_00035	98.3 / -	10 / -	3rht / -
Regucalcin-like calcium binding regulatory protein	C-CRS01_00037	85.6 / -	20 / -	2ghs / -
Phage endosialidase	C-CRS01_00039	96.3 / -	24 / -	1v0e / -
Major capsid protein	C-CRS01_00040	100 / -	22 / -	3j4u / -
Helical protein	C-CRS01_00041	97 / -	12 / -	1y4c / -
Tail tape-measure protein	C-CRS01_00048	- / 5×10^{-72}	- / 55	- / YP_007674092.1
NblA-like protein	C-CRS01_00050	88.8 / -	14 / -	1ojh / -
Major tail protein	C-CRS01_00053	93.9 / -	19 / -	2k4q / -
DNA methyltransferase	C-CRS01_00056	100 / -	24 / -	3swr / -
Rubisco-like fold	C-CRS01_00057	84.1 / -	21 / -	1ir1 / -
Minor head protein	C-CRS01_00062	- / 8×10^{-23}	- / 26	- / ACY75734.1
Portal protein	C-CRS01_00063	99.5 / -	13 / -	2jes / -
RNAP rpoD-like sigma factor	C-CRS01_00064	100 / -	22 / -	4igc / -
Terminase large subunit	C-CRS01_00065	100 / -	16 / -	4bij / -
Spore photoproduct lyase	C-CRS01_00067	100 / -	15 / -	4fhe / -
DNA-binding protein	C-CRS01_00068	99.9 / -	17 / -	4rsf / -
Repressor protein	C-CRS01_00071	87.4 / -	50 / -	3b73 / -
HTH-domain-containing DNA-binding protein	C-CRS01_00075	87.6 / -	31 / -	2cob / -
Tc3-like transposase	C-CRS01_00081	89.8 / -	24 / -	1u78 / -
ssDNA-binding protein	C-CRS01_00082	100 / -	22 / -	1qvc / -
DNA interstrand crosslink repair nuclease	C-CRS01_00083	99.6 / -	17 / -	4qbn / -

Table 5.2: ORF annotations for C-CRS01. Annotations for ORFs identified in C-CRS01 were made by transferring annotations from homologs identified by structure-guided alignment in Phyre2 and primary amino-acid sequence alignment in BLASTP.

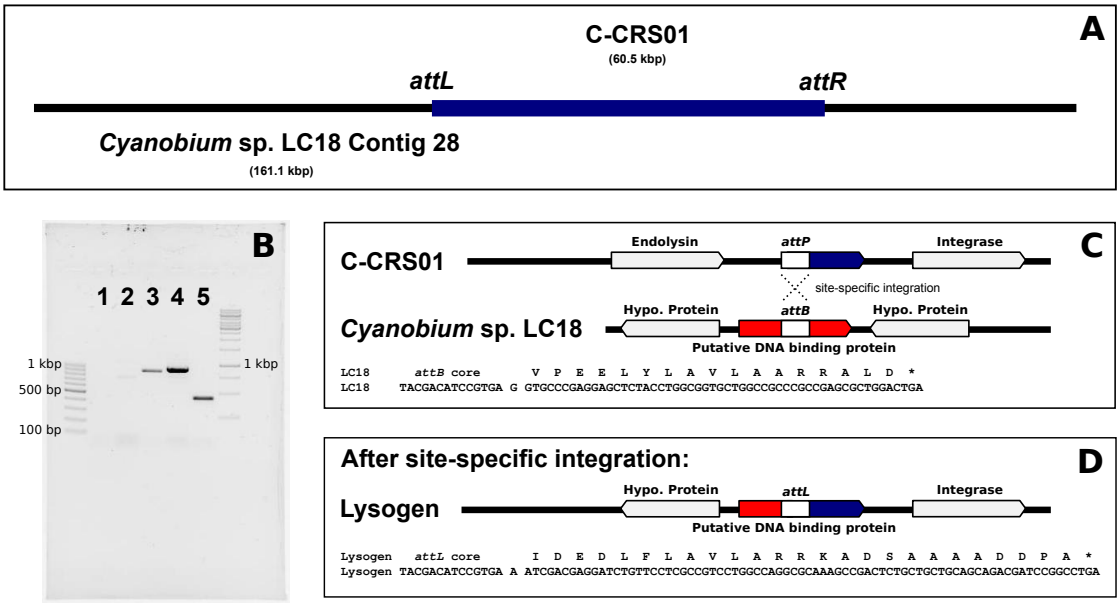

Figure 5.5: Integration of C-CRS01 genome into the host *Cyanobium* sp. LC18 genome. A) Contig 28 from the assembly of a putative *Cyanobium* sp. LC18 lysogen indicates that the C-CRS01 genome does indeed integrate into the host genome (to scale). B) PCR verification of the prophage junctions in the *Cyanobium* sp. LC18 lysogen: lanes 1 and 2 are amplifications of the *attL* and *attR* junctions (respectively) from a non-lysogenic *Cyanobium* sp. LC18 strain, showing no detection of either junction, lanes 3 and 4 are amplifications of the *attL* and *attR* junctions (respectively) from a *Cyanobium* sp. LC18 lysogen, showing detection of both junctions as expected, lane 5 shows amplification of the *attP* junction in C-CRS01 genomic DNA (no bacterial DNA) C) Diagram (not to scale) showing the site of integration for C-CRS01 in *Cyanobium* sp. LC18. C-CRS01 integrates into a putative DNA binding protein with a helix-turn-helix domain. The amino acid and DNA sequence beginning from the *attB* region of the putative protein to the C-terminus of the protein is shown. D) After integration, C-CRS01 fuses the disrupted protein at the *attL* site with a homologous amino acid sequence encoded in its genome that includes an 8-amino-acid addition at the C terminus: AAAADDPA. The sequence of the fusion protein formed is shown at the bottom.

