METAGENOMIC CHARACTERIZATION OF BLOOM AND NON-BLOOM SITES

COMPARATIVE METAGENOMICS OF FRESHWATER CYANOBACTERIA BLOOM AND NON-BLOOM SITES IN ONTARIO AND THE INVESTIGATION INTO THE UTILIZATION OF CONSERVED SIGNATURE PROTEINS FOR IDENTIFICATION OF CYANOBACTERIA

By RACHELLE ATRACHE, HBSc

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
Master of Science (2017)
Hamilton, Ontario **(Department of Biology)**

TITLE: Comparative Metagenomics of Freshwater Cyanobacteria Bloom and Non-Bloom Sites in Ontario and the Investigation into the Utilization of Conserved Signature Proteins for Identification of Cyanobacteria
AUTHOR: Rachelle ATRACHE (McMaster University)
SUPERVISOR: Dr. Herb E. Schellhorn
NUMBER OF PAGES: **ix, 89**

Abstract

Cyanobacterial algal blooms have been increasing in frequency and severity over the past few years in Ontario. Depending on the presence of toxigenic Cyanobacteria, these blooms have the potential to release toxins into the water, posing a public and environmental risk to humans and animals. Although traditional methods of studying Cyanobacteria provide important information regarding the microbial community, metagenomic sequencing allows for a more comprehensive examination of microbial diversity and functional capacities as limitations in cultivating organisms is circumvented. Therefore, to gain insight into the community composition of freshwater blooms and to compare them to non-bloom sites within Ontario, we collaborated with the Ministry of the Environment and Climate Change (MOECC) to undergo a high throughput DNA sequencing approach for a comparative metagenomic analysis. In 2015, 108 bloom and non-bloom samples were collected and sent for 16S rRNA sequencing and a subset of these were chosen for shotgun metagenomic sequencing. Our study focuses on comparing community structure and functional differences that may exist between bloom and non-bloom sites as well as analyzing differences in cyanobacterial communities across bloom sites. Our findings reveal differences in the microbial communities between these two environments. At the functional level, large-scale functionalities were conserved across the two groups but differences in specialized functions were revealed. Overall, our results show that metagenomics is a powerful tool for delineating functional and taxonomic analysis of bloom and non-bloom sites across Ontario. The second part of this work studied the utilization of the molecular marker, Conserved Signature Proteins (CSPs), as a valid method for identifying Cyanobacteria to facilitate the problem of cyanobacterial taxonomic classification. It was found that CSPs proved to be reliable in identifying Cyanobacteria within environmental samples when compared to amplicon and shotgun metagenomic sequencing approaches.

Acknowledgements

I would like to start off by thanking Dr. Schellhorn for this great opportunity to work in his lab and under his supervision. You have been a great mentor in the field of research as you nurtured my love for science, motivated and inspired me. The intellect and lessons that you have provided me with will be ones that I will be able to apply outside of the lab and in my career. I would also like to thank Dr. Gupta for providing me with excellent data, direction, and insight with signature proteins. I would like to thank Dr. Brian Golding for being the Chair on my defense. I would like to thank Michelle Palmer and colleagues from the Ministry of the Environment and Climate Change for their collaboration with us on this project and for sending the water samples.

I want to thank all the members of Dr. Schellhorn's lab both past and present. I want to give a special thanks to Bansri, Alex, Cathy, Vivian, Ashish, Paul, Isaac, and Deepinder. You have brought so much laughter and joy to the lab as well as helped to make my time in the lab an enjoyable and memorable experience. I would also like to thank Yasser for all of his help with the bioinformatics and for putting up with my constant questions. Sakis, I cannot thank you enough for all of the guidance that you have provided me in the lab. You were always there to facilitate my research and provided me with great insight with my work.

Last but certainly not least, I would like to thank all my family and friends for their endless love, support, and guidance. I certainly would not have been able to complete this milestone in my life without you all. To my parents, sister and teta, you have been amazing in supporting me and motivating me throughout these past few years. You have always believed in me and made me strive to be the best version of myself and without you, I would not have made it this far. I would especially like to thank Sarah for being an amazing friend who constantly pushed me to be great and who was always there for me when I needed her.

Table of Contents

List of Figures

List of Tables

Declaration of Authorship

I, Rachelle ATRACHE, declare that this thesis titled “Comparative Metagenomics of Freshwater Cyanobacteria Bloom and Non-Bloom Sites in Ontario and the Investigation into the Utilization of Conserved Signature Proteins for Identification of Cyanobacteria” and the work presented in it are my own.

Chapter 1. Introduction

1.1 Algal Blooms

Algal blooms are natural phenomena that are caused by a mass proliferation of phytoplankton in water bodies. Although algal blooms can be caused by different organisms such as green algae and chrysophytes, the three main organisms generating algal blooms are dinoflagellates, diatoms and Cyanobacteria. Dinoflagellates and diatoms are eukaryotic algae that are primarily associated with blooms in seawater, whereas Cyanobacteria are prokaryotic phytoplankton that most commonly produce blooms in freshwater (Sanseverino *et al.*, 2016). Dinoflagellates are unicellular microalgae that are responsible for "red tides" worldwide. Dinoflagellates can form dormant cells called cysts, that settle to bottom sediments where they persist for years to resist environmental and nutritional stresses (Mohamed & Al Shehri, 2011). With the presence of favorable conditions such as warm temperature, cysts will germinate and produce algal blooms (Mohamed & Al Shehri, 2011). Diatoms are also unicellular algae and include approximately 100,000 species worldwide. They have developed several strategies to cope with different environmental stressors and this includes regulation at the proteasome level for modification of metabolism (Muhseen *et al.*, 2015). Both dinoflagellates and diatoms can produce toxins that affect the nervous and intestinal system in humans and are therefore a global concern (Sanseverino *et al.*, 2016). The samples that were collected and analyzed in this work include some eukaryotic blooms but the majority is of cyanobacterial origin and therefore the focus of this work will be on cyanobacterial algal blooms.

1.2 Cyanobacteria

Cyanobacteria are a phylum of oxygenic photosynthesis performing prokaryotes that are often referred to as blue-green algae. It is estimated that Cyanobacteria evolved 2.6-3.5 billion years ago, thereby making them the oldest microorganisms that perform oxygenic photosynthesis and the major contributors of introducing free oxygen into Earth's anoxic atmosphere (Kauff & Budel, 2010, Rasmussen *et al.*, 2008). Cyanobacteria primarily use three multi-subunit light harvesting complexes that convert light into energy-rich compounds: photosystem I, photosystem II and phycobilisomes (Bryant & Frigaard, 2006). Chlorophyll is a photosensitizer found in the thylakoid membrane and is the main photosynthetic pigment used for absorption of light wavelengths (Heinz *et al.*, 2016). Aside from chlorophyll, Cyanobacteria produce the accessory pigments phycoerythrin, phycocyanin and allophycocyanin that are embedded in phycobilisomes to absorb light of different wavelengths (Heinz *et al.*, 2016, Tang *et al.*, 2015).

Cyanobacteria are ubiquitous in nature, inhabiting both terrestrial and aquatic environments as well as thriving in extreme niches such as deserts, hot springs, and in polar regions (Warren-Rhodes *et al.*, 2006, Ionescu *et al.*, 2010, Pandey *et al.*, 2004). Worldwide, Cyanobacteria inhabit aquatic environments that include oceans, rivers, streams, ponds and lakes. In terrestrial environments, Cyanobacteria can be primarily found on the surfaces of rocks, soil and tree bark (Tripathi *et al.*, 2007). They can also occur as endosymbionts in corals, ferns, diatoms, sponges and a number of other organisms (Rai *et al.*, 2002). Cyanobacteria are a diverse group that possess cell envelopes with a combination of gram-positive and gram-negative features. Although they have an overall gram-negative structure, their peptidoglycan layer is thicker than other gram-negative bacteria (Hoiczyk & Hansel, 2000). Additionally, teichoic acid, a constituent of gram-positive cells, is missing in their cell walls (Hoiczyk & Hansel, 2000). Despite all known Cyanobacteria lacking a true flagella, several cyanobacterial species have been shown to exhibit motility either through gliding or twitching across surfaces (Rippka *et al.*, 1979). In freshwater environments, Cyanobacteria are found in one of three morphological groups: 1. Unicells which can be associated together in colonies or may be solitary; 2. Solitary or aggregated undifferentiated, and non-heterocystous filaments; and 3. Differentiated cells called heterocysts (Paerl *et al.*, 2001).

1.3 Heterocysts

In response to nitrogen starvation, a portion of the vegetative cyanobacterial cells in the filament will differentiate into nitrogen fixing cells called heterocysts (Fogg, 1969). Heterocysts are specialized cells that create a microoxic environment for nitrogen fixation and whose formation is inhibited by the availability of nitrate or ammonia (Cai & Wolk, 1997). The only two groups of Cyanobacteria which can differentiate into heterocysts include the two cyanobacterial orders of *Nostocales* and *Stigonematales* (Howard-Azzeh *et al.*, 2014). Nitrogen starvation leads to differentiation of approximately 5 to 10% of the vegetative cells into heterocysts, creating a pattern of single heterocysts that are separated by approximately ten vegetative cells (Gerdtzen *et al.*, 2009). This forms a multicellular organism that is composed of two interdependent cells types (Gerdtzen *et al.*, 2009). Heterocysts within the filaments will provide fixed nitrogen in the form of glutamine and other amino acids to the neighbouring vegetative cells who in turn supply fixed carbon as sucrose through photosynthesis to the heterocysts (Wolk *et al.*, 1976, Cumino *et al.*, 2007). Moreover, heterocyst formation results in morphological and structural changes. Since the nitrogenase enzyme responsible for nitrogen fixation is inactivated by molecular oxygen, heterocysts have evolved to lack a functional photosystem II which is responsible for oxygen production (Thomas, 1970). Overall, heterocysts are distinguished from vegetative cells by their larger and rounder shape, thicker cell envelopes, and diminished pigmentation.

1.4 Cyanobacterial Algal Blooms

Under certain environmental conditions, Cyanobacteria have the ability to rapidly and vigorously proliferate to form an algal bloom. Depending on whether toxigenic cyanobacterial species are present, these blooms have the potential to release toxins (cyanotoxins) into the water and are thereafter called harmful algal blooms (HABs). HABs can occur in marine, estuarine and freshwater ecosystems. The duration and timing of a bloom season depends on the climate of the region. Temperate zones have prominent blooms that can last for 2-4 months during the late summer and early autumn period (van Apeldoorn *et al.*, 2007). Algal bloom seasons start earlier and last longer in subtropical and Mediterranean climates (van Apeldoorn *et al.*, 2007). To this day, the factors that contribute to the formation of blooms are debatable. However, a complex interplay of conditions and nutrients are known to play a major role in their formation.

