Investigating Cyanobacterial Communities Across Canada and Characterizing the Seasonality of Cyanobacterial Toxin and Nitrogen Metabolism Genes in the Turkey Lakes Watershed Using Metagenomics

by

Hadeel Abed

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology (Water)

Waterloo, Ontario, Canada, 2023

© Hadeel Abed 2023

Author's Declaration

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

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

Statement of Contributions

The samples obtained for sequencing were obtained by several *for*Water Network highly qualified personnel (HQP). The following is a list of the personnel responsible for obtaining environmental samples and/or sequencing eDNA used for this thesis from the Müller Lab at the University of Waterloo: Ellen Cameron obtained and sequenced Boreal Shield (TLW) samples, Timothy Shardlow sequenced Atlantic Maritime, Pacific Maritime, and Montane Cordillera (Glenmore Reservoir) samples, Ghadeer Aljoudi sequenced Montane Cordillera (Glenmore Reservoir) samples, and Arnold Kwok obtained and sequenced Boreal Plains (URSA) samples. For further information regarding the origins of environmental samples, please refer to the corresponding theses available on UWSpace. The V4 region of the 16S rRNA gene from these environmental samples were sequenced by Dr. Trevor Charles, Dr. Michael Lynch, and Dr. Jiujun Cheng of Metagenom Bio Inc. (Waterloo, Ontario).

Pauline Wang and colleagues of the Centre for the Analysis of Genome Evolution and Function (CAGEF) performed whole genome shotgun sequencing for environmental samples from Turkey Lakes Watershed. Raw data was assembled via Metagenome ATLAS with the assistance of Cody Collis (MSc) and Dr. Briallen Lobb (PhD).

Abstract

The long evolutionary history of gram-negative cyanobacteria has enabled them to tolerate extreme conditions, succeed in numerous habitats, and exploit special nutrients, all while resisting harmful anthropogenic and climate change exacerbated influences. Increases in harmful cyanobacterial blooms (cHABs) pose considerable challenges to the provision of safe drinking water globally. Cyanobacterial blooms are most often recorded in eutrophic lakes where they have sufficient nutrients for proliferation, however, this has led to the misunderstanding that they do not pose risks in oligotrophic lakes because of the lower nutrient availability needed to form harmful blooms. Yet, the absence of visual surface blooms does not correlate to the absence of cyanobacteria as they can be abundant at different depths within the water column. Cyanobacteria may release toxic metabolites into water sources and if left untreated, these toxins threaten drinking water security, including those from oligotrophic lakes, hence, the high demand for research in drinking water treatability sectors and sustainable forested watershed management.

The distribution of cyanobacterial communities across Canada was investigated in this study (Chapter 2); samples were collected by various *for*Water HQP from five major forested ecozones: Atlantic and Pacific Maritime, Boreal Shield and Plains, and the Montane Cordillera. Unique cyanobacterial ASVs obtained from V4-16S rRNA region gene sequencing of all samples were phylogenetically analysed using Maximum Likelihood tree models to observe cyanobacterial taxonomy and biogeography within the various research platforms. Understanding the composition of cyanobacterial communities is a valuable tool in the proactive monitoring of threats to water systems because by observing microbial communities, it can help identify major toxin producing species and illustrate the overall quality of the source water. This research demonstrated that cyanobacteria across Canada are highly diverse and spread across the country, cyanobacteria are unaffected by geographical restrictions as the same taxa can be observed in any of the five ecozones, and there is a significant gap in sequenced picocyanobacteria (0.2-2 µm) in genomic libraries.

Whole genome shotgun metagenomic sequencing uses untargeted sequencing to characterize the functional gene diversity and biochemical potential of entire environmental DNA microbial communities. Its application is a valuable tool in observing the potential for toxin synthesis and taste/odour compounds that may negatively impact the state of drinking/recreational water systems. As the only toxin regulated in Canada (1.5 µg/L drinking water, 10 µg/L recreational waters), microcystin damages the liver and kidneys by acting as hepatocytes and disrupting intracellular protein regulation. Lipid A is the toxic component of lipopolysaccharides (endotoxins produced by gram-negative bacteria) which initiate dangerous immune responses and is the culprit for most gram-negative diseases. Alternatively, nitrogen-fixation by cyanobacteria converts atmospheric nitrogen into usable intracellular N_2 supplies, allowing for optimal nitrogen availability that is crucial for proliferation and gene expression. The potential for organisms to produce such toxins or be capable of fixing nitrogen is dependent on the presence of biosynthetic gene clusters encoding the necessary genes.

Several gene markers were targeted that represent the presence of potential biosynthesis of toxins and N_2-metabolic abilities. This research (Chapter 3) characterized seasonal distribution of two toxin gene families: microcystin (*mcy*) and lipoxygenase (*lpx*), and three nitrogen metabolism genes families: cyanase (*cyn*S), nitrogenase (*nif*), and a global nitrogen regulator (*ntc*A). Genes within the biosynthetic *mcy* gene cluster signify the production of microcystin, and *lpx* genes encode for lipid A endotoxins; the *cyn*S gene is involved in fixing cyanate (an organic nitrogen source), the *nif* gene cluster encodes the primary nitrogen fixation enzyme family, nitrogenase, and *ntc*A is a global nitrogen regulator crucial for all nitrogen assimilation and acquisition within cyanobacteria. The target toxin genes (*mcy, lpx*) were chosen for their potential to harm individuals infected with contaminated water, and the target nitrogen-metabolism genes (*cyn*S, *nif, ntc*A) were chosen due to their capability of providing nitrogen to cyanobacteria which allows them to proliferate, thrive, and express genes even when nitrogen levels in the water are not optimal. The target gene abundances were observed over six sampling points, including under an ice-covered period, using metagenomic analysis, to identify if there is a link among seasonal gene observations.

Acknowledgements

First and foremost, I would like to express my gratitude towards my supervisor, Dr. Kirsten Müller, who took the chance by taking me on amid a global pandemic. As it was unprecedented territory, all your patience and guidance were greatly appreciated, and I have learned so much under your supervision – I could not be more grateful for this opportunity. To my committee members, Dr. Monica Emelko and Dr. Trevor Charles, thank you for all your contributions and providing me with excellent guidance and feedback throughout my study.

I would like to acknowledge that this research was funded by the *for*Water NSERC Network and would not have been achievable without its contributions. And a special thank you to Dr. Ellen Cameron, whose work in the Turkey Lakes Watershed trailblazed the way for the rest of us in the Müller Lab, and was the fundamental basis for my thesis, without which my introduction to metagenomic and bioinformatic research would not have been possible.

To the members of the Müller Lab, past and present, I thank you for all your support and guidance. Timothy Shardlow, thank you for answering my endless questions as I attempted to learn new programs and concepts within a virtual vacuum, and for listening to me complain about metagenomics, R, and hundreds of phylogenetic tree revisions. Harshina Brijlall and Anjali Krishna, thank you for letting me join your fieldwork in TLW where we not only gained valuable field experience, but made long lasting memories during the 10-hour drives, our personalized playlists, and the countless lab inside jokes. Honourable mention to Anna Valero as well – our paths almost did not cross, and I am forever glad they did.

A tremendous thank you to my fellow grad student in crime and closest friend – Chelsea Lyons. Starting graduate studies during a national lockdown was a unique and incredibly stressful experience we will always share, and I could never have done it without you.

To the rest of my friends and family, thank you all for your ongoing support.

Finally, I'd like to acknowledge my siblings, Yazen and Rana, and our mother Lubna, whose tenacity in life taught me how to be resilient, strong, and stubborn enough to pursue big dreams.

Dedication

I dedicate this work, and all my future successes, to my late father, Yahia Abed (B.S. Civil Engineering, 1985), who I know without a shadow of a doubt is looking down on me, smiling from ear to ear.

None of this would have been possible without you, Dad.

I am not afraid of tomorrow, for I have seen yesterday, and I love today.

Table of Contents

Author's Declaration ..... **ii**

Statement of Contributions ..... **iii**

Abstract ..... **iv**

Acknowledgements ..... **vi**

Dedication ..... **vii**

List of Figures ..... **xiii**

List of Tables ..... **xv**

List of Abbreviations ..... **xvi**

Chapter 1 Cyanobacterial Harmful Algal Blooms **1**

1.1 Introduction 1

1.2 Algal Blooms 1

1.2.1 Factors Influencing Bloom Formations 2

1.2.2 Harmful Algal Blooms 3

1.3 Cyanobacteria 5

1.3.1 Cyanobacterial Harmful Algal Blooms (cHABs) 6

1.3.2 Cyanobacterial Prevalence across Canada 8

1.3.3 Obstacles to Cyanobacterial Identification 10

1.4 Water Quality and Conditions 13

1.4.1 Nutrient Availability 13

1.4.2 Nitrogen-Phosphorus Controversy 15

1.5 Research Goals 17

Chapter 2 Cyanobacterial Community Compositions across Five Forested Ecozones in Canada 19

2.1 Introduction 19

2.1.1 Forested Watersheds 19

2.1.2 16S rRNA Gene Sequencing 33

2.2 Objectives 35

2.3 Methods 35

2.3.1 Sample Collection and DNA Extraction 35

2.3.2 V4-16S rRNA Gene Sequencing and Analysis 36

2.3.3 Phylogenetic Analysis 37

2.4 Results 38

2.4.1 Taxonomy of Cyanobacteria 49

2.4.2 Biogeography of Cyanobacteria 51

2.5 Discussion 53

2.5.1 History of Cyanobacterial Classification Systems 53

2.5.2 Taxonomy of Cyanobacteria Across Five Major Forested Ecozones 54

2.5.3 Biogeography of Cyanobacteria Across Canada 59

2.6 Conclusions 61

2.7 Implications 63

Chapter 3 Seasonality of Cyanobacterial Toxin and Nitrogen Metabolism Genes in Turkey Lakes Watershed using Metagenomics 66

3.1 Introduction 66

3.1.1 Metagenomic Sequencing 66

3.1.2 Functional Gene Diversity (Toxins) 68
3.1.2.1 Microcystin (*mcy*) 71
3.1.2.2 Lipoxygenase (*lpx*) 74
3.1.3 Functional Gene Diversity (Nitrogen Metabolism) 77
3.1.3.1 Cyanase (*cyn*S) 78
3.1.3.2 Nitrogenase (*nif*) 80
3.1.3.3 Global Nitrogen Regulator (*ntc*A) 82
3.2 Objectives 83
3.3 Methods 84
3.3.1 Study Site: Turkey Lakes Watershed 84
3.3.1.1 Sample Collection 86
3.3.2 Whole Genome Shotgun (WGS) Sequencing 87
3.3.2.1 Illumina Sequencing 88
3.3.2.2 Bioinformatic Analysis 88
3.3.3 Phylogenetic Analysis 91
3.4 Results 92
3.4.1 Community Compositions 92
3.4.2 Metagenomic Gene Observations 95
3.4.3 Phylogenetic Analysis 103
3.5 Discussion 120
3.5.1 Cyanobacterial Taxonomy of Target Genes in TLW 120
3.5.2 Metagenomic Analysis of Target Genes 122
3.5.2.1 Functional Toxin Genes 123
3.5.2.2 Functional Nitrogen Metabolism Genes 128
3.5.3 Phylogenetic Analysis of Target Genes 133

3.5.4 Weather and Sampling Conditions 137
3.6 Conclusions 138
3.7 Implications 140
Chapter 4 Research Summary, Implications, and Future Directions 143
4.1 Research Summary 143
4.1.1 16S rRNA Gene Sequencing 144
4.1.2 Whole Genome Shotgun Metagenomic Sequencing 146
4.2 Implications 147
4.2.1 Cyanobacteria are distributed everywhere and can be unpredictable 147
4.2.2 Cyanobacteria and functional genes pose potential threats in oligotrophic systems 148
4.2.3 Standard water sampling protocols to observe accurate and reliable trends 149
4.2.4 Establish baselines for long-term source water monitoring amid a changing climate 151
4.3 Future Research 152
4.3.1 Multiple loci should be examined for higher resolution phylogenies 152
4.3.2 Specialized cyanobacterial biological mechanisms should be considered 153
4.3.3 Additional sequencing methods should be employed to observe gene expression 154
References 156
Appendix A V4-16S rRNA Region Gene ASV Community Composition 179
Appendix B Turkey Lakes Watershed Environmental Conditions and Sampling Summary 190

Appendix C Turkey Lakes Watershed Metagenomic NCBI Blast Reference Sequences 191

Appendix D Turkey Lakes Watershed Metagenomic NCBI Taxonomy Reference Sequences 203

Appendix E Metagenomic Shotgun Sequencing Supplementary Information 226

List of Figures

Figure 2.1 Map of the major forested ecozones and watersheds across Canada and their respective research platform sites **..... 21**

Figure 2.2 Maximum likelihood phylogenetic tree featuring V4-16S rRNA cyanobacterial ASVs from five forested watersheds across Canada (A-H) part of the *for*Water Network**.... 48**

Figure 3.1 Microcystin biosynthetic gene cluster with *mcy* encoding genes from *Microcystis aeruginosa* PCC7806 including NRPS, PKS, Tailoring, and ABC transporters**..... 74**

Figure 3.2 Lipopolysaccharide (LPS) gene cluster with biosynthetic *lpx* encoding genes**.... 77**

Figure 3.3 Cyanate assimilation pathway gene cluster from *Prochlorococus* sp. MED4 **..... 80**

Figure 3.4 Nitrogenase biosynthetic gene cluster from *Anabaena variabilis* ATCC29413 and *Leptolyngbya borya*na dg5**..... 82**

Figure 3.5 Map of Turkey Lakes Watershed, Sault Ste. Marie, Ontario (Canada)**..... 86**

Figure 3.6 Relative abundance of TLW microbial communities containing nitrogen metabolism and lipoxygenase gene markers**..... 93**

Figure 3.7 Heatmap representing seasonal abundances of target genes observations in Turkey Lakes Watershed microbial communities **..... 101**

Figure 3.8 Heatmap representing the seasonal abundances of target genes observed in Turkey Lakes Watershed cyanobacterial communities **..... 102**

Figure 3.9 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*A, based on identified proteins by Metagenome ATLAS eggNOG annotations **..... 105**

Figure 3.10 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*C, based on identified proteins by Metagenome ATLAS eggNOG annotations **..... 106**

Figure 3.11 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*H, based on identified proteins by Metagenome ATLAS eggNOG annotations **..... 107**

Figure 3.12 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*A, based on identified proteins by Metagenome ATLAS eggNOG annotations**..... 108**

Figure 3.13 The maximum likelihood phylogenetic tree for functional gene of interest *lpx*B, based on identified proteins by Metagenome ATLAS eggNOG annotations**..... 109**

Figure 3.14 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*C, based on identified proteins by Metagenome ATLAS eggNOG annotations **110**
Figure 3.15 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*D, based on identified proteins by Metagenome ATLAS eggNOG annotations **111**
Figure 3.16 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*K, based on identified proteins by Metagenome ATLAS eggNOG annotations **112**
Figure 3.17 The maximum likelihood phylogenetic tree for the functional gene of interest *cyn*S, based on identified proteins by Metagenome ATLAS eggNOG annotations **114**
Figure 3.18 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*J, based on identified proteins by Metagenome ATLAS eggNOG annotations **115**
Figure 3.19 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*K, based on identified proteins by Metagenome ATLAS eggNOG annotations **116**
Figure 3.20 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*S, based on identified proteins by Metagenome ATLAS eggNOG annotations **117**
Figure 3.21 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*U, based on identified proteins by Metagenome ATLAS eggNOG annotations **118**
Figure 3.22 The maximum likelihood phylogenetic tree for the functional gene of interest *ntc*A, based on identified proteins by Metagenome ATLAS eggNOG annotations **119**

List of Tables

Table 1.1 Major cyanotoxins and their corresponding bloom-forming genera..... **8**
Table 2.1 Number of unique cyanobacterial ASVs obtained from community V4-16S rRNA gene profiling from each ecozone included for phylogenetic analysis..... **39**
Table 3.1 Summary of sampling information and DNA extraction details from the Turkey Lakes Watershed, Ontario..... **87**
Table 3.2 Toxin biosynthesis gene cluster families present in TLW samples..... **90**
Table 3.3 Biosynthetic genes involved in nitrogen metabolism present in TLW samples **91**
Table 3.4 Assembled contigs generated through Quality Control and Assembly summaries as designated by the Metagenome ATLAS standard workflow pipeline..... **96**
Table 3.5 Microcystin gene markers observed from metagenomic eggNOG annotated data with assigned cyanobacteria taxonomy **97**
Table 3.6 Total abundances of target genes observed in microbial communities in TLW.... **98**
Table 3.7 Total abundances of target genes observed in Cyanobacteria in TLW..... **99**

List of Abbreviations

AM	Atlantic Maritime
ASV	Amplicon Sequence Variant
AXT-a	Anatoxin-a
BLAST	Basic Local Alignment Search Tool
bp	Base pairs (nucleotides)
BP	Boreal Plains
BS	Boreal Shield
C	Carbon
CAGEF	Centre for the Analysis of Genome Evolution and Function
CRD	Capital Regional District
cHAB	Cyanobacterial Harmful Algal Bloom
***cyn*S**	Gene encoding cyanate lyase/hydratase (cyanase)
DFO	Department of Fisheries and Oceans (Canada)
DNA	Deoxyribonucleic acid
ECCC	Environment and Climate Change Canada
eDNA	Environmental DNA
fdx	Gene encoding ferredoxin
feo	Gene encoding ferrous iron transport
geo	Gene encoding geosmin (T/O compound)
GR	Glenmore Reservoir
HAB	Harmful Algal Bloom
HRM	Halifax Regional Municipality
HQP	Highly Qualified Personnel
kb	Kilobase (1000 base pairs of DNA)
LPS	Lipopolysaccharide
LPX	Lipoxygenase
***lpx*A**	Gene encoding UDP-N-acetylglucosamine acyltransferase
***lpx*B**	Gene encoding lipid-A-disaccharide synthetase
***lpx*C**	Gene encoding UDP-3-0-acyl-N-acetylglucosamine deacetylase
***lpx*D**	Gene encoding UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acetyltransferase
***lpx*K**	Gene encoding tetraacyldisaccharide 4’-kinase
MAG	Metagenomic assembled genome
MC	Montane Cordillera
MC-LR	Microcystin-LR (leucine-arginine)
MCY	Microcystin
***mcy*A**	Gene encoding nonribosomal peptide synthetase
***mcy*C**	Gene encoding nonribosomal peptide synthetase
***mcy*D**	Gene encoding polyketide synthase

***mcy*E**	Gene encoding nonribosomal peptide and polyketide synthase fusion protein
***mcy*H**	Gene encoding ABC transporter ATP-binding-like protein
***mcy*J**	Gene encoding methyltransferase
MEGA ©	Molecular Evolutionary Genetic Analysis
MPB	Mountain Pine Beetle
N	Nitrogen
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NIF	Nitrogenase
***nif*J**	Gene encoding pyruvate-ferredoxin (flavodoxin) oxireductase
***nif*K**	Gene encoding nitrogenase molybdenum-iron
***nif*S**	Gene encoding cysteine desulferase
***nif*U**	Gene encoding Fe-S cluster assembly
***ntc*A**	Gene encoding global transcriptional nitrogen regulator
NRPS	Non-ribosomal peptide synthase
NRCAN	Natural Resources Canada
NSERC	Natural Sciences and Engineering Research Council (Canada)
P	Phosphorus
PCR	Polymerase Chain Reaction
PLW	Pockwock Lake Watershed
PKS	Polyketide synthase
PM	Pacific Maritime
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
TLW	Turkey Lakes Watershed
T/O	Taste and odour
URSA	Utikuma Region Study Area
V4	Variable region 4 of the 16S rRNA gene
WBF	Western Boreal Forest
WTF	Water Treatment Facility
WGS	Whole Genome Shotgun
WHO	World Health Organization

Chapter 1
Cyanobacterial Harmful Algal Blooms

1.1 Introduction

Access to clean and safe drinking water should be a human right and not a luxury, and yet, at least three billion people do not know the quality of their source waters due to a lack of monitoring (United Nations, 2022). All water requires treatment (Emelko et al., 2019), yet this is becoming more difficult as climate change is threatening the quality, quantity, and security of even the most regulated drinking water resources by triggering a plethora of climate-change-exacerbated disturbances such as floods, hurricanes, and wildfires (Arnell et al., 2015; Chorus et al., 2021; Pick, 2016; Robinne et al., 2019; Wells et al., 2020). Global warming is also influencing human migration, population distribution and urbanization as more rural areas are being populated, causing an increase in anthropogenic disturbances – ultimately altering water chemistry and quality by shifting nutrient inputs and disrupting the balance of water systems (Arnell et al., 2015; Chorus et al., 2021; Pick, 2016). The necessity of safe drinking water has led to a rise in research to better understand the interconnected synergies at work beneath the surface of water bodies, and specifically oligotrophic (nutrient poor) systems as there are still significant gaps within this research area. One expanding area of interest in forest management and water treatability practices, and the principal focus for this thesis, is the study of cyanobacterial blooms and their presence and distribution in oligotrophic lakes.

1.2 Algal Blooms

Algal blooms have many definitions, one of them (i.e., the one referred to in this thesis) being the rapid accumulation of phytoplankton within a water supply, often – but not exclusively – recognizable as visually unappealing discoloured scums on the surface of lakes and other water bodies (Paerl, 2008). Typically, algal blooms accumulate in eutrophic (nutrient rich) lakes because of anthropogenic eutrophication – the process by which water systems are increasingly enriched with nutrients and minerals – often due to agricultural and residential

fertilizer and irrigation runoff (Morrissey et al., 2018). Eutrophication leads to algal blooms by providing phytoplankton with excess nitrogen (N) and phosphorus (P) which are important nutrients for their proliferation (Paerl, 2018; Paerl et al., 2011a). Climate change also interferes with the hydrological cycle and promotes algal proliferation by increasing water temperatures, water column stability, growing season lengths, and higher incidences of invasive species (Chorus et al., 2021; O'Neil et al., 2012; Pick, 2016). Climate change-exacerbated disturbances such as wildfires frequently involve increased erosion and associated runoff, which can contain high levels of nutrients that can persist for decades or longer in some regions (Emelko et al., 2011; 2016; Blackburn et al., 2023), even at large basin scales (Emmerton et al., 2020). Thus, it is not surprising increases in the occurrence of algal blooms have been observed in oligotrophic lakes, rivers, and streams across the North American continent and attributed to changing climate (Stoddard et al., 2016). Collectively, these factors are leading to an increased prevalence of algal blooms across water bodies during longer seasonal windows and in response to shifting nutrient inputs.

1.2.1 Factors Influencing Bloom Formations

The rapid accumulation of algal blooms can occur differently depending on the algal species and the environment in which they grow in. Many algal species can proliferate so efficiently that they can outpace the rate of consumption and grazing by predatory species, such as some amoeba that have been documented to prey on cyanobacteria (Havens, 2008; Paerl et al., 2001), and some species (mainly toxin producing strains of cyanobacteria) which are not grazed upon at all (Ma et al., 2017; Paerl, 2008). Aside from production-consumption dynamics (i.e., how quickly phytoplankton can proliferate before being grazed on), many phytoplankton can passively accumulate by a unique form of "swimming" where they deform their cell body in a way that promotes underwater movements, regulating buoyancy, and using special adaptations that decrease sinking rates, as well as the combined efforts of each (Barsanti & Gualtieri, 2014; Reinl et al., 2021). This allows phytoplankton to move throughout the water column to depths that are beneficial to them, such as shifting for optimal light and nutrient

availability or avoiding predators, much like any other biological organism, all of which influence how successful they are in their given environment.

The optimal nutrient loading rates that influence the onset of algal blooms has been an area of controversy for over forty years due to conflicting results from research studies stating phosphorus as the limiting agent (Sterner, 2008), nitrogen as the limiting agent (Howarth & Marino, 2006), or that both nutrients are equally important for algal growth (Paerl et al., 2016). While eutrophication is a well-known driving force for algal bloom formations, the precise conditions required is still not as clear (Morrissey et al., 2018). Although many algae show an affinity for nitrogen (N) and phosphorus (P) inputs, the N:P ratio that best predicts the onset of blooms is still argued among the scientific community (Chorus et al., 2021). The root of the debate lies in the capability of some algal species to produce their own nitrogen sources through nitrogen-fixation. Nitrogen fixation is the process of transforming free atmospheric nitrogen into a more usable source of nitrogen, such as ammonia or nitrate, to use in metabolic processes (Morrissey et al., 2018; Watanabe & Horiike, 2021). As early as 1975, the effects of eutrophication have been of serious concern, especially regarding how changes in biodiversity can ultimately affect water quality (Chorus et al., 2021; Favot et al., 2019; Paerl, 1997; Schindler, 1975; Watson, 2004). More specifically, the controversy around algal blooms and eutrophication focuses on which nutrient (N and P) is more linked to high nutrient systems, and which is the more limiting agent responsible for the degree to which blooms can thrive (Chorus et al., 2021; Paerl, 2008; Paerl et al., 2016).

1.2.2 Harmful Algal Blooms

Aside from exuding displeasing odours and displaying visually unappealing scums, algal blooms are also capable of creating harmful environments for humans, animals, and other phytoplankton communities, thereby coining the phrase harmful algal blooms (HABs) (Barsanti & Gualtieri, 2014). Although HABs form naturally, climate change and anthropogenic factors are causing them to increase in prevalence and distribution among water systems of varying trophic levels (Chorus et al., 2021; Chorus & Welker, 2021).

Anthropogenic influences in industrial, agricultural, and urban sectors have induced many environmental changes leading to more nutrients seeping into the water and in turn, more phytoplankton proliferation (Chorus & Welker, 2021; Davis et al., 2009; Favot et al., 2023; Paerl, 1997; Piehler, 2008). Phytoplankton proliferation in HABs depletes the water of nutrients, dissolved oxygen, and inorganic carbon (C), leaving remaining organisms with very little to survive on, let alone thrive (Chorus et al., 2021; Paerl et al., 2001). The surface scum produced by HABs can become dense enough to reduce water clarity by obstructing light penetration into the water column, thereby shading photosynthetic species from sunlight, and reducing visibility for animals foraging in the area (Barsanti & Gualtieri, 2014; Morrissey et al., 2018; Paerl et al., 2001). These scums are also increasingly likely to harbor microbes, pathogens, and (potentially harmful) secondary metabolites – all of which can deplete oxygen from the water column causing hypoxic events that can be fatal to microbial communities and aquatic inhabitants like fish and molluscs (Paerl et al., 2001; Paerl, 2018).

Natural conditions influencing bloom formation include sedimentation patterns within the water column – referring to how nutrients stratified in the water column can alter turbidity and clarity, subsequently increasing UV ray absorption, water temperature, and thermocline status (Paerl, 2008; Paerl, 2018). Hydrological events (i.e., the movement of water) impacts bloom formation through residence time and flushing rates (required time for water in a source to completely exchange) (Morrissey et al., 2018; Romanis et al., 2021) that determines how often and how radically nutrients shift within the water column (Chorus & Welker, 2021; Paerl et al., 2001; Paerl, 2008). Algal blooms have complex impacts on the environment and can have several influences on the other organisms in source waters, many of which are still currently under investigation, yet the ambiguity clearly demonstrates the danger they pose, and why more research on HABs is still required (Chorus et al., 2021; Morissey et al., 2018; Paerl, 2018).

1.3 Cyanobacteria

Cyanobacteria, one of the most notorious producers of HABs, are the oldest known oxygenic phototrophic organisms to inhabit Earth and have been represented in the fossil record dating back approximately 3.5 billion years (Chorus & Welker, 2021; Paerl et al., 2001; Paerl, 2018). They are gram-negative bacteria, also known as blue-green algae, whose evolutionary history allowed them to develop unique ecological, morphological, and physiological adaptations to help them survive in changing environments (Chorus et al., 2021; Paerl et al., 2001; Valadez-Cano et al., 2022). These distinctive adaptations enabled cyanobacteria to dominate in marine and freshwater environments, particularly under (but not limited to) eutrophic conditions (Paerl & Otten, 2013). As a result, these organisms can expand their population size by proliferating highly efficiently, resulting in cyanobacterial harmful algal blooms (cHABs) that dominate nutrient and oxygen consumption, and prevent light penetration to deeper depths in the water column (Paerl, 2018; Paerl & Otten, 2013). In addition to adapting distinctive properties over-time, varying genera of cyanobacteria can fix atmospheric nitrogen (i.e., diazotrophs), store phosphorus, and isolate iron (Fe) among other trace metals from the water column – all of which influence their success rates (Detoni et al., 2022; Howarth et al., 1988; Nash et al., 2018; Paerl & Otten, 2013; Sadvakasova et al., 2022). For this study, when harmful algal blooms and blooms of any kind are discussed, the focus will be specific to cyanobacterial HABs.

The long evolutionary history of gram-negative cyanobacteria has enabled them to tolerate extreme conditions, succeed in numerous habitats, and exploit limited nutrients, all while resisting harmful anthropogenic influences (Paerl & Otten, 2013; Singh et al., 2016). Left undisturbed, cyanobacteria are natural biofertilizers by providing nitrogen to their surrounding environment, and act as biocontrol agents by producing active biological compounds with the potential for antiviral, antifungal, antibacterial, and antialgal properties (Singh et al., 2016). Additional applications have been used towards bioremediation, plant growth promoters, and bioenergy, however, the most widely known application of

cyanobacteria is for its use as a food supplement (Singh et al., 2016). The genus *Spirulina* is the most commercialized strain of cyanobacteria due to its outstanding nutritional value, high protein content, and its vibrant blue-green coloration which works excellently as natural food coloring (Singh et al., 2016). Although cyanobacteria do have several beneficial applications in industrial settings, left unmonitored or untreated in their own in the environment they can wreak havoc on aquatic systems and threaten the provision of safe drinking water (Chorus et al., 2021; Singh et al., 2016).

1.3.1 Cyanobacterial Harmful Algal Blooms (cHABs)

Harmful algal blooms dominated by cyanobacteria are the most challenging and nuisance species in freshwater environments because of the adverse health effects they can have on humans and animals interacting with surface waters contaminated with harmful cyanobacteria, many of which can produce potent toxins (Kelly et al., 2019; Paerl et al., 2001). Although cHABs have been around for decades, they are becoming a larger concern because increased eutrophication is leading to more widespread, frequent, and severe toxic blooms (Chorus et al., 2021; Chorus & Welker, 2021; Favot et al., 2023; Kelly et al., 2019), and enhanced molecular analysis has provided better insight to the unique taxonomy within the Cyanobacteria phylum (Strunecký et al., 2023). The research conducted in this thesis is particularly concerned with cyanotoxins because of their potential to enter water treatment facilities (WTF) clogging filtration processes and potentially leading to service disruptions or even outages when toxins are produced because most WTFs do not have the capacity to treat toxins (Emelko et al., 2011). By better understanding the mechanisms behind cyanobacterial blooms, proactive improvements to monitoring programs can be made to prevent the risk of toxins entering water systems and mitigate threats to clean and safe drinking water.

Cyanobacteria can produce secondary metabolites, products synthesized by organisms that most likely provides some type of benefit as additional energy is expended for their production but are not crucial to their survival (Calteau et al., 2014; Chen et al., 2020; Cornet et al., 2021). Many secondary metabolites released by cyanobacteria are identified as naturally

occurring toxins that can cause serious damage to the environment as well as to humans interacting with contaminated water (Chen et al., 2020). The three main toxins associated with cyanobacteria (referred to hereon out as cyanotoxins) are all classified according to the organ being targeted – hepatotoxins (liver), neurotoxins (nervous system), and dermatotoxins (skin irritant) (Chorus & Welker, 2021; Health Canada, 2022; Paerl et al., 2001; Pearson et al., 2010; Sivonen & Jones, 1999). Several well-known toxin-producing cyanobacteria genera and their common toxin names are listed in **Table 1.1**. Depending on the type of toxin and how it is released from cells there are many routes of exposure that cyanotoxins can utilize, among the most prevalent routes being dermal absorption, water ingestion, and inhalation (Health Canada, 2022; Sivonen & Jones, 1999). This makes cyanobacteria blooms especially dangerous because their toxins can harm individuals (humans and animals alike) swimming in infected waters, consuming toxin-contaminated water and/or food, and sometimes through airborne inhalation (Curren & Leong, 2020; Health Canada, 2022; Wiśniewska et al., 2020). While the reasons for cyanobacteria to synthesize cyanotoxins is still unknown and widely debated, vigilant awareness has been established worldwide to monitor the production of the toxic metabolites and prevent their arrival into source waters.

In addition to direct contact with cyanotoxins in recreational or drinking water sources, several studies have also shown that cyanotoxins can enter the food chain downstream from a bloom if contaminated waters are used for irrigation or enter riverbeds, creating a domino effect through bioaccumulation (Paerl et al., 2001; Romanis et al., 2021; Zanchett & Oliveira-Filho, 2013). There have been reported cases of livestock drinking from recycled water contaminated with the cyanotoxin microcystin and demonstrating symptoms of fatigue, lethargy, weakness, and other gastrointestinal issues (Romanis et al., 2021). How instances like this may be currently affecting the food web and trophic levels is still greatly unknown, and studies observing bioaccumulation of cyanotoxins are very sparce yet (Romanis et al., 2021). With the rise of cyanobacteria around the world, the potential for these toxins to spread within the environment is progressively alarming, and a cause for immediate research and the implementation of improved proactive mitigation practices.

Table 1.1 Major cyanotoxins and their corresponding bloom-forming genera

Toxicity	Common Name	Bloom-Forming Genera
Hepatotoxin	Microcystin	*Anabaena* (benthic), *Aphanizomenon, Anabaenopsis, Dolichospermum* (planktonic), *Hapalosiphon, Leptobylynga, Microcystis, Nostoc, Oscillatoria, Phormidium, Planktothrix, Pseudoanabaena, Spirulina*
	Nodularin	*Iningainema, Nodularia*
	Cylindrospermopsin	*Anabaena* (benthic), *Aphanizomenon, Cylindrospermopsis, Dolichospermum* (planktonic), *Lyngbya, Oscillatoria, Raphidiopsis, Umezakia*
Neurotoxin	Anatoxin-(a)	*Anabaena* (benthic), *Aphanizomenon, Cylindrospermopsis, Dolichospermum* (planktonic), *Lyngbya, Microcystis, Oscillatoria, Phormidium, Planktothrix, Raphidiopsis*
	Saxitoxin	*Anabaena* (benthic), *Aphanizomenon, Cylindrospermopsis, Lyngbya, Oscillatoria, Planktothrix*
Dermatotoxin	Aplysiatoxin	*Lyngbya, Oscillatoria, Schizothrix*
	Lyngbyatoxin	*Lyngbya, Oscillatoria, Schizothrix*

Adapted from Du et al., 2019, Paerl and Otten, 2013, and Chorus and Welker, 2021.

1.3.2 Cyanobacterial Prevalence across Canada

In North America alone, toxic cyanobacteria have already been identified in no less than 27 of the 50 United States and are most likely present in water bodies across most of the US and Canada, and especially among the Great Lakes (Hudnell & Dortch, 2008; McKindles et al., 2019; Pearson et al., 2010). Substantial cHABs across the Great Lakes were significant enough to warrant the US-Canada Great Lakes Water Quality Agreement in 1968 (amended 1972) to restore the health of Laurentian Great Lakes (McKindles et al., 2019). The agreement improved water quality and diminished many cHAB events, one of the most notable being the

Lake Erie blooms that lasted throughout the 1950s–1970s (McKindles et al., 2019). However, while some progress was made to restore the Great Lakes, invasive dreissenid mussels in the 1990s in Lake Erie quickly altered the ecosystem and composition of nutrients (McKindles et al., 2019); while these mussels are hypothesized to support *Microcystis* proliferation, not enough monitoring was conducted to adequately test the theories (McKindles et al., 2019). And with a very similar water composition to Lake Erie, Lake Winnipeg (Manitoba) is another example where severe cHABs dominate – momentous enough to be picked up by satellite imagery and are rapidly increasing in magnitude and frequency (McKindles et al., 2019). The aftermath of cHAB identification in Canadian lakes ignited a drive towards improved monitoring of blooms and cyanotoxins using molecular methods and an increase in reporting of such events.

The presence of cyanobacteria in Canada has led to the construction and redefining of numerous guidelines laid out by Health Canada as an attempt to monitor cyanotoxin levels, especially in recreational and drinking water sites (Health Canada, 2020; 2022). As of 2015, the only cyanotoxin guidelines set by the Canadian government were outlined for microcystin (nation-wide), and anatoxins in Quebec (Pick, 2016). Cyanobacteria have been prevalent in most Canadian provinces and territories since the early 1900's, however, the extensive number of lakes across Canada makes monitoring and reporting very challenging (Kotak & Zurawell, 2007; McKindles et al., 2019; O'Keeffe, 2019). In 2018, for example, there were at least 150 occurrences of cHABs in Canada reported by provincial governments, media, and/or public health sites (O'Keeffe, 2019) – a vastly under-reported number considering the extent of water bodies across Canada (Government of Canada, 2013). The most prominent toxic cyanobacteria genera found across Canada includes *Dolichospermum* sp. (previously known as planktonic *Anabaena* sp.), *Aphanizomenon* sp., *Microcystis* sp., and *Worochinia* sp., although the list keeps growing (Favot et al., 2023; Health Canada, 2020; McKindles et al., 2019; Reinl et al., 2021).

Conversely, the wide range of biomes across Canada provide cHABs with varying conditions that may lead to longer or shorter bloom seasons, or different timeframes altogether; the drastic differences in biomes and how it relates to bloom seasons is especially important for this research and will be discussed in detail in Chapter 2. Because cHABs typically prefer warm water, there have been new cases of longer bloom seasons recorded across Canada, most likely the result of warming conditions from climate change (Favot et al., 2023; O'Keeffe, 2019; Winter et al., 2011). Recent data in Canada has shown that there are increasing trends for blooms to begin earlier in the spring and extend later into the year (Favot et al., 2023; Health Canada, 2020; Reinl et al., 2021; Winter et al., 2011). The general trend for most cyanobacterial bloom seasons in Canada has tended to fall near the end of spring, through into summer, and usually extending well into autumn, although some have the potential to last much longer (Favot et al., 2023; Havens, 2008; Kotak & Zurawell, 2007; O'Keeffe, 2019; Reinl et al., 2021; Winter et al., 2011). In some regions in Canada, cyanobacterial bloom seasons start as early as April and last as late into December (much longer than the usual season), with some blooms even appearing during ice covered months (suggesting year-round seasonality), and the highest density of blooms are typically observed in August and September (Favot et al., 2023; O'Keeffe, 2019; Winter et al., 2011). The concern, however, is these statistics are based on monitoring when visual observations of algal blooms on water surfaces are observed, yet ongoing research has pointed out that cyanobacteria can be present at varying depths in the water column without visual blooms (Cameron et al., 2022a; Cameron et al., 2022b; Cameron, 2021); this notion suggests that testing the water only when a bloom is observed may be leading to a crucial gap in knowledge regarding cyanobacteria presence year-round when sampling may not be conducted.

1.3.3 Obstacles to Cyanobacterial Identification

Cyanobacteria have been directly observed and reported on every continent and in at least 54 countries, with suspected cases of cyanotoxins infiltrating water bodies across Europe, the Middle East, Africa, South America, Australia, and Asia (Du et al., 2019; Pearson et al.,

2010; Sivonen & Jones, 1999). Although cyanobacteria have been reported virtually anywhere that samples could be taken (Sivonen & Jones, 1999), records are difficult to maintain because a national database does not currently exist (Hudnell & Dortch, 2008; Orihel et al., 2012). One of the major issues preventing a national database are the disagreements regarding taxonomic characterization of algae, and specifically cyanobacteria as a member of the algal community (Du et al., 2019; Hudnell & Dortch, 2008; Strunecký et al., 2023). There are additional issues in the type of approaches used to classify cyanobacteria – ranging from morphological observations, genetic fingerprinting, and physiological characterizations to molecular analyses, biogeochemical tests, and bioinformatics – all of which may result in different algal classifications (Du et al., 2019; Hudnell & Dortch, 2008; Komárek et al., 2014; Strunecký et al., 2023). The ambiguity of algae and a lack of proper criterion means creating and organizing a nationally recognized database to record cyanobacteria detections is incredibly difficult and complicated. The complexity regarding cyanobacterial classification is a topic of dispute among the scientific community, and as such, can be an area that undergoes heavy and frequent revisions.

The detection and identification of cyanobacteria is an ongoing obstacle for researchers due to the wide array of analytical and systematic methods that can be used. Some cyanobacteria have distinct physical characteristics such as visual (e.g., shape, colour), odours, and bioindicators (e.g., micro/macroinvertebrates), however, the overlap across algal taxa are too widespread to rely solely on physical features (Komárek et al., 2014; Strunecký et al., 2023). Yet many researchers still rely on morphology due its relative simplicity compared to analytical genetic and biochemical testing such as molecular analyses and electron microscopy (Cordeiro et al., 2020; Komárek, 2016). More significantly, difficulties arise when morphological and molecular identifications conflict with each other. This often occurs because morphology can be too simple or too variable for identification purposes and does not consider the probability of cyanobacteria altering their morphologies to adapt to their ecological conditions (Komárek, 2016; Strunecký et al., 2023). One of the most significant mistakes made in the classification process are when morphological identifications are

arbitrarily selected from traits in isolated cyanobacterial strains and then used to identify cyanobacteria in the field, not taking into consideration the ecological habitats and how that may alter morphology (Cordeiro et al., 2020). Although extremely valuable, the most significant downfall of morphological classification is the misconception that morphology can determine toxicity in cyanobacteria, which is highly inaccurate and untrue and is a mistake many researchers continue to make when classifying cyanobacteria (Komárek, 2016; Komárek et al., 2020; Strunecký et al., 2023).

The most reliable solution for modern taxonomy, and the most currently used approach (Cordeiro et al., 2020), is in fact a union of all available methods – morphological, ecological, genetic, biochemical, and molecular – known collectively as the *polyphasic approach* (Komárek, 2016). The polyphasic approach combines and compares different identification criteria based on genetic and molecular foundations to remove unsuspected cyanobacteria and eventually narrow the list down (Komárek, 2016). However, molecular methods are not without pitfalls as they only cluster samples based on sequence similarities – meaning the results can highly vary based on the availability and accessibility of data sets (Komárek, 2016; Strunecký et al., 2023), and is contingent upon computational analysis and statistical error, often requiring further analyses to validate results. The polyphasic approach is highly regarded because it considers the gaps in each individual methodology (morphological, molecular, ecological, etc.) and demonstrates how ambiguous taxa can be clarified and highly resolved by using multiple approaches (e.g., some species are morphologically identical but are determined to be taxonomically distinct when molecular methods are incorporated in addition to morphology) (Komárek, 2016; Strunecký et al., 2023). Nevertheless, molecular and genetic components combined are more reliable and accessible than morphological assessments alone, and therefore all samples for the purpose of this research study have been identified using V4-16S ribosomal gene RNA (rRNA) or whole genome shotgun (WGS) metagenomic sequencing.

1.4 Water Quality and Conditions

A key component affecting the likelihood of cHAB formations is the temperature and thermocline within the water column. Generally, the highest growth rates of cyanobacteria are seen at temperatures exceeding 20-25 ˚C (Chorus et al., 2021; Havens, 2008; Paerl & Otten, 2013). At these temperatures, cHABs can outcompete eukaryotic algae in growth (Chorus et al., 2021), allowing them to occupy more space and consume nutrients faster. As more cyanobacteria occupy the surface waters, they continue to increase the temperature of the water by absorbing more heat and light energy from the sun's UV rays (Paerl & Otten, 2013). Temperature is considered equally as important to bloom proliferation as nutrient availability is and, along with wind currents and other environmental conditions, affects water quality by disrupting by hydrological processes taking place within the water body, such as thermal stratification (Favot et al., 2023). Water columns with longer residence times (i.e., the amount of time water remains within a water body before being flushed out) will amass nutrients when there is little mixing within the water column, further contributing to rising water temperatures due to increasing light and heat absorption (Ibelings et al., 2021; Paerl & Otten, 2013). Low rates of flushing also supply cHABs with more readily available nutrients because the lack of water mixing allows nutrients to remain stagnant on the surface, whereas high flushing rates likely dilute cyanobacteria as they form blooms by frequently turning over the water and moving the cyanobacteria throughout the water column and across a water body (Ibelings et al., 2021; Paerl & Otten, 2013).

1.4.1 Nutrient Availability

Cyanobacterial HABs are most often observed in eutrophic lakes where they have sufficient nutrients to acquire for bloom formations (Chorus et al., 2021; Ibelings et al., 2021; Paerl & Otten, 2013). Because the composition of nutrients within lakes can vary greatly, studies conducted on which nutrient ratios provide the optimal conditions for cHABs is still a very controversial area of research. Micronutrients are a fundamental requirement for algae –

used to build their cells and perform all metabolic and basic cellular functions – and comprise of three main categories: macronutrients, trace metals, and vitamins (Zhang et al., 2019). The longest studied category, macronutrients, was first proposed by Alfred Redfield (1958) in which he described the composition of nutrient ratios for phosphorus, nitrogen, and carbon in particulate organic matter (POM, originally retained on 0.7 μm filters) in oceans and lakes with high levels of algal blooms to be approximately 1:15:105 (P:N:C), respectively (Sharoni & Halevy, 2020). Moving forward, the accepted understanding was that phosphorus should be the limiting nutrient for algae because not only was it found in the smallest quantity, but it is also the only nutrient that could not be supplemented by the organisms themselves (i.e., via physiological mechanisms) (Chorus et al., 2021; Chorus & Welker, 2021; Favot et al., 2023; Sterner, 2008). As further research has continued, less researchers agree with the coined "Redfield's ratio" because there is much more variability in N:P ratios depending on the water source and its relative environment (Sharoni & Halevy, 2020). Due to this phosphorus limitation paradigm, phosphorus became widely accepted as the major limiting agent in primary production in aquatic environments, controlling the extent to which plants and phytoplankton could prosper (Howarth & Marino, 2006; Ibelings et al., 2021; Paerl & Otten, 2013; Paerl et al., 2016; Redfield, 1958; Sterner, 2008).

The second category, trace metals, is also important aside from N, P, and C nutrients; some cyanobacteria can isolate iron (Fe), molybdenum (Mo), and zinc (Zn), among other trace metals which may have additional nutrient loading properties (McRose et al., 2017; Paerl & Otten, 2013; Sadvakasova et al., 2022; Watanabe & Horiike, 2021). The most notable metal used by cyanobacteria is iron in its soluble form (Fe^{2+}) because of its crucial role in maintaining nitrogen assimilatory enzymes, without which the algae could not fix nitrogen (Esteves-Ferreira et al., 2017; Ibelings et al., 2021; McRose et al., 2017; Perovich et al., 2008; Thiel, 2019). Other metals found to play some role in bloom proliferation or thought to contribute to HABs include vanadium (V), manganese (Mn), cobalt (Co), copper (Cu), and selenium (Se), although their scope is not well known but predicted to also interact with nitrogen assimilation enzymes or act as limiting cofactors – still, there is not enough evidence to confidently support

these claims (Barsanti & Gualtieri, 2014; Ibelings et al., 2021; Paerl & Otten, 2013; Thiel, 2019; Watanabe & Horiike, 2021). Trace metals play vital roles in nitrogen fixation and the nitrogenase enzyme complex which directly correlates to the success of cyanobacteria in their environment and will be discussed in further detail in Chapter 3.

1.4.2 Nitrogen-Phosphorus Controversy

Typically, nitrogen is believed to be the nutrient responsible for encouraging the initiation of new primary productivity, whereas phosphorus acts as the nutrient that limits the amount of primary productivity after it is underway (Paerl & Otten, 2013). While this signifies that both nitrogen and phosphorus are essential for cHABs, it also indicates that excess availability of phosphorus is more likely to encourage algal blooms than an excess of nitrogen would (Howarth et al., 1988; Paerl & Otten, 2013). A supporting factor that perpetuates the phosphorus limitation paradigm is that of nitrogen fixation among diazotrophic cyanobacterial communities in which atmospheric nitrogen is converted to ammonium (the main N source utilized) and the assumption that N_2-fixation provides an adequate source of nitrogen (Detoni et al., 2022; Ibelings et al., 2021). Nitrogen-fixation is the root of the paradigm because along with generally higher amounts of nitrogen in water sources, any deficits are believed to be counter-balanced by diazotrophic cyanobacteria, and therefore is often overlooked as a limiting nutrient for cHABs (Howarth & Marino, 2006; Ibelings et al., 2021). There are many inconsistencies with this theory though because not all cyanobacteria can fix nitrogen and are therefore excluded from supplementing nitrogen themselves, relying solely on the amount bioavailable in the water (Ibelings et al., 2021; Paerl et al., 2011b). These inconsistencies are further supported by freshwater cyanobacterial species whose nitrogen fixation abilities can differ depending on the nutrient loads and ratios available to them. While diazotrophic cyanobacteria can supplement themselves with free nitrogen and thrive under low N:P ratios, high nitrogen loads (high N:P ratios) have been seen to encourage the proliferation of non-nitrogen-fixing cyanobacteria instead (Howarth & Marino, 2006; Paerl et al., 2011b).

More recently however, the argument against phosphorus being the critical limiting nutrient has gained traction in the scientific community. There is a growing list of research studies finding that having water nutrient levels high in phosphorus and low in nitrogen are not universal triggers for cHABs (Paerl et al., 2008; 2011b; 2016; Paerl & Otten, 2013). One of the main arguments is against the concept of nitrogen fixation. Although N_2-fixation in cyanobacteria can supplement some species with metabolic nitrogen, the amount produced physiologically is not sufficient to meet the demands of the ecosystem; this is especially important when considering the high energy expenditure required by the organisms, and that only a limited portion of the phytoplanktonic community constitutes diazotrophic bacteria and cyanobacteria (Moisander et al., 2002; Paerl & Otten, 2013). Furthermore, as N_2-fixation is an anaerobic process, occurring in specialized cells called heterocysts, the oxygen produced through photosynthesis can act as an antagonist and inhibit the nitrogen fixation process which can also be disrupted by natural disturbances such as turbulence in the water or strong gusts of wind (Ibelings et al., 2021; Paerl & Otten, 2013).

Cyanobacterial scums have been found to occur in waters with high N:P ratios in regions where nitrogen inputs from urban, agricultural, and/or industrial sectors were found to exceed phosphorus levels on a frequent basis (Paerl & Otten, 2013). A consequence of a history of researchers claiming that phosphorus is the limiting agent is that the government, scientists, and environmentalists alike have all made it a priority to control phosphorus monitoring programs (Ibelings et al., 2021) and keep levels below the maximum daily loads in water sources at risk for cHABs – however, the same cannot be said for nitrogen (Paerl & Otten, 2013). This is most likely due to the difference between nitrogen being "unlimited" because it can be imported into the environment via N_2-fixation, and exported via denitrification, whereas phosphorus remains in the environment until it is taken up by organisms and cannot be imported back into the atmosphere, recycling itself between the water column and sediment until it is consumed (Ibelings et al., 2021). The most recent studies in this area have found that cHAB formations are most strongly stimulated by the addition of equal levels of both nitrogen and phosphorus, as opposed to one over the other, suggesting co-limitation may be the

underlying factor (Paerl & Otten, 2013; Paerl et al., 2016). Unfortunately, the controversy regarding N:P ratios are still very much on-going, and more research is being conducted to specify more accurate findings and a stronger consensus. Aside from the conflicts between nitrogen and phosphorus among the scientific community, nitrogen will be the nutrient of focus for this study to observe how cyanobacterial biological responses (i.e., nitrification) vary seasonally and will be explored further in Chapter 3.

1.5 Research Goals

How and when cyanobacterial harmful algal blooms appear in response to nutrient inputs continues to be a major area of research, yet most of the work being done focuses on eutrophic water systems impacted by anthropogenic disturbances (despite knowing cyanobacterial blooms can occur in all trophic systems). Less is understood about blooms in oligotrophic systems unimpacted by external disruptions and therefore does not reflect cyanobacteria in undisturbed habitats (Brown et al., 2021; Favot et al., 2019; Reinl et al., 2021). As climate change continues to shift environmental conditions, microbial communities are shifting and adapting as well. The shift in microbial communities is especially significant in oligotrophic systems that have fewer monitoring programs in place, hence a better understanding of the taxonomy and functional diversity of cyanobacteria is crucial to improve and maintain mitigation practices.

The work conducted in this thesis builds upon previous research on oligotrophic systems across Canada including the Turkey Lakes Watershed (Ontario), Glenmore Reservoir (Alberta), hundreds of shallow ponds across the Utikuma Region Study Area (Alberta), Pockwock Lake Watershed (Nova Scotia), and the Comox Valley Watershed (British Columbia). Much of the work in these watersheds have focused on the composition of cyanobacteria communities in water systems with unique and distinctive environmental biomes, biodiversity, and varying levels of anthropogenic interference (Aljoudi, 2021; Cameron, 2021; Kwok, unpublished; Shardlow, 2021). Previous work conducted in the Turkey

Lakes Watershed by Cameron (2021) identified the knowledge gaps in present day literature regarding how cyanobacteria are stratified in the water column throughout the day, how their community composition varies seasonally, and highlighted how these factors can impact monitoring programs, while also suggesting practical solutions that can be applied to all types of water systems.

Expanding on previous research, the overarching aim of this study is to gain a comprehensive whole understanding of cyanobacteria distribution observed across the country and gain a deeper insight into what cyanobacterial biochemical and gene potentials may look like in undisturbed ultraoligotrophic systems to help mitigate monitoring protocols in source waters. Research in this thesis expands on Cameron's research (2021) to demonstrate the functional gene potential in cyanobacterial communities in a relatively undisturbed watershed, free of anthropogenic interactions, pollution, and natural disasters. This research provides deeper insight as to how cyanobacteria are distributed across great distances (and if they are limited by geographic restrictions) and demonstrate what the baseline gene potential in microbial communities looks like when there are little to no disturbances in the region. In doing so, describing the potential threats of cyanobacteria these water systems may face even in the absence of eutrophication will hopefully drive the importance of proactive research to prevent toxic blooms as opposed to responding to them, as well as by highlighting potentially harmful cyanobacteria in each forested ecozone (or the lack thereof), it emphasizes the potential threats individual watersheds could encounter. Funding for the research conducted in this thesis was provided by the pan-Canadian *for*Water NSERC transdisciplinary network for Forested Drinking Water Source Protection Technologies (https://www.forwater.ca/).

Chapter 2
Cyanobacterial Community Compositions across Five Forested Ecozones in Canada

2.1 Introduction

This chapter aims to identify the biogeographic trends among cyanobacterial communities from five forested watersheds in ecozones across Canada (as outlined by the *for*Water NSERC Network): Pockwock Lake watershed from the Atlantic Maritime (AM), Utikuma Region Study Area (URSA) in the Boreal Plains (BP), Turkey Lakes Watershed (TLW) in the Boreal Shield (BS), Glenmore Reservoir (GR) in the Montane Cordillera (MC), and the Capital Regional District (CRD) in the Pacific Maritime (PM). Each watershed was chosen for its unique location and physiographic regions that represent the major forested ecozones of Canada that experience differing climates, latitudes, altitudes, and vegetation – all of which might affect the composition of microbial communities present. With the application of next-generation sequencing (NGS), 16S rRNA gene amplicon sequencing has been the marker gene of choice since 1985 and is the most widely used gene in phylogenetic classification systems for several microorganisms, including bacteria and cyanobacteria (Tsukuda et al., 2017; Winand et al., 2019; Yang et al., 2016). The purpose of this chapter is to compare cyanobacteria amplified sequence variants (ASVs) of the V4 region of the16S rRNA gene (V4-16S rRNA region) obtained from each of the selected watersheds to view how dispersed cyanobacteria communities are across Canada, and to determine if they exhibit biogeographic trends between ecozones.

2.1.1 Forested Watersheds

Forested watersheds play a critical role in sustaining most of the world's freshwater supplies by acting as a natural water filtration system (Emelko et al., 2011; Furniss et al., 2010). The forest cover intercepts rainfall before reaching the ground (preventing and reducing soil

erosion) and in doing so, water is absorbed from the soil, removing stormwater pollutants, sediment, and other contaminants from the soil, groundwater, and runoff streams (Furniss et al., 2010). Forested watersheds are immensely effective at filtering water from the forest to downstream headwater sources where they are often used as recreational source waters or as the output drainage for water treatment facilities (WTF) (Furniss et al., 2010). Most usable source waters, like those significantly used for supplying communities with clean, safe, and potable drinking water, are supplied by upstream headwater sources contained within forested watersheds (Emelko et al., 2011). An issue with these environments, however, is that the water quality is susceptible and vulnerable to the impacts of climate change and natural disturbances (e.g., wildfires, insect infestations, disease outbreaks) (Emelko et al., 2011). These disturbances can drastically diminish the quantity and quality of source waters, leading to higher amounts of pollutants, nutrients, and trace metals to remain in the water where the substrates are rarely filtered out (Emelko et al., 2011).

Canada is home to a wide range of forested ecozones and watersheds (see **Figure 2.1**), representing approximately 10% of the world's forested landscapes (Coops et al., 2018). For this study, five major ecozones of Canada and the northern United States were chosen to study the impacts forest management practices have on drinking water quality and treatability. The ecozones and all selected sample sites (Atlantic Maritime [PLW], Boreal Plains [URSA], Boreal Shield [TLW], Montane Cordillera [GR], and Pacific Maritime [CRD]) were determined by collaborators from the *for*Water NSERC Network based on the climate-change exacerbated disturbances affecting each region and the subsequent impacts the disturbances have on source water protection and sustainable forested watershed management practices.

Each research site was chosen to highlight key disturbances brought on by climate change, with each watershed experiencing varying degrees and types of disturbances due to the distinct biomes. Freshwater lakes in the AM are susceptible to hurricanes and strong wind events that threaten the landscape of forested regions and thereby source watersheds, as well as high acidification of freshwater bodies due to the proximity to the Atlantic Ocean (Korosi

et al., 2013; Steenberg et al., 2013). Within the BP, source waters are affected by rising concentrations of carbon as the forests remove carbon from the atmospheres, effectively acting as large masses of carbon sinks (Kurz et al., 2013). The BS is an example of temperate forests experiencing a steady increase in air temperatures, longer and shifting growing seasons, and a growing number of days temperate lakes in the region remain ice-free (Webster et al., 2021). Influenced by the forest-dominated terrain and climate change, the MC has been experiencing dramatic increases in uncontrolled wildfires, thereby affecting sediment patterns, nutrient runoff into source waters, biodiversity, and insect infestations (Government of Canada, 2021). Finally, the PM has been experiencing growing tree mortality rates due to the native Mountain Pine Beetle (MPB) that attacks most *Pinus* tree species along western North America; because development of the MPB is in response to temperature, climate change has increased the development and life cycle of these pests, causing epidemics within forests and leading to dead, dry trees that ignite during wildfires (Mitton & Ferrenberg, 2012; Talucci & Krawchuck, 2019).



Figure 2.1 Map of the major forested ecozones and watersheds across Canada and their respective research platform sites

Adapted from the *for*Water Grant Proposal outlining the research platforms in the major forested regions in Canada (Emelko et al., 2016). Map generated using ArcGIS Pro [v. 2.6] (ESRI, 2020), based on the target ecozones outlined by the *for*Water NSERC network. The

five forested ecozones studied (left to right) are Pacific Maritime, Montane Cordillera, Boreal Plains, Boreal Shield, and Atlantic Maritime. The forested watersheds of interest, all highlighted in red, (left to right) are the Capital Regional District consisting of Comox Lake, Leech River, and Sooke Lake of the PM, Utikuma Region Study Area of the BP, Glenmore Reservoir of the MC, Turkey Lakes Watershed of the BS, and Pockwock Lake Watershed of the AM.

2.1.1.1 Atlantic Maritime

On the eastern coast of Canada, situated around 17-231 meters above sea level on un-weathered bedrock (e.g., metamorphic rocks including granite and slate) (Tropea et al., 2007) and directly on the Atlantic Ocean the AM has a cool, moist environment that receives approximately 1400mm of total annual precipitation (Steenberg et al., 2013; Wiken, 2015). Climate change-induced trends in the region have identified an increase in air and sea-surface temperatures, increase in ocean acidity, and a decrease in sea-ice cover (Lemmen et al., 2016). The changes in sea-ice cover and sea-level rise are major indicators for the rate of coastal erosions which is expected to continue increasing, and although physical structures have been installed to delay erosion rates, many coastal processes can be disrupted by artificial constructions (Lemmen et al., 2016).

On the southeast coast of Nova Scotia (Canada), hundreds of oligotrophic lakes (based on phosphorus measurements) are managed by the Halifax Region Municipality (HRM) – many of which have served as primary or backup drinking water sources and timber supplies for adjacent communities since as early as 1977 (Dunnington et al., 2018; Steenberg et al., 2013; Tropea et al., 2007). A forested watershed of interest, spanning over 50 km^2 in land area within the AM ecozone is the Pockwock Lake Watershed (44.7966816, -63.8442498) (Dunnington et al., 2018; Steenberg et al., 2013). The watersheds in this region experience a mean annual temperature of 5.8˚C and a growing season lasting on average about 196 days (Steenberg et al., 2013). Freshwater lakes in the Maritimes experience unique conditions due to the proximity to the marine environment which results in high dissolved organic carbon and

significant sulphate contributions in the water bodies (Korosi et al., 2013). Because of this, as of 2013 Atlantic Canada comprised 70% of acidic lakes throughout Canada, representing the greatest proportion of acid deposition in the country (Korosi et al., 2013).

The watersheds managed by the HRM have been of interest to better understand the range of impacts forested ecosystems may experience in response to anthropogenic and/or natural disturbances (Steenberg et al., 2013). Moreover, the research historically conducted in HRM watersheds focused on reversing the gap in ecological data from pre-disturbance periods by using paleolimnology to extrapolate if environmental changes to the landscape could be more attributed to natural variation or from direct observations of anthropogenic impacts (Tropea et al., 2007). The proximity to the ocean puts the AM at higher risks for natural climate disturbances (e.g., insects, disease, windthrow) (Steenberg et al., 2013). Of particular significance to the AM are the higher chances for hurricanes and severe windstorms associated with climate change (Steenberg et al., 2013) which threaten to disrupt natural ecosystems. Of most concern, however, in the AM watersheds are the impacts acid deposition have on source waters (Dunnington et al., 2018; Korosi et al., 2013; Tropea et al., 2007). The acidification of lakes in the AM ecozone have been caused by low pH buffering capacity due to acid deposition from climate disturbances, road salt applications increasing salinity levels, and watershed deforestations ((Dunnington et al., 2018; Tropea et al., 2007). Eutrophication has increased in HRM watersheds due to inputs from stormwater, developments of residential lawns and septic systems that caused an influx of phosphorus into waterways (Dunnington et al., 2018). Other than socio-economic concerns regarding treatment of impacted water sources, acidic lakes are also a cause for concerns from aluminum, sulphate anions, and other potentially toxigenic metals that can be a health hazard if consumed in drinking water (Tropea et al., 2007).

Being that HAB monitoring is not federally established in Canada, each province has its own authority for how to handle algal blooms and most provinces choose to handle HABs on a case-by-case basis, often only conducting water quality testing when blooms are physically reported (Rashidi et al., 2021). In recent years 19 bodies of water were reported in

the AM for having cyanobacterial blooms, with 14 of those cHABs occurring only in the last five years (Valadez-Cano et al., 2022), and in 2018-2019, 17 water bodies exceeded the federal guidelines for cyanobacterial blooms (Rashidi et al., 2021). Although the AM has seen an increase in cyanobacterial blooms, there is a clear lack of recorded data, especially regarding toxigenic properties and species. This is largely due to the inconsistencies in taxonomic identification methods because most AM provinces still depend on morphological and microscopic techniques in identifying cyanobacteria (Valadez-Cano et al., 2022), however these methods cannot differentiate between toxic or toxin-producing species, for which molecular methods are most accurate. The AM is also observing seasonal cyanobacterial blooms for which they have been reported annually in New Brunswick for the past six years in two lakes, one oligotrophic and the other mesotrophic (Valadez-Cano et al., 2022). In these same waters, there have also been incidents of dogs ingesting waters infected with the algal scums containing the cyanotoxin anatoxin-a (AXT-a), for which at least three cases have proved to be fatal (Valadez-Cano et al., 2022). While Nova Scotia monitors cyanobacteria more closely than its neighbouring provinces, HRM and Halifax Water still only monitor beaches that have lifeguard supervision and do not release any annual public reports describing HAB occurrences throughout the province (Rashidi et al., 2021). In general, reports of cyanobacterial blooms are on the rise in the AM, however, monitoring methods are not yet consistent and reliable enough to effectively prevent the contamination of cHABs and cyanotoxins in drinking or recreational water sources.

2.1.1.2 Boreal Plains

Of significance in the BP, north of Utikuma Lake (northern Alberta) (55.8525760, -115.3992159), approximately 400 km north of Edmonton, is the URSA (Devito et al., 2016). This study area was initiated with the intention to better grasp and comprehend the impacts anthropogenic land use and natural disturbances have on the quantity and quality of source water in the Western Boreal Plains (WBP) ecozone (Devito et al., 2016), especially regarding the consumption of carbon from the atmosphere (Kurz et al., 2013). The research at URSA

spanned several multi-year projects that covered both wet and dry seasons to have a holistic view of the impact's disturbances have on landscape ecology and the environment as the source waters in the region act as carbon sinks (Devito et al., 2016; Kurz et al., 2013). At an elevation of 640-695 m, URSA has a drier climate with cold winters and moderately warm summers, encountering a total mean of 483 mm of precipitation per year, and a mean annual temperature of 1.2 ˚C (Devito et al., 2016; Wiken, 2015). The URSA is comprised of low-relief, poorly defined and drained topographic watersheds, filled with central mixed-wood forests, which cover three dominant superficial geological landforms, all draining into the Wabasca and Peace River drainage outputs (Devito et al., 2016; Prepas et al., 2001).

Currently, the largest natural disturbance for the Western Boreal Forest (WBF) in the BP are wildfires, and due to climate change, fires are predicted to continue to increase in frequency, range, and severity (Coops et al., 2018; Devito et al., 2016). The loss of forestry in forested watersheds removes the natural filtration for water sources and can lead to runoff filled with high concentrations of pollutants, nutrients, and trace metals – all of which can contaminate drinking water supplies (Emelko et al., 2011). Determining the impacts of disturbances can be relatively difficult in the BP because variations in superficial geology, connectivity between ground and surface water, and the long residence times in ponds all combine to mask any initial impacts (Devito et al., 2016). Forest fires across the BP have been of great concern because of their negative impacts and the greater frequency to which they are occurring is only increasing with climate change. Declining water quality is among one of the major impacts of forest fires due to the increase in nitrogen inputs that can take five to six years to recover back to base levels, and up to twenty times higher levels of phosphorus within the water column and entering from transported sediments – both of which the recovery rate has not yet been determined (Silins et al., 2014; 2016). The fires also result in elevated concentrations of trace metals in the water quality in nearby ecozones (e.g., lead, nickel, cobalt, manganese, aluminum, molybdenum) which can be detrimental for drinking water treatability and have shown to remain elevated even five years post fire event (Silins et al., 2014; 2016). Regions such as URSA are of great importance because the threat of forest disturbances is only

becoming more prevalent in the BP, and the risks to drinking water quality and treatment are becoming increasingly more complicated as source waters continue to be impacted.