Chapter 6 Conclusion

This body of work expands the number of freshwater cyanobacterial genomes and cyanobacterial phage genomes available in the public databases, contributing to a firm foundation for future metagenomics studies in freshwater habitats. Three novel *Nostocales* genomes, *Anabaena* sp. WA102, *Aphanizomenon* sp. WA102, and *Anabaena* sp. AL93, provide insight into genome evolution, variation in the anatoxin-a biosynthesis gene cluster, and variation in core metabolic pathways among the *Nostocales*. Large-scale genome rearrangement was observed between closely related *Anabaena* sp. WA102 and *Anabaena* sp. 90, indicating that *Anabaena* genome rearrangement may have a role in adaptation to changing conditions. Finishing additional closely related *Anabaena* genomes will allow recombination, point mutation, and gene gain/loss rates to be quantified. Comparing the anatoxin-a biosynthesis gene cluster in closely related *Anabaena* sp. WA102 and *Anabaena* sp. AL93 highlighted a triplication of the *anaB* promoter region in *Anabaena* sp. WA102. This - likely unstable - triplication is hypothesized to be under selective pressure and increase expression of the putative *anaBCD* operon in *Anabaena* sp. WA102. Testing this hypothesis by comparing *anaBCD* transcription and anatoxin-a production levels in *Anabaena* sp. WA102 and *Anabaena* sp. AL93 is a promising future research direction. Interestingly, although *Anabaena* sp. WA102 and *Aphanizomenon* sp. WA102 share the same habitat and

high sequence similarity (88.7% gANI), they swap dominance in Anderson Lake at two different time points. Comparing the core metabolic pathways encoded in their genomes shows that they have major differences in light response strategies and sulfur metabolism pathways. *Anabaena* sp. WA102 is capable of sensing and responding to changes in red light levels via the *cph1/rcp1* phytochrome two-component system and is capable of positive phototaxis because of encoding the entire *pix* operon, whereas *Aphanizomenon* sp. WA102 encodes neither system. Additionally, *Anabaena* sp. WA102 is capable of assimilating organic sulfur, whereas *Aphanizomenon* sp. WA102 is not. Both of these predicted differences in core metabolism may be clues to how *Anabaena* sp. WA102 outcompetes other similar *Nostocales* to dominate Anderson Lake during the warm summer months. A more complete time series detailing *Nostocales* population dynamics in Anderson Lake alongside environmental parameters such as temperature, light levels, day length, and nutrient levels would provide a framework for understanding how these two species compete in their natural environment.

The utility of long-read sequencing for finishing high-quality cyanobacterial genomes is demonstrated by finishing the *Anabaena* sp. WA102 reference genome. This is especially important for cyanobacterial genomes, which are refractory to assembly from short-read sequencing data. Questions about recombination and horizontal gene transfer are easiest to address with finished genomes, and since the most common mode of bacterial evolution is recombination rather than point mutation [102], these questions are paramount for understanding how bacteria evolve. Bacteria such as *Anabaena* that undergo dramatic boom/bust bloom cycles

may be highly recombinogenic during blooms due to an increase in the number of opportunities for horizontal transfer and recombination of homologous sequences between individuals in the large population of bloom bacteria [143]. Finishing more *Anabaena* genomes would allow for measuring the rate of recombination and determining the frequency of recombination versus mutation events, which would indicate how recombinogenic this genus is relative to other bacterial genera. The biosynthetic gene clusters responsible for secondary metabolite synthesis, such as anatoxin-a, are modular and hypothesized to be horizontally transferred and recombine with each other to generate novel metabolites [91]. This is the first complete genome sequence of an anatoxin-a-producing cyanobacterium, placing the anatoxin-a biosynthetic gene cluster in context with the rest of the *Anabaena* sp. WA102 genome. If more genomes from anatoxin-a-producing strains can be finished, then the relative position of the anatoxin-a biosynthesis gene clusters in each finished genome can been compared between strains for signs of wholesale horizontal gene transfer of the gene cluster and synteny within the gene clusters can be compared for signs of horizontal gene transfer.

Sequencing technology, sampling strategies, and computational tools have recently been developed that enable high-quality bacterial population genomes to be extracted from assemblies of deep shotgun metagenomes [164, 134, 1]. By extracting the population genome of *Anabaena* sp. WA102 from the 2012 Anderson Lake metagenome, we were able to compare it to our cultured *Anabaena* sp. WA102 isolate. Extracting the sequencing reads that mapped to the population genome and then mapping them to the finished *Anabaena* sp. WA102 isolate genome re-

vealed that the *Anabaena* sp. WA102 population genome from the metagenome was nearly complete, only lacking 0.11% (6,199 nt) of the full genome, and nearly clonal, with 36 polymorphisms (representing a total difference of 107 nt) detected from reads at an average read coverage depth of 399. With the major caveat that reads from the *Anabaena* sp. WA102 population genome did not map to 6,199 nt of the finished *Anabaena* sp. WA102 isolate genome, there were very few differences between the *Anabaena* sp. WA102 isolate and the *Anabaena* sp. WA102 population genome. The near clonality of the *Anabaena* sp. WA102 population genome might mean that the 2012 Anderson Lake bloom began from a severe genetic bottleneck, with one or a few closely related strains responsible for seeding the bloom. There may be interesting implications for the evolution of *Anabaena* sp. WA102 if it annually undergoes a population bottleneck. Annually sequencing the bloom strain in Anderson Lake may reveal interesting patterns of evolution related to this potential cyclic evolutionary force. Synthesizing results from the 3rd and 4th chapters, it may be possible to sequence the finished genome of the bloom strain directly from a carefully prepared lake sample using long-read sequencing technology since the bloom strain is the most abundant bacterium and relatively clonal. Further, the near clonality of the *Anabaena* sp. WA102 population genome supports the hypothesis that some freshwater cyanobacterial blooms may be susceptible to rapid termination by bacteriophage infection in a kill-the-winner event. This hypothesis needs to be tested by either isolating candidate phages on *Anabaena* sp. WA102 and tracking both phage and host populations over time with quantitative molecular methods such as qPCR or with culture-independent metagenomic

methods that can track relative abundances of both phage and host.