1.4.1 Temperature, Carbon Dioxide and Nutrients

One important driver to the formation of algal blooms is increased temperature. At approximately 25°C, the photosynthetic capacity, specific respiration rate and growth rate of Cyanobacteria are optimized (Robarts & Zohary, 1987). These elevated temperatures consequently lead to greater effective cyanobacterial competition with eukaryotic primary producers such as chlorophytes and diatoms (Elliott *et al.*, 2006, Johnk *et al.*, 2008). Additionally, unlike eukaryotic phytoplankton species, some species of bloom-forming Cyanobacteria form hollow gas-filled structures called gas vesicles (Walsby, 1994). These gas vesicles regulate cyanobacterial buoyancy by moving the bacteria within a water column either upwards or downwards. A greater advantage to buoyancy is achieved when water is stagnant as opposed to turbulent waters that evenly distribute Cyanobacteria across the water column. With both a temperature increase and stagnant waters, buoyant Cyanobacteria float upwards, leading to a dense biomass of Cyanobacteria at the water surface. These stable stratification conditions in turn lead to the formation of dense surface blooms that shade deep eukaryotic phytoplankton and provide Cyanobacteria with a competitive advantage (Huisman *et al.*, 2004).

Increased water temperatures will additionally decrease the viscosity of water, and thereby decreases water's resistance to vertical migration to enhance cyanobacterial movement upward to optimize photosynthesis (Reynolds, 1987). An increase in light absorption at water surfaces by Cyanobacteria increases surrounding water temperatures to create a positive feedback loop that further propagates bloom dominance (Sonntag & Hense, 2011, Kahru *et al.*, 1993). Moreover, an increase in temperature in freshwater polar ecosystems has been found to alter cyanobacterial diversity where in some cases diversity shifts to toxin-producing species or to elevated toxin production from pre-existing species (Kleinteich *et al.*, 2012).

Dense algal blooms require a high supply of CO_2 to support photosynthetic growth since carbon dioxide is the preferred carbon source (Paerl & Ustach, 1982). The high demand for carbon dioxide increases the pH of the surrounding water and at times, the rate of CO_2 supply can limit rate of algal bloom formation (Paerl & Ustach, 1982). Buoyant Cyanobacteria have a competitive advantage over subsurface phytoplankton populations since the buoyant surface-dwelling Cyanobacteria directly obtain CO_2 from the atmosphere, overcoming the carbon limitation of photosynthetic growth (Paerl & Ustach, 1982).

Anthropogenic-induced nutrient loading of nitrogen and phosphorus due to agricultural, urban and industrial development has led to eutrophication of many freshwater and brackish ecosystems. In turn, eutrophication of the ecosystems promotes the dominance of algal blooms. Phosphorus availability is usually the limiting factor for cyanobacterial growth since some genera of Cyanobacteria can differentiate into heterocysts and bypass the nitrogen requirements (Schindler *et al.*, 2008). It is currently debated whether nitrogen availability is also a limiting factor for growth. For instance, under high phosphorus and low nitrogen conditions, biovolumes of some taxa of Cyanobacteria such as *Cylindrospermopsis raciborskii* and *Aphanizomenon gracile* continually increased with nitrogen concentration indicating a potential for nitrogen limitation (Dolman *et al.*, 2012, Chislock *et al.*, 2014).

1.5 Cyanotoxins

During these cyanobacterial bloom events, some cyanobacterial species are able to produce cyanotoxins which can be toxic to both animals and humans. Over the years, these toxin-producing blooms have been increasing in prevalence worldwide. Within toxic blooms there are distinct cyanotoxin producing species and nonproducing species. However, within cyanotoxin producing species non-toxic strains exist, and not all species that contain these cyanotoxin genes will produce the toxins. Therefore, it is difficult to predict which species will synthesize cyanotoxins and which will not. In general, cyanotoxins are divided into four structurally and functionally diverse groups: neurotoxins, hepatotoxins, cytotoxins and dermatoxins.

1.5.1 Microcystin Hepatotoxin

Microcystin is a cyclic hepatotoxin that is the largest and most structurally diverse cyanotoxin. There are approximately 90 known isoforms of microcystin that vary by degree of hydroxylation, epimerization and methylation (Pearson *et al.*, 2010). Different levels of toxicity have been reported for each microcystin isoform. For instance, microcystin-LR has an LD50 of 50µg/kg in mice (Krishnamurthy *et al.*, 1986) whereas microcystin-RR has an LD50 of 600µg/kg (Watanabe *et al.*, 1988). Microcystin is produced nonribosomally by the microcystin synthetase enzyme complex by multiple

genera of Cyanobacteria including *Anabaena*, *Microcystis*, *Oscillatoria*, and *Chroococcus*. The microcystin biosynthesis gene cluster is approximately 55kb in length and is composed of 10 genes arranged in two divergently transcribed operons: *mcyA-C* and *mcyD-J* **(Figure 1)**. On one end, *mcyD-J* codes for: a polyketide synthase (*mcyD*), two hybrid enzymes composed of polyketide synthase and non-ribosomal peptide synthetase domains (*mcyE* and *mcyG*), tailoring enzymes (*mcyJ*, *mcyF*, and *mcyI*) and a transport enzyme (*mcyH*)(Pearson *et al.*, 2010). On the other hand, *mcyA-C* encodes three non-ribosomal peptide synthetases (*mcyA-C*) (Pearson *et al.*, 2010).

Microcystins exert their toxicity through inhibition of eukaryotic serine/threonine protein phosphatases 1 and 2A, thereby increasing the overall level of phosphorylation in hepatocytes (Runnegar *et al.*, 1995, Eriksson *et al.*, 1990). Increased phosphorylation of cytoskeletal components results in a reorganization of the microfilament network, a breakdown of hepatic ultra structure and an induction of hepatocyte deformation (Eriksson *et al.*, 1990, Runnegar *et al.*, 1981). Consequently, blood begins to collect in liver tissues as hepatocytes begin to retract from neighbouring cells ultimately resulting in tissue damage, liver failure and hemorrhagic shock (Pearson *et al.*, 2010).

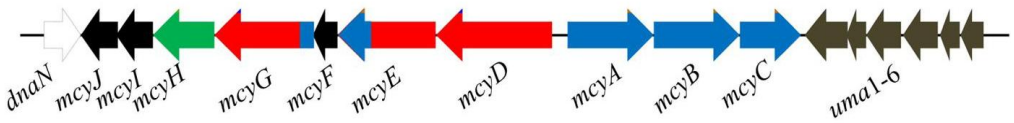


Figure 1. The Microcystin Gene Operon. This figure is adapted from (Rastogi *et al.*, 2015). Black represents tailoring enzymes, red represents polyketide synthases, blue represents non-ribosomal peptide synthetases, dark grey represents non-microcystin synthetase, green is an ABC transporter.

1.6 Public and Ecological Health Effects

1.6.1 Human Exposure to Cyanotoxins

The type and concentration of toxin as well as the route and length of exposure will dictate the health effects and symptoms associated with harmful algal bloom exposure. There are three different routes of cyanotoxin exposure to humans. The first is through the ingestion of shellfish that have bioaccumulated toxins from contaminated water or through cyanobacteria-based food ingredients (Rellan *et al.*, 2009, Bautista *et al.*, 2015). The second route of exposure is through dermal contact and accidental ingestion and inhalation of contaminated waters during recreational activities (Kozdeba *et al.*, 2014, Genitsaris *et al.*, 2011). Finally, ingestion of contaminated drinking water is another route

of exposure and is one that is able to affect a large population of individuals (Hawkins *et al.*, 1985).

1.6.2 Human Health Concerns

Worldwide, cyanotoxins have caused several health-related problems and human poisonings. More specifically, chronic exposure to low concentrations of microcystin through ingestion of contaminated food and water, dermal contact and inhalation has affected the liver, kidneys and colon. Therefore, illnesses related to microcystin poisoning are liver disease, gastroenteritis, as well as allergic and irritation reactions. The first confirmed report of microcystin poisoning in humans was in 1996 in the town of Caruaru, Brazil. Patients were receiving renal dialysis treatment when they began complaining of eye pain, nausea, headaches, blurred vision and vomiting. 76 of these patients died from liver complications that were attributed to cyanotoxin contamination in drinking water (Carmichael *et al.*, 2001). Analysis found microcystins in the liver and blood of the patients and the presence of cylindrospermopsin in the hospital's water purification system (Carmichael *et al.*, 2001). Microcystin exposure has additionally been linked with primary colorectal and liver cancers in human populations (Campos & Vasconcelos, 2010). In 2010, the International Agency for Research on Cancer (IARC) classified microcystin-LR as a possible human carcinogen (Group 2B) (Zegura *et al.*, 2011). Epidemiological studies have shown a direct association between the increased incidence of colon and liver cancer and contaminated drinking water with microcystin in certain parts of China (Zanchett & Oliveira-Filho, 2013). Microcystins are considered to act as tumor promoters through their inhibition of protein phosphatases 1 and 2A, which is directly linked to their tumor-promoting activity (Zegura *et al.*, 2011).

1.6.3 Drinking Water Concerns

Cyanotoxins released in drinking water pose a potential risk to public water supplies and can cause adverse health effects. If blooms affect a population's main or only water supply, this can have detrimental effects on their drinking water. An example of this was the drinking water crisis in May 2007 that took place in Wuxi, Jiangsu Province, China (Qin *et al.*, 2010). A large bloom dominated by *Microcystis spp*. in Lake Taihu left approximately two million people without drinking water for approximately a week (Qin *et al.*, 2010). Additionally, cyanobacterial blooms can release taste and odor compounds that create undesirable drinking water. Most of these taste and odor related events are caused by 2-methylisoborneal and geosmin (Graham *et al.*, 2010). The production of these compounds is strain, rather than species specific (Graham *et al.*, 2017). Some cyanobacterial strains may produce toxins and taste-and-odor compounds simultaneously, whereas other strains may not and the amount and presence of one does not reliably predict the amount and presence of the other (Chorus & Bartram, 1999).

1.6.4 Ecological Impacts

Cyanobacterial bloom events exert their effects at all taxonomic levels including plants, algae and bacteria in aquatic ecosystems. When blooms occur, irradiance and water transparency is reduced resulting in light limitation to plants, benthic algae and phytoplankton (Havens, 2008). Increased photosynthetic activity during intense algal blooms depletes free carbon dioxide from lake water and pH is consequently elevated. This elevated pH results in sub-lethal and lethal impacts on the fish population whereas the reduced carbon dioxide depletes the carbon source for other phytoplankton (Havens, 2008). Aquatic organisms can be affected by cyanotoxins either through direct contact with the toxin, or through ingestion of contaminated water. For instance, when rainbow trout are exposed intraperitoneally or by immersion with algae in water that have microcystin present, the trout dies within 96 hours when 1.440 mg of lyophilized algae/kg of body weight passes through the gills for 18 hours (Tencalla *et al.*, 1994). Cyanotoxins also have adverse effects on wildlife and livestock. The first reported poisoning of livestock was in the 19th century and by the 20th century, cyanotoxin related poisonings in wildlife and livestock had been reported in all continents (Stewart *et al.*, 2008). There have been several mass mortality events attributed to cyanotoxin poisoning. One example was in 2001 that involved 55 cattle in Queensland, Australia that had 1.0mg/L of cylindrospermopsin concentration in rumen and water samples (Saker *et al.*, 1999).