The shifts in source water quality from fire disturbances, observed in nearby ecozones (e.g., Montane Cordillera), have caused a rise in cyanobacterial proliferation in response to changes in light availability, increasing water temperatures, and an inflow of nutrients associated with increased precipitation (Silins et al., 2014; 2016; Wagner et al., 2014; Williams et al., 2019). Thick algal scums are often observed in watersheds that have experienced recent fires or other disturbances (e.g., Fort McMurray), and the rise of algae subsequently attracts invertebrates (such as the Mountain Pine Beetle as seen in the PM); these changes in organism communities shifts entire aquatic habitats, community compositions, and trophic structures (e.g., trout in disturbed lakes are found to be larger than average and grow at much faster rates) (Silins et al., 2014; 2016; Talucci & Krawchuck, 2019; Wagner et al., 2014; Williams et al., 2019). Historically, algal communities in lakes from BP have mainly been characterized by cyanobacteria, heavily dominated by N_2-fixing species, and typically seen in highest abundance during summer months (Prepas et al., 2001). Currently, a monitoring program in Alberta (the province largely governing the BP) that tests for cyanobacteria and microcystin presence in 40 public and recreational sites during the summer reported 25 and 21 cyanobacterial HABs across the province in 2018 and 2019, respectively (Rashidi et al., 2021). While the short-term impacts of forest disturbances on algal and cyanobacterial communities is well understood, the knowledge regarding long-term legacy effects of fire disturbances is much needed and is one of the emerging areas of research in the Boreal Plains.

2.1.1.3 Boreal Shield

As the largest ecozone in Canada, the Boreal Shield's biome consists of short, warm summers, long, cold winters, and widespread rainfall between 400-1000mm of total annual precipitation which are often influenced by masses of cold air from Hudson Bay (Wiken, 2015). A major research site for the BS, north of Sault St Marie (northern Ontario), the Turkey Lakes Watershed (TLW) Study (47.0467651, -84.4236758), was established in 1979 by

Environment Canada as a study site for the effects of acid rain on terrestrial and aquatic ecosystems in the region (Government of Canada, 2019). Located on the Canadian shield at a total relief of 300 m (elevation between 330 to 625 m) from the highest point atop Batchawana mountain, TLW is comprised almost completely of Precambrian silicate greenstone (with some felsic igneous present as well) (Buttle et al., 2019; Jeffries et al., 1988; Leach et al., 2020; Webster et al., 2021).

With over 37 years of environmental monitoring, TLW is one of Canada's longest running watershed-based ecosystem studies (Webster et al., 2021). The mean annual precipitation across that period was 1203 mm, with a third of that precipitating as snow, and the mean annual temperature was 4.5 ˚C (Buttle et al., 2019; Webster et al., 2021). Spanning 10.5 km^2, TLW contains four lakes (Batchawana Lake, Big Turkey Lake, Little Turkey Lake, Wishart Lake), drains from Eastern temperate mixed hardwood forestry, and is now the focus for more environmental issues such as toxic contaminants and fire disturbance (Government of Canada, 2019; Webster et al., 2021). The research site was originally established to better understand the impacts acid rain deposition has on aquatic and terrestrial habitats. The site has been utilized in collaboration with Environment and Climate Change Canada (ECCC), the Department of Fisheries and Oceans Canada (DFO), Natural Resources Canada (NRCAN), and several Canadian universities (Webster et al., 2021). The BS has and continues to be impacted by climate change-associated risks of rising temperatures, forest fires, and acid rain (Wiken, 2015), mainly due to the high relief of the watershed and its position relative to Lake Superior (Lindsay et al., 2004). Other than logging activities in the early 1950s where sugar maple, yellow birch, and eastern white pine trees were logged from areas of the watershed (Jeffries et al., 1988; Webster et al., 2021) and a controlled forest harvest impact study in 1997, TLW has been relatively undisturbed by anthropogenic factors (Lindsay et al., 2004). Other than the occasional small forest fire or insect outbreak, there have been no natural disturbances to the watershed (Webster et al., 2021), and along with the lack of anthropogenic disturbances, TLW is an ideal ecosystem study site to observe the effects of natural environmental variability over a prolonged period.

The Turkey Lakes Watershed has been the focus of many environmental-based studies, especially ones focused on the forestry, atmospheric chemistry, drainage and hydrological systems, and whole-ecosystem analyses (Webster et al., 2021), however, little focus has been put on microbial communities within these source waters. In Ontario, one of the provinces making up a large portion of the BS, the Ministry of Environment, Conservation, and Parks (MECP) have confirmed that there has been a significant increase in the number of HABs from 1994 to 2009, and specifically an increase in cyanobacteria taxa being present (Favot et al., 2019). Over this same period, the seasonal length where blooms are reported has been observed to have increased, specifically that cyanobacterial blooms reports are seen occurring much later into the fall months compared to when cHABs were prevalent a decade ago (Favot et al., 2019), suggesting the impacts of climate change and natural or anthropogenic disturbances.

One of the few studies on cyanobacteria in TLW, conducted by Cameron (2021), explored how the composition of cyanobacterial communities changed across the vertical distribution in the water column, changing spatiotemporal shifts, as well as seasonally in the oligotrophic lakes of TLW (Big and Little Turkey Lake, Wishart Lake). This study investigated how cyanobacteria are present in undisturbed oligotrophic lakes throughout the water column (although visual blooms are not present) and seasonally throughout the year (Cameron, 2021). Environmental samples from Cameron's study represent the Boreal Shield sequences used in the phylogenetic tree in **Figure 2.2** and are the focus for metagenomic analyses in Chapter 3 of this thesis.

2.1.1.4 Montane Cordillera

The Montane Cordillera, located in mid-western Canada, is one of Canada's most complex ecozones as it has an exceptionally diverse topography and climate over a relatively short area (Government of Canada, 2021). The very mountainous terrain, running entirely north to south of the ecozone, results in long, cold winters, and short, moderately warm summers (Government of Canada, 2021; Wiken, 2015). Because of the size and terrain variability within this ecozone, the average precipitation per year ranges between 500 mm in

the interior plains (considered Canada's only desert), and up to 1200 mm along the Alberta-British Columbia border (Government of Canada, 2021; Wiken, 2015). There are over 11,000 lakes and seven major river systems across the MC (Government of Canada, 2021).

Located along Elbow and Bow River in Calgary, Alberta, the Glenmore Reservoir (50.9881416, -114.1125211) is an artificial oligotrophic water body that was constructed to provide ample supplies of clean, safe drinking water to 50% of Calgary (City of Calgary, n.d.; Watson, 2004). In addition to being the source for half of Calgary's water supply, the GR also acts as a recreational water body and is constantly being monitored for risk factors to public safety, often including surveying for cHAB presence (City of Calgary, n.d.). It has undergone rapid developments in the urban and environmental sectors since its construction and is known to typically support some assemblage of algal communities (Watson, 2004). The Glenmore Reservoir has been a part of a collaborative research program for the last decade in which physicochemical and biological variables are continually being monitored (Watson et al., 2001).

The Montane Cordillera is a heavily fire dominated ecosystem with 65% of the terrain being covered by forestry (Government of Canada, 2021). Due to the fire suppression methods frequently employed in this ecozone, the environment in the MC has become more disposed to insect infestations and wildfires of higher magnitudes (Government of Canada, 2021). In lower elevation areas this ecozone experiences landscape disruptions due to wildfires (Silins et al., 2014; 2016; Wiken, 2015), and based on previous research on wildfire trends, the MC is expected to see significant increases in fire disturbances (Coops et al., 2018). Further impacts to the MC involve the drying of entire creeks in the Okanagan region due to the combination of fire disturbance and climate change influences, as well as interruptions from dam failures (Silins et al., 2014; 2016; Tannant & Skermer, 2013). Changing hydrological and climate conditions is threatening the success of many of the dams in the region, which should they completely fail, could seriously impact watershed level systems (most of which are not considered during dam construction) or in the least could affect the rate of water discharge

entering and exiting reservoirs and source waters due to the erodibility of the earth and sediment around the dam (Tannant & Skermer, 2013). While there are many layers to the impacts occurring across the Montane Cordillera, there appears to be misguided communication between sectors and a lack of source water protection.

Although there is a great number of lake and river systems throughout the Montane Cordillera, the composition of microbial life is an ongoing study. While it is understood that fire disturbances have the potential to impact aquatic and microbial communities, this is still an area of ongoing research. The fire disturbances occurring upstream of source waters result in fine sediments, contaminants, and any other material present in the water to remain in riverbeds longer than usual due to bio-stabilization effects influenced by associated disturbances (Silins et al., 2014; 2016; Stone et al., 2011; Wagner et al., 2015; Williams et al., 2019). Following the accumulation of material during bio-stabilization events, more wildfire-associated biofilm communities are being identified as being indicative for having a predisposition for T/O events that is highly correlated with the downstream transport of microorganisms from sediment turnover post fire disturbances (Silins et al., 2014; 2016; Stone et al., 2011; Wagner et al., 2015; Williams et al., 2019). As for the Glenmore Reservoir, despite having low nutrient concentrations and high flushing rates, this reservoir sometimes experiences random outbreaks of algal blooms with species responsible for taste and odour events (mainly caused by chrysophytes) (Watson et al., 2001; Watson, 2004). Because of this, there has been sporadic reports of fish-like odours throughout the spring and winter seasons over the last ten years (Watson, 2004). While the typical assemblage of algal communities in Glenmore Reservoir has consisted of diatoms and small flagellates, the increase in fish-like odours suggests there may be seasonal and inter-annual changes occurring that are not yet understood (Watson, 2004). Additionally, taste and odour events are not uncommon to Glenmore Reservoir even under ice cover, where some events have lasted up to two months (Watson et al., 2001). As fire disturbances continue to increase in severity and frequency, the shift in nutrient and sediment composition is sure to alter the composition of microbial and algal communities.

2.1.1.5 Pacific Maritime

On the west coast of Canada, the Pacific Maritime spans over 27,000 km of coastline adjacent to large mountains and the Pacific Ocean (Lemmen et al., 2016). The proximity to the ocean and mountainous topography creates very wet winters, dry summers, and mild temperatures with about 600-3500mm of annual precipitation, mainly falling as rain (opposed to snowfall) (Lemmen et al., 2016; Wiken, 2015). Unique to the region, the Pacific Maritime tends to have quite stable and predictable weather throughout the year, where mean monthly temperatures generally range from 4-6 ˚C during the winter, and only increasing to about 12-18 ˚C during the summer (Wiken, 2015), not usually differing more than 15 ˚C between summer and winter months (Lemmen et al., 2016). Over 50% of the Pacific Maritime is forested land, including many old growth forests, which largely consists of (over 90%) coniferous trees (Government of Canada, 2021). Because of the location and environment, much of the economy in the Pacific Maritime is linked to the ecosystems, making the impacts of climate change (such as the decline in forestry and fisheries) a larger threat.

Three interconnected watersheds on Vancouver Island, British Columbia (BC) – Comox Lake (49.6184857, -125.1695425), Sooke Lake (48.55061123, -123.6956512), and Leech River (48.52940275, -123.786897) – all sectors of the Capital Regional District (CRD), are important drinking water supplies to the nearby communities. Comox Lake is 97% natural area that covers approximately 2100 hectares of surface area, large enough to provide drinking and recreational source waters to multiple Comox Valley communities (>38,000 people) near Cumberland, BC (Capital Regional District, n.d.; Chandran & Mazumder, 2015; Epps & Phippen, 2011). As populations in the region continue to climb, the Comox Lake Watershed is a key resource to members of the communities and as such, any issues to the watershed present a dangerous situation should threats to clean drinking water arise.

Although the climate has historically been relatively stable in the Pacific Maritime, the effects of climate change have been causing unexpected rise in temperatures leading to more frequent wildfires, rises in sea levels, and warming waters – all of which have cascading

impacts on the rest of the ecosystem (Coops et al., 2018; Lemmen et al., 2016). Rising temperatures have led to an increase in the native MPB which attacks trees along the northwest coast of North America, significantly killing forested regions and impacting the success of wildfires as there is more dried trees to act as kindling (Mitton and Ferrenberg, 2012; Talucci and Krawchuck, 2019). Sea-level rises can result in storm-surge floods which damage infrastructures, cause upwelling within the water column and are generally a danger to communities living near or on the coasts (Lemmen et al., 2016); a project was initiated in 2012 by the Capital Regional District called the Coastal Sea-Level Rise Risk Assessment to study the early consequences of sea-level rise on the immediate ecosystems (Lemmen et al., 2016). As water temperatures increase as well, ecosystems are witnessing total shifts of organism communities as more organisms move north to cooler waters in response to the warming of their current habitats possibly impacting reproductive and offspring success (Lemmen et al., 2016).

Although climate change may not be influencing mean precipitation levels in the Pacific Maritime, the patterns for when it precipitates, in what form (rain or snow), and the amount (how much during a given precipitation event) are significant as it impacts water availability (Lemmen et al., 2016). Water availability is impacted because during high precipitation events floods are a greater risk, and during decreased precipitation seasons rivers run low causing the water temperatures in drier rivers to run higher than average (Lemmen et al., 2016). Dry seasons are the greatest threat to the CRD as summer water shortages related to droughts are often seen in freshwater reserves, significant enough to instigate a Climate Action Program in 2008 that worked on mitigation practices, policies, and environmental and infrastructure planning all to fight the negative impacts of climate change on CRD's watersheds (Lemmen et al., 2016).

The provincial guidelines for HABs in British Columbia are explained in the report "Decision Protocols for Cyanobacterial Toxins in British Columbia Drinking Water and Recreational Water" (Rashidi et al., 2021). The protocol in BC suggests that only vulnerable

water bodies (i.e., ones with a history or susceptibility for blooms) should be tested regularly, and that any findings related to HABs be disclosed to the public (Rashidi et al., 2021). Being that a large portion of the Pacific Maritime ecozone is along an ocean coast, much of the research regarding microbial and algal communities focuses on marine species such as shellfish and marine algae, and not as much on the composition of communities within freshwaters. Much of algal bloom monitoring in the Pacific Maritime emphasizes on the impacts HABs have on the shellfish industry as it is major economic revenue for the province (McIntyre et al., 2013). Information on freshwater algal blooms tends to appear when unusual (i.e., outside of expected seasons) algal blooms occur, such as a bloom outbreak in an interior lake that appeared as late as December in 2013 (Pick, 2016). Of the tests done looking at phytoplankton communities, some species of algae and cyanobacteria have been recorded, but within the usual range for oligotrophic systems and is thus not considered a cause for concern (Epps & Phippen, 2011). While not specifically algae, Comox Lake has shown to test regularly for *Escherichia coli* in the drinking water intake in a study by Chandran and Mazumder (2015) that ran for three years in which many months over the study period indicated *E. coli* concentrations exceeded the limits set by the province for raw water sources. Although cyanobacterial blooms are infrequent, if not rare, within the Comox Lake Watershed, the impacts of climate change can shift phytoplankton communities and should be continually monitored.

2.1.2 16S rRNA Gene Sequencing

The traditional classification systems of cyanobacteria based on morphological similarities has been slowly fading out of practice as phenotypic characteristics are increasingly being observed to be shared in distantly related taxa (Clarridge, 2004; Strunecký et al., 2023; Walter et al., 2017). The introduction of classifying taxonomy based on comparing a stable part of the genetic code changed the course of bacterial identification history (Clarridge, 2004). The new standard of identification and taxonomy based on genomic sequencing was supported by the high conservation of the 16S rRNA gene which is believed

to suggest its critical role in cell function (Clarridge, 2004), and the assumption that there is no better description for an organism than its genome (Walter et al., 2017). Individual genes are not considered accurate markers because not all bacteria contain the same genes, and because genes that are not crucial to the viability of cells are much more likely to undergo mutations (Clarridge, 2004; Tsukuda et al., 2017). After its introduction to the scientific community, the 16S rRNA gene quickly became the most used genetic marker for bacterial profiling because of its identification accuracy, and its evolutionary neutrality as it is assumed to rarely experience horizontal gene transfers (HGT) and therefore serves as a stable marker gene (Tsukuda et al., 2017; Yang et al., 2016).

Amplification and sequencing of the 16S rRNA gene is one of the most common practices among the scientific community because it is a highly conserved phylogenetic marker present among all bacteria and can identify isolates down to the genus (97% similarity) and species (99% similarity) levels (Escapa et al., 2020; Janda & Abbott, 2007; Walter et al., 2017; Winand et al., 2019). The 16S rRNA gene is approximately 1600 bp in length and consists of nine (hyper) variable regions (V1-V9), with nine conserved regions interspersed throughout the gene (Johnson et al., 2019; Winand et al., 2019; Yang et al., 2016). Although sequencing the entire gene provides better taxonomic resolution, most studies only target a small subset of the gene due to cost restrictions, and limits with technological advances (Johnson et al., 2019; Kim et al., 2011). Depending on the scope of the research or the type of sequencing platforms used, smaller regions of the 16S rRNA gene may be better suited than others. Specifically, the fourth variable region of the gene, hereby referred to as the V4 region (~400 bp long) has been observed to characterize bacteria well, best differentiating genus and species within phyla such as the Cyanobacteria, Firmicutes, Proteobacteria, Planctomycetes, and Tenericutes (Zhang et al., 2018). As no true universal primer set is available for the entirety of the 16S rRNA gene, specific primers for the V4-16S rRNA gene region are the most useful in classifying bacterial communities, which, along with downstream bioinformatic analysis, can illuminate important information as to microbial community compositions and identifying taxa.

2.2 Objectives

Currently, there is little research available in literature on the distribution of cyanobacteria across Canada that focuses on the biogeographical distribution of species to determine risk factors. The research in this chapter is novel in that aims to inspect the taxonomy and biogeography of cyanobacterial community compositions throughout five major forested ecozones across the nation. As each ecozone has its own unique microbiome, it is possible that each region may contribute to (or inhibit) the proliferation of cHABs and be home to unique cyanobacterial communities. Identifying the taxonomy and biogeography of cyanobacteria across source waters in Canada can help mitigate monitoring protocol by highlighting significant species in each forested ecozone (or the lack thereof) and signify the potential threats cyanobacteria pose on individual watersheds.

This research utilized V4-16S rRNA region amplicon sequenced water samples obtained and filtered by various *for*Water HQP from watersheds in the Atlantic Maritime, Boreal Plains, Boreal Shield, Montane Cordillera, and Pacific Maritime forested ecozones – all of which are regions experiencing some type of significant alterations to the ecosystem due to anthropogenic or climate-change-exacerbated landscape disturbances. Findings from this research aims to examine the diversity and biogeographic trends of cyanobacterial species across a large land mass as Canada.

2.3 Methods

2.3.1 Sample Collection and DNA Extraction

Numerous samples from each of the respective watersheds have been obtained and analyzed by various *for*Water highly qualified personnel (HQP). Outlined in **Table 2.1** are the samples whose taxonomy has been assigned using V4-16S rRNA amplicon gene sequencing and workflow pipelines, which classify ASVs against genomic libraries, and consolidate with appropriate phylogenies. Of the environmental samples sequenced, ones that had little to no composition of cyanobacteria have been removed and were not included in the analysis. The

DNA extraction and sequencing for this objective were conducted by previous members of the Müller Lab at the University of Waterloo and further information regarding methodology and supplementary information can be found in their corresponding theses; AM and PM samples were sequenced by Timothy Shardlow (2021), BS samples were obtained, filtered, and sequenced by Ellen Cameron (2021), BP samples were obtained, filtered, and sequenced by Arnold Kwok (unpublished), and the samples from the MC were sequenced by Ghadeer Aljoudi (2021) and Timothy Shardlow (2021).

Sample collection in the AM ecozone were taken by members of the Johnson Lab from Dalhousie University from Pockwock Lake Watershed (PLW) between June to October 2019 in collaboration with Halifax Water (Shardlow, 2021). From the BP ecozone, samples were collected from 34 ponds in the URSA between July 4th to 9th, 2018 (Kwok, unpublished). Samples from the BS ecozone were collected frequently from the TLW between 2018-2020 from the four interconnected lake network made up of Big and Little Turkey Lake, Wishart Lake, and Batchawana Lake (Cameron, 2021). From Glenmore Reservoir in the MC, sample sets were collected on July 27, 2017, and again on September 21-23, 2020, both during taste and odour (T/O) events from four sections that make up the Glenmore Reservoir – Weaselhead, Heritage Cove, Mid-Lake, and Head Pond (Aljoudi, 2021; Shardlow, 2021). All samples from the PM (CRD) were collected by previous *for*Water HQP from multiple sites within Comox Lake, Sooke Lake, and Leech River watersheds (Shardlow, 2021). For more information regarding sample collection, DNA extraction, and V4-16S rRNA gene sequencing, please refer to the respective theses.

2.3.2 V4-16S rRNA Gene Sequencing and Analysis

DNA extractions using the DNeasy PowerSoil Kit © (QIAGEN Inc, Venlo, Netherlands) and primary analysis conducted by HQP followed all manufacturer's protocol and quantified using the NanoDrop 2000 spectrophotometer (Thermo Scientific, DE, USA) when permissible. The extracted microbial DNA (~20 µL) were sent to a commercial environmental sequencing company (Metagenom Bio Inc., Waterloo, Ontario) to conduct 16S

rRNA gene amplicon sequencing targeting the V4 gene region using specified primers and the Illumina MiSeq 2x250 platform (Illumina Inc., San Diego, California). Bioinformatic processing employed the software QIIME2 © (v.2022.8; Bolyen et al., 2019) to assign taxonomy, followed by downstream analysis, trimming, and filtering out sequences not assigned to cyanobacteria phylum.

2.3.3 Phylogenetic Analysis

Environmental sample sequences from each forested ecozone watershed (**Table 2.1**) were compiled, aligned, and trimmed using the computational tool Geneious © (v. 2022.1.1) followed by phylogenetic tree construction using the software Molecular Evolutionary Genetics Analysis Version X (MEGA v. X ©; Kumar et al., 2018) . Before aligning, all V4-16S rRNA region gene sequences were first searched against the NCBI (National Center for Biotechnology Information) database using BLAST (Basic Local Alignment Search Tool) to check the quality and taxonomy assignment of the sequences. Sequences that were not representative of cyanobacteria or did not have high identity and query support (> 70%) were excluded from the final alignments to maintain high confidence in the acquired sequences.

A Multiple Sequence Alignment (MSA) constructed using Geneious © aligned all remaining sequences against each other based on nucleotide similarity to study the conservation and/or variability of the sequences. Samples with a significant portion of missing sequence data were removed entirely, and all samples were trimmed to similar base pair lengths to create a uniform alignment. From the MSA, a comprehensive phylogenetic tree clustered the samples for each gene by sequence similarity against documented cyanobacterial reference sequences for which the species or genus is confirmed. Reference sequences were obtained using the NCBI GenBank database, specifically targeting the V4 conserved region within cyanobacterial phyla. All aligned V4-16S rRNA region sequences were exported into MEGA v. X where phylogenetic analysis constructed Maximum Likelihood (ML) trees using the best model chosen by the software resulting in the K2 + G + I (Kimura 2-parameter and Gamma distributed with invariant sites; Kimura, 1980) algorithm, and the Nearest-Neighbor-

Interchange (NNI) ML heuristic method. The bootstrap support was set to 1000 replicates, indicating the confidence value for the number of times the same branches were observed throughout the tree after the sequences underwent resampling with random data sets 1000 times. The site coverage cut-off was set to 95%, the tree was rooted with standard bacterial outgroups (e.g., *Escherichia coli*), and the final alignment and tree included 522 total sequences (399 ASVs, 8 outgroups, and 115 reference sequences) all measuring up to 259 base pairs in length.

2.4 Results

From the 75 water samples compiled across the five forested watersheds (see **Fig 2.1**), a total of 399 uncultured environmental V4-16S rRNA region cyanobacterial sequences were obtained. A list of the sampling sites and dates associated with these amplicon sequence variants (ASVs) are in Appendix A (Table A1). These sequences were included in a phylogenetic analysis along with eight outgroup sequences and 115 sequences obtained from GenBank that were from peer-reviewed publications (**Table A2**). The number of sequences for each ecozone are specified in **Table 2.1** below; within the entire tree 19 ASVs are from AM (labelled ASV#), 158 are from BP (labelled CYAN#), 123 are from BS (labelled ASV#), 21 are from MC (labelled Cyano#), and 78 are from PM (labelled ASV#).

The final full maximum likelihood phylogenetic tree (K2+I+G, NNI model) (**Figure 2.2,** below) is too large to fit onto one page and as such spans over eight pages, labeled A-H, for clarity purposes. Starting at the very top of the tree (page A) is the most derived clade, and the top of each page directly corresponds to the very bottom of the previous page (e.g., the clade beginning at the top of page B would connect directly after the clade at very end of the tree on page A). The orders labelled on the right-hand side of the tree correspond to the clustered clades defined by reference sequences and are not strictly attributed to a solid beginning or ending point as reference sequences may span multiple clades. The ASVs within the phylogenetic tree are labeled via coloured shading according to the forested ecozones they

were collected from. A legend, present on each page, depicts AM sequences shaded in red, BS sequences shaded in purple, BP shaded in blue, MC in brown, and PM in yellow.

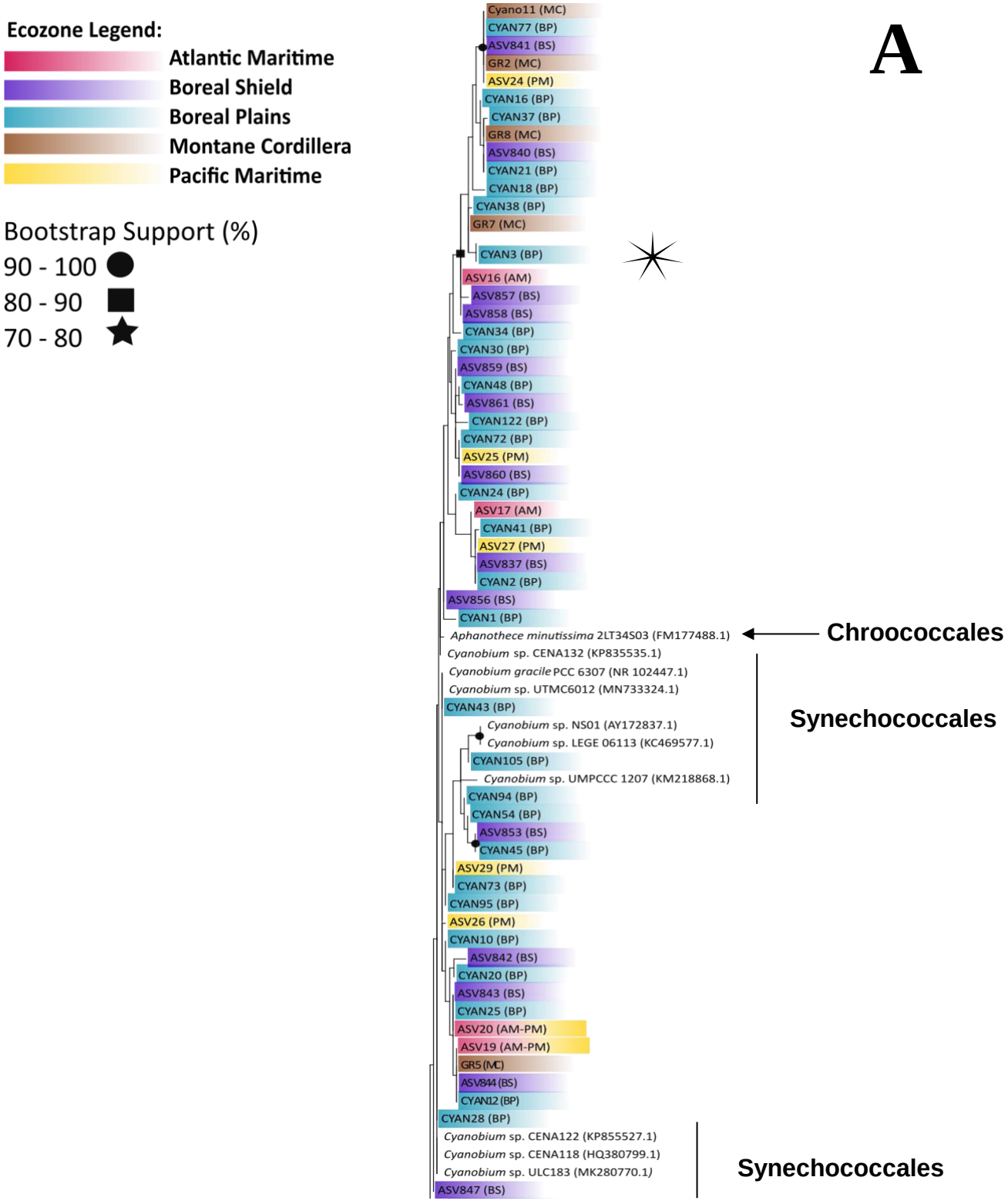
Bootstrap support confidence intervals (1000 replicates) are labeled on corresponding clades with specific symbols to represent the percentage that clade appeared in the bootstrapping analysis. Bootstrap values are represented on each node and refer to the proportion of replicates (in this case out of 1000) that supports the monophyly of each clade (Soltis & Soltis, 2003). A good or typical bootstrap value of at least 70% is often considered a good indicator of support (Soltis & Soltis, 2003), hence, only well supported clades (>70%) are labelled in **Figure 2.2.** Nodes with a star represent bootstrap support of 70-80%, a square represents 80-90% support, and a clade signified by a circle represents the highest bootstrap confidence of 90-100% phylogeny estimation. Branches with <70% support were left unlabelled. The results of the phylogenetic tree can be regarded with high confidence due to the 1000 bootstrap support because statistically, the branches observed with high bootstrap values were repeatedly observed in the same location on the tree and among the same clades against random data resampling; thereby bootstrap values refer to the likelihood that a particular phylogeny would be estimated (Soltis & Soltis, 2003).

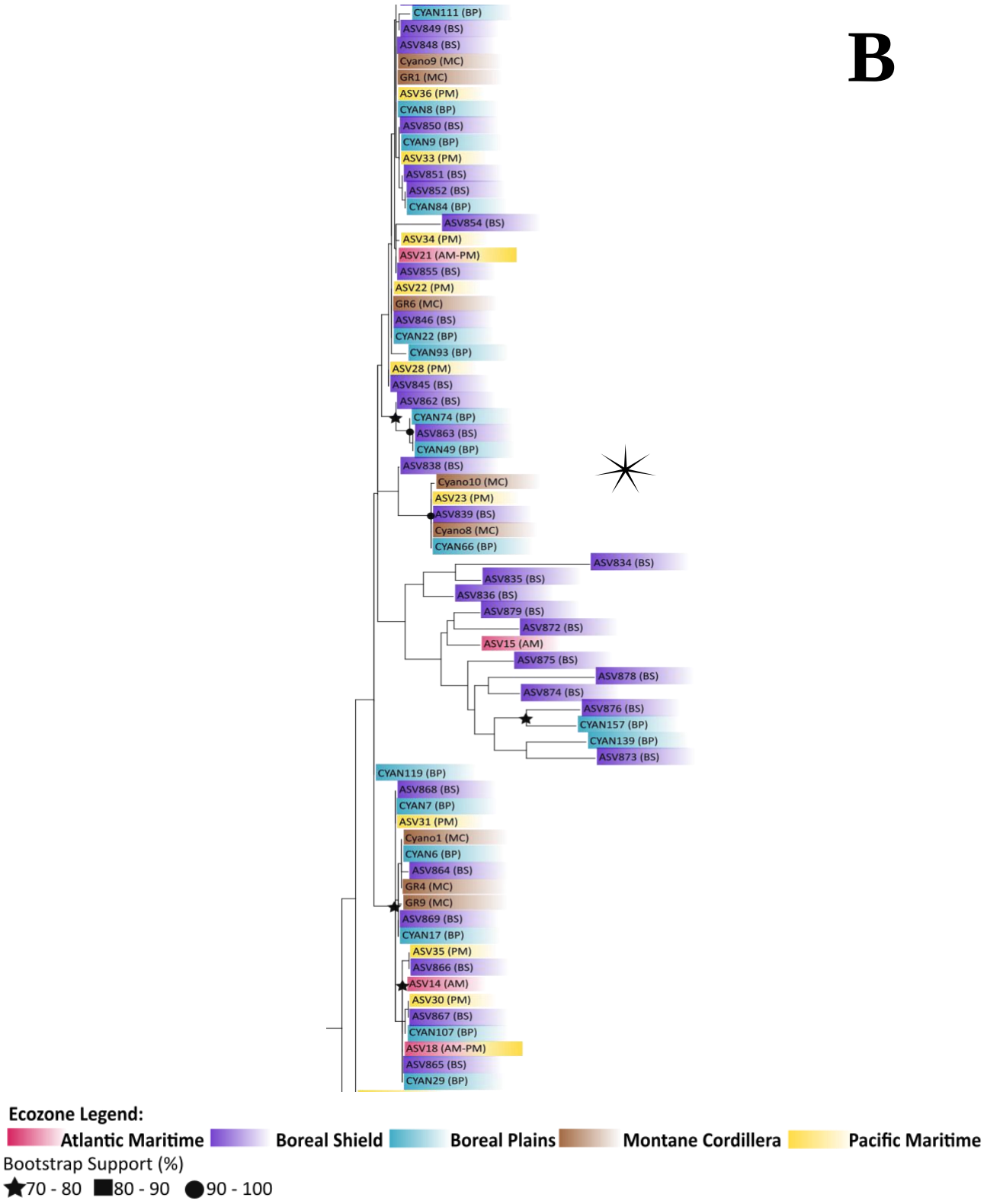
Table 2.1 Number of unique cyanobacterial ASVs obtained from community V4-16S rRNA gene profiling from each ecozone included for phylogenetic analysis

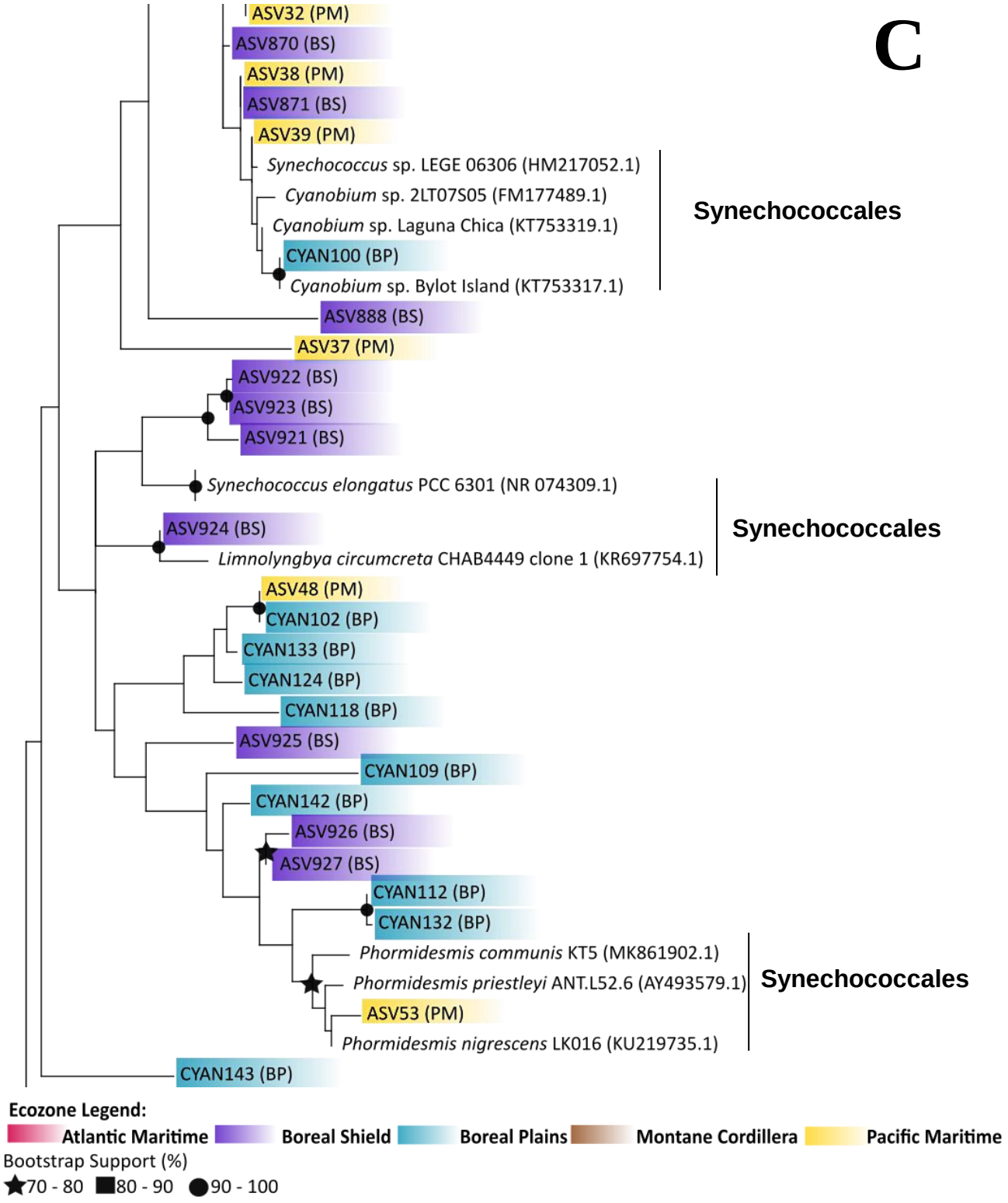
Ecozone	Watershed	Cyanobacterial ASVs
Atlantic Maritime	Pockwock Lake Watershed	19
Boreal Plains	Utikuma Region Study Area (URSA)	158
Boreal Shield	Turkey Lakes Watershed	123
Montane Cordillera	Glenmore Reservoir	21
Pacific Maritime	Capital Regional District	78

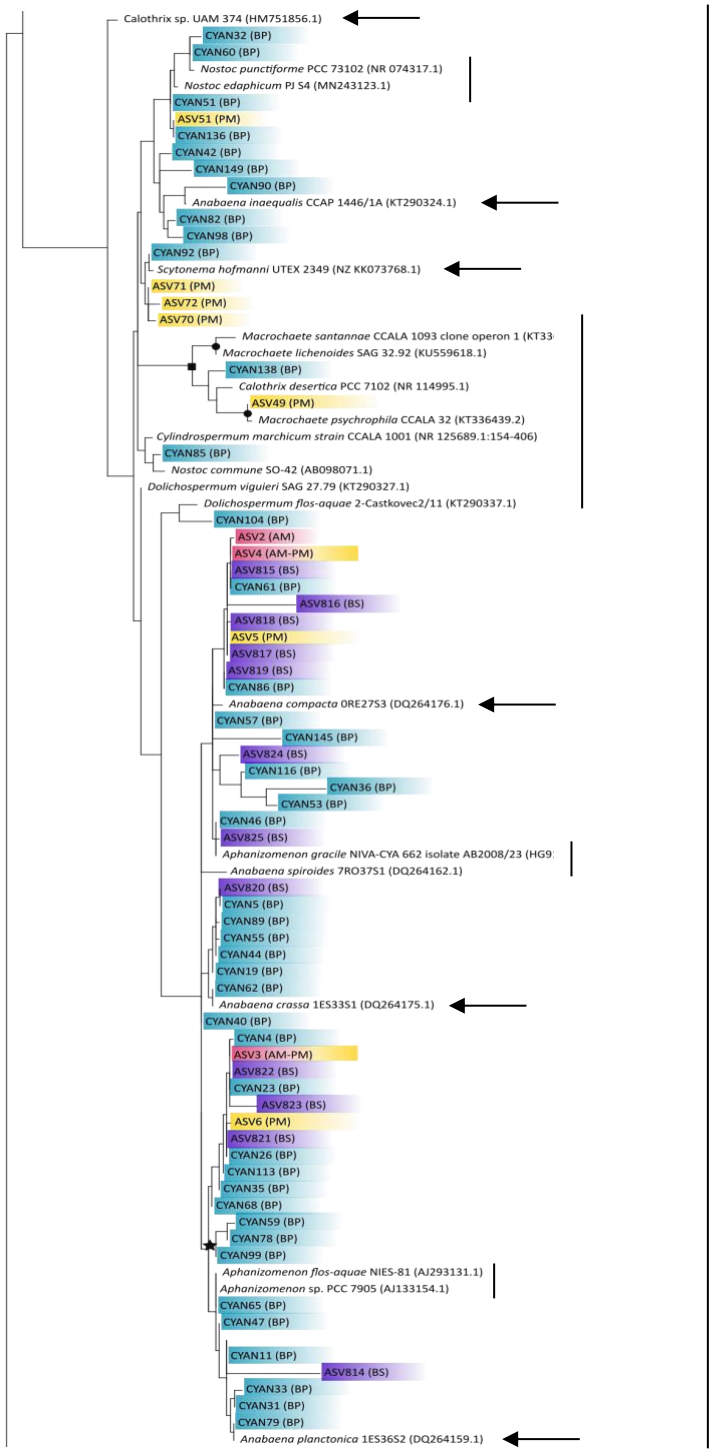
Total table containing detailed information about sampling dates and specific sites is indicated in Appendix A, Table A1.

As taxonomy assignments are often a source of controversy because nomenclature is continually revised, it can be difficult to differentiate between designations of cyanobacteria taxonomy, especially when they are under review; because of this, some species can be classified into different orders depending on the database used. In addition, it is recognized that taxonomists do not have a true estimate of all species on earth, and many are difficult to culture and grow within laboratories (Strunecký et al., 2023). Therefore, to avoid contradicting classifications, all the taxonomy for the orders indicated in the phylogenetic tree below follow the literature published within the AlgaeBase database as it focuses solely on phycological organisms, which includes approximately 169,761 species names specific to algae groups (Guiry & Guiry, 2023). Some examples of species with contradicting classifications include *Leptolyngbya* and *Pseudanabaena* which are classified under the order *Pseudanabaena*les by the NCBI database, but under the order Synechococcales according to AlgaeBase, and so the species are grouped in clusters specific to Synechococcales in **Figure 2.2 (next several pages)**.


Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
90 - 100
80 - 90
70 - 80
A
Cyano11 (MC)
CYAN77 (BP)
ASV841 (BS)
GR2 (MC)
ASV24 (PM)
CYAN16 (BP)
CYAN37 (BP)
GR8 (MC)
ASV840 (BS)
CYAN21 (BP)
CYAN18 (BP)
CYAN38 (BP)
GR7 (MC)
CYAN3 (BP)
ASV16 (AM)
ASV857 (BS)
ASV858 (BS)
CYAN34 (BP)
CYAN30 (BP)
ASV859 (BS)
CYAN48 (BP)
ASV861 (BS)
CYAN122 (BP)
CYAN72 (BP)
ASV25 (PM)
ASV860 (BS)
CYAN24 (BP)
ASV17 (AM)
CYAN41 (BP)
ASV27 (PM)
ASV837 (BS)
CYAN2 (BP)
ASV856 (BS)
CYAN1 (BP)
Aphanothece minutissima 2LT34S03 (FM177488.1)
Cyanobium sp. CENA132 (KP835535.1)
Cyanobium gracile PCC 6307 (NR 102447.1)
Cyanobium sp. UTMC6012 (MN733324.1)
CYAN43 (BP)
Cyanobium sp. NS01 (AY172837.1)
Cyanobium sp. LEGE 06113 (KC469577.1)
CYAN105 (BP)
Cyanobium sp. UMPCCC 1207 (KM218868.1)
CYAN94 (BP)
CYAN54 (BP)
ASV853 (BS)
CYAN45 (BP)
ASV29 (PM)
CYAN73 (BP)
CYAN95 (BP)
ASV26 (PM)
CYAN10 (BP)
ASV842 (BS)
CYAN20 (BP)
ASV843 (BS)
CYAN25 (BP)
ASV20 (AM-PM)
ASV19 (AM-PM)
GR5 (MC)
ASV844 (BS)
CYAN12 (BP)
CYAN28 (BP)
Cyanobium sp. CENA122 (KP855527.1)
Cyanobium sp. CENA118 (HQ380799.1)
Cyanobium sp. ULC183 (MK280770.1)
ASV847 (BS)
Chroococcales
Synechococcales
Synechococcales


B
CYAN111 (BP)
ASV849 (BS)
ASV848 (BS)
Cyano9 (MC)
GR1 (MC)
ASV36 (PM)
CYAN8 (BP)
ASV850 (BS)
CYAN9 (BP)
ASV33 (PM)
ASV851 (BS)
ASV852 (BS)
CYAN84 (BP)
ASV854 (BS)
ASV34 (PM)
ASV21 (AM-PM)
ASV855 (BS)
ASV22 (PM)
GR6 (MC)
ASV846 (BS)
CYAN22 (BP)
CYAN93 (BP)
ASV28 (PM)
ASV845 (BS)
ASV862 (BS)
CYAN74 (BP)
ASV863 (BS)
CYAN49 (BP)
ASV838 (BS)
Cyano10 (MC)
ASV23 (PM)
ASV839 (BS)
Cyano8 (MC)
CYAN66 (BP)
ASV834 (BS)
ASV835 (BS)
ASV836 (BS)
ASV879 (BS)
ASV872 (BS)
ASV15 (AM)
ASV875 (BS)
ASV878 (BS)
ASV874 (BS)
ASV876 (BS)
CYAN157 (BP)
CYAN139 (BP)
ASV873 (BS)
CYAN119 (BP)
ASV868 (BS)
CYAN7 (BP)
ASV31 (PM)
Cyano1 (MC)
CYAN6 (BP)
ASV864 (BS)
GR4 (MC)
GR9 (MC)
ASV869 (BS)
CYAN17 (BP)
ASV35 (PM)
ASV866 (BS)
ASV14 (AM)
ASV30 (PM)
ASV867 (BS)
CYAN107 (BP)
ASV18 (AM-PM)
ASV865 (BS)
CYAN29 (BP)
Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
70 - 80
80 - 90
90 - 100


C
ASV32 (PM)
ASV870 (BS)
ASV38 (PM)
ASV871 (BS)
ASV39 (PM)
Synechococcus sp. LEGE 06306 (HM217052.1)
Cyanobium sp. 2LT07S05 (FM177489.1)
Cyanobium sp. Laguna Chica (KT753319.1)
CYAN100 (BP)
Cyanobium sp. Bylot Island (KT753317.1)
Synechococcales
ASV888 (BS)
ASV37 (PM)
ASV922 (BS)
ASV923 (BS)
ASV921 (BS)
Synechococcus elongatus PCC 6301 (NR 074309.1)
ASV924 (BS)
Limnolyngbya circumcreta CHAB4449 clone 1 (KR697754.1)
Synechococcales
ASV48 (PM)
CYAN102 (BP)
CYAN133 (BP)
CYAN124 (BP)
CYAN118 (BP)
ASV925 (BS)
CYAN109 (BP)
CYAN142 (BP)
ASV926 (BS)
ASV927 (BS)
CYAN112 (BP)
CYAN132 (BP)
Phormidesmis communis KT5 (MK861902.1)
Phormidesmis priestleyi ANT.L52.6 (AY493579.1)
ASV53 (PM)
Phormidesmis nigrescens LK016 (KU219735.1)
Synechococcales
CYAN143 (BP)
Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
70 - 80
80 - 90
90 - 100



D

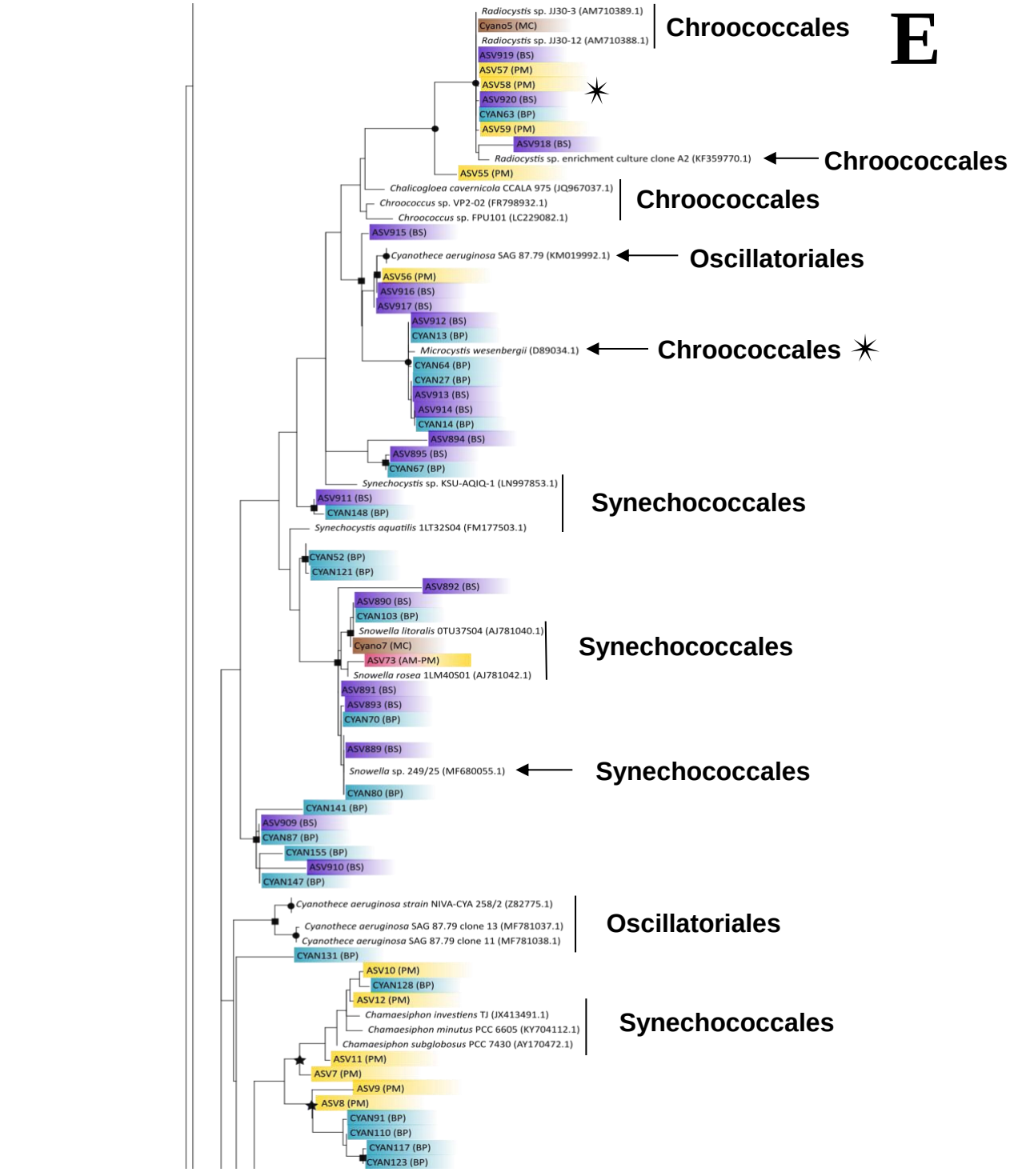
Nostocales

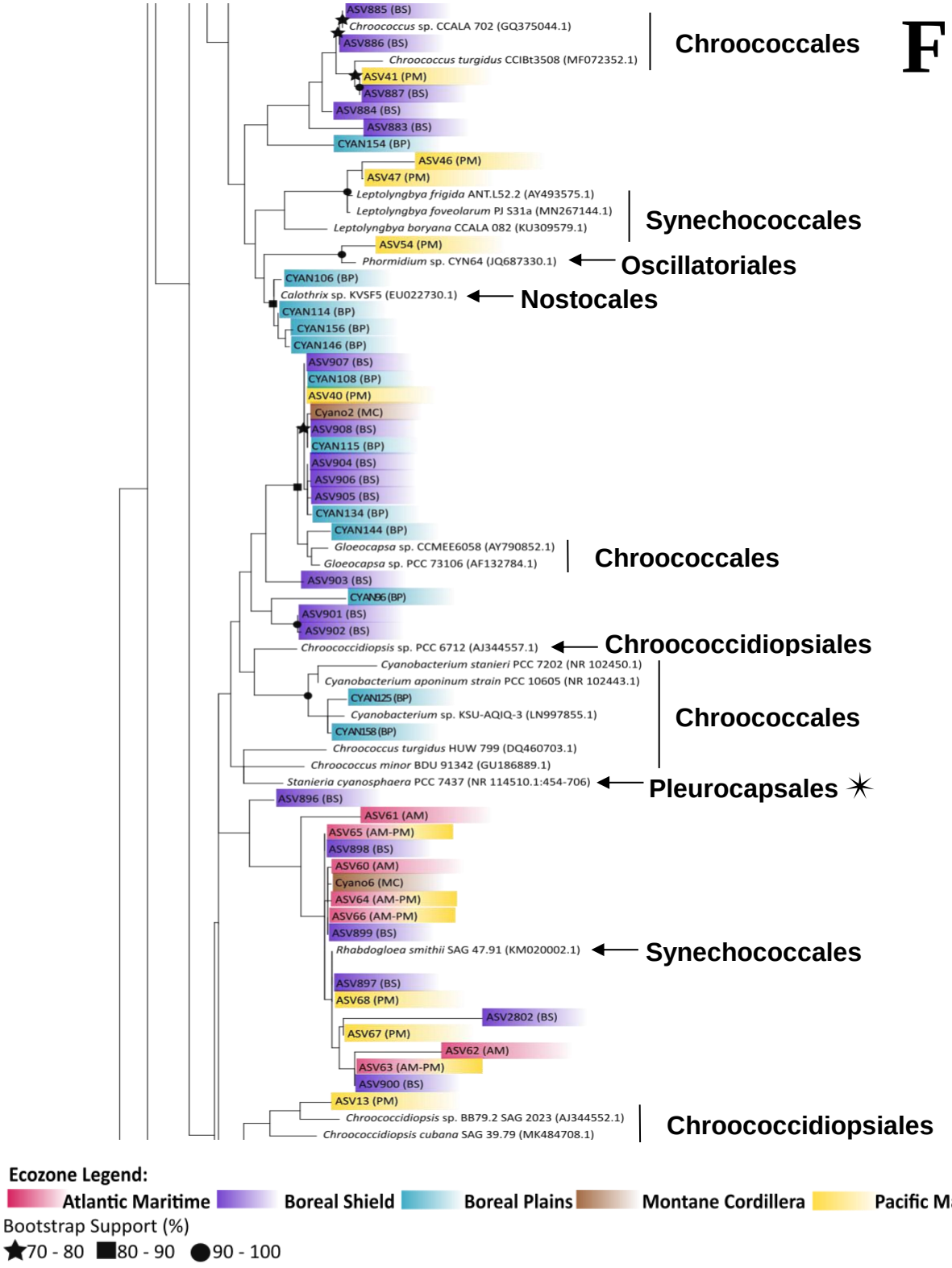
Ecozone Legend:

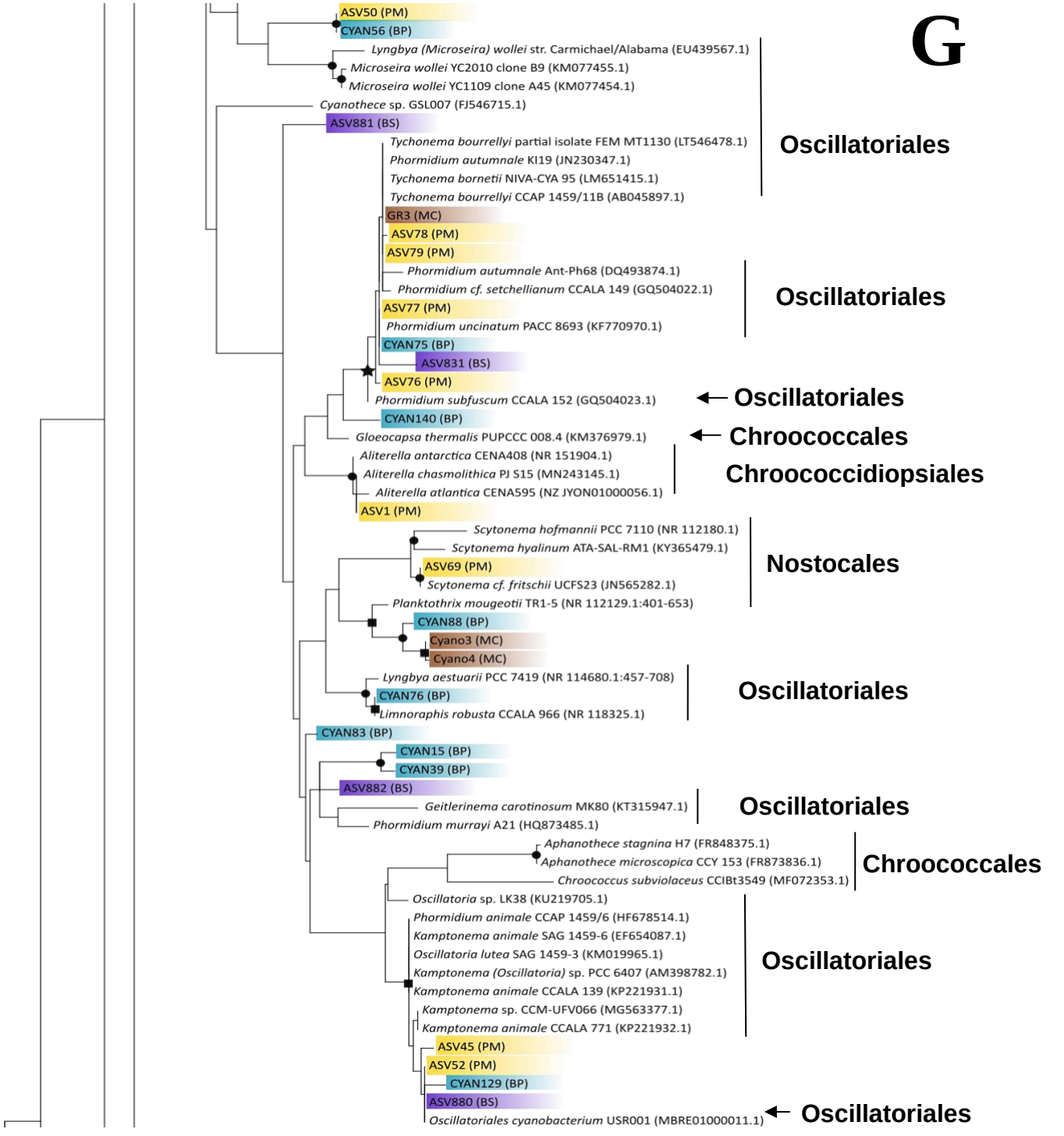
Atlantic Maritime **Boreal Shield** **Boreal Plains** **Montane Cordillera** **Pacific Maritime**

Bootstrap Support (%)

★70 - 80 ■80 - 90 ●90 - 100


E
Radiocystis sp. JJ30-3 (AM710389.1)
Cyano5 (MC)
Radiocystis sp. JJ30-12 (AM710388.1)
Chroococcales
ASV919 (BS)
ASV57 (PM)
ASV58 (PM)
ASV920 (BS)
CYAN63 (BP)
ASV59 (PM)
ASV918 (BS)
Radiocystis sp. enrichment culture clone A2 (KF359770.1)
Chroococcales
ASV55 (PM)
Chalicogloea cavernicola CCALA 975 (JQ967037.1)
Chroococcus sp. VP2-02 (FR798932.1)
Chroococcus sp. FPU101 (LC229082.1)
Chroococcales
ASV915 (BS)
Cyanothece aeruginosa SAG 87.79 (KM019992.1)
Oscillatoriales
ASV56 (PM)
ASV916 (BS)
ASV917 (BS)
ASV912 (BS)
CYAN13 (BP)
Microcystis wesenbergii (D89034.1)
Chroococcales
CYAN64 (BP)
CYAN27 (BP)
ASV913 (BS)
ASV914 (BS)
CYAN14 (BP)
ASV894 (BS)
ASV895 (BS)
CYAN67 (BP)
Synechocystis sp. KSU-AQIQ-1 (LN997853.1)
Synechococcales
ASV911 (BS)
CYAN148 (BP)
Synechocystis aquatilis 1LT32S04 (FM177503.1)
CYAN52 (BP)
CYAN121 (BP)
ASV892 (BS)
ASV890 (BS)
CYAN103 (BP)
Snowella litoralis 0TU37S04 (AJ781040.1)
Synechococcales
Cyano7 (MC)
ASV73 (AM-PM)
Snowella rosea 1LM40S01 (AJ781042.1)
ASV891 (BS)
ASV893 (BS)
CYAN70 (BP)
ASV889 (BS)
Snowella sp. 249/25 (MF680055.1)
Synechococcales
CYAN80 (BP)
CYAN141 (BP)
ASV909 (BS)
CYAN87 (BP)
CYAN155 (BP)
ASV910 (BS)
CYAN147 (BP)
Cyanothece aeruginosa strain NIVA-CYA 258/2 (Z82775.1)
Cyanothece aeruginosa SAG 87.79 clone 13 (MF781037.1)
Cyanothece aeruginosa SAG 87.79 clone 11 (MF781038.1)
Oscillatoriales
CYAN131 (BP)
ASV10 (PM)
CYAN128 (BP)
ASV12 (PM)
Chamaesiphon investiens TJ (JX413491.1)
Chamaesiphon minutus PCC 6605 (KY704112.1)
Chamaesiphon subglobosus PCC 7430 (AY170472.1)
Synechococcales
ASV11 (PM)
ASV7 (PM)
ASV9 (PM)
ASV8 (PM)
CYAN91 (BP)
CYAN110 (BP)
CYAN117 (BP)
CYAN123 (BP)
Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
70 - 80
80 - 90
90 - 100


F
ASV885 (BS)
Chroococcus sp. CCALA 702 (GQ375044.1)
ASV886 (BS)
Chroococcus turgidus CCIBt3508 (MF072352.1)
ASV41 (PM)
ASV887 (BS)
ASV884 (BS)
ASV883 (BS)
CYAN154 (BP)
ASV46 (PM)
ASV47 (PM)
Leptolyngbya frigida ANT.L52.2 (AY493575.1)
Leptolyngbya foveolarum PJ S31a (MN267144.1)
Leptolyngbya boryana CCALA 082 (KU309579.1)
ASV54 (PM)
Phormidium sp. CYN64 (JQ687330.1)
CYAN106 (BP)
Calothrix sp. KVSF5 (EU022730.1)
CYAN114 (BP)
CYAN156 (BP)
CYAN146 (BP)
ASV907 (BS)
CYAN108 (BP)
ASV40 (PM)
Cyano2 (MC)
ASV908 (BS)
CYAN115 (BP)
ASV904 (BS)
ASV906 (BS)
ASV905 (BS)
CYAN134 (BP)
CYAN144 (BP)
Gloeocapsa sp. CCMEE6058 (AY790852.1)
Gloeocapsa sp. PCC 73106 (AF132784.1)
ASV903 (BS)
CYAN96 (BP)
ASV901 (BS)
ASV902 (BS)
Chroococcidiopsis sp. PCC 6712 (AJ344557.1)
Cyanobacterium stanieri PCC 7202 (NR 102450.1)
Cyanobacterium aponinum strain PCC 10605 (NR 102443.1)
CYAN125 (BP)
Cyanobacterium sp. KSU-AQIQ-3 (LN997855.1)
CYAN158 (BP)
Chroococcus turgidus HUW 799 (DQ460703.1)
Chroococcus minor BDU 91342 (GU186889.1)
Stanieria cyanosphaera PCC 7437 (NR 114510.1:454-706)
ASV896 (BS)
ASV61 (AM)
ASV65 (AM-PM)
ASV898 (BS)
ASV60 (AM)
Cyano6 (MC)
ASV64 (AM-PM)
ASV66 (AM-PM)
ASV899 (BS)
Rhabdogloea smithii SAG 47.91 (KM020002.1)
ASV897 (BS)
ASV68 (PM)
ASV2802 (BS)
ASV67 (PM)
ASV62 (AM)
ASV63 (AM-PM)
ASV900 (BS)
ASV13 (PM)
Chroococcidiopsis sp. BB79.2 SAG 2023 (AJ344552.1)
Chroococcidiopsis cubana SAG 39.79 (MK484708.1)
Chroococcales
Synechococcales
Oscillatoriales
Nostocales
Chroococcales
Chroococcidiopsiales
Chroococcales
Pleurocapsales
Synechococcales
Chroococcidiopsiales
Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
70 - 80
80 - 90
90 - 100


G
ASV50 (PM)
CYAN56 (BP)
Lyngbya (Microseira) wollei str. Carmichael/Alabama (EU439567.1)
Microseira wollei YC2010 clone B9 (KM077455.1)
Microseira wollei YC1109 clone A45 (KM077454.1)
Cyanothece sp. GSL007 (FJ546715.1)
ASV881 (BS)
Tychonema bourrellyi partial isolate FEM MT1130 (LT546478.1)
Phormidium autumnale KI19 (JN230347.1)
Tychonema bornetii NIVA-CYA 95 (LM651415.1)
Tychonema bourrellyi CCAP 1459/11B (AB045897.1)
Oscillatoriales
GR3 (MC)
ASV78 (PM)
ASV79 (PM)
Phormidium autumnale Ant-Ph68 (DQ493874.1)
Phormidium cf. setchellianum CCALA 149 (GQ504022.1)
Oscillatoriales
ASV77 (PM)
Phormidium uncinatum PACC 8693 (KF770970.1)
CYAN75 (BP)
ASV831 (BS)
ASV76 (PM)
Phormidium subfuscum CCALA 152 (GQ504023.1)
Oscillatoriales
CYAN140 (BP)
Gloeocapsa thermalis PUPCCC 008.4 (KM376979.1)
Chroococcales
Aliterella antarctica CENA408 (NR 151904.1)
Aliterella chasmolithica PJ S15 (MN243145.1)
Aliterella atlantica CENA595 (NZ JYON01000056.1)
Chroococcidiopsiales
ASV1 (PM)
Scytonema hofmannii PCC 7110 (NR 112180.1)
Scytonema hyalinum ATA-SAL-RM1 (KY365479.1)
Nostocales
ASV69 (PM)
Scytonema cf. fritschii UCFS23 (JN565282.1)
Planktothrix mougeotii TR1-5 (NR 112129.1:401-653)
CYAN88 (BP)
Cyano3 (MC)
Cyano4 (MC)
Lyngbya aestuarii PCC 7419 (NR 114680.1:457-708)
Oscillatoriales
CYAN76 (BP)
Limnoraphis robusta CCALA 966 (NR 118325.1)
CYAN83 (BP)
CYAN15 (BP)
CYAN39 (BP)
ASV882 (BS)
Geitlerinema carotinosum MK80 (KT315947.1)
Oscillatoriales
Phormidium murrayi A21 (HQ873485.1)
Aphanothece stagnina H7 (FR848375.1)
Aphanothece microscopica CCY 153 (FR873836.1)
Chroococcus subviolaceus CCIBt3549 (MF072353.1)
Chroococcales
Oscillatoria sp. LK38 (KU219705.1)
Phormidium animale CCAP 1459/6 (HF678514.1)
Kamptonema animale SAG 1459-6 (EF654087.1)
Oscillatoria lutea SAG 1459-3 (KM019965.1)
Kamptonema (Oscillatoria) sp. PCC 6407 (AM398782.1)
Kamptonema animale CCALA 139 (KP221931.1)
Kamptonema sp. CCM-UFV066 (MG563377.1)
Kamptonema animale CCALA 771 (KP221932.1)
Oscillatoriales
ASV45 (PM)
ASV52 (PM)
CYAN129 (BP)
ASV880 (BS)
Oscillatoriales cyanobacterium USR001 (MBRE01000011.1)
Oscillatoriales
Ecozone Legend:
Atlantic Maritime
Boreal Shield
Boreal Plains
Montane Cordillera
Pacific Maritime
Bootstrap Support (%)
70 - 80
80 - 90
90 - 100

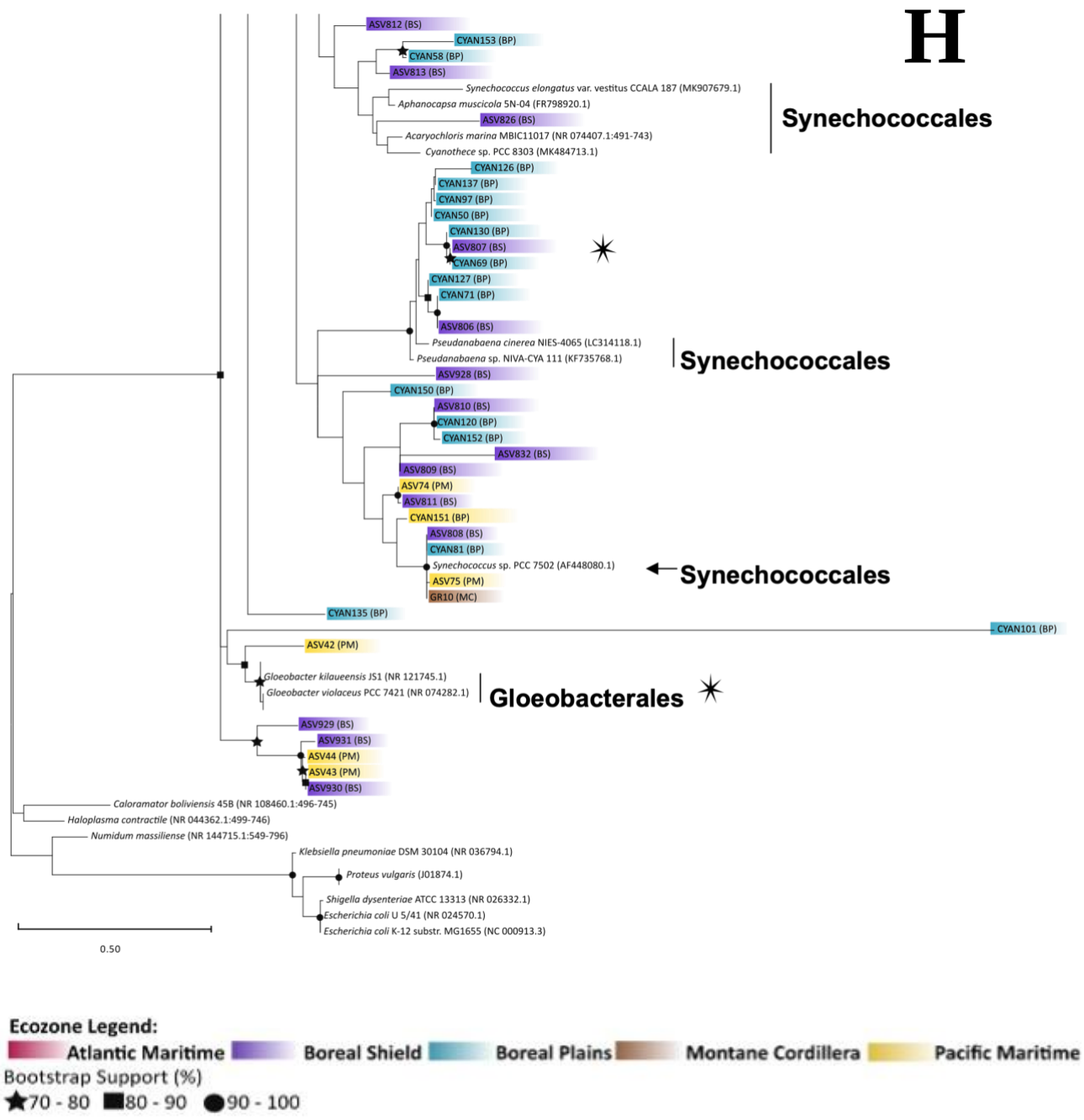


Figure 2.2 Maximum likelihood phylogenetic tree featuring V4-16S rRNA cyanobacterial ASVs from five forested watersheds across Canada (A-H) part of the *for*Water Network

The compilation of cyanobacterial samples sequenced from five Canadian forested ecozones. Each highlighted sample represents the ecozone it was collected from: AM (red), BS (purple), BP (blue), MC (brown), PM (yellow). The ML phylogenetic tree (522 sequences, 249 sites) constructed in MEGA X ©, (log likelihood -13806.90) includes bootstrap support percent of 1000 replicates. Clusters of samples belonging to the same orders are also labeled to distinguish the conservation or variability of cyanobacterial species. Bootstrap support branches distinguish highly supported clades with 90-100% (circle), 80-90% (square), and 70-80% (star) and branches lacking values indicate less than 70% bootstrap support.