Finally, a new temperate cyanobacterial phage and freshwater *Cyanobium* host system has been established. *Cyanobium* sp. LC18 is a newly isolated species of freshwater picocyanobacteria with a high-quality draft genome. Although this is only one genome, it is an important contribution, considering that there are only 11 freshwater picocyanobacterial genomes of 183 total picocyanobacterial genomes deposited in Genbank and only 3 *Cyanobium* genomes (as of February 2016). Additional picocyanobacterial genome sequences are essential for resolving the early cyanobacterial phylogenetic lineage and correlating the evolution of cyanobacterial physiological innovations with biogeochemical events on ancient Earth [7]. The temperate cyanobacterial siphovirus C-CRS01 was isolated on *Cyanobium* sp. LC18, its genome was sequenced, and its genome was shown to integrate into an ORF on the *Cyanobium* sp. LC18 genome. As evidence mounts that freshwater cyanobacterial blooms are susceptible to bacteriophage lysis [111, 49], establishing cyanobacterial host/phage model systems is important for being able to track host/phage population quantities and dynamics in freshwater lakes. The *Cyanobium* sp. LC18/C-CRS01 system offers specific DNA sequences that can be tracked with qPCR to quantify their population dynamics during bloom lifecycles in Upper Klamath Lake.

This body of work includes multiple incremental advances that will help to move the field of limnology forward, particularly as focus shifts to the biotic drivers of bloom emergence and collapse and methodology relies more heavily on high-throughput -omics methods.

Chapter 7 Contributions from authors

7.1 Chapter 3: Structural and Functional Analysis of the Closed Genome of the Recently Isolated Toxic *Anabaena* sp. WA102

Nathan M. Brown and Theo W. Dreher conceived and designed the experimental plan, with input from F. Joan Hardy and Ryan S. Mueller, and wrote the manuscript with input from other authors. Nathan M. Brown conducted most of the experiments. Ryan S. Mueller provided bioinformatic advice and analysis. Jonathan W. Shepardson assisted with experiments. Zachary C. Landry conducted the phylogenomic analysis; Claudia S. Maier and Jeffrey T. Morré conducted the mass spectrometry analysis.

7.2 Chapter 4: Identification of the major anatoxin-a producing cyanobacterium in Anderson Lake, its dynamics, and its distribution in the Puget Sound region

Nathan M. Brown and Theo W. Dreher conceived and designed the experimental plan, with input from F. Joan Hardy and Ryan S. Mueller, and wrote the manuscript with input from other authors. Nathan M. Brown conducted most of the experiments. Claudia S. Maier and Souyun Ahn conducted the mass spec-

trometry analysis.

7.3 Chapter 5: The genome of a novel freshwater picocyanobacterium, *Cyanobium* sp. LC18, and lysogenization by of one of its temperate phages, C-CRS01

Nathan M. Brown and Theo W. Dreher conceived and designed the experimental plan and wrote the manuscript. Nathan M. Brown conducted all experiment and carried out most analysis. Zachary C. Landry conducted the phylogenomic analysis. Jeff H. Chang sequenced and assembled the C-CRS01 genome.

Bibliography

[1] Mads Albertsen, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538, 2013.

[2] R Philip Anderson and John R Roth. Tandem genetic duplications in phage and bacteria. *Annual Reviews in Microbiology*, 31(1):473–505, 1977.

[3] Gabriele Beckers, Anne K Bendt, Reinhard Krämer, and Andreas Burkovski. 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, 2004.

[4] SD Bentley, KF Chater, A-M Cerdeno-Tarraga, GL Challis, NR Thomson, KD James, DE Harris, MA Quail, H Kieser, D Harper, et al. Complete genome sequence of the model actinomycete streptomyces coelicolor a3 (2). *Nature*, 417(6885):141–147, 2002.

[5] Xin Bi and Leroy F Liu. DNA rearrangement mediated by inverted repeats. *Proceedings of the National Academy of Sciences*, 93(2):819–823, 1996.

[6] Roger Bivand and Nicholas Lewin-Koh. maptools: Tools for reading and handling spatial objects. *R package version 0.8–27*, 2013.

[7] CE Blank and P SÁNCHEZ-BARACALDO. Timing of morphological and ecological innovations in the cyanobacteria–a key to understanding the rise in atmospheric oxygen. *Geobiology*, 8(1):1–23, 2010.

[8] Christopher JS Bolch, Susan I Blackburn, Brett A Neilan, and Peter M Grewe. Genetic characterization of strains of cyanobacteria using pcr-rflp of the cpcba intergenic spacer and flanking regions1. *Journal of phycology*, 32(3):445–451, 1996.

[9] G Bratbak, JK Egge, and M Heldal. Viral mortality of the marine alga emiliania huxleyi (haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series*, 1993.

[10] Justin D Brookes, Cayelan C Carey, et al. Resilience to blooms. *Science*, 334(6052):46–47, 2011.

[11] Nathan M Brown, Ryan S Mueller, Jonathan W Shepardson, Zachary C Landry, Claudia S Maier, Jeffrey T Morré, F Joan Hardy, and Theo W Dreher. Structural and functional analysis of the closed genome of the recently isolated toxic *Anabaena* sp. wa102. *BMC Genomics*, 2016.

[12] Sabrina Cadel-Six, Isabelle Iteman, Caroline Peyraud-Thomas, Stéphane Mann, Olivier Ploux, and Annick Méjean. Identification of a polyketide synthase coding sequence specific for anatoxin-a-producing *Oscillatoria* cyanobacteria. *Applied and Environmental Microbiology*, 75(14):4909–4912, 2009.

[13] Cristiana Callieri. Single cells and microcolonies of freshwater picocyanobacteria: a common ecology. *Journal of Limnology*, 69(2):257–277, 2010.

[14] Cristiana Callieri, Manuela Coci, Gianluca Corno, Miroslav Macek, Beatriz Modenutti, Esteban Balseiro, and Roberto Bertoni. Phylogenetic diversity of nonmarine picocyanobacteria. *FEMS microbiology ecology*, 85(2):293–301, 2013.

[15] Alexandra Calteau, David P Fewer, Amel Latifi, Thérèse Coursin, Thierry Laurent, Jouni Jokela, Cheryl A Kerfeld, Kaarina Sivonen, Jörn Piel, and Muriel Gugger. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in cyanobacteria. *BMC Genomics*, 15(1):977, 2014.

[16] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[17] Wayne W Carmichael. Health effects of toxin-producing cyanobacteria:the cyanohabs. *Human and Ecological Risk Assessment: An International Journal*, 7(5):1393–1407, 2001.