Chapter 2. Comparative Metagenomics of Freshwater Blooms and Non-Blooms in Ontario

2.1 Introduction

There are a wide range of methods that have been developed and are employed for the detection of Cyanobacteria and their associated cyanotoxins. Traditional methods such as microscopic counting have become supplemented with molecular techniques and these have recently become popular in monitoring cyanobacterial blooms. Polymerase-chain reaction (PCR) analysis is the most frequently used molecular technique and it consists of an in-vitro amplification of a DNA sequence that is targeted by specific primers. PCR has a high specificity and sensitivity that warrants its employment in early detection of cyanobacterial blooms and evaluation of toxicity. There have been numerous primers and PCR protocols designed for cyanobacterial and cyanotoxin detection. For instance, PCR primers have been developed to amplify 16S ribosomal RNA (rRNA) genes in Cyanobacteria to detect their presence in a microbial community (Nubel *et al.*, 1997). Similarly, primers have been developed for the toxin synthesizing gene clusters such as those involved in synthesizing microcystins (Singh *et al.*, 2015), nodularin (Moffitt & Neilan, 2001), anatoxin-a (Rantala-Ylinen *et al.*, 2011), cylindrospermopsin (Baron-Sola *et al.*, 2012) and saxitoxin (Kellmann *et al.*, 2009). For microcystin, primers have been developed for the amplification of six gene fragments of the microcystin synthetase mcy cluster: three for the nonribosomal peptide synthetase genes *mcyA*, *mcyB* and *mcyC*, and three to the polyketide synthase genes *mcyD*, *mcyE* and *mcyG* (Hisbergues *et al.*, 2003, Rantala *et al.*, 2006, Vaitomaa *et al.*, 2003). The use of these primers allow for the identification of microcystin-producing strains of the *Anabaena*, *Planktothrix* and *Microcystis* genera. PCR-based approaches have considerable potential in cyanobacterial research since new primers are continuously being developed that become more species and even strain specific. It is important to note that not every toxin producing species will produce toxins, and the presence of toxin genes does not necessarily indicate toxin production. Therefore, PCR-based methods will primarily provide insight into the potential for toxin production within a bloom sample.

Since over 99% of microbes in environments including algal bloom sites cannot be cultured, it is difficult to directly study microbial communities. However, in recent years there have been remarkable advances in Next Generation Sequencing (NGS) technologies. With the use of metagenomic research, culturing microbes has been circumvented and the study of microbial communities, their phylogenetic composition, as well as their metabolic and functional capacity can now be studied more extensively. 16S rRNA sequencing has been widely utilized in analyzing complex bacterial populations since the mid-1990s and is one of the elementary steps used in a metagenomics project

(Muyzer *et al.*, 1993). In this method, a 16S rRNA variable region is amplified using PCR with primers that recognize highly conserved regions of the gene where the amplicons are then sent for sequencing (Sanschagrin & Yergeau, 2014). 16S rRNA genes contain 9 hypervariable regions (V1-V9) that are able to elucidate differential sequence diversity among various bacteria within a community. Despite there being no single hypervariable region that is able to distinguish among all bacteria, hypervariable regions V2, V3 and V4 have the greatest heterogeneity and provide the maximum discrimination between bacterial groups (Chakravorty *et al.*, 2007). Therefore, 16S rRNA sequencing produces hundreds of thousands of 16S rRNA fragments that enable taxonomic analysis into microbial communities simultaneously. However, it is important to be aware that 16S rRNA analysis of metagenomic samples has the limitation of less precise analysis at the genus and species level (Ranjan *et al.*, 2016, Quince *et al.*, 2009). An alternative and supplementary approach to 16S rRNA amplicon sequencing is shotgun metagenomic sequencing. Instead of relying on a single diagnostic gene, this approach provides a global analysis of the microbial community, allowing for a more in-depth analysis of phylogenetic diversity, metabolic pathways and functional capacities (Chen & Pachter, 2005). This allows for the discovery of new genes that are recently diverged and that cannot be amplified using PCR (Chen & Pachter, 2005). The major advantage of shotgun metagenomic sequencing is that taxa can be more precisely and accurately classified at the genus and species level (Ranjan *et al.*, 2016). Additionally, it can be used to characterize functional and metabolic pathways associated with samples and microbial communities. However, it is more expensive than amplicon sequencing and requires a more comprehensive analysis. It is important to note that amplicon sequencing and shotgun metagenomic sequencing utilize different databases for classification and therefore there will be some variation in the results from both methods (Ranjan *et al.*, 2016). Overall, amplicon sequencing and shotgun metagenomic sequencing are important tools that have allowed for a more extensive analysis of microbial communities in several different environments.

One environment that is of upmost importance is freshwater ecosystems. Freshwater ecosystems have been increasing in incidence with cyanobacterial blooms over the years. These cyanobacterial blooms have been linked to the deterioration of freshwater ecosystems since they have been known to be associated with numerous public and ecological health concerns that affect both wildlife and humans. Therefore, the increasing incidence of cyanobacterial blooms has led to an urgent need for efficient and reliable monitoring. The majority of microorganisms such as those in cyanobacterial bloom communities that cannot be cultured can be studied utilizing DNA sequencing approaches. Therefore, in collaboration with the Ministry of the Environment and Climate Change (MOECC), a comparative metagenomics study was initiated to examine blooms and non-blooms from Ontario freshwater lakes and ecosystems. The primary objective of this project was to identify differences that may exist on a taxonomic and functional level between bloom and non-bloom sites as well as to analyze bloom communities to build upon and advance our previous knowledge of cyanobacterial communities.

2.2 Questions

1. What unique microbial consortia are associated with bloom events and non-bloom sites?
2. Are there any differences in functional capacities between bloom and non-bloom sites? For instance, any differences in Nitrogen and Phosphorus metabolism?
3. Within bloom sites, is there temporal variation for cyanobacterial abundance?
4. Within cyanobacterial bloom sites, what are the predominant genera that comprise the cyanobacterial community and are responsible for bloom formation?
5. Is there microcystin toxin producing potential at bloom and non-bloom sites?

2.3 Methods

2.3.1 Study Sites and Sample Collection

Between February and October 2015, 108 water samples from across Ontario, Canada were collected by the Ministry of the Environment and Climate Change (MOECC). Of these samples, 101 were sampled from freshwater lakes, 3 from ponds, 3 from rivers and 1 sample was taken from a creek. These samples were comprised of 60 water samples that were taken from bloom sites and 48 water samples taken from sites that had no blooms present and are therefore referred to as non-blooms throughout this work. It is important to mention that "bloom" throughout this work refers to the presence of an algal bloom that can be caused by eukaryotic organisms or Cyanobacteria. The water samples were collected in 500mL sampling bottles and once all the sampling was completed, the MOECC sent the water samples in coolers in October 2015. Upon arrival, samples were stored at -80°C until DNA extraction. **Table 1** and **Table 2** outline the metadata for the bloom and non-bloom samples, respectively.

There were no samples that were replicates and all samples were taken from different locations. Some samples were taken from the same body of water but were not sampled in the same exact location. Therefore, sites that were sampled on the same day were not taken from the same location and each sample is represented as being a single sample for that particular location.

It is important to note that the samples have been coded and the precise locations are not disclosed in order to maintain anonymity in accordance with the MOECC's regulations. Additionally, the MOECC underwent a preliminary microscopic analysis on each of the samples to determine whether the bloom was caused by Cyanobacteria or eukaryotic organisms. Not all bloom samples were cyanobacterial in origin. Therefore, blooms that are not caused by Cyanobacteria will be briefly examined, but only cyanobacterial blooms will be studied extensively.

Table 1. Metadata of Bloom Samples

Sample Code	Sampling Date	Sample Code	Sampling Date
ON6	June 4, 2015	ON59	July 29, 2015
ON8	June 16, 2015	ON64	July 31, 2015
ON9	June 22, 2015	ON65	August 5, 2015
ON10	June 19, 2015	ON66	August 5, 2015
ON14	June 25, 2015	ON68	August 7, 2015
ON15	June 25, 2015	ON70	August 7, 2015
ON16	June 29, 2015	ON71	August 11, 2015
ON17	June 30, 2015	ON72	August 12, 2015
ON19	June 29, 2015	ON73	August 12, 2015
ON20	July 5, 2015	ON76	August 18, 2015
ON21	July 6, 2015	ON81	August 25, 2015
ON22	July 6, 2015	ON84	August 25, 2015
ON23	July 6, 2015	ON85	August 31, 2015
ON24	July 8, 2015	ON86	August 31, 2015
ON25	July 8, 2015	ON87	August 31, 2015
ON26	July 8, 2015	ON88	August 31, 2015
ON27	July 8, 2015	ON89	September 2, 2015
ON29	July 9, 2015	ON92	September 8, 2015
ON30	July 8, 2015	ON93	September 8, 2015
ON33	July 13, 2015	ON97	September 10, 2015
ON35	July 13, 2015	ON98	September 17, 2015
ON39	July 15, 2015	ON99	September 21, 2015
ON44	July 17, 2015	ON100	September 21, 2015
ON46	July 20, 2015	ON101	September 22, 2015
ON47	July 22, 2015	ON103	September 21, 2015
ON48	July 22, 2015	ON104	September 28, 2015
ON49	July 22, 2015	ON105	September 28, 2015
ON54	July 27, 2015	ON106	October 5, 2015
ON55	July 27, 2015	ON107	October 6, 2015
ON56	July 28, 2015	ON108	October 6, 2015

Table 2. Metadata of Non-Bloom Samples

Sample Code	Sampling Date
ON1	February 25, 2015
ON2	March 25, 2015
ON3	April 13, 2015
ON4	April 15, 2015
ON5	May 28, 2015
ON7	June 9, 2015
ON11	June 24, 2015
ON12	June 24, 2015
ON13	June 24, 2015
ON18	June 30, 2015
ON28	July 9, 2015
ON31	July 7, 2015
ON32	July 7, 2015
ON34	July 13, 2015
ON36	July 10, 2015
ON37	July 8, 2015
ON38	July 13, 2015
ON40	July 16, 2015
ON41	July 16, 2015
ON42	July 16, 2015
ON43	July 15, 2015
ON45	July 20, 2015
ON50	July 21, 2015
ON51	July 21, 2015
ON52	July 21, 2015
ON53	July 21, 2015
ON57	July 28, 2015
ON58	July 28, 2015
ON60	July 30, 2015
ON61	July 30, 2015
ON62	July 30, 2015
ON63	July 24, 2015
ON67	July 29, 2015
ON69	August 10, 2015
ON74	August 12, 2015
ON75	August 17, 2015
ON77	August 18, 2015
ON78	August 19, 2015
ON79	July 27, 2015
ON80	July 27, 2015
ON82	August 25, 2015
ON83	August 25, 2015
ON90	August 31, 2015
ON91	September 3, 2015
ON94	September 8, 2015
ON95	September 8, 2015
ON96	September 8, 2015
ON102	September 22, 2015

2.3.2 DNA Extraction

Approximately 5.0 mL of each water sample was centrifuged at 10,000 rpm for 15 minutes. The supernatant was then discarded and the pellet was used for DNA extraction. DNA extractions were performed using the Norgen Biotek Corp Soil DNA Isolation Kit (Norgen Biotek Corporation, Canada). Manufacturer's protocol was followed, with the addition of 10µg/mL of lysozyme and 200mM of β-mercaptoethanol to maximize lysis of cyanobacterial cells (Mehta *et al.*, 2015). Approximately 150µL of DNA was extracted for each sample. Extracted DNA was stored at -20°C until further use for library preparation, sequencing and polymerase-chain reactions.