2.4.1 Taxonomy of Cyanobacteria

The orders identified on the right side of **Figure 2.2** throughout the phylogenetic tree are specific to clusters in which the reference sequences define the classification rather than being based off the taxonomy attributed by the SILVA database. Across all sampled ecozones, many of the sequences largely clustered into one of four cyanobacterial orders – Synechococcales, Nostocales, Chroococcales, and Oscillatoriales. Additional orders that are represented by reference sequences to a lesser extent are the orders Gloeobacterales, Chroococcidiopsiales, and Pleurocapsales. The orders with the highest number of clusters, Synechococcales and Chroococcales, are represented eight times each throughout the tree, followed by six clusters of Oscillatoriales, two clusters of Nostocales, three clusters of Chroococcidiopsiales, and one cluster each of Pleurocapsales and Gloeobacterales. It is important to note that because the clusters are defined by reference sequences that could span over separate clades, there is not always a clear divide on where a cluster may begin and end. Clusters containing sequences that did not group with any reference sequences and are not well-supported within a cluster containing a reference sequence are considered unclassified cyanobacterial sequences.

Although the orders are represented repeatedly across the tree and at varying levels of support, clades noted with higher bootstrap values are more strongly supported. Synechococcales clusters (**Figure 2**) are highly supported on page 41, (**A**, 90-100%), page 43

(**C,** 90-100%), page 45 (**E**, 80-90%), two clusters on page 46 (**F**) one supported between 70-80% and the other at 90-100%, and lastly a strongly supported cluster on page 48 (**H**) which has several variable clades spanning across 70-100% bootstrap support. As for Chroococcales clusters, there is strong support (90-100%) on page 45 (**E**) that contains two highly supported clades, two variable clades on page 46 (**F**, 70-100%, 90-100%), and on page 47 (**G**, 90-100%). Oscillatoriales is highly supported at three points on the ML tree with bootstrap supports ranging from 80-100% on pages 45 (**E**), 46 (**F**), and 47 (**G**). Lastly, Nostocales is well supported on page 44 (**D**, 70-100%), page 46 (**F**, 80-90%), and page 47 (**G**, 90-100%). Although Chroococcidiopsiales, Pleurocapsales, and Gloeobacterales are present as well, these clades are not as well supported (<70% bootstrap support). Many other clades throughout the tree are supported at the highest bootstrap percentage, however, there are instances, such as the very first clade on page A where clades are supported without validation from any reference sequences; alternatively, like the Oscillatoriales clade on page 45 (**E**) consisting of three *Cyanothece aeruginosa* (Z82775.1, MF781037.1, MF781038.1) reference sequences, this clade is highly supported but does not contain any environmental sample sequences.

Although full clades are not always highly supported, there are instances where smaller clades within larger clusters are individually supported with high confidence. Such as on page 46 (**F**), an Oscillatoriales cluster with ASV54 (PM) and *Phormidium* sp. (JQ687330.1), and page 44 (**D**)**,** a Nostocales cluster with ASV49 (PM) and *Macrochaete psychrophile* (KT336439.2) are all supported at the highest confidence. On page 43, (**C**) one clade contains CYAN100 (BP) and *Cyanobium* sp. (KT753317.1), and another clade contains ASV921, ASV922, and ASV923 (BS) – all supported with 90-100% bootstrap confidence (**Fig 2.2,** pg. 43, **C**). Directly below on the same page 43 (**C**), ASV924 (BS) and *Limnolynbya circumcreta* (KR697754.1) are also supported at the highest bootstrap confidence. On page 45 (**E**) the sequences Cyano5 (MC), ASV918, 919, 920 (BS), ASV57, 58 59 (PM), CYAN63 (BP) are all present on a highly supported clade with reference sequences for *Radiocystis* spp. (AM710389.1, AM710388.1, KF359770.1).

Being that the orders represented by reference sequences appear several times throughout the entire tree, Synechococcales, Chroococcales, Oscillatoriales, Nostocales, Chroococcidiopsiales, and Gloeobacterales are not monophyletic. There are numerous clusters that do not have any affinity to reference sequences (e.g., pg. 42 (**B**) mainly consists of clades with no defined reference sequences), yet still have well supported bootstrap confidence. Approximately 25% of the sample sequences did not have valid blast results and are therefore considered to be unknown cyanobacterial taxa until determined to be new species or genera.

2.4.2 Biogeography of Cyanobacteria

From a whole perspective, the phylogenetic tree in **Figure 2.2** does not show any immediate biogeographic trends. Overall, there tends to be very few patterns between the geographical distribution and the type of cyanobacterial orders present, despite the ecozones having distinctive microbiomes. From an in-depth perspective, there are some, albeit scarce, well supported biogeographic groups present throughout the tree. On page 44 (**D**) there are several sequences from the BP (blue) all grouping in the center of a Nostocales cluster – CYAN5, 89, 55, 44, 19, and 62 – however, no bootstrap support is present. Just below that grouping, CYAN59, 78, and 99 also cluster together with 70-80% bootstrap support. There is a grouping of highly supported (90-100%) BS sequences (purple) on page 43 (**C**) in which ASV921-923 all congregate within a Synechococcales cluster. On page 45 (**E**) sequences from BS and BP group together in a clade supported by 90-100% bootstrap confidence within a Chroococcales cluster – ASV 912-914 (BS) and CYAN13, 14, 27, 64 (BP); within that same region, several more BS and BP sequences group together, some with little support, and others with 80-90% bootstrap support. Some sequences from the PM – ASV7, 8, 9, 11 – grouped together (80-90% support) on page 45 (**E**) within a Synechococcales cluster. On the next page (46, **F**), both AM and PM sequences grouped together (ASV64, 65, 66) in an unsupported clade between ambiguously defined orders, however, some sequences from the BS and MC are also present within this group (ASV898, 899, Cyano6), lowering the biogeographic

significance. There are no apparent groupings of sequences from the MC present in **Figure 2.2**.

In addition to observing the biogeographic distribution of general cyanobacterial taxonomy across Canada, it is also very informative to observe what potential toxin producing strains could present in harmful blooms. While the exact strains require further analysis, the tree in **Figure 2.2** does exhibit sequences clustering in clades with well-known toxin producing genera. The large Nostocales cluster on page 44 (**D**) contains several clades including sequences from each of the five ecozones clustering with toxin producing genera *Aphanizomenon* (e.g., *Aphanizomenon flos-aquae* NIES-81 [AJ293131.1]), *Anabaena (*e.g., *Anabaena compacta* 0RE2753 [DQ264176.1]), and *Dolichospermum* (ex., *Dolichospermum flos-aquae* 2-Caskovec2/11 [KT290337.1]). On page 45 (**E**), there is a highly supported clade (>90% bootstrap support) containing microcystin producer *Microcystis wesenbergii* (D89034.1) and sequences from BS and BP. Several *Phormidium* reference sequences are clustering on page 44 (>70% bootstrap, **G**) with sequences from BS, BP, MC, and PM. And finally, on page 48 **(H**), within a Synechococcales cluster *Pseudanabaena* reference sequences such as *Pseudanabaena cinerea* NIES-4065 (LC314118.1) are contained in >90% confidence clades with sequences from BP and BS.

While there are some areas of the tree where sequences from the same ecozone clustered together, sometimes with high bootstrap support as well, there are exceedingly more sequences scattered across the tree showing no biogeographic trends. This is clear across pages 41-42 (**A**-**B**) where all the sequences are distributed from each of the five ecozones and there appears to be no trend of any kind, even within well-supported bootstrap clades. In all, based on the spread of sequences in the phylogenetic tree (**Fig 2.2**) the environmental sequences across major forested watersheds in Canada do not show clusters according to location and biogeographic groupings.

2.5 Discussion

Cyanobacteria are photosynthetic gram-negative organisms that primarily inhabit aquatic habitats, though many genera can also survive in arid deserts, high salinity environments, and even in snow or ice resulting in a diversity of species (Cordeiro et al., 2020; Johansen & Casamatta, 2005). Species diversity within this phylum is considerable and as a result cyanobacterial classification systems have undergone extensive revisions over the past century to have cyanobacterial taxonomy reflect its phylogeny. This taxonomy continues to rapidly change with the advent of new sequencing technologies such as using 16S rRNA gene as a marker in next generation sequencing of environmental samples (Cheng et al., 2022; Cordeiro et al., 2020; Strunecký et al., 2023). Along with the next generation sequencing, the expansion of bioinformatic workflows and improved polyphasic approaches, between 2014 and 2021, 273 species in 140 genera were described, mostly classified into previously defined orders but included new families as well (Strunecký et al., 2023). More recently, a polyphasic approach that combines molecular targets (16S rRNA gene, whole genomes), ecological, physiological, and morphological characteristics, and biochemical potential (i.e., gene diversity) have been used to establish classification systems (Cheng et al., 2022; Komárek, 2016). The ideal outcome for a cyanobacterial naming system is to establish monophyletic groupings that show the taxonomy reflects the phylogeny, however, even the most updated classification systems still find many challenges with this (Cheng et al., 2022; Strunecký et al., 2023).

2.5.1 History of Cyanobacterial Classification Systems

Previously, cyanobacteria were separated into distinct orders based on morphological characters prior to sequencing data becoming more accessible in cost and use (Rippka et al., 1979; Strunecký et al., 2023). The use of morphological characteristics to identify species has many challenges due to the lack of defining characteristics in the bacteria and the algae. In addition, morphology can be influenced by environmental and/or culturing conditions resulting

in cryptic species (Cheng et al., 2022; Johansen & Casamatta, 2005; Komárek, 2016; MacKeigan et al., 2022). Over time, classification systems for cyanobacteria have undergone several dramatic reconstructions, starting with Geitler in 1925 which was referred to for around 80 years before being re-examined by Rippka et al. (1979). This system was later adopted into Bergey's Manual for Systematic Bacteriology, which was soon thereafter altered into more traditional taxonomic classifications to comply with the International Code of Nomenclature (ICN) for algae, fungi, and plants (Komárek et al., 2014; Strunecký et al., 2023). These changes utilized 16S rRNA gene sequenced data that was available at the time and began considering thylakoid arrangement into cyanobacterial classifications (Strunecký et al., 2023). One of the most recent classification systems by Komárek et al. (2014) introduced phylogenetic analysis that suggested classifying some genera into new orders but was unable to fully address polyphyly in some groups (Komárek et al., 2014; Strunecký et al., 2023). The current proposal for cyanobacterial classification systems by Strunecký et al. (2023) uses the developments from all previous system revisions, with new 16S rRNA gene phylogenies and genomes for phylogenetic analysis from reference strains based on polyphasic identification methods proposed by Komárek et al. (2016). Still, with many unknown taxa yet identified and many orders falling into polyphyletic categories, the current system acknowledges that it is a checkpoint in cyanobacterial taxonomy, needing far more identification for improved taxonomic resolution (Strunecký et al., 2023).

2.5.2 Taxonomy of Cyanobacteria Across Five Major Forested Ecozones

The samples used in the current study for the phylogenetic tree (**Figure 2.2**) were sequenced by previous members of the Müller Lab at the University of Waterloo and hence a more descriptive analysis on the phylogeny from each of the ecozones separately can be found in their respective theses (Aljoudi, 2021; Cameron, 2021; Kwok, unpublished; Shardlow, 2021). Within each respective ecozone (excluding the BP) Synechococcales, Chroococcales, and Oscillatoriales were the most abundant cyanobacterial orders in all the studied watersheds, organized into non-monophyletic clades in each maximum likelihood tree (Aljoudi, 2021;

Cameron, 2021; Shardlow, 2021). The phylogenetic tree in this study (**Figure 2.2**) brings together all the previously extracted sequences from the forested ecozones across Canada into a single combined analysis.

As portrayed at the very top of the phylogenetic tree in **Figure 2.2** (ex., pages 41-42, **A-B**), phylogenetic analysis of cyanobacterial amplicon sequence variants (ASVs) from across Canada indicates there are gaps in sequenced and identified cyanobacterial genomes according to what is available in NCBI GenBank. This finding is unsurprising as the phylum Cyanobacteria has a large amount of diversity and this diversity is not reflected in the available reference sequences for entire cyanobacterial genomes, as well as constant new research leading to frequent revisions and updates to taxonomic classifications of the phylum (MacKeigan et al., 2022; Strunecký et al., 2023). The first starred clade, on pg. 41 (**Fig 2.2, A**), as well as the entirety of page 42 (**B**) reveals several clades containing ASVs from each of the five ecozones clustering together; although no reference sequences are present in these clades, they still have very high bootstrap support, signifying how many unknown sequences are available that do not align with known taxonomy groups and further demonstrating the incredible diversity of unknown taxa in environmental samples. This is to be expected to some degree because there are still many taxa of cyanobacteria, especially at the family level, that have not been sequenced and determined by taxonomy (Strunecký et al., 2023). As of 2018, only about 400 cyanobacterial genomes were completely sequenced and available publicly, compared to the over 30,000 complete genomes from bacterial phyla (Ribeiro et al., 2018). In addition, while the most recent classification revision by Strunecký et al. (2023) proposed ten new orders and fifteen new families within the Cyanobacteria, the phylum is still underrepresented in genomic libraries because of the lack in environmental, taxonomical, and geographical diverse sequences (Ribeiro et al., 2018).

The polyphyletic order, Synechococcales, contains many highly diverse families and genera with genetically and distinct taxonomic lineages (Komárek et al., 2020; Strunecký et al., 2023). Synechococcales are small unicellular cocci, with oval or elongated rod-shaped cells

2-5 µm long and 1.5-3 µm wide, and parietal thylakoids (cells that contain chlorophyll for photosynthesis) (Komárek et al., 2020; Strunecký et al., 2023). There are both marine and freshwater groups, living in benthic and planktonic habitats, either living solitary or joining together in large masses (Strunecký et al., 2023). There are over 100 classified genera within this order, with at least 40 new genera (includes 88 species) being added since 2014 due to the improvements in sequencing using the 16S rRNA gene (Strunecký et al., 2023). Traditionally, small cocci cyanobacteria with thin filamentous simple morphologies were grouped under Synechococcales, however, the extreme polyphyly in this order, along with high morphological diversity and unclear phylogenetic assignments keeps pushing for newly revised definitions for this group (Strunecký et al., 2023).

Appearing in multiple clades in **Figure 2.2**, Synechococcales represents the most abundant order across all five ecozones. Reference sequences from Synechococcales are observed to form frequent clades with a variety of well-supported ASVs, clustering together from distant ecozone regions. This is unsurprising as many cyanobacterial community analyses in Canada find that the dominating order, Synechococcales, encompasses 50-75% of the relative abundances, more specifically to *Synechococcus* complexes (e.g., *Aphanocapsa* and *Cyanobium*) in mesotrophic and oligotrophic lakes (Brown et al., 2021; MacKeigan et al., 2022). Many of the reference genera observed in the phylogenetic tree, notably *Cyanobium* and *Synechococcus*, are picocyanobacteria, very small cyanobacteria species measuring 0.2-2 µm in diameter (Komárek et al., 2020; MacKeigan et al., 2022; Strunecký et al., 2023). While much of the sequenced cyanobacteria in Canada are these picocyanobacteria, they account for a significantly low percent of identification rates because of inconsistencies of taxonomy in incomplete reference libraries (MacKeigan et al., 2022). Because of this, picocyanobacteria account for many of the unknown ASVs observed in **Figure 2.2**, such as the starred clade on page 48 (**H**) of unknown ASVs from the BP and BS clustering in between recognized Synechococcales clades. While the highlighted ASVs are not clustered with known taxonomic groups, the high bootstrap supports the validity of the sequences, and re-establishes the amount of unknown cyanobacteria waiting to be sequenced and identified. While sequences from the

Synechococcales are abundant throughout the tree, this is further evidence that the picocyanobacteria are highly diverse and widespread across very different habitats in Canada.

The order Nostocales consists of cyanobacteria that are phylogenetically distinct due to their ability to form nitrogen-fixation cells called heterocysts, which are ancestral to this group (Strunecký et al., 2023). The taxa belonging to Nostocales tend to be monophyletic, however many of the relationships within this order have low 16S rRNA gene diversity and high morphological resemblances (Strunecký et al., 2023). Most cyanobacteria in this group have either true or falsely branched filaments, and it is these traits associated with filamentous morphology, along with the presence of heterocytes and akinetes (spore-like cells that allow cyanobacteria to enter dormant stages during unfavourable environmental conditions), that make Nostocales one of the most phenotypically diverse (and challenging) group of cyanobacteria (Strunecký et al., 2023). A distinct Nostocales clade in **Figure 2.2** (pg. 44, **D**) signifies that this order is not well-supported, yet still form very large clades containing multiple reference sequences and ASVs from all five ecozones. Nostocales contains bloom forming and potential toxin-producing species in freshwater sources, including *Dolichospermum*, *Aphanizomenon*, *Anabaena*, and *Cylindrospermum* (family Aphanizomenonaceae) (Brown et al., 2021; MacKeigan et al., 2022; Strunecký et al., 2023), all of which are observed in **Figure 2.2** from all over Canada. The observation of further Nostocales clades (pg. 46 **F**, pg. 47 **G**) challenges the typical monophyly of this order. This could be the result of low phylogenetic resolution and incorrect reference sequences that are known to be misidentified with sequencing methods due to the taxonomic complexity of Nostocales (Strunecký et al., 2023). Nostocalean genera require better resolved reference sequences and a stringent framework that observes genotypic, phenotypic, and taxonomic diversity in unison to untangle the current state of classification within this order, of which one has not been developed yet (Strunecký et al., 2023).

Chroococcales, one of the first cyanobacterial orders to classified by Geitler (1925), has undergone many fundamental restructurings in its taxonomic classification, largely due to

the difficulty of culturing members of this order (Strunecký et al., 2023). Morphologically, this order is highly variable, contains parietal thylakoids arranged into simple spherical cells, which might develop into colonies that string together to form complex configurations (Komárek et al., 2020; Strunecký et al., 2023). Another common order observed in **Figure 2.2**, Chroococcales, is present in all ecozones, excluding the Atlantic Maritime. At the top of page 45 (**E**) is a starred example of *Radiocystis* sp. JJ30-3 (AM710389.1) and *Radiocystis* sp. JJ30-12 (AM710388.1) reference sequences forming a clade with well supported ASVs from four of the forested watersheds. Below that, another starred clade with high bootstrap support aligns *Microcystis wesenbergii* (D89034.1) with ASVs from the Boreal Shield and Boreal Plains, significant for its potential for microcystin production. Although *Microcystis* only appears once in the phylogenetic tree, it is not unexpected because across over 379 lakes in Canada, *Microcystis* and other Chroococcales genera showed a clear trophic gradient in which their ASVs scarcely appeared in oligotrophic and mesotrophic lakes, increasing in eutrophic lakes, and spiking in hypereutrophic lakes (MacKeigan et al., 2022). Since the samples used in this research all come from oligotrophic watersheds, eutrophication would be low, and members of this order were not as likely to persist.

Due to the lack of data among the Chroococcales, there are several suggested reorganizations currently being considered that may alter the organization of **Figure 2.2**, one of them being reintegrating Chroococcales and Pleurocapsales into a single order (Strunecký et al., 2023). Because many taxa have been classified back and forth between both orders, integrating them is reinforced by multi-locus analysis to support monophyletic organization (Strunecký et al., 2023). A clear example is present in the phylogenetic tree (**Fig 2.2**, pg. 46, **F**), where *Stanieria cyanosphaera* PCC 7437 (NR 114510.1), a reference sequence belonging to the order Pleurocapsales, is starred. The position of *Stanieria* directly below several strongly supported Chroococcales clades, its direct incorporation in a clade containing two Chroococcales genera, *Chroococcus minor* BDU 91342 (GU186889.1) and *Chroococcus turgidus* HUW 799 (DQ460703.1), as well as a lack of further Pleurocapsales sequences, supports the claims of reintegrating the two orders. At the time of publication of this thesis,

both orders are considered valid within **Figure 2.2**, and the potential for a taxonomic shift in the future is acknowledged.

The remaining orders observed in **Figure 2.2** were Oscillatoriales, Chroococcidiopsiales, and Gloeobacterales. Oscillatorean cyanobacteria have a widely variable cell diameter ranging between 2-115 µm, are typically filamentous (Strunecký et al., 2023), and have irregular thylakoids. In past classification systems, filamentous genera that did not belong to Synechococcales were considered Oscillatoriales, however, phylogenetic analyses shows that many of these ambiguous genera fall much further apart than their morphologies indicate (Strunecký et al., 2023). Chroococcidiopsiales, possibly the closest identified relative to heterocystous cyanobacteria, still lacks considerable genetic data, and therefore only contains a single family of coccoid colonial species with irregular thylakoid architecture (Strunecký et al., 2023). The position of Chroococcidiopsiales clades near Chroococcales and Pleurocapsales is expected as they are morphologically similar and are orders requiring more sequencing to resolve their relatedness to each other (Strunecký et al., 2023). The last order, Gloeobacterales, located in a single clade at the base of the tree (**Fig 2.2**, pg. 48, **H**), is positioned exactly as expected in cyanobacterial phylogeny (Strunecký et al., 2023). The Gloeobacterales are rod-shaped cells between 2-10 µm long and are the only known cyanobacteria to not contain thylakoids – instead containing their respiratory and photosynthetic complexes elsewhere in the cell (Komárek et al., 2020; Strunecký et al., 2023). The lack of thylakoids along with their basal position in phylogenies support the notion that the Gloeobacterales are a primitive group of cyanobacteria (Komárek et al., 2020; Strunecký et al., 2023).

2.5.3 Biogeography of Cyanobacteria Across Canada

Microbial biogeography describes how species are distributed and investigates the forces driving their dispersal patterns (Ribeiro et al., 2018). Forested watersheds across Canada are ideal study sites to observe biogeography of cyanobacteria because there is considerable distance between the ecozones, and each region has its own vast unique climate, geology, and

ecology (MacKeigan et al., 2022). The findings of the phylogenetic tree in **Figure 2.2** illustrates that cyanobacteria are distributed evenly across all five major forested ecozones in Canada. While there are some instances of local clustering within orders, such as the high percentage of BP ASVs in the Nostocales clade on page 44 (**D**), and the BS and BP ASV cluster in the Chroococcales clade (starred) on page 45 (**E**), there is not enough local clustering to suggest concrete biogeographic trends. There are even several instances of similar ASVs from the AM and PM clustering together, despite their ecozones being a great distance away from one another. The lack of biogeographic trends suggests that cyanobacteria are not constrained by geographic restrictions as they are capable of dispersal to some degree; instead, the success of cyanobacterial dispersal relies more on the organism's resiliency to find and withstand transport and tolerate the conditions for the new geographic region they reach (Curren & Leong, 2020).

Cyanobacterial dispersal is not a new area of research, however, far less is understood abut dispersal between freshwater species compared to marine types (Curren & Leong, 2020; Sheath, 2006). The main sources of freshwater cyanobacterial dispersal involve moving throughout connected waterways, attaching to biological vectors (e.g., fish, insects, terrestrial animals), through the air (i.e., transported by wind as condensation and precipitation droplets), and finally by human interference (e.g., boat pollution) (Curren & Leong, 2020; Sheath, 2006; Wiśniewska et al., 2019). Although airborne transportation of cyanobacteria is still the least understood dispersal mechanism, the suggested mode of travel is through bioaerosols – a collection of biological material such as bacteria, viral matter, fungi, microalgae, etc., entering the atmosphere and moving via wind and precipitation patterns (Curren & Leong, 2020; Wiśniewska et al., 2019; Wiśniewska et al., 2020). Though the results of **Figure 2.2** corroborate the theories of cyanobacterial dispersal, bioaerosols are heavily influenced by several factors including the weather, size of water bodies, and organism abundances in the water, and thus requires much more research to better understand just how successful this mode of transport is (Curren & Leong, 2020; Sheath, 2006; Wiśniewska et al., 2019; Wiśniewska et al., 2020).

As identified in section 2.2.2, there are several cases of toxin-producing reference sequences clustering with samples from each of the five ecozones, such as *Aphanizomenon*, *Microcystis*, and *Dolichospermum*, which are also the three most commonly dominant cyanobacterial taxa observed in blooms across Ontario (Favot et al., 2023; Reinl et al., 2021; Winter et al., 2011). Numerous species identified in the V4-16S rRNA region tree, including the previously mentioned three, as well as *Planktothrix* and *Pseudanabaena* are some of the more problematic toxic cyanobacteria present across Canada (Health Canada, 2022). While it is not possible to determine if the toxin-producing strains identified in **Figure 2.2** contain the biosynthetic genes required for cyanotoxin synthesis without further analysis, it is still imperative to understand these threats might be present. For example, although one of the most well-known microcystin producers, *Microcystis*, is only present in one clade clustered with samples from the BS and BP (pg. 45, **E**), this does not directly relate to the presence of the toxin itself – only the potential for its synthesis. In the next chapter, sequences from the Boreal Shield will be further analyzed using metagenomic whole genome shotgun sequencing to explore biochemical potential and functional gene diversity within an oligotrophic forested watershed (TLW).

2.6 Conclusions

The taxonomy of cyanobacteria is widely distributed across Canada as shown in the results of **Figure 2.2**, with many ASVs clustering together from each of the five forested ecozones. The taxonomy spans eight orders, majority of which are non-monophyletic, with taxa appearing in different watersheds and including a high number of unknown ASVs. Although classification systems for cyanobacteria aim to construct monophyletic groupings, the findings from the V4-16S rRNA region gene phylogenetic tree illustrate the degree of polyphyly occurring among these orders and represents the vast diversity of cyanobacteria (Cheng et al., 2022; Komárek et al., 2014; Strunecký et al., 2023). Misidentification and incorrectly deposited sequences are one of the largest challenges in gene sequencing methods because as cyanobacteria classification is reorganized, many reference databases do not update

as often as the taxonomy changes (MacKeigan et al., 2022). While **Figure 2.2** highlights the diverse distribution of cyanobacteria, the proportion of unknown sequences signifies how much diversity is yet unknown, reinforcing the need for more sequencing and reference strains to bridge the gap in missing diversity data.

Based on the position of many of these unknown ASVs in relation to Synechococcales clades, it is likely that there is a significant percent of picocyanobacteria that have yet to be sequenced, as they may be passing through standard pore filters or more likely being misidentified as other genera (Komárek et al., 2020; MacKeigan et al., 2022). As for the biogeography of cyanobacteria in Canada**,** there was no distinct geographic grouping of any cyanobacterial taxa, and instead ASVs from all ecozones were observed clustered together in well supported clades. Broadly, cyanobacteria are suggested to be widely distributed worldwide in both fresh and marine systems (MacKeigan et al., 2022; Sheath, 2006), perhaps through airborne transportation (Wiśniewska et al., 2019), and the results from the V4-16S rRNA region gene tree support the prevalent circulation of cyanobacteria taxa, demonstrating successful dispersal between freshwater bodies across Canada.

While modern molecular methods are more convenient and user-friendly, molecular phylogenies often do not align with morphological classifications (Soo et al., 2014; Strunecký et al., 2023; Zhaxybayeva et al., 2006). There are many limitations and challenges to phylogenetic analyses of cyanobacteria, many of which were observed in this research. As clearly highlighted, cyanobacterial classification systems are always undergoing revisions and reconstructions to the current proposals. The main obstacle is the lack in diverse sequenced cyanobacteria and a need for more (Cheng et al., 2022; Komárek et al., 2014; Strunecký et al., 2023). Within the cyanobacteria, there are still a significant number of species missing reference sequences, as well as a lack in whole genome sequenced data for reference strains (Strunecký et al., 2023). It is important to note that the phylogenetic analysis conducted in this research only looked at taxonomy to the order level due to the low resolution of sequencing at the family level, making it apparent that there may be unknown genera or families present that

have not been considered (Cheng et al., 2022; Strunecký et al., 2023). Another limitation is the use of the V4 region of the 16S rRNA gene; although highly conserved and useful for identification purposes, using a small subset of the gene limits the amount of sequencing to analyse and future research may consider resequencing using the entire 16S rRNA gene to resolve unknown ASVs. Cyanobacteria are undergoing morphological and phylogenetic diversification at an incredible rate, and the frequency of morphologically similar taxa having distant phylogenetic positions continues to increase and complicate polyphasic methods (Cheng et al., 2022; Komárek et al., 2014; Strunecký et al., 2023). What is clear from this research and similar studies, is that V4-16S rRNA region analysis alone is not always sufficient enough to construct clear monophyletic phylogenies without including all ecophysiological and morphological factors in congruence with molecular analyses, and especially when there are still such significant knowledge gaps in unidentified taxa (Cheng et al., 2022; Komárek et al., 2014; Strunecký et al., 2023).

2.7 Implications

The results identified from this research have significant implications in that they provide well supported reasons behind why water treatment professionals and scientists should consider all cyanobacteria as potential threats when developing monitoring protocols in their watersheds. This research clearly illustrated how cyanobacteria are evenly distributed across the five major ecozones in Canada, and that it cannot be assumed that cyanobacteria are absent in their water systems simply because regular blooms are not observed. Current monitoring programs in Canada for harmful algal blooms and cyanobacteria vary by province and are most often only implemented when visual surface blooms are reported (Rashidi et al., 2021). Many WTFs and source water protection services are not concerned with regular water monitoring unless the source frequently experiences recurring algal blooms, has blooms that are significant enough to be a nuisance (e.g., visually unpleasant), or if the blooms that do occur are known to contain toxic cyanobacteria species (Rashidi et al., 2021). However, what is apparent from the phylogenetic analysis in this chapter are that cyanobacteria have an astonishing diversity

across Canada, specifically referring to the widespread distribution of species across five different ecozones and biomes. This directly implicates how source waters are monitored and how/if microbial communities are researched within a system; testing for one specific cyanobacteria species, such as *Microcystis* which is often screened for due to its association with the toxin microcystin, does not provide sufficient information as other species are likely present and can present their own biochemical risk potentials (i.e., many cyanobacteria can produce toxins, see **Table 1.1**). Considering the size of Canada, the cyanobacterial distribution and lack of biogeography imply cyanobacteria can appear virtually anywhere, at any time, and are far more widespread than initially presumed, and water monitoring programs moving forward must begin to reflect this information.

The vast diversity of cyanobacteria highlighted in **Figure 2.2** also signifies the most likely mode of transportation for cyanobacteria as being via aerial transport. In some instances, identical ASVs were observed on opposite ends of the country (in the Atlantic and Pacific Maritime ecozones, for example), a distribution most likely explained by aerial transportation as opposed to water or human transport. The absence of any local biogeography of the cyanobacteria communities supports their transport via wind through airborne droplets, as well as by being evaporated into clouds and precipitated back down to land (likely far from their origins), moving long distances throughout the atmosphere (Curren & Leong, 2020; Sheath, 2006; Wiśniewska et al., 2019; Wiśniewska et al., 2020). It is also important to highlight that there could potentially be more instances of identical ASVs spread across the country, embedded as the unidentified sequences. In short, the implications of having aerial cyanobacteria suggests the immense diversity of cyanobacteria species that are still yet unidentified by taxonomists. From a water management point of view, this means that specific water bodies should not be overlooked due to the lack of visual evidence of blooms or for not having a history of toxic cyanobacteria proliferation, because any species of cyanobacteria, toxin-producing or not, could very well be present within the water system and given the right conditions, they could proliferate. It is imperative that taxonomists continue to identify and classify cyanobacteria so that the true range of taxonomy across Canada can be determined,

especially regarding the true potential for toxin synthesis within these water systems. It is difficult to meaningfully estimate potential risks without having all the information, and the only way to achieve that is to go out and identify and classify more species.

Chapter 3

Seasonality of Cyanobacterial Toxin and Nitrogen Metabolism Genes in Turkey Lakes Watershed using Metagenomics

3.1 Introduction

Improvements to environmental DNA sequencing has vastly improved scientific knowledge and awareness of microbial community compositions in water sources and is increasingly employed to understand how organisms are adapting to climate change and anthropogenic stressors (Edge et al., 2020; Pérez-Cobas et al., 2020). In Chapter 2, one of the primary sequencing techniques, V4-16S rRNA gene amplicon sequencing, was used to characterize bacterial community compositions to genus and species levels. While 16S rRNA gene sequencing is highly effective for characterizing species and can subsequently be used to identify toxin-producing species in source waters, the presence of toxin-producing or harmful bacterial taxa does not directly correlate to an immediate threat because not all species can produce the same toxins (Romanis et al., 2021). Toxin-production is directly dependent on functional gene diversity and biochemical potential based on the expression of biosynthetic gene clusters (Hoff, 2009). Although 16S rRNA gene sequencing can target genes of interest, it is only applicable when target-specific primers are employed for the gene markers and does not provide any additional information about biochemical potential for any other genes (Pereira et al., 2018). To achieve a holistic understanding of the status of water sources, especially in a changing environment, metagenomic analysis is highly unique in that whole genome shotgun (WGS) sequencing is untargeted and provides an unbiased description of microbial communities and their biochemical potential (Pereira et al., 2018).

3.1.1 Metagenomic Sequencing

The term metagenomics refers simply to the analysis of a collection of similar items that are unalike from one another (i.e., genes are similar items, but their functions are unalike)

(Sharma et al., 2020). Also known as community or environmental genomics, metagenomic sequencing aims to analyse the entirety of genomic material observed within a given environmental sample in an efficient and unbiased manner (Pereira et al., 2018; Sharma et al., 2020). The boom in metagenomics is largely based on novel opportunities to study microbial populations in their natural environments, as opposed to using laboratory-based axenic cultures particularly since it is estimated that >99% of marine and freshwater bacteria cannot be conventionally cultured (Kennedy et al., 2011; Sharma et al., 2020). Metagenomics works by extracting and cloning genomic DNA from environmental samples, comparing sequences to other genomes from a metagenomic library (libraries are generated as more genomic DNA is sequenced and added), and screening the metadata for predicted genes, taxonomic classifications, and/or constructing new metagenomic assembled genomes (MAGs) from unassigned sequences (Kennedy et al., 2011).

There are two approaches to metagenomic studies: sequence-based and function-based, however, both utilize the same techniques for genomic DNA extractions, DNA library construction, analysis of taxonomic compositions, shotgun sequencing, and statistical analysis of uncultured samples (Kennedy et al., 2011; Pérez-Cobas et al., 2020; Sharma et al., 2020). Sequence-based (i.e., taxonomy-independent) approached metagenomics uses PCR primer pairs to target individual genes within conserved regions and recognised protein families (Kennedy et al., 2011); function-based metagenomics (i.e., taxonomy-dependent) sequence entire samples and screen against genomic libraries for all gene expressions and biochemical potential, and because there are no target sequences, functional-based methods have the option to discover novel gene classes and functions (Kennedy et al., 2011). Either strategy is widely applicable as it provides extensive detail on biodiversity, taxonomic structure, and functional gene diversity within microbial communities in eDNA and continues to progress along with next-generation sequencing the more clone libraries are constructed and bioinformatic processes improve (Pérez-Cobas et al., 2020; Sharma et al., 2020).

The primary mode of analysing WGS metagenomics is by observing the presence and abundance of genes, specifically in comparison between environmental samples (i.e., differences in gene types and abundances varying by location, season, etc.) (Pereira et al., 2018). The advent discoveries of metagenomics include advancing knowledge of macroevolution and highlighting the sheer amount of diversity that has yet to be explored. With metagenomics, partial and sometimes complete microbial genomes can be retrieved from mixed and non-axenic samples (Alvarenga et al., 2017). New and improved binning algorithms during the assembly process is advanced enough to identify and retrieve whole cyanobacterial genomes from a mix of microbial genomes which allows researchers to bypass the regular requirements needed for axenic cultures (Alvarenga et al., 2017; Pérez-Cobas et al., 2020); however, a downfall to this is that cyanobacteria are extremely underrepresented throughout genomic databases – limiting the amount of environmental, taxonomic, and geographical diversity metagenomic pipelines can access (Alvarenga et al., 2017). Limitations in databases and libraries could also result in unassigned or incorrectly selected sequences, and inaccurate gene predictions and functional assignments even after assembly and binning (Alvarenga et al., 2017; Cornet et al., 2021; Hoff, 2009; Pérez-Cobas et al., 2020). As more research is conducted and new eDNA is added into genomic libraries, NGS and shotgun metagenomics will soon play a crucial role in discovering new genes and genomes, identifying new threats, and overall expanding scientific knowledge of functional gene diversity and biochemical potential within individual communities.

3.1.2 Functional Gene Diversity (Toxins)

In the last decade, the frequency and severity of harmful algal blooms has increased globally in both fresh and marine water sources (Wells et al., 2020). With the rise of HABs, there has also been a rise in focus and research on cyanobacteria and as a result, more genera are constantly being discovered as being toxin producers for known and newly described unidentified toxins (O'Neil et al., 2012). Cyanobacteria can produce many secondary metabolites which includes toxins (any type of compound that can poison and/or cause

significant health risks to animals and humans) and other types of bioactive compounds with cytotoxic and antimicrobial properties (e.g., geosmin is a taste and odour compound produced by some cyanobacteria) (Calteau et al., 2014; Valadez-Cano et al., 2022), and it is estimated that 25-75% of cHABs are toxic (Bláha et al., 2009; O'Neil et al., 2012). Secondary metabolites are compounds produced by organisms but are not essential for their viability, growth, and/or metabolism (Bláha et al., 2009; Calteau et al., 2014; Chen et al., 2020; Memon et al., 2013; O'Neil et al., 2012). Although many cyanobacterial toxins are well known, it is also recognized that there is a substantial number of currently unknown toxins produced by cyanobacteria, including secondary metabolites that have observed toxic effects but are not yet associated with an identified compound or gene family (Romanis et al., 2021), in addition to metabolites that may only exhibit toxicity when interacting synergistically, additively, or antagonistically with other metabolites (Bláha et al., 2009). While many sequencing techniques can identify secondary metabolites, most of them require specified primers to target one protein family at a time, whereas because metagenomics is unspecified, it is highly useful in identifying a multitude of gene families within a given sample (Pereira et al., 2018).

Understanding the composition of cyanobacterial communities is a useful tool in identifying harmful species (such as ones that can produce toxins), however, the threat of toxin/bioactive compound release is very much dependent on functional gene diversity – all of which are details that taxonomy alone cannot distinguish (Affe et al., 2018). The only true way to know if secondary metabolite production is possible is to identify if functional genes are present to synthesize toxic proteins (Affe et al., 2018); even then, further research using approaches like transcriptomics or substrate-induced gene expression (SIGEX) would be necessary to determine if the functional genes are being expressed (i.e., functional genes are required to produce secondary metabolites, and activator genes are necessary for expression of encoding genes) (Kennedy et al., 2011). Cyanobacterial genomes are exceptionally complicated to evaluate because their genomes contain a high number of repeated sequences, thereby limiting the bioinformatic assembling process even with sequencing depth and normalization standards, effectively impacting genome completeness even after all algorithms

are expended (Affe et al., 2018). Even axenic cultures sequenced at high depth and using several library strategies might still prove insufficient for chromosome and plasmid reconstruction or for meeting the minimum acceptance assembly quality required by NCBI reference sequence databases (Kennedy et al., 2011). Despite difficulty assembling cyanobacterial genomes, metagenomic analysis has increased the accuracy for quantifying information of cyanobacterial taxonomic and biochemical diversity, opening more opportunities to explore taxonomic compositions, metabolic profiles, and relative gene abundances (Affe et al., 2018).

As outlined above, sequencing whole genomes from environmental samples is a valued tool that can illustrate the metabolic potential within entire communities. For this study, because the genes of interest were highly dependent upon the biochemical potential present in the metadata, specific gene pathways were not chosen prior to sequencing. Following assembly of the functional gene diversity in the metagenomic data, several gene families and pathways that could be linked to environmental or health issues within water bodies were investigated within each sample. A full list of the genes screened throughout the metagenomic samples is indicated in **Appendix E**, **Table E4**. Biosynthetic gene clusters that were not included in the analysis of this thesis were omitted due to lack of observation in the samples or an inconsistency of gene identification in published literature (e.g., *cyr* and *cyn* are sometimes both used to depict cylindrospermopsin gene families). Only genes that were confidently attributed to one gene family and had sufficient observations throughout the samples were retained for the study.

The final target genes selected for exploration were two toxin gene families: microcystin (*mcy*) and lipoxygenase (*lpx*), and three genes related to nitrogen-fixation: cyanase (*cyn*S), nitrogenase (*nif*), and a global nitrogen regulator *(ntc*A). The target toxin genes (*mcy, lpx)* were chosen for their toxic potential to contaminate source waters, and the target nitrogen-fixation genes were chosen due to their capacity of altering nitrogen inputs in source waters. Microcystin (*mcy*) was chosen for this research because of the detrimental health effects it can

have on humans and animals when ingested (Health Canada, 2020), and lipoxygenase (*lpx*) was chosen for its prevalent role in gram-negative diseases and antibiotic resistance (Moosavian et al., 2020); cyanase (*cyn*S) was chosen as cyanate is hypothesized to be a significant nitrogen source for cyanobacteria in oligotrophic systems (Mao et al., 2022), nitrogenase (*nif*), a crucial nitrogen acquisition complex was targeted as it is a crucial enzyme complex in nitrogen-fixation pathways (Detoni et al., 2022), and finally, a global nitrogen regulator (*ntc*A) was selected as it is the primary gene responsible for nitrogen assimilation (Esteves-Ferreira et al., 2017; Flores et al., 2019; Liu et al., 2022). By observing these five biosynthetic pathways, this research aims to view a holistic depiction of the seasonal distribution of toxin gene potential and nitrogen fixation in an oligotrophic watershed.

3.1.2.1 Microcystin (*mcy*)

Microcystin, the only cyanotoxin regulated in Canada, is among the most structurally diverse and largest group of cyanotoxins known to date (Health Canada, 2020; Pearson et al., 2010; Pearson et al., 2016). Produced globally, there are at least 279 variant isoforms that have already been identified (Bouaïcha et al., 2019; Du et al., 2019). Most isoforms differ by two variable X and Y amino acids, with the most common one also being the most toxic form – microcystin-LR (MC-LR) – in which leucine (L) and arginine (R) hold the variable X and Y amino acid positions (Health Canada, 2020; Jungblut & Neilan, 2006; Pearson et al., 2010; Pearson et al., 2016). As a potent hepatotoxin, microcystin can cause severe damage to the liver and kidneys, peripheral damage to the small intestines and lungs, and has the potential (although rare) to lead to fatalities via haemorrhagic shock and organ failure if chronically exposed (Health Canada, 2020; Pearson et al., 2010; Sivonen & Jones, 1999; WHO, 2003). Humans exposed to acute toxicities or short-term exposures to microcystin can become sick with gastric-enteritis which is an infection in the gastrointestinal tract that displays symptoms such as stomach cramping, diarrhea, vomiting, frequent headaches, and general joint and muscle pain (Kotak & Zurawell, 2007; WHO, 2003; Zanchett & Oliveira-Filho, 2013). The liver is the primary affected organ because 70% of microcystin and other hepatotoxins are

transported there by hepatocytes and are therefore most readily localized in the liver for detoxification (Falconer, 2008; Nishiwaki-Matsushima et al., 1992; WHO, 2003). It is so frequently found globally that the World Health Organization (WHO) implemented provisional guidelines for MC-LR in drinking water at 1.5 µg/liter, making it the only cyanotoxin to have universal guidelines set (Health Canada, 2020; Kotak & Zurawell, 2007; WHO, 2003), and many countries, including Canada (10 µg/L), have guidelines set in recreational water as well (Health Canada, 2020).

The long-term effects of microcystin have the potential to be much more serious due to *in-vitro* laboratory studies in rodents showing the possibility of tumour promotion in the liver and hepatocellular carcinoma, although less is known about these effects in humans (Falconer, 2008; Pearson et al., 2010; WHO, 2003). In 2010 it was classified as a Group 2B potential human carcinogen by the International Agency for Research on Cancer (IARC) based on the mentioned studies exhibiting a relationship between microcystin and tumour promotion (Kotak & Zurawell, 2007; WHO, 2003). It is also suggested that there may be a link between populations that have been chronically exposed to microcystin (and other cyanotoxins) and higher incidences of cancer (Falconer, 2008; Nishiwaki-Matsushima et al., 1992; Zanchett & Oliveira-Filho, 2013). It is very difficult to properly observe the connection between microcystin and cancer rates however, because it would require chronic ingestion of non-lethal doses over a long period of time and there has not been sufficient evidence to confidently state these claims (Pearson et al., 2010).

The cyclic heptapeptide structure of microcystin results in its stability and high resistance to chemical, biological, and environmental breakdown, allowing them to persist longer in water sources without being detected or removed (Falconer, 2008; Health Canada, 2020; Sivonen & Jones, 1999; WHO, 2003). The main routes of exposure for microcystin are through recreational exposure (i.e., swimming), direct consumption of drinking waters contaminated with the cyanotoxin, and inhalation (O'Keeffe, 2019; WHO, 2003; Wiśniewska et al., 2020; Zanchett & Oliveira-Filho, 2013). Because microcystin do not readily penetrate

lipid cell membranes, they do not enter tissues, suggesting that microcystin poisonings through skin absorption is not a likely source of exposure (Pearson et al., 2010; Sivonen & Jones, 1999; WHO, 2003). Instead, what makes microcystin so harmful is that they are actively transported into hepatocytes via organic anion transport proteins where they bind and inhibit protein phosphatases 1 and 2A (Falconer, 2008; Kuiper-Goodman et al., 1999; Pearson et al., 2010; WHO, 2003; Zanchett & Oliveira-Filho, 2013). The inhibition of these two protein phosphatases is the critical mechanism of toxicity because it results in hyperphosphorylation, disrupting cellular processes like intracellular regulation of structural and regulatory proteins which alter cytoskeletal structures (Falconer, 2008; Kuiper-Goodman et al., 1999; Pearson et al., 2010; WHO, 2003; Zanchett & Oliveira-Filho, 2013). Microcystin also increases the production of reactive oxygen species (ROS), thereby altering intracellular and mitochondrial enzymes (Zanchett & Oliveira-Filho, 2013); all of which create oxidative stress that are critical components to microcystin-induced apoptosis and DNA damage – two factors integral to the initiation and production of tumours and excess cell proliferation (Zanchett & Oliveira-Filho, 2013).

The microcystin gene cluster (**Fig 3.1**) spans approximately 55 kb with 10 genes organised among two divergently transcribed clusters, ranging from *mcy*A to *mcy*J (Harke & Gobler, 2015; Pearson et al., 2004). One half of the operon (*mcy*ABCDEG) contains non-ribosomal peptide synthase/polyketide synthase (NRPS/PKS) genes, comprising genes encoding peptide synthesis (*mcy*ABC), a modular peptide synthetase (*mcy*D), and two hybrid peptide and polyketide synthase modules (*mcy*EG) (Harke & Gobler, 2015; Pearson et al., 2004). The second cluster of the operon contains genes encoding tailoring enzymes (*mcy*FIJ) and is completed by the gene *mcy*H which encodes for toxin transport (Harke & Gobler, 2015; Pearson et al., 2004). Based on correspondence between the stable phylogenies of microcystin housekeeping and synthetase genes, the *mcy* gene operon most likely represents the ancestral state of many lineages because of vertical gene transfer (Tooming-Klunderud et al., 2008). Still, natural variation within the gene cluster is unavoidable and is mainly caused by spontaneous recombination (Mikalsen et al., 2003). The region within the *mcy* gene cluster

seen to experience the highest site variation is in the peptide synthetase genes for *mcy*A and *mcy*B which houses the N-methyltransferase (NMT) compound (Tooming-Klunderud et al., 2008); as an example, the *mcy*A gene in the genus *Planktothrix* has been found to not contain NMT at all, effectively changing the isoform outcome enough that MC-RR is produced instead of the most common MC-LR (arginine is present in both the X and Y position of the isoform in place of leucine in the X position) (Tooming-Klunderud et al., 2008). Phylogenetic studies on the distribution of *mcy* genes have found that since the genes cluster by function, strong selection forces have acted on the genes rather than arising via convergent evolution, suggesting the peptide synthetases are the result of horizontal transfer (Mikalsen et al., 2003).



Figure 3.1 Microcystin biosynthetic gene cluster with *mcy* encoding genes from *Microcystis aeruginosa* PCC7806 including NRPS, PKS, Tailoring, and ABC transporters
Adapted from Tooming-Klunderud et al., 2008 and Liu et al., 2022. Non-ribosomal peptide synthetase (NRPS) encoding genes (green) include *mcy*ABC, polyketide synthase (PKS) encoding genes (orange) include *mcy*DEG, tailoring enzyme genes (blue) for *mcy*FIJ, and an ABC (toxin) transporter (purple) gene *mcy*H.

3.1.2.2 Lipoxygenase (*lpx*)

Gram-negative bacteria are highly studied organisms for their regular appearance as being the cause for many public health concerns (Moosavian et al., 2020; Raetz & Whitfield, 2002; Raetz et al., 2009). The structure of gram-negative bacteria consists of an inner and outer membrane, and a thin peptidoglycan wall (Moosavian et al., 2020; Wang & Quinn, 2010). Gram-negative bacteria are notorious for infections because they are endotoxins as opposed to exotoxins. Exotoxins are highly immunogenic toxic proteins secreted by living bacteria that can be neutralized by antibodies. Endotoxins, in contrast are weakly immunogenic toxins

contained within lipopolysaccharide-protein complexes (LPS) on the bacteria themselves and are released when the cell dies and are not neutralized by antibodies (Wang & Quinn, 2010). The LPS on the outer membranes are the main culprit for gram-negative diseases as they contain several toxic components (Moosavian et al., 2020; Wang & Quinn, 2010). The composition of LPS can vary from bacteria to bacteria and affects the type and severity of toxicity elicited (Wang & Quinn, 2010). Lipopolysaccharides are composed of three distinct regions: a hydrophobic lipid A layer on the outer membrane (key toxic component and target for antibiotics), a core oligosaccharide, and O-antigen repeats (Moosavian et al., 2020; Wang & Quinn, 2010). While research on LPS is prevalent in the literature, information regarding the enzyme families that synthesize each component are less understood.

Antibiotic resistance is an increasing threat to standard medical practices, and bacterial families such as *Enterobacteriaceae* including *Escherichia coli* are multidrug-resistant (MDR) bacteria linked with life-threatening conditions ((Moosavian et al., 2020). Bacterial infections are often treated using broad spectrum carbapenem antibiotics, responsible for inhibiting cell wall synthesis, however, this being the first line of defence has caused an ongoing rise in drug resistance to this group of antibiotics (Moosavian et al., 2020). To avoid increasing the number of drug-resistant strains, colistin (polymyxin E), one of the main polymyxin class antibiotics, is designated as the last resort for MDR gram negative bacterial infections,75ncludeing those resistant to carbapenem class drugs (Moosavian et al., 2020). Polymyxin drugs work by displacing divalent cations (specifically Ca^{2+} and Mg) within lipopolysaccharides, thereby disrupting electrostatic interactions between the positively charged surface peptides and the negatively charged lipid A endotoxin (Moffatt et al., 2010; Moosavian et al., 2020). Antibiotic resistance is one of the most critical threats humans deal with today, and as more bacteria gain resistance, scientists are working tirelessly to find and produce drugs that target toxins directly at the source.

Lipid A – the chief endotoxic proponent in gram-negative bacteria and the main target for polymyxin and other antibiotic treatments – is encoded by the lipoxygenase (*lpx*) gene

cluster (Moosavian et al., 2020; Raetz & Whitfield, 2002). There are nine *lpx* genes incorporated for biosynthesis of lipid A, located across two distinct regions of the genome separated by 349 kb (Moffatt et al., 2010; Moosavian et al., 2020; Wang & Quinn, 2010). The genes *lpx*ABD are all clustered near each other **(Fig 3.2)** and function as an operon in conjunction with 11 adjoining genes, separated from *lpx*C by 71 genes (93 kb) (Moosavian et al., 2020). In order of enzymatic activity, the first reaction in lipid A biosynthesis is carried out by *lpx*A, followed by *lpx*C, *lpx*D, *lpx*H, *lpx*B, *lpx*K, and the final two genes *lpx*L and *lpx*M incorporate residues and further modifications (Moosavian et al., 2020). Interruptions, mutations, or deletions to several *lpx* genes, but specifically to *lpx*ACD, has been found to result in a loss of lipid A and therefore LPS, which in turn can lead to polymyxin or other antibiotic resistant strains (Moffatt et al., 2010; Moosavian et al., 2020; Raetz & Whitfield, 2002). This is due to *lpx*ACD being the most essential genes for lipid A synthesis and ultimately LPS; because polymyxin acts on the electric charges between lipid A and the drug, mechanisms of resistance act on modifying the net charge of lipid A (either reducing or neutralizing the charge) with correlates to interrupting the antibiotic interactions (Moffatt et al., 2010).

While the pathway of lipid A toxicity from bacteria to animals is variable depending on the species carrying lipid A, there is genetic evidence that a lipid transfer protein in plasma delivers the toxin from bacterial membranes to a signal receptor (CD14) on the surface of animal cells, which is then recognized by an innate immune system receptor, the toll-like receptor 4 (TLR4) (Raetz et al., 2009; Wang & Quinn, 2010). TLR4 interacts with lipid A and subsequently involves other accessory proteins, but the direct biochemical process at this stage is still widely unknown (Raetz et al., 2009). As more gram-negative bacterial strains are acquiring resistance mechanisms to polymyxin and other last resort antibiotics, it is crucial to learn and understand more about how lipid A works in its entirety so that new developed drugs could target different pathways and avoid increasing the number of MDR strains.



Figure 3.2 Lipopolysaccharide (LPS) gene cluster with biosynthetic *lpx* encoding genes
Adapted from Moosavian et al., 2020. Gene cluster analysed using data obtained from *Escherichia coli* K-12, including gene positions and directions.

3.1.3 Functional Gene Diversity (Nitrogen Metabolism)

Being one of the most important nutrients and biological element, the ability to utilize nitrogen from the atmosphere was a fundamental advancement in bacterial evolutionary history. Diazotrophs (bacteria capable of fixing nitrogen) account for the largest contribution of non-anthropogenic nitrogen inputs into the environment (Ininbergs et al., 2011); this is especially important in environments where nitrogen is the limiting factor for primary production (Detoni et al., 2022). Despite the importance of N_2-fixing organisms and the crucial need for nitrogen nutrients, very little is known about diazotrophic bacteria/cyanobacteria distribution in marine waters, and even less can be said for freshwater bodies. Since diazotrophic cyanobacteria can induce active systems to take up and convert different nitrogen compounds to ammonium when nitrogen is limited, this calls for specialized biosynthetic pathways with genes encoded for nitrogen-fixation (Harke & Gobler, 2015). Although ammonium is the preferred nitrogen source, cyanobacteria can also utilize nitrate, nitrite, urea, cyanate, and other amino acids high in nitrogen (Giner-Lamia et al., 2017; Luque-Almagro et al., 2008). When nitrogen supplies are critical, cyanobacteria can prolong their survival by downregulating photosynthesis, activating N_2-fixation assimilation pathways, and rely on internal N supplies (Harke & Gobler, 2015), all of which is controlled by a global nitrogen regulator (*ntc*A) and utilizes a core nitrogen assimilation complex (nitrogenase). However, the variety of nitrogen sources also means there are also enzyme complexes specific to the nitrogen source targeted. Incidentally, the same gene markers affiliated with assimilation pathways can also be used for detection of diazotrophic cyanobacteria and provides researchers a wealth of information pertaining to where N_2-fixing organisms are found and what the nitrogen inputs in

source waters is like (Detoni et al., 2022). Although the process of nitrogen-fixation is not inherently a danger to water systems or drinking water treatability, nitrogen fixation activity can be directly correlated to the success of precarious microorganisms, including toxic cyanobacteria.

The expression of nitrogen assimilation genes is controlled by the balance of carbon to nitrogen ratios within specialized cells in cyanobacteria called heterocyst's and is especially dependent on the presence of ammonium in the environment (Esteves-Ferreira et al., 2017). As the most reduced inorganic form of nitrogen, ammonium indirectly represses nitrogenase (*nif*) gene expression by blocking transcription of the global nitrogen regulator gene (*ntc*A) (Esteves-Ferreira et al., 2017). Heterocyst's evolved due to nitrogenase activity being very sensitive to the presence of oxygen, prompting the need to spatially separate the nitrogenase complex from any oxygen sources (Esteves-Ferreira et al., 2017). In heterocystous cyanobacteria, the physical separation of heterocyst's in the cell allows nitrogen fixation and photosynthesis to occur simultaneously and independently of the time of day (Sadvakasova et al., 2022); whereas, in non-heterocystous cyanobacteria, both processes must occur separately, photosynthesis during the daytime and nitrogen fixation only at night (Sadvakasova et al., 2022). Heterocyst's are differentiated by *ntc*A (see 3.1.3.3 for details) and provide nitrogen to the cells by converting ammonium to glutamine through a process called the glutamine-synthetase glutamate oxoglutarate aminotransferase cycle (GS-GOGAT) (Esteves-Ferreira et al., 2017; Harke & Gobler, 2015) before being incorporated into 2-oxoglutarate (2-OG) (Flores et al., 2019; Giner-Lamia et al., 2017).

3.1.3.1 Cyanase (*cyn*S)

Originating from the oxidation of cyanide, cyanate is an organic compound that is toxic in industrial settings, but naturally produced by physicochemical decomposition of metabolites in aquatic environments (e.g., urea) (Luque-Almagro et al., 2008; Mao et al., 2022; Palatinszky et al., 2015). Cyanate can enter the environment in several ways, mostly anthropogenically induced, including through herbicide use, and as industrial waste runoff from polymer

production (i.e., synthetic fibers) (Kebeish & Al-Zoubi, 2017), as well as by spontaneous cyanide photooxidation or oxidative treatment of cyanide discard (Luque-Almagro et al., 2008; Sáez et al., 2019). While cyanate concentrations are typically difficult to capture because it is not chemically stable, new analytical methods are becoming more accessible, such as a chromatographic method that can now detect cyanate concentrations down to nanomolar ranges (Palatinszky et al., 2015). The findings from the new detection methods are that cyanate levels are near very similar ranges to that of ammonium in oligotrophic environments and is estimated to be a significant nitrogen source for cyanobacteria under nitrogen-stress (Palatinszky et al., 2015).

The enzyme cyanase (*cyn*S) is responsible for assimilating cyanate into N sources by converting it to carbon dioxide and ammonium (Mao et al., 2022; Palatinszky et al., 2015; Sáez et al., 2019). Originally found in *E. coli*, cyanase enzymatic activity is now known to be present in many microorganisms including plants and fungi, Proteobacteria and Cyanobacteria, as well as other gram-negative bacteria (Kamennaya & Post, 2011). Its operon is light-dependent, negatively regulated by ammonium, positively regulated by cyanate or cyanide, and requires the global nitrogen regulator *(ntc*A) for activation (Kamennaya & Post, 2011; Luque-Almagro et al., 2008). Cyanase has three defined physiological roles: nitrogen assimilation being the primary role, cyanate detox as the secondary result of N assimilation, and additional metabolic regulation (Luque-Almagro et al., 2008). Depending on the organism, the cyanase encoding gene (*cyn*S) clusters differently within the operon; even among cyanobacterial strains, ABC-type cyanate transporter genes (encoding *cyn*ABD) could be separated by other genes, or be situation directly upstream from *cyn*S, and the ABC-type transporter genes in some strains can also have capacity for nitrate transport (Sáez et al., 2019).

Most research on cyanate assimilation gene clusters are conducted on marine strains of cyanobacteria and show there is a great amount of flexibility regarding genomic context of *cyn*S. **Figure 3.3** represents a standard gene pathway for cyanate transport in cyanobacterial species, based on the biosynthetic gene cluster from *Prochlorococcus* sp. MED4, in which

*cyn*S is flanked by similar *cyn* genes as well as variable unidentified genes. One area of cyanate acquisition that is universal among bacteria and cyanobacteria, however, is the control of *cyn*S by *ntc*A, though, how *ntc*A exerts its control and regulation, as well as its position on the operon, is still undefined (Kamennaya & Post, 2011). Transcript studies have found that *cyn*S levels tend to follow the same trends as *ntc*A under nitrogen-stress periods and is usually observed immediately following nitrogen reduction and decreases after enough nitrogen is depleted from the environment (Harke & Gobler, 2015; Kamennaya & Post, 2011). Although cyanase knowledge is widespread across the marine environment, much more research is required to better understand how cyanase genes cluster among freshwater cyanobacterial species, considering cyanate is hypothesized to be prevalent in oligotrophic systems (Mao et al., 2022).

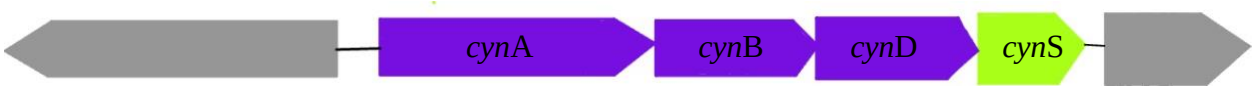


Figure 3.3 Cyanate assimilation pathway gene cluster from *Prochlorococcus* sp. MED4
Adapted from Sáez et al., 2019 and Kamennaya and Post, 2011. Gene positions and directions are included, indicating the cyanate transport genes (purple), cyanase is (lime green), and unidentified genes (grey).

3.1.3.2 Nitrogenase (*nif*)

Nitrogen assimilation occurs in over 100 genera of prokaryotic organisms in bacterial and archaeal domains, including cyanobacteria, and is largely associated with molybdenum-dependent nitrogenase enzyme complexes, encoded by *nif* genes (Detoni et al., 2022; Esteves-Ferreira et al., 2017; Latysheva et al., 2012; Li & Chen, 2020). Nitrogenase complexes comprise of two metalloproteins: a molybdenum-iron (MoFe) protein encoded by *nif*D and *nif*K, and an iron (Fe) protein encoded by *nif*H (Detoni et al., 2022; Latysheva et al., 2012; Li & Chen, 2020). While the genetic component of *nif* genes can vary greatly across diazotrophs (Detoni et al., 2022), genetic sequencing analysis have shown that nearly all diazotrophic

bacteria contain at least 6 conserved genes with the operon– *nif*BDEHKN – which is also coincidentally the minimum number of *nif* genes required to assemble the MoFe complex in in-vitro lab studies (Flores et al., 2019; Li & Chen, 2020). The genes *nif*HDK encode for nitrogenase structural subunits, often clustered together among other N assimilation genes, are expressed by a promotor upstream from *nif*B (Flores et al., 2019). The nitrogenase complex also needs at minimum 12 expressed genes to function and is activated only after the heterocyst is microoxic (low oxygen) (Thiel, 2019).