[18] Wayne W Carmichael, David F Biggs, and Paul R Gorham. Toxicology and pharmacological action of *Anabaena flos-aquae* toxin. *Science*, 187(4176):542–544, 1975.

[19] Wayne W Carmichael, David F Biggs, and Marge A Peterson. Pharmacology of anatoxin-a, produced by the freshwater cyanophyte *Anabaena flos-aquae* NRC-44-1. *Toxicon*, 17(3):229–236, 1979.

[20] JW Chase and CC Richardson. *Escherichia coli* mutants deficient in exonuclease VII. *Journal of Bacteriology*, 129(2):934–947, 1977.

[21] Kai Cheng, Yijun Zhao, Xiuli Du, Yaran Zhang, Shubin Lan, and Zhengli Shi. Solar radiation-driven decay of cyanophage infectivity, and photoreactivation of the cyanophage by host cyanobacteria. *Aquatic microbial ecology*, 48(1):13, 2007.

[22] Kathryn L Cottingham, Holly A Ewing, Meredith L Greer, Cayelan C Carey, and Kathleen C Weathers. Cyanobacteria as biological drivers of lake nitrogen and phosphorus cycling. *Ecosphere*, 6(1):1–19, 2015.

[23] Tal Dagan, Mayo Roettger, Karina Stucken, Giddy Landan, Robin Koch, Peter Major, Sven B Gould, Vadim V Goremykin, Rosmarie Rippka, Nicole Tandeau de Marsac, et al. Genomes of stigonematalean cyanobacteria (subsection v) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome biology and evolution*, 5(1):31–44, 2013.

[24] Roberto Danovaro, Antonio DellAnno, Cinzia Corinaldesi, Mirko Magagnini, Rachel Noble, Christian Tamburini, and Markus Weinbauer. Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature*, 454(7208):1084–1087, 2008.

[25] Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165, 2011.

[26] Simanti Datta, Nina Costantino, Xiaomei Zhou, et al. Identification and analysis of recombineering functions from gram-negative and gram-positive bacteria and their phages. *Proceedings of the National Academy of Sciences*, 105(5):1626–1631, 2008.

[27] Daniel De Palmenaer, Patricia Siguier, and Jacques Mahillon. IS4 family goes genomic. *BMC evolutionary biology*, 8(1):18, 2008.

[28] Daniel E Deatherage and Jeffrey E Barrick. 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, 2014.

[29] C Lisa Dent, Graeme S Cumming, and Stephen R Carpenter. Multiple states in river and lake ecosystems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1421):635–645, 2002.

[30] Céline Desnous, Dominique Guillaume, and Pascale Clivio. Spore photoproduct: a key to bacterial eternal life. *Chemical reviews*, 110(3):1213–1232, 2009.

[31] Bhavjinder K Dhillon, Terry A Chiu, Matthew R Laird, Morgan GI Langille, and Fiona SL Brinkman. IslandViewer update: improved genomic island discovery and visualization. *Nucleic Acids Research*, 41(W1):W129–W132, 2013.

[32] Elke Dittmann, Muriel Gugger, Kaarina Sivonen, and David P Fewer. Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. *Trends in microbiology*, 23(10):642–652, 2015.

[33] Evan Dobrowski and Michael Dawson. Jefferson county toxic cyanobacteria project. Technical Report G1400003, Jefferson County Public Health, Port Townsend, Washington, June 2015.

[34] Alexis Dufresne, Martin Ostrowski, David J Scanlan, Laurence Garczarek, Sophie Mazard, Brian P Palenik, Ian T Paulsen, N Tandeau De Marsac, Patrick Wincker, Carole Dossat, et al. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol*, 9(5):R90, 2008.

[35] Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[36] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[37] Eric Eichhorn, Jan R Van Der Ploeg, and Thomas Leisinger. Deletion analysis of the *Escherichia coli* taurine and alkanesulfonate transport systems. *Journal of Bacteriology*, 182(10):2687–2695, 2000.

[38] Jonathan A Eisen, John F Heidelberg, Owen White, Steven L Salzberg, et al. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol*, 1(6):1–0011, 2000.

[39] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[40] Ron Feiner, Tal Argov, Lev Rabinovich, Nadejda Sigal, Ilya Borovok, and Anat A Herskovits. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nature Reviews Microbiology*, 13(10):641–650, 2015.

[41] Ralf Flachmann, Genhai Zhu, Richard G Jensen, and Hans J Bohnert. Mutations in the small subunit of ribulose-1, 5-bisphosphate carboxylase/oxygenase increase the formation of the misfire product xylulose-1, 5-bisphosphate. *Plant physiology*, 114(1):131–136, 1997.

[42] Julia M Flynn, Igor Levchenko, Meredith Seidel, Sue H Wickner, Robert T Sauer, and Tania A Baker. Overlapping recognition determinants within the ssra degradation tag allow modulation of proteolysis. *Proceedings of the National Academy of Sciences*, 98(19):10584–10589, 2001.

[43] Claudine Franche. Isolation and characterization of a temperate cyanophage for a tropical anabaena strain. *Archives of microbiology*, 148(3):172–177, 1987.

[44] Michael Y Galperin, Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, page gku1223, 2014.

[45] Miroslav Gantar and Zorica Svirčev. Microalgae and cyanobacteria: food for thought1. *Journal of Phycology*, 44(2):260–268, 2008.

[46] E-Bin Gao, Jian-Fang Gui, and Qi-Ya Zhang. A novel cyanophage with a cyanobacterial nonbleaching protein a gene in the genome. *Journal of virology*, 86(1):236–245, 2012.

[47] Stephen J Giovannoni, H James Tripp, Scott Givan, Mircea Podar, Kevin L Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H Richardson,

Michiel Noordewier, et al. Genome streamlining in a cosmopolitan oceanic bacterium. *science*, 309(5738):1242–1245, 2005.

[48] James W Golden and Ho-Sung Yoon. Heterocyst development in *Anabaena*. *Current Opinion in Microbiology*, 6(6):557–563, 2003.