2.3.3 Preparation of Libraries and Sequencing

DNA extracted from all water samples were subjected to amplicon sequencing and a subset of these samples were subjected to shotgun metagenomic sequencing. For amplicon sequencing, the V3-V4 region of the 16S rRNA gene was amplified using the primer pair P5 and P7 (Illumina Inc., San Diego, CA. USA). 25µL PCR reaction mixes were prepared into 96 well plates (BioRad) with the following: 2.5µL of 1.0µM 500S forward and 700S reverse primers that had unique adapter sequences attached to them (Illumina), 0.5µL of 100mM dNTP solution (Invitrogen), 2.5µL of 10X Reaction Buffer (Invitrogen), 1.0µL of 10mg/mL BSA in ddH_2O (UV irradiated for 10 minutes), 0.25µL Taq recombinant polymerase (Invitrogen) and 2.0µL of template DNA. The PCR reactions were run with the CF96 Touch Real-Time PCR Detection System (BioRad) with the following conditions: Initial denaturation at 95°C for 3 minutes, 40 cycles of denaturation at 94°C for 30 seconds, annealing at 50°C for 30 seconds, and elongation at 72°C for 60 seconds that was followed by a final extension at 72°C for 10 minutes. All PCR products were loaded and run on a 1.0% agarose gel to ensure presence of expected amplicons. Once confirmed, 5.0µL of each amplicon product was pooled into a single 1.7mL microcentrifuge tube and sent to the Farncombe Metagenomics Facility (McMaster University, Hamilton, Ontario, Canada) for amplicon sequencing with the Illumina MiSeq platform.

For whole-genome shotgun sequencing, 24 samples were selected where 12 were bloom samples and 12 were non-bloom samples. Within the 12 bloom samples, 6 were assigned as heavy blooms due to the blue-green nature of the sample indicating high cell density and the other 6 were assigned as light blooms due to their transparent color indicating low cell density. All 12 bloom samples were determined to be caused by Cyanobacteria by the MOECC using microscopy. Within the 12 non-bloom samples, 6 were chosen that had greater than 10% cyanobacterial abundance according to 16S rRNA sequencing data and 6 were chosen that had less than 10% cyanobacterial abundance. This method of choosing which samples to send for shotgun metagenomic sequencing was utilized to capture the microbial diversity as accurately as possible and to avoid any

biases that may exist within the two environments. DNA was diluted to 0.2ng/µL and prepared for sequencing using the Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA). The samples were sent to the Farncombe Metagenomics Facility (McMaster University, Hamilton, Ontario, Canada) for 250bp paired-end sequencing on the Illumina HiSeq 1500 platform.

2.3.4 Bioinformatic Analysis

Prokaryotic 16S rRNA reads were processed using the software package Quantitative Insights into Microbial Ecology (QIIME) v.1.9.1 using default parameters except where specified (Caporaso *et al.*, 2010). A custom script developed for QIIME by a previous student in the lab, Steven Botts, was used (Botts, 2016). Reads were first quality trimmed with a quality score cut-off of 25. Quality filtered sequences were then clustered into OTUs according to a 97% sequence similarity threshold against the reference database. In this case, the May 2013 release of the Greengenes reference database was used (DeSantis *et al.*, 2006). The QIIME script was also utilized to perform statistical analyses to visualize microbial communities. In addition to the custom script, further analyses were conducted using QIIME. The 2D principal coordinate (PCoA) plot was calculated using the beta_diversity.py and principal_coordinates.py python scripts.

Raw reads obtained from the Illumina HiSeq platform were quality screened using FastQC (v.0.11.3)(Andrews, 2010) to detect adapter sequences and low-quality reads. Trimmomatic (v.0.36) (Bolger *et al.*, 2014) was used for adapter removal and quality trimming. Trimmomatic's *Nextera-PE* template was used for adapter clipping and a quality score of 25 was selected as the quality score threshold. Overlapping paired-ended reads were then merged using Fast Length Adjustment of Short Reads (FLASh v.1.2.11) (Magoc & Salzberg, 2011). Default parameters were used for merging with the exception of the minimum overlap set to 15bp. Processed reads were again quality tested using FastQC and reads that passed were retained for further analysis. For annotation, the processed reads were aligned against the non-redundant NCBI protein database using the program DIAMOND (v.0.8.38)(Buchfink *et al.*, 2015) with an e-value cutoff of $1x10^{-5}$. Read alignments obtained from DIAMOND were parsed and classified using MEGAN (v.6.7.0)(Huson *et al.*, 2007). Default parameters were used for MEGAN and both functional and taxonomic classification were used for assignment. For functional assignment, EggNOG (October 2016 Release)(Jensen *et al.*, 2008) and SEED subsystems (May 2015 Release)(Bao *et al.*, 2011) databases were used and for taxonomic assignment the NCBI taxonomy (Sayers *et al.*, 2009) was used.

2.3.5 Statistical Analysis

Statistical analyses of shotgun metagenomic sequencing data were done in R (v.3.2.3). Principal Coordinate Analysis (PCoA) plots and rarefaction curves were prepared using the VEGAN package (v.2.4-2)(Dixon, 2003). DESeq2 (v.1.14.1)(Love *et al.*, 2014) was used for testing differential abundance of taxonomic and functional count data.

2.3.6 Toxin Polymerase Chain Reaction

Microcystin synthetase gene E (*mcyE*) forward primer and reverse primers specific to *Microcystis spp.* **(Table 3)** were used to test all 108 samples for the presence/absence of the *mcyE* toxin gene. The PCR was performed with 1.0µL of extracted DNA, 1X ThermoPol Reaction Buffer (New England BioLabs), 250µM concentration of deoxynucleotide solution mix (New England BioLabs), 0.5µM concentration of primers (Integrated DNA Technologies), and 0.5 U of Taq DNA Polymerase (New England BioLabs) in a final volume of 20µL. PCR amplification was performed as follows (Vaitomaa *et al.*, 2003). The first step was an initial denaturation step of 3 minutes at 95°C followed by 25 cycles of PCR, with 1 cycle consisting of 30 seconds at 94°C, 30 seconds at 60°C and 60 seconds at 72°C. It was followed by a final extension of 10 minutes at 72°C. The PCR products were then run on a 1.5% agarose gel to test the presence or absence of the *mcyE* gene product. The expected band size was 247 bp. The bands were stained with ethidium bromide for visualization and bands were manually checked. PCR was performed in duplicate to reduce false-positives and false-negatives.

Table 3. Microcystin Synthetase Gene E Primers

	Primer Sequence (5' →3')	Primer Length (bp)	Reference
***mcyE* forward primer**	GAA ATT TGT GTA GAA GGT GC	20	(Vaitomaa *et al.*, 2003)
***mcyE* reverse primer**	CAA TGG GAG CAT AAC GAG	18	(Vaitomaa *et al.*, 2003)

2.4 Results

2.4.1 Amplicon Sequencing Analysis of 108 Bloom and Non-Bloom Sites

Taxonomic Composition

To examine if any differences exist in the microbial communities between bloom and non-bloom sites, the taxonomic composition at the phylum level was first examined **(Figure 2).** In general, heterogeneity was exhibited between and among the two environments but some trends were noted. In both environments, Proteobacteria were the most abundant phyla. Other predominant phyla in the two groups were Bacteroidetes, Actinobacteria, Cyanobacteria, Firmicutes and Verrucomicrobia. Together, the Proteobacteria, Bacteroidetes and Actinobacteria comprised 80-90% of the abundance in most of the samples in both groups. More specifically, there were a few samples within both environments that stood out and are of interest. There were eight bloom samples that had greater than 10% relative abundance of Verrucomicrobia in contrast to the one non-bloom sample that had greater than 10% relative abundance. This substantiates previous literature where Verrucomicrobia is associated with algal blooms and high-nutrient environments (Haukka *et al.*, 2006, Kolmonen *et al.*, 2004, Louati *et al.*, 2015). The occurrence of Verrucomicrobia with bloom sites suggests that Verrucomicrobia thrive in nutrient-rich environments. Since many Verrucomicrobia are prosthecate, this provides an advantage in nutrient uptake allowing for their prevalence in eutrophic bloom conditions (Zwart *et al.*, 1998, Haukka *et al.*, 2006). Therefore, this finding provides the initial support that bloom and non-bloom environments differ in their microbial communities. Moreover, five bloom samples had greater than 10% relative abundance of Planctomycetes whereas none of the non-bloom samples had above 10% relative abundance. It has been previously shown that Planctomycetes abundance is positively correlated with algal blooms (Pizzetti *et al.*, 2011, Eiler & Bertilsson, 2004). Planctomycetes are found in diverse environments such as lakes and soil in abundances of approximately 5% of total bacterial counts (Eiler & Bertilsson, 2004). These heterotrophic aerobic bacterial cells are often attached to particulate matter or in cell clusters (Eiler & Bertilsson, 2004). A possible ecological role of Planctomycetes in blooms is their degradation of sulfated polysaccharides produced by Cyanobacteria (Cai *et al.*, 2013). Therefore, the association of Planctomycetes abundance with bloom sites further illustrates the different microbial communities associated with bloom and non-bloom environments.

It is important to point out that there is an absence of metadata that could have an impact on microbial community structure. Such examples of missing parameters include nutrient levels, proximity to agricultural runoff, weather forecast, as well as depth and

region of water body. Nonetheless, it is imperative that an analysis into the microbial communities be completed despite the lack of in-depth metadata.

In terms of Cyanobacteria, bloom sites had a greater overall relative abundance of Cyanobacteria when compared with non-bloom sites. The relative abundance range of Cyanobacteria was 58.68-0.03% and 32.34-0.03% in bloom and non-bloom sites, respectively. Variability in cyanobacterial abundance in bloom samples can possibly be explained due to different causative agents of the blooms. Not all bloom samples were caused by Cyanobacteria and other taxa could have initiated the formation of the bloom. Examples of these taxa include chrysophytes, cryptophytes, dinoflagellates, euglenophytes, xanthophyceae and diatoms. These can be seen in sampling sites with minimal cyanobacterial relative abundance closer to the right side of the figure. However, on the left side of the bloom figure, the high relative abundance of Cyanobacteria reveals the causative agents of those blooms to possibly be of cyanobacterial origin. Surprisingly, there were some non-bloom sites that had a high proportion of Cyanobacteria present. This could be explained due to the sample being collected in close proximity to a bloom but being labeled as a non-bloom site.