The organization of nitrogenase pathways among cyanobacterial species often differ in the order of genes and presence of accessory genes (see **Fig 3.4**); many follow a similar pattern as *Anabaena* spp. with a large operon in the order of *nif*BSUHDKENXW, interrupted only by a ferredoxin encoding gene (*fdx*N) situated between *nif*B and *nif*S, and several more genes located downstream of the operon (Flores et al., 2019; Watanabe & Horiike, 2021). Many cyanobacteria also present their *nif* gene cluster in the same or similar orders but separated over two operons with accessory proteins located between (Watanabe & Horiike, 2021). Moreover, nitrogenase activity has been found to be uninterrupted by *nif* gene mutations, suggesting that homologs of *nif* genes are present elsewhere in the genome and provoke at least some substitution functions to prevent an all or nothing pathway (Li & Chen, 2020), and that most *nif* gene mutations are neutral and do not change amino acid composition (Bolhuis et al., 2010). Phylogenetic analysis of nitrogenase genes by Bolhius et al. (2010) found that the *nif*BUS genes were closely related to other cyanobacterial homologs, yet the cluster containing *nif*HDKENB genes tended to cluster with other bacterial families such as Deltaproteobacteria and Chlorobiaceae; these findings suggest that horizontal gene transfer is a more likely explanation for the evolution of *nif*HDKENB than recombination or conjugation (Bolhuis et al., 2010).

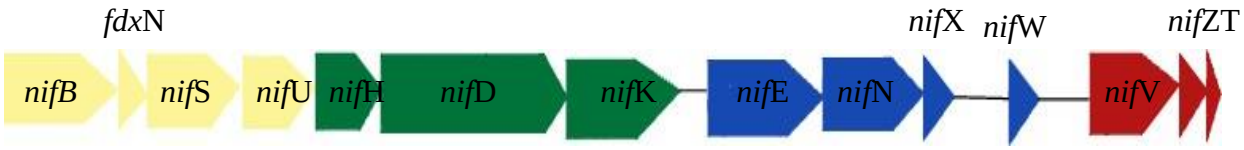


Figure 3.4 Nitrogenase biosynthetic gene cluster from *Anabaena variabilis* ATCC29413 and *Leptolyngbya borya*na dg5

Adapted from Esteves-Ferreira et al., 2017 and Thiel, 2019. Different colours represent the individual operons: *nif*B-*fdx*N-*nif*SU (yellow), *nif*HKD (green), *nif*ENXW (blue), and *nif*VZT (red). The black lines between genes represent genes that encode for proteins with unknown functions.

3.1.3.3 Global Nitrogen Regulator (*ntc*A)

The *ntc*A gene is a global transcriptional regulator that controls all nitrogen assimilation within cyanobacteria and other N_2-fixing taxa (Esteves-Ferreira et al., 2017). It is basally present in the presence of ammonium because as the preferred source of nitrogen for cyanobacteria, ammonium availability signifies there are adequate nitrogen supplies available in the environment that needs to be taken up by assimilation enzymes. When complex nitrogen sources are present or the organism is in a state of nitrogen-starvation, *ntc*A transcriptionally activates genes responsible for the nitrogenase operon complex (i.e., *nif*, *fdx*, *feo* and many more accessory genes), in addition to genes responsible for developing heterocyst's (e.g., *hep*, *hgl*, *dev*, etc.), and for nitrogen mobilization within cells (Esteves-Ferreira et al., 2017; Flores et al., 2019; Thiel, 2019). Specifically, if ammonium is present, *ntc*A promotes expression of nitrogenase complexes, and if ammonium is absent, *ntc*A activates alternative nitrogen-source assimilation gene complexes (i.e., nitrite/nitrate assimilation, cyanate, urea transport), even in cyanobacteria that do not contain heterocyst's (Esteves-Ferreira et al., 2017; Herrero et al., 2004).

Belonging to the cAMP receptor protein (CRP) family, *ntc*A transcriptionally activates or represses nitrogen assimilation pathways based on the supplies of nitrogen in the cell and

environment relative to carbon (Flores et al., 2019; Giner-Lamia et al., 2017; Thiel, 2019). Global cell responses mediated by *ntc*A are initiated by high levels of 2-oxoglutarate sensing changes in N:C ratios (i.e., low fixed nitrogen levels), prompting the binding of 2-OG and *pip*X to *ntc*A (Flores et al., 2019; Thiel, 2019). The binding of 2-OG and *pip*X initiate a cascade of regulatory proteins that activate N deprivation genes, heterocyst differentiation genes, and express nitrogenase enzyme complexes (Flores et al., 2019; Thiel, 2019). Because *ntc*A is a transcriptional regulator, its location within the operon differs by taxa, if the organism contains heterocyst's or not, and which nitrogen assimilation complex it is activating; in general though, when *ntc*A is promoting transcriptional activation it tends to be located upstream of the enzyme complex, and when it represses transcriptional function, *ntc*A is located downstream on the operon (Flores et al., 2019; Giner-Lamia et al., 2017; Herrero et al., 2004; Thiel, 2019).

3.2 Objectives

The objectives of this study were to characterize the seasonal distribution of cyanobacteria and quantify gene abundances of functional biosynthetic gene pathways for toxins microcystin (*mcy*) and lipoxygenase (*lpx*), and nitrogen-metabolism gene pathways for cyanase (*cyn*S), nitrogenase (*nif*), and a global nitrogen regulator (*ntc*A). This research was conducted with the purpose of identifying functional gene diversity among cyanobacterial phytoplankton in a thermally stratified oligotrophic water system over different seasonal periods and different depths to highlight seasonal distribution of toxin and/or nitrogen related genes. This information is critical in better understanding the potential for toxin synthesis in oligotrophic systems that are often left unmonitored, despite being systems closest to those used for drinking water treatability and recreational purposes. Toxin potential directly implicates the health of the entire water system, as well as anyone encountering contaminated water; while not inherently threatening, nitrogen metabolism affects nutrient inputs within the water column that indirectly implicates the success of phytoplankton communities, and their energy potential to form aggregates such as harmful algal blooms, and/or synthesize secondary metabolites, including those for toxin production. The low interference of anthropogenic

disturbances makes the TLW an ideal research platform to observe natural variation in phytoplankton communities and their biochemical potential in response to a changing and warming climate.

This research explored metagenomic whole genome shotgun sequencing as a tool to observe and quantify the abundance of functional genes and biochemical potential for six environmental DNA samples from Turkey Lakes Watershed (Ontario, Canada) in the Boreal Shield ecozone. The samples were all obtained and filtered by Ellen Cameron (2021) from Big Turkey Lake and sequenced (excluding one, see 3.3.1.1 for details) by the University of Toronto's Center for the Analysis of Genome Evolution and Function (Toronto, ON). To date, this research is novel as it is the first of its kind to use metagenomic sequencing to observe the seasonality of functional gene diversity among cyanobacterial communities in Big Turkey Lake.

3.3 Methods

3.3.1 Study Site: Turkey Lakes Watershed

Located approximately 50 km north of Sault Ste Marie, Ontario is a vast forested watershed primarily used as a core research site for studying the natural effects of water systems and climate change on an undisturbed environment (Jeffries et al., 1988). Beginning in 1979, the Turkey Lakes Watershed was used as a study site to research the impact of acid deposition on the aquatic and terrestrial environment. Since the conclusion of the acid deposition research, TLW has continued to be an excellent research site to observe the effects of natural climate change exacerbated disturbances in a watershed that has very little interaction with human populations. It has contributed a wealth of information regarding long term ecological assessments, sustainable forest management practices, water treatability monitoring programs (e.g., water quality observations using molecular methods), in addition to serving as a core site for training highly qualified future scientists and researchers (Webster et al., 2021). Using a site like TLW to research forested watershed management and water

treatability exemplifies how sensitive even the most isolated systems can be to natural disturbances and drives the demand for proactive research and preventative mitigation programs as an alternative to reacting after a problem (e.g., toxic bloom) has occurred.

The TLW was chosen as a research platform due to several reasons, including that there is a comprehensive description of the physical composition of the area, comprised of the four oligotrophic to mesotrophic lakes and the Precambrian silicate greenstone lining the watersheds (Jeffries et al., 1988), as well as the ongoing research that has been conducted over the last four decades which has depicted a steady increase in temperature across the region (Webster et al., 2021). The physical characteristics of the watershed results in the lakes (excluding Wishart) to thermally stratify annually and are well known to contain phytoplankton communities dominated by cyanobacteria (Cameron et al., 2022a; 2022b). Big Turkey Lake, the water body researched in detail in this study was the focus for Cameron's research (2021), where water samples were collected from various depths at different times throughout the day to observe the spatial and temporal distribution of cyanobacterial communities in a deep stratified lake.

Following the identification of potential toxin-producing cyanobacterial strains from Cameron's study (2021), the potential for cHABs was characterized by the seasonality of cyanobacteria presence. Of the samples collected by Cameron, six unused filtered samples from Turkey Lakes Watershed in the Boreal Shield were selected for metagenomic shotgun sequencing (see **Table 3.1**) based on the month and season samples were obtained, taking into consideration environmental factors, and especially focusing on how these factors relate to the presence of cyanobacteria and harmful algal blooms. This chapter focuses on which seasons cyanobacteria might be most present in, and which, if any, toxin, or nitrogen-fixation gene markers were observed at those times as well to better characterize the composition of functional genes throughout the year.

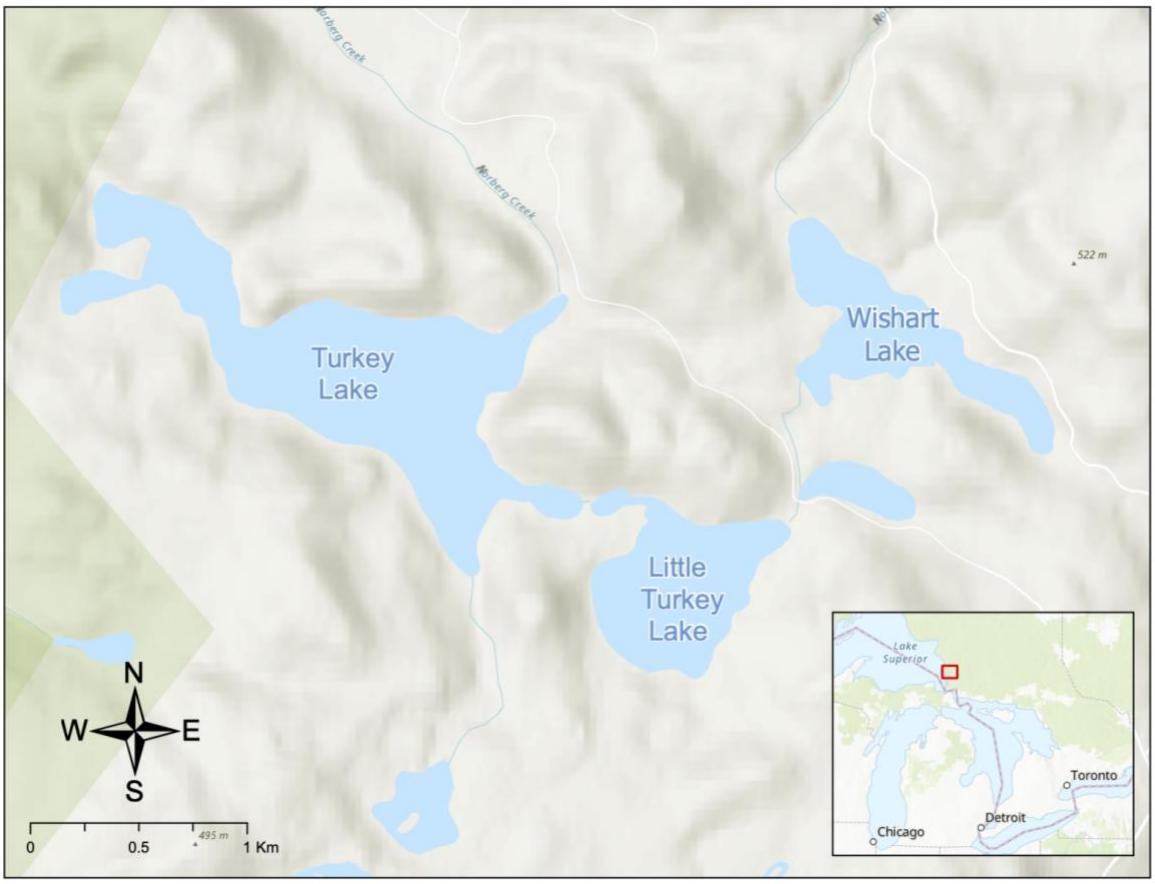


Figure 3.5 Map of Turkey Lakes Watershed, Sault Ste. Marie, Ontario (Canada)
Map generated using ArcGIS Pro [v. 3.0.1] (ESRI, 2023). The three main lakes – (Big) Turkey Lake, Little Turkey Lake, and Wishart Lake – are present approximately 50 km north of Sault Ste, Marie (Ontario, CA) and labelled within the forested Turkey Lakes Watershed.

3.3.1.1 Sample Collection

Six filtered samples from Cameron's study (2021) were chosen and reassessed for metagenomic sequencing – one sample from each summer month from May to August, and one winter sample from January. Samples (**Table 3.1**) were taken periodically from July 18, 2018, to January 23, 2020. One-liter samples were collected using a Masterflex E/S portable sampler peristaltic pump, from Secchi depth (exact depth varied), apart from sample TLW43 (July 2018) which was taken at one meter below Secchi depth and sample TLW410 (January

2020) which was taken at 0.76 m below surface due to ice cover. The water samples were vacuum filtered onto 47 mm 1.2 µm GF/C Whatman filters (Whatman plc, Buckinghamshire, United Kingdom) and stored at -20 ˚C until DNA extraction was conducted (Cameron, 2021).

Indicated in **Table 3.1** are the sample identification codes attributed to each sample date. The depth category (cat) refers to the surface (1), Secchi depth (2), and one meter below Secchi depth (3), and the specific measured depth value (m) for each sample is noted, along with the measure of Secchi depth during the time of sampling (as an indicator of water clarity). The volume of water run through vacuum filtration onto GF/C filters is stated, as well as the DNA concentrations obtained from extraction (quantified by a NanoDrop Spectrophotometer).

Table 3.1 Summary of sampling information and DNA extraction details from the Turkey Lakes Watershed, Ontario

Sample ID	Sample Date	Depth (category)	Depth (m)	Secchi (m)	Volume Filtered	DNA concentration
TLW 43	July 18, 2018	3	8	7	800 mL	11.7 ng/µL
TLW 233	May 23, 2019	2	5.25	5.25	1000 mL	1.6 ng/µL
TLW 278	June 28, 2019	2	5.75	5.75	1000 mL	5.1 ng/µL
TLW 323	July 24, 2019	2	5.25	5.25	1000 mL	3.7 ng/µL
TLW 368	Aug 21, 2019	2	5	5	1000 mL	4.9 ng/µL
TLW 410	Jan 23, 2020	N/A (ice)	Surface (0.76)	N/A (ice)	1000 mL	3.5 ng/µL

Note: DNA extraction for the TLW43 sample was conducted by Cameron (2021) and sent to a commercial laboratory (Metagenome Bio Inc., Waterloo, ON) for sequencing; the remaining samples all underwent DNA extractions and sequencing by the University of Toronto's Center for Analysis and Genome Evolution and Function (CAGEF, Toronto, ON).

3.3.2 Whole Genome Shotgun (WGS) Sequencing

The five environmental water samples (excluding TLW43) were filtered, purified, and the DNA was isolated following a modified protocol with the Qiagen DNeasy PowerSoil Kit

© (Qiagen Inc., Venlo, Netherlands) by the University of Toronto's (UofT) Centre for Analysis of Genome Evolution and Function (CAGEF). Modifications of the protocol include aseptically cutting the GF/C filters in half with 70% ethanol sterilized tools, placing them into a Qiagen PowerBead Tube with PowerBead Solution and vortexed to mix the contents, up to five times until the contents of the tube were pulverized. All following steps followed the PowerSoil kit protocol, apart from using double the reagent volumes (higher yields), and the final DNA was eluted in 100 µL of solution.

3.3.2.1 Illumina Sequencing

Illumina libraries were prepared with the Illumina DNA Prep kit (formerly known as Nextera Flex) (Illumina, San Diego, CA, USA), and the final library size was approximately 550 base pairs (including adapters). Sequencing used the Illumina MiSeq V2 chemistry with 250 x 2 PE reads. The data files obtained from the University of Toronto were raw and zipped fastq files containing a sequence identifier, raw DNA sequence, and quality scores, all of which required further processing to make the data accessible.

3.3.2.2 Bioinformatic Analysis

The raw fastq filters were processed through the Metagenome ATLAS © workflow pipeline based on Snakemate (Kieser et al., 2020). Within ATLAS, default parameters were implemented in which the raw sequences underwent standard quality control which included removal of duplicates and contamination plus sequence trimming, then assembled into high-quality scaffolds using the simple taxonomy-independent binning approach with metaSPAdes and MEGAHIT, and genomes underwent a post-binning quality assessment (checkM). Metagenomic reads were binned and assembled into contigs, final contigs were binned with MetaBAT, refined with DAStool using the search engine diamond (v2.0.4.142), and gene and taxonomic abundances were quantified (Kieser et al., 2020). Taxonomy was assigned based on the Genome Taxonomy Database (GTDB) and functional genes were predicted by Prodigal (Hyatt et al., 2010). All features were organized by eggNOG (evolutionary genealogy of genes,

non-supervised orthologous groups) respectively, and output as a detailed fna and csv file containing all functional genes and diversity, as well as additional descriptive information (Kieser et al., 2020).

Within the eggNOG output file, functional genes of interest were chosen based on toxin and/or nitrogen assimilation relevance, and contig files were extracted using a pipeline workflow in RStudio © (v. 4.1.3 "One Push-Up"; R Core Team, 2020) which produced a readable fasta sequence file for each gene. In the R environment, extracted sequences were output as fasta sequences in an accessible text (.txt) file. The annotations of genes of interest were validated by conducting homology searches using the extracted fasta sequences against the National Center for Biotechnology Information (NCBI) nonredundant (nr) database. The search results from NCBI determine and/or confirm the validity and accuracy of the sample sequences and the metagenomic annotations against valid reference sequences. All blast homology searches were done within NCBI Genbank using megablast for highly similar sequences, filtered for cyanobacteria taxid with a minimum Percent Identity of 70%, and a minimum Query coverage of 70%. Any blast results lower than the filter parameters were not included in the analysis. Of the sequences that met the percent identity and query coverage range, they were then screened again for publication source, and only blast results that belonged to peer-reviewed published literature were included. The reputable reference sequences downloaded from NCBI Blast for each gene are present in Appendix D, **Tables D1-D12**, along with the NCBI-generated accession number, protein ID, and peer-reviewed citation.

The genes selected for analysis were specific to the toxins microcystin and lipoxygenase, and nitrogen assimilation genes. For microcystin, the genes *mcy*ACDEHJ were selected, *lpx*ABCDK for lipoxygenase, cyn*S* for cyanase, *nif*JKSU for nitrogenase, and *ntc*A for the global transcriptional nitrogen regulator. The list of each toxin and nitrogen assimilation functional genes targeted for this research are indicated in **Tables 3.2** and **3.3**, respectively, including the specific functions they encode for. The functional genes were chosen based on

their abundant observations within metagenomic eggNOG database and their relevance to affecting water quality and treatment.

Table 3.2 Toxin biosynthesis gene cluster families present in TLW samples

Name	Gene	Function	References
Microcystin	*mcy*A	Encodes for non-ribosomal peptide synthetases (NRPS)	Rouhiainen et al. 2004 Zhou et al. 2021 Liu et al. 2022
	*mcy*C	Encodes for non-ribosomal peptide synthetases (NRPS)	
	*mcy*D	Encodes for polyketide synthase (PKS) type 1	
	*mcy*E	Encodes for non-ribosomal peptide and polyketide synthase fusion protein	
	*mcy*H	ABC transporter ATP-binding-like protein	
	*mcy*J	Encodes for methyltransferase	
Lipoxygenase	*lpx*A	Encodes for acyl-ADP-UDP-N-acetylglucosamine acyltransferase	Moosavian et al. 2020
	*lpx*B	Encodes for lipid-A-disaccharide synthetase	
	*lpx*C	Encodes for UDP-3-O-acyl N-acetylglucosamine deacetylase	
	*lpx*D	Encodes for UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase	
	*lpx*K	Encodes for tetraacyldisaccharide 4’-kinase	

Table 3.3 Biosynthetic genes involved in nitrogen metabolism present in TLW samples

Name	Gene	Function	Reference
Cyanase	*cyn*S	Encodes cyanate lyase/hydratase, converts cyanate to ammonium and carbon dioxide	Mao et al. 2022
Nitrogenase	*nif*J	Encodes for pyruvate-ferredoxin (flavodoxin) oxidoreductase	Dai et al. 2014 Esteves-Ferreira et al. 2017 Thiel, 2019
	*nif*K	Encodes for nitrogenase molybdenum-iron (Mo-Fe) protein subunit beta	
	*nif*S	Encodes for cysteine desulfurase	
	*nif*U	Encodes for Fe-S cluster assembly protein	
Nitrogen Transcriptional Regulator	*ntc*A	Global transcriptional nitrogen regulator	Mo et al. 2017

3.3.3 Phylogenetic Analysis

All reference sequences from peer-reviewed publications were downloaded as fasta files from NCBI GenBank and aligned against the sample sequences as a Multiple Sequence Alignment (MSA) for phylogenetic analysis. Multiple sequence alignments for each individual gene were created using Geneious © (v. 2022.1.1). Only samples and reference sequences that significantly aligned to each other were included in the MSAs, and sequences that did not align well were excluded from the final alignment and phylogenetic trees. The phylogenetic trees were generated using MEGA v. X ©, using the best predicted model for Maximum Likelihood (ML) trees resulting in the K2 + G + I (Kimura 2-parameter and Gamma distributed with invariant sites) algorithm and the Nearest-Neighbor-Interchange (NNI) ML heuristic method. The analysis resampled using bootstrap support of 1000 (i.e., resampled the sequences 1000 times and labelled the branches based on how many times the same branch was observed), and base pair lengths all differed by gene and are indicated in each figure summary. All trees were

rooted at the midpoint to best represent the evolutionary relationship among the genes and maintain uniformity among each gene analysis.

3.4 Results

3.4.1 Community Compositions

The diversity and abundance of functional genes within the eggNOG annotated file was observed to determine which functional genes were a good representation of toxin and nitrogen assimilation analyses. Immediate observations showed that the samples only contained toxin marker genes for microcystin and did not show the presence of any biosynthesis markers for any other known cyanotoxins such as saxitoxin (*sxt*), nodularins (*nda*), aplysiatoxins (*apx*) and lyngbyatoxin (*ltx*). Instead, toxin gene markers for microcystin (*mcy*) were the only cyanotoxin genes chosen for this research; the second gene marker family specific to toxin production chosen was lipoxygenase genes (*lpx*), as they were present in many cyanobacterial and bacterial taxa among all the samples and are directly involved in production of the endotoxin lipid A. Genes associated with nitrogen assimilation and metabolism were also targeted based on prevalence and abundance in the metagenomic samples; those chosen were selected due to either the capability of utilising a specific type of nitrogen source (e.g., cyanate) available in oligotrophic systems (cyanase, *cyn*S), the overall required function of nitrogen acquisition enzymes (*nif*) necessary for all nitrogen metabolism roles, and the role as the gene controlling and regulating the transcription of all nitrogen-related genes (*ntc*A).

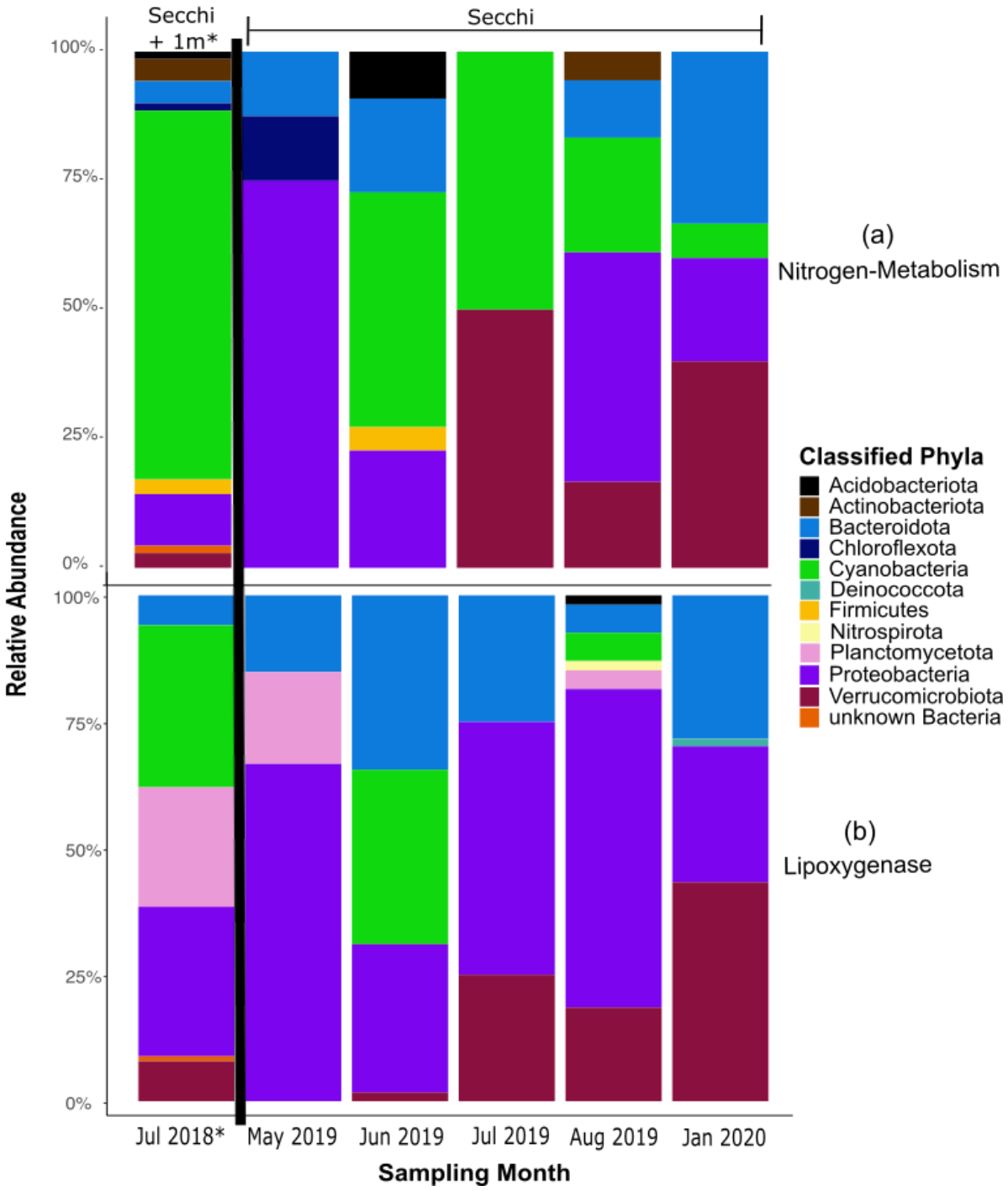


Figure 3.6 Relative abundance of TLW microbial communities containing nitrogen metabolism and lipoxygenase gene markers

From the metagenomic data, genes of interest were observed in several classified phyla, observed in the figure above each denoted by a separate coloration. All samples were collected at Secchi depth, except for the July 2018 sample which was collected at one metre below Secchi. Cyanobacteria are highlighted in lime green to distinguish itself from other bacteria. In general, the composition of bacterial taxa varies, with the highest abundances shown in purple depicting Proteobacteria, and dark red depicting Verrucomicrobiota. Note: although differentially abundant genes were observed in classified phyla during May 2019, none were observed within any cyanobacterial taxa.

Of the chosen genes of interest pathways, relative community compositions were compiled to illustrate the relative abundance of cyanobacterial taxa the genes were observed in. **Figure 3.6** describes the composition of microbial communities specific to nitrogen-assimilation and lipoxygenase gene pathways. Microcystin gene abundances were too low to represent in a figure and is instead depicted textually in **Table 3.5**. In the nitrogen-fixation pathway (**Fig 3.6a**), cyanobacteria were observed to encompass the majority percent of taxa containing nitrogen-related genes in July 2018 and July 2019, are abundant in August 2019 and January 2020, and do not contain any nitrogen-related genes in May 2019. Second to the Cyanobacteria phylum, nitrogen-related genes are primarily observed in Proteobacteria (purple), Verrucomicrobiota (dark red), and Bacteroidota (light blue). As for the lipoxygenase pathway (**Fig 3.6b**), the community composition indicates that the *lpx* genes were primarily observed in other bacterial microorganisms opposed to cyanobacterial taxa, however, some were observed in cyanobacteria in July 2018, July 2019, and in very small abundance in August 2019. Most lipoxygenase biosynthesis genes were observed in Proteobacteria (purple), followed by Verrucomicrobiota (dark red), Bacteroidota (light blue), and Planctomycetota (pink).

It is important to mention that as nomenclature and taxonomy classifications are continually edited and revised, some phyla mentioned in this report may not align with all database classifications. This includes the Proteobacteria which were reclassified as

Pseudomonadota (Garrity et al., 2021), and the Firmicutes which were reclassified to Bacillota (Gibbons & Murray, 2021); the nomenclature used in this report aligns with the classifications available at the time the research was conducted, and therefore Proteobacteria and Firmicutes are the chosen phyla names used in the community composition bar plot in **Figure 3.6**.

3.4.2 Metagenomic Gene Observations

The results of the Illumina MiSeq platform sequencing and Metagenome ATLAS © workflow pipeline assembly varied in each sample, indicating the biomass and genetic diversity within each sample fluctuated by depth and sampling month. The quality control and assembly, as determined by the shotgun sequencing and Metagenome ATLAS platform, is specified in **Table 3.4**. Within each sample, the numerical value for the number of total assembled reads (unfiltered), the number of predicted genes, the percent of assembled reads (post quality control), maximum contig base pair length, and the final number of assembled contigs is included in **Table 3.4**. Compared to the DNA concentrations specified from each sample in **Table 3.1**, the extent of genomic diversity corresponds to the amount of DNA determined by the DNA extraction protocol and the NanoDrop Spectrophotometer analysis (i.e., the highest DNA concentration was observed in TLW43, and the highest number of reads and contigs is also observed in this sample). Please refer to **Appendix E** for details regarding metagenomic sequencing parameters – a full table of metagenomic quality control and assembly supplementary information is indicated in **Table E2**, and each Gene ID with its respective sequencing depth and contig length is indicated in **Table E3**, as well a list of all possible genes and their functions that were screened for prior to analysis (**Table E4**).

Table 3.4 Assembled contigs generated through Quality Control and Assembly summaries as designated by the Metagenome ATLAS standard workflow pipeline

Sample	Assembled Reads	Predicted Genes	% Assembled Reads	Longest contig (bp)	Assembled Contigs
TLW43	3545185	193828	39.33439906	146757372	82018
TLW233	492744	27260	6.859789076	19092615	15221
TLW278	641723	45853	11.91258551	29765964	23257
TLW323	173001	9601	7.394478621	6554774	5547
TLW368	801558	38057	12.38752929	25422331	19214
TLW410	1168717	46608	15.97506411	34613319	20765

In contrast to **Figure 3.6**, relative abundance plots were not produced for microcystin gene markers due to the low abundance of genes in the metagenomic database that would skew results in a relative abundance plot (i.e., one gene hit in a sample would appear as 100% abundance). Instead, **Table 3.5** (below) identifies which microcystin gene was observed in each sample, how many gene hits were observed (abundance of gene contigs in the samples), and from which taxa they were observed in (cyanobacterial taxa are directly indicated, any gene contigs assigned to bacteria or had unknown classifications are indicated by "--"). July 2018 appeared to have the highest abundance of microcystin genes, with 20 *mcy*H genes observed in cyanobacteria, and January 2020 appeared to not contain any microcystin biosynthesis genes from under the ice surface. Aside from *mcy*D observed twice in July 2018, and *mcy*H observed six times in June 2019, all other microcystin biosynthesis genes were observed only once in various taxa (see below).

Table 3.5 Microcystin gene markers observed from metagenomic eggNOG annotated data with assigned cyanobacteria taxonomy

Sampling Date	Gene	Number of gene observations	Classified phyla	Assigned cyanobacteria taxonomy
July 2018	*mcy*D	2	Unknown taxa	--
	*mcy*E	1	Unknown taxa	--
	*mcy*H	20	Cyanobacteria	*Cyanobium*, *Synechococcus*
	*mcy*H	1	Unknown taxa	--
	*mcy*J	1	Unknown taxa	--
May 2019	*mcy*D	1	Unknown taxa	--
Jun 2019	*mcy*E	1	Unknown taxa	--
	*mcy*H	6	Cyanobacteria	*Cyanobium*
July 2019	*mcy*A	1	Cyanobacteria	Nostocales
Aug 2019	*mcy*C	1	Cyanobacteria	Nostocales
	*mcy*D	1	Unknown taxa	--
	*mcy*H	1	Cyanobacteria	*Synechococcus*
Jan 2020	--	--	--	--

The abundance of functional gene hits observed in the metagenomic data per sample season are indicative of the number of gene sequences found within cyanobacterial and bacterial communities. A description of the function of each of the target toxin (*mcy*ACDEHJ and *lpx*ABCDK) and N_2 metabolism genes (*cyn*S, *nif*JKSU, and *ntc*A) can be found in **Table 3.2 and 3.3**, respectively. The precise number of gene observations for each gene in all communities is noted in **Table 3.6**, and gene abundances for cyanobacterial communities is indicated in **Table 3.7**. The total abundance of genes in an entire sample, and the total abundance of each gene across all samples are also included in **Tables 3.6** and **3.7** for comparative purposes. The same information is further visualized across heat maps in **Figures 3.6 and 3.7** to better represent the dynamics between gene abundances and seasonality.

Table 3.6 Total abundances of target genes observed in microbial communities in TLW

	July 2018	**May 2019**	**June 2019**	**July 2019**	**August 2019**	**January 2020**	**Total**
*mcy*A	0	0	0	1	0	0	**1**
*mcy*C	0	0	0	0	1	0	**1**
*mcy*D	2	1	0	0	1	0	**4**
*mcy*E	1	0	1	0	0	0	**2**
*mcy*H	21	0	6	0	1	0	**28**
*mcy*J	1	0	0	0	0	0	**1**
*lpx*A	37	9	9	1	8	20	**84**
*lpx*B	48	1	15	2	12	12	**90**
*lpx*C	29	9	6	0	11	4	**59**
*lpx*D	55	10	17	0	14	19	**115**
*lpx*K	34	0	11	1	9	12	**67**
*cyn*S	8	3	4	0	1	0	**16**
*nif*J	9	2	2	0	2	3	**18**
*nif*K	5	0	1	0	0	0	**6**
*nif*S	16	1	6	1	4	6	**34**
*nif*U	21	2	8	1	10	6	**48**
*ntc*A	11	0	1	0	1	0	**13**
Total	**298**	**38**	**87**	**7**	**75**	**82**	

From all microbial communities (see **Table 3.6** above), the highest prevalence of toxin and nitrogen metabolism genes were found in July 2018, and the fewest observations were in July 2019. On average, the lipoxygenase and nitrogen metabolism genes appeared the most throughout all six samples; microcystin observations remained low across all samples – specifically, one hit was predominantly observed for each *mcy* gene, with a few exceptions (i.e., *mcy*H was observed 21 times in July 2018, and six times in June 2019, and *mcy*D was observed twice in July 2019). The genes synthesizing LPS were present in high abundances across all samples, except for July 2019 which had only four total hits (one *lpx*A, two *lpx*B, one *lpx*K). The highest abundance of *lpx* genes was observed in July 2018, reaching a total of 55 hits for *lpx*D, and 48 hits for *lpx*B. From the nitrogen-fixation gene pathways in all bacterial communities, the highest observations were of *nif*U, *nif*S, and *ntc*A that had 21, 16, and 11

hits, respectively, all in July 2018. In total, the most abundant gene was *lpx*D with 115 total gene hits, and the highest abundance of genes in any given sample was July 2018 with 298 total gene hits; the least abundant genes observed were *mcy*ACJ each having only one gene sequence present throughout all samples, and the sample with the least abundant genes was July 2019 with only seven total gene hits.

Table 3.7 Total abundances of target genes observed in Cyanobacteria in TLW

	July 2018	**May 2019**	**June 2019**	**July 2019**	**August 2019**	**January 2020**	**Total**
*mcy*A	0	0	0	1	0	0	**1**
*mcy*C	0	0	0	0	1	0	**1**
*mcy*D	0	0	0	0	0	0	**0**
*mcy*E	0	0	0	0	0	0	**0**
*mcy*H	20	0	6	0	1	0	**27**
*mcy*J	0	0	0	0	0	0	**0**
*lpx*A	9	0	2	0	0	0	**11**
*lpx*B	18	0	10	0	1	0	**29**
*lpx*C	10	0	1	0	1	0	**12**
*lpx*D	16	0	5	0	1	0	**22**
*lpx*K	11	0	2	0	0	0	**13**
*cyn*S	7	0	2	0	0	0	**9**
*nif*J	6	0	0	0	2	1	**9**
*nif*K	5	0	1	0	0	0	**6**
*nif*S	12	0	5	0	1	0	**18**
*nif*U	8	0	1	1	0	0	**10**
*ntc*A	11	0	1	0	1	0	**13**
Total	**133**	**0**	**36**	**2**	**9**	**1**	

Of the cyanobacterial communities exclusively, the highest gene abundances were still observed in July 2018 with a total of 133 gene hits, however no genes of interest from any biosynthetic pathway were observed in May 2019 (although cyanobacteria communities were present). The second fewest gene abundances were observed in the winter sample, in which the only gene observed beneath ice in January 2020 was the nitrogenase biosynthesis gene *nif*J. Across all samples *mcy*H was again observed the most (20 times) within the microcystin

pathway genes, *lpx*B (observed 29 times) was the most abundant within the lipoxygenase gene family, and *nif*S (observed 18 times) was the most abundant among the nitrogen-related genes.

Visual representation for the gene abundances in all microbial communities (**Figure 3.7**), and cyanobacteria-exclusive communities (**Figure 3.8**) are presented below. Warmer colours (red, orange) represent higher gene abundances compared to the cooler colour tones (blue, green) that signify low gene abundances. In July 2018 *lpx*D has the highest gene abundance, indicated by the bright red colouration, followed by a bright orange *lpx*B demonstrating the second highest gene abundance.

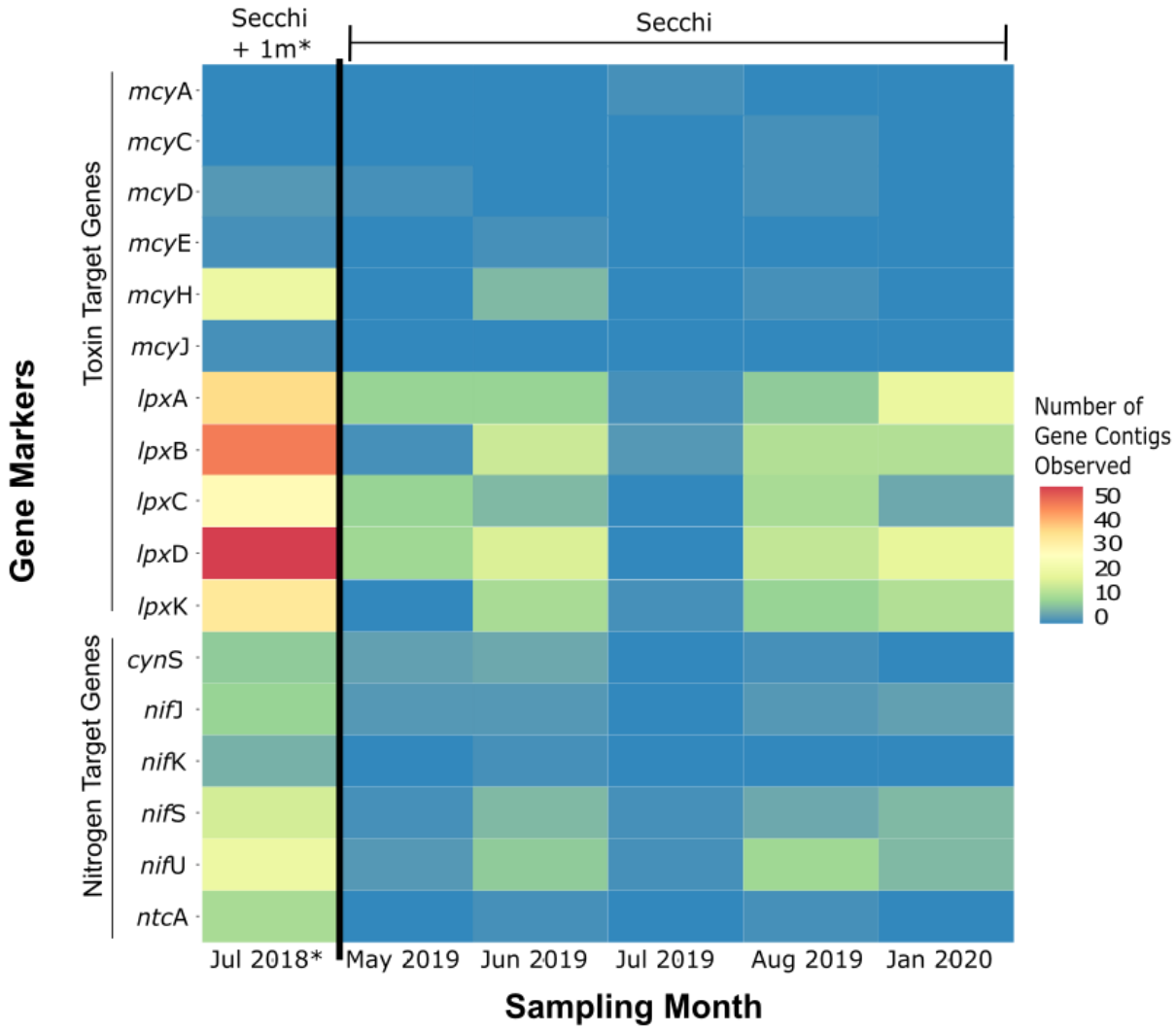


Figure 3.7 Heatmap representing seasonal abundances of target genes observations in Turkey Lakes Watershed microbial communities

The number of genes in all microbial communities (including bacteria and cyanobacteria) observed is indicated by the color gradient on the right side of the plot, with the warmer colors (orange - red) representing a higher number of gene observations within the metagenomic dataset, and cooler colors (green - blue) symbolizing fewer observations. The genes are listed on the left column and the sampling months are at the bottom of the graph. All samples were

taken at Secchi depth (except July 2018* which was taken one meter below Secchi). Please refer to **Table 3.6** for specific number of gene observations.

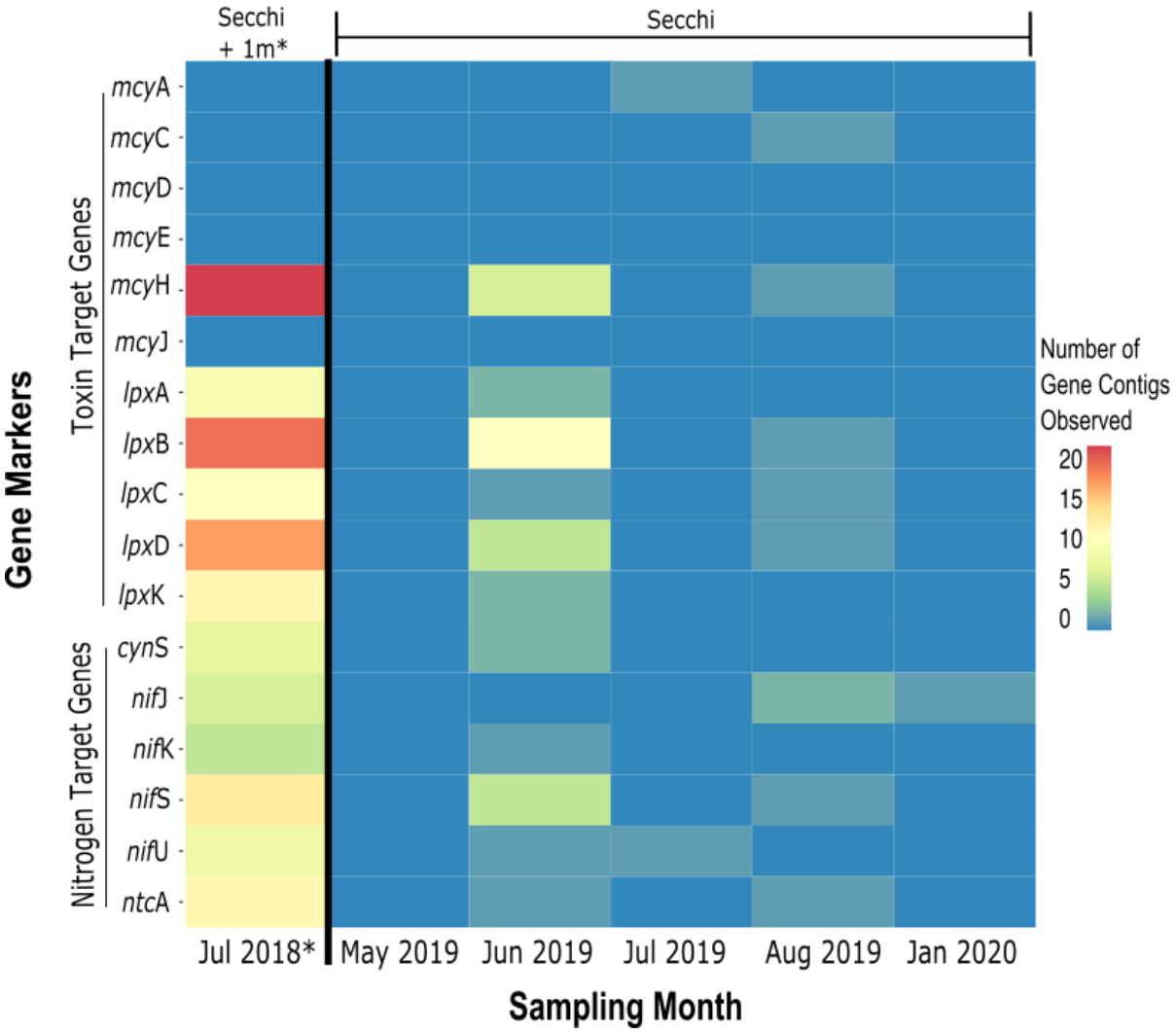


Figure 3.8 Heatmap representing the seasonal abundances of target genes observed in Turkey Lakes Watershed cyanobacterial communities

The number of genes in cyanobacterial communities observed is indicated by the color gradient on the right side of the plot, with the warmer colors (orange - red) representing a higher number of gene observations within the metagenomic dataset, and cooler colors (green - blue)

symbolizing fewer observations. The genes are listed on the left column and the sampling months are at the bottom of the graph. All samples were taken at Secchi depth, except July 2018* which was taken one meter below Secchi. Please refer to **Table 3.7** for specific number of gene observations.

3.4.3 Phylogenetic Analysis

A phylogenetic analysis for each gene of interest was constructed to validate the annotated proteins from the metagenomic dataset. Sequences all underwent multiple sequence alignments and final alignments were exported to MEGA X ©. Each tree was constructed using the maximum likelihood method and 1000 bootstrap resamples, with the Nearest-Neighbor-Interchange (NNI) Heuristic ML method. The scale at the bottom of each tree indicates the genetic distance as calculated by the default K2 + G + I algorithm set by MEGA X ©. Gene ID # for genes of interest were all arbitrarily generated by Metagenome ATLAS, and the coloured circles at each node indicate the percent of bootstrap support, as labelled by the coloured gradient to the left of each distance tree. The results of the phylogenetic trees can be regarded with high confidence due to the bootstrap analysis because higher bootstrap values indicate the consistency to which nodes were observed in the same location and among the same clades in the tree when the data was randomly resampled 1000 times (Soltis & Soltis, 2003). Nodes signified with high bootstrap percentages (70-100%) are likely good estimates of the phylogeny based on repeated random analysis of the data set (Soltis & Soltis, 2003).

Although all sequences from observed genes from the total microbial communities were included in the previous analyses, only genes observed in cyanobacterial communities were focused on for the phylogenetic analyses. Because of this, genes *mcy*DEJ are not included because they were annotated as bacterial taxa and did not have any valid blast results in NCBI. Furthermore, although all gene sequences of interest observed in cyanobacteria were aligned with reference sequences from NCBI GenBank, only sequences with significant alignment were included in the final tree. Any sequences with missing portions and/or did not align well

were excluded in the final tree. The total number of sequences and the final alignment length for each gene alignment and tree is included with each corresponding figure.

Figures 3.9 to 3.22 (below) all represent the phylogenetic analysis for the valid cyanobacterial sequences against peer-reviewed reference sequences. The bootstrap support range is present on the left side of each tree, in which low bootstrap supports are indicated by a blue tone (~0-40%), moderate support is indicated by purple (~40-70%), and high bootstrap supports (~70-100%) are indicated as red tones, and the distance matrix for each tree is designated by a black line at the bottom left-hand side of the trees.

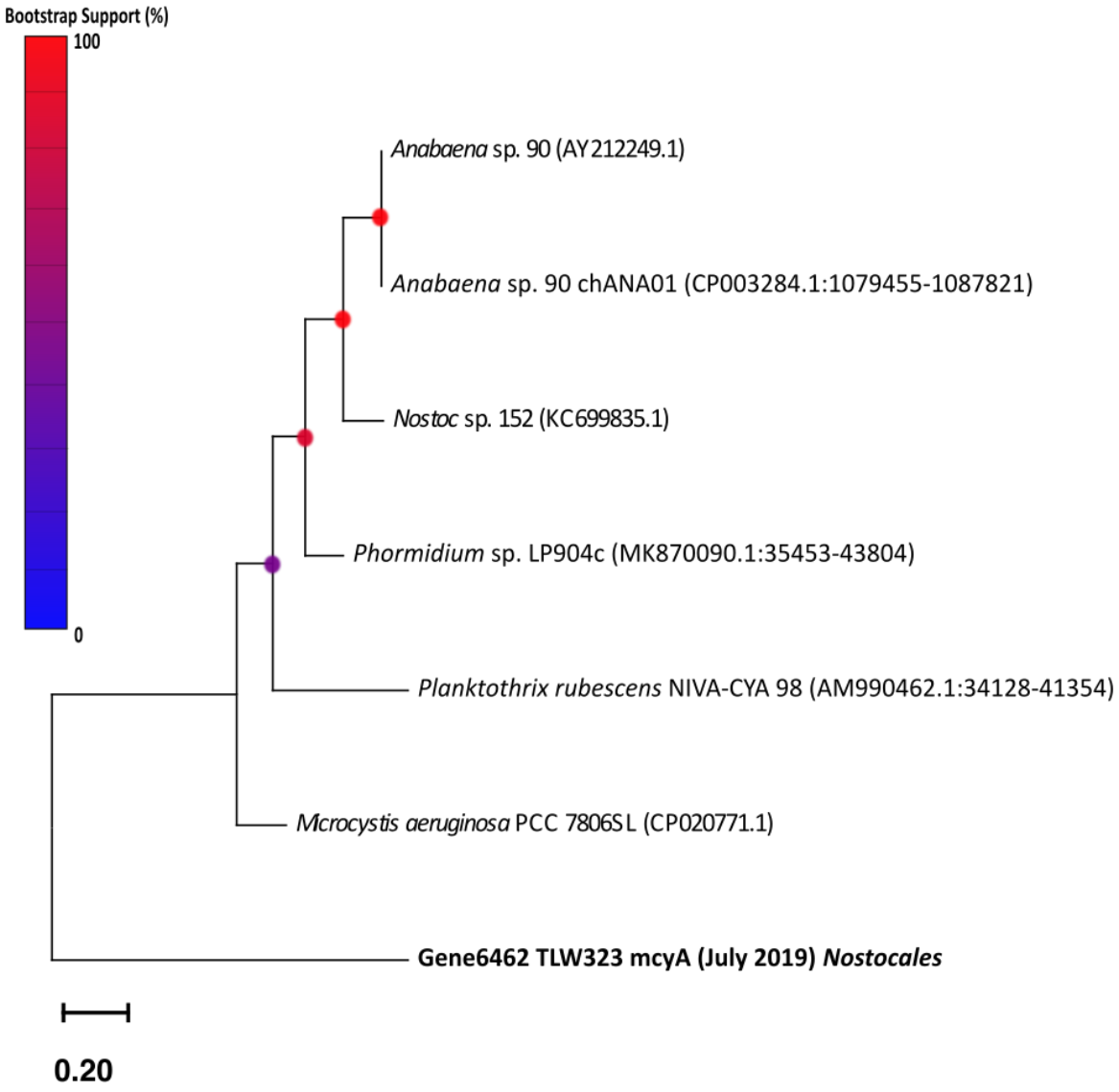


Figure 3.9 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*A, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendix D, Table D7. The length of the alignment was 1029 bp, and the final alignment included 7 sequences (log likelihood -4475.29).

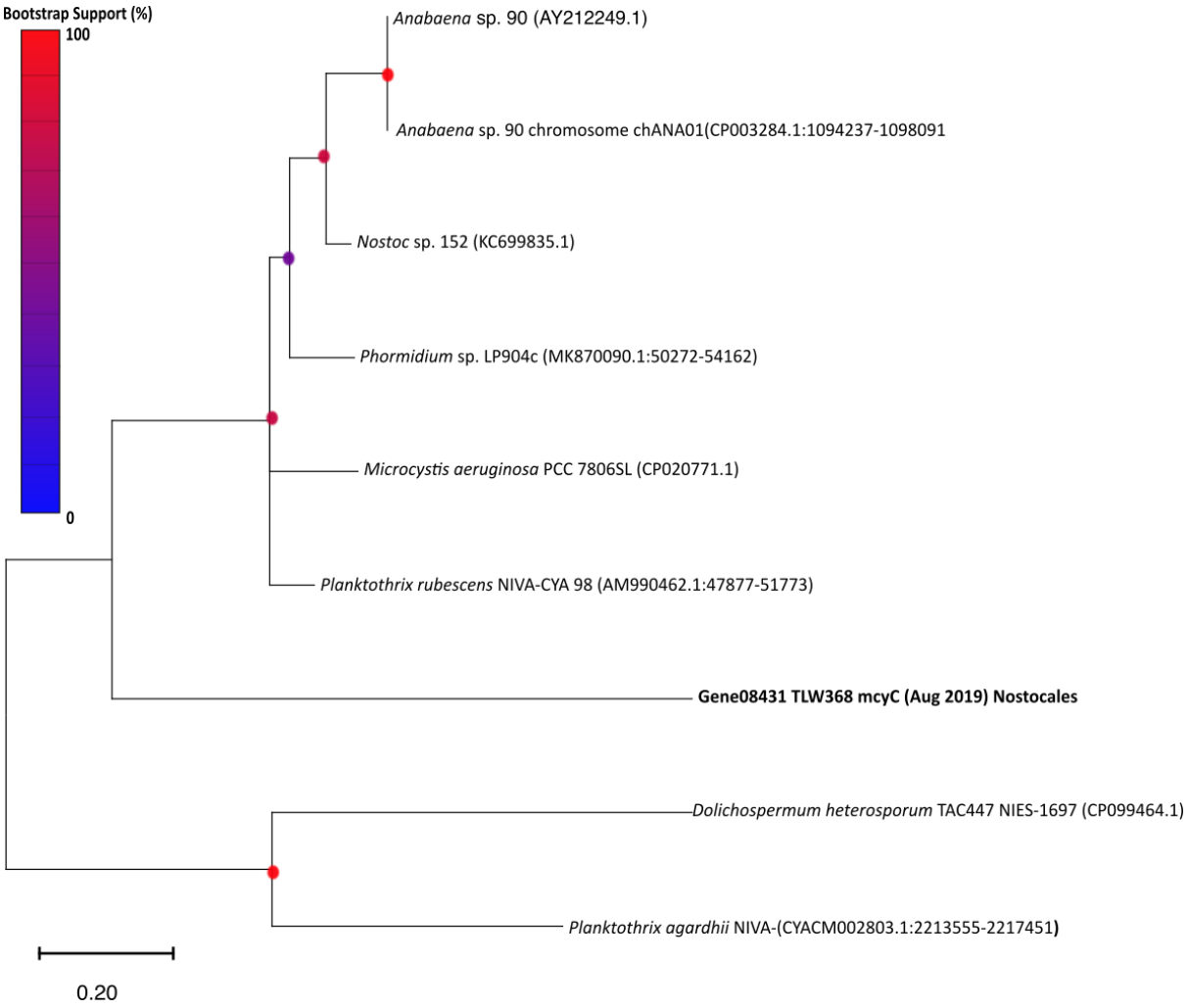


Figure 3.10 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*C, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C7/D8. The length of the alignment was 395 bp, and the final alignment included 9 sequences (log likelihood -2990.79).

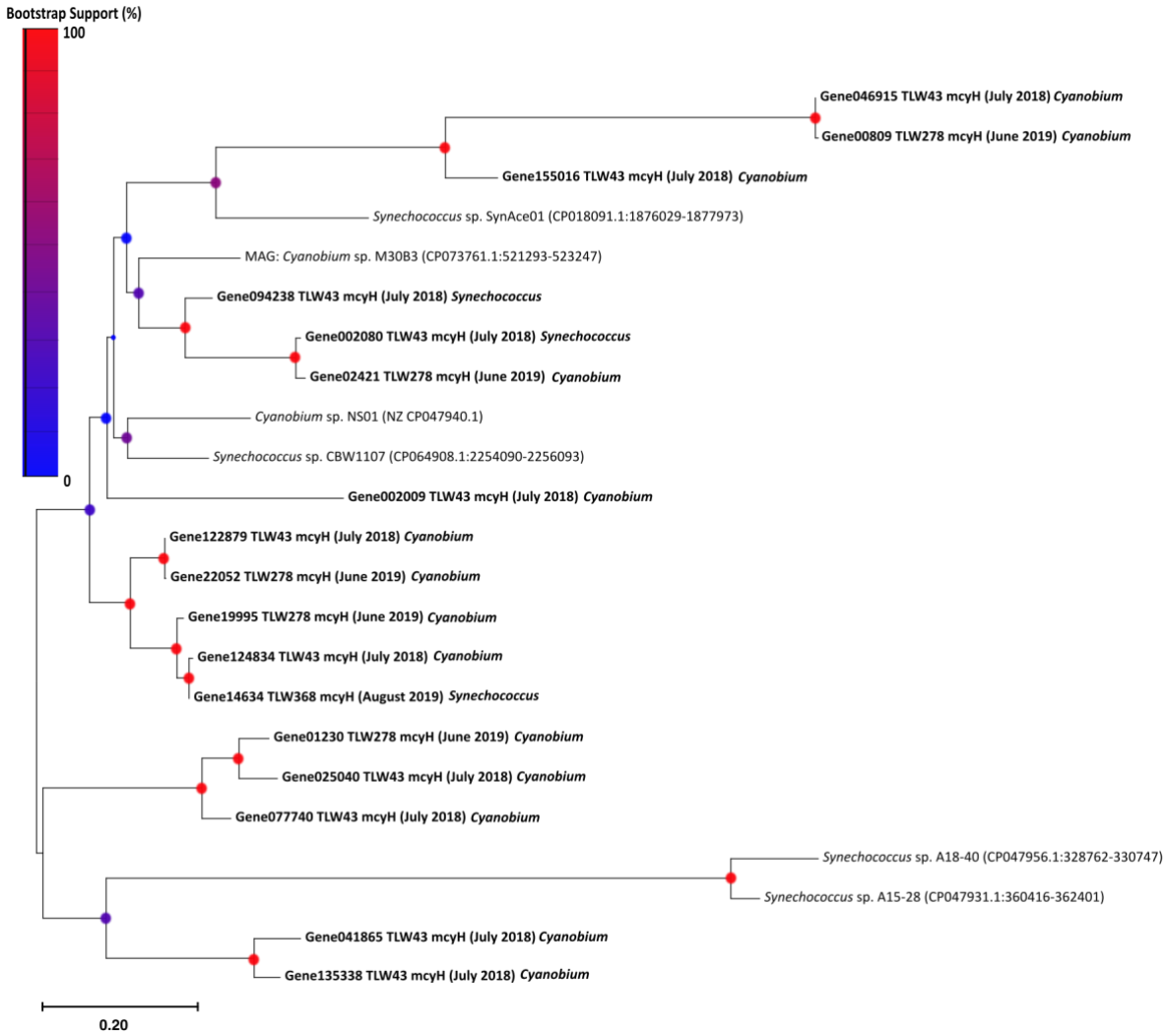


Figure 3.11 The maximum likelihood phylogenetic tree for the functional gene of interest *mcy*H, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C8/D9. The length of the alignment was 1437 bp, and the final alignment included 23 sequences (log likelihood -7184.18).

For the microcystin genes of interest, both *mcy*A (**Fig 3.9**) and *mcy*C (**Fig 3.10**) each only had one gene of interest in the phylogenetic analyses, and both did not cluster with any reference sequences, regardless of having highly supported clades in the trees; and although

there were up to 20 *mcy*H gene sequences in the phylogenetic tree (**Fig 3.11**), the genes belonged to highly supported clades while not clustering with reference sequences.

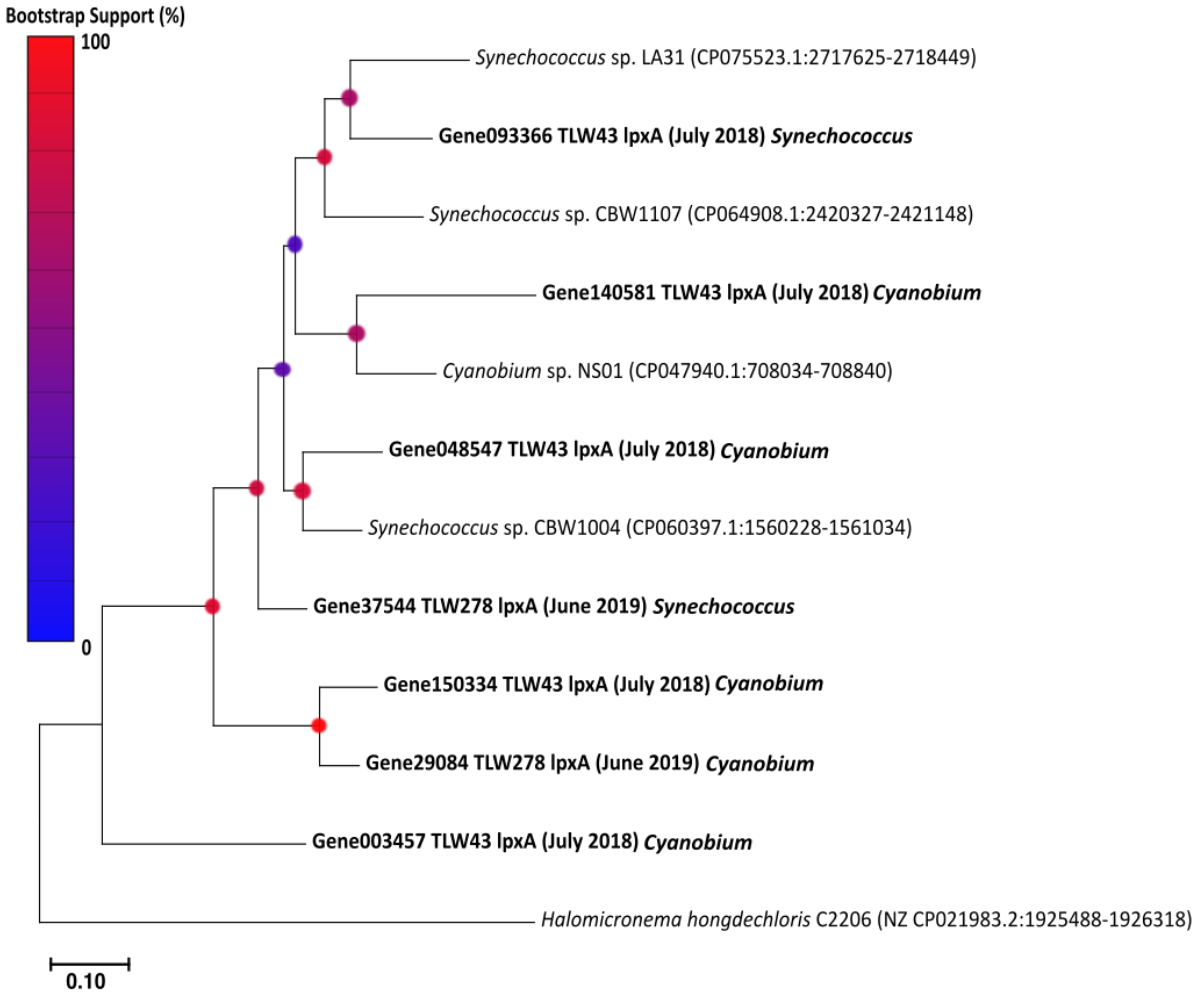


Figure 3.12 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*A, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C2/D2. The length of the alignment was 822 bp, and the final alignment included 12 sequences (log likelihood -4807.64).

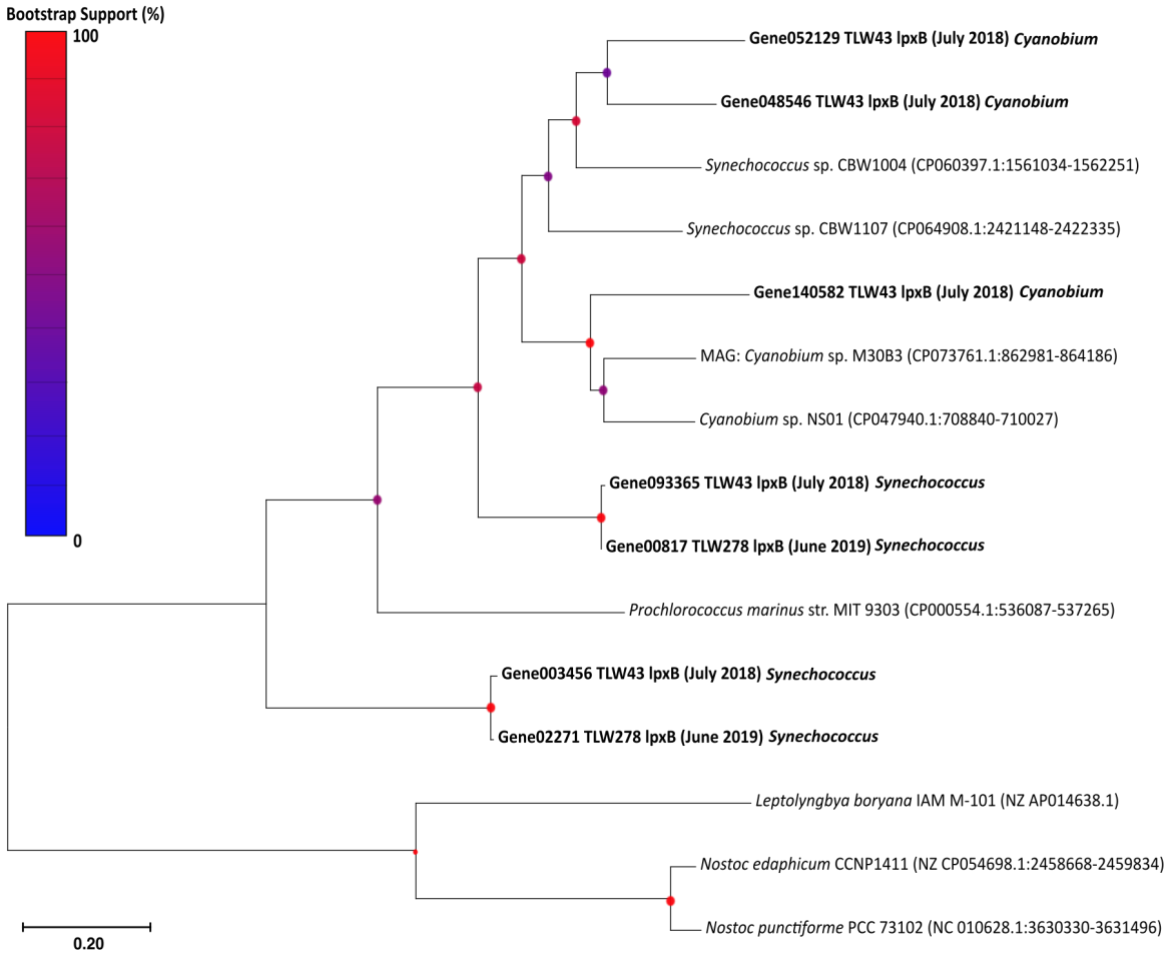


Figure 3.13 The maximum likelihood phylogenetic tree for functional gene of interest *lpx*B, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C3/D3. The length of the alignment was 1156 bp, and the final alignment included 15 sequences (log likelihood -12089.84).