[49] Herman J Gons, Jeannine Ebert, Hans L Hoogveld, Linda van den Hove, Roel Pel, Wijnand Takkenberg, and Conrad J Woldringh. Observations on cyanobacterial population collapse in eutrophic lake water. *Antonie Van Leeuwenhoek*, 81(1-4):319–326, 2002.

[50] Johan Goris, Konstantinos T Konstantinidis, Joel A Klappenbach, Tom Coenye, Peter Vandamme, and James M Tiedje. Dna–dna hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, 57(1):81–91, 2007.

[51] I Gregor, J Dröge, M Schirmer, C Quince, and AC McHardy. Phylopythias+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *arXiv preprint arXiv:1406.7123*, 2014.

[52] Marco Griese, Christian Lange, and Jörg Soppa. Ploidy in cyanobacteria. *FEMS microbiology letters*, 323(2):124–131, 2011.

[53] Andrei Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic acids research*, 26(10):2286–2290, 1998.

[54] Andrian Gutu, Richard M Alvey, Sami Bashour, Daniel Zingg, and David M Kehoe. Sulfate-driven elemental sparing is regulated at the transcriptional and posttranscriptional levels in a filamentous cyanobacterium. *Journal of bacteriology*, 193(6):1449–1460, 2011.

[55] Ken-ichi Harada. Production of secondary metabolites by freshwater cyanobacteria. *Chemical and Pharmaceutical bulletin*, 52(8):889–899, 2004.

[56] Jef Huisman, Hans CP Matthijs, and Petra M Visser. *Harmful cyanobacteria*, volume 3. Springer Science & Business Media, 2006.

[57] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1, 2010.

[58] Bas W Ibelings, Rob Portielje, Eddy HRR Lammens, Ruurd Noordhuis, Marcel S van den Berg, Willemien Joosse, and Marie Louise Meijer. Resilience of alternative stable states during the recovery of shallow lakes from eutrophication: Lake veluwe as a case study. *Ecosystems*, 10(1):4–16, 2007.

[59] Stéphan Jacquet, Mikal Heldal, Debora Iglesias-Rodriguez, Aud Larsen, William Wilson, and Gunnar Bratbak. Flow cytometric analysis of an emiliana huxleyi bloom terminated by viral infection. *Aquatic Microbial Ecology*, 27(2):111–124, 2002.

[60] Thibaut Jombart, François Balloux, and Stéphane Dray. Adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics*, 26(15):1907–1909, 2010.

[61] Adam C Jones, Liangcai Gu, Carla M Sorrels, David H Sherman, and William H Gerwick. New tricks from ancient algae: natural products biosynthesis in marine cyanobacteria. *Current opinion in chemical biology*, 13(2):216–223, 2009.

[62] Thomas Junier and Evgeny M Zdobnov. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics*, 26(13):1669–1670, 2010.

[63] BP Jupp and DHN Spence. Limitations on macrophytes in a eutrophic lake, loch leven: I. effects of phytoplankton. *The Journal of Ecology*, pages 175–186, 1977.

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

[65] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[66] Takakazu Kaneko, Nobuyoshi Nakajima, Shinobu Okamoto, Iwane Suzuki, Yuuhiko Tanabe, Masanori Tamaoki, Yasukazu Nakamura, Fumie Kasai, Akiko Watanabe, Kumiko Kawashima, et al. Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research*, 14(6):247–256, 2007.

[67] Nadav Kashtan, Sara E Roggensack, Sébastien Rodrigue, Jessie W Thompson, Steven J Biller, Allison Coe, Huiming Ding, Pekka Marttinen, Rex R Malmstrom, Roman Stocker, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science*, 344(6182):416–420, 2014.

[68] Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael JE Sternberg. The phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–858, 2015.

[69] Fyodor A Kondrashov, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. Selection in the evolution of gene duplications. *Genome Biol*, 3(2):8–1, 2002.

[70] Sergey Koren, Gregory P Harhay, TP Smith, James L Bono, Dayna M Harhay, Scott D Mcvey, Diana Radune, Nicholas H Bergman, and Adam M Phillippy. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol*, 14(9):R101, 2013.

[71] Sergey Koren and Adam M Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, 23:110–120, 2015.

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

[73] James J Kozich, Sarah L Westcott, Nielson T Baxter, Sarah K Highlander, and Patrick D Schloss. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Applied and environmental microbiology*, 79(17):5112–5120, 2013.

[74] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[75] Michael Lawrence, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8):e1003118, 2013.

[76] Pedro N Leão, Margarida Costa, Vitor Ramos, Alban R Pereira, Virgínia C Fernandes, Valentina F Domingues, William H Gerwick, Vitor M Vasconcelos, and Rosário Martins. Antitumor activity of hierridin b, a cyanobacterial secondary metabolite found in both filamentous and unicellular marine strains. *PloS one*, 8(7):e69562, 2013.

[77] Lee H Lee, Doris Lui, Patricia J Platner, Shi-Fang Hsu, Tin-Chun Chu, John J Gaynor, Quinn C Vega, and Bonnie K Lustigman. Induction of temperate cyanophage as-1 by heavy metal-copper. *BMC microbiology*, 6(1):1, 2006.

[78] Ivica Letunic, Richard R Copley, Steffen Schmidt, Francesca D Ciccarelli, Tobias Doerks, Jörg Schultz, Chris P Ponting, and Peer Bork. SMART 4.0: towards genomic data integration. *Nucleic Acids Research*, 32(suppl 1):D142–D144, 2004.

[79] Ivica Letunic, Takuji Yamada, Minoru Kanehisa, and Peer Bork. ipath: interactive exploration of biochemical pathways and networks. *Trends in biochemical sciences*, 33(3):101–103, 2008.

[80] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[81] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[82] X Li, Theo W Dreher, and R Li. An overview of diversity, occurrence and toxin production of bloom-forming *Dolichospermum* (*Anabaena*) species. *Harmful Algae*, in press, 2015.

[83] Xia Liu, Xiaohua Lu, and Yuwei Chen. The effects of temperature and nutrient ratios on microcystis blooms in lake taihu, china: an 11-year investigation. *Harmful Algae*, 10(3):337–343, 2011.