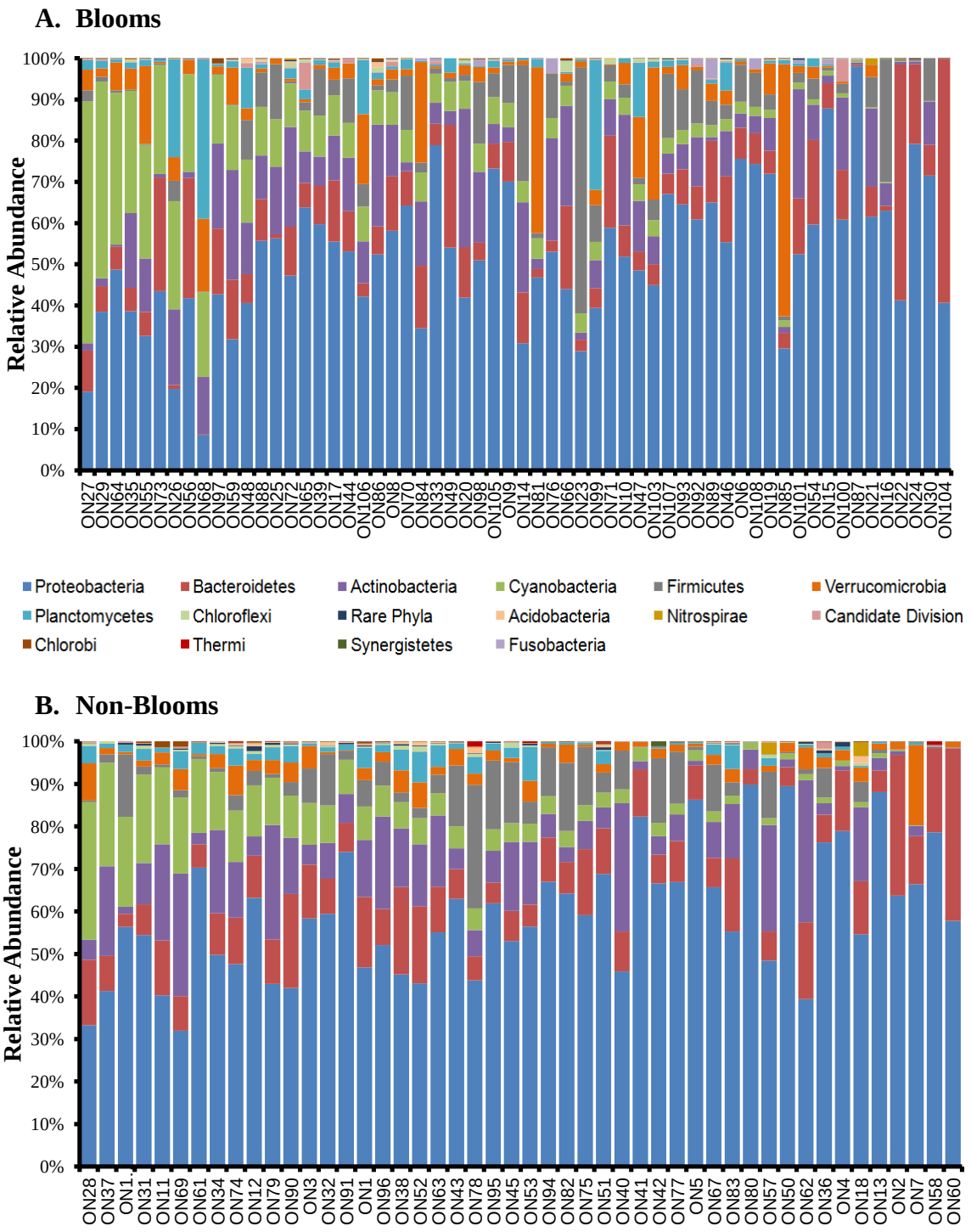


Figure 2. Relative Abundance of Microbial Community Phyla. Abundance of bacterial phyla in **A.** Bloom Sites and **B.** Non-Bloom Sites. Samples were ranked from highest to lowest cyanobacterial relative abundance. Phyla that were under 1% abundance were labeled as Rare Phyla.

A further examination into the cyanobacterial group was conducted by comparing the cyanobacterial community at the order level between the two environments (**Figure 3).** Overall, there was a lot of cyanobacterial diversity with the three cyanobacterial orders Oscillatoriales, Chroococcales, and Synechococcales as the most abundant in both environments. However, across bloom sites Oscillatoriales had the greatest proportion whereas Synechococcales had the greatest proportion across non-bloom sites. Although not all blooms were caused by Cyanobacteria, it is still important to dissect some of the differences that are seen in the samples with high cyanobacterial abundance.

There were four bloom sites (ON73, ON35, ON56, ON81) that had greater than 20% of reads assigned to Oscillatoriales as opposed to the one non-bloom site (ON102). Toxin-producing filamentous *Planktothrix spp.* and *Lyngbya spp.* belong to the Oscillatoriales group and this could be an indication of possible toxin production within those bloom samples that had a high proportion of Oscillatoriales. There were two bloom sites that had greater than 10% of reads assigned to Synechococcales (ON68, ON26). Although across the non-bloom sites the Synechococcales group was the most abundant, none of the sites had greater than 10% of reads assigned to it. The Synechococcales are a picophytoplankton group that have over 70 genera and are a group of unicellular Cyanobacteria. They are widespread in marine environments but are also found in freshwater environments. Most studies have focused on marine lineages, leaving freshwater strains to be poorly characterized (Coutinho *et al.*, 2016). Nonetheless, Synechococcales dominance in freshwater ecosystems is gaining attention and this study provides an initial insight into the high abundance of Synechococcales in freshwater ecosystems. Additionally, there were three bloom sites and two non-bloom sites that had greater than 20% abundance of Chroococcales. This is interesting since the toxin producing and bloom forming *Microcystis spp.* belongs to the Chroococcales group and its presence at a high proportion at some sites is a possible indicator of toxicity. Although Stramenopiles, Cryptophyta, Chlorophyta, Streptophyta, Euglenozoa and Haptophyceae are eukaryotic and belong to the Chloroplast class, they do contain 16S rRNA and that is the explanation for their appearance in the figures. Some of these eukaryotes such as Chlorophyta are divisions of green algae. Therefore, bloom samples such as ON97 that have a high proportion of Chlorophyta reads represent a possibility that this site's bloom was dominated by the eukaryote rather than by Cyanobacteria.

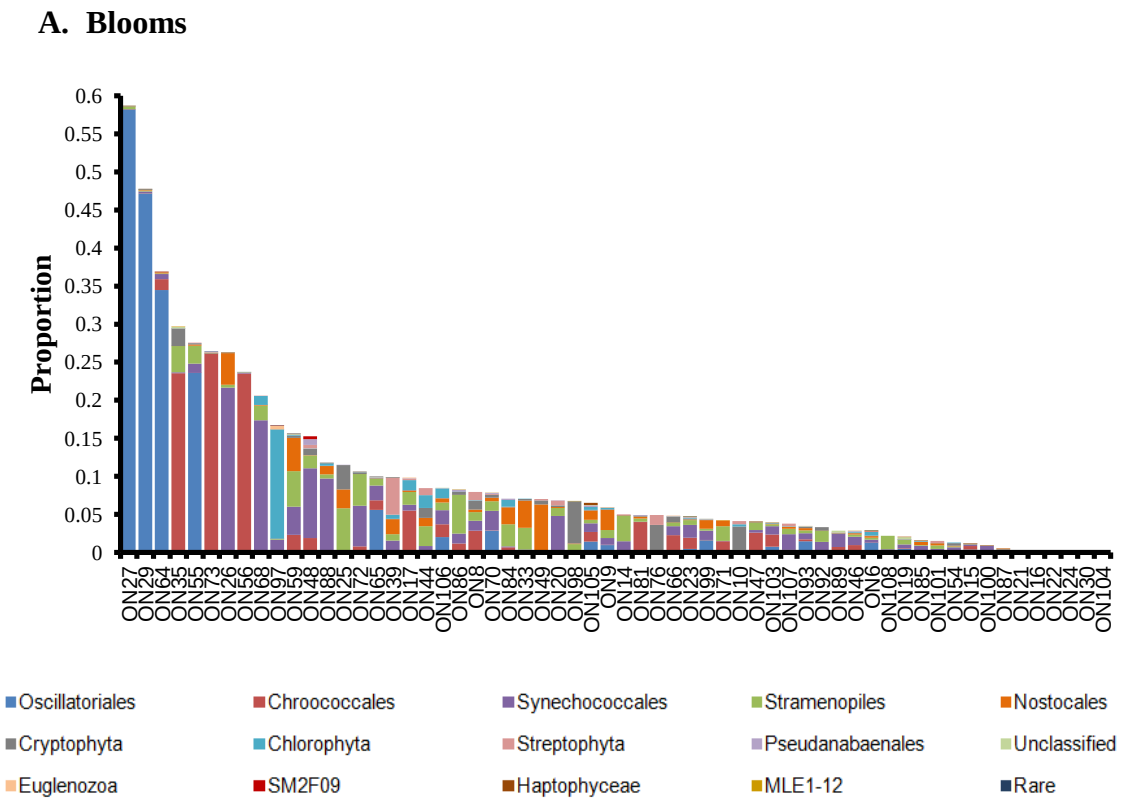


Figure 3. Proportion of Reads Assigned to the Cyanobacterial and Chloroplast Order. Reads were annotated at the cyanobacterial order level at **A.** Bloom Sites and **B.** Non-Bloom Sites. Rare represents <1% abundance of cyanobacterial order reads.

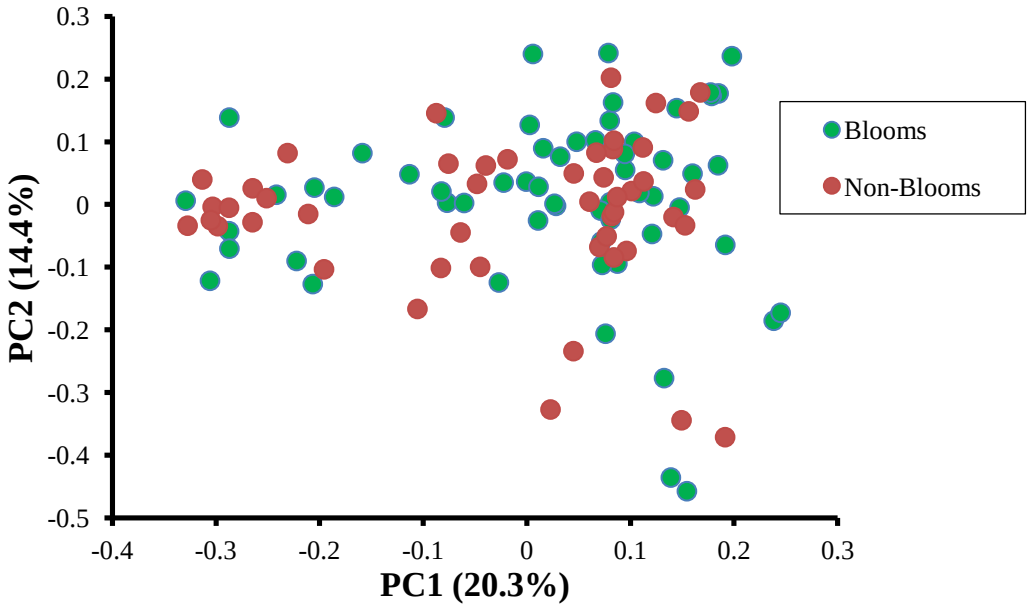
Diversity Analysis

To further address whether differences in the microbial communities exist between the two groups, a Principal Coordinates Analysis (PCoA) plot was calculated. PCoA is a metric multidimensional scaling method that places the samples on a Euclidean coordinate space. The purpose of the plot is to represent each sample as a point in space so that the distance between any two points represents the dissimilarity between the samples that they represent. Depending on the distance matrix calculated, a graphical representation is created in a low-dimensional (usually two or three) Euclidean space so that the distances between the points in the plot will reflect the original distances between the samples. There are several different distance matrices that can be used for a PCoA plot. Some examples include the Bray-Curtis dissimilarity and UniFrac measure.