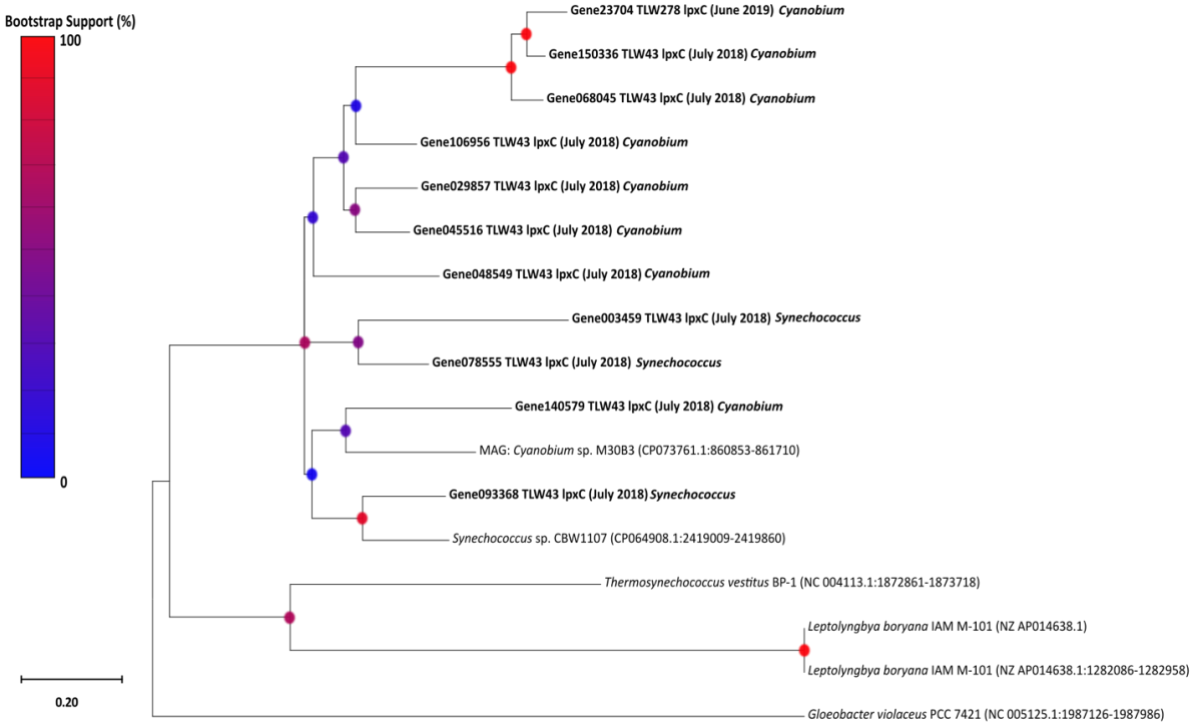


Figure 3.14 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*C, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C4/D4. The length of the alignment was 713 bp, and the final alignment included 17 sequences (log likelihood -5401.55).

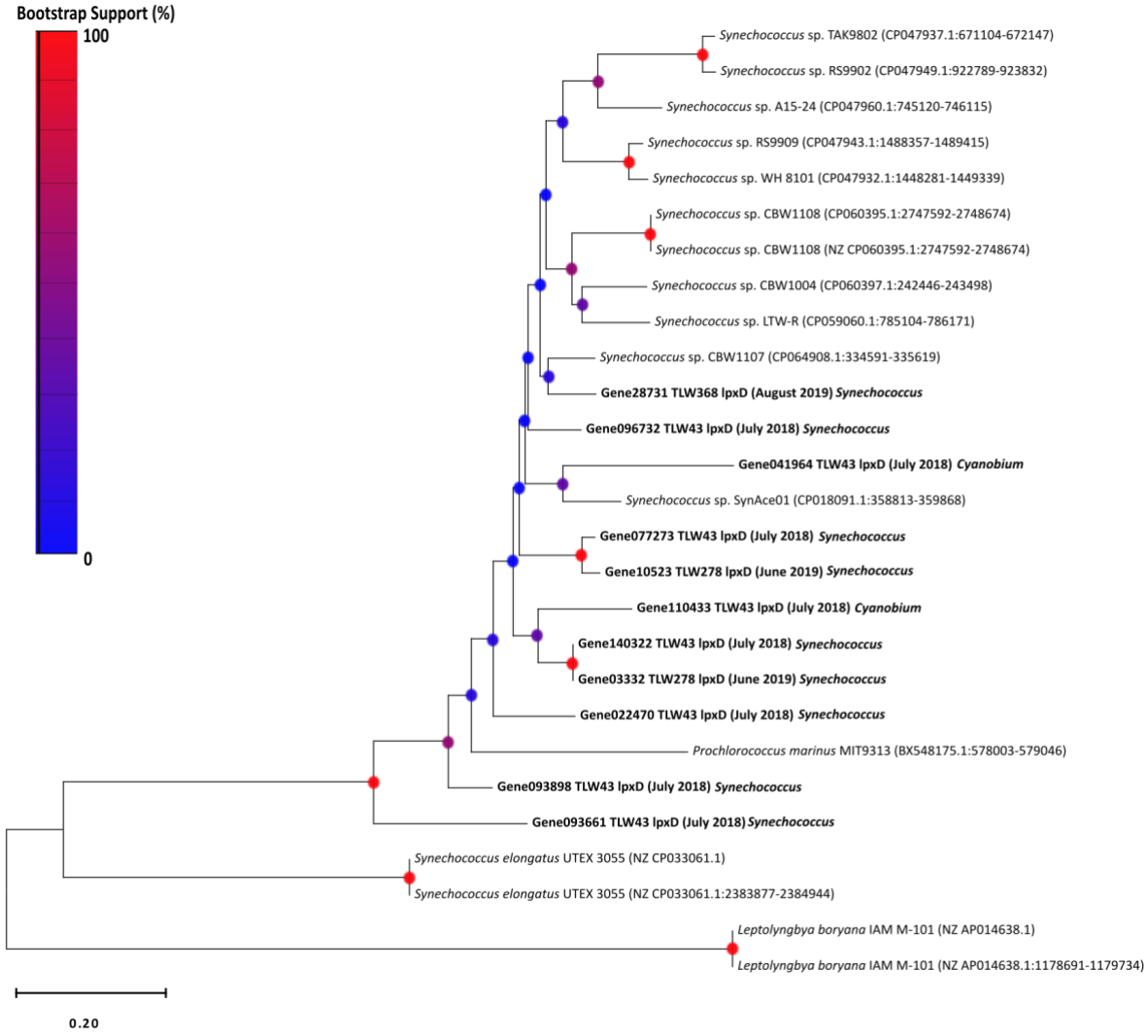


Figure 3.15 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*D, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C5/D5. The length of the alignment was 406 bp, and the final alignment included 28 sequences (log likelihood -4666.54).

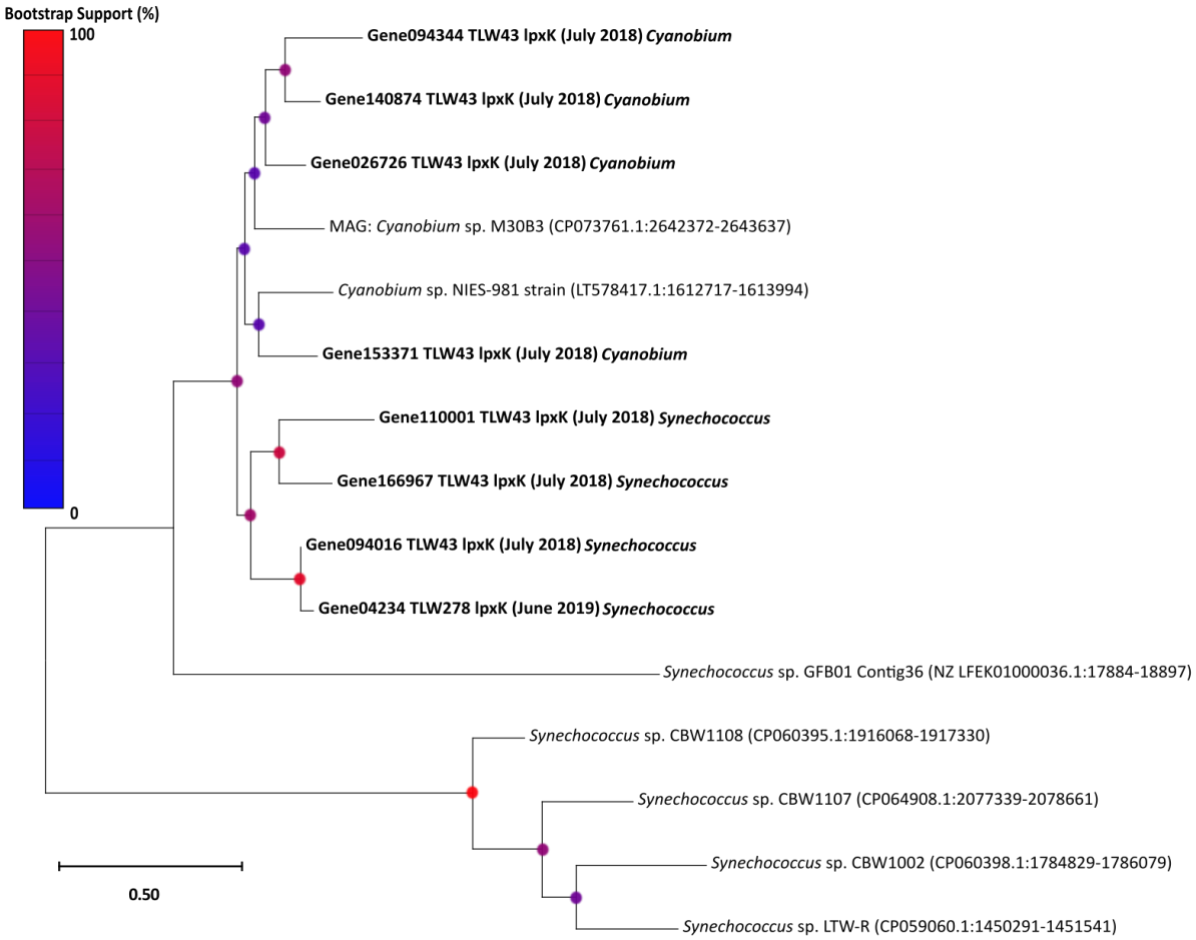


Figure 3.16 The maximum likelihood phylogenetic tree for the functional gene of interest *lpx*K, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C6/D6. The length of the alignment was 1196 bp, and the final alignment included 15 sequences (log likelihood -4847.24).

The gene sequences for *lpx*A (**Fig 3.12**) are more evenly spread out across the tree; some highly supported clades include Gene093366 clustering with *Synechococcus* sp. LA31 (CP075523.1), Gene140581 with *Cyanobium* sp. NS01 (CP047940.1), Gene048547 with *Synechococcus* sp. CBW1004 (CP060397.1), as well as Genes 37544, 150334, and 29084 which all clustered within highly supported clades but did not group with reference sequences.

Similarly, for *lpx*B (**Fig 3.13**), many of the genes belonged to highly supported clades but did not group with any reference sequences, except for Gene140582 which clustered in a high bootstrap supported clade with MAG *Cyanobium* sp. M30B3 (CP073761.1) and *Cyanobium* sp. NS01 (CP047940.1). There are only two *lpx*C gene sequences that clustered with reference sequences (**Fig 3.14**): Gene140579 had moderate bootstrap support (between 30-40%) in a clade with MAG *Cyanobium* sp. M30B3 (CP073761.1), and Gene093368 had high support with *Synechococcus* sp. CBW1107 (CP064908.1). For *lpx*D (**Fig 3.15**), the only two clades clustered with reference sequences were weakly supported, Gene28731 with *Synechococcus* sp. CBW1107 (CP64908.1), and Gene041964 with *Synechococcus* sp. SynAce01 (CP018091.1). The last gene of interest for lipoxygenase, *lpx*K (**Fig 3.16**), showed no clustering with reference sequences except for a weakly supported clade involving Gene153371 and *Cyanobium* NIES-981 (LT578417.1).

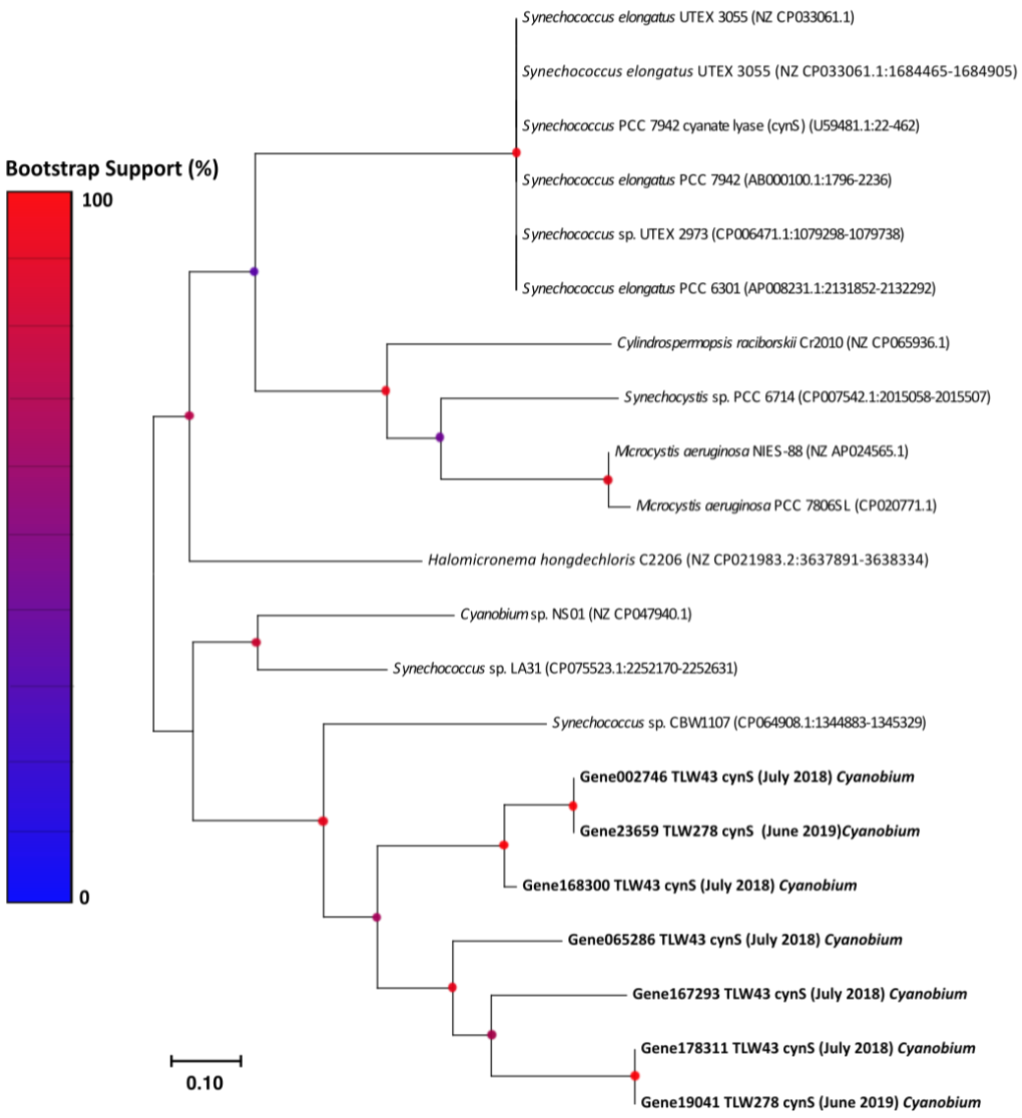


Figure 3.17 The maximum likelihood phylogenetic tree for the functional gene of interest *cyn*S, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C1/D1. The length of the alignment was 459 bp, and the final alignment included 21 sequences (log likelihood -3047.72).

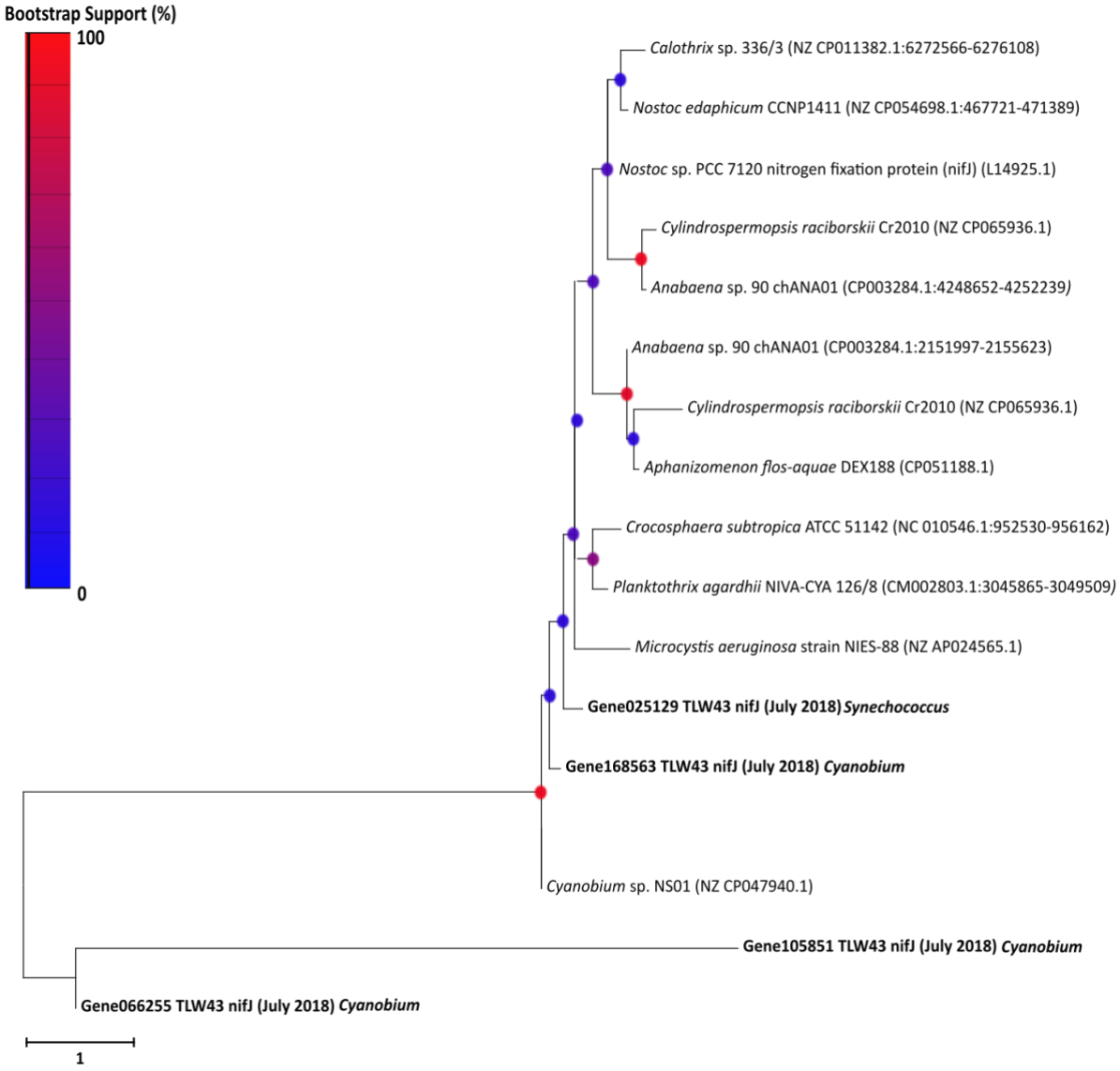


Figure 3.18 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*J, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C9/D10. The length of the alignment was 1470 bp, and the final alignment included 17 sequences (log likelihood -1090.23).

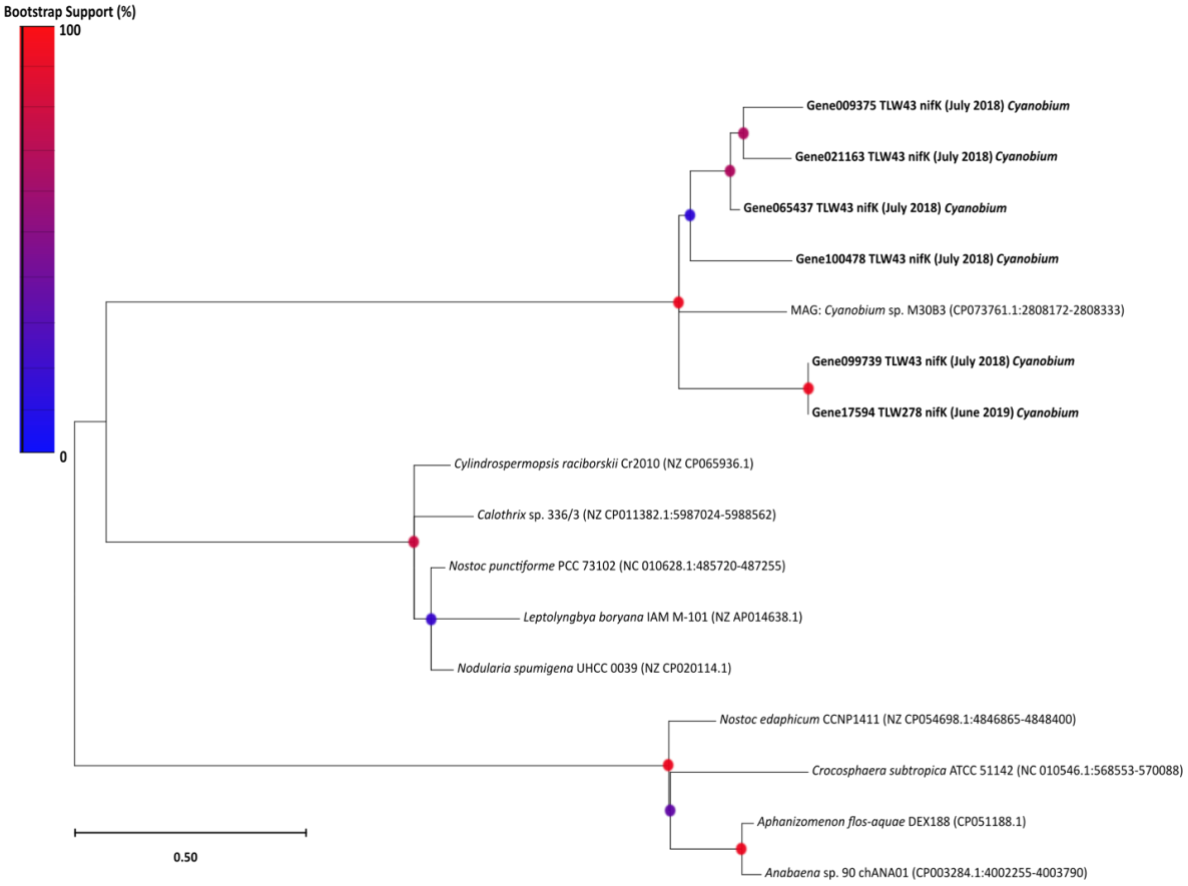


Figure 3.19 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*K, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C10/D11. The length of the alignment was 147 bp, and the final alignment included 16 sequences (log likelihood -1445.54).

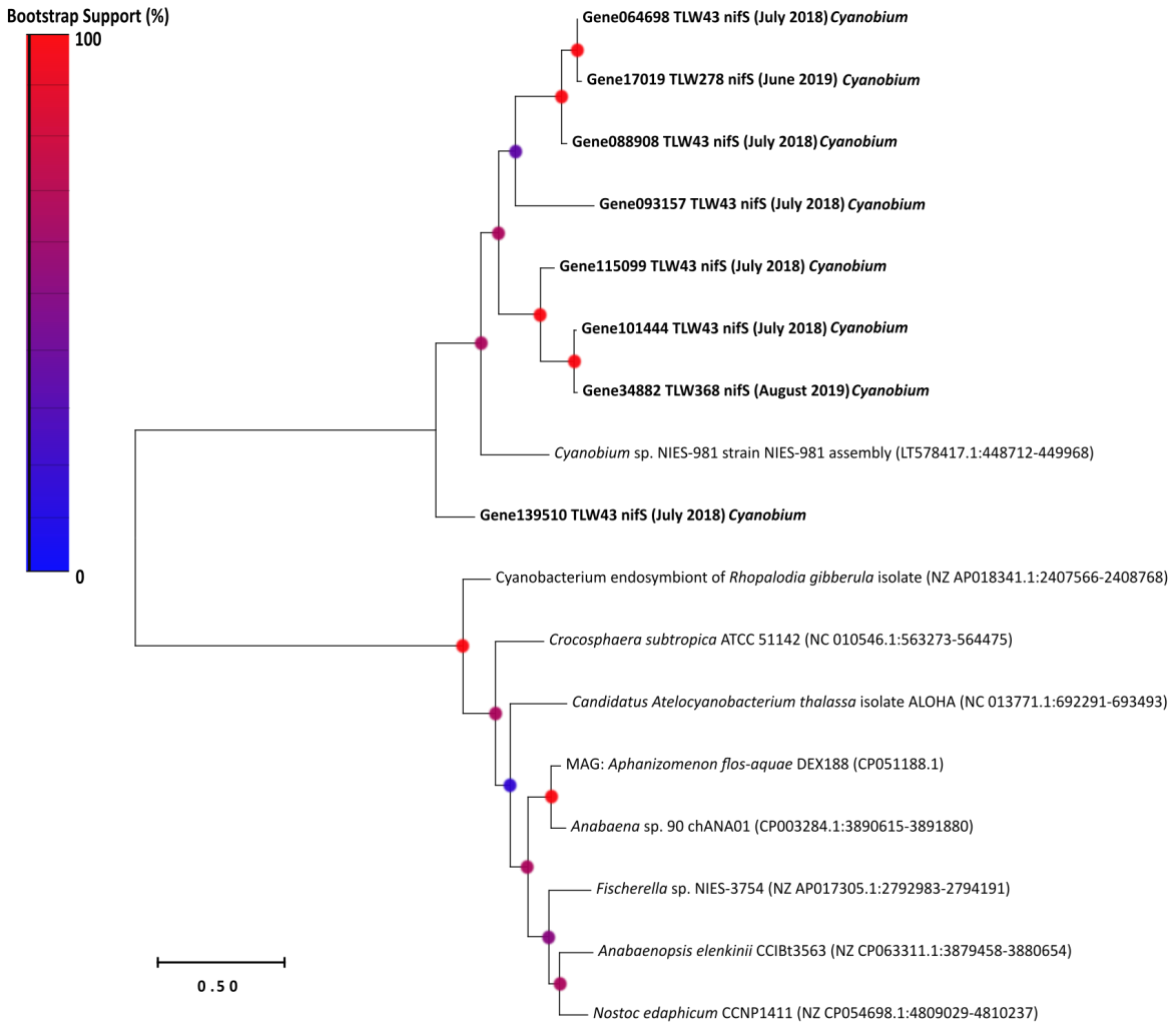


Figure 3.20 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*S, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C11/D12. The length of the alignment was 574 bp, and the final alignment included 17 sequences (log likelihood -6518.06).

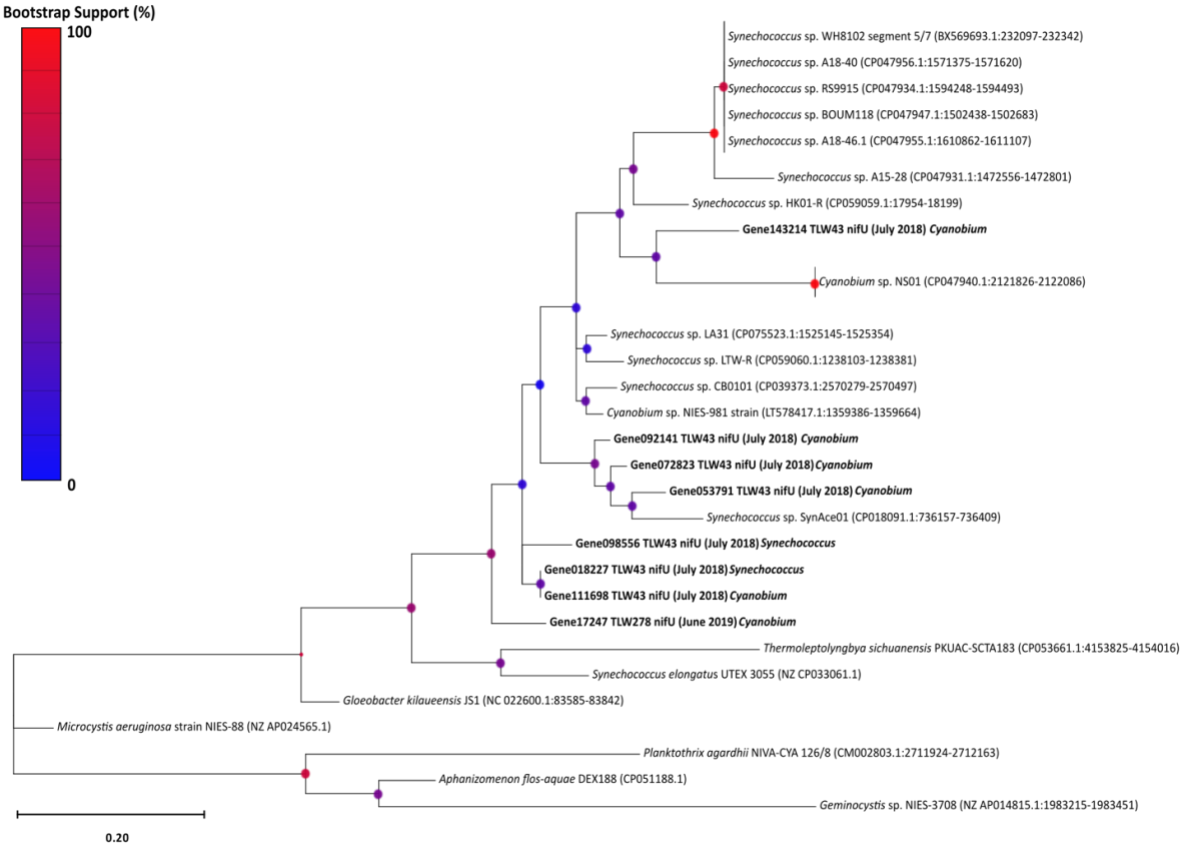


Figure 3.21 The maximum likelihood phylogenetic tree for the functional gene of interest *nif*U, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C12/D13. The length of the alignment was 172 bp, and the final alignment included 28 sequences (log likelihood -1127.44).

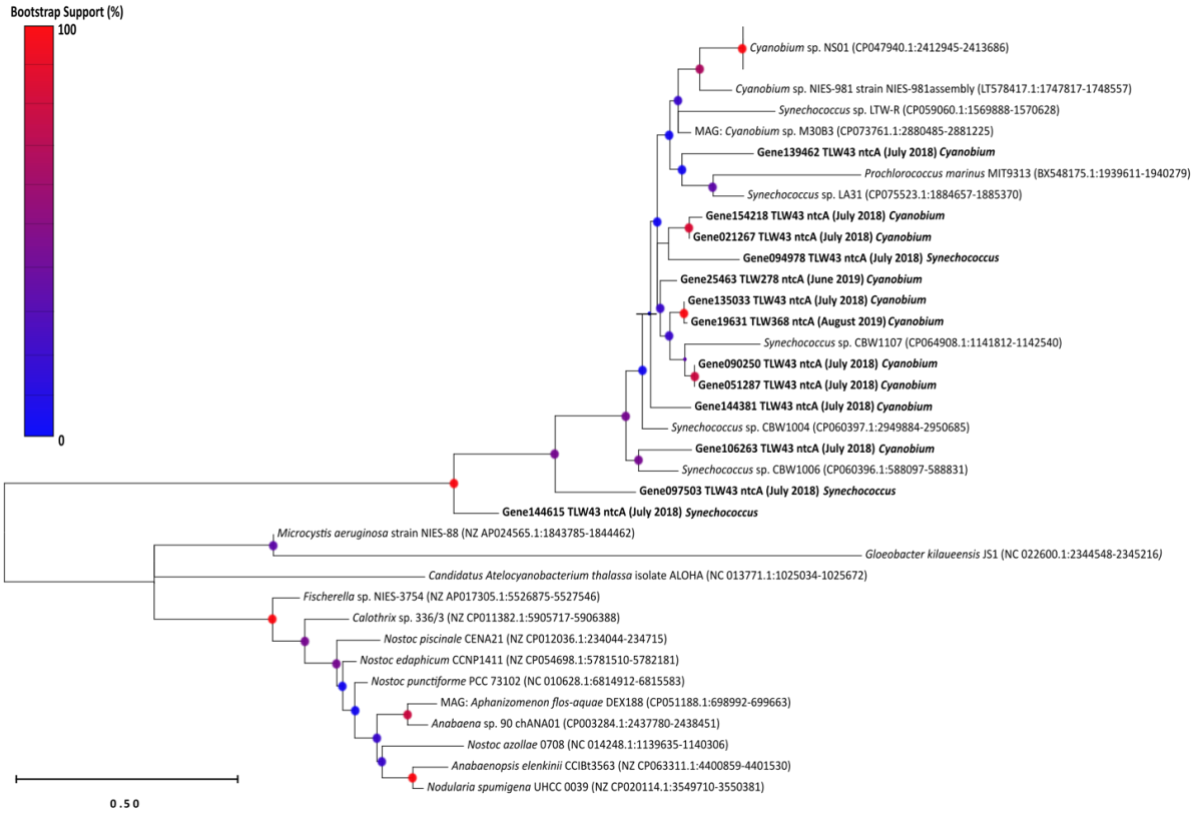


Figure 3.22 The maximum likelihood phylogenetic tree for the functional gene of interest *ntc*A, based on identified proteins by Metagenome ATLAS eggNOG annotations

Reference sequence accession numbers in NCBI GenBank databases are in parentheses and are listed in Appendices C and D, Tables C13/D14. The length of the alignment was 617 bp, and the final alignment included 35 sequences (log likelihood -2487.24).

For cyanase (*cyn*S), all the metagenomic gene sequences clustered together in highly supported clades at the bottom of the tree (**Fig 3.17**) but did not cluster with reference sequences. The phylogenetic analyses of three nitrogenase genes of interest – *nif*J (**Fig 3.18**), *nif*K (**Fig 3.19**), and *nif*S (**Fig 3.20**) – all have highly supported clades, yet no gene sequences cluster with any reference sequences. The only nitrogenase gene to cluster with reference sequences is *nif*U (**Fig 3.21**) in which Gene143214 and Gene053791 cluster with *Cyanobium* sp. NS01 (CP047940.1) and *Synechococcus* sp. SynAce01 (CP018091.1), respectively, both in weakly supported clades (<50%). The last gene of interest, *ntc*A (**Fig 3.22**) includes a few

genes clustered with references sequences. Gene090250 and Gene051287 clustered (in a clade with high bootstrap support) with *Synechococcus* sp. CBW1107 (CP064908.1), and Gene106263 clustered with *Synechococcus* sp. CBW1006 (CP060396.1) in a clade with moderate bootstrap support.

3.5 Discussion

3.5.1 Cyanobacterial Taxonomy of Target Genes in TLW

Algal and cyanobacterial blooms growth seasons have increased globally, with most research showing significant results in bloom frequency and severity trending upwards (Cameron, 2021; Wells et al., 2020). Across temperate lakes in Canada (and three oligotrophic shield lakes in Ontario specifically) increasing relative abundances of *Aphanizomenon*, *Dolichospermum*, *Microcystis*, *Gloeotrichia*, among other toxic cyanobacteria, are prompting improved bloom forecasting and overall better reporting and research on bloom formations in local freshwater lakes (Favot et al., 2023; Persaud et al., 2015; Reinl et al., 2021; Winter et al., 2011). Typical bloom protocols call for investigations to suspected lakes when taste and odour events and/or visible blooms are reported, and most often the first step is microscopic analysis of a water sample to rule out the presence of cyanobacteria (Winter et al., 2011). As mentioned in the previous chapter, there are many challenges to cyanobacteria classifications, so not properly taking the time to identify the type of bloom can be a cause for concern when trying to quickly rule out the potential for toxin presence, particularly when most response protocols do not continue if cyanobacteria is not believed to be the culprit (Winter et al., 2011). As observed in a multi-year seasonality analysis of bacteria and cyanobacteria in the temperate Turkey Lakes Watershed, cyanobacteria persist in the water column throughout the entire year, even without any visible bloom proliferations and even beneath ice-covered surfaces (Cameron et al., 2022a; 2022b; Cameron, 2021). This thesis utilizes six seasonal samples to go one step further and observe the distribution of biosynthesis genes specific to gene production and nitrogen-fixation pathways.

As mentioned previously, an overview of the complete community composition in the Turkey Lakes Watershed can be found in Cameron's dissertation (2021), including an in-depth composition of the breakdown of cyanobacteria specifically observed. For this thesis, the community composition is specific to the bacteria and cyanobacteria from which the target genes were observed in, and do not describe the whole picture of all the populations present in the samples. The breakdown of classified phyla composition specific to the nitrogen-fixation and lipoxygenase biosynthesis pathways (**Fig 3.6**) demonstrates the high relative abundance of Cyanobacteria, Proteobacteria, Bacteroidota, and Verrucomicrobiota, among others. The phyla comprising the highest proportions of abundance align with expectations as the total community composition from the Turkey Lakes Watershed as researched by Cameron (2021) also found many of the same classified phyla including Proteobacteria and Cyanobacteria within Big and Little Turkey Lake, and Wishart Lake. The phyla representing the highest proportion in Big Turkey Lake is indicative of community population analyses in other oligotrophic watersheds in Canada, such as the Atlantic and Pacific Maritime, and the Montane Cordillera, as they too have observed the same phyla from vacuum filtered water samples, and in similar proportions (Aljoudi, 2021; Cameron, 2021; Shardlow, 2021).

Due to the low abundance of microcystin biosynthesis genes, short sequence length, and sequence diversity, assigned taxonomy remained low as well. The genes that were assigned as cyanobacteria and validated through the blast search tool were attributed to one of two cyanobacterial orders: Synechococcales (*Cyanobium* or *Synechococcus*) or Nostocales (see **Table 3.5**). Many of the genes, assigned as unknown taxa in **Table 3.5** were annotated by Prodigal as belonging to various bacterial and fungal taxa such as Methylococcales and Planctomycetes; however, these contig sequences did not return any valid reference sequences upon blasting for similar sequences against NCBI GenBank and were not included in the final phylogenetic analysis due to the inconclusively of sequence taxonomy. The only instance where the eggNOG annotation was found to be incorrect is for *mcy*C that was assigned to *Ruminococcaceae* (Firmicutes), however the blast tool resulted in high sequence similarity to *Dolichospermum*, validating a cyanobacteria species opposed to bacteria and was thus included

for phylogenetic analysis. Similar studies researching the presence of toxin genes (e.g., microcystin) in oligotrophic lakes found high proportions of sequences either slipped through smaller pore size glass fiber filters (1.2 µm), or the ones that remained on the filters did not have taxonomic or functional assignments and therefore were considered unknown (Brown et al., 2021); however, additional comparative studies have found that GF/F 1.6 µm filters captured more ASVs compared to smaller filter units (e.g., 0.2 µm Sterivex filters), suggesting that majority of microbes are in fact being captured on the filters (Aljoudi, 2021). The samples in this thesis were filtered onto 1.2 µm GF/C filters, suggesting that while pore size could be a potential reason for the low taxonomic resolution within the metagenomic data, it is more likely that the unknown contigs are due to low functional annotation references in genomic libraries. Based on what is available in metagenomic reference libraries, there is a high dominance of picocyanobacteria belonging to *Synechococcus* and *Prochlorococcus* that are annotated in the bioinformatic pipelines (Affe et al., 2018; MacKeigan et al., 2022); because of the bias towards picocyanobacteria and the lack in diverse reference strains, most annotation databases label cyanobacteria as Synechococcales, however, phylogenetic analyses often suggest otherwise.

3.5.2 Metagenomic Analysis of Target Genes

Functional metagenomic approaches are the ideal mechanism for discovering novel enzymes and isolating genes based on their function rather than comparative sequence analysis (Cheng et al., 2017). Using metagenomic workflows, it is possible to identify entire biosynthetic gene clusters (i.e., a cluster involving two or more genes locally grouped together to encode a secondary product) that could be novel or well known in either cultured or uncultured strains (Chen et al., 2020). The secondary metabolites often produced by biosynthetic gene clusters do not directly affect the growth of the microorganism and are therefore not essential components, however, they can provide additional benefits to microorganism success (Calteau et al., 2014; Chen et al., 2020); instead, secondary metabolites can protect against harmful bacteria, transport agents within cells, quorum sense compounds in the environment, as well as release toxic molecules (Chen et al., 2020). The structures of

biosynthetic gene clusters form based on the secondary metabolite being encoded, including non-ribosomal peptide synthetases (NRPS), polyketide synthases (PKS), bacteriocins (inhibit harmful bacteria), and terpenes (medicinal properties) (Calteau et al., 2014; Chen et al., 2020). The most understood structures, NRPS and PKS, are well known for synthesizing pharmaceutical products (i.e., antibiotics, immunosuppressants) (Chen et al., 2020). This current study observed several biosynthetic gene clusters within cyanobacteria and bacteria communities to gain an insight to the potential secondary metabolites that can be produced, as well as the influences they may have on the microorganisms and their environment.

3.5.2.1 Functional Toxin Genes

In total, six genes from the microcystin biosynthesis gene cluster (*mcy*ACDEHJ) were identified and annotated as either NRPS, PKS, and/or ABC transporters related to the biosynthesis of microcystin within the metagenomic data (Rouhiainen et al., 2004). The entire gene cluster in **Figure 3.1** depicts the individual genes related to microcystin production, their sequencing order and directions, and their relative gene sizes. Although there are slight differences in the MC biosynthesis pathways of different species, microcystin is bidirectionally synthesized by ten proteins, encoded by a gene cluster spanning 55 kb divided onto two operons (*mcy*A-C and *mcy*D-G), through a series of complex catalytic reactions (Harke & Gobler, 2015; Liu et al., 2022; Pearson et al., 2016; Zhou et al., 2021). In phonetic order (not the order the genes appear on operons), *mcy*A (~8300 bp) and *mcy*B (~6300 bp) encode conserved proteins that are essential for microcystin synthesis (in *Microcystis mcy*A protein encodes NRPS, in *Anabaena mcy*B encodes a polypeptide) (Zhou et al., 2021); *mcy*A is positively influenced by nitrogen levels, whereas *mcy*B has higher expression levels in higher temperatures (Scherer et al., 2017) and nitrogen stress (Zhou et al., 2021). Next, *mcy*C (~3800 bp) activates amino acids, and *mcy*D (~11,700 bp) encodes a polypeptide like PKS, which along with *mcy*E (~10,400 bp), encodes another polypeptide that directly affects the synthesis of Adda (an amino acid with hepatotoxic activity) that influences the overall structural integrity and toxicity of MC (Pearson et al., 2010; Zhou et al., 2021). An amino acid racemase (catalyzes

the conversion of L-amino acids to D-amino acids) is encoded by *mcy*F (~700 bp) (Liu et al., 2022), and the beginning of the biosynthetic pathway, containing the open reading frame, is encoded by *mcy*G (Zhou et al., 2021). The last biosynthesis genes include *mcy*H (~1600 bp) – which belongs to an ABC-transporter superfamily and is believed to be responsible to transporting MC out of the cell (Harke & Gobler, 2015; Pearson et al., 2004; Zhou et al., 2021), *mcy*I (number of bp unknown) – a gene not well studied and hypothesized to encode several different proteins, and finally the last gene in the cluster, *mcy*J (~900 bp) – encodes for methyltransferase and is located at the end of the operon immediately before the stop codon (Liu et al., 2022; Zhou et al., 2021).

As shown in **Figure 3.7** (all microbial communities) and **Figure 3.8** (only cyanobacteria), there are microcystin biosynthesis genes present in the metagenomic data, albeit at low abundances. While this does not confirm the presence of MC in the watershed, it is a good indicator for the capacity of toxin synthesis in an oligotrophic watershed. Many studies have researched the mechanism behind microcystin synthesis, as well as its function and importance to the microorganisms, and although many theories have circulated, it continues to be a strong area of focus. On a cellular level, microcystin is suggested to play critical roles in intracellular functions by acting as a stabilizing agent in times of oxidative stress, keeping critical enzymes and proteins from denaturing during high light or temperature periods (Scherer et al., 2017; Zhou et al., 2021). This supports previous studies that found several MC regulatory genes were often upregulated in response to environmental stressors including increasing temperature (*mcy*B), viral plankton (*mcy*BD), changing light intensities (*mcy*AD), phosphorus deficiency (*mcy*D), and nitrogen starvation (*mcy*A-J) (Harke & Gobler, 2015; Liu et al., 2022; Scherer et al., 2017; Zhou et al., 2021). Although it is difficult to determine which, if any, of these factors were occurring over the 6-month sampling period at TLW, it is possible to extrapolate theories about environmental conditions based on the observations of specific *mcy* genes.

Looking at the *mcy* genes observed in cyanobacterial communities (**Figure 3.8**) in TLW over 6 months, *mcy*A was observed once in July 2019, *mcy*C was observed once in August 2019, and *mcy*H was observed a total of 27 times over July 2018, June 2019, and August 2019. Although it is quite possible the *mcy* genes observed in TLW could be in response to environmental factors, it is also possibly due to the stability of individual genes, and the availability and length of gene sequences in metagenomic libraries resulting in identifiable contigs. Many studies on the *mcy*H gene have suggested it is highly stable, and that there is a link between its function and the release of microcystin into the environment (Harke & Gobler, 2015; Pearson et al., 2004). As an ABC transport protein, *mcy*H supports the transition of microcystin from the cell into the environment (Harke & Gobler, 2015; Liu et al., 2022; Pearson et al., 2004; Rouhiainen et al., 2004; Zhou et al., 2021). The deletion of *mcy*H under standard lab conditions resulted in the abrupt elimination of microcystin production, and while the exact mechanisms of *mcy*H are not well understood (whether it exports the toxin outside of the cell or to the inner membrane), it is speculated to be a stabilizing anchor between the NRPS/PKS complex and the membrane (Harke & Gobler, 2015; Liu et al., 2019; Pearson et al., 2004). The observations of *mcy*H in TLW, in congruence with *lpx* and nitrogen metabolism genes, suggests potential links between microcystin biosynthesis and the environment, especially regarding nitrogen availability at those sampling points.

The microcystin compound contains seven nitrogen atoms within its chemical structure, making it a nitrogen expensive molecule to produce, suggesting its dependence on available nitrogen in the environment to be synthesized (Harke & Gobler, 2015; Liu et al., 2022). Transcriptomic studies on the *mcy* gene cluster in response to nitrogen levels found that transcription rates were directly influenced by N supplies (Harke & Gobler, 2015). In N depleting conditions, transcription for genes responsible for peptide synthesis (*mcy*ACDE) were downregulated, while genes involved in tailoring and transport (*mcy*HIJ) were upregulated (Harke & Gobler, 2015). Because microcystin is a nitrogen-rich molecule, low availability of N affects the amount of microcystin that can be synthesized. Harke and Gobler (2015) theorized that tailoring/transport proteins increased in low N conditions because the

cells could be modifying the microcystin isomer to one that requires less nitrogen atoms or could be disassembling the entire compound to preserve nitrogen for the cells (all of which could affect MC toxicity). Many of these theories highly depend on the reasoning behind microcystin function, whether that be because it is a quorum sensing compound that can detect grazing and induce an offensive response, or it is an internal indicator for oxidative stress that induces a protective defensive mode (Harke & Gobler, 2015; Liu et al., 2022; Zhou et al., 2021); there are likely important functions microcystin performs that have not yet been discovered, solely on the presumption that exporting microcystin during low nitrogen availability seems counterintuitive because a significant amount of nitrogen leaves the cell at a time when the nutrient is in high demand (Harke & Gobler, 2015).

Five genes from the lipoxygenase gene family (*lpx*ABCDK) were annotated and identified (**Figure 3.7, 3.8**) as genes responsible for the biosynthesis of the lipid A endotoxin (Moosavian et al., 2020). The lipoxygenase biosynthetic pathway (see **Figure 3.2** for entire cistern) is strongly understudied as less than 0.5% of sequenced bacteria genomes found *lpx* genes (as of 2015) and of those found only a handful have been looked at in further detail (Horn et al., 2015; Kurakin et al., 2020). Approximately 13,000 genomes have been searched for lipoxygenase genes, and only about 60 similar sequences have been identified, generally from the Firmicutes, Proteobacteria, as well as cyanobacteria species (Horn et al., 2015; Kurakin et al., 2020). Of the genomes that lipoxygenase was identified in, there were no strongly conserved nucleotide sequences for any of the *lpx* genes within the biosynthetic cluster, likely leading to false positive/negative hits when comparing contig sequences to reference databases (Horn et al., 2015). Furthermore, lipoxygenase genes that have been identified still lack complete characterization regarding the functional diversity of the genes (Horn et al., 2015). The low percentage of identification rates of lipoxygenase in bacterial genomes suggest it is not expected to be a vital enzyme for organism physiology, and instead provides supplementary purposes (Horn et al., 2015). One known function for lipoxygenase is its detoxification of polyunsaturated fatty acids (PUFA) into lipoxins (anti-inflammatory and immunosuppressor molecules), as protection to the organism against harmful compounds

and/or bacteria (Chandrasekharan & Sharma-Walia, 2015; Horn et al., 2015; Kurakin et al., 2020). Other hypothesized functions for lipoxygenase activity include cell-cell signaling and oxygen sensors, though there is no current experimental data available to validate the theories (Kurakin et al., 2020).

The lipoxygenase genes targeted in this study were observed in each sampling period from all microbial communities in TLW (**Figure 3.7**). The higher abundance of *lpx* genes in **Fig 3.7** than **Fig 3.8** is attributed to the bacterial groups in the water column containing lipoxygenase gene clusters (aside from cyanobacteria) among which include species from Proteobacteria and Verrucomicrobiota. Being that algae is considered an ancestral producer of PUFAs, the observations of fatty acid metabolizing enzymes (e.g., lipoxygenase) is probable and well supported by its observations in algal and cyanobacterial genomes, as well as in bacteria that likely evolved from ancestral algal species (Horn et al., 2015). Lipoxygenase studies within cyanobacteria found the *lpx* gene clusters were primarily found in Oscillatoriales, followed by Nostocales, Synechococcales, and Chroococcales, and its prevalence in multicellular organisms suggest potential roles in helping form multicellular structures within the cells (Horn et al., 2015; Kurakin et al., 2020). From the cyanobacterial communities in TLW, the *lpx* genes were exclusively observed and assigned to Synechococcales, unsurprising as that was the most prevalent cyanobacterial order in the water column (Cameron et al., 2022b; Cameron, 2021). Especially apparent within the cyanobacterial communities (**Fig 3.8**) are the similar observations between the gene abundances for *lpx*, *mcy*H, and the various nitrogen metabolism genes at the same sampling points (July 2018, June 2019, August 2019). There have been some studies linking the incidence of lipoxygenase in nitrogen fixing bacteria, suggesting preliminary theories regarding lipoxygenase's association with nitrogen metabolism activities (Kurakin et al., 2020). The findings from this study corroborate potential links to toxin-related biosynthetic gene clusters and nitrogen metabolism genes, for which more research in necessary to determine concrete associations.

3.5.2.2 Functional Nitrogen Metabolism Genes

The ability of diazotrophic organisms to utilize and convert atmospheric nitrogen into available nitrogen for non-diazotrophs established them as crucial proponents to advancing nitrogen storages in the environment, and subsequently advancing the success of heterotrophic organisms (Nash et al., 2018). A field of study that continues to a strong focus in cyanobacteria and other diazotrophic organisms is understanding the basis of how nitrogen fixation works on a genetic level, and what environmental conditions are prompting these responses. Most research attempts to understand the mechanism of nitrogen acquisition by observing various cyanobacteria under nitrogen depleted conditions such as Harke and Gobler (2015) whose research observed gene expression levels exposed to different nitrogen environments. Observing genes related to nitrogen fixation pathways, heterocyst differentiation, and nitrogen sensors and transcriptional activators can be a great tool in understanding why and how nitrogen is utilized by various microorganisms (Affe et al., 2018; Harke & Gobler, 2015; Sadvakasova et al., 2022). Three biosynthetic gene clusters (cyanase, nitrogenase, nitrogen regulator) were identified and annotated as being involved in nitrogen fixation within cyanobacteria species in this study; of the three gene families, six genes (*cyn*S, *nif*JKSU, *ntc*A) were extracted and looked at further. Although related, the three gene clusters observed in this study are unique in their functionality and when they are expressed by the organism.

Cyanate, an organic toxic compound used by many microbes, is a waste by-product that is introduced to the environment via anthropogenic pollution and can be detoxified by cyanase (*cyn*S) which converts it into carbon dioxide (CO_2) and ammonium (NH_3) (Kebeish & Al-Zoubi, 2017; Mao et al., 2022; Palatinszky et al., 2015; Sáez et al., 2019). Of the three nitrogen fixation gene families, cyanase was least observed in the metagenomic data, only appearing in July 2018 and June 2019 within cyanobacterial communities, and within entire bacterial communities the gene was observed in all but two sampling points (albeit at low abundances). Although first fully characterized in *Escherichia coli*, cyanase is a conserved gene (see **Figure 3.3** for biosynthetic gene cluster) that has now been characterized in several

organisms including cyanobacteria (e.g., *Synechococcus elongatus*), as well as Proteobacteria and some fungi (Kamennaya & Post, 2011; Kebeish & Al-Zoubi, 2017; Mao et al., 2022). The aim of cyanase in all organisms is to detoxify cyanate for organic nitrogen sources, yet the exact mechanism and outcomes varies by what the species requires (Kebeish & Al-Zoubi, 2017); specifically, whether the CO_2 is utilized (i.e., in photosynthetic cyanobacteria), or if ammonium is the only product reintroduced in assimilation pathways (Kebeish & Al-Zoubi, 2017). In a study observing how nitrogen influences transcription rates of N acquisition, assimilation, and transport genes, *cyn*S expression was detected to have a small increase in transcript abundance on days two and three of nitrogen deprivation and was not detected at any other points (neither increasing nor decreasing), even after nitrogen levels were restored (Harke & Gobler, 2015). The reasons are unknown regarding its low transcription rates; however, it can be assumed that if no cyanate is present in the environment, then there would be no need to express the enzyme (Harke & Gobler, 2015). Being that cyanate is introduced via anthropogenic waste, as urbanization and anthropogenic disturbances increase, a rise in cyanate concentrations would not be surprising, and thereby potentially cause a prominent rise in cyanase activity in response. Though cyanate has been identified as a key nitrogen source for microbial communities, and the enzyme responsible for its assimilation (*cyn*S) has been identified in numerous species, its diversity, expression, and prevalence are still understudied and has mainly been focused on marine communities (Mao et al., 2022). This study establishes cyanase observations in cyanobacterial and microbial communities in a freshwater oligotrophic watershed, scarce of anthropogenic interactions and disturbances.

The main enzyme, nitrogenase, encoded by the *nif* biosynthetic gene cluster (see **Figure 3.4**), is a highly efficient enzyme superfamily responsible for the bulk of nitrogen fixation among diazotrophs (Detoni et al., 2022; Esteves-Ferreira et al., 2017; McRose et al., 2017; Thiel, 2019; Watanabe & Horiike, 2021). The majority of diazotrophs utilize Mo-nitrogenase (molybdenum-dependent), however, the alternative Fe- (iron) and V- (vanadium) dependent nitrogenases are sometimes employed if Mo is limited, and if the diazotroph can express the genes specific to these pathways (McRose et al., 2017). Although the amount of

nitrogen biologically available in the environment is the key instigator for nitrogen metabolism gene transcription and translation, the type of nitrogenase that is activated is dependent on the availability of molybdenum (or lack thereof) as it comprises the key proteins in the *nif* gene cluster (Nash et al., 2018). Most research regarding nitrogen fixation refers to Mo-nitrogenase because its gene cluster (*nif*HDK) is encoded by all diazotrophs, whereas the diversity and distribution of the alternative nitrogenase enzymes is widely unknown as they are more organism and environmentally specific (Esteves-Ferreira et al., 2017; McRose et al., 2017). Regardless of which nitrogenase pathway is employed, all nitrogen fixation rates are dependent on the availability of Mo, Fe, and V in the environment, as the individual encoding genes require these trace metals to function (Renaudin et al., 2022); furthermore, as the concentrations of these metal cofactors have been linked to human interference via industrial pollution and atmospheric alterations, nitrogen fixation rates and cyanobacterial community growth have been increasing since the start of the Industrial Revolution and is expected to continue rising as diazotrophs have more access to key nutrients (Renaudin et al., 2022).

The nitrogenase encoding genes observed in the data (*nif*JKSU) are each responsible for serving an individual purpose in expressing nitrogen fixation enzymes (see **Table 3.3** for detailed functions). As described in **Figures 3.7/3.8,** at least one *nif* gene was observed at all sampling points within all microbial and cyanobacterial communities (excluding May 2019 as no cyanobacterial genes of interest were detected). The observations of *nif* genes within bacterial and cyanobacterial communities was unsurprising as nitrogenase is a crucial enzyme for nitrogen fixation (Detoni et al., 2022), and *nif* gene expression has been detected in as little as 18-24 hours after cyanobacteria were exposed to nitrogen depleted environments (Sadvakasova et al., 2022). In similar studies observing transcript levels of nitrogenase genes in nitrogen deprived conditions, *nif*J (required for nitrogen fixation when iron is limited) (Thiel, 2019), had increased transcription levels throughout the entire duration of the experimental nitrogen-depleted conditions (Harke & Gobler, 2015).

Of particular interest for the *nif* gene cluster was the observations of genes under ice cover in January 2020 (*nif*JSU), of which *nif*J represented a cyanobacterial contig. Unfortunately, the cyanobacterial contig sequence for *nif*J in January 2020 did not result in significant reference sequences from the NCBI Blast tool so its taxonomic assignment is unknown, however the major orders observed in TLW (as well as across Canada) have been known and reported to occur under ice cover (e.g., Nostocales, Chroococcales, and Oscillatoriales primarily) (Reinl et al., 2023). Some cyanobacteria species have adapted to colder water temperatures by accumulating PUFAs to develop more fluid membranes and evolving a host of anti-freeze and cold shock enzymes (Reinl et al., 2023). The presence of cyanobacterial blooms under ice cover is hypothesized to occur in one of three manners: (1) blooms that were initiated in cold water conditions and persisted, (2) blooms that occurred at a different thermocline (i.e., lower down the water column where the water temperature is warmer) and were brought to the surface, and (3) blooms that were initiated in warmer temperatures but persisted as the water temperatures dropped (Reinl et al., 2023). Although the observation of a nitrogenase gene under ice cover does not signify a bloom, it does, along with advancing research, support the claims of cyanobacterial blooms being active in colder water conditions, and demonstrates the demand for future studies of cyanobacteria outside of the typical warmer seasons.

Cyanobacterial communities under ice in this sample were observed to comprise less than 1% of the microbial communities analyzed by Cameron (2021), suggesting there could be a substantial lack in references sequences from ice covered environments either due to reduced detection or reduced sampling efforts (i.e., more dangers during winter sampling) (Powers & Hampton, 2016). Although studies have shown that some cyanobacterial species have a tolerance for freezing temperatures, sometimes even exhibiting toxic winter blooms in frozen lakes (e.g., *Planktothrix* sp.) (Reinl et al., 2023; Wejnerowski et al., 2018), research in this field is still underrepresented because of the assumptions that microbial communities become dormant under ice cover, not warranting the risks associated with winter sampling (Powers & Hampton, 2016). Evading winter and ice-covered sampling efforts is resulting in

substantial knowledge gaps regarding the forces that drive under-ice microbial developments, such as nitrification-associated anoxic events caused by the oxygen-consumptive nitrogen fixation process (Powers & Hampton, 2016).

As mentioned, there are many different pathways for nitrogen availability to affect bacterial and growth. Looking at cell differentiation, a recent study found that when heterocystous cyanobacteria are cultured for 24 hours in a nitrogen absent media, at least 5-10% of the filamentous cells will differentiate into heterocyst's to supply the organism with nitrogen for survival (Sadvakasova et al., 2022). Ultimately, the bulk of global nitrogen control in cyanobacteria is managed by one crucial gene – *ntc*A, a nitrogen assimilation transcriptional regulator (Esteves-Ferreira et al., 2017; Flores et al., 2019; Giner-Lamia et al., 2017; Herrero et al., 2004; Mo et al., 2017). While nitrogen acquisition and assimilation are regulated by transcriptional activation of *ntc*A, *ntc*A itself and the abundance of nitrogen metabolism genes in general is largely driven by the availability of ammonium in the environment and hence ruled by the genes encoding ammonium and nitrate/nitrite transporters (Affe et al., 2018; Esteves-Ferreira et al., 2017; Flores et al., 2019; Giner-Lamia et al., 2017). In this study, *ntc*A was observed in July 2018, June 2019, and August 2019 – and was only observed in cyanobacterial species. Apart from a *nif*U observation in July 2019 and *nif*J in January 2020 (**Fig 3.8**), the bulk of nitrogen fixation genes tend to be observed in unison with *ntc*A; although the genes observed do not represent gene transcription and expression, it can be speculated that if *ntc*A is active, then it may be controlling the upregulation of nitrogenase and other nitrogen assimilation genes, depending on environmental conditions and the amount and types of nitrogen sources available. Under experimental conditions, *ntc*A transcription was observed to rapidly increase when nitrogen levels decreased (Giner-Lamia et al., 2017; Harke & Gobler, 2015) and remained elevated until nitrogen levels were restored and transcription downregulated (Harke & Gobler, 2015). While it cannot be determined without nitrogen nutrient levels for each sampling point, it is possible the samples in which *ntc*A and nitrogenase genes were observed may be associated with a lack in nitrogen, or ammonium, specifically in Big Turkey Lake.

In cyanobacteria, *ntc*A is highly conserved, and aside from regulating expression of nitrogen assimilation gene pathways, *ntc*A is also involved in regulating secondary pathways, such as carbon and nitrogen balance in cells, stress responses, photosynthesis, as well as DNA and central metabolism (Esteves-Ferreira et al., 2017; Flores et al., 2019; Liu et al., 2022). It is also observed in non-diazotrophic cyanobacteria where it functions to activate the uptake of various nitrogen sources (e.g., urea, nitrate, ammonium) rather than activating nitrogen-fixation pathways (Esteves-Ferreira et al., 2017). Several putative binding sites for *ntc*A have been discovered in cyanobacterial DNA and it is predicted that this gene is involved with regulating at least 2000 individual genes, not just ones related to nitrogen assimilation (Flores et al., 2019), of which, at least some of the binding boxes are present on *mcy* operons for microcystin synthesis (Liu et al., 2022; Scherer et al., 2017). The *mcy* operon is polycistronically transcribed and *ntc*A is suggested to be a bidirectional regulator, either inhibiting or initiating the biosynthetic gene cluster (Liu et al., 2022; Scherer et al., 2017), however, the exact mechanisms and links between nitrogen metabolism and microcystin production are still unknown (Liu et al., 2022). The trends observed in **Figure 3.8**, regarding how most target genes are observed when *ntc*A is also observed, excluding the three outliers (*mcy*A, *nif*U in July 2019, and *nif*J in January 2020), is suggestive that *ntc*A could potentially be involved in more regulative processes than originally believed. In addition, these trends are not observed in **Figure 3.7** which further supports *ntc*A activity is highly conserved in cyanobacterial species (Giner-Lamia et al., 2017) and may not function similarly or be present in other bacterial communities. While the trends are suggestive, far more research is required to better understand the relationship between *ntc*A, nitrogen metabolism, and biosynthetic gene families, especially about transcriptional regulation of these operons.

3.5.3 Phylogenetic Analysis of Target Genes

The rapid expansion of sequencing methods and applications within the last decade has greatly enhanced the accuracy of genomic-based phylogenetic analyses (Parks et al., 2018). Taxonomy is a necessary component to interpreting metagenomic data and describing

microbial diversity within samples (Parks et al., 2018; Strunecký et al., 2023). To ensure validity of the metagenomic gene annotations used in this study, each target functional gene underwent phylogenetic analysis to gain further insight to the taxonomic assignment and correctness of gene predictions. The outcome of the blast searches established that some of the annotated genes labelled by Prodigal in the eggNOG file did not have any valid NCBI reference sequences; while this could be the result of the gene contig length being too short to have a complete or existing blast response, the more likely reason is that many gene predictors tend to predict too many genes and end up producing many false positives (Hyatt et al., 2010). This is potentially one of the reasons that several target genes – namely *mcy*DEJ – did not have accurate taxonomic predictions by Prodigal (all were predicted generically as bacteria) and did not have any similar sequences in NCBI GenBank. Because of this, no phylogenetic analysis was performed for these three genes of interest. While this does not negate the observations of these genes (i.e., they are still considered acceptable observations in the genomic data), it does suggest that some sequences are still unknown and require more research and a wider range of sequences to compare to in genomic libraries.

Whole genome shotgun (WGS) sequencing relies primarily on sequence identification and corresponding functional annotations as it does not rely on specific primers to target genes of interest (Pérez-Cobas et al., 2020). While this has its benefits to discovering novel enzymes and genomes, there are limitations as with any other sequencing application. In this study, the metagenomic assembly pipeline employed taxonomy-independent annotations that does not rely on a reference genome for functional gene annotations, but instead contigs are binned by either nucleotide similarities or abundance-based categorizations (Pérez-Cobas et al., 2020). Nucleotide similarities assumes that oligonucleotide composition within same genomes is more similar than from those of different genomes, and abundance similarities assumes that contigs belonging to the same genome should be observed at similar rates within a sample compared to other samples (Pérez-Cobas et al., 2020). While this method does not have the limitations of depending on the availability of sequenced genomes, nor must wait the time-consuming process of aligning contigs against reference genomes, taxonomy-independent

annotations instead strongly rely on the length of reads/contigs (Pérez-Cobas et al., 2020); therefore, fragments of DNA that are not long enough to extract such information (i.e., nucleotide/abundance similarity) will result in low confidence annotations and separation at low taxonomic levels (Pérez-Cobas et al., 2020).