[84] Susan T Lovett and Vladimir V Feschenko. Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. *Proceedings of the National Academy of Sciences*, 93(14):7120–7124, 1996.

[85] Luo, Weijun, Brouwer, and Cory. Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[86] Nicholas H Mann, Martha RJ Clokie, Andrew Millard, Annabel Cook, William H Wilson, Peter J Wheatley, Andrey Letarov, and HM Krisch. The genome of s-pm2, a photosynthetic t4-type bacteriophage that infects marine synechococcus strains. *Journal of bacteriology*, 187(9):3188–3200, 2005.

[87] Nicholas H Mann, Annabel Cook, Andrew Millard, Shaun Bailey, and Martha Clokie. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*, 424(6950):741–741, 2003.

[88] Staci Matlock. Toxic algae blamed for elk deaths in northeastern new mexico. *Santa Fe New Mexican*, October 2013.

[89] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.

[90] Marnix H Medema, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(suppl 2):W339–W346, 2011.

[91] Marnix H Medema, Peter Cimermancic, Andrej Sali, Eriko Takano, and Michael A Fischbach. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput Biol*, 10(12):e1004016, 2014.

[92] Annick Méjean, Guillaume Paci, Valérie Gautier, and Olivier Ploux. Biosynthesis of anatoxin-a and analogues (anatoxins) in cyanobacteria. *Toxicon*, 91:15–22, 2014.

[93] Daniel R Mende, Alison S Waller, Shinichi Sunagawa, Aino I Järvelin, Michelle M Chan, Manimozhiyan Arumugam, Jeroen Raes, and Peer Bork. Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one*, 7(2):e31386, 2012.

[94] Melinda L Micallef, Paul M DAgostino, Deepti Sharma, Rajesh Viswanathan, and Michelle C Moffitt. Genome mining for natural product biosynthetic gene clusters in the subsection v cyanobacteria. *BMC genomics*, 16(1):1, 2015.

[95] Markus Mikulic. *Knock-out mutants of respiratory terminal oxidases in the cyanobacterium* Anabaena *sp. strain PCC 7120.* PhD thesis, Universitt Wien, 2013.

[96] Joshua A Mosberg, Marc J Lajoie, and George M Church. Lambda red recombineering in escherichia coli occurs through a fully single-stranded intermediate. *Genetics*, 186(3):791–799, 2010.

[97] Richard Münch, Karsten Hiller, Andreas Grote, Maurice Scheer, Johannes Klein, Max Schobert, and Dieter Jahn. Virtual footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, 21(22):4187–4189, 2005.

[98] Ginji Nakamura, Shigeko Kimura, Yoshihiko Sako, and Takashi Yoshida. Genetic diversity of microcystis cyanophages in two different freshwater environments. *Archives of microbiology*, 196(6):401–409, 2014.

[99] Joakim Näsvall, Lei Sun, John R Roth, and Dan I Andersson. Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338(6105):384–387, 2012.

[100] Rie Nishiwaki-Matsushima, Tetsuya Ohta, Shinji Nishiwaki, Masami Suganuma, Kiyomi Kohyama, Takatoshi Ishikawa, Wayne W Carmichael, and Hirota Fujiki. Liver tumor promotion by the cyanobacterial cyclic peptide toxin microcystin-lr. *Journal of cancer research and clinical oncology*, 118(6):420–424, 1992.

[101] Rachel T Noble and Jed A Fuhrman. Use of sybr green i for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology*, 14(2):113–118, 1998.

[102] Pavel S Novichkov, Yuri I Wolf, Inna Dubchak, and Eugene V Koonin. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *Journal of bacteriology*, 191(1):65–73, 2009.

[103] Nancy Obeng, Akbar Adjie Pratama, and Jan Dirk van Elsas. The significance of mutualistic phages for bacterial ecology and evolution. *Trends in Microbiology*, 2016.

[104] Timothy G Otten, Joseph R Crosswell, Sam Mackey, and Theo W Dreher. Application of molecular tools for microbial source tracking and public health risk assessment of a microcystis bloom traversing 300km of the klamath river. *Harmful algae*, 46:71–81, 2015.

[105] Etana Padan, Moshe Shilo, and Amos B Oppenheim. Lysogeny of the blue-green alga plectonema boryanum by lpp2-spi cyanophage. *Virology*, 47(2):525–526, 1972.

[106] Hans W Paerl, Rolland S Fulton, Pia H Moisander, and Julianne Dyble. Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *The Scientific World Journal*, 1:76–113, 2001.

[107] Hans W Paerl, Jef Huisman, et al. Blooms like it hot. *SCIENCE-NEW YORK THEN WASHINGTON-*, 320(5872):57, 2008.

[108] Hans W Paerl and Timothy G Otten. Duelling cyanohabs: unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-n2-fixing harmful cyanobacteria. *Environmental microbiology*, 2015.

[109] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[110] F Partensky, Jean Blanchot, and D Vaulot. Differential distribution and ecology of prochlorococcus and synechococcus in oceanic waters: a review. *BULLETIN-INSTITUT OCEANOGRAPHIQUE MONACO-NUMERO SPECIAL-*, pages 457–476, 1999.

[111] Peter Peduzzi, Martin Gruber, Michael Gruber, and Michael Schagerl. The virus's tooth: cyanophages affect an african flamingo population in a bottom-up cascade. *The ISME journal*, 8(6):1346, 2014.

[112] Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012.

[113] Jasper John Lobl Pengelly. *Molecular Characterisation of Membrane Transporters Associated with Saxitoxin Biosynthesis in Cyanobacteria: A Dissertation Submitted in Partial Fulfilment of the Requirements for the Award of Doctor of Philosophy (Ph. D).* PhD thesis, UNSW, School of Biotechnology and Biomolecular Sciences, 2008.

[114] Adam M Phillippy, Michael C Schatz, Mihai Pop, et al. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol*, 9(3):R55, 2008.

[115] Gur Pines, Emily F Freed, James D Winkler, and Ryan T Gill. Bacterial recombineering: genome engineering via phage-based homologous recombination. *ACS synthetic biology*, 4(11):1176–1185, 2015.

[116] Alkes L Price, Neil C Jones, and Pavel A Pevzner. *De novo* identification of repeat families in large genomes. *Bioinformatics*, 21(suppl 1):i351–i358, 2005.