To measure the diversity that may exist between bloom and non-bloom sites, beta-diversity was calculated using the beta-diversity measure UniFrac and visualized with a PCoA plot. UniFrac is widely applied in microbial ecology and uses phylogenetic information to compare environmental samples (Lozupone & Knight, 2005). It measures the phylogenetic distance between groups of taxa within a phylogenetic tree as the fraction of the branch length of the tree that results in descendants from not both environments, but from either one or the other environment (Lozupone & Knight, 2005). There are two standard UniFrac calculations: Weighted and Unweighted. Unweighted UniFrac considers only the presence or absence of the taxa between samples whereas Weighted UniFrac takes into account the differences in the abundance of taxa between samples (Lozupone & Knight, 2005, Lozupone *et al.*, 2007). Since UniFrac fulfills the requirements as a distance metric, it is used in multivariate statistics such as in PCoA (Lozupone & Knight, 2005). UniFrac is an advantageous diversity measure since it analyzes sequence similarities that may exist between the samples.

Weighted UniFrac PCoA plots were created to compare bloom and non-bloom sites at the taxonomic level. **Figure 4A** highlights the lack of sample clustering based on bloom and non-bloom sites. This suggests that phylogenetically, there is no clear distinction between the two environments and that there are other underlying factors that might contribute to differences. Therefore, a weighted UniFrac PCoA plot that looked at monthly variations in both groups was created to see if a temporal basis could account for any diversity based on sampling month and irrespective of its environment **(Figure 4B).** There were no clear differences on a monthly basis further supporting the presence of other factors that might explain variations between the samples such as nutrient levels, depth of sampling, weather forecast, and geographical location. In subsequent sections of this work, another distance matrix was utilized to analyze if the distance matrix has an effect on the results as well.

A. Bloom vs. Non-Bloom Sites



B. Monthly

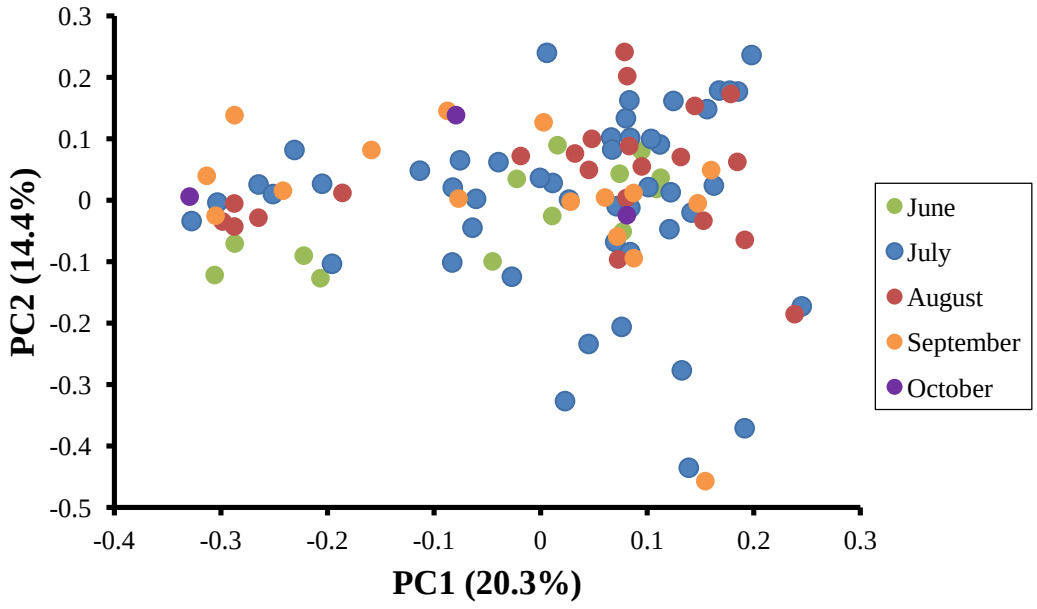


Figure 4. PCoA Plot Utilizing Weighted UniFrac Measure for Bloom and Non-Bloom Sites. Principal Coordinate Analysis calculated with Weighted UniFrac measure **A.** between Bloom and Non-Bloom Sites and **B.** on a Monthly Basis. Months that had less than 2 samples were omitted from the figure.

Temporal Variation: Bloom Sites

Next, only bloom sites were examined and compared with each other to analyze any temporal variation that may exist within these sites. The taxonomic phylum level was compared and samples were divided into their respective months **(Figure 5).** February, March, April and May had two or less samples and therefore were not included in the figure. Overall, there was a high percent abundance of Proteobacteria across the months. In July and August there was a high abundance of Cyanobacteria present in some of the samples. This provides insight into cyanobacterial abundance being the greatest in July and August and that most of these sites contained blooms that were caused by Cyanobacteria. This coincides with the fact that cyanobacterial blooms on average tend to occur near the end of summer months in Ontario as previously described (van Apeldoorn *et al.*, 2007).

Temporal variation within bloom sites was analyzed using a PCoA plot next **(Figure 6).** The weighted UniFrac PCoA plot highlighted a lack of variation on a monthly basis between the sites illustrating no phylogenetic differences between the sites based on month of sampling and the possibility of underlying factors to explain any variation.

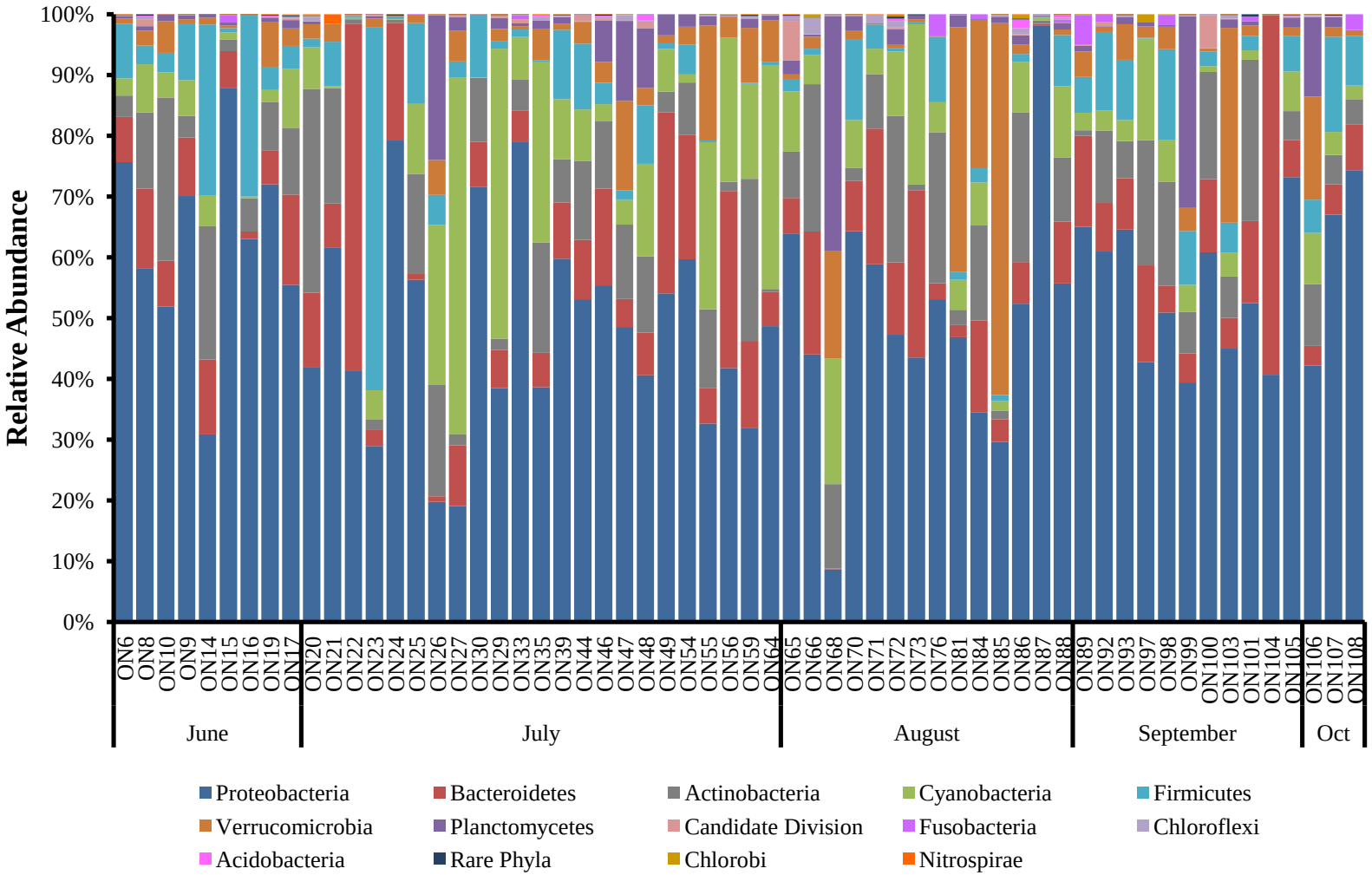


Figure 5. Taxonomic Composition of Bloom Samples on a Monthly Basis. Bloom samples were divided into their respective month of sampling. Rare Phyla represents phyla that are less than 1% abundance across all samples.

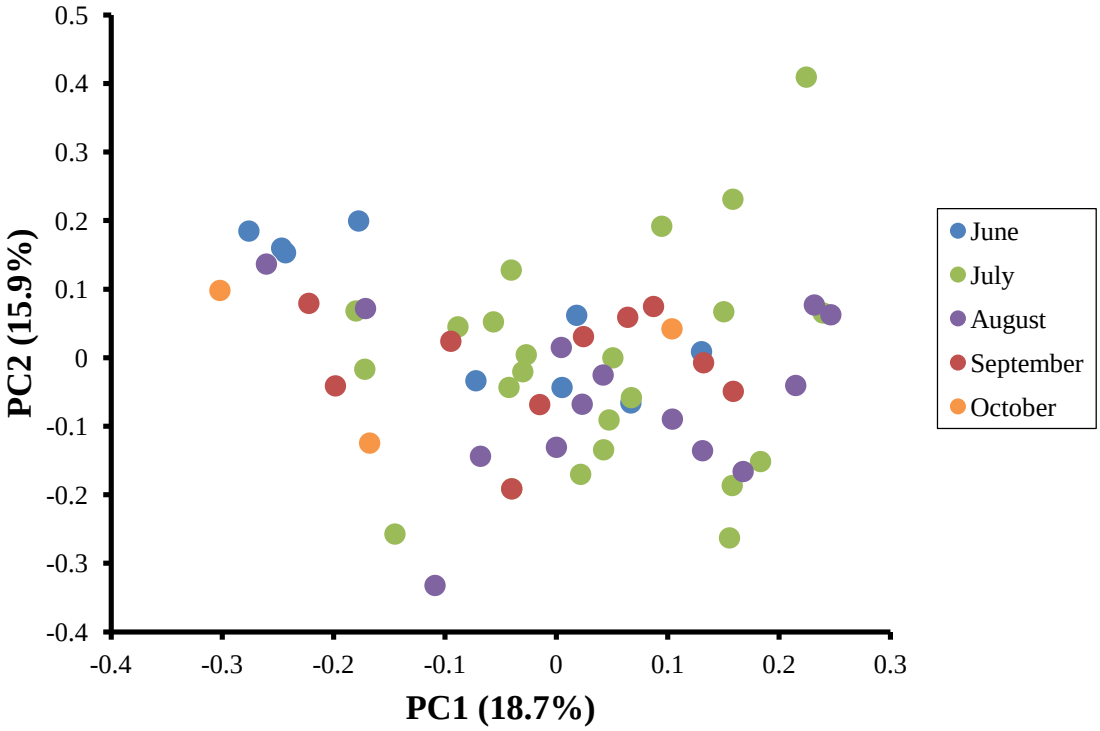


Figure 6. PCoA Plot Utilizing Weighted Unifrac Measure of Bloom Sites. Weighted UniFrac PCoA plot of bloom sites on a monthly basis. Months that had less than 2 samples were omitted from analysis.