Many of the individual target genes observed in the phylogenetic gene trees clustered with other sample gene sequences (in moderate to highly supported clades) yet remained spatially separated from the reference sequences – namely the genes *cyn*S, *lpx*B, *mcy*A, *mcy*C, *mcy*H, *nif*J, *nif*K, *nif*S (**Figures 3.9, 3.15-3.20**, respectively). The high bootstrap confidence supports the validity of the gene sequences, yet the lack of reference sequences available in GenBank as well as an even shallower pool of ones that strongly align with the metagenomic sequences confirms there is a significant gap in sequencing specific to functional gene contigs. The *lpx* phylogenies, as well as *nif*U and *ntc*A, confirmed that of the gene sequences that aligned with similar reference sequences, all clades were taxonomically assigned to either *Synechococcus* or *Cyanobium* taxa. While this confirms the taxonomic assignment by Prodigal and the eggNOG annotation algorithm for these genes, it also suggests potential sequencing biases, and a shortage of diverse gene sequences accessible in GenBank.

Studies targeting functional genes have been employed by many researchers to better understand a particular conception, whether that pertain to an organism of interest, community structure, toxin potential, or water quality. One example, a study by Bukowska et al. (2017) targeted the microcystin biosynthetic gene *mcy*A as a marker to determine the abundance of potentially toxic cyanobacteria species in a large lake system in Poland (Bukowska et al., 2017). A finding from this study, one that is not unique to studies of the same breadth, observed that although the abundance of cyanobacteria in these lakes were high, and known toxin-producing species were present (determined by microscopy), the abundance of *mcy*A genes in the water system did not correlate to the proportion of cyanobacteria (Bukowska et al., 2017). These findings relate to the results in this study as the abundance of cyanobacterial related target genes does not correlate to the relative abundance of cyanobacteria phyla as determined

by Cameron (2021) (i.e., the abundance of target genes is low relative to the abundance of cyanobacteria in Big Turkey Lake in the samples). In Bukowska et al. (2017), *mcy*A was only observed in two cyanobacterial species – *Microcystis* and *Planktothrix* – suggesting that either the primers used in the study were not universal, or that the gene sequences for *mcy*A are not yet known for all taxa; while the issue of primers does not apply to this thesis because metagenomic sequencing was applied, the more likely reasoning for low gene abundance in both Bukowska et al. (2017) and this study is the gap in gene sequencing for all cyanobacterial taxa, as well as a limit in available GenBank sequences. While this does not invalidate the significance of this research or any research like it, it does suggest that the low diversity in annotated genes and gene abundance is likely to be an underestimation of the true abundance of types and number of genes.

Phylogenies focused on a single locus from genomes can sometimes lack statistical support (i.e., low sequence identities, bootstrap, sample size, etc.) and resolution because depicting an entire representation of evolutionary history is difficult and can be inaccurate from the viewpoint of a single locus (Cornet et al., 2021). Although in some cases, it can lead to a broader understanding of evolutionary history when multiple copies of the locus are discovered; such as the case in phylogenetic studies focusing on nitrogen fixation gene clusters that detected several nitrogenase gene copies present in bacterial genomes, suggesting they represent different types of nitrogenases (i.e., Mo, Fe, or V-dependent) and are a result of HGT during cyanobacterial evolution (Esteves-Ferreira et al., 2017). Considering the significant reliance on nitrogen fixation cyanobacterial species had, having evolved before oxygen levels were sufficiently available, it is understandable that a selective pressure on nitrogenases would have occurred to utilize all available nitrogen storages, leading to the specification of at least five nitrogenases in cyanobacteria sharing conserved regions (Esteves-Ferreira et al., 2017). In contrast, phylogenetic studies using genome wide datasets can also present complications as foreign DNA is more likely to be introduced via environmental contamination or due to novel HGT sequences (Cornet et al., 2021). Because cyanobacteria have frequently shown to undergo high incidences of horizontal gene transfers, especially among the diazotrophic taxa, foreign

DNA is a common occurrence (i.e., misaligned sequences in MSAs), often due to non-cyanobacterial sequences being incorrectly selected, most likely resulting from the complex trophic relationships cyanobacteria has with other bacterial taxa (Cornet et al., 2021).

3.5.4 Weather and Sampling Conditions

One finding observed throughout the seasonal variation of relative microbial and functional gene abundance is the stark difference between the two samples taken in July (2018 and 2019). Both samples were collected only a year apart, followed identical collection standards and protocols, and collected at similar times of the day (midday) – with the only exception being TLW43 (2018) was collected at one meter below Secchi depth (8 m) and TLW323 (2019) was collected at Secchi depth (5.25 m) (Cameron, 2021). Despite the similarities in collection methods and season, the relative abundance of cyanobacteria (**Figure 3.6**) and thereby of the target gene abundances (**Figure 3.8**) significantly varied from one another. Although the findings from Cameron (2021) highlighted the vertical migration of cyanobacteria in the water column throughout seasonal and diurnal timeframes, it is likely that the sampling depth difference was not the sole cause for the disparities (however it is important to note that Secchi depth varied by 1.75 m naturally, potentially suggesting fluctuation in water clarity). An additional influence on the relative abundances could be a local weather event during July 2019 that occurred prior to the sample collection (Cameron, 2021). Currently, there is still very little research available that investigates how storm events affect microbial community composition and functional gene abundances (Kan, 2018; Tang et al., 2022). At present, it is unclear how abundances in the water column will fluctuate (i.e., increase or decrease) following a storm event such as heavy rain or wind because variables like the type and strength of storm, and the hydrology and physical characteristics (e.g., soil, bedrock) of the watershed all play crucial roles in how a system will react and how long it will take to re-establish itself (Kan, 2018; Tang et al., 2022). In general, however, storm events do tend to cause a microbial influx to water streams by disturbing communities deep in the soil and sediment, causing them to shift into the water column (Kan, 2018; Tang et al., 2022). Considering it is not possible to predict where microbial communities will shift within the

water column and how these fluctuations may affect their functional gene biosynthesis, it is essential that water sampling programs implement delayed sampling times following weather events to allow source waters to stabilize and the communities within them to redistribute naturally back to the desired depths (Cameron, 2021).

3.6 Conclusions

The cyanobacterial communities at Turkey Lakes Watershed, dominated by members of Synechococcales (primarily picocyanobacteria), are indicative of cyanobacterial communities in forested watersheds across Canada (Aljoudi, 2021; Cameron, 2021; Favot et al., 2023; Reinl et al., 2021; Shardlow, 2021; Winter et al., 2011). Although knowing the community composition in a water system is very useful in hypothesizing toxin potential based on toxin-producing species, it is still very speculative and does not confirm the functional potential of species present (Romanis et al., 2021). Whole genome shotgun sequencing of the samples provided a deeper insight to the functional gene diversity and biochemical potential (**Figure 3.7, 3.8**) among the cyanobacterial and bacterial communities in Turkey Lakes Watershed over a six-month seasonal period. The findings illustrated in **Figure 3.8** highlight the coincidence that target genes are observed in all-or-nothing patterns when *ntc*A is observed. Whether these observations are interconnected requires further research to confirm, however, *ntc*A has been noted to transcriptionally regulate several secondary pathways in cyanobacteria, with putative binding sites that serve to potentially control thousands of individual genes (Esteves-Ferreira et al., 2017; Flores et al., 2019; Liu et al., 2022).

Although the research conducted in this study yielded significant findings, there were several factors that were limiting throughout the research. These limitations include the number of samples sequenced, the amount of available sequenced data in reference libraries, as well as the sampling depth eDNA was extracted from. The number of samples sequenced can be a limitation in that the more samples observed provides a wider depiction of the activities of cyanobacterial communities in the watershed as well as having a longer seasonal breadth.

Recently, cyanobacteria communities in oligotrophic lakes are trending more frequently towards blooms earlier in the spring and later into the autumn season (Cameron et al., 2022a; Cameron, 2021; O'Keeffe, 2019; Winter et al., 2011), both of which include months that are not available for observation in this study. The observation of *nif*J in the January sample suggests that there may be more activity in the winter months than initially hypothesized (Powers & Hampton, 2016), however, additional winter samples are necessary to truly understand the community compositions and activities under ice cover. The situation of *nif*J also not having any significant blast results leads into the second limitation in this study regarding the number of available sequences in genomic libraries, including sequences specific to gene contigs. Especially when researching phylogeny, the resolution of phylogenetic analyses is strongly dependent on the number of similar sequences available to align against target sequences (Cornet et al., 2021), and like the case of the *nif*J contig, the lack of sampling during winter months lowers the diversity of available sequenced data (Powers & Hampton, 2016). To combat some of these limitations, future researchers should analyse a wider range of samples throughout the year, and as more samples are sequences, reference libraries will continue to grow, in turn increasing the resolution of phylogenetic analyses.

An unexpected outcome of this research recognized the significance sampling depth can have on community composition and functional gene diversity. The highest abundance of cyanobacteria and genes are observed in July 2018, which is coincidentally the only sample to be collected from one metre below Secchi depth. And although Secchi depth itself varies by day and season, this research, in harmony with Cameron (2021), suggests that one metre below Secchi contained higher abundances of microbial and cyanobacterial communities during some sampling months. Alternatively, based on Cameron's findings of microbial communities from different depths on different days and months, the abundances from each depth can highly vary as cyanobacteria can regulate their buoyancy and stratify themselves within the water column (Barsanti & Gualtieri, 2014; Cameron, 2021). Unless constant and specific sampling times are incorporated in sampling efforts (i.e., when cyanobacteria are known to be at a particular depth, although this can vary by season, weather conditions, etc.), integrated samples may be a better

option for a comprehensive view of communities and gene diversity in the water column; depth integrative samples avoids the challenging biases of fluctuating Secchi depth (depending on water clarity), sampler bias (e.g., Secchi depth depends on visual cues of the Secchi disk and vision strength varies by person), as well as accounts for microbial communities from all depths and vertical variability within a single sample (Cameron, 2021). Regardless of sampling and bioinformatic limitations, the findings from this study reveal the gaps in knowledge of cyanobacterial activity within the water column and signifies the need for continued research using WGS sequencing applications.

3.7 Implications

Applying next-generation sequencing to water quality monitoring programs is an immensely valuable tool in observing all potential threats within a water system, including those of harmful algal blooms. Observing an isolated forested watershed, distant from anthropogenic disturbances with little to no history of natural disturbances – such as the Turkey Lakes Watershed – establishes a baseline for microbial community composition and functional gene diversity in simple oligotrophic systems to assess changes in diversity (Edge et al., 2020). Far more focus has been applied to eutrophic systems that are known to be home to wide ranges of microbial communities that can create HABs due to the high availability of nutrients, yet there is less attention on how these dynamics are observed when nutrients are scarce, despite documentation that cyanobacteria and other HABs prevalence are increasing in oligotrophic water bodies (Reinl et al., 2021). The research conducted in this study provides a basis for prospective microbial communities and their biochemical potential in undisturbed low nutrient systems and illuminates the effects climate change alone can have on the synergies beneath the surface. By understanding how functionally diverse cyanobacteria can be when they are exposed to very little nutrients and thereby have relatively low energy potential, this study aims to emphasize the importance on proactive monitoring for the provision of drinking and recreational watersheds because the potential for threats is always present and it is much easier

to prevent toxic blooms through efficient sampling programs as opposed to defensively reacting to them.

The observations of toxic biosynthetic gene clusters and nitrogen metabolism related genes within the cyanobacteria communities in the TLW signifies the universal and constant potential for cyanobacterial HABs to occur in any type of system regardless of nutrient input. An especially important implication discovered from this study, is the observations of the microbial communities and their functional genes several meters deep into the water column. Majority of water quality monitoring conducted in Canada only occurs at the surface level and only after a possible bloom is reported (Rashidi et al., 2021); however, this study highlights how doing so severely underestimates cyanobacteria prevalence since the presence and productivity of cyanobacteria in an oligotrophic water system was observed throughout the water column despite showing no observable blooms on the surface. The implications of observing toxin and nitrogen metabolism related functional genes in the Turkey Lakes Watershed in and around Secchi depth in the absence of a visual bloom insinuates that monitoring programs should not be overlooking frequent and deeper water sampling protocols because it is likely that cyanobacteria and other hypothetically harmful microbial species are present and with them, the potential for threatening blooms (Cameron, 2021). From a water system management viewpoint, the potential for cyanobacteria to be present can disrupt the health and safety of the water system which may directly implicate local communities utilizing these source waters for drinking and recreational provisions. Ecologically, the productivity of cyanobacteria in the water column at different depths and throughout the year can influence the structure of the water system by driving nutrient inputs of nitrogen through nitrogen metabolism biosynthesis, effectively providing energy sources to diazotrophic organisms which may result in increased proliferation and the initiation of secondary metabolite production (Detoni et al., 2022). Regardless of the use (or lack thereof) of source waters, the only way to better understand the effects climate change exacerbated disturbances have on water quality, biodiversity, and biochemical potential is to maintain strict and effective monitoring of such environments.

Lastly, the findings observed in both July samples taken a year apart illuminated the importance of following strict water sampling protocols when quantifying microbial and functional gene relative abundances. Whether the disparities should be attributed toward the sampling difference of one meter or due to the rain event in the latter sample, this research signifies the importance that little variations in the environment, natural or otherwise, can meaningfully influence eDNA sampling. Moving forward, water monitoring programs should be keen to maintain uniformity among their sampling protocols which includes taking samplings at similar times of the day, from similar sampling depths, and allowing a generous stabilization period before sampling source waters that have experienced any type of weather events (Cameron, 2021). While it is not feasible to eliminate all potential drivers that may externally influence sampling results, it can at least limit researcher bias and/or lower the disparities from natural impacts, especially as weather events and climate change exacerbated disturbances are predicted to increase (Tang et al., 2022). As well, having longer term data sets can alleviate the discrepancies that naturally occur in environmental monitoring programs, and provide researchers and water management industries with a better view of seasonal variation in microbial communities and functional gene abundances in response to the natural environment amid a changing climate.

Chapter 4
Research Summary, Implications, and Future Directions

4.1 Research Summary

Cyanobacterial harmful algal blooms are an emerging threat to freshwater systems globally (Reinl et al., 2021). These blooms often proliferate in water bodies rich in nitrogen and phosphorus nutrients and are highly influenced by anthropogenic and climate change exacerbated disturbances (Chorus et al., 2021; Paerl, 2018; Reinl et al., 2021; Wells et al., 2020). Notwithstanding advanced research and monitoring programs, cyanobacterial blooms still dominate many systems (both eutrophic and oligotrophic), though far less research has focused on cyanobacteria in oligotrophic systems, despite their continued persistence even when nutrients are low (Reinl et al., 2021). The skewed research focus has led to an underestimation of the gravity of oligotrophic cyanobacterial blooms, including knowledge gaps in bloom ecology (e.g., stratification, nutrient inputs, etc.) and ineffective monitoring and remediation programs, despite the awareness of cyanobacteria relative abundance rising in oligotrophic systems (Reinl et al., 2021; Winter et al., 2011). These blooms are increasing in frequency and severity due to climate change and human interference in the environment, and the threats of toxin release or water quality deterioration is of utmost concern, especially in water systems that are relied on by communities (Kelly et al., 2019; Stoddard et al., 2016). The research conducted in this thesis aimed to better understand the composition of cyanobacterial communities in oligotrophic forested watersheds across Canada (including some drinking/recreation sources), as well as observe the functional gene diversity and biochemical potential within the selected Turkey Lakes Watershed.

When disturbed, freshwater systems undergo disruptions to nutrient cycling within the water column, providing or limiting access of nutrients such as carbon, nitrogen, and phosphorus to microbial communities (Edge et al., 2020). These changes in nutrient availability, as well as stratification shifts of phytoplankton communities can lead to harmful

algal blooms that spread and contaminate water bodies, causing potential shutdowns of recreation waters (e.g., swimming, fishing), drinking waters, and even trigger boil water advisories, in addition to adversely impacting animal and plant health in the surrounding environment (Edge et al., 2020). The highly disastrous 2014 Lake Erie toxic blooms attracted global publicity of the dangers toxic cyanobacteria pose when residents of Toledo, Ohio awoke to a three-day water system shut off to prevent the use of toxin-contaminated water (Favot et al., 2023). On a more local scale, implementation of an algal bloom response protocol in 2006 by the province of Ontario increased awareness across Canada to the threats harmful algal and cyanobacterial blooms pose to drinking water reservoirs, recreational sources, and to water treatment facilities (Favot et al., 2023; Romanis et al., 2021; Winter et al., 2011). Although algal blooms are not new to the scientific community, increased awareness to their threats has resulted in better response protocols, and improved detailed reporting of visible blooms, correlating to a rise in research in the field (Favot et al., 2023; Winter et al., 2011). Home to the most lake area in the world (Government of Canada, 2013), the Government of Canada is continually motivated to protect these environments from contamination and disruption, especially in water bodies that communities interact with either for drinking or recreational purposes (Health Canada, 2022).

4.1.1 16S rRNA Gene Sequencing

Though incredibly useful and widespread in the scientific community, 16S rRNA gene amplicon sequencing has limitations like any other sequencing analysis, yet its conserved nature and resistance to horizontal gene transfer is unprecedented in identifying bacterial genera and species (Escapa et al., 2020; Winand et al., 2019; Zhang et al., 2018). Due to their vital functions in organisms, conserved operons are assumed to have evolved early on via vertical gene transfer, such as is the case for the conserved regions within the 16S rRNA taxonomic marker gene that make it ideal for characterizing microbial communities (Memon et al., 2013; Romanis et al., 2021; Zhang et al., 2018). The limitations observed in this study regarding 16S rRNA gene sequencing have less to do with the decision of choosing the 16S

rRNA gene to amplify, but rather being limited by only analysing the V4 subsection of the gene. While the V4 region of the 16S rRNA gene is widely applied in taxonomic identification studies because the conserved region is easily identified (yet variable enough to differentiate among cyanobacteria) (Zhang et al., 2018), being constrained to a smaller section of the gene immediately limits the length of sequences to compare among reference libraries; this factor may have resulted in lower taxonomic resolutions for the ASVs examined in this study, specifically regarding the number of unknown amplicon sequences observed in the maximum likelihood tree in Chapter 2. Based on the proximity of unknown ASVs near clades belonging to Synechococcales, it is likely that many of the unassigned sequences may be groups of picocyanobacteria, however, this cannot be confirmed without additional cyanobacterial sequencing and identification. The observations of several large clades with high bootstrap support that did not cluster with any valid reference sequences drives home the necessity for more diverse and continuous cyanobacterial sequencing, identification and classification, and publishing of accessible referenced sequences to genomic libraries.

Despite the unknown ASVs, the ML phylogenetic analysis depicted that the cyanobacteria taxa across Canada are polyphyletic and widespread across all five major ecozones, showing very little biogeographic clustering of species. The main cyanobacterial orders observed in this study – Synechococcales, Chroococcales, Nostocales, and Oscillatoriales – are coincidentally the dominating phytoplankton communities observed during blooms reported in oligotrophic lakes (Favot et al., 2023; Reinl et al., 2021; Winter et al., 2011). The toxic species primarily observed in this study – *Microcystis*, *Dolichospermum*, and *Aphanizomenon* – align with expectations as well as these are the most widely observed genera of toxic cyanobacteria in blooms reported from water bodies across Ontario (Favot et al., 2023; Winter et al., 2011). Biogeographically, the cyanobacterial taxa observed across Canada did not cluster specifically to any ecozone, but rather taxa were observed throughout all ecozone and biome regions. These findings support the theories of cyanobacterial dispersal occurring through airborne (aerial plankton), insect, or animal vectors, in addition to the standard dispersal throughout linked waterways and by anthropogenic introduction (Curren &

Leong, 2020; Wiśniewska et al., 2019; Wiśniewska et al., 2020). The diverse taxonomy and widespread biogeography of cyanobacteria distributed across Canada, as observed in this study, indicates the rigorous and efficient nature of cyanobacteria to adapt and thrive in most environments, and amid a changing climate.

4.1.2 Whole Genome Shotgun Metagenomic Sequencing

Whole genome shotgun sequencing is a vastly beneficial tool in environmental assessment applications as it provides an expansive output of information from environmental samples including but not limited to taxonomic assignments, biochemical potential, and functional gene diversity within given community populations (Alvarenga et al., 2017; Cheng et al., 2017; Pérez-Cobas et al., 2020; Sharma et al., 2020). The untargeted sequencing used in metagenomics leads to higher taxonomic resolution (i.e., all regions of the 16S rRNA gene are sequenced to determine taxonomy), however, it is still limited by the availability of identified and published referenced sequences and genomes to compare to, as well as the efficiency of bioinformatic pipelines to accurately annotate gene regions (Pérez-Cobas et al., 2020; Romanis et al., 2021). Untargeted sequencing is especially suitable for discovering novel enzymes and secondary metabolites by focusing on gene functionalities (Cheng et al., 2017). There are obstacles with this, however, because the functions of secondary metabolites are not crucial to survival and are therefore more likely to be lost during evolution or not present in all species, resulting in genes that are not conserved among organisms and still lack accurate functional annotations (Calteau et al., 2014; Chen et al., 2020; Memon et al., 2013). As of 2014, only about 20% of biosynthetic gene clusters with known product compounds had designated functional assignments, and the products of the remaining 80% of gene clusters were still unknown (Calteau et al., 2014). While these numbers have surely increased since, there are still a significant amount of unknown secondary metabolite products, and even less that have accurate annotated assignments (Calteau et al., 2014), prompting a need for further research and improvements upon annotation pipelines and reference libraries.

This study aimed to observe aspects of cyanobacteria and their biochemical potential in understudied oligotrophic systems using WGS sequencing; the findings of the study highlight the demand to better understand the factors influencing cyanobacterial toxin and nitrogen-metabolism gene expression and how their functional gene diversity may be supporting their adaptations to a changing climate. The outcomes of this thesis observed seasonal detections of biosynthetic gene clusters associated with the pathways for microcystin and lipoxygenase production, as well as gene families involved with nitrogen metabolism and acquisition within cyanobacterial communities. The observations of microcystin and lipoxygenase toxin genes in tandem with nitrogen metabolism genes at several sampling points support potential associations between biosynthesis of secondary metabolites and the activities of nitrogen assimilation (Kurakin et al., 2020). Although still largely conjectural, the observations of *mcy* and *lpx* toxin biosynthesis genes in concurrence with observations of the global nitrogen regulator (*ntc*A), is highly suggestive of a link between the pathways, supporting the current belief that *ntc*A bidirectionally regulates hundreds, if not thousands, of genes involved in mechanisms outside of nitrogen assimilation (Flores et al., 2019; Giner-Lamia et al., 2017). The use of metagenomics in this thesis yielded a substantial amount of data that provided meaningful insights to cyanobacterial communities, their secondary metabolites, and their seasonality, while also raising additional questions that further studies can continue to build upon.

4.2 Implications

4.2.1 Cyanobacteria are distributed everywhere and can be unpredictable

Across Canada, existing cyanobacteria documentation in source water blooms are often not reported nor published in peer-reviewed journals (Reinl et al., 2021). Although studies have shown that the dominant taxa observed in the Turkey Lakes Watershed are indicative of toxic species among temperate oligotrophic shield lakes in Ontario (as observed over the last 20 years), many of these were only determined through direct communication with water

resource managers from across the Great Lakes (Favot et al., 2023; Reinl et al., 2021; Winter et al., 2011). There is still a large gap in cyanobacteria identification across the nation, calling for a dynamic restructuring of reporting and publishing of bloom events (Reinl et al., 2021). As exemplified in Chapter 2 of this thesis, cyanobacteria are widely distributed across the country, across many distinct biomes, and do not tend to cluster based on biogeography or be affected by geographic restrictions (Curren & Leong, 2020; Sheath, 2006; Wiśniewska et al., 2020). The implications of these findings are that cyanobacterial species can thrive in many environments, and as such it is not possible to predict which species may be present in any source waters, which also encompasses the unpredictability of where toxin-producing species may prosper. This is especially vital information to water management and water treatability sectors because it means that oligotrophic or source waters with little history of cHABs should not be overlooked as not being threats because the potential for harmful cyanobacteria communities in a water system is always present and the only way to mitigate risks to drinking or recreational waters is through diligent mitigation and monitoring programs.

4.2.2 Cyanobacteria and functional genes pose potential threats in oligotrophic systems

Climate metrics affecting bloom formations in water systems are profoundly changing and are currently on a trajectory estimated to be beneficial for cyanobacterial proliferation (Favot et al., 2019). While the trajectory for remote oligotrophic systems (such as TLW) are less understood, an increase in frequency and severity of blooms is anticipated even in lieu of anthropogenic interference (Favot et al., 2019), as well as a shift in bloom seasons that extend later into autumn months (Favot et al., 2023; Winter et al., 2011). Although visible blooms are less observed and recorded in oligotrophic systems, the findings from Chapter 3 of this study exhibit the potential for cyanobacteria and harmful secondary metabolites to pose threats to the provision of clean and safe water sources even in the most isolated and undisturbed lakes. Being that many drinking water treatment facilities and recreational source waters utilize oligotrophic water bodies, continual monitoring of the entire water column is a fundamental measure that should be taken by these facilities to mitigate the potential for these risks. In the

case of drinking and recreational source waters, conducting frequent integrative samples of the water column over different seasons and using molecular analyses to observe the microbial communities and functional genes is an undervalued tool. Previously (and currently), these types of mitigation efforts are not employed by water management sectors because of the misconception that oligotrophic systems are at low risk for cyanobacterial blooms or toxin production, yet the community analysis of Big Turkey Lake showed by Cameron (2021) and the metagenomic screening for functional genes exhibited in this thesis all disprove the absence of threats in low nutrient systems and during ice covered seasons. To further validate the need for improved sampling, climate change exacerbated land disturbances, such as wildfires, are actively appearing more frequently and doing more damage (Favot et al., 2019; Tang et al., 2022), increasing the risks of nutrient runoff and cyanobacteria proliferation, and toxic winter blooms have been documented under frozen lakes (Reinl et al., 2023; Wejnerowski et al., 2018). Monitoring and mitigation efforts are the only proactive tools to screen for cyanobacteria and toxin gene potential in all trophic water systems and is much more advisable than defensively trying to control active blooms and prevent them from spreading and/or releasing toxic compounds.

4.2.3 Standard water sampling protocols to observe accurate and reliable trends

As demonstrated in Chapter 3, microbial community and functional gene abundances varied greatly in two samples taken at the same lake, month, and sampling time but during different years and may have ultimately been the result of sampling efforts. Without further long-term studies it cannot be ascertained if the variation was a natural fluctuation of microbial communities or due to the researcher sampling bias. The differences could very well be due to taking one sample at Secchi and the other sample one meter below Secchi, as well as the natural interference of weather events. It is possible that the heavy precipitation event prior to water collection may have redistributed the cyanobacterial communities deeper into the water column, potentially resulting in the underestimation of community and gene abundance in that sample (July 2019). The potential for these events to alter the stability of the water column and

affect cyanobacteria and microbial distribution is likely to under or overestimate population dynamics within a system (Kan, 2018; Tang et al., 2022), thus skewing seasonal trend research. Therefore, a standard for water collection protocols among science, environmental, and WTF researchers is crucial for accurate and reliable statistical analysis and observing long term trends, specifically regarding post mixing events (such as those caused by heavy precipitation or high winds). Especially among water management practices, water sampling should be taken at similar increments either throughout the day, week, month, or year (depending on accessibility) to establish what is considered "typical" for the specific water system, in addition to implementing delayed sampling protocols post weather events to allow the water column time to restabilize and for the microbial communities to naturally redistribute to their desired depths (Cameron, 2021). By doing so, observations of microbial communities and functional gene diversity and abundances will likely represent the most accurate depiction of seasonal variations throughout the water system.

Aside from the specificities of water sampling protocols, studies observing algal blooms should also be keen to maintain some standards for tracking algal activity as statistically speaking more populated areas will tend to have more bloom reporting's solely based on the higher chance of an observation compared to rural water bodies that are less likely to be observed. As an example, although there have been ongoing algal bloom studies across Ontario water sources for the last twenty years, the reporting's for blooms have only focused on visual surface blooms and rely on not only the blooms to be directly observed, but also recorded and available to the public (Favot et al., 2023; Winter et al., 2011). This poses many challenges as (1) the potential for cyanobacteria to inhabit lower depths is not at all considered (which in of itself disregards community composition throughout the water column) but also (2) can skew public perception of cyanobacteria and algal distribution because the data is contingent on self-reporting's and not actually depicting all water bodies within the study region (Favot et al., 2023; Winter et al., 2011). As the prevalence of cHABs have risen in the media, water researchers, WTF managers, and the public are now more aware of the threats blooms have on the provision of safe and clean water sources, resulting in increased vigilance

for bloom reporting's (Favot et al., 2023; Winter et al., 2011). What this has led to is a steadier influx of bloom reporting's in the recent years (mainly in urbanized settings and eutrophic systems), potentially skewing the data as to whether higher observations are the direct result of climate change impacting algal activity or the subsequent result of higher attentiveness to bloom events (Favot et al., 2023; Winter et al., 2011). By implementing standard water sampling protocols that are not dependent on visual blooms or self-reporting, the trends for cyanobacteria, water quality, microbial community composition, and functional genes could all be better compared across the nation, and there is less potential for misrepresentation for the threats of cyanobacterial activity.

4.2.4 Establish baselines for long-term source water monitoring amid a changing climate

Because oligotrophic systems are heavily understudied, beginning a basis framework of a water system analysis such as that delivered in this study can teach researchers and water management partners how their source waters and the diversity within them may currently look, and how they may be impacted by climate change moving forward. Long term studies are essential in understanding how systems fluctuate over time, and if there are concerns present that should continue to be monitored should conditions take a turn for the worse. More so, environmental protection assessments are particularly important in freshwater systems as they have higher likelihoods to degrade in response to adversities compared to other stabler ecosystems (Edge et al., 2020). The degree of freshwater systems across Canada are considerably at risk should unforeseen disasters occur, pushing for improved monitoring criteria to mitigate risks, such as programs like the Canadian Aquatic Biomonitoring Network (CABIN) and the Ecobiomics Project (est. 2016) (Edge et al., 2020; Environment and Climate Change Canada, 2019). Programs such as CABIN (developed by the Government of Canada) evaluate the conditions of various water bodies across the country and compare them to reference sites (i.e., water bodies that have not been disturbed by humans – such as the Turkey Lakes Watershed) (Environment and Climate Change Canada, 2019). Currently Canada is in the beginning stages of advancing the level of monitoring and remediation assessments to

improve the current standards for water quality testing (Edge et al., 2020; Environment and Climate Change Canada, 2019), and the findings from this research support the necessity and implications of applying metagenomics to improve water quality research. Conducting regular water sampling collections to observe microbial community and functional gene abundance, especially in oligotrophic source waters (as are the objectives of the Ecobiomics Project) can (1) to determine ideal standards for soil, water, and invertebrate collection and testing methods so that all Canadian agencies follow the same protocols, (2) establish a baseline for metagenomic observations so that future studies can observe changes to quality and biodiversity, and (3) fill in knowledge gaps and better characterize aquatic microbiomes using NGS (Edge et al., 2020; Environment and Climate Change Canada, 2019). Parallelly, the research conducted in this thesis and Cameron (2021) served as a baseline of metagenomic observations in the isolated, oligotrophic TLW which can be built upon moving forward to observe water quality and biodiversity fluctuations in response to a changing and warming climate with the hopes that it will encourage additional water quality managers and scientific researchers to do the same.

4.3 Future Research

4.3.1 Multiple loci should be examined for higher resolution phylogenies

The application of V4-16S rRNA region amplicon and WGS sequencing for determining water quality and the microorganisms or secondary metabolites potentially responsible for it are invaluable tools for source water management. However, aside from the technological aspect of sequencing, there are also aspects of research that can be improved on by the researchers themselves. Although 16S rRNA gene and metagenomic sequencing are widely used and highly accurate for bacterial identification, the time and monetary cost of sequencing limits most researchers into targeting single gene markers (such as the V4 region). This may not have been a significant issue previously, yet as more taxa and their functional genes are continually being discovered, reidentified, or added to reference databases,

phylogenetic analyses based on single marker genes (e.g., V4-16S rRNA gene, or targeted functional genes) are getting larger, and the resolution of the analyses is gradually lowering as a result (Cornet et al., 2021). This has led to a stronger focus on multilocus phylogenetic analyses that sequence and compare several conserved genes, strengthening the accuracy of phylogenetic analysis by having multiple genes that can distinguish between highly recombinant species (Dvořák et al., 2015; Strunecký et al., 2023). While this does not negate the value of 16S rRNA gene amplicon sequencing (because it does not undergo HGT) (Tsukuda et al., 2017; Yang et al., 2016), it does suggest that studies targeting genes of interest, such as the phylogenetic analyses of functional genes from Chapter 3, may benefit from combined sequencing efforts for several loci and quantitative molecular methods, if applicable, to increase the resolution.

4.3.2 Specialized cyanobacterial biological mechanisms should be considered

Regarding bloom and cyanobacterial proliferation research in source waters, consideration has typically focused on the importance and management of nutrient levels (i.e., monitoring nitrogen and phosphorus levels), and far less consideration has been on the cyanobacterial traits that utilize such nutrients (Reinl et al., 2021). Researched motivated by the specialized cyanobacteria features, such as the mechanisms responsible for buoyancy and dormancy (Barsanti & Gualtieri, 2014), akinete and heterocyst production (Esteves-Ferreira et al., 2017), specialized light adaptations (Beecraft, 2018), nutrient and carbon uptake/concentration means, as well as how hydrodynamic properties interplay with biological features that all give rise to harmful blooms is far more limited in literature (Reinl et al., 2021). Understanding these features may be vital to better understanding cyanobacterial behaviour within the water column, which can subsequently improve monitoring and sampling efforts (e.g., knowing which depth to sample according to buoyancy or dormancy states) (Cameron et al., 2022b; Reinl et al., 2021). Understanding more about cyanobacteria behaviour may explain which physical, chemical, and biological mechanisms drive their growth and maintenance across trophic states, (i.e., how they can dominate in both eutrophic and oligotrophic systems

where nutrients levels vary greatly), and may also provide some explanation as to how they are adapting so rapidly with climate change and manage to thrive even in heavily monitored water systems such as water treatment plants (Brown et al., 2021; Favot et al., 2019; Reinl et al., 2021; Romanis et al., 2021). Cyanobacterial research must shift its focus from the typical nutrient paradigm within eutrophic systems to the biological and hydrological mechanisms that exist in all systems, which will provide a holistic illustration of what environmental features give rise to cyanobacteria proliferation, and what biological features drive the onset of harmful blooms.

4.3.3 Additional sequencing methods should be employed to observe gene expression

Using metagenomics to observe the abundance of functional genes in cyanobacterial communities (and all microbial communities) in TLW demonstrated the depth to which researchers can observe biosynthetic gene clusters and therefore the biochemical potential for toxins and other troublesome secondary metabolites in water systems. A major limitation, however (aside from the need for more reference genes in libraries and improved annotation pipelines as discussed in Chapter 3), is that metagenomics provides a glimpse of observed genes within microbial communities but cannot clarify if the genes are being expressed at the time. Other molecular methods like metatranscriptomics and metaproteomics can used in addition or alternatively to metagenomics to determine if the biosynthetic genes in question are being expressed, and if the protein produced by the clusters are observed as well (Harke & Gobler, 2015; Romanis et al., 2021). A limitation encountered in Chapter 3 demonstrated that although genes of interest were observed at unexpected sampling points, (e.g., nitrogen metabolism genes under ice cover), it was not possible to determine if the enzymes encoded by the genes were being synthesized (e.g., it could not be determined if nitrogenase was active under ice cover) and thus requires further research to clarify. Understanding if genes are expressed versus observing them can be especially important when applying this type of research to drinking or recreational water sources that are actively in use by the public because the observations of genes does not directly imply a present danger (Romanis et al., 2021).

In the case of water treatment facilities (WTF), the difference between expressed genes producing and releasing a toxin into the water system versus cyanobacteria containing biosynthetic gene clusters specific for toxin production but are not actively being expressed can be very costly in the cases when facilities are required to shut down assembly; thus, the more specific molecular methods are in identifying active threats as opposed to potential ones is far more effective and efficient (Romanis et al., 2021). Rather than wait for visible blooms, or test for a specific toxin or secondary metabolite at a time, metagenomic, transcriptomic, and proteomic methods can identify which dominant taxa are present, define their biochemical and metabolic potential (i.e., toxicity potential), and link the ecophysiological and environmental factors that may be contributing/impacting bloom trends (Romanis et al., 2021). Although more costly, these molecular methods can provide more information about the entire water system and the potential and active threats within them all at once from one integrated sample collection.

References

Affe, H. M. d. J., Rigonato, J., Nunes, J. M. d. C., & Menezes, M. (2018). Metagenomic analysis of cyanobacteria in an oligotrophic tropical estuary, South Atlantic. *Frontiers in Microbiology, 9,* 1393. https://doi.org/10.3389/fmicb.2018.01393

Aljoudi, G. (2021). *Bacterial communities in the Glenmore Reservoir with an emphasis on the Cyanobacteria* http://hdl.handle.net/10012/17315

Alvarenga, D. O., Fiore, M. F., & Varani, A. M. (2017). A metagenomic approach to cyanobacterial genomics. *Frontiers in Microbiology, 8,* 809. https://doi.org/10.3389/fmicb.2017.00809

Arnell, N. W., Halliday, S. J., Battarbee, R. W., Skeffington, R. A., & Wade, A. J. (2015). The implications of climate change for the water environment in England. *Progress in Physical Geography: Earth and Environment, 39*(1), 93-120. https://doi.org/10.1177/0309133314560369

Barsanti, L., & Gualtieri, P. (2014). *Algae: Anatomy, biochemistry, and biotechnology* (2nd ed.). Taylor and Francis Group.

Beecraft, L. (2018). *Application and assessment of multi-wavelength measurements of variable chlorophyll a fluorescence for analysis of irradiance stress, physiological state and community composition in freshwater phytoplankton* http://hdl.handle.net/10012/13172

Blackburn, E. A. J., Dickson-Anderson, S. E., Anderson, W. B., & Emelko, M. B. (2023). Biological Filtration is Resilient to Wildfire Ash-Associated Organic Carbon Threats to Drinking Water Treatment. *ACS ES&T Water, 3*(3), 639–649. https://doi.org/10.1021/acsestwater.2c00209

Bláha, L., Babica, P., & Maršálek, B. (2009). Toxins produced in cyanobacterial water blooms - toxicity and risks. *Interdisciplinary Toxicology, 2*(2), 36-41. https://doi.org/10.2478/v10102-009-0006-2

Bolhuis, H., Severin, I., Confurius - Guns, V., Wollenzien, U. I. A., & Stal, L. J. (2010). Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. *The ISME Journal, 4*(1), 121-130. https://doi.org/10.1038/ismej.2009.99

Bolyen, E., Rideout, J. R., Dillon, M. R., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Asnicar, F., Bai, Y., Bisanz, J. E., Brejnrod, A., Brislawn, C. J., Brown, T. C., Caraballo-Roíguez, A. M., Chase, J., Cope, E. K., da Silva, R., Diener, C.,

Dorrestein, P. C., . . . Knight, R. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology, 37*(8), 852-857. https://doi.org/10.1038/s41587-019-0209-9

Bouaïcha, N., Miles, C., Beach, D., Labidi, Z., Djabri, A., Benayache, N., & Nguyen-Quang, T. (2019). Structural diversity, characterization and toxicology of microcystins. *Toxins, 11*(12), 714. https://doi.org/10.3390/toxins11120714

Brown, J., Hawkes, K., Calvaruso, R., Reyes-Prieto, A., Lawrence, J., & Palenik, B. (2021). Seasonality and distribution of cyanobacteria and microcystin toxin genes in an oligotrophic lake of Atlantic Canada. *Journal of Phycology, 57*(6), 1768-1776. https://doi.org/10.1111/jpy.13210

Bukowska, A., Kaliński, T., Koper, M., Kostrzewska-Szlakowska, I., Kwiatowski, J., Mazur-Marzec, H., & Jasser, I. (2017). Predicting blooms of toxic cyanobacteria in eutrophic lakes with diverse cyanobacterial communities. *Scientific Reports, 7*(1), 8342-12. https://doi.org/10.1038/s41598-017-08701-8

Buttle, J. M., Webster, K. L., Hazlett, P. W., & Jeffries, D. S. (2019). Quickflow response to forest harvesting and recovery in a northern hardwood forest landscape. *Hydrological Processes, 33*(1), 47-65. https://doi.org/10.1002/hyp.13310

Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., Kerfeld, C. A., Sivonen, K., Piel, J., & Gugger, M. (2014). Phylum-wide comparative genomics unravel the diversity of secondary metabolism in cyanobacteria. *BMC Genomics, 15*(1), 977. https://doi.org/10.1186/1471-2164-15-977

Cameron, E. S. (2021). *Spatiotemporal shifts in cyanobacterial communities in a northern temperate watershed – applications of next-generation sequencing and implications for monitoring and climate change adaptation* http://hdl.handle.net/10012/17185

Cameron, E., Emelko, M., & Müller, K. (2022b). *Characterization of cyanobacterial communities in lakes requires consideration of diurnal and spatial variation.* (). Cold Spring Harbor: Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/2022.01.27.478008 Retrieved from Publicly Available Content Database https://search.proquest.com/docview/2623505861

Cameron, E., Muller, K., Stone, M., Buttle, J., Leach, J., Webster, K., & Emelko, M. (2022a). *Early seasonal increases and persistence in relative abundance of potentially toxic cyanobacteria: Concerning impacts of extended ice-free periods in northern temperate lakes.* (). Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

https://doi.org/10.1101/2022.12.20.521158 Retrieved from Publicly Available Content Database https://search.proquest.com/docview/2756496001

Capital Regional District. (n.d.). *Greater Victoria Water Supply Area.* https://www.crd.bc.ca/service/drinking-water/watershed-protection/greater-victoria-water-supply-area

Chandran, A., & Mazumder, A. (2015). Investigation on the temporal variation and source tracking of faecal bacteria in a forest dominated watershed (Comox Lake), British Columbia, Canada. *Journal of Applied Microbiology, 119*(6), 1718-1728. https://doi.org/10.1111/jam.12969

Chandrasekharan, J. A., & Sharma-Walia, N. (2015). Lipoxins: Nature's way to resolve inflammation. *Journal of Inflammation Research, 2015*(default), 181-192. https://doaj.org/article/fe128f70ad644fa3b6879d1909baae30

Chen, R., Wong, H. L., Kindler, G. S., MacLeod, F. I., Benaud, N., Ferrari, B. C., & Burns, B. P. (2020). Discovery of an abundance of biosynthetic gene clusters in shark bay microbial mats. *Frontiers in Microbiology, 11*, 1950. https://doi.org/10.3389/fmicb.2020.01950

Cheng, J., Romantsov, T., Engel, K., Doxey, A. C., Rose, D. R., Neufeld, J. D., & Charles, T. C. (2017). Functional metagenomics reveals novel β-galactosidases not predictable from gene sequences. *PLoS ONE, 12*(3), e0172545. https://doi.org/10.1371/journal.pone.0172545

Cheng, Y., Geng, R., Shan, L., Liu, Y., Zhang, H., Xiao, P., Ma, Z., & Li, R. (2022). Taxonomic discussion on cyanobacterial systematics at family level, with special regards to *Phormidiaceae* by using the strains of Chinese newly recorded genera *Ancylothrix* and *Potamolinea*. *Diversity (Basel), 14*(4), 301. https://doi.org/10.3390/d14040301

Chorus, I., Fastner, J., & Welker, M. (2021). Cyanobacteria and cyanotoxins in a changing environment: Concepts, controversies, challenges. *Water (Basel), 13*(18), 2463. https://doi.org/10.3390/w13182463

Chorus, I., & Welker, M. (2021). *Toxic cyanobacteria in water : A guide to their public health consequences, monitoring and management (edition 2)* (2nd ed.). CRC Press. https://doi.org/10.1201/9781003081449

City of Calgary. (n.d.). *Glenmore reservoir.* https://www.calgary.ca/csps/recreation/sailing-school/glenmore-reservoir.html

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews, 17*(4), 840-862. https://doi.org/10.1128/CMR.17.4.840–862.2004

Coops, N. C., Hermosilla, T., Wulder, M. A., White, J. C., & Bolton, D. K. (2018). A thirty year, fine-scale, characterization of area burned in Canadian forests shows evidence of regionally increasing trends in the last decade. *Plos One, 13*(5), e0197218. https://doi.org/10.1371/journal.pone.0197218

Cordeiro, R., Luz, R., Vasconcelos, V., Goncalves, V., & Fonseca, A. (2020). Cyanobacteria phylogenetic studies reveal evidence for polyphyletic genera from thermal and freshwater habitats. *Diversity (Basel), 12*(8), 298. https://doi.org/10.3390/d12080298

Cornet, L., Magain, N., Baurain, D., & Lutzoni, F. (2021). Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: A case study with cyanobacteria and cyanolichens. *Molecular Phylogenetics and Evolution, 162*, 107100.

Curren, E., & Leong, S. C. Y. (2020). Natural and anthropogenic dispersal of cyanobacteria: A review. *Hydrobiologia, 847*(13), 2801-2822. https://doi.org/10.1007/s10750-020-04286-y

Davis, T. W., Berry, D. L., Boyer, G. L., & Gobler, C. J. (2009). The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms. *Harmful Algae, 8*(5), 715-725. https://doi.org/https://doi.org/10.1016/j.hal.2009.02.004

Detoni, A. M. S., Subramaniam, A., Haley, S. T., Dyhrman, S. T., & Calil, P. H. R. (2022). Cyanobacterial diazotroph distributions in the western South Atlantic. *Frontiers in Marine Science, 9*https://doi.org/10.3389/fmars.2022.856643

Devito, K. J., Mendoza, C., Petrone, R. M., Kettridge, N., & Waddington, J. M. (2016). Utikuma Region Study Area (URSA) – part 1: Hydrogeological and ecohydrological studies (HEAD). *The Forestry Chronicle, 92*(01), 57-61. https://doi.org/10.5558/tfc2016-017

Du, X., Liu, H., Yuan, L., Wang, Y., Ma, Y., Wang, R., Chen, X., Losiewicz, M. D., Guo, H., & Zhang, H. (2019). The diversity of cyanobacterial toxins on structural characterization, distribution and identification: A systematic review. *Toxins, 11*(9), 530. https://doi.org/10.3390/toxins11090530

Dunnington, D. W., Spooner, I. S., Krkošek, W. H., Gagnon, G. A., Cornett, R. J., Kurek, J., White, C. E., Misiuk, B., & Tymstra, D. (2018). Anthropogenic activity in the Halifax region, Nova Scotia, Canada, as recorded by bulk geochemistry of lake sediments. *Lake and Reservoir Management, 34*(4), 334-348. https://doi.org/10.1080/10402381.2018.1461715

Dvořák, P., Poulíčková, A., Hašler, P., Belli, M., Casamatta, D. A., & Papini, A. (2015). Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodiversity and Conservation, 24*(4), 739-757. https://doi.org/10.1007/s10531-015-0888-6

Edge, T. A., Baird, D. J., Bilodeau, G., Gagné, N., Greer, C., Konkin, D., Newton, G., Séguin, A., Beaudette, L., Bilkhu, S., Bush, A., Chen, W., Comte, J., Condie, J., Crevecoeur, S., El-Kayssi, N., Emilson, E. J. S., Fancy, D., Kandalaft, I., . . . Macklin, J. (2020). The ecobiomics project: Advancing metagenomics assessment of soil health and freshwater quality in Canada. *The Science of the Total Environment, 710*, 135906. https://doi.org/10.1016/j.scitotenv.2019.135906

Emelko, M. B., Schmidt, P. J., & Borchardt, M. A. (2019). Confirming the need for virus disinfection in municipal subsurface drinking water supplies. *Water Research, 157*, 356–364. https://doi.org/10.1016/j.watres.2019.03.057

Emelko, M. B., Silins, U., Bladon, K. D., & Stone, M. (2011). Implications of land disturbance on drinking water treatability in a changing climate: Demonstrating the need for "source water supply and protection" strategies. *Water Research, 45*(2), 461-472. https://doi.org/10.1016/j.watres.2010.08.051

Emelko, M. B., Stone, M., Silins, U., Allin, D., Collins, A. L., Williams, C. H. S., Martens, A. M., & Bladon, K. D. (2016). Sediment-phosphorus dynamics can shift aquatic ecology and cause downstream legacy effects after wildfire in large river systems. *Global Change Biology*, *22*(3), 1168–1184. https://doi.org/10.1111/gcb.13073

Emmerton, C. A., Cooke, C. A., Hustins, S., Silins, U., Emelko, M. B., Lewis, T., Kruk, M. K., Taube, N., Zhu, D., Jackson, B., Stone, M., Kerr, J. G., & Orwin, J. F. (2020). Severe western Canadian wildfire affects water quality even at large basin scales. *Water Research, 183*, 116071. https://doi.org/10.1016/j.watres.2020.116071

Environment and Climate Change Canada. (2019). *CABIN Canadian Aquatic Biomonitoring Network*. Open Data Canada.

Epps, D., & Phippen, B. (2011). *Water quality assessment and objectives for Comox Lake.* (). British Columbia: Ministry of Environment, Environmental Sustainability Division, Environmental Protection Division.

Escapa, I. F., Huang, Y., Chen, T., Lin, M., Kokaras, A., Dewhirst, F. E., & Lemon, K. P. (2020). Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome, 8*(1), 65. https://doi.org/10.1186/s40168-020-00841-w

ESRI. (2020). *ArcGIS pro*

ESRI. (2023). *ArcGIS pro*

Esteves-Ferreira, A. A., Cavalcanti, J. H. F., Vaz, M. G. M. V., Alvarenga, L. V., Nunes-Nesi, A., & Araújo, W. L. (2017). Cyanobacterial nitrogenases: Phylogenetic diversity, regulation and functional predictions. *Genetics and Molecular Biology, 40*(1 Suppl 1), 261-275. https://doi.org/10.1590/1678-4685-GMB-2016-0050

Falconer, I. R. (2008). Health effects associated with controlled exposures to cyanobacterial toxins. In K. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 607-612). Springer.

Favot, E. J., Holeton, C., DeSellas, A. M., & Paterson, A. M. (2023). Cyanobacterial blooms in Ontario, Canada: Continued increase in reports through the 21st century. *Lake and Reservoir Management, ahead-of-print*, 1-20. https://doi.org/10.1080/10402381.2022.2157781

Favot, E. J., Rühland, K. M., DeSellas, A. M., Ingram, R., Paterson, A. M., & Smol, J. P. (2019). Climate variability promotes unprecedented cyanobacterial blooms in a remote, oligotrophic Ontario Lake: Evidence from paleolimnology. *Journal of Paleolimnology, 62*(1), 31-52. https://doi.org/10.1007/s10933-019-00074-4

Flores, E., Picossi, S., Valladares, A., & Herrero, A. (2019). Transcriptional regulation of development in heterocyst-forming cyanobacteria. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms, 1862*(7), 673-684. https://doi.org/10.1016/j.bbagrm.2018.04.006

Furniss, M. J., Staab, B. P., Hazelhurst, S., Clifton, C. F., Roby, K. B., Ilhadrt, B. L., Larry, E. B., Todd, A. H., Reid, L. M., Hines, S. J., Bennett, K. A., Luce, C. H., & Edwards, P. J. (2010). *Water, climate change, and forests: Watershed stewardship for a changing climate.* (). Portland, OR: https://www.fs.fed.us/pnw/pubs/pnw_gtr812.pdf

Giner-Lamia, J., Robles-Rengel, R., Hernández-Prieto, M. A., Muro-Pastor, M. I., Florencio, F. J., & Futschik, M. E. (2017). Identification of the direct regulon of *Ntc*A during early acclimation to nitrogen starvation in the cyanobacterium *Synechocystis* sp. PCC 6803. *Nucleic Acids Research, 45*(20), 11800-11820. https://doi.org/10.1093/nar/gkx860

Government of Canada. (2013). *Water Sources: Lakes.* https://www.canada.ca/en/environment-climate-change/services/water-overview/sources/lakes.html

Government of Canada. (2019). *Turkey Lakes Watershed Study.* Government of Canada. https://www.canada.ca/en/environment-climate-change/services/turkey-lakes-watershed-study.html

Government of Canada. (2021). *Statistical Data.* Government of Canada. https://cfs.nrcan.gc.ca/statsprofile/inventory/atlanticmaritime

Guiry, M. D., & Guiry, G. M. (2023). *AlgaeBase: An on-line resource for algae.* Cryptogamie, Algologie. http://www.bioone.org/doi/abs/10.7872/crya.v35.iss2.2014.105

Harke, M. J., & Gobler, C. J. (2015). Daily transcriptome changes reveal the role of nitrogen in controlling microcystin synthesis and nutrient transport in the toxic cyanobacterium, *Microcystis aeruginosa*. *BMC Genomics, 16*(1066), 1068. https://doi.org/10.1186/s12864-015-2275-9

Havens, K. E. (2008). Cyanobacteria blooms: Effects on aquatic ecosystems. In K. H. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 733-747). Springer.

Health Canada. (2020). *Cyanobacteria and their toxins in recreational water: Guideline technical document for public consultation.* ().Government of Canada.

Health Canada. (2022). *Guidelines for Canadian recreational water quality - cyanobacteria and their toxins.* ().Government of Canada. https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidance-canadian-recreational-water-quality-cyanobacteria-toxins.html

Herrero, A., Muro-Pastor, A. M., Valladares, A., & Flores, E. (2004). Cellular differentiation and the *Ntc*A transcription factor in filamentous cyanobacteria. *FEMS Microbiology Reviews, 28*(4), 469-487. https://doi.org/10.1016/j.femsre.2004.04.003

Hoff, K. J. (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics, 10*(1), 520. https://doi.org/10.1186/1471-2164-10-520

Horn, T., Adel, S., Schumann, R., Sur, S., Kakularam, K. R., Polamarasetty, A., Redanna, P., Kuhn, H., & Heydeck, D. (2015). Evolutionary aspects of lipoxygenases and genetic diversity of human leukotriene signaling. *Progress in Lipid Research; Prog Lipid Res, 57*, 13-39. https://doi.org/10.1016/j.plipres.2014.11.001

Howarth, R. W., Marino, R., & Cole, J. J. (1988). Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 2. biogeochemical controls1. *Limnology and Oceanography, 33*(4), 688-701. https://doi.org/https://doi.org/10.4319/lo.1988.33.4part2.0688

Howarth, R. W., & Marino, R. (2006). Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: Evolving views over three decades. *Limnology and Oceanography, 51*(1part2), 364-376.

Hudnell, H. K., & Dortch, Q. (2008). A synopsis of research needs identified at the interagency, international symposium on cyanobacterial harmful algal blooms (ISOC-HAB). In H. K. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 17-42). Springer.

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

Ibelings, B. W., Kurmayer, R., Azevedo, S. M. F. O., Wood, S. A., Chorus, I., & Welker, M. (2021). Understanding the occurrence of cyanobacteria and cyanotoxins. *Toxic cyanobacteria in water* (2nd ed., pp. 213-294). CRC Press. https://doi.org/10.1201/9781003081449-4

Ininbergs, K., Bay, G., Rasmussen, U., Wardle, D. A., & Nilsson, M. (2011). Composition and diversity of *nif*H genes of nitrogen-fixing cyanobacteria associated with boreal forest feather mosses. *The New Phytologist, 192*(2), 507-517. https://doi.org/10.1111/j.1469-8137.2011.03809.x

Janda, M. J., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology, 45*(9), 2761-2764. https://doi.org/10.1128/JCM.01228-07

Jeffries, D. S., Kelso, J. R. M., & Morrison, I. K. (1988). Physical, chemical, and biological characteristics of the Turkey Lakes Watershed, central Ontario, Canada. *Canadian Journal of Fisheries and Aquatic Sciences, 45*, s3-s13. https://doi.org/10.1139/f88-262

Johansen, J. R., & Casamatta, D. A. (2005). Recognizing cyanobacterial diversity through adoption of a new species paradigm. *Algological Studies/Archiv Für Hydrobiologie, Supplement Volumes, 117*, 71-93. https://doi.org/10.1127/1864-1318/2005/0117-0071

Johnson, J. S., Spakowicz, D. J., Hong, B., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and

strain-level microbiome analysis. *Nature Communications, 10*(1), 5029-11. https://doi.org/10.1038/s41467-019-13036-1

Jungblut, A., & Neilan, B. (2006). Molecular identification and evolution of the cyclic peptide hepatotoxins, microcystin and nodularin, synthetase genes in three orders of cyanobacteria. *Archives of Microbiology, 185*, 107-14. https://doi.org/10.1007/s00203-005-0073-5

Kamennaya, N. A., & Post, A. F. (2011). Characterization of cyanate metabolism in marine *Synechococcus* and *Prochlorococcus* spp. *Applied and Environmental Microbiology, 77*(1), 291-301. https://doi.org/10.1128/AEM.01272-10

Kan, J. (2018). Storm Events Restructured Bacterial Community and Their Biogeochemical Potentials. *Journal of Geophysical Research: Biogeosciences*, *123*(7), 2257–2269. https://doi.org/10.1029/2017JG004289

Kebeish, R., & Al-Zoubi, O. (2017). Expression of the cyanobacterial enzyme cyanase increases cyanate metabolism and cyanate tolerance in *Arabidopsis*. *Environmental Science and Pollution Research, 24*, 11825–11835. https://doi.org/10.1007/s11356-017-8866-z

Kelly, N. E., Javed, A., Shimoda, Y., Zastepa, A., Watson, S., Mugalingam, S., & Arhonditsis, G. B. (2019). A bayesian risk assessment framework for microcystin violations of drinking water and recreational standards in the Bay of Quinte, Lake Ontario, Canada. *Water Research, 162*, 288-301. https://doi.org/https://doi.org/10.1016/j.watres.2019.06.005

Kennedy, J., O'Leary, N. D., Kiran, G. S., Morrissey, J. P., O'Gara, F., Selvin, J., & Dobson, A. D. W. (2011). Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *Journal of Applied Microbiology, 111*(4), 787-799. https://doi.org/10.1111/j.1365-2672.2011.05106.x

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

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution, 16*(2), 111–120. https://doi.org/10.1007/BF01731581

Komárek, J. (2016). A polyphasic approach for the taxonomy of cyanobacteria: Principles and applications. *European Journal of Phycology, 51*(3), 346-353. https://doi.org/10.1080/09670262.2016.1163738

Komárek, J., Johansen, J. R., Šmarda, J., & Strunecký, O. (2020). Phylogeny and taxonomy of *Synechococcus*-like cyanobacteria. *Fottea (Praha), 20*(2), 171-191. https://doi.org/10.5507/fot.2020.006

Komárek, J., Kaštovský, J., Mareš, J., & Johansen, J. R. (2014). Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia 86: 295–335*,

Korosi, J. B., Ginn, B. K., Cumming, B. F., & Smol, J. P. (2013). Establishing past environmental conditions and tracking long-term environmental change in the Canadian maritime provinces using lake sediments. *Environmental Reviews, 21*(1), 15-27. https://doi.org/10.1139/er-2012-0041

Kotak, B. G., & Zurawell, R. W. (2007). Cyanobacterial toxins in Canadian freshwaters: A review. *Null, 23*(2), 109-122. https://doi.org/10.1080/07438140709353915

Kuiper-Goodman, T., Falconer, I., & Fitzgerald, J. (1999). Human health aspects. In I. Chorus, & J. Bartram (Eds.), *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management* (). World Health Organization. E & FN Spon.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution, 35*(6), 1547-1549. https://doi.org/10.1093/molbev/msy096

Kurakin, G. F., Samoukina, A. M., & Potapova, N. A. (2020). Bacterial and protozoan lipoxygenases could be involved in cell-to-cell signaling and immune response suppression. *Biochemistry (Moscow), 85*(9), 1048-1063. https://doi.org/10.1134/S0006297920090059

Kurz, W. A., Shaw, C. H., Boisvenue, C., Stinson, G., Metsaranta, J., Leckie, D., Dyk, A., Smyth, C., & Neilson, E. T. (2013). Carbon in Canada's boreal forest—A synthesis. *Environmental Reviews, 21*(4), 260–292. https://doi.org/10.1139/er-2013-0041

Kwok, A. (unpublished). *Examination of the interdependence of microbial communities and physiochemical characteristics of freshwater ponds within the Canadian Western Boreal Plain*

Latysheva, N., Junker, V. L., Palmer, W. J., Codd, G. A., & Barker, D. (2012). The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics, 28*(5), 603-606.

Leach, J. A., Buttle, J. M., Webster, K. L., Hazlett, P. W., & Jeffries, D. S. (2020). Travel times for snowmelt-dominated headwater catchments: Influences of wetlands and forest harvesting, and linkages to stream water quality. *Hydrological Processes, 34*(10), 2154-2175. https://doi.org/10.1002/hyp.13746

Lemmen, D. S., Warren, F. J., James, T. S., & Mercer Clarke, C. S. L. (2016). *Canada's marine coasts in a changing climate*. Government of Canada.

Li, Q., & Chen, S. (2020). Transfer of nitrogen fixation (*nif*) genes to non-diazotrophic hosts. *Chembiochem : A European Journal of Chemical Biology, 21*(12), 1717-1722. https://doi.org/10.1002/cbic.201900784

Lindsay, J. B., Creed, I. F., & Beall, F. D. (2004). Drainage basin morphometrics for depressional landscapes. *Water Resources Research, 40*(9), W09307-n/a. https://doi.org/10.1029/2004WR003322

Liu, J., Chen, L., & Zhang, X. (2022). Current research scenario for biological effect of exogenous factors on microcystin synthesis. *Environmental Science and Pollution Research International; Environ Sci Pollut Res Int, 29*(18), 26190-26201. https://doi.org/10.1007/s11356-021-18256-x

Liu, T., Mazmouz, R., Pearson, L. A., & Neilan, B. A. (2019). Mutagenesis of the microcystin tailoring and transport proteins in a heterologous cyanotoxin expression system. *ACS Synthetic Biology, 8*(5), 1187-1194. https://doi.org/10.1021/acssynbio.9b00068

Luque-Almagro, V. M., Huertas, M., Sáez, L. P., Luque-Romero, M. M., Moreno-Vivián, C., Castillo, F., Roldán, M. D., & Blasco, R. (2008). Characterization of the pseudomonas *Pseudoalcaligenes* CECT5344 cyanase, an enzyme that is not essential for cyanide assimilation. *Applied and Environmental Microbiology, 74*(20), 6280-6288. https://doi.org/10.1128/AEM.00916-08

Ma, A. T., Beld, J., & Brahamsha, B. (2017). An amoebal grazer of cyanobacteria requires cobalamin produced by heterotrophic bacteria. *Applied and Environmental Microbiology, 83*(10), E00035. https://doi.org/10.1128/AEM.00035-17

MacKeigan, P. W., Garner, R. E., Monchamp, M., Walsh, D. A., Onana, V. E., Kraemer, S. A., Pick, F. R., Beisner, B. E., Agbeti, M. D., da Costa, N. B., Shapiro, B. J., & Gregory-Eaves, I. (2022). Comparing microscopy and DNA metabarcoding techniques for

identifying cyanobacteria assemblages across hundreds of lakes. *Harmful Algae, 113*, 102187. https://doi.org/10.1016/j.hal.2022.102187

Mao, X., Chen, J., van Oosterhout, C., Zhang, H., Liu, G., Zhuang, Y., & Mock, T. (2022). Diversity, prevalence, and expression of cyanase genes (*cyn*S) in planktonic marine microorganisms. *The ISME Journal, 16*(2), 602-605. https://doi.org/10.1038/s41396-021-01081-y

McIntyre, L., Cassis, D., & Haigh, N. (2013). Formation of a volunteer harmful algal bloom network in British Columbia, Canada, following an outbreak of diarrhetic shellfish poisoning. *Marine Drugs, 11*(11), 4144-4157. https://doi.org/10.3390/md11114144

McKindles, K. M., Zimba, P. V., Chiu, A. S., Watson, S. B., Gutierrez, D. B., Westrick, J., Kling, H., & Davis, T. W. (2019). A multiplex analysis of potentially toxic cyanobacteria in Lake Winnipeg during the 2013 bloom season. *Toxins, 11*(10)https://doi.org/10.3390/toxins11100587

McRose, D. L., Zhang, X., Kraepiel, A. M. L., & Morel, F. M. M. (2017). Diversity and activity of alternative nitrogenases in sequenced genomes and coastal environments. *Frontiers in Microbiology, 8*, 267. https://doi.org/10.3389/fmicb.2017.00267

Memon, D., Singh, A. K., Pakrasi, H. B., & Wangikar, P. P. (2013). A global analysis of adaptive evolution of operons in cyanobacteria. *Antonie Van Leeuwenhoek, 103*(2), 331-346. https://doi.org/10.1007/s10482-012-9813-0

Mikalsen, B., Boison, G., Skulberg, O. M., Fastner, J., Davies, W., Gabrielsen, T. M., Rudi, K., & Jakobsen, K. S. (2003). Natural variation in the microcystin synthetase operon *mcy*ABC and impact on microcystin production in *Microcystis* strains. *Journal of Bacteriology, 185*(9), 2774-2785. https://doi.org/10.1128/JB.185.9.2774-2785.2003

Mitton, J. B., & Ferrenberg, S. M. (2012). Mountain Pine Beetle Develops an Unprecedented Summer Generation in Response to Climate Warming. *The American Naturalist, 179*(5), E163–E171. https://doi.org/10.1086/665007

Mo, H., Xie, X., Zhu, T., & Lu, X. (2017). Effects of global transcription factor *Ntc*A on photosynthetic production of ethylene in recombinant *Synechocystis* sp. PCC 6803. *Biotechnology for Biofuels, 10*(1), 145. https://doi.org/10.1186/s13068-017-0832-y

Moffatt, J. H., Harper, M., Adler, B., Nation, R. L., Jian, L. I., Boyce, J. D., Harrison, P., Hale, J. D. F., Vinogradov, E., Seemann, T., Henry, R., Crane, B., Michael, F. S., & Cox, A. D. (2010). Colistin resistance in *Acinetobacter baumannii* is mediated by

complete loss of lipopolysaccharide production. *Antimicrobial Agents and Chemotherapy, 54*(12), 4971-4977. https://doi.org/10.1128/AAC.00834-10

Moisander, P. H., McClinton, E., & Paerl, H. W. (2002). Salinity effects on growth, photosynthetic parameters, and nitrogenase activity in estuarine planktonic cyanobacteria. *Microbial Ecology, 43*(4), 432-442. https://doi.org/10.1007/s00248-001-1044-2

Moosavian, M., Emam, N., Pletzer, D., & Savari, M. (2020). Rough-type and loss of the LPS due to *lpx* genes deletions are associated with colistin resistance in multidrug-resistant clinical *Escherichia coli* isolates not harbouring *mcr* genes. *PloS One, 15*(5), e0233518. https://doi.org/10.1371/journal.pone.0233518

Morrissey, J. F., Sumich, J. L., & Pinkard-Meier, D. R. (2018). *Introduction to the biology of marine life* (11th ed.). Jones & Bartlett Learning.