[117] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

[118] Pere Puigbò, Alexander E Lobkovsky, David M Kristensen, Yuri I Wolf, and Eugene V Koonin. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC biology*, 12(1):66, 2014.

[119] Arumugham Raghunathan, Harley R Ferguson, Carole J Bornarth, Wanmin Song, Mark Driscoll, and Roger S Lasken. Genomic dna amplification from a single bacterium. *Applied and environmental microbiology*, 71(6):3342–3347, 2005.

[120] Pirjo Rajaniemi, Pavel Hrouzek, Klara Kaštovska, Raphael Willame, Anne Rantala, Lucien Hoffmann, Jiří Komárek, and Kaarina Sivonen. Phylogenetic and morphological evaluation of the genera anabaena, aphanizomenon, trichormus and nostoc (nostocales, cyanobacteria). *International Journal of Systematic and Evolutionary Microbiology*, 55(1):11–26, 2005.

[121] Anne Rantala-Ylinen, Suvi Känä, Hao Wang, Leo Rouhiainen, Matti Wahlsten, Ermanno Rizzi, Katri Berg, Muriel Gugger, and Kaarina Sivonen. Anatoxin-a synthetase gene cluster of the cyanobacterium *Anabaena* sp. strain 37 and molecular methods to detect potential producers. *Applied and environmental microbiology*, 77(20):7271–7278, 2011.

[122] Andrew B Reams and Ellen L Neidle. Selection for gene clustering by tandem duplication. *Annu. Rev. Microbiol.*, 58:119–142, 2004.

[123] Andrew B Reams and John R Roth. Mechanisms of gene duplication and amplification. *Cold Spring Harbor perspectives in biology*, 7(2), 2015.

[124] Wolf-Dieter Reiter, Peter Palm, and Siobhan Yeats. Transfer rna genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic acids research*, 17(5):1907–1914, 1989.

[125] Rodrigo Reyes-Lamothe, Emilien Nicolas, and David J Sherratt. Chromosome replication and segregation in bacteria. *Annual review of genetics*, 46:121–143, 2012.

[126] Paul A Rochelle et al. *Environmental Molecular Microbiology: Protocols and Applications.* Horizon Scientific Press, 2001.

[127] HD Rodger, T Turnbull, C Edwards, and GA Codd. Cyanobacterial (blue-green algal) bloom associated pathology in brown trout, salmo trutta l., in loch leven, scotland. *Journal of Fish Diseases*, 17(2):177–181, 1994.

[128] Luis M Rodriguez-R and Konstantinos T Konstantinidis. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, 30(5):629–635, 2014.

[129] P Sanchez-Baracaldo, PK Hayes, and CE Blank. Morphological and habitat evolution in the cyanobacteria using a compartmentalization approach. *Geobiology*, 3(3):145–165, 2005.

[130] Patricia Sanchez-Baracaldo, Barbara A Handley, and Paul K Hayes. Picocyanobacterial community structure of freshwater lakes and the baltic sea revealed by phylogenetic analyses and clade-specific quantitative pcr. *Microbiology*, 154(11):3347–3357, 2008.

[131] Ueli Schibler and Felipe Sierra. Alternative promoters in developmental gene expression. *Annual review of genetics*, 21(1):237–257, 1987.

[132] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[133] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, page btu153, 2014.

[134] Itai Sharon, Michael J Morowitz, Brian C Thomas, Elizabeth K Costello, David A Relman, and Jillian F Banfield. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome research*, 23(1):111–120, 2013.

[135] Patrick M Shih, Dongying Wu, Amel Latifi, Seth D Axen, David P Fewer, Emmanuel Talla, Alexandra Calteau, Fei Cai, Nicole Tandeau de Marsac, Rosmarie Rippka, et al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences*, 110(3):1053–1058, 2013.

[136] Patricia Siguier, Jonathan Filée, and Michael Chandler. Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology*, 9(5):526–531, 2006.

[137] Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl 1):D32–D36, 2006.

[138] Shailendra P Singh and Beronda L Montgomery. Determining cell shape: adaptive regulation of cyanobacterial cellular differentiation and morphology. *Trends in microbiology*, 19(6):278–285, 2011.

[139] Tony A Slieman, Roberto Rebeil, and Wayne L Nicholson. Spore photoproduct (sp) lyase from bacillus subtilisspecifically binds to and cleaves sp (5-thyminyl-5, 6-dihydrothymine) but not cyclobutane pyrimidine dimers in uv-irradiated dna. *Journal of bacteriology*, 182(22):6412–6417, 2000.

[140] George Sorensen, Andrea C Baker, Matthew J Hall, Colin B Munn, and Declan C Schroeder. Novel virus dynamics in an emiliania huxleyi bloom. *Journal of plankton research*, 31(7):787–791, 2009.

[141] Brian G Spratt, William P Hanage, and Edward J Feil. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current opinion in microbiology*, 4(5):602–606, 2001.

[142] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[143] Bärbel Stecher, Lisa Maier, and Wolf-Dietrich Hardt. 'blooming'in the gut: how dysbiosis might contribute to pathogen evolution. *Nature Reviews Microbiology*, 11(4):277–284, 2013.

[144] Lisa M Steenhauer, Peter C Pollard, Corina PD Brussaard, and Christin Säwström. Lysogenic infection in sub-tropical freshwater cyanobacteria cultures and natural blooms. *Marine and Freshwater Research*, 65(7):624–632, 2014.

[145] Alain Stintzi, Carmen Barnes, Jide Xu, and Kenneth N Raymond. Microbial iron transport via a siderophore shuttle: a membrane ion transport paradigm. *Proceedings of the National Academy of Sciences*, 97(20):10691–10696, 2000.

[146] John Stockner, Cristiana Callieri, and Gertrud Cronberg. Picoplankton and other non-bloom-forming cyanobacteria in lakes. In *The ecology of cyanobacteria*, pages 195–231. Springer, 2000.

[147] John G Stockner and Ken S Shortreed. Response of *Anabaena* and *Synechococcus* to manipulation of nitrogen: phosphorus ratios in a lake fertilization experiment. *Limnology and Oceanography*, 33(6):1348–1361, 1988.