Microcystin Toxin Production Potential

The potential for all examined sites to produce the microcystin toxin was analyzed by looking at the presence of the *mcyE* gene of the microcystin operon. Overall, only four samples were positive for the *mcyE* gene and all four were bloom samples **(Table 4 & Table 5).** On the other hand, none of the non-bloom sites tested positive for the *mcyE* gene. A temporal variation was observed with July samples having the greatest toxin production potential. February, March, April, May, June, and September had no samples test positive for the *mcyE* gene. This does not imply that the toxin gene is not present at other sites. It could be that the *mcyE* gene is not present in the operon in some species and therefore is not detected (Vaitomaa *et al.*, 2003) or were in concentrations too low to be detected. Additionally, the primers were designed specific to *Microcystis spp.* and would not detect the potential of *Planktothrix spp.*, and *Anabaena spp.* to produce the toxin.

Table 4. PCR Toxin Gene Results at Bloom Sites

Sample	Month of Sampling	*mcyE* Presence
ON6	June	-
ON8	June	-
ON9	June	-
ON10	June	-
ON14	June	-
ON15	June	-
ON16	June	-
ON17	June	-
ON19	June	-
ON20	July	-
ON21	July	+
ON22	July	-
ON23	July	-
ON24	July	-
ON25	July	-
ON26	July	-
ON27	July	-
ON29	July	-
ON30	July	-
ON33	July	-
ON35	July	-
ON39	July	-
ON44	July	-
ON46	July	-
ON47	July	+
ON48	July	-
ON49	July	+
ON54	July	-
ON55	July	-
ON56	July	-

Sample	Month of Sampling	*mcyE* Presence
ON59	July	-
ON64	July	-
ON65	August	-
ON66	August	-
ON68	August	-
ON70	August	-
ON71	August	+
ON72	August	-
ON73	August	-
ON76	August	-
ON81	August	-
ON84	August	-
ON85	August	-
ON86	August	-
ON87	August	-
ON88	August	-
ON89	September	-
ON92	September	-
ON93	September	-
ON97	September	-
ON98	September	-
ON99	September	-
ON100	September	-
ON101	September	-
ON103	September	-
ON104	September	-
ON105	September	-
ON106	October	-
ON107	October	-
ON108	October	-

Table 5. PCR Toxin Gene Results at Non-Bloom Sites

Sample	Month Of Sampling	*mcyE* Presence
ON1	February	-
ON2	March	-
ON3	April	-
ON4	April	-
ON5	May	-
ON7	June	-
ON11	June	-
ON12	June	-
ON13	June	-
ON18	June	-
ON28	July	-
ON31	July	-
ON32	July	-
ON34	July	-
ON36	July	-
ON37	July	-
ON38	July	-
ON40	July	-
ON41	July	-
ON42	July	-
ON43	July	-
ON45	July	-
ON50	July	-
ON51	July	-

Sample	Month of Sampling	*mcyE* Presence
ON52	July	-
ON53	July	-
ON57	July	-
ON58	July	-
ON60	July	-
ON61	July	-
ON62	July	-
ON63	July	-
ON67	July	-
ON79	July	-
ON80	July	-
ON69	August	-
ON74	August	-
ON75	August	-
ON77	August	-
ON78	August	-
ON82	August	-
ON83	August	-
ON90	August	-
ON91	September	-
ON94	September	-
ON95	September	-
ON96	September	-
ON102	September	-

2.4.2 Shotgun Metagenomic Sequencing Analysis of 24 Bloom and Non-Bloom Sites

Sample Statistics

The primary focus of this work was to analyze if any differences may exist between bloom and non-bloom environments. As such, it was imperative to ensure that there were no sequencing biases that may have occurred that could affect downstream analyses and taxonomic and functional classification. **Table 6** was devised to examine the number of reads at each of the processing and classification stages.

Table 6. Summary Statistics of Shotgun Metagenomic Sequencing Data

Read Processing	Blooms	Non- Blooms	Average
Pre-Trim Reads	5,846,086	8,833,478	7,339,782
Post-Trim Reads	5,420,303	8,121,162	6,770,732
Post-Merge Reads	2,845,031	4,284,222	3,564,626
Taxonomic Classification Reads	1,262,314	1,377,925	1,320,119
Functional Assignment (EggNOG) Reads	659,547	839,033	749,290
Functional Assignment (SEED Subsystems) Reads	485,321	594,863	540,092

On average there were 7,339,782 reads before processing and 6,770,732 reads after trimming of adapter sequences and low quality reads. This constituted an approximate 7.75% loss of reads after trimming of sequences, illustrating a relatively high quality of reads overall. After merging of pair-ended reads, there was on average 52.6% of reads that merged and combined into larger read lengths. Despite a lower average of reads before and after processing in bloom samples, there was no significant difference between the two environments (t-test, $p > 0.05$). **Table 7** illustrates taxonomic and functional classification proportions that were assigned. Although there is a slightly greater proportion of reads assigned to the taxa level in bloom samples, there was no significant difference in the proportion of reads that were assigned taxonomically or functionally between the two environments (t-test, $p > 0.05$).

Table 7. Statistics of Taxonomic and Functional Proportions

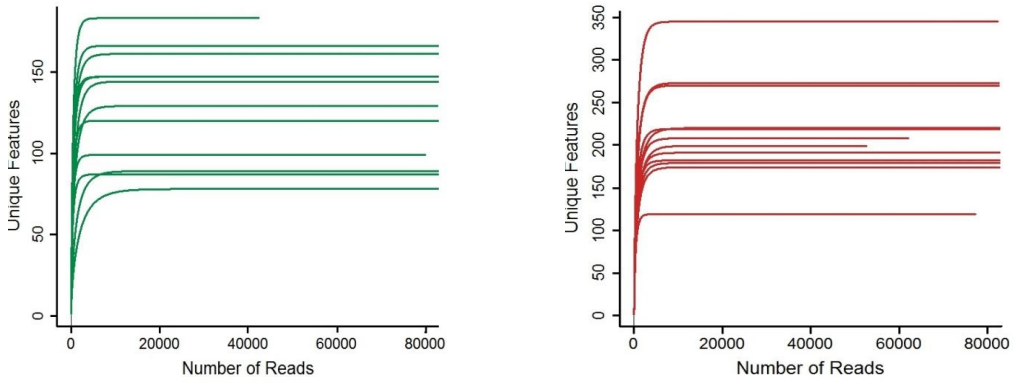
	Blooms	Non-Blooms	Average
Proportion Assigned Taxa	0.44	0.32	0.38
Proportion Assigned Functions (EggNOG)	0.23	0.20	0.22
Proportion Assigned Functions (SEED Subsystems)	0.17	0.14	0.16

Rarefaction Curves

To estimate coverage, a rarefaction curve is widely utilized. A rarefaction curve plots the number of unique features (for instance, operational taxonomic units, predicted genes, functional categories) as a function of the number of reads within a sample (Rodriguez & Konstantinidis, 2014). The curve begins with a steep slope and is followed by a point that begins to flatten out as fewer unique features are being detected. If the curve begins to plateau, this illustrates that the sample is close to saturation. However, if the curve does not plateau, this indicates that the sample has not been sufficiently sequenced since new unique features are being detected with a greater number of reads. Therefore, rarefaction curves assess whether the environment has been sufficiently sequenced to detect all unique features and taxa. Rarefaction curves are typically utilized in amplicon sequencing data (such as 18S or 16S rRNA) but are also applied to shotgun metagenomic sequencing data.

The taxonomic rarefaction plots for this study display saturation in terms of discovery of new taxonomic unique features **(Figure 7A & 7B).** The plots also illustrate that there is no bias in saturation between the bloom and non-bloom sites. All samples from both environments reached saturation at a relatively low number of reads, depicting that minimal sequencing effort was required to attain adequate taxonomic information. However, it is important to recall that shotgun metagenomic sequencing is utilized for functional capacity and a greater sequencing effort is required for saturation at the functional level. This is illustrated in **Figure 7C & 7D** since a greater number of reads does not result in a plateau, indicating that saturation has not been reached and a greater sequencing effort is required.

A. Blooms **B. Non-Blooms**



C. Blooms **D. Non-Blooms**

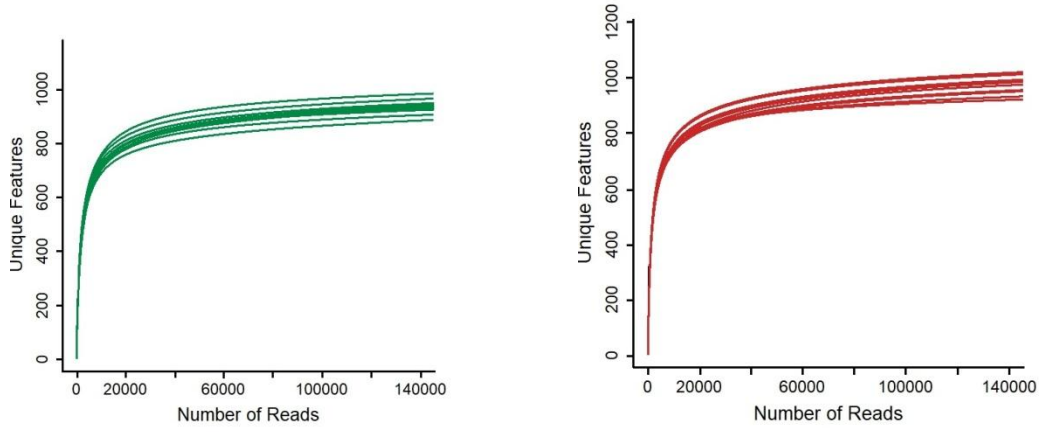


Figure 7. Rarefaction Curves of Taxonomic and Functional Features. Rarefaction curves calculated for taxonomic features in **A** and **B** as well as for functional features in **C** and **D**.

Taxonomic Classification

Taxonomic classification at the Domain level was first examined between bloom and non-bloom sites **(Figure 8).** All sites from both environments had greater than 90% bacterial composition. Only four bloom sites had greater than 2% of reads annotated to the Eukaryota whereas eight non-bloom sites had greater than 2% of reads assigned to the Eukaryota. All samples from both groups had less than 1% of reads assigned to the Archaea except for the non-bloom sample, ON36. Overall, the greatest proportion of reads were annotated to bacterial taxa and less than 10% of reads were annotated to the Eukaryota. Therefore, this validates investigation into the bacterial communities with a focus on cyanobacterial blooms. Eukaryotic algal blooms will not be investigated in detail since all bloom samples in this part of the work were determined to be caused by Cyanobacteria as indicated by microscopic analysis conducted by the MOECC.