Nash, M. V., Anesio, A. M., Barker, G., Tranter, M., Varliero, G., Eloe-Fadrosh, E. A., Nielsen, T., Turpin-Jelfs, T., Benning, L. G., & Sánchez-Baracaldo, P. (2018). Metagenomic insights into diazotrophic communities across arctic glacier forefields. *FEMS Microbiology Ecology, 94*(9)https://doi.org/10.1093/femsec/fiy114

Nishiwaki-Matsushima, R., Ohta, T., Nishiwaki, S., Suganuma, M., Kohyama, K., Ishikawa, T., Carmichael, W. W., & Fujiki, H. (1992). Liver tumor promotion by the cyanobacterial cyclic peptide toxin microcystin-LR. *Journal of Cancer Research and Clinical Oncology, 118*(6), 420-424. https://doi.org/10.1007/BF01629424

O'Neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae, 14*, 313-334. https://doi.org/10.1016/j.hal.2011.10.027

O'Keeffe, J. (2019). *Cyanobacteria and drinking water: Occurrence, risks, management and knowledge gaps for public health* . (). Vancouver, BC:

Orihel, D. M., Bird, D. F., Brylinsky, M., Chen, H., Donald, D. B., Huang, D. Y., Giani, A., Kinniburgh, D., Kling, H., Kotak, B. G., Leavitt, P. R., Nielson, C. C., Reedyk, S., Rooney, R. C., Watson, S. B., Zurawell, R. W., & Vinebrooke, R. D. (2012). High microcystin concentrations occur only at low nitrogen-to-phosphorus ratios in nutrient-rich Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, https://doi.org/10.1139/f2012-088

Paerl, H., Fulton, R. S. I., Moisander, P. H., & Dyble, J. (2001). Harmful freshwater algal blooms, with an emphasis on cyanobacteria. *The Scientific World Journal, 1*, 76-113. https://doi.org/10.1100/tsw.2001.16

Paerl, H. (2008). Nutrient and other environmental controls of harmful cyanobacterial blooms along the freshwater–marine continuum. In K. H. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 217-237). Springer.

Paerl, H. (2018). Mitigating toxic planktonic cyanobacterial blooms in aquatic ecosystems facing increasing anthropogenic and climatic pressures. *Toxins; Toxins (Basel), 10*(2), 76. https://doi.org/10.3390/toxins10020076

Paerl, H. W. (1997). Coastal eutrophication and harmful algal blooms: Importance of atmospheric deposition and groundwater as "new" nitrogen and other nutrient sources. *Limnology and Oceanography, 42*(5), 1154-1165. https://doi.org/https://doi.org/10.4319/lo.1997.42.5_part_2.1154

Paerl, H. W., Hall, N. S., & Calandrino, E. S. (2011a). Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change. *Science of the Total Environment, 409*(10), 1739-1745. https://doi.org/https://doi.org/10.1016/j.scitotenv.2011.02.001

Paerl, H. W., Joyner, J. J., Joyner, A. R., Arthur, K., Paul, V., O'Neil, J. M., & Heil, C. A. (2008). Co-occurrence of dinoflagellate and cyanobacterial harmful algal blooms in southwest Florida coastal waters: Dual nutrient (N and P) input controls. *Marine Ecology Progress Series, 371*, 143-153. https://doi.org/10.3354/meps07681

Paerl, H. W., & Otten, T. G. (2013). Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microbial Ecology; Microb Ecol, 65*(4), 995-1010. https://doi.org/10.1007/s00248-012-0159-y

Paerl, H. W., Scott, J. T., McCarthy, M. J., Newell, S. E., Gardner, W. S., Havens, K. E., Hoffman, D. K., Wilhelm, S. W., & Wurtsbaugh, W. A. (2016). It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. *Environmental Science & Technology; Environ.Sci.Technol, 50*(20), 10805-10813. https://doi.org/10.1021/acs.est.6b02575

Paerl, H. W., Xu, H., McCarthy, M. J., Zhu, G., Qin, B., Li, Y., & Gardner, W. S. (2011b). Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The need for a dual nutrient (N & P) management strategy. *Water Research, 45*(5), 1973-1983. https://doi.org/https://doi.org/10.1016/j.watres.2010.09.018

Palatinszky, M., Herbold, C., Jehmlich, N., Pogoda, M., Han, P., von Bergen, M., Lagkouvardos, I., Karst, S. M., Galushko, A., Koch, H., Berry, D., Daims, H., & Wagner, M. (2015). Cyanate as an energy source for nitrifiers. *Nature, 524*(7563), 105-108. https://doi.org/10.1038/nature14856

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology, 36*(10), 996-1004. https://doi.org/10.1038/nbt.4229

Pearson, L. A., Dittmann, E., Mazmouz, R., Ongley, S. E., D'Agostino, P. M., & Neilan, B. A. (2016). The genetics, biosynthesis and regulation of toxic specialized metabolites of cyanobacteria. *Harmful Algae, 54*, 98-111. https://doi.org/10.1016/j.hal.2015.11.002

Pearson, L. A., Hisbergues, M., Börner, T., Dittmann, E., & Neilan, B. A. (2004). Inactivation of an ABC transporter gene, *mcy*H, results in loss of microcystin production in the cyanobacterium *Microcystis aeruginosa* PCC 7806. *Applied and Environmental Microbiology, 70*(11), 6370-6378. https://doi.org/10.1128/AEM.70.11.6370-6378.2004

Pearson, L., Mihali, T., Moffitt, M., Kellmann, R., & Neilan, B. (2010). On the chemistry, toxicology and genetics of the cyanobacterial toxins, microcystin, nodularin, saxitoxin and cylindrospermopsin. *Marine Drugs; Mar Drugs, 8*(5), 1650-1680. https://doi.org/10.3390/md8051650

Pereira, M. B., Wallroth, M., Jonsson, V., & Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics, 19*(1), 274. https://doi.org/10.1186/s12864-018-4637-6

Pérez-Cobas, A. E., Gomez-Valero, L., & Buchrieser, C. 2. (2020). Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. *Microbial Genomics, 6*(8), e000409. https://doi.org/10.1099/mgen.0.000409

Perovich, G., Dortch, Q., Goodrich, J., Berger, P. S., Brooks, J., Evens, T. J., Gobler, C. J., Graham, J., Hyde, J., Karner, D., O'Shea, D., Paul, V., Paerl, H., Piehler, M., Rosen, B. H., Santelmann, M., Tester, P., & Westrick, J. (2008). Causes, prevention, and mitigation workgroup report. In K. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 186-215). Springer.

Persaud, A. D., Paterson, A. M., Dillon, P. J., Winter, J. G., Palmer, M., & Somers, K. M. (2015). Forecasting cyanobacteria dominance in Canadian temperate lakes. *Journal of Environmental Management, 151*, 343-352. https://doi.org/10.1016/j.jenvman.2015.01.009

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

Piehler, M. F. (2008). Watershed management strategies to prevent and control cyanobacterial harmful algal blooms. In K. H. Hudnell (Ed.), *Cyanobacterial harmful algal blooms: State of the science and research needs* (pp. 259-273). Springer.

Powers, S. M., & Hampton, S. E. (2016). Winter limnology as a new frontier. *Limnology and Oceanography Bulletin, 25*(4), 103-108. https://doi.org/10.1002/lob.10152

Prepas, E. E., Planas, D., Gibson, J. J., Vitt, D. H., & et al. (2001). Landscape variables influencing nutrients and phytoplankton communities in Boreal Plain lakes of northern Alberta: A comparison of wetland- and upland-dominated catchments. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(7), 1286-1299. http://search.proquest.com.proxy.lib.uwaterloo.ca/scholarly-journals/landscape-variables-influencing-nutrients/docview/219336093/se-2

Raetz, C. R. H., Guan, Z., Ingram, B. O., Six, D. A., Song, F., Wang, X., & Zhao, J. (2009). Discovery of new biosynthetic pathways: The lipid A story. *Journal of Lipid Research, 50*, S103-S108. https://doi.org/10.1194/jlr.R800060-JLR200

Raetz, C. R. H., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annual Review of Biochemistry, 71*(1), 635-700. https://doi.org/10.1146/annurev.biochem.71.110601.135414

Rashidi, H., Baulch, H., Gill, A., Bharadwaj, L., & Bradford, L. (2021). Monitoring, managing, and communicating risk of harmful algal blooms (HABs) in recreational resources across Canada. *Environmental Health Insights, 15*(1), 11786302211014401. https://doi.org/10.1177/11786302211014401

Redfield, A. C. (1958). The biological control of chemical factors in the environment. *American Scientist, 46*(3), 230A-221. http://www.jstor.org.proxy.lib.uwaterloo.ca/stable/27827150

Reinl, K. L., Harris, T. D., North, R. L., Almela, P., Berger, S. A., Bizic, M., Burnet, S. H., Grossart, H., Ibelings, B. W., Jakobsson, E., Knoll, L. B., Lafrancois, B. M., McElarney, Y., Morales-Williams, A., Obertegger, U., Ogashawara, I., Paule-Mercado, M., Peierls, B. L., Rusak, J. A., . . . Yokota, K. (2023). Blooms also like it cold. *Limnology and Oceanography Letters, n/a*https://doi.org/10.1002/lol2.10316

Reinl, K. L., Brookes, J. D., Carey, C. C., Harris, T. D., Ibelings, B., Morales-Williams, A. M., de Senerpont Domis, L., Atkins, K. S., Isles, P. D. F., Mesman, J. P., North, R. L., Rudstam, L. G., Stelzer, J. A. A., Venkiteswaran, J. J., Yokota, K., & Zhan, Q. (2021). Cyanobacterial blooms in oligotrophic lakes: Shifting the high-nutrient paradigm. *Freshwater Biology, 66*(9), 1846-1859. https://doi.org/10.1111/fwb.13791

Renaudin, M., Blasi, C., Bradley, R. L., & Bellenger, J. (2022). New insights into the drivers of moss-associated nitrogen fixation and cyanobacterial biomass in the eastern Canadian boreal forest. *The Journal of Ecology, 110*(6), 1403-1418. https://doi.org/10.1111/1365-2745.13881

Ribeiro, K. F., Duarte, L., & Crossetti, L. O. (2018). Everything is not everywhere: A tale on the biogeography of cyanobacteria. *Hydrobiologia, 820*(1), 23-48. https://doi.org/10.1007/s10750-018-3669-x

Rippka, R., Derueles, J., Waterbury, J. B., & Herdman, M. S. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Journal of General Microbiology, 111*(1), 1-61. https://doi.org/10.1099/00221287-111-1-1

Robinne, F.-N., Bladon, K. D., Silins, U., Emelko, M. B., Flannigan, M. D., Parisien, M.-A., Wang, X., Kienzle, S. W., & Dupont, D. P. (2019). A Regional-Scale Index for Assessing the Exposure of Drinking-Water Sources to Wildfires. *Forests, 10*(5). https://doi.org/10.3390/f10050384

Romanis, C. S., Pearson, L. A., & Neilan, B. A. (2021). Cyanobacterial blooms in wastewater treatment facilities: Significance and emerging monitoring strategies. *Journal of Microbiological Methods; J Microbiol Methods, 180*, 106123. https://doi.org/10.1016/j.mimet.2020.106123

Rouhiainen, L., Vakkilainen, T., Simer, B. L., Buikema, W., Haselkorn, R., & Sivonen, K. (2004). Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium anabaena strain 90. *Applied and Environmental Microbiology, 70*(2), 686-692. https://doi.org/10.1128/AEM.70.2.686-692.2004

Sadvakasova, A. K., Kossalbayev, B. D., Token, A. I., Bauenova, M. O., Wang, J., Zayadan, B. K., Balouch, H., Alwasel, S., Leong, Y. K., Chang, J., & Allakhverdiev, S. I. (2022). Influence of Mo and Fe on photosynthetic and nitrogenase activities of nitrogen-fixing cyanobacteria under nitrogen starvation. *Cells (Basel, Switzerland); Cells, 11*(5), 904. https://doi.org/10.3390/cells11050904

Sáez, L. P., Cabello, P., Ibáñez, M. I., Luque-Almagro, V. M., Roldán, M. D., & Moreno-Vivián, C. (2019). Cyanate assimilation by the alkaliphilic cyanide-degrading bacterium

Pseudomonas pseudoalcaligenes CECT5344: Mutational analysis of the *cyn* gene cluster. *International Journal of Molecular Sciences, 20*(12), 3008. https://doi.org/10.3390/ijms20123008

Scherer, P. I., Raeder, U., Geist, J., & Zwirglmaier, K. (2017). Influence of temperature, mixing, and addition of microcystin-LR on microcystin gene expression in *Microcystis aeruginosa*. *MicrobiologyOpen, 6*(1), e00393-n/a. https://doi.org/10.1002/mbo3.393

Schindler, D. W. (1975). Whole-lake eutrophication experiments with phosphorus, nitrogen and carbon. *Null, 19*(4), 3221-3231. https://doi.org/10.1080/03680770.1974.11896436

Shardlow, T. (2021). *Identification and characterization of toxic cyanobacteria in two forested maritime watersheds in North America* http://hdl.handle.net/10012/17055

Sharma, J., Sharma, S., Sharma, I., Chopra, C., & Sharma, V. (2020). The coming together of sciences: Metagenomics for microbial biochemistry . In Reena Singh Chopra, Chirag Chopra & Neeta Raj Sharma (Eds.), *Metagenomics: Techniques, applications, challenges and opportunities* (pp. 23-36). Springer. https://doi.org/10.1007/978-981-15-6529-8_2

Sharoni, S., & Halevy, I. (2020). Nutrient ratios in marine particulate organic matter are predicted by the population structure of well-adapted phytoplankton. *Science Advances, 6*(29), eaaw9371. https://doi.org/10.1126/sciadv.aaw9371

Sheath, R. (2006). *Biogeography of freshwater algae*. John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0004278

Silins, U., Anderson, A., Bladon, K. D., Emelko, M. B., Stone, M., Spencer, S. A., Williams, C. H. S., Wagner, M. J., Martens, A. M., & Hawthorn, K. (2016). Southern Rockies Watershed Project. *Forestry Chronicle, 92*(1), 39-42. https://doi.org/10.5558/tfc2016-012

Silins, U., Bladon, K. D., Kelly, E. N., Esch, E., Spence, J. R., Stone, M., Emelko, M. B., Boon, S., Wagner, M. J., Williams, C. H. S., & Tichkowsky, I. (2014). Five-year legacy of wildfire and salvage logging impacts on nutrient runoff and aquatic plant, invertebrate, and fish productivity. *Ecohydrology*, *7*(6), 1508–1523. https://doi.org/10.1002/eco.1474

Singh, J. S., Kumar, A., Rai, A. N., & Singh, D. P. (2016). Cyanobacteria: A precious bio-resource in agriculture, ecosystem, and environmental sustainability. *Frontiers in Microbiology*, *7*, 529. https://doi.org/10.3389/fmicb.2016.00529

Sivonen, K., & Jones, G. (1999). Cyanobacterial toxins. In I. Chorus, & J. Bartram (Eds.), *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management* (). World Health Organization. E & FN Spon.

Soltis, P. S., & Soltis, D. E. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Statistical Science*, *18*(2), 256–267. JSTOR.

Soo, R. M., Skennerton, C. T., Sekiguchi, Y., Imelfort, M., Paech, S. J., Dennis, P. G., Steen, J. A., Parks, D. H., Tyson, G. W., & Hugenholtz, P. (2014). An expanded genomic representation of the phylum Cyanobacteria. *Genome Biology and Evolution, 6*(5), 1031-1045. https://doi.org/10.1093/gbe/evu073

Steenberg, J. W. N., Duinker, P. N., & Bush, P. G. (2013). Modelling the effects of climate change and timber harvest on the forests of central Nova Scotia, Canada. *Annals of Forest Science, 70*(1), 61-73. https://doi.org/10.1007/s13595-012-0235-y

Sterner, R. W. (2008). On the phosphorus limitation paradigm for lakes. *International Review of Hydrobiology, 93*(4), 433-445. https://doi.org/https://doi.org/10.1002/iroh.200811068

Stoddard, J. L., Van Sickle, J., Herlihy, A. T., Brahney, J., Paulsen, S., Peck, D. V., Mitchell, R., & Pollard, A. I. (2016). Continental-Scale Increase in Lake and Stream Phosphorus: Are Oligotrophic Systems Disappearing in the United States? *Environmental Science & Technology*, *50*(7), 3409–3415. https://doi.org/10.1021/acs.est.5b05950

Stone, M., Emelko, M. B., Droppo, I. G., & Silins, U. (2011). Biostabilization and erodibility of cohesive sediment deposits in wildfire-affected streams. *Water Research (Oxford), 45*(2), 521-534. https://doi.org/10.1016/j.watres.2010.09.016

Strunecký, O., Ivanova, A. P., & Mareš, J. (2023). An updated classification of cyanobacterial orders and families based on phylogenomic and polyphasic analysis. *Journal of Phycology*, https://doi.org/https://doi.org/10.1111/jpy.13304

Talucci, A. C., & Krawchuk, M. A. (2019). Dead forests burning: The influence of beetle outbreaks on fire severity and legacy structure in sub-boreal forests. *Ecosphere*, *10*(5), e02744. https://doi.org/10.1002/ecs2.2744

Tang, W., Liu, Y., Li, Q., Chen, L., Li, Q., Li, P., & Xia, S. (2022). The Impact of Extreme Weather Events on Bacterial Communities and Opportunistic Pathogens in a Drinking Water Treatment Plant. *Water*, *14*(1). https://doi.org/10.3390/w14010054

Tannant, D. D., & Skermer, N. (2013). Mud and debris flows and associated earth dam failures in the Okanagan region of British Columbia. *Canadian Geotechnical Journal, 50*(8), 820-833. https://doi.org/10.1139/cgj-2012-0206

Thiel, T. (2019). Organization and regulation of cyanobacterial *nif* gene clusters: Implications for nitrogenase expression in plant cells. *FEMS Microbiology Letters, 366*(7) https://doi.org/10.1093/femsle/fnz077

Tooming-Klunderud, A., Mikalsen, B., Kristensen, T., & Jakobsen, K. S. (2008). The mosaic structure of the *mcy*ABC operon in *Microcystis*. *Microbiology (Society for General Microbiology), 154*(7), 1886-1899. https://doi.org/10.1099/mic.0.2007/015875-0

Tropea, A. E., Ginn, B. K., Cumming, B. F., & Smol, J. P. (2007). Tracking long-term acidification trends in Pockwock Lake (Halifax, Nova Scotia), the water supply for a major eastern Canadian city. *Lake and Reservoir Management, 23*(3), 279-286. https://doi.org/10.1080/07438140709354016

Tsukuda, M., Kitahara, K., & Miyazaki, K. (2017). Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16S rRNAs. *Scientific Reports, 7*(1), 9993-8. https://doi.org/10.1038/s41598-017-10214-3

United Nations. (2022). *The sustainable development goals report 2022.* (). Retrieved from Aquatic Science & Fisheries Abstracts (ASFA) Professional https://search.proquest.com/docview/2739735175

Valadez-Cano, C., Hawkes, K., Calvaruso, R., Reyes-Prieto, A., & Lawrence, J. (2022). Amplicon-based and metagenomic approaches provide insights into toxigenic potential in understudied Atlantic Canadian lakes. *Facets, 7*, 194-214. https://doi.org/10.1139/facets-2021-0109

Wagner, M. J., Bladon, K. D., Silins, U., Williams, C. H. S., Martens, A. M., Boon, S., MacDonald, R. J., Stone, M., Emelko, M. B., & Anderson, A. (2014). Catchment-scale stream temperature response to land disturbance by wildfire governed by surface–subsurface energy exchange and atmospheric controls. *Journal of Hydrology, 517*, 328–338. https://doi.org/10.1016/j.jhydrol.2014.05.006

Walter, J. M., Coutinho, F. H., Dutilh, B. E., Swings, J., Thompson, F. L., & Thompson, C. C. (2017). Ecogenomics and taxonomy of Cyanobacteria phylum. *Frontiers in Microbiology, 8*, 2132. https://doi.org/10.3389/fmicb.2017.02132

Wang, X., & Quinn, P. J. (2010). Lipopolysaccharide: Biosynthetic pathway and structure modification. *Progress in Lipid Research, 49*(2), 97-107. https://doi.org/10.1016/j.plipres.2009.06.002

Watanabe, T., & Horiike, T. (2021). The evolution of molybdenum dependent nitrogenase in cyanobacteria. *Biology (Basel, Switzerland), 10*(4), 329. https://doi.org/10.3390/biology10040329

Watson, S. B., Satchwill, T., Dixon, E., & Mccauley, E. (2001). Under-ice blooms and source-water odour in a nutrient-poor reservoir: Biological, ecological and applied perspectives. *Freshwater Biology, 46*(11), 1553-1567. https://doi.org/10.1046/j.1365-2427.2001.00769.x

Watson, S. (2004). Aquatic taste and odor: A primary signal of drinking-water integrity. *Journal of Toxicology and Environmental Health. Part A, 67*(20-22), 1779-1795. https://doi.org/10.1080/15287390490492377

Webster, K. L., Leach, J. A., Hazlett, P. W., Fleming, R. L., Emilson, E. J. S., Houle, D., Chan, K. H. Y., Norouzian, F., Cole, A. S., O'Brien, J. M., Smokorowski, K. E., Nelson, S. A., & Yanni, S. D. (2021). Turkey Lakes Watershed, Ontario, Canada: 40 years of interdisciplinary whole-ecosystem research. *Hydrological Processes, 35*(4), n/a. https://doi.org/10.1002/hyp.14109

Wejnerowski, Ł, Rzymski, P., Kokociński, M., & Meriluoto, J. (2018). The structure and toxicity of winter cyanobacterial bloom in a eutrophic lake of the temperate zone. *Ecotoxicology (London), 27*(6), 752-760. https://doi.org/10.1007/s10646-018-1957-x

Wells, M. L., Karlson, B., Wulff, A., Kudela, R., Trick, C., Asnaghi, V., Berdalet, E., Cochlan, W., Davidson, K., De Rijcke, M., Dutkiewicz, S., Hallegraeff, G., Flynn, K. J., Legrand, C., Paerl, H., Silke, J., Suikkanen, S., Thompson, P., & Trainer, V. L. (2020). Future HAB science: Directions and challenges in a changing climate. *Harmful Algae, 91*, 101632. https://doi.org/10.1016/j.hal.2019.101632

WHO. (2003). *Cyanobacterial toxins: Microcystin-LR in drinking-water background document for development of WHO guidelines for drinking-water quality.* (). Geneva:

Wiken, E. (2015). *Natural regions.* The Canadian Encyclopedia. https://www.thecanadianencyclopedia.ca/en/article/natural-regions

Williams, C. H. S., Silins, U., Spencer, S. A., Wagner, M. J., Stone, M., & Emelko, M. B. (2019). Net precipitation in burned and unburned subalpine forest stands after wildfire in the northern Rocky Mountains. *International Journal of Wildland Fire*, *28*(10), 750–760.

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Braekel, J. V., Fu, Q., Roosens, N. H., Keersmaecker, S. C. D., & Vanneste, K. (2019). Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (Illumina) and third (oxford nanopore technologies) generation sequencing technologies. *International Journal of Molecular Sciences, 21*(1), 298. https://doi.org/10.3390/ijms21010298

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

Wiśniewska, K., Lewandowska, A. U., & Śliwińska-Wilczewska, S. (2019). The importance of cyanobacteria and microalgae present in aerosols to human health and the environment – review study. *Environment International, 131,* 104964. https://doi.org/10.1016/j.envint.2019.104964

Wiśniewska, K. A., Śliwińska-Wilczewska, S., & Lewandowska, A. U. (2020). The first characterization of airborne cyanobacteria and microalgae in the Adriatic Sea region. *Plos One, 15*(9), e0238808. https://doi.org/10.1371/journal.pone.0238808

Yang, B., Wang, Y., & Qian, P. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics, 17*(1), 135. https://doi.org/10.1186/s12859-016-0992-y

Zanchett, G., & Oliveira-Filho, E. C. (2013). Cyanobacteria and cyanotoxins: From impacts on aquatic ecosystems and human health to anticarcinogenic effects. *Toxins, 5*(10), 1896-1917. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813918/

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., & Lu, Z. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *The Science of the Total Environment, 618,* 1254-1267. https://doi.org/10.1016/j.scitotenv.2017.09.228

Zhang, X., Li, B., Xu, H., Wells, M., Tefsen, B., & Qin, B. (2019). Effect of micronutrients on algae in different regions of Taihu, a large, spatially diverse, hypereutrophic lake. *Water Research, 151,* 500–514. https://doi.org/10.1016/j.watres.2018.12.023

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., & Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Research; Genome Res, 16*(9), 1099-1108. https://doi.org/10.1101/gr.5322306

Zhou, C., Chen, H., Zhao, H., & Wang, Q. (2021). Microcystin biosynthesis and toxic effects. *Algal Research, 55*, 102277. https://doi.org/10.1016/j.algal.2021.102277

Appendix A

V4-16S rRNA Region Gene ASV Community Composition

Table A1. Summary of samples isolated and sequenced for V4-16S rRNA region gene analysis from five major forested watersheds in Canada. Environmental water samples were obtained by various *for*Water HQP at designated watersheds in each ecozone. The sampling dates and associated sample IDs and sample sites are all included and the number of cyanobacterial ASVs obtained from the samples and used for phylogenetic analysis is recorded.

Ecozone	Watershed	Sample Site	Sample ID	Sample Date	Number of cyano ASVs
Pacific Maritime	Comox Lake Watershed	Boston Creek	M-BOC	May 2019	42
		Cruikshank River	M-CSR	May 2019	
		Upper Puntledge	M-UPL	May 2019	
		Lake Outlet	M-LOT	May 2019	
		Boston Creek	S-BOC	Sept 2019	
		Cruikshank River	S-CSR	Sept 2019	
		Upper Puntledge	S-UPL	Sept 2019	
		Lake Outlet	S-LOT	Sept 2019	
	Sooke River & Leech River Watershed	Jarvis Lake	Jy-JL	July 2019	36
		Jarvis Lake	A-JL	Aug 2019	
		Weeks Lake	Jy-WL	July 2019	
		Weeks Lake	A-WL	Aug 2019	
		Deception Reservoir	Jy-DR	July 2019	
		Deception Reservoir	A-DR	Aug 2019	
Atlantic Maritime	Pockwock Lake Watershed	Pockwock Lake	J-PL	June 2019	19
		Pockwock Lake	A-PL	Aug 2019	
		Pockwock Lake	O-PL	Oct 2019	

		Island Lake	J-IL	June 2019	
		Island Lake	S-IL	Sept 2019	
		Island Lake	O-IL	Oct 2019	
Boreal Shield	Turkey Lakes Watershed	Big Turkey Lake	TLW37	July 18, 2018	123
			TLW42	July 18, 2018	
			TLW43	July 18, 2018	
			TLW82	Aug 13, 2018	
			TLW85	Aug 13, 2018	
			TLW88	Aug 13, 2018	
			TLW171	Oct 25, 2018	
			TLW181	Feb 19, 2019	
			TLW232	May 23, 2019	
			TLW277	Jun 28, 2019	
			TLW322	July 24, 2019	
			TLW325	July 24, 2019	
			TLW367	Aug 21, 2019	
Montane Cordillera	Glenmore Reservoir	Heritage Cove	HC1A	July 27, 2017	11
		Weaselhead	WH2A	July 27, 2017	
		Mid-Lake	ML3A	July 27, 2017	
		Head Pond	HP4A	July 27, 2017	
		Bearspaw West	BW	Sept 21, 2020	10
		Bearspaw East	BE	Sept 21, 2020	
		Glenmore - Raw Monitoring Station	RMS	Sept 22, 2020	
		Bearspaw - Raw 1	R1	Sept 22, 2020	
		Headpond	HP	Sept 23, 2020	
		Weaselhead	WH	Sept 23, 2020	
			Pond 1	July 8, 2018	
			Pond 2	July 8, 2018	
			Pond 5	July 7, 2018	
			Pond 7	July 8, 2018	
			Pond 11	July 8, 2018	
			Pond 15	July 8, 2018	
			Pond 16	July 7, 2018	
			Pond 17	July 8, 2018	
			Pond 27	July 8, 2018	

Boreal Plains	Utikuma Region Study Area (URSA)	Water bodies within URSA	Pond 39	July 5, 2018	158
			Pond 40	July 5, 2018	
			Pond 42	July 5, 2018	
			Pond 43	July 5, 2018	
			Pond 48	July 5, 2018	
			Pond 59	July 8, 2018	
			Pond 111	July 8, 2018	
			Pond 112	July 8, 2018	
			Pond 118	July 7, 2018	
			Pond 121	July 6, 2018	
			Pond 122	July 7, 2018	
			Pond 162	July 6, 2018	
			Pond 168	July 6, 2018	
			Pond 169	July 6, 2018	
			Pond 171	July 6, 2018	
			Pond 201	July 7, 2018	
			Pond 205	July 9, 2018	
			Pond 206	July 7, 2018	
			Pond 208	July 7, 2018	
			Pond 300	July 6, 2018	
			Pond 301	July 6, 2018	
			Pond 303	July 6, 2018	

Table A2. Cyanobacterial reference sequences and accession numbers. Documented reference sequences were obtained from NCBI GenBank, AlgaeBase, and Cyanotype databases and used in the V4-16S rRNA gene phylogenetic tree in Chapter 2 to validate cyanobacteria communities.

Scientific Name	Accession #	Reference
Acaryochloris marina MBIC11017	NR 074407.1:491-743	Swingley et al., 2008 10.1073/pnas.0709772105
Aliterella atlantica CENA595	NZ JYON01000056.1	Rigonato et al., 2015 10.1128/genomeA.00384-15
Aliterella antarctica strain CENA408	NR 151904.1	Rigonato et al., 2016 10.1099/ijsem.0.001066
Aliterella chasmolithica PJ S15	MN243145.1	Jung et al., 2020 10.1111/jpy.13025
Aphanizomenon flos-aquae NIES81	AJ293131.1	Gugger et al., 2002 10.1099/00207713-52-5-1867
Aphanizomenon gracile NIVA-CYA 662 isolate AB2008/23	HG917857.1	Casero et al., 2014 10.1016/j.hal.2014.05.006
Aphanizomenon sp. PCC 7905	AJ133154.1	Lyra et al., 2001 10.1099/00207713-51-2-513
Calothrix sp. KVSF5	EU022730.1	N/A
Calothrix sp. UAM 374	HM751856.1	Berrendero et al., 2011 10.1099/ijs.0.027581-0
Calothrix desertica PCC 7102	NR 114995.1	Sihvonen et al., 2007 10.1111/j.1574-6941.2007.00321.x
Chamaesiphon investiens	JX413491.1	Loza et al., 2013 10.1128/AEM.03351-12
Chamaesiphon subglobosus PCC 7430	AY170472.1	N/A
Chamaesiphon minutus PCC 6605	KY704112.1	Kurmayer et al., 2018 10.1007/s10750-017-3295-z
Chroococcidiopsis sp. BB79.2	AJ344552.1	Fewer et al., 2002 10.1006/mpev.2001.1075
Chroococcidiopsis cubana SAG 39.79	MK484708.1	De Wever et al., 2019 10.1111/gbi.12358
Chroococcidiopsis sp. PCC 6712	AJ344557.1	Fewer et al., 2002 10.1006/mpev.2001.1075
Cyanobium gracile PCC 6307	NR102447.1	N/A
Cyanobium sp. CENA122	KP835527.1	
Cyanobium sp. CENA118	HQ380799.1	Genuário et al., 2016

		10.1007/s10750-016-2802-y
Cyanothece aeruginosa SAG 87.79	KM019992.1	N/A
Cyanothece sp. PCC 8303	MK484713.1	De Wever et al., 2019 10.1111/gbi.12358
Cyanothece aeruginosa strain NIVA-CYA 258/2	Z82775.1	Rudi et al., 1997 10.1128/aem.63.7.2593-2599.1997
Caloramator boliviensis strain 45B	NR 108460.1:496-745	Crespo et al., 2012 10.1099/ijs.0.032664-0
Cylindrospermum marchicum strain CCALA 1001	NR 125689.1:154-406	Johansen et al., 2014 10.1111/jpy.12150
Gloeocapsa sp. PCC 73106 small subunit	AF132784.1	Turner et al., 1999 10.1111/j.1550-7408.1999.tb04612.x
Gloeocapsa sp. CCMEE6058	AY790852.1	Norris and Castenholz, 2006 10.1111/j.1574-6941.2006.00134.x
Gloeocapsa thermalis PUPCCC 008.4	KM376979.1	Singh et al., 2018 10.1016/j.algal.2017.11.023
Gloeobacter kilaueensis JS1	NR 121745.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloeobacter violaceus PCC 7421	NR 074282.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Haloplasma contractile SSD-17B	NR 044362.1:499-746	Antunes et al., 2008 10.1128/JB.01860-07
Kamptonema animale CCALA 139	KP221931.1	N/A
Kamptonema (Oscillatoria) sp. PCC 6407	AM398782.1	Marquardt and Palinska, 2007 10.1007/s00203-006-0204-7
Kamptonema sp. CCM-UFV066	MG563377.1	Obuekwe et al., 2019 10.1016/j.algal.2019.101589
Leptolyngbya sp. SAG 2411	KF417652.1	Brinkmann et al., 2015 https://doi.org/10.1080/01490451.2014.901438
Leptolyngbya frigida ANT.L52.2	AY493575.1	Taton et al., 2006 10.1111/j.1529-8817.2006.00278.x
Leptolyngbya foveolarum PJ S31a	MN267144.1	Jung et al., 2019 10.1002/mbo3.729
Lyngbya aestuarii PCC 7419	NR 114680.1:457-708	Nübel et al., 1997

		10.1128/aem.63.8.3327-3332.1997
Macrochaete psychrophila CCALA	KT336439.2	Gómez et al., 2016 10.1111/jpy.12425
Macrochaete santannae CCALA 1093 clone operon 1	KT336440.1	Gómez et al., 2016 10.1111/jpy.12425
Macrochaete lichenoides SAG 32.92	KU559618.1	Gómez et al., 2016 10.1111/jpy.12425
Lyngbya (Microseira) wollei str. Carmichael/Alabama	EU439567.1	Kellmann et al., 2008 10.1128/AEM.00353-08
Microseira wollei YC2010 clone B9	KM077455.1	McGregor and Sendall, 2015 10.1111/jpy.12256
Microseira wollei YC1109 clone A45	KM077454.1	McGregor and Sendall, 2015 10.1111/jpy.12256
Nostoc punctiforme PCC 73102	NR 074317.1	N/A
Nostoc edaphicum PJ S4	MN243123.1	Jung et al., 2019 10.1002/mbo3.729
Nostoc commune for strain:SO-42	AB098071.1	Arima et al., 2012 10.1111/j.1574-6941.2011.01195.x
Numidum massiliense strain mt3	NR 144715.1:549-796	Alou et al., 2016 10.1016/j.nmni.2016.05.009
Oscillatoriales cyanobacterium USR001	MBRE01000011.1	Te et al., 2016 10.1128/genomeA.01115-16
Oscillatoria lutea SAG 1459-3	KM019965.1	N/A
Phormidium animale CCAP 1459/6	HF678514.1	N/A
Kamptonema animale SAG 1459-6	EF654087.1	Siegesmund et al., 2008 10.1111/j.1529-8817.2008.00604.x
Phormidesmis priestleyi ANT.L52.6	AY493579.1	Taton et al., 2006 10.1111/j.1529-8817.2006.00278.x
Phormidesmis nigrescens LK016	KU219735.1	Raabová et al., 2019 10.11646/phytotaxa.395.1.1
Phormidesmis communis KT5	MK861902.1	Strunecký et al., 2020 10.1093/femsec/fiz189
Phormidium sp. CYN64	JQ687330.1	N/A
Phormidium autumnale Ant-Ph68	DQ493874.1	Comte et al., 2007 10.1111/j.1574-6941.2006.00257.x

Phormidium uncinatum PACC 8693	KF770970.1	Stoyanov et al., 2014 10.4137/EBO.S13748
Planktothrix mougeotii TR1-5	NR 112129.1:401-653	Suda et al., 2002 10.1099/00207713-52-5-1577
Radiocystis sp. JJ30-12	AM710388.1	N/A
Radiocystis sp. enrichment culture clone A2	KF359770.1	N/A
Radiocystis sp. JJ30-3	AM710389.1	N/A
Rhabdogloea smithii SAG 47.91	KM020002.1	N/A
Stanieria cyanosphaera PCC 7437	NR 114510.1:454-706	Turner et al., 1999 10.1111/j.1550-7408.1999.tb04612.x
Scytonema hofmannii PCC 7110	NR 112180.1	Tomitani et al., 2006 10.1073/pnas.0600999103
Scytonema cf. fritschii UCFS23	JN565282.1	Smith et al., 2012 https://doi.org/10.2216/11-84.1
Scytonema hyalinum ATA-SAL-RM1	KY365479.1	Johansen et al., 2017 10.1371/journal.pone.0186393
Scytonema hofmanni UTEX 2349	NZ KK073768.1	N/A
Snowella litoralis 0TU37S04	AJ781040.1	Rajaniemi-Wacklin et al., 2006 10.1111/j.1529-8817.2006.00179.x
Snowella rosea 1LM40S01	AJ781042.1	Rajaniemi-Wacklin et al., 2006 10.1111/j.1529-8817.2006.00179.x
Snowella sp. 249/25	MF680055.1	Shishido et al., 2017 10.1021/acschembio.7b00570
Synechococcus sp. PCC 7502	AF448080.1	N/A
Synechococcus elongatus PCC 6301	NR 074309.1	Maeda et al., 2006 10.1074/jbc.M513196200
Synechococcus sp. LEGE 06306	HM217052.1	Lopes et al., 2012 10.1016/j.marenvres.2011.10.005
Tychonema bourrellyi CCAP 1459/11B	AB045897.1	Suda et al., 2002 10.1099/00207713-52-5-1577
Tychonema bornetii NIVA-CYA 95	LM651415.1	Shams et al., 2015 10.1016/j.watres.2014.11.006
Tychonema bourrellyi isolate FEM MT1130	LT546478.1	Salmaso et al., 2016 10.1093/femsec/fiw155
Anabaena inaequalis CCAP 1446/1A	KT290324.1	N/A

Aphanocapsa muscicola 5N-04	FR798920.1	Cuzman et al., 2010 10.1007/s00248-010-9672-z
Aphanothece stagnina H7	FR848375.1	Komárek et al., 2011 10.1080/09670262.2011.606373
Aphanothece microscopica CCY 153	FR873836.1	Komárek et al., 2011 10.1080/09670262.2011.606373
Anabaena planctonica 1ES36S2	DQ264159.1	Willame et al., 2005 10.1111/j.1529-8817.2006.00284.x
Anabaena spiroides 7RO37S1	DQ264162.1	Willame et al., 2005 10.1111/j.1529-8817.2006.00284.x
Anabaena crassa 1ES33S1	DQ264175.1	Willame et al., 2005 10.1111/j.1529-8817.2006.00284.x
Anabaena compacta 0RE27S3	DQ264176.1	Willame et al., 2005 10.1111/j.1529-8817.2006.00284.x
Aphanothece minutissima 2LT34S03	FM177488.1	N/A
Cyanobium sp. UTMC6012	MN733324.1	N/A
Cyanothece aeruginosa strain NIVA-CYA 258/2	Z82775.1	Rudi et al., 1997 10.1128/aem.63.7.2593-2599.1997
Cyanobium sp. NS01	AY172837.1	Fuller et al., 2003 10.1128/AEM.69.5.2430-2443.2003
Chroococcus turgidus HUW 799	DQ460703.1	Palinska et al., 2006 10.1099/ijs.0.64417-0
Cyanothece sp. GSL007	FJ546715.1	N/A
Cyanobium sp. 2LT07S05	FM177489.1	Mugnai et al., 2008 10.1127/1864-1318/2008/0128-0037
Chroococcus sp. VP2-02 VP2-02	FR798932.1	Cuzman et al., 2010 10.1007/s00248-010-9672-z
Chroococcus sp. CCALA 702	GQ375044.1	Komárková et al., 2010 https://doi.org/10.1007/s10750-009-0015-3
Chroococcus minor BDU 91342	GU186889.1	N/A
Chalicogloea cavernicola CCALA 975	JQ967037.1	Roldán et al., 2013 10.1099/ijs.0.045468-0

Cyanobium sp. LEGE 06113	KC469577.1	Leão et a., 2013 10.1371/journal.pone.0069562
Cyanobacterium aponinum strain PCC 10605	NR 102443.1	N/A
Cyanobacterium stanieri PCC 7202	NR 102450.1	N/A
Cyanothece aeruginosa SAG 87.79	KM019992.1	N/A
Cyanobium sp. UMPCCC 1207	KM218868.1	Blank and Hinman, 2016 10.1016/j.algal.2016.02.014
Cyanobium sp. CENA132	KP835535.1	Genuário et al., 2016 10.1007/s10750-016-2802-y
Cyanobium sp. Bylot Island	KT753317.1	Pessi et al., 2016 10.1111/jpy.12399
Cyanobium sp. Laguna Chica	KT753319.1	Pessi et al., 2016 10.1111/jpy.12399
Cyanobacterium sp. KSU-AQIQ-3	LN997855.1	N/A
Chroococcus sp. FPU101	LC229082.1	N/A
Chroococcus turgidus CCIBt3508	MF072352.1	Gama et al., 2019 10.1080/09670262.2018.1563913
Chroococcus subviolaceus CCIBt3549	MF072353.1	Gama et al., 2019 10.1080/09670262.2018.1563913
Cyanothece aeruginosa SAG 87.79 clone 13	MF781037.1	Mareś et al., 2019 10.1111/jpy.12853
Cyanothece aeruginosa SAG 87.79 clone 11	MF781038.1	Mareś et al., 2019 10.1111/jpy.12853
Cyanobium sp. ULC183	MK280770.1	N/A
Dolichospermum viguieri SAG 27.79	KT290327.1	Kozlíková-Zapomelova et al., 2016 10.5507/fot.2015.026
Dolichospermum flos-aquae 2-Castkovec2/11	KT290337.1	Kozlíková-Zapomelova et al., 2016 10.5507/fot.2015.026
Gloeobacter kilaueensis JS1	NR 121745.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloeobacter violaceus PCC 7421	NR 074282.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Geitlerinema carotinosum MK80	KT315947.1	Strunecký et al., 2017 10.5507/fot.2016.025

Kamptonema animale CCALA 771	KP221932.1	Strunecký et al., 2014 https://www.researchgate.net
Limnoraphis robusta strain CCALA 966	NR 118325.1	Komárek et al., 2013 10.5507/fot.2013.004
Limnolyngbya circumcreta CHAB4449 clone 1	KR697754.1	N/A
Leptolyngbya boryana CCALA 082	KU309579.1	N/A
Microcystis wesenbergii	D89034.1	Kondo et al., 2000 10.2331/suisan.64.291
Oscillatoria sp. LK38	KU219705.1	N/A
Phormidium cf. setchellianum CCALA 149	GQ504022.1	Strunecký et al., 2010 10.1007/s00300-010-0834-8
Phormidium subfuscum CCALA 152	GQ504023.1	Strunecký et al., 2010 10.1007/s00300-010-0834-8
Phormidium murrayi A21	HQ873485.1	Strunecký et al., 2011 10.5507/fot.2011.007
Phormidium autumnale KI19	JN230347.1	Strunecký et al., 2012 10.1111/j.1574-6941.2012.01426.x
Pseudanabaena sp. NIVA-CYA 111	KF735768.1	Kutovaya & Watson, 2014 10.1016/j.jglr.2014.03.016
Pseudanabaena cinerea NIES-4065	LC314118.1	Tuji & Niiyama, 2018 10.1111/pre.12327
Synechocystis sp. KSU-AQIQ-1	LN997853.1	N/A
Snowella sp. 249/25	MF680055.1	Shishido et al., 2017 10.1021/acschembio.7b00570
Synechococcus elongatus var. vestitus CCALA 187	MK907679.1	Komárek et al., 2020 10.5507/fot.2020.006
Synechococcus elongatus PCC 6301	NR 074309.1	Sugita et al., 2007 10.1007/s11120-006-9122-4
Rhabdogloea smithii SAG 47.91	KM020002.1	N/A
Synechocystis aquatilis 1LT32S04	FM177503.1	Mugnai et al., 2008 10.1127/1864-1318/2008/0128-0037
Escherichia coli strain U 5/41	NR 024570.1	Cilia et al., 1996 10.1093/oxfordjournals.molbev.a025606
Klebsiella pneumoniae strain DSM 30104	NR 036794.1	Ludwig et al., 1995 10.1016/S0723-2020(11)80388-7

Escherichia coli str. K-12 substr. MG1655	NC 000913.3	Riley et al., 2006 10.1093/nar/gkj405
Shigella dysenteriae strain ATCC 13313	NR 026332.1	Wang et al., 1997 10.1006/mcpr.1997.0136
Proteus vulgaris	J01874.1	Carbon et al., 1981 10.1093/nar/9.10.2325

Appendix B

Turkey Lakes Watershed Environmental Conditions and Sampling Summary

Table B1. Sampling collection, conditions, and extraction details for all samples chosen for metagenomic sequencing. Weather conditions provided by Environment and Climate Change Canada and Natural Resources Canada field technicians as detailed by Cameron (2021) (conditions not provided were due to missing field data sheets). Sampling depth and volume filtered were provided by Ellen Cameron. DNA concentration for TLW43 was provided by Ellen Cameron; DNA concentrations for remaining samples were provided by University of Toronto's CAGEF.

Sample ID	Sampling Date	Sampling Depth	Weather Conditions	Volume Filtered (mL)	DNA Concentration
TLW43	July 18, 2018	8 m	23.4 ˚C 10% cloud coverage Light wind	800	11.7 ng/µL
TLW233	May 23, 2019	5.25	N/A	1000	1.6 µg/mL
TLW278	June 28, 2019	5.75	19.3 ˚C 95% cloud coverage No wind	1000	5.1 µg/mL
TLW323	July 24, 2019	5.25	20 ˚C 95% cloud coverage No wind	1000	3.7 µg/mL
TLW368	August 21, 2019	5	21.2 ˚C 10% cloud coverage Moderate wind	1000	4.9 µg/mL
TLW410	January 23, 2020	surface	N/A	1000	3.5 µg/mL

Appendix C

Turkey Lakes Watershed Metagenomic NCBI Blast Reference Sequences

Table C1. Blasted Reference Sequence Results for *cyn*S. Corresponding reference sequences from NCBI Genbank database for the *cyn*S gene – annotated function encodes cyanase/cyanate hydratase.

Scientific Name	Accession #	Protein ID	Reference
Leptolyngbya sp. NIES-3755	AP017308.1	BAU14591.1	Hirose et al., 2016 10.1128/genomeA.00090-16
Synechococcus sp. A18-25c	CP047957.1	QNJ21301.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. PROS-U-1	CP047951.1	QNJ04711.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-60	CP047933.1	QNI49694.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. Minos11	CP047953.1	QNJ10109.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. LA31	CP075523.1	QVV69196.1	Kling et al., 2020 10.1101/2020.05.27.119842
Synechococcus sp. WH 8101	CP047932.1	QNI46663.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. NOUM97013	CP047941.1	QNI74978.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. UTEX 2973	CP006471.1	AJD57390.1	Yu et al., 2015 10.1038/srep08132
Synechococcus sp. BIOS-U3-1	CP047936.1	QNI60283.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus elongatus PCC 6301	AP008231.1	BAD80179.1	Sugita et al., 2007 10.1007/s11120-006-9122-4
Synechococcus elongatus PCC 7942 = FACHB-805	AB000100.1	BAA19515.1	Harano et al., 199710.1128/jb.179.18.5744-5750.1997
Synechococcus elongatus PCC 7942 = FACHB-805	U59481.1	AAB02940.1	Espie et al., 2006 10.1128/JB.01328-06

Synechococcus sp. CB0101	CP039373.1	QCH16230.1	Marsan et al., 2014 10.1128/genomeA.01111-13
Synechococcus sp. MIT S9220	CP047958.1	QNJ24324.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechocystis sp. PCC 6714	CP007542.1	AIE74569.1	Kopf et al., 2014 10.1128/genomeA.00757-14
Synechococcus sp. MEDNS5	CP047952.1	QNJ07451.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. PROS-7-1	CP047945.1	QNI86633.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Pseudanabaena sp. ABRG5-3	AP017565.1	BBC27153.1	Tajima et al., 2018 10.1128/genomeA.01608-17
Synechococcus sp. CBW1107	CP064908.1	QPN57801.1	Fucich et al., 2021 10.1128/MRA.01399-20

Table C2. Blasted Reference Sequence Results for *lpx*A. Corresponding reference sequences from NCBI Genbank database for the *lpx*A gene – annotated function encodes an acyl-ADP-UDP-N-acetylglucosamine acyltransferase.

Scientific Name	**Accession**	**Protein ID**	**Reference**
Synechococcus sp. SynAce01	CP018091.1	APD48512.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. LA31	CP075523.1	QVV69254.1	Kling et al., 2022 10.1128/mra.00775-21
Cyanobium sp. NS01	CP047940.1	QNI69901.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1107	CP064908.1	QPN55966.1	Fucich et al., 2021b 10.1128/MRA.01399-20
Cyanobium sp. NIES-981	LT578417.1	SBO44476.1	Yamaguchi et al., 2016 10.1128/genomeA.00736-16
Synechococcus sp. CBW1006	CP060396.1	QPN68241.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1002	CP060398.1	QPN61405.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1004	CP060397.1	QPN64967.1	Fucich et al., 2021a 10.1093/gbe/evab009

Synechococcus sp. CB0101	CP039373.1	QCH16205.1	Marsan et al., 2014 10.1128/genomeA.01111-13

Table C3. Blasted Reference Sequence Results for *lpx*B. Corresponding reference sequences from NCBI Genbank database for the *lpx*B gene – annotated function encodes lipid-A-disaccharide synthetase.

Scientific Name	Accession	Protein ID	Reference
Synechococcus sp. RS9907	CP047944.1	QNI83206.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-60	CP047933.1	QNI47538.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A18-25c	CP047957.1	QNJ19163.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1108	CP060395.1	QPN70890.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. A15-28	CP047931.1	QNI41610.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. PROS-7-1	CP047945.1	QNI86001.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. BMK-MC-1	CP047939.1	QNI68550.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-44	CP047938.1	QNI65498.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. MEDNS5	CP047952.1	QNJ06849.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. PROS-U-1	CP047951.1	QNJ04147.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. SynAce01	CP018091.1	APD49498.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. SynAce01	CP018091.1	APD48511.1	Tang et al., 2019 10.3390/ijms20010152
Cyanobium sp. M30B3	CP073761.1	QVL53619.1	Chen et al., 202 10.1038/s42003-021-02510-6
Cyanobium sp. M30B3	CP073761.1	QVL52720.1	Chen et al., 2021 10.1038/s42003-021-02510-6
Synechococcus sp. A15-62	CP047950.1	QNJ01087.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9902	CP047949.1	QNI96658.1	Dore et al., 2020 10.3389/fmicb.2020.567431

Prochlorococcuus marinus str. MIT 9313	BX548175.1	CAE21584.1	Rocap et al., 2003 10.1038/nature01947
Prochlorococcuus marinus str. MIT 9303	CP000554.1	ABM77305.1	Kettler et al., 2007 10.1371/journal.pgen.0030231
Synechococcus sp. CB0101	CP039373.1	QCH16373.1	Marsan et al., 2014 10.1128/genomeA.01111-13
Synechococcus sp. CB0101	CP039373.1	QCH14611.1	Marsan et al., 2014 10.1128/genomeA.01111-13
Cyanobium sp. NS01	CP047940.1	QNI69902.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Cyanobium sp. NS01	CP047940.1	QNI71814.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-28	CP047931.1	QNI41610.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-24	CP047960.1	QNJ28302.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. BOUM118	CP047947.1	QNI90948.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A18-46.1	CP047955.1	QNJ13265.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9915	CP047934.1	QNI50339.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Parasynechococcus marenigrum WH 8102	BX569690.1	CAE07074.1	Palenik et al., 2003 10.1038/nature01943
Synechococcus sp. A18-40	CP047956.1	QNJ16136.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1004	CP060397.1	QPN64574.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. A15-127	CP047948.1	QNI95312.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. Minos11	CP047953.1	QNJ09414.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1107	CP064908.1	QPN55967.1	Fuchich et al., 2021 10.1128/MRA.01399-20
Synechococcus sp. WH 8101	CP035914.1	QBE69874.1	Marston and Polson, 2020 10.1128/MRA.01593-19
Synechococcus sp. WH 8101	CP047932.1	QNI46133.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9909	CP047943.1	QNI80237.1	Dore et al., 2020 10.3389/fmicb.2020.567431

Table C4. Blasted Reference Sequence Results for *lpx*C. Corresponding reference sequences from NCBI Genbank database for the *lpx*C gene – annotated function encodes UDP-3-O-acyl N-acetylglucosamine deacetylase.

Scientific Name	Accession	Protein ID	Reference
Synechococcus sp. CBW1107	CP064908.1	QPN55965.1	Fucich et al., 2021 10.1128/MRA.01399-20
Cyanobium sp. M30B3	CP073761.1	QVL53617.1	Chen et al., 2021 10.1038/s42003-021-02510-6

Table C5. Blasted Reference Sequence Results for *lpx*D. Corresponding reference sequences from NCBI Genbank database for the *lpx*D gene – annotated function encodes a UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acetyltransferase.

Scientific Name	Accession	Protein ID	Reference
Synechococcus sp. CBW1006	CP060396.1	QPN66490.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW1002	CP060398.1	QPN59669.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. SynAce01	CP018091.1	APD47265.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. CBW1004	CP060397.1	QPN63522.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. Minos11	CP047953.1	QNJ08536.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Cyanobium sp. NIES-981	LT578417.1	SBO41972.1	Yamaguchi et al., 2016 10.1128/genomeA.00736-16
Cyanobium sp. M30B3	CP073761.1	QVL54021.1	Chen et al., 2021 https://rdcu.be/cHeb7
Synechococcus sp. A15-62	CP047950.1	QNJ00812.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9907	CP047944.1	QNI82939.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-44	CP047938.1	QNI65059.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. TAK9802	CP047937.1	QNI61092.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. M16.1	CP047954.1	QNJ11883.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9902	CP047949.1	QNI96924.1	Dore et al., 2020 10.3389/fmicb.2020.567431

Cyanobium sp. NS01	CP047940.1	QNI70204.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9909	CP047943.1	QNI79560.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. WH 8101	CP047932.1	QNI45395.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1108	CP060395.1	QPN69742.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. A18-25c	CP047957.1	QNJ19893.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. A15-60	CP047933.1	QNI48260.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CB0101	CP039373.1	QCH16082.1	Marsan et al., 2014 10.1128/genomeA.01111-13
Synechococcus sp. LTW-R	CP059060.1	QNG30370.1	Xia et al., 2017 10.3389/fmicb.2017.01254
Synechococcus sp. A15-24	CP047960.1	QNJ28561.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1107	CP064908.1	QPN56934.1	Fucich et al., 2021 10.1093/gbe/evab009

Table C6. Blasted Reference Sequence Results for *lpx*K. Corresponding reference sequences from NCBI Genbank database for the *lpx*K gene – annotated function encodes a tetraacyldisaccharide 4'-kinase sugar synthetase.

Scientific Name	**Accession**	**Protein ID**	**Reference**
Synechococcus sp. CBW1002	CP060398.1	QPN61267.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW1107	CP064908.1	QPN58291.1	Fucich et al., 2021 10.1128/MRA.01399-20
Synechococcus sp. CBW1004	CP060397.1	QPN62955.1	Fucich et al., 2021 10.1093/gbe/evab009
Cyanobium sp. M30B3	CP073761.1	QVL52488.1	Chen et al., 2021 https://doi.org/10.1038/s42003-021-02510-6
Synechococcus sp. CBW1108	CP060395.1	QPN69054.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. LTW-R	CP059060.1	QNG28752.1	Xia et al., 2017 https://doi.org/10.3389/fmicb.2017.01254

Cyanobium sp. NIES-981	LT578417.1	SBO43519.1	Yamaguchi et al., 2016 https://doi.org/10.1128/genomeA.00736-16

Table C7. Blasted Reference Sequence Results for *mcy*C. Corresponding reference sequences from NCBI Genbank database for the *mcy*C gene – annotated function encodes a nonribosomal peptide synthetase (NRPS).

Scientific Name	Accession #	Protein ID	Reference
Dolichospermum heterosporum TAC447	CP099464.1	UUO14007.1	Phan et al., 2022 10.1021/acs.jnatprod.2c00382

Table C8. Blasted Reference Sequence Results for *mcy*H. Corresponding reference sequences from NCBI Genbank database for the *mcy*H gene – annotated function encodes an ABC transporter ATP-binding protein/permease.

Scientific Name	Accession #	Protein ID	Reference
Synechococcus sp. SynAce01	CP018091.1	APD48431.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. CB0101	CP039373.1	QCH15884.1	Fucich et al., 2019 10.1128/MRA.00484-19
Cyanobium sp. NS01	CP047940.1	QNI69675.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. LTW-R	CP059060.1	QNG28599.1	Xia et al., 2015 10.1128/AEM.01895-15
Synechococcus sp. CBW1108	CP060395.1	QPN70897.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1004	CP060397.1	QPN62736.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1107	CP064908.1	QPN55823.1	Fucich et al., 2021b 10.1128/MRA.01399-20
Cyanobium sp. M30B3	CP073761.1	QVL53321.1	Chen et al., 2021 10.1038/s42003-021-02510-6
Cyanobium sp. NIES-981	LT578417.1	SBO44686.1	Yamaguchi et al., 2016 10.1128/genomeA.00736-16

Table C9. Blasted Reference Sequence Results for *nif*J. Corresponding reference sequences from NCBI Genbank database for the *nif*J gene – annotated function encodes pyruvate-ferredoxin (flavodoxin) oxidoreductase.

Scientific Name	Accession #	Protein ID	Reference
Synechococcus sp. CBW1006	CP060396.1	QPN65438.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1002	CP060398.1	QPN58703.1	Fucich et al., 2021a 10.1093/gbe/evab009
Cyanobium sp. NIES-981	LT578417.1	SBO43385.1	Yamaguchi et al., 2016 10.1128/genomeA.00736-16
Cyanobium sp. M30B3	CP073761.1	QVL52315.1	Chen et al., 2021 10.1038/s42003-021-02510-6
Cyanobium sp. NS01	CP047940.1	QNI71537.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CBW1107	CP064908.1	QPN55764.1	Fucich et al., 2021b 10.1128/MRA.01399-20
Synechococcus sp. WH 8101	CP047932.1	QNI45953.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. SynAce01	CP018091.1	APD48841.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. CBW1108	CP060395.1	QPN69176.1	Fucich et al., 2021a 10.1093/gbe/evab009
Synechococcus sp. CBW1004	CP060397.1	QPN63461.1	Fucich et al., 2021a 10.1093/gbe/evab009

Table C10. Blasted Reference Sequence Results for *nif*K. Corresponding reference sequences from NCBI Genbank database for the *nif*K gene – annotated function encodes nitrogenase molybdenum-iron (Mo-Fe) protein subunit beta.

Scientific Name	Accession #	Protein ID	Reference
Cyanobium sp. M30B3	CP073761.1	QVL52629.1	Chen et al., 2021 10.1038/s42003-021-02510-6

Table C11. Blasted Reference Sequence Results for *nif*S. Corresponding reference sequences from NCBI Genbank database for the *nif*S gene – annotated function encodes cysteine desulfurase.

Scientific Name	Accession #	Protein ID	Reference
Cyanobium sp. NIES-981	LT578417.1	SBO42241.1	Yamaguchi et al., 2016 10.1128/genomeA.00736-16
Synechococcus sp. CBW1006	CP060396.1	QPN67878.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW1006	CP060396.1	QPN66365.1	Fucich et al., 2021 10.1093/gbe/evab009
Cyanobium sp. M30B3	CP073761.1	QVL54221.1	Chen et al., 2021 10.1038/s42003-021-02510-6

Table C12. Blasted Reference Sequence Results for *nif*U. Corresponding reference sequences from NCBI Genbank database for the *nif*U gene – annotated function encodes Fe-S cluster assembly protein.

Scientific Name	Accession #	Protein ID	Reference
Synechococcus sp. A18-40	CP047956.1	QNJ17409.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. A18-46.1	CP047955.1	QNJ14594.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. BOUM118	CP047947.1	QNI92190.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. RS9915	CP047934.1	QNI51674.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Parasynechococcus marenigrum WH 8102	BX569693.1	CAE08202.1	Palenik et al., 2003 10.1038/nature01943
Cyanobium sp. M30B3	CP073761.1	QVL52354.1	Chen et al., 2021 10.1038/s42003-021-02510-6
Synechococcus sp. LTW-R	CP059060.1	QNG28543.1	Xia et al., 2017 doi.org/10.3389/fmicb.2017.01254
Synechococcus sp. CB0101	CP039373.1	QCH16370.1	Fucich et al., 2019 doi.org/10.1128/MRA.00484-19

Synechococcus sp. CBW1004	CP060397.1	QPN64735.1	Fucich et al., 2021 doi.org/10.1093/gbe/evab009
Synechococcus sp. CBW1108	CP060395.1	QPN69256.1	Fucich et al., 2021 doi.org/10.1093/gbe/evab009
Synechococcus sp. LA31	CP075523.1	QVV69111.1	Kling et al., 2022 https://doi.org/10.1128/mra.00775-21
Cyanobium sp. NIES-981	LT578417.1	SBO43210.1	Yamaguchi et al., 2016 doi.org/10.1128/genomeA.00736-16
Synechococcus sp. HK01-R	CP059059.1	QNG27119.1	Xia et al., 2017 https://doi.org/10.3389/fmicb.2017.01254
Synechococcus sp. BMK-MC-1	CP047939.1	QNI66821.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. CBW1107	CP064908.1	QPN58310.1	Fucich et al., 2021 doi.org/10.1128/MRA.01399-20
Synechococcus sp. Minos11	CP047953.1	QNJ08072.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Cyanobium sp. NS01	CP047940.1	QNI71438.1	Dore et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Thermoleptolyngbya sichuanensis PKUAC-SCTA183	CP053661.1	QKD82748.1	Tang et al., 2021 10.3389/fmicb.2021.696102
Synechococcus sp. SynAce01	CP018091.1	APD49283.1	Tang et al., 2021 10.3389/fmicb.2021.696102
Synechococcus sp. CBW1006	CP060396.1	QPN68506.1	Fucich et al., 2021 doi.org/10.1128/MRA.01399-20
Synechococcus sp. CBW1002	CP060398.1	QPN61668.1	Fucich et al., 2021 doi.org/10.1093/gbe/evab009

Table C13. Blasted Reference Sequence Results for *ntc*A. Corresponding reference sequences from NCBI Genbank database for the *ntc*A gene – annotated function encodes a global transcriptional nitrogen regulator.

Scientific Name	Accession	Protein ID	Reference
Synechococcus sp. SynAce01	CP018091.1	APD48201.1	Tang et al., 2019 10.3390/ijms20010152
Synechococcus sp. HK01-R	CP059059.1	QNG28007.1	Xia et al., 2017 https://doi.org/10.3389/fmicb.2017.01254
Synechococcus sp. MIT S9220	CP047958.1	QNJ21664.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. CC9311	CP000435.1	ABI46137.1	Palenik et al., 2006 10.1073/pnas.0602963103
Synechococcus sp. CBW1004	CP060397.1	QPN62751.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. WH 8101	CP035914.1	QBE67962.1	Marston and Polson, 2020 10.1128/MRA.01593-19
Synechococcus sp. WH 8101	CP047932.1	QNI44167.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. LA31	CP075523.1	QVV66867.1	Kling et al., 2022 10.1128/mra.00775-21
Synechococcus sp. CBW1006	CP060396.1	QPN67199.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW1002	CP060398.1	QPN61146.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. LTW-R	CP059060.1	QNG28872.1	Xia et al., 2017 https://doi.org/10.3389/fmicb.2017.01254
Synechococcus sp. RS9902	CP047949.1	QNI96196.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. TAK9802	CP047937.1	QNI60579.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9907	CP047944.1	QNI81132.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. RS9909	CP047943.1	QNI78424.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. SYN20	CP047959.1	QNJ24750.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Cyanobium sp. M30B3	CP073761.1	QVL52663.1	Chen et al., 2021 10.1038/s42003-021-02510-6

Cyanobium sp. NS01	CP047940.1	QNI71768.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Cyanobium sp. NIES-981	LT578417.1	SBO43662.1	Yamaguchi et al., 2016 https://doi.org/10.1128/genomeA.00736-16
Synechococcus sp. CBW1108	CP060395.1	QPN68907.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW1107	CP064908.1	QPN57653.1	Fucich et al., 2019 doi.org/10.1128/MRA.00484-19
Synechococcus sp. A18-46.1	CP047955.1	QNJ12955.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Parasynechococcus marenigrum WH 8102	BX569689.1	CAE06790.1	Palenik et al., 2006 10.1073/pnas.0602963103
Synechococcus sp. A15-62	CP047950.1	QNI99158.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. ROS8604	CP047946.1	QNI87175.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. MVIR-18-1	CP047942.1	QNI75376.1	Dore et al., 2020 10.3389/fmicb.2020.567431
Synechococcus sp. PROS-9-1	CP047961.1	QNJ30758.1	Dore et al., 2020 10.3389/fmicb.2020.567431

Appendix D

Turkey Lakes Watershed Metagenomic NCBI Taxonomy Reference Sequences

Table D1. Taxonomic Reference Sequence Results for *cynS*.