[148] David J Studholme, Selena G Ibanez, Daniel MacLean, Jeffery L Dangl, Jeff H Chang, and John P Rathjen. A draft genome sequence and functional screen reveals the repertoire of type iii secreted proteins of pseudomonas syringae pathovar tabaci 11528. *Bmc Genomics*, 10(1):395, 2009.

[149] Assaf Sukenik, Ora Hadas, Aaron Kaplan, and Antonio Quesada. Invasion of nostocales (cyanobacteria) to subtropical and temperate freshwater lakes–physiological, regional, and global driving forces. *Frontiers in Microbiology*, 3:86, 2012.

[150] Matthew B Sullivan, Maureen L Coleman, Peter Weigele, Forest Rohwer, and Sallie W Chisholm. Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*, 3(5):e144, 2005.

[151] Mikita Suyama, David Torrents, and Peer Bork. Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34(suppl 2):W609–W612, 2006.

[152] RV Swanson, R De Lorimier, and AN Glazer. Genes encoding the phycobilisome rod substructure are clustered on the *Anabaena* chromosome: characterization of the phycoerythrocyanin operon. *Journal of bacteriology*, 174(8):2640–2647, 1992.

[153] Lik Tong Tan. Filamentous tropical marine cyanobacteria: a rich source of natural products for anticancer drug discovery. *Journal of applied phycology*, 22(5):659–676, 2010.

[154] Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology*, 6(9):938–947, 2004.

[155] Luke R Thompson, Qinglu Zeng, Libusha Kelly, Katherine H Huang, Alexander U Singer, JoAnne Stubbe, and Sallie W Chisholm. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences*, 108(39):E757–E764, 2011.

[156] Claire S Ting, Gabrielle Rocap, Jonathan King, and Sallie W Chisholm. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends in microbiology*, 10(3):134–142, 2002.

[157] Neha J Varghese, Supratim Mukherjee, Natalia Ivanova, Konstantinos T Konstantinidis, Kostas Mavrommatis, Nikos C Kyrpides, and Amrita Pati. Microbial species delineation using whole genome sequences. *Nucleic acids research*, page gkv657, 2015.

[158] Pirjo Wacklin, Lucien Hoffmann, and Jiří Komárek. Nomenclatural validation of the genetically revised cyanobacterial genus *Dolichospermum* (Ralfs ex Bornet et Flahault) comb. nova. *Fottea*, 9(1):59–64, 2009.

[159] AE Walsby. Gas vesicles. *Microbiological reviews*, 58(1):94, 1994.

[160] Hao Wang, David P Fewer, Liisa Holm, Leo Rouhiainen, and Kaarina Sivonen. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences*, 111(25):9259–9264, 2014.

[161] Hao Wang, Kaarina Sivonen, Leo Rouhiainen, David P Fewer, Christina Lyra, Anne Rantala-Ylinen, Johanna Vestola, Jouni Jokela, Kaisa Rantasärkkä, Zhijie Li, et al. Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. *BMC Genomics*, 13(1):613, 2012.

[162] Steven W Wilhelm and Charles G Trick. Iron-limited growth of cyanobacteria: Multiple siderophore production is a common response. *Limnology and Oceanography*, 39(8):1979–1984, 1994.

[163] Jason Nicholas Woodhouse, Andrew Stephen Kinsela, Richard Nicholas Collins, Lee Chester Bowling, Gordon L Honeyman, Jon K Holliday, and Brett Anthony Neilan. Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. *The ISME journal*, 2015.

[164] Kelly C Wrighton, Brian C Thomas, Itai Sharon, Christopher S Miller, Cindy J Castelle, Nathan C VerBerkmoes, Michael J Wilkins, Robert L Hettich, Mary S Lipton, Kenneth H Williams, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102):1661–1665, 2012.

[165] Martin Wu, Sourav Chatterji, and Jonathan A Eisen. Accounting for alignment uncertainty in phylogenomics. *PLoS One*, 7(1):e30288, 2012.

[166] Haruyo Yamaguchi, Shigekatsu Suzuki, Yuuhiko Tanabe, Yasunori Osana, Yohei Shimura, Ken-ichiro Ishida, and Masanobu Kawachi. Complete genome sequence of *Microcystis aeruginosa* NIES-2549, a bloom-forming cyanobacterium from lake kasumigaura, japan. *Genome Announcements*, 3(3):e00551–15, 2015.

[167] Yoshimasa Yamamoto and Hiroyuki Nakahara. The formation and degradation of cyanobacterium aphanizomenon flos-aquae blooms: the importance of ph, water temperature, and day length. *Limnology*, 6(1):1–6, 2005.

[168] Kuo-Chen Yeh, Shu-Hsing Wu, John T Murphy, and J Clark Lagarias. A cyanobacterial phytochrome two-component light sensory system. *Science*, 277(5331):1505–1508, 1997.

[169] Takashi Yoshida, Keizo Nagasaki, Yukari Takashima, Yoko Shirai, Yuji Tomaru, Yoshitake Takao, Shigetaka Sakamoto, Shingo Hiroishi, and Hiroyuki Ogata. Ma-lmm01 infecting toxic microcystis aeruginosa illuminates diverse cyanophage genome strategies. *Journal of bacteriology*, 190(5):1762–1772, 2008.

[170] Shizue Yoshihara and Masahiko Ikeuchi. Phototactic motility in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. *Photochemical & Photobiological Sciences*, 3(6):512–518, 2004.

[171] Kevin D Young. Bacterial shape: two-dimensional questions and possibilities. *Annual review of microbiology*, 64:223, 2010.

[172] Eliška Zapomělová, Klára Řeháková, Jitka Jezberová, and Jaroslava Komárková. Polyphasic characterization of eight planktonic anabaena strains (cyanobacteria) with reference to the variability of 61 anabaena populations observed in the field. *Hydrobiologia*, 639(1):99–113, 2010.

[173] You Zhou, Yongjie Liang, Karlene H Lynch, Jonathan J Dennis, and David S Wishart. PHAST: a fast phage search tool. *Nucleic Acids Research*, page gkr485, 2011.