Whole Genome Organism	Accession	Protein ID	Reference
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061430355.1	Phan et al., 2021 10.1021/jacs.1c05732
Planktothrix agardhii NIES-204	AP017991	BBD56279.1	Shimura et al., 2020 10.1016/j.hal.2020.101942
Microcystis aeruginosa PCC 7806SL	CP020771	ARI83859.1	Zhao et al., 2018 10.1016/j.hal.2018.04.005
Cyanobium sp. NS01	NZ_CP047940	WP_186700865.1	Doré et al., 2020 10.3389/fmicb.2020.567431
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_115539406.1	Martin et al., 2019 0.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011244299.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017289966.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_026148836.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67904.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039725296.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15

Microcystis aeruginosa NIES-87	NZ_BFAC01000002	WP_104395869.1	Yamaguchi et al., 2018 10.1128/genomeA.01596-17
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011244299.1	Yang et al., 2015 10.1073/pnas.1812871115
Fischerella sp. NIES-3754	NZ_AP017305	WP_231936156.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc piscinale CENA21	NZ_CP012036	WP_062290431.1	Leão et al., 2016 10.1128/genomeA.00189-16
Acaryochloris marine MBIC11017	NC_009925	WP_012165389.1	Swingley et al., 2008 10.1073/pnas.0709772105
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_017296379.1	Shih et al., 2013 10.1073/pnas.1217107110
Halomicronema hongdechloris C2206	NZ_CP021983	WP_080812301.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066343191.1	Hirose et al., 2015 10.1128/genomeA.00357-15

Table D2. Taxonomic Reference Sequence Results for *lpx*A

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_002797088.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_107806154.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_072149089.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011377794.1	Yang et al., 2018 10.1073/pnas.1812871115
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011244044.1	Yang et al., 2018 10.1073/pnas.1812871115

Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290208.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI66582.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200989464.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197151339.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011377794.1	Yang et al., 2015 10.1073/pnas.1812871115
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011244044.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035156181.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtrpica ATCC 51142	NC_010546	WP_009545641.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062248574.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013192295.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181931570.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062294312.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012409453.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacterium thalassa isolate ALOHA	NC_013771	WP_041487706.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290208.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Acaryochloris marine MBIC11017	NC_009925	WP_012161232.1	Swingley et al., 2008 10.1073/pnas.0709772105
Gloebacter kilaueensis JS1	NC_022600	WP_041243904.1	Saw et al., 2013 10.1371/journal.pone.0076376

Gloebacter violaceus PCC 7421	NC_005125	WP_197530156.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Halomicronema hongdechloris C2206	NZ_CP021983	WP_088431601.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Halomicronema hongdechloris C2206	NZ_CP021983	WP_088430858.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Phormidium sp. ETS-05	NZ_CP051168	WP_199247618.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_164921189.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Thermosynechococcus vestitus BP-1	NC_004113	WP_164921002.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066348401.1	Hirose et al., 2015 10.1128/genomeA.00357-15
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119259950.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D3. Taxonomic Reference Sequence Results for *lpx*B

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061431751.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_107806153.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_115538782.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_196794065.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_036046524.1	Hiraide et al., 2015 10.1093/pcp/pcu165

Prochlorococcuus marinus MIT9313	BX548175	CAE21584.1	Rocap et al., 2003 10.1038/nature01947
Nostoc sp. HK-01	AP018318	BBD61175.1	Hirose et al., 2021 10.1093/dnares/dsab024
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI68403.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200989463.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197151338.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197147861.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_196794065.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035159394.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009545642.1	Welsh et al., 2008 10.1073/pnas.0805418105
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009545643.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062248576.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013192294.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181931571.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062294311.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012409454.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacterium thalassa isolate ALOHA	NC_013771	WP_012953540.1	Tripp et al., 2010 10.1038/nature08786

Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_036046524.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Acaryochloris *marine* MBIC11017	NC_009925	WP_012161231.1	Swingley et al., 2008 10.1073/pnas.0709772105
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_026102039.1	Shih et al., 2013 10.1073/pnas.1217107110
Gloebacter kilaueensis JS1	NC_022600	WP_023173749.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter violaceus PCC 7421	NC_005125	WP_011141866.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Halomicronema hongdechloris C2206	NZ_CP021983	WP_080811986.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Phormidium sp. ETS-05	NZ_CP051168	WP_199247619.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_011056176.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066345387.1	Hirose et al., 2015 10.1128/genomeA.00357-15
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119259949.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D4. Taxonomic Reference Sequence Results for *lpx*C

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061433096.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_006196760.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_006276945.1	Martin et al., 2019 10.1111/1462-2920.14425

Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290210.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Prochlorococcuus marinus MIT9313	BX548175	CAE21587.1	Rocap et al., 2003 10.1038/nature01947
Nostoc sp. HK-01	AP018318	BBD61172.1	Hirose et al., 2021 10.1093/dnares/dsab024
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI66580.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200989466.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197151344.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126146674.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035156186.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Fischerella sp. NIES-3754	NZ_AP017305	WP_062248571.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013192297.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181931568.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062294316.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012409451.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacterium thalassa isolate ALOHA	NC_013771	WP_012953774.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290210.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Acaryochloris marine MBIC11017	NC_009925	WP_071819929.1	Swingley et al., 2008 10.1073/pnas.0709772105
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_017293706.1	Shih et al., 2013

			10.1073/pnas.1217107110
Gloebacter kilaueensis JS1	NC_022600	WP_051382789.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter violaceus PCC 7421	NC_005125	WP_011141863.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Halomicronema hongdechloris C2206	NZ_CP021983	WP_080811989.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Phormidium sp. ETS-05	NZ_CP051168	WP_233258433.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_011057627.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066344044.1	Hirose et al., 2015 10.1128/genomeA.00357-15
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119260709.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D5. Taxonomic Reference Sequence Results for *lpx*D

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061431306.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_006198165.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_040007867.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126148012.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290318.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Prochlorococcuus marinus MIT9313	BX548175	CAE20705.1	Rocap et al., 2003 10.1038/nature01947

Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI66422.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039726601.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039726602.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200988052.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197149093.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126148012.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035156701.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009546956.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062246326.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013191297.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181928682.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062290101.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012410761.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacterium thalassa isolate ALOHA	NC_013771	WP_012953811.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290318.1	Hiraide et al., 2015 10.1093/pcp/pcu165

Acaryochloris marine MBIC11017	NC_009925	WP_083776700.1	Swingley et al., 2008 10.1073/pnas.0709772105
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_017293155.1	Shih et al., 2013 10.1073/pnas.1217107110
Gloebacter kilaueensis JS1	NC_022600	WP_023171915.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter kilaueensis JS1	NC_022600	WP_023173744.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter violaceus PCC 7421	NC_005125	WP_011141862.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Gloebacter violaceus PCC 7421	NC_005125	WP_011142009.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Gloebacter violaceus PCC 7421	NC_005125	WP_011142707.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Halomicronema hongdechloris C2206	NZ_CP021983	WP_080807088.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Phormidium sp. ETS-05	NZ_CP051168	WP_199249920.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_011055887.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066347808.1	Hirose et al., 2015 10.1128/genomeA.00357-15
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119259741.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D6. Taxonomic Reference Sequence Results for *lpx*K

Whole Genome Organism	Accession	Protein ID	Reference
Cyanobium sp. NS01	NZ_CP047940	WP_186697414.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197147860.1	Fucich et al., 2021 10.1093/gbe/evab009
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_017295523.1	Shih et al., 2013 10.1073/pnas.1217107110
Synechococcus sp. GFB01 Contig 36	NZ_LFEK01000036	WP_048017329.1	Guimarães et al., 2015 10.1128/genomeA.00876-15
Nostoc sp. CHAB 5834	NZ_JAIVFU010000260	WP_229475576.1	Huo et al., 2021 10.3390/microorganisms9102015

Table D7. Taxonomic Reference Sequence Results for *mcy*A

Whole Genome Organism	Accession	Protein ID	Reference
Anabaena sp. 90	AY212249	AAO62586.1	Rouhiainen et al., 2004 10.1128/AEM.70.2.686-692.2004
Nostoc sp. 152	KC699835	AGZ05275.1	Fewer et al., 2013 10.1016/j.chembiol.2013.04.020
Microcystis aeruginosa PCC 7806SL	CP020771	ARI82521.1	Zhao et al., 2018 10.1016/j.hal.2018.04.005
Anabaena sp. 90 chromosome chANA01	CP003284	AFW93775.1	Wang et al., 2012 10.1186/1471-2164-13-613
Planktothrix rubescens NIVA-CYA98	AM990462	CAQ48245.1	Rounge et al., 2009 10.1186/1471-2164-10-396
Phormidium sp. LP904c	MK870090	QCQ67879.1	Shishido et al., 2019 10.3390/md17050271
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67167.1	Christiansen et al., 2014 10.1128/AEM.01188-14

Table D8. Taxonomic Reference Sequence Results for *mcy*C

Whole Genome Organism	Accession	Protein ID	Reference
Anabaena sp. 90	AY212249	AAO62588.1	Rouhiainen et al., 2004 10.1128/AEM.70.2.686-692.2004
Nostoc sp. 152	KC699835	AGZ05277.1	Fewer et al., 2013 10.1016/j.chembiol.2013.04.020
Microcystis aeruginosa PCC 7806SL	CP020771	ARI82523.1	Zhao et al., 2018 10.1016/j.hal.2018.04.005
Anabaena sp. 90 chromosome chANA01	CP003284	AFW93777.1	Wang et al., 2012 10.1186/1471-2164-13-613
Planktothrix rubescens NIVA-CYA98	AM990462	CAQ48247.1	Rounge et al., 2009 10.1186/1471-2164-10-396
Phormidium sp. LP904c	MK870090	QCQ67881.1	Shishido et al., 2019 10.3390/md17050271
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67165.1	Christiansen et al., 2014 10.1128/AEM.01188-14

Table D9. Taxonomic Reference Sequence Results for *mcy*H

Whole Genome Organism	Accession	Protein ID	Reference
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061432154.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Planktothrix agardhii NIES-204	AP017991	BBD53407.1	Shimura et al., 2020 10.1016/j.hal.2020.101942
Anabaena sp. 90	AY212249	AAO62579.1	Rouhiainen et al., 2004 10.1128/AEM.70.2.686-692.2004
Anabaena sp. 90	AJ536156	CAD60093.1	Rouhiainen et al., 2004 10.1128/AEM.70.2.686-692.2004
Nostoc sp. 152	KC699835	AGZ05268.1	Fewer et al., 2013 10.1016/j.chembiol.2013.04.020

Microcystis aeruginosa PCC 7806SL	CP020771	ARI82516.1	Zhao et al., 2018 10.1016/j.hal.2018.04.005
Microcystis aeruginosa PCC 7806SL	NZ_CP020771	WP_002743209.1	Zhao et al., 2018 10.1016/j.hal.2018.04.005
Cyanobium sp. NS01	NZ_CP047940	WP_186698580.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. RS9909	CP047943	QNI80400.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. A15-28	CP047931	QNI41433.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. A18-40	CP047956	QNJ15897.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. BIOS-E4-1	CP047935	QNI56408.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_181884057.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126146833.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017288477.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Prochlorococcuus marinus MIT9313	BX548175	CAE21770.1	Rocap et al., 2003 10.1038/nature01947
Anabaena sp. 90 chromosome chANA01	CP003284	AFW93768.1	Wang et al., 2012 10.1186/1471-2164-13-613
Planktothrix rubescens NIVA-CYA98	AM990462	CAQ48244.1	Rounge et al., 2009 10.1186/1471-2164-10-396
Phormidium sp. LP904c	MK870090	QCQ67878.1	Shishido et al., 2019 10.3390/md17050271
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67168.1	Christiansen et al., 2014 10.1128/AEM.01188-14

Microcystis aeruginosa NIES-87	NZ_BFAC01000002	WP_104395676.1	Yamaguchi et al., 2018 10.1128/genomeA.01596-17

Table D10. Taxonomic Reference Sequence Results for *nif*J

Whole Genome Organism	Accession	Protein ID	Reference
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_061431802.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Aphanizomenon flos-aquae DEX188	CP051188	QSV67358.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Aphanizomenon flos-aquae DEX188	CP051188	QSV67499.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_107806858.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Planktothrix agardhii NIES-204	AP017991	BBD56354.1	Shimura et al., 2020 10.1016/j.hal.2020.101942
Cyanobium sp. NS01	NZ_CP047940	WP_186697192.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_057177031.1	Martin et al., 2019 10.1111/1462-2920.14425
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_115538487.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126147448.1	Yang et al., 2018 10.1073/pnas.1812871115
Anabaena sp. 90 chromosome chANA01	CP003284	AFW94704.1	Wang et al., 2012 10.1186/1471-2164-13-613
Anabaena sp. 90 chromosome chANA01	CP003284	AFW96422.1	Wang et al., 2012 10.1186/1471-2164-13-613
Nostoc sp. PCC 7120	L14925	AAC36818.1	Bauer et al., 1993 10.1073/pnas.90.19.8812

Dolichospermum planctonicum NIES-80	BJCF01000043	GCL43539.1	Suzuki et al., 2019 10.1128/MRA.00605-19
Nostoc sp. HK-01	AP018318	BBD57499.1	Hirose et al., 2021 10.1093/dnares/dsab024
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67827.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_084607113.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200989242.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197148152.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_126147448.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035158219.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009547149.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062246773.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181930306.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_235526475.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc piscinale CENA21	NZ_CP012036	WP_062290409.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_041565536.1	Ekman et al., 2013 10.1104/pp.112.213116
Phormidium sp. ETS-05	NZ_CP051168	WP_199247173.1	Gris et al., 2020 10.3390/microorganisms8101590

Table D11. Taxonomic Reference Sequence Results for *nif*K

Whole Genome Organism	Accession	Protein ID	Reference
Aphanizomenon flos-aquae DEX188	CP051188	QSV67173.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_006194550.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_006278576.1	Martin et al., 2019 10.1111/1462-2920.14425
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017289031.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Anabaena sp. 90 chromosome chANA01	CP003284	AFW96195.1	Wang et al., 2012 10.1186/1471-2164-13-613
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_044151256.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Calothrix sp. 336/3	NZ_CP011382	WP_035157993.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009546641.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_009459214.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013190626.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181928152.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062292547.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012407198.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacteriu	NC_013771	WP_012954004.1	Tripp et al., 2010 10.1038/nature08786

m thalassa isolate ALOHA			
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017289031.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119260547.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D12. Taxonomic Reference Sequence Results for *nif*S

Whole Genome Organism	Accession	Protein ID	Reference
Aphanizomenon flos-aquae DEX188	CP051188	QSV67142.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_006194563.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_040009945.1	Martin et al., 2019 10.1111/1462-2920.14425
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_026148746.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Prochlorococcuus marinus MIT9313	BX548175	CAE20892.1	Rocap et al., 2003 10.1038/nature01947
Anabaena sp. 90 chromosome chANA01	CP003284	AFW94149.1	Wang et al., 2012 10.1186/1471-2164-13-613
Anabaena sp. 90 chromosome chANA01	CP003284	AFW96083.1	Wang et al., 2012 10.1186/1471-2164-13-613
Nostoc sp. HK-01	AP018318	BBD60695.1	Hirose et al., 2021 10.1093/dnares/dsab024
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI68097.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039727382.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200987978.1	Delbaje et al., 2021 10.1099/ijsem.0.004648

Calothrix sp. 336/3	NZ_CP011382	WP_035158000.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009546637.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062246906.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013190630.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181928122.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062292559.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012407223.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus *Atelocyanobacterium thalassa* isolate ALOHA	NC_013771	WP_012954000.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_026148746.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119260543.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D13. Taxonomic Reference Sequence Results for *nif*U

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_002735229.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Aphanizomenon flos-aquae DEX188	CP051188	QSV66845.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Aphanizomenon flos-aquae DEX188	CP051188	QSV67143.1	Dreher et al., 2021

			https://doi.org/10.1016/j.hal.2021.102005
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_107807030.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Planktothrix agardhii NIES-204	AP017991	BBD56686.1	Shimura et al., 2020 10.1016/j.hal.2020.101942
Cyanobium sp. NS01	NZ_CP047940	WP_186697118.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. RS9909	CP047943	QNI78695.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. A15-28	CP047931	QNI42845.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. A18-40	CP047956	QNJ17409.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Synechococcus sp. BIOS-E4-1	CP047935	QNI53286.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_006278579.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011243379.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017289028.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Anabaena sp. 90 chromosome chANA01	CP003284	AFW96084.1	Wang et al., 2012 10.1186/1471-2164-13-613
Nostoc sp. HK-01	AP018318	BBD60694.1	Hirose et al., 2021 10.1093/dnares/dsab024
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI67540.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039727381.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15

Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200987979.1	Delbaje et al., 2021 10.1099/ijsem.0.004648
Synechococcus sp. CBW 1108	NZ_CP060395	WP_197148226.1	Fucich et al., 2021 10.1093/gbe/evab009
Calothrix sp. 336/3	NZ_CP011382	WP_035157999.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_009546638.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_062246908.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013190629.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_181928123.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_062292556.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_012407222.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus Atelocyanobacterium thalassa isolate ALOHA	NC_013771	WP_012954001.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017289028.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Gloebacter kilaueensis JS1	NC_022600	WP_023171319.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter violaceus PCC 7421	NC_005125	WP_011142871.1	Nakamura et al., 2003 10.1093/dnares/10.4.181
Phormidium sp. ETS-05	NZ_CP051168	WP_199248242.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_197524159.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066345667.1	Hirose et al., 2015 10.1128/genomeA.00357-15

Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119260544.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Table D14. Taxonomic Reference Sequence Results for *ntc*A

Whole Genome Organism	**Accession**	**Protein ID**	**Reference**
Microcystis aeruginosa NIES-88	NZ_AP024565	WP_002793163.1	Phan et al., 2021 https://doi.org/10.1021/jacs.1c05732
Aphanizomenon flos-aquae DEX188	CP051188	QSV66046.1	Dreher et al., 2021 https://doi.org/10.1016/j.hal.2021.102005
Nodularia spumigena UHCC 0039	NZ_CP020114	WP_006198916.1	Teikari et al., 2018 https://doi.org/10.3389/fmicb.2018.00356
Cyanobium sp. NS01	NZ_CP047940	WP_186697377.1	Doré et al., 2020 https://doi.org/10.3389/fmicb.2020.567431
Cylindrospermopsis raciborskii Cr2010	NZ_CP065936	WP_040008442.1	Martin et al., 2019 10.1111/1462-2920.14425
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011243689.1	Yang et al., 2018 10.1073/pnas.1812871115
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290537.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Prochlorococcuus marinus MIT9313	BX548175	CAE22006.1	Rocap et al., 2003 10.1038/nature01947
Anabaena sp. 90 chromosome chANA01	CP003284	AFW94915.1	Wang et al., 2012 10.1186/1471-2164-13-613
Planktothrix rubescens NIVA-CYA 126/8	CM002803	KEI66820.1	Christiansen et al., 2014 10.1128/AEM.01188-14
Lyngbya confervoides BDU141951	NZ_JTHE02000003	WP_039727480.1	Chandrababunaidu et al., 2015 10.1128/genomeA.00066-15
Anabaenopsis elenkinii CCIBt3563	NZ_CP063311	WP_200988347.1	Delbaje et al., 2021 10.1099/ijsem.0.004648

Synechococcus sp. CBW 1108	NZ_CP060395	WP_197147893.1	Fucich et al., 2021 10.1093/gbe/evab009
Synechococcus elongatus UTEX 3055	NZ_CP033061	WP_011243689.1	Yang et al., 2015 10.1073/pnas.1812871115
Calothrix sp. 336/3	NZ_CP011382	WP_035157823.1	Isojärvi et al., 2015 10.1128/genomeA.01474-14
Crocosphaera subtropica ATCC 51142	NC_010546	WP_008277872.1	Welsh et al., 2008 10.1073/pnas.0805418105
Fischerella sp. NIES-3754	NZ_AP017305	WP_009459526.1	Hirose et al., 2015 10.1016/j.jbiotec.2016.01.011
Nostoc azollae 0708	NC_014248	WP_013190440.1	Ran et al., 2010 10.1371/journal.pone.0011486
Nostoc edaphicum CCNP1411	NZ_CP054698	WP_010998529.1	Fidor et al., 2020 10.3390/md18090442
Nostoc piscinale CENA21	NZ_CP012036	WP_015113299.1	Leão et al., 2016 10.1128/genomeA.00189-16
Nostoc punctiforme PCC 73102	NC_010628	WP_010998529.1	Ekman et al., 2013 10.1104/pp.112.213116
Candidatus *Atelocyanobacterium thalassa* isolate ALOHA	NC_013771	WP_236608151.1	Tripp et al., 2010 10.1038/nature08786
Leptolyngbya boryana IAM M-101	NZ_AP014638	WP_017290537.1	Hiraide et al., 2015 10.1093/pcp/pcu165
Acaryochloris marine MBIC11017	NC_009925	WP_012163575.1	Swingley et al., 2008 10.1073/pnas.0709772105
Geminocystis herdmanii PCC 6308	NZ_CM001775	WP_017293890.1	Shih et al., 2013 10.1073/pnas.1217107110
Gloebacter kilaueensis JS1	NC_022600	WP_023173632.1	Saw et al., 2013 10.1371/journal.pone.0076376
Gloebacter violaceus PCC 7421	NC_005125	WP_011143275.1	Nakamura et al., 2003 10.1093/dnares/10.4.181

Halomicronema hongdechloris C2206	NZ_CP021983	WP_080804858.1	Yaqiong et al., 2015 10.1016/j.bbabio.2015.10.009
Phormidium sp. ETS-05	NZ_CP051168	WP_199246430.1	Gris et al., 2020 10.3390/microorganisms8101590
Thermosynechococcus vestitus BP-1	NC_004113	WP_164920910.1	Nakamura et al., 2002 10.1093/dnares/9.4.123
Geminocystis sp. NIES-3708	NZ_AP014815	WP_066345608.1	Hirose et al., 2015 10.1128/genomeA.00357-15
Cyanobacterium endosymbiont of *Rhopaladia gibberula* isolate	NZ_AP018341	WP_119260215.1	Nakayama and Inagaki, 2017 10.1038/s41598-017-13578-8

Appendix E

Metagenomic Shotgun Sequencing Supplementary Information

Table E1. Normalization values for genes of interest. Values are log-transformed to account for sampling variability between and within samples and signify sampling depth and coverage for each gene sequence.

	TLW43	TLW233	TLW278	TLW323	TLW368	TLW410
***cyn*S**	-9.096	-8.288	-8.398	--	-7.837	--
***lpx*A**	-9.198	-8.342	-8.234	-7.717	-8.355	-8.663
***lpx*B**	-9.190	-7.910	-8.358	-7.834	-8.382	-8.683
***lpx*C**	-9.166	-8.296	-8.320	--	-8.386	-8.523
***lpx*D**	-9.197	-8.297	-8.398	--	-8.358	-8.612
***lpx*K**	-9.266	-8.086	-8.271	-7.818	-8.404	-8.658
***mcy*A**	--	--	--	-6.885	--	--
***mcy*C**	--	--	--	--	-8.486	--
***mcy*D**	-9.351	-8.339	--	--	-8.462	--
***mcy*E**	-9.338	--	-8.049	--	--	--
***mcy*H**	-9.053	--	-8.189	--	-8.362	--
***mcy*J**	-9.231	--	--	--	--	--
***nif*J**	-9.299	-8.287	-8.218	--	-8.501	-8.597
***nif*K**	-9.268	--	-8.385	--	--	--
***nif*S**	-9.228	-8.190	-8.321	-7.657	-8.565	-8.595
***nif*U**	-9.129	-8.389	-8.406	-7.833	-8.396	-8.689
***ntc*A**	-9.122	--	-8.174	--	-8.480	--

Normalization calculation (averaged per gene):

$$\sum \frac{(coverage/contig\ length)}{total\ \#\ assembled\ reads}$$

Table E2. Quality Control and Assembly Summary from Metagenome ATLAS workflow pipeline.

Sample	Assembled Reads	Predicted Genes	% Assembled Reads	contig_bp	L50	L90	N50	N90	max	gc_avg	gc_std	Assembled contigs
TLW43	3545185	193828	39.3343990	146757372	1947	1050	19592	62735	84424	0.57486	0.10153	82018
TLW233	492744	27260	6.85978907	19092615	1217	945	5411	12654	14842	0.44488	0.10315	15221
TLW278	641723	45853	11.9125855	29765964	1215	855	7554	19537	26597	0.49609	0.13153	23257
TLW323	173001	9601	7.39447862	6554774	1134	835	1867	4621	20632	0.46521	0.10207	5547
TLW368	801558	38057	12.3875292	25422331	1281	881	6224	16051	17457	0.52369	0.11381	19214
TLW410	1168717	46608	15.9750641	34613319	1666	1002	5776	16935	92276	0.52556	0.10742	20765

Table E3. Sequencing depth, coverage, and IDs for individual functional gene contigs extracted from metagenomic dataset

Sample	Gene ID	Taxa	Gene	Coverage	Contig Length
TLW43	Gene008235	Burkholderiaceae	*cyn*S	3.3666	1263
	Gene002746	*Cyanobium*	*cyn*S	5.9827	10855
	Gene065286	*Cyanobium*	*cyn*S	3.5201	2013
	Gene167293	*Cyanobium*	*cyn*S	3.014	1425
	Gene168300	*Cyanobium*	*cyn*S	2.8172	1384
	Gene178311	*Cyanobium*	*cyn*S	13.8489	1079
	Gene040726	*Cyanobium*	*cyn*S	4.3936	1034
	Gene091166	Oscillatoriales	*cyn*S	3.3218	634
	Gene114309	Caulobacterales	*lpx*A	4.8624	1802
	Gene143787	Comamonadaceae	*lpx*A	4.8279	6974
	Gene107554	Comamonadaceae	*lpx*A	6.0462	2511
	Gene017696	Comamonadaceae	*lpx*A	3.8635	2132
	Gene110675	Comamonadaceae	*lpx*A	4.6049	2126
	Gene114646	Comamonadaceae	*lpx*A	2.0288	1773
	Gene140581	*Cyanobium*	*lpx*A	21.2796	21301
	Gene048547	*Cyanobium*	*lpx*A	9.8871	12199
	Gene003457	*Cyanobium*	*lpx*A	25.3873	8677
	Gene052130	*Cyanobium*	*lpx*A	12.8812	5405
	Gene150334	*Cyanobium*	*lpx*A	6.3109	3194
	Gene151291	*Cyanobium*	*lpx*A	5.1626	2976
	Gene016880	Hydrogenophilales	*lpx*A	3.2571	2217
	Gene105880	Opitutae	*lpx*A	5.8795	2764
	Gene063481	Opitutae	*lpx*A	6.4487	2191
	Gene038693	Opitutae	*lpx*A	2.5716	1076
	Gene094071	Planctomycetes	*lpx*A	9.4047	20280
	Gene002839	Planctomycetes	*lpx*A	9.9428	10509
	Gene096404	Planctomycetes	*lpx*A	8.3332	8551
	Gene011530	Planctomycetes	*lpx*A	4.9586	3044
	Gene152899	Planctomycetes	*lpx*A	4.626	2695
	Gene107966	Planctomycetes	*lpx*A	3.2771	2458
	Gene073148	Planctomycetes	*lpx*A	12.1215	1473
	Gene119973	Planctomycetes	*lpx*A	3.7058	1458
	Gene122708	Planctomycetes	*lpx*A	2.5183	1341
	Gene085768	Planctomycetes	*lpx*A	7.4709	1066
	Gene044413	Planctomycetes	*lpx*A	5.8607	639
	Gene055167	Sphingobacteriia	*lpx*A	5.1356	3769
	Gene093366	*Synechococcus*	*lpx*A	48.7849	31264
	Gene182842	*Synechococcus*	*lpx*A	11.6619	772

Gene043920	*Synechococcus*	*lpx*A	11.2471	684
Gene137788	*Synechococcus*	*lpx*A	6.0274	621
Gene064322	unclassified	*lpx*A	4.3044	2109
Gene084902	Betaproteobacteria unclassified	*lpx*A	1.6581	1085
Gene024116	Betaproteobacteria unclassified	*lpx*A	2.5695	1612
Gene086456	Verrucomicrobia Verrucomicrobia	*lpx*A	1.5323	1052
Gene114233	Verrucomicrobiae	*lpx*A	2.6794	1809
Gene163623	Caulobacterales	*lpx*B	7.5603	1617
Gene143788	Comamonadaceae	*lpx*B	4.8279	6974
Gene107553	Comamonadaceae	*lpx*B	6.0462	2511
Gene015809	Comamonadaceae	*lpx*B	2.785	2353
Gene017697	Comamonadaceae	*lpx*B	3.8635	2132
Gene017698	Comamonadaceae	*lpx*B	3.8635	2132
Gene157282	Comamonadaceae	*lpx*B	4.6049	2126
Gene157283	Comamonadaceae	*lpx*B	4.6049	2126
Gene162693	Comamonadaceae	*lpx*B	4.5006	1678
Gene024271	Comamonadaceae	*lpx*B	2.1301	1599
Gene177222	Comamonadaceae	*lpx*B	2.933	1105
Gene046279	*Cyanobium*	*lpx*B	20.9834	64759
Gene140582	*Cyanobium*	*lpx*B	21.2796	21301
Gene094245	*Cyanobium*	*lpx*B	9.8544	19260
Gene048546	*Cyanobium*	*lpx*B	9.8871	12199
Gene144810	*Cyanobium*	*lpx*B	20.0041	5784
Gene052129	*Cyanobium*	*lpx*B	12.8812	5405
Gene123065	*Cyanobium*	*lpx*B	13.2647	1326
Gene182060	*Cyanobium*	*lpx*B	3	857
Gene182841	*Cyanobium*	*lpx*B	11.6619	772
Gene045685	*Cyanobium*	*lpx*B	7.0449	534
Gene177971	Cytophagia	*lpx*B	1.8511	1088
Gene064791	Opitutae	*lpx*B	5.6663	2056
Gene086342	Opitutae	*lpx*B	1.3302	1054
Gene096358	Planctomycetes	*lpx*B	9.0105	8755
Gene053093	Planctomycetes	*lpx*B	10.2382	4689
Gene057664	Planctomycetes	*lpx*B	5.6947	3072
Gene107828	Planctomycetes	*lpx*B	5.6087	2474
Gene062561	Planctomycetes	*lpx*B	15.5193	2305
Gene019102	Planctomycetes	*lpx*B	2.3691	2002

Gene120029	Planctomycetes	*lpx*B	4.3623	1452
Gene169748	Planctomycetes	*lpx*B	4.5	1328
Gene125781	Planctomycetes	*lpx*B	4.5511	1232
Gene082175	Planctomycetes	*lpx*B	2.6236	1153
Gene178980	Planctomycetes	*lpx*B	2.1745	1066
Gene089137	Planctomycetes	*lpx*B	4.7193	830
Gene099985	Sphingobacteriia	*lpx*B	5.9588	4582
Gene093365	*Synechococcus*	*lpx*B	48.7849	31264
Gene046907	*Synechococcus*	*lpx*B	29.4319	26043
Gene002088	*Synechococcus*	*lpx*B	50.9931	12406
Gene003456	*Synechococcus*	*lpx*B	25.3873	8677
Gene151292	*Synechococcus*	*lpx*B	5.1626	2976
Gene026732	*Synechococcus*	*lpx*B	11.2156	1470
Gene175523	*Synechococcus*	*lpx*B	3.1397	1145
Gene134916	*Synechococcus*	*lpx*B	8.9304	920
Gene064321	unclassified	*lpx*B	4.3044	2109
Gene016881	Betaproteobacteria unclassified	*lpx*B	3.2571	2217
Gene066664	Burkholderiales unclassified	*lpx*B	2.5003	1893
Gene022231	Burkholderiales Burkholderiaceae	*lpx*C	2.3224	1737
Gene052817	Caulobacterales	*lpx*C	4.957	4909
Gene024710	Caulobacterales	*lpx*C	3.7668	1574
Gene015163	Comamonadaceae	*lpx*C	3.2016	2441
Gene157725	Comamonadaceae	*lpx*C	2.1055	2076
Gene159574	Comamonadaceae	*lpx*C	4.3996	1912
Gene116845	Comamonadaceae	*lpx*C	4.7049	1623
Gene079074	Comamonadaceae	*lpx*C	3.0355	1240
Gene035632	Comamonadaceae	*lpx*C	2.3206	1151
Gene089614	Comamonadaceae	*lpx*C	2.6208	778
Gene140579	*Cyanobium*	*lpx*C	21.2796	21301
Gene048549	*Cyanobium*	*lpx*C	9.8871	12199
Gene150336	*Cyanobium*	*lpx*C	6.3109	3194
Gene106956	*Cyanobium*	*lpx*C	5.2912	2593
Gene068045	*Cyanobium*	*lpx*C	4.7683	1778
Gene029857	*Cyanobium*	*lpx*C	6.4145	1334
Gene045516	*Cyanobium*	*lpx*C	10.4635	548
Gene115393	Nitrosomonadales	*lpx*C	2.8983	1721
Gene002838	Planctomycetes	*lpx*C	9.9428	10509

Gene160539	Planctomycetes	*lpx*C	4.1427	1836
Gene073147	Planctomycetes	*lpx*C	12.1215	1473
Gene119972	Planctomycetes	*lpx*C	3.7058	1458
Gene078689	Planctomycetes	*lpx*C	9.1708	1253
Gene131795	Planctomycetes	*lpx*C	5.6781	1078
Gene134670	Planctomycetes	*lpx*C	4.8247	947
Gene093368	*Synechococcus*	*lpx*C	48.7849	31264
Gene003459	*Synechococcus*	*lpx*C	25.3873	8677
Gene078555	*Synechococcus*	*lpx*C	3.3007	1257
Gene070168	unclassified Betaproteobacteria	*lpx*C	3.2425	1637
Gene122530	Burkholderiaceae	*lpx*D	3.8858	1348
Gene176280	Burkholderiaceae	*lpx*D	2.8865	1128
Gene036668	Burkholderiaceae	*lpx*D	2.6332	1126
Gene040277	Capnocytophaga	*lpx*D	3.9155	1042
Gene114307	Caulobacterales	*lpx*D	4.8624	1802
Gene084652	Colwelliaceae	*lpx*D	1.9129	1091
Gene143785	Comamonadaceae	*lpx*D	4.8279	6974
Gene107556	Comamonadaceae	*lpx*D	6.0462	2511
Gene114648	Comamonadaceae	*lpx*D	2.0288	1773
Gene024815	Comamonadaceae	*lpx*D	2.8853	1569
Gene077239	Comamonadaceae	*lpx*D	2.0299	1304
Gene124276	Comamonadaceae	*lpx*D	3.5523	1282
Gene172482	Comamonadaceae	*lpx*D	3.502	1231
Gene045085	Comamonadaceae	*lpx*D	2.8559	583
Gene092621	Comamonadaceae	*lpx*D	2.3021	523
Gene110121	*Cyanobium*	*lpx*D	6.5007	2183
Gene110433	*Cyanobium*	*lpx*D	3.5381	2154
Gene083302	*Cyanobium*	*lpx*D	3.5582	1125
Gene180858	*Cyanobium*	*lpx*D	2.831	1012
Gene041785	*Cyanobium*	*lpx*D	9.9455	917
Gene041964	*Cyanobium*	*lpx*D	2.7508	895
Gene045190	*Cyanobium*	*lpx*D	1.9913	576
Gene159874	Deltaproteobacteria	*lpx*D	3.9671	1886
Gene030797	Deltaproteobacteria	*lpx*D	3.4174	1301
Gene159427	Flavobacteriia	*lpx*D	3.9465	1927
Gene007295	Opitutae	*lpx*D	7.7145	4466
Gene045000	Opitutae	*lpx*D	1.3695	590
Gene003258	Planctomycetes	*lpx*D	6.8303	9245

Gene097789	Planctomycetes	*lpx*D	10.0394	6274
Gene098280	Planctomycetes	*lpx*D	5.4623	5830
Gene015602	Planctomycetes	*lpx*D	4.366	2377
Gene021573	Planctomycetes	*lpx*D	3.6907	1788
Gene123130	Planctomycetes	*lpx*D	2.9563	1326
Gene077092	Planctomycetes	*lpx*D	1.8037	1309
Gene031800	Planctomycetes	*lpx*D	4.6893	1265
Gene088488	Planctomycetes	*lpx*D	8.1705	903
Gene090949	Planctomycetes	*lpx*D	1.4593	651
Gene045738	Planctomycetes	*lpx*D	1.0918	534
Gene001466	Sphingobacteriia	*lpx*D	7.5658	14444
Gene055169	Sphingobacteriia	*lpx*D	5.1356	3769
Gene156950	Sphingobacteriia	*lpx*D	5.16	2156
Gene119597	Sphingobacteriia	*lpx*D	2.7549	1473
Gene093661	*Synechococcus*	*lpx*D	28.2464	25023
Gene140322	*Synechococcus*	*lpx*D	48.0148	23488
Gene093898	*Synechococcus*	*lpx*D	19.3033	22335
Gene096732	*Synechococcus*	*lpx*D	10.0489	7860
Gene053591	*Synechococcus*	*lpx*D	3.6493	4448
Gene022470	*Synechococcus*	*lpx*D	4.0041	1722
Gene077273	*Synechococcus*	*lpx*D	13.1131	1300
Gene036065	*Synechococcus*	*lpx*D	15.6219	1140
Gene132390	*Synechococcus*	*lpx*D	1.4347	1065
Gene118943	unclassified	*lpx*D	3.3643	1507
Gene121467	Burkholderiales unclassified	*lpx*D	3.6336	1392
Gene032038	Burkholderiales Verrucomicrobia	*lpx*D	3.1258	1256
Gene139223	Verrucomicrobia	*lpx*D	2.2144	513
Gene046280	Bacteria	*lpx*K	20.9834	64759
Gene113978	Burkholderiaceae	*lpx*K	3.9256	1827
Gene086575	Burkholderiaceae	*lpx*K	1.7216	1049
Gene053203	Caulobacterales	*lpx*K	4.0796	4637
Gene053978	Comamonadaceae	*lpx*K	4.1637	4270
Gene055721	Comamonadaceae	*lpx*K	3.9888	3578
Gene168517	Comamonadaceae	*lpx*K	2.7297	1376
Gene140874	*Cyanobium*	*lpx*K	11.1639	19253
Gene094344	*Cyanobium*	*lpx*K	19.5563	18919
Gene153371	*Cyanobium*	*lpx*K	6.8944	2614
Gene157094	*Cyanobium*	*lpx*K	17.1118	2137

Gene026726	*Cyanobium*	*lpx*K	8.715	1470
Gene088815	*Cyanobium*	*lpx*K	10.3972	866
Gene089189	*Cyanobium*	*lpx*K	6.9698	828
Gene135663	*Cyanobium*	*lpx*K	4.6513	826
Gene048920	Planctomycetes	*lpx*K	8.4489	10829
Gene099955	Planctomycetes	*lpx*K	3.0504	4586
Gene010642	Planctomycetes	*lpx*K	11.2293	3249
Gene011309	Planctomycetes	*lpx*K	3.5889	3104
Gene151882	Planctomycetes	*lpx*K	4.4146	2870
Gene114293	Planctomycetes	*lpx*K	8.8696	1802
Gene120794	Planctomycetes	*lpx*K	3.9492	1416
Gene124678	Rhodospirillales	*lpx*K	1.3932	1269
Gene001177	Sphingobacteriia	*lpx*K	7.0934	15646
Gene163567	Sphingobacteriia	*lpx*K	3.6724	1621
Gene120485	Sphingobacteriia	*lpx*K	2.318	1431
Gene094016	*Synechococcus*	*lpx*K	52.1291	21385
Gene110001	*Synechococcus*	*lpx*K	23.9653	2193
Gene166967	*Synechococcus*	*lpx*K	2.9126	1441
Gene060002	unclassified Bacteria	*lpx*K	3.1605	2661
Gene034101	Verrucomicrobia	*lpx*K	1.8581	1191
Gene010742	Verrucomicrobiae	*lpx*K	4.0527	3227
Gene079487	Verrucomicrobiae	*lpx*K	2.0636	1227
Gene132101	Verrucomicrobiae	*lpx*K	3.0701	1070
Gene060512	Methylococcales	*mcy*D	3.4591	2579
Gene018127	Methylococcales	*mcy*D	3.8842	2090
Gene113917	Planctomycetes	*mcy*E	2.9864	1835
Gene046915	*Cyanobium*	*mcy*H	29.4319	26043
Gene002009	*Cyanobium*	*mcy*H	19.5128	12548
Gene149698	*Cyanobium*	*mcy*H	4.6756	3363
Gene152057	*Cyanobium*	*mcy*H	4.8681	2844
Gene155016	*Cyanobium*	*mcy*H	4.3977	2394
Gene155142	*Cyanobium*	*mcy*H	7.9933	2373
Gene021866	*Cyanobium*	*mcy*H	7.8511	1766
Gene025040	*Cyanobium*	*mcy*H	5.2776	1556
Gene168357	*Cyanobium*	*mcy*H	2.9689	1383
Gene122879	*Cyanobium*	*mcy*H	11.829	1333
Gene077740	*Cyanobium*	*mcy*H	5.1034	1286
Gene124834	*Cyanobium*	*mcy*H	10.7555	1264
Gene177050	*Cyanobium*	*mcy*H	6.0009	1108

Gene041865	*Cyanobium*	*mcy*H	5.511	906
Gene135338	*Cyanobium*	*mcy*H	3.7161	863
Gene184277	*Cyanobium*	*mcy*H	1.8842	639
Gene086068	Methylococcales	*mcy*H	1.8585	1060
Gene094238	*Synechococcus*	*mcy*H	9.8544	19260
Gene002080	*Synechococcus*	*mcy*H	50.9931	12406
Gene042571	*Synechococcus*	*mcy*H	13.0708	819
Gene184471	*Synechococcus*	*mcy*H	2.6758	620
Gene160144	Pseudonocardiales	*mcy*J	3.8907	1867
Gene141470	Acidobacteria	*nif*J	7.4141	15988
Gene141469	Chloroflexia	*nif*J	7.4141	15988
Gene026344	Comamonadaceae	*nif*J	5.5221	1490
Gene007542	*Cyanobium*	*nif*J	4.9611	4373
Gene105850	*Cyanobium*	*nif*J	6.7732	2765
Gene105851	*Cyanobium*	*nif*J	6.7732	2765
Gene066255	*Cyanobium*	*nif*J	7.4403	1926
Gene168563	*Cyanobium*	*nif*J	6.2153	1375
Gene025129	*Synechococcus*	*nif*J	2.9446	1551
Gene099739	*Cyanobium*	*nif*K	5.6364	4681
Gene100478	*Cyanobium*	*nif*K	8.6711	4342
Gene009375	*Cyanobium*	*nif*K	4.8595	3601
Gene065437	*Cyanobium*	*nif*K	8.2461	2003
Gene021163	*Cyanobium*	*nif*K	3.489	1820
Gene062233	Bradyrhizobiaceae	*nif*S	2.8713	2346
Gene016191	Caulobacterales	*nif*S	4.4767	2299
Gene139510	*Cyanobium*	*nif*S	19.1877	45382
Gene093157	*Cyanobium*	*nif*S	10.9951	37011
Gene101444	*Cyanobium*	*nif*S	8.8036	3895
Gene148046	*Cyanobium*	*nif*S	6.6835	3871
Gene064698	*Cyanobium*	*nif*S	5.778	2068
Gene115099	*Cyanobium*	*nif*S	4.7724	1744
Gene088908	*Cyanobium*	*nif*S	6.2832	858
Gene044408	*Cyanobium*	*nif*S	7.1377	639
Gene024940	Desulfovibrionales	*nif*S	3.5506	1562
Gene139623	*Synechococcus*	*nif*S	50.416	38978
Gene142083	*Synechococcus*	*nif*S	27.9825	13739
Gene181629	*Synechococcus*	*nif*S	1.5658	912
Gene136714	*Synechococcus*	*nif*S	5.3329	715
Gene128430	Verrucomicrobia	*nif*S	2.8418	1157

	Gene102418	Acidimicrobiia	*nif*U	5.8279	3550
	Gene061894	Acidimicrobiia	*nif*U	5.6959	2397
	Gene135683	Acidimicrobiia	*nif*U	4.1864	826
	Gene001107	Alicyclobacillaceae	*nif*U	9.0406	15872
	Gene111506	Bacteria	*nif*U	4.0387	2042
	Gene031290	Caulobacterales	*nif*U	4.6139	1282
	Gene053288	Cyanobacteria	*nif*U	39.7998	4595
	Gene143214	*Cyanobium*	*nif*U	27.9963	7835
	Gene053791	*Cyanobium*	*nif*U	5.7482	4365
	Gene111698	*Cyanobium*	*nif*U	27.5816	2027
	Gene072823	*Cyanobium*	*nif*U	10.8649	1488
	Gene035039	*Cyanobium*	*nif*U	7.4087	1167
	Gene092141	*Cyanobium*	*nif*U	1.9497	557
	Gene144318	Rhizobiaceae	*nif*U	5.0189	6233
	Gene049761	Sphingobacteriia	*nif*U	8.0449	8459
	Gene021928	Sphingobacteriia	*nif*U	2.6633	1758
	Gene034877	Sphingobacteriia	*nif*U	1.7543	1172
	Gene098556	*Synechococcus*	*nif*U	13.4226	5532
	Gene018227	*Synechococcus*	*nif*U	39.6104	2079
	Gene001109	Thermoanaerobacterales	*nif*U	9.0406	15872
	Gene067331	Verrucomicrobia	*nif*U	3.6761	1840
	Gene139462	*Cyanobium*	*ntc*A	20.9834	64759
	Gene144381	*Cyanobium*	*ntc*A	10.5712	6226
	Gene051287	*Cyanobium*	*ntc*A	11.4962	6123
	Gene106263	*Cyanobium*	*ntc*A	8.7188	2703
	Gene154218	*Cyanobium*	*ntc*A	14.2397	2499
	Gene021267	*Cyanobium*	*ntc*A	11.9602	1811
	Gene135033	*Cyanobium*	*ntc*A	3.4807	909
	Gene090250	*Cyanobium*	*ntc*A	6.1196	719
	Gene094978	*Synechococcus*	*ntc*A	46.5025	16032
	Gene097503	*Synechococcus*	*ntc*A	31.2379	6613
	Gene144615	*Synechococcus*	*ntc*A	5.4492	5999
TLW233	Gene01197	Comamonadaceae	*cyn*S	2.8091	1247
	Gene18998	Burkholderiaceae	*cyn*S	2.3047	965
	Gene02547	Burkholderiaceae	*cyn*S	2.8837	946
	Gene16550	Comamonadaceae	*lpx*A	4.131	2129
	Gene10259	Salmonella	*lpx*A	1.5162	1577
	Gene07009	Comamonadaceae	*lpx*A	2.1991	1562
	Gene07383	Chromatiales	*lpx*A	2.6689	1365

	Gene14251	unclassified Betaproteobacteria	*lpx*A	2.3226	1237
	Gene24598	Burkholderiaceae	*lpx*A	2.7741	1160
	Gene24599	unclassified Betaproteobacteria	*lpx*A	2.7741	1160
	Gene22149	unclassified Burkholderiales	*lpx*A	4.2349	992
	Gene12676	Burkholderiaceae	*lpx*A	4.0405	667
	Gene12677	Burkholderiaceae	*lpx*B	4.0405	667
	Gene03253	Nitrosomonadales	*lpx*C	3.5987	2208
	Gene03334	Planctomycetes	*lpx*C	7.9538	1993
	Gene20260	Comamonadaceae	*lpx*C	4.0755	1642
	Gene10260	Planctomycetes	*lpx*C	1.5162	1577
	Gene17552	Comamonadaceae	*lpx*C	3.7401	1266
	Gene24858	Oxalobacteraceae	*lpx*C	4.3971	1093
	Gene22148	Planctomycetes	*lpx*C	4.2349	992
	Gene15692	unclassified Betaproteobacteria	*lpx*C	2.6333	968
	Gene15976	Alcaligenaceae	*lpx*C	1.6099	933
	Gene09885	Comamonadaceae	*lpx*D	4.131	2129
	Gene03335	Flavobacteriia	*lpx*D	7.9538	1993
	Gene06897	Deltaproteobacteria	*lpx*D	6.265	1653
	Gene06916	Flavobacterium	*lpx*D	3.9165	1652
	Gene04624	Flavobacterium	*lpx*D	6.4454	1154
	Gene04640	Desulfuromonadales	*lpx*D	1.7288	1154
	Gene24640	Comamonadaceae	*lpx*D	2.1134	1138
	Gene24716	Comamonadaceae	*lpx*D	3.6123	1135
	Gene22012	Planctomycetes	*lpx*D	2.4757	1028
	Gene22534	Flavobacterium	*lpx*D	1.0503	955
	Gene03538	Flavobacterium	*lpx*K	5.7119	1711
	Gene02003	Planctomycetes	*lpx*K	7.2027	1031
	Gene15612	Comamonadaceae	*lpx*K	3.4059	983
	Gene05696	Planctomycetes	*lpx*K	3.1913	967
	Gene15651	Methylococcales	*mcy*D	2.1889	969
	Gene07007	Comamonadaceae	*nif*J	2.7599	1562
	Gene04470	Chloroflexia	*nif*J	4.4066	1205
	Gene20003	Desulfovibrionales	*nif*S	3.1591	993
	Gene04254	Sphingobacteriia	*nif*U	2.8885	1273
	Gene09224	unclassified Alphaproteobacteria	*nif*U	1.6709	936
TLW278	Gene23659	*Cyanobium*	*cyn*S	3.5921	1856

Gene19041	*Cyanobium*	*cyn*S	4.3377	1380
Gene19170	Rhodospirillales	*cyn*S	3.108	1306
Gene42003	Comamonadaceae	*cyn*S	3.2035	1071
Gene29084	*Cyanobium*	*lpx*A	6.0526	1843
Gene18924	Comamonadaceae	*lpx*A	2.1378	1422
Gene31867	Cryomorphaceae	*lpx*A	3.1228	945
Gene26632	Cytophagia	*lpx*A	4.9492	926
Gene10076	Comamonadaceae	*lpx*A	3.6104	924
Gene37544	*Synechococcus*	*lpx*A	6.7492	921
Gene05166	Flavobacterium	*lpx*A	2.1322	847
Gene22119	Sphingobacteriia	*lpx*A	6.2345	840
Gene16956	Sphingobacteriia	*lpx*A	2.5215	813
Gene11606	*Synechococcus*	*lpx*B	11.6228	13393
Gene00817	*Synechococcus*	*lpx*B	3.7127	2283
Gene29085	*Synechococcus*	*lpx*B	6.0526	1843
Gene01678	Flavobacterium	*lpx*B	5.0858	1561
Gene18923	Comamonadaceae	*lpx*B	2.1378	1422
Gene13707	Flavobacteriia	*lpx*B	4.1304	1311
Gene02271	*Synechococcus*	*lpx*B	5.3608	1297
Gene08375	*Synechococcus*	*lpx*B	2.4775	1179
Gene31054	*Synechococcus*	*lpx*B	4.1612	1061
Gene10125	*Cyanobium*	*lpx*B	4.3814	923
Gene37543	*Synechococcus*	*lpx*B	6.7492	921
Gene21511	Comamonadaceae	*lpx*B	1.9398	897
Gene16260	*Synechococcus*	*lpx*B	3.1369	869
Gene16446	*Cyanobium*	*lpx*B	2.3469	859
Gene17030	Burkholderiaceae	*lpx*B	2.4594	812
Gene23704	*Cyanobium*	*lpx*C	6.9238	1851
Gene30866	Comamonadaceae	*lpx*C	3.5829	1098
Gene31655	Comamonadaceae	*lpx*C	3.2438	972
Gene37344	Rhodospirillales	*lpx*C	2.1186	944
Gene32926	Comamonadaceae	*lpx*C	4.5782	844
Gene16977	Nitrosomonadales	*lpx*C	1.3776	813
Gene23428	Sphingobacteriia	*lpx*D	3.2189	2179
Gene07367	Flavobacteriia	*lpx*D	4.3399	1568
Gene35078	Comamonadaceae	*lpx*D	6.2294	1495
Gene01936	*Cyanobium*	*lpx*D	3.495	1412
Gene24454	unclassified Alphaproteobacteria	*lpx*D	1.3912	1360

Gene24573	Comamonadaceae	*lpx*D	2.6553	1320
Gene02744	Sphingobacteriia	*lpx*D	4.9638	1160
Gene41689	Cyanobium	*lpx*D	3.7811	1124
Gene36337	Sphingobacteriia	*lpx*D	4.2819	1096
Gene03332	*Synechococcus*	*lpx*D	2.3796	1051
Gene31546	unclassified Alphaproteobacteria	*lpx*D	1.9624	985
Gene37046	Bacteroidaceae	*lpx*D	2.2431	983
Gene20929	Bacteroidetes	*lpx*D	2.7827	962
Gene21203	Bacteroidetes	*lpx*D	2.9644	928
Gene10523	Synechococcus	*lpx*D	2.733	884
Gene27009	*Cyanobium*	*lpx*D	1.6701	882
Gene05333	Comamonadaceae	*lpx*D	2.7485	831
Gene34430	Comamonadaceae	*lpx*K	7.2597	724
Gene01183	Flavobacterium	*lpx*K	7.4214	1896
Gene40706	Flavobacterium	*lpx*K	4.7112	1430
Gene35966	Comamonadaceae	*lpx*K	2.4492	1182
Gene08698	Sphingobacteriia	*lpx*K	6.4576	1119
Gene15208	*Synechococcus*	*lpx*K	4.6563	992
Gene04234	*Synechococcus*	*lpx*K	2.0727	935
Gene21487	Verrucomicrobia	*lpx*K	2.4939	897
Gene32674	Flavobacteriia	*lpx*K	2.7761	862
Gene10799	Sphingobacteriia	*lpx*K	1.7529	858
Gene38439	Sphingobacteriia	*lpx*K	2.2734	834
Gene36396	Myxococcales	*mcy*E	6.2706	1094
Gene00809	*Cyanobium*	*mcy*H	3.3767	945
Gene01230	*Cyanobium*	*mcy*H	11.8462	1820
Gene02421	*Cyanobium*	*mcy*H	3.8417	1238
Gene24879	*Cyanobium*	*mcy*H	8.2387	1219
Gene19995	*Cyanobium*	*mcy*H	3.1434	1102
Gene22052	*Cyanobium*	*mcy*H	3.1439	848
Gene36862	Acidobacteria	*nif*J	4.8208	1010
Gene43438	Acidobacteria	*nif*J	2.7648	876
Gene17594	*Cyanobium*	*nif*K	7.4131	2803
Gene39284	*Cyanobium*	*nif*S	3.2091	1387
Gene07991	*Cyanobium*	*nif*S	3.3129	1275
Gene13870	*Synechococcus*	*nif*S	2.7141	1245
Gene14683	Rickettsiales	*nif*S	5.0214	1075
Gene41981	*Synechococcus*	*nif*S	3.0411	1071

	Gene17019	*Cyanobium*	*nif*S	3.8744	812
	Gene17247	*Cyanobium*	*nif*U	6.5144	5655
	Gene12183	Flavobacterium	*nif*U	6.7218	2678
	Gene12645	Rhodospirillales	*nif*U	6.7727	1971
	Gene01294	Flavobacterium	*nif*U	4.8409	1747
	Gene07034	Thermoanaerobacterales	*nif*U	3.2325	1738
	Gene30990	Sphingobacteriia	*nif*U	4.6052	1079
	Gene03711	Sphingomonadales	*nif*U	3.4167	996
	Gene09847	Cryomorphaceae	*nif*U	2.0898	958
	Gene25463	Cyanobium	*ntc*A	4.6707	1087
TLW323	Gene9021	Verrucomicrobia	*lpx*A	2.7554	830
	Gene3144	Maribacter	*lpxB*	1.6291	922
	Gene7382	unclassified Alphaproteobacteria	*lpxB*	4.1917	1153
	Gene8630	Comamonadaceae	*lpx*K	2.6709	1015
	Gene6462	Nostocales	*mcy*A	21.8929	971
	Gene2190	Verrucomicrobia	*nif*S	3.2323	848
	Gene0353	Cyanobacteria	*nif*U	3.5948	1414
TLW368	Gene01007	Beijerinckiaceae	*cyn*S	21.2917	1824
	Gene22618	Rhodospirillales	*lpx*A	3.7607	844
	Gene18334	Verrucomicrobia	*lpx*A	7.5192	2708
	Gene09701	Oxalobacteraceae	*lpx*A	6.603	2025
	Gene14873	Verrucomicrobiae	*lpx*A	3.7884	1484
	Gene24449	Deltaproteobacteria	*lpx*A	1.9012	1316
	Gene34501	unclassified Verrucomicrobia	*lpx*A	3.5167	1169
	Gene21377	Acidobacteria	*lpx*A	7.2878	952
	Gene31057	Rhodocyclales	*lpx*A	6.6092	916
	Gene22620	unclassified Alphaproteobacteria	*lpx*B	1.9313	844
	Gene05041	unclassified Betaproteobacteria	*lpx*B	5.7646	2230
	Gene05404	Verrucomicrobia	*lpx*B	4.0972	1842
	Gene33450	Rhodospirillales	*lpx*B	3.381	1593
	Gene01340	Nitrospirae	*lpx*B	3.3407	1579
	Gene06599	Planctomycetes	*lpx*B	3.2452	1203
	Gene34516	Verrucomicrobia	*lpx*B	4.3253	1168
	Gene30342	Synechococcus	*lpx*B	3.1504	1017
	Gene16874	Burkholderiaceae	*lpx*B	4.5384	938
	Gene31056	Rhodocyclales	*lpx*B	6.6092	916

Gene17386	Burkholderiaceae	*lpx*B	6.2189	877
Gene17635	Rhodospirillales	*lpx*B	3.4633	844
Gene00178	unclassified	*lpx*C	8.0989	4511
Gene18228	Betaproteobacteria unclassified	*lpx*C	12.951	3039
Gene14840	Alphaproteobacteria Chromatiales	*lpx*C	8.3318	1486
Gene19522	Rhodospirillales	*lpx*C	4.1931	1414
Gene19587	Burkholderiaceae	*lpx*C	5.2087	1409
Gene15115	Cyanobium	*lpx*C	2.8531	1348
Gene06833	unclassified	*lpx*C	5.8568	1124
Gene25544	Betaproteobacteria Rhodospirillales	*lpx*C	3.1832	1026
Gene03040	Planctomycetes	*lpx*C	3.0845	994
Gene25832	Rhodospirillales	*lpx*C	2.536	972
Gene22136	Rhodospirillales	*lpx*C	3.2669	858
Gene22616	unclassified	*lpx*D	2.8116	844
Gene09703	Alphaproteobacteria Rhodocyclales	*lpx*D	4.394	571
Gene05649	Burkholderiaceae	*lpx*D	3.7901	1615
Gene28648	Rhodospirillales	*lpx*D	4.6229	1546
Gene28731	Synechococcus	*lpx*D	1.1046	1491
Gene05953	Rhodospirillales	*lpx*D	2.8644	1431
Gene10792	unclassified	*lpx*D	6.6582	1296
Gene24531	Betaproteobacteria Verrucomicrobia	*lpx*D	7.105	1276
Gene02164	Burkholderiaceae	*lpx*D	5.9951	1214
Gene20902	Dokdonia	*lpx*D	2.5474	1023
Gene20976	Arenibacter	*lpx*D	4.2086	1021
Gene11867	Sphingobacteriia	*lpx*D	3.396	1010
Gene21362	Gammaproteobacteria	*lpx*D	8.6113	952
Gene31474	Verrucomicrobiae	*lpx*D	3.2944	873
Gene27427	unclassified	*lpx*K	11.0648	3628
Gene00457	Alphaproteobacteria unclassified	*lpx*K	7.3295	3059
Gene13956	Betaproteobacteria Alcaligenaceae	*lpx*K	8.0401	2392
Gene05690	Verrucomicrobia	*lpx*K	5.1732	1611
Gene34525	Verrucomicrobia	*lpx*K	3.4957	1168
Gene11415	Brucellaceae	*lpx*K	3.1457	1119
Gene07131	Verrucomicrobia	*lpx*K	6.2686	1050

	Gene08060	Rhodospirillales	*lpx*K	1.8661	911
	Gene08388	Rhodospirillales	*lpx*K	3.3824	876
	Gene08431	Ruminococcaceae	*mcy*C	2.2644	866
	Gene25548	Clostridia	*mcy*D	2.8411	1026
	Gene14634	Synechococcus	*mcy*H	5.6494	1623
	Gene19346	Cyanothece	*nif*J	2.6984	1512
	Gene08164	Cyanobacteria	*nif*J	3.2247	899
	Gene00285	Verrucomicrobia	*nif*S	5.7392	3738
	Gene18440	Rhodospirillales	*nif*S	9.1616	2445
	Gene19016	Caulobacterales	*nif*S	1.9631	1708
	Gene34882	Cyanobium	*nif*S	3.693	1075
	Gene17993	unclassified Alphaproteobacteria	*nif*U	2.0439	934
	Gene14406	Verrucomicrobia	*nif*U	4.8006	1851
	Gene34053	unclassified	*nif*U	10.0622	1318
	Gene15347	Alphaproteobacteria unclassified	*nif*U	6.2036	1272
	Gene20217	Alphaproteobacteria Rhodospirillales	*nif*U	3.0246	1177
	Gene02323	Verrucomicrobiae	*nif*U	3.1615	1158
	Gene21161	Caulobacterales	*nif*U	4.2924	985
	Gene35940	Sphingobacteriia	*nif*U	1.3283	920
	Gene22244	Verrucomicrobia	*nif*U	2.0682	850
	Gene31846	Acidimicrobiia	*nif*U	4.4036	835
	Gene19631	*Cyanobium*	*ntc*A	3.6407	1372
TLW410	Gene00062	Opitutae	*lpx*A	6.3787	9638
	Gene39975	Opitutae	*lpx*A	12.5689	7339
	Gene11516	Opitutae	*lpx*A	9.1764	7070
	Gene23159	Opitutae	*lpx*A	6.8727	5469
	Gene05668	Verrucomicrobia	*lpx*A	6.6279	4915
	Gene17734	Nitrosomonadales	*lpx*A	7.2706	3433
	Gene35180	unclassified	*lpx*A	6.3764	2723
	Gene06872	Verrucomicrobia unclassified	*lpx*A	7.2361	2109
	Gene24919	Verrucomicrobia unclassified	*lpx*A	4.3422	1768
	Gene07728	Verrucomicrobia Flavobacteriia	*lpx*A	7.4991	1635
	Gene13716	Opitutae	*lpx*A	3.1109	1623
	Gene13979	Opitutae	*lpx*A	6.7364	1487
	Gene37051	Nitrosomonadales	*lpx*A	8.3778	1416

Gene43461	Citrobacter	*lpx*A	1.2433	1196
Gene43705	unclassified Betaproteobacteria	*lpx*A	3.2487	1154
Gene09439	Comamonadaceae	*lpx*A	4.2315	1149
Gene15701	Sphingobacteriia	*lpx*A	6.6967	1078
Gene27510	Methylobacteriaceae	*lpx*A	2.4644	1025
Gene22072	unclassified Verrucomicrobia	*lpx*A	5.0598	986
Gene11083	Deinococcus-Thermus	*lpx*A	4.3757	708
Gene22752	Opitutae	*lpx*B	15.6954	46419
Gene05558	Opitutae	*lpx*B	8.1498	5709
Gene11936	Verrucomicrobiae	*lpx*B	6.4567	3858
Gene17735	Nitrosomonadales	*lpx*B	7.2706	3433
Gene23954	Aquimarina	*lpx*B	8.9463	2662
Gene36007	Verrucomicrobiae	*lpx*B	4.6602	1919
Gene14626	Sphingobacteriia	*lpx*B	9.4496	1299
Gene37743	Sphingobacteriia	*lpx*B	2.7764	1212
Gene09220	Flavobacterium	*lpx*B	6.7392	1185
Gene43704	unclassified Betaproteobacteria	*lpx*B	3.2487	1154
Gene27511	Methylobacteriaceae	*lpx*B	2.4644	1025
Gene33494	Cryomorphaceae	*lpx*B	4.4862	981
Gene23814	Nitrosomonadales	*lpx*C	8.5735	2804
Gene06654	Nitrosomonadales	*lpx*C	8.1473	2363
Gene09753	Nitrosomonadales	*lpx*C	4.1629	1099
Gene33497	Caulobacterales	*lpx*C	3.7176	981
Gene39818	Opitutae	*lpx*D	14.4678	22529
Gene17782	Porphyromonadaceae	*lpx*D	7.8446	3416
Gene35188	Opitutae	*lpx*D	7.6886	2720
Gene01182	Opitutae	*lpx*D	8.4776	2590
Gene01467	Cytophagia	*lpx*D	6.252	2226
Gene18588	Verrucomicrobiae	*lpx*D	3.8852	2178
Gene30024	Verrucomicrobia	*lpx*D	5.9217	2157
Gene01521	Sphingobacteriia	*lpx*D	7.1043	2139
Gene19412	Sphingobacteriia	*lpx*D	4.512	1666
Gene37053	Nitrosomonadales	*lpx*D	8.3778	1416
Gene20215	Cytophagia	*lpx*D	3.2233	1348
Gene20237	Verrucomicrobia	*lpx*D	3.8762	1325
Gene42963	unclassified Alphaproteobacteria	*lpx*D	3.197	1320

Gene20365	Sphingobacteriia	*lpx*D	3.651	1298
Gene09441	Comamonadaceae	*lpx*D	4.2315	1149
Gene21363	Flavobacteriia	*lpx*D	4.1193	1090
Gene44194	Flavobacterium	*lpx*D	6.7798	1067
Gene33208	Cryomorphaceae	*lpx*D	3.423	1019
Gene10967	Flavobacterium	*lpx*D	2.968	938
Gene11198	Opitutae	*lpx*K	13.6087	92276
Gene17228	Opitutae	*lpx*K	7.8843	8763
Gene40565	Verrucomicrobiae	*lpx*K	7.5292	3120
Gene29445	Cytophagia	*lpx*K	9.1575	2908
Gene29601	Verrucomicrobiae	*lpx*K	6.4557	2587
Gene41732	Comamonadaceae	*lpx*K	5.4562	1813
Gene20396	Opitutae	*lpx*K	3.9869	1297
Gene43634	Verrucomicrobiae	*lpx*K	4.9922	1155
Gene27238	Nitrosomonadales	*lpx*K	3.5766	1070
Gene10462	Verrucomicrobiae	*lpx*K	6.7251	993
Gene04739	Sphingobacteriia	*lpx*K	5.1225	963
Gene33818	Opitutae	*lpx*K	6.434	947
Gene31635	Opitutae	*nif*J	3.3398	1339
Gene09584	Oscillatoriales	*nif*J	3.2979	1118
Gene16035	Opitutae	*nif*J	3.6177	1028
Gene35329	Verrucomicrobiae	*nif*S	5.6562	2551
Gene25445	Rickettsiales	*nif*S	3.7758	1521
Gene42515	Verrucomicrobia	*nif*S	3.6857	1438
Gene31512	Verrucomicrobia	*nif*S	3.826	1368
Gene37221	Desulfovibrionales	*nif*S	5.3776	1356
Gene16452	Verrucomicrobia	*nif*S	4.2391	966
Gene23151	Sphingobacteriia	*nif*U	7.4275	5471
Gene05756	Sphingobacteriia	*nif*U	7.126	4356
Gene24033	Cytophagia	*nif*U	7.295	2512
Gene07635	Bacteroidetes	*nif*U	6.7329	1685
Gene13693	Bacteroidetes	*nif*U	4.336	1625
Gene42365	unclassified Alphaproteobacteria	*nif*U	4.1273	1516

Table E4. Additional functional genes searched for within metagenomic eggNOG data.

Gene Family	Target Gene	Observed in metadata	Function	Reference
Anatoxin	*ana*	N/A	Synthesizes neurotoxin anatoxin-a	Rantala-Ylinen et al., 2011 10.1128/AEM.06022-11
	axt	N/A	Anatoxin-a synthesis	N/A
Aplasyiatoxin	*apx*	TLW43	Involved in biosynthesis of dermatotoxin aplasyiatoxin	Dittman et al., 2013 10.1111/j.1574-6976.2012.12000.x
BMAA	Unknown	N/A	Synthesizes BMAA toxin	Dittman et al., 2013 10.1111/j.1574-6976.2012.12000.x
Cyanase	*cyn*	TLW43, 233, 278, 368	Involved in cyanate metabolism	Sáez et al., 2019 10.3390/ijms20123008
Cylindrospermopsin	*aoa*	N/A	Involved in cylindrospermopsin biosynthesis	Kaplan et al., 2012 10.3389/fmicb.2012.00138
	cyr	N/A	Synthesizes hepatotoxin cylindrospermopsin	Mazmouz et al., 2010 10.1128/AEM.00717-10
Cytochrome P450	*cyp*	All samples	Contribute to production of secondary metabolites	Kumalo et al., 2020 10.3390/ijms21020656
Lipopolysaccharide	*lps*	TLW43, 233, 278, 323	Biosynthesis of lipopolysaccharide wall	N/A
Lyngbyatoxin	*ltx*	N/A	Synthesizes dermatotoxins lyngbyatoxin	Dittman et al., 2013 10.1111/j.1574-6976.2012.12000.x
Geosmin	*geo*	N/A	Synthesizes T/O compound geosmin	Churro et al., 2020

				10.1038/s41598-020-64774-y
Nitrogen Metabolism	*pip*X	N/A	Coactivates nitrogen regulator *ntc*A	Espinosa et al., 2014 https://doi.org/10.1073/pnas.1404097111
Nodularin	*nda*	N/A	Synthesizes potent nodularin toxin	Dittman et al., 2013 10.1111/j.1574-6976.2012.12000.x
Saxitoxin	*sxt*	N/A	Involved in saxitoxin biosynthesis	Dittman et al., 2013 10.1111/j.1574-6976.2012.12000.x