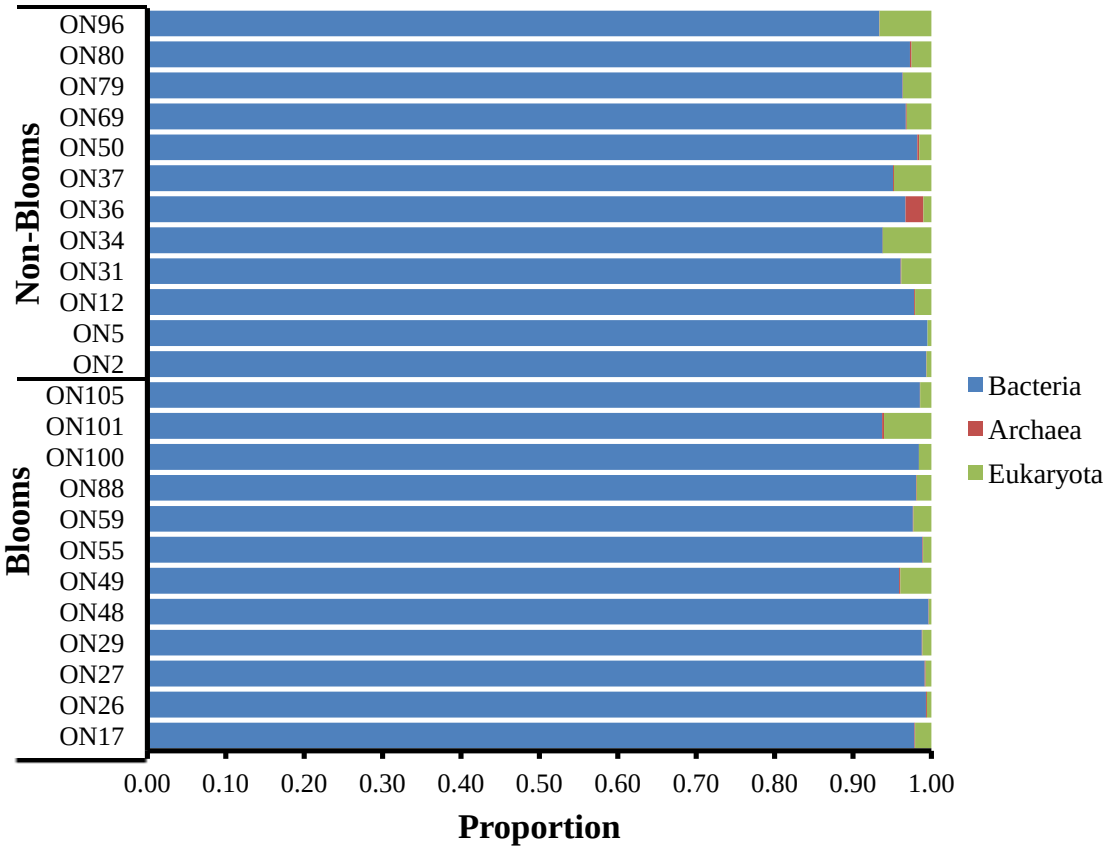


Figure 8. Proportion of Reads Assigned at the Domain Level. Reads were assigned at the domain level in bloom and non-bloom samples.

At the phylum level, taxonomic classification between the two groups was investigated using shotgun metagenomic sequencing to gain a more in-depth analysis of the two environments and their associated microbial communities **(Figure 9).** The only eukaryotic phylum that was greater than 1% abundance in both groups was Bacillariophyta; the rest were below 1% and therefore aggregated with the Rare Phyla group. It is interesting that the only eukaryote greater than 1% abundance in both groups was Bacillariophyta since these organisms are the diatoms that are known to produce blooms. At both bloom and non-bloom sites the most predominant phyla were Proteobacteria, Cyanobacteria and Bacteroidetes. The cumulative proportion of Proteobacteria, Cyanobacteria, Bacteroidetes and Actinobacteria comprised 80-90% of all phyla at all sites in both groups except for the two non-bloom sites ON12 and ON36. This was due to ON12 site's high proportion of Firmicutes and ON36 site's high proportion of Rare Phyla.

A closer examination into the taxonomic differences at the phyla level led to some interesting results. There was only one bloom site that had greater than 20% abundance of Bacteroidetes as opposed to the nine non-bloom sites that had greater than 20% abundance. This was an unexpected result since it has been previously shown that Bacteroidetes abundance is positively correlated with algal bloom sites (Berry *et al.*, 2017, Zhu *et al.*, 2016, Te *et al.*, 2017). However, this further validates findings that blooms contain different microbial communities than non-blooms. Additionally, eight non-bloom sites had greater than 5% Actinobacteria as opposed to the one bloom site. This finding was expected since abundances of Actinobacteria are inversely correlated to Cyanobacteria associated with blooms (Ghai *et al.*, 2014). These findings were further illustrated in **Figure 10.**

Upon examination into cyanobacterial abundance in the two groups, there was a greater abundance of Cyanobacteria present at the bloom sites compared to the non-bloom sites. The high abundance of Cyanobacteria at the bloom sites was expected since the MOECC listed the causative agents of these blooms as Cyanobacteria. The relative abundance of Cyanobacteria at the bloom sites ranged from 13.7 – 86.8 % whereas non-bloom sites had a range of 0.6 - 26.4%. Interestingly, not all bloom sites had a certain abundance threshold of Cyanobacteria that would lead to a cyanobacterial bloom. There were bloom sites that had as little as 13.7% of cyanobacterial relative abundance present at the site whereas non-bloom sites would have more than 20% cyanobacterial relative abundance but a bloom was not present. Therefore, this illustrated the variation in cyanobacterial abundance present at bloom sites and how there was no exact percent abundance threshold that dictated the presence or absence of an algal bloom. This can be explained in part due to cell density of Cyanobacteria being the measure of whether there is or is not a bloom present and that percent abundance is relative to the microbial community.

The cyanobacterial genera between the two groups were examined next to gain a greater insight into the cyanobacterial community and what predominant genera were associated with cyanobacterial blooms **(Figure 11).** Non-bloom sample ON2 was omitted

from the cyanobacterial genus level since there was no classification at that level. Overall, both groups exhibited variation between and within the samples. In the bloom samples, the three main predominant genera were *Microcystis spp.*, *Anabaena spp.*, and *Planktothrix spp.* Three bloom sites had greater than 80% *Microcystis spp.*, four bloom sites had greater than 80% *Anabaena spp.*, and three bloom sites had greater than 90% *Planktothrix spp.* On the other hand, there were two non-bloom sites, ON96 and ON69 that had greater than 80% *Microcystis spp.* This might be explained by the site being sampled in close proximity to a bloom that was *Microcystis spp.* dominated. In contrast to bloom sites, *Anabaena spp.* and *Planktothrix spp.* were in low relative abundance at non-bloom sites. Comparing the two groups, the bloom sites had less cyanobacterial diversity whereas the non-bloom sites had greater cyanobacterial diversity. This reveals that bloom sites had few cyanobacterial genera that would lead to the formation of the algal blooms. Although non-bloom samples had a greater average of *Microcystis spp.*, **Figure 12** further validates the high abundance of *Microcystis spp.*, *Anabaena spp.*, and *Planktothrix spp.* at bloom sites.

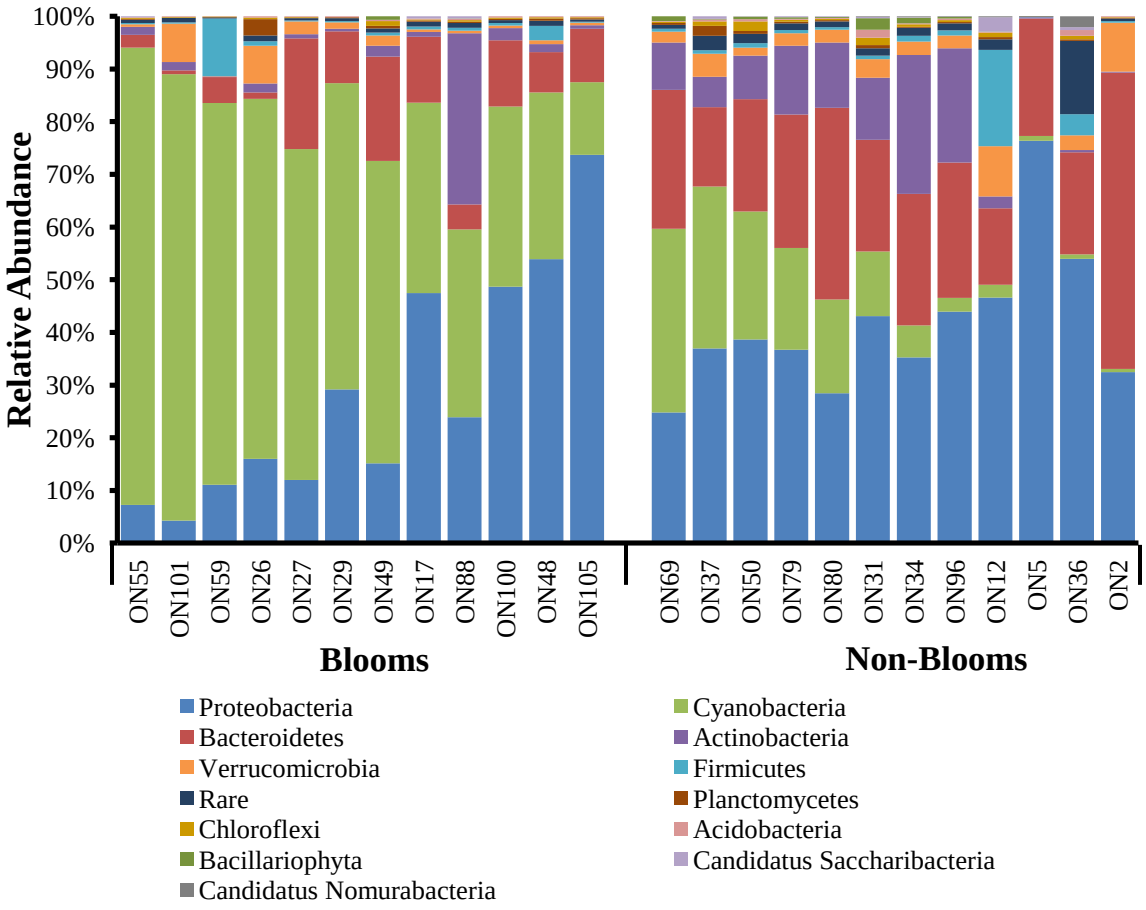


Figure 9. Relative Abundance of Phyla at Bloom and Non-Bloom Sites. Reads were annotated at the phylum level at bloom and non-bloom sites and abundance was calculated relative to each site. Samples within each environment were ranked from highest to lowest cyanobacterial relative abundance. Rare represent phyla <1% in abundance in both groups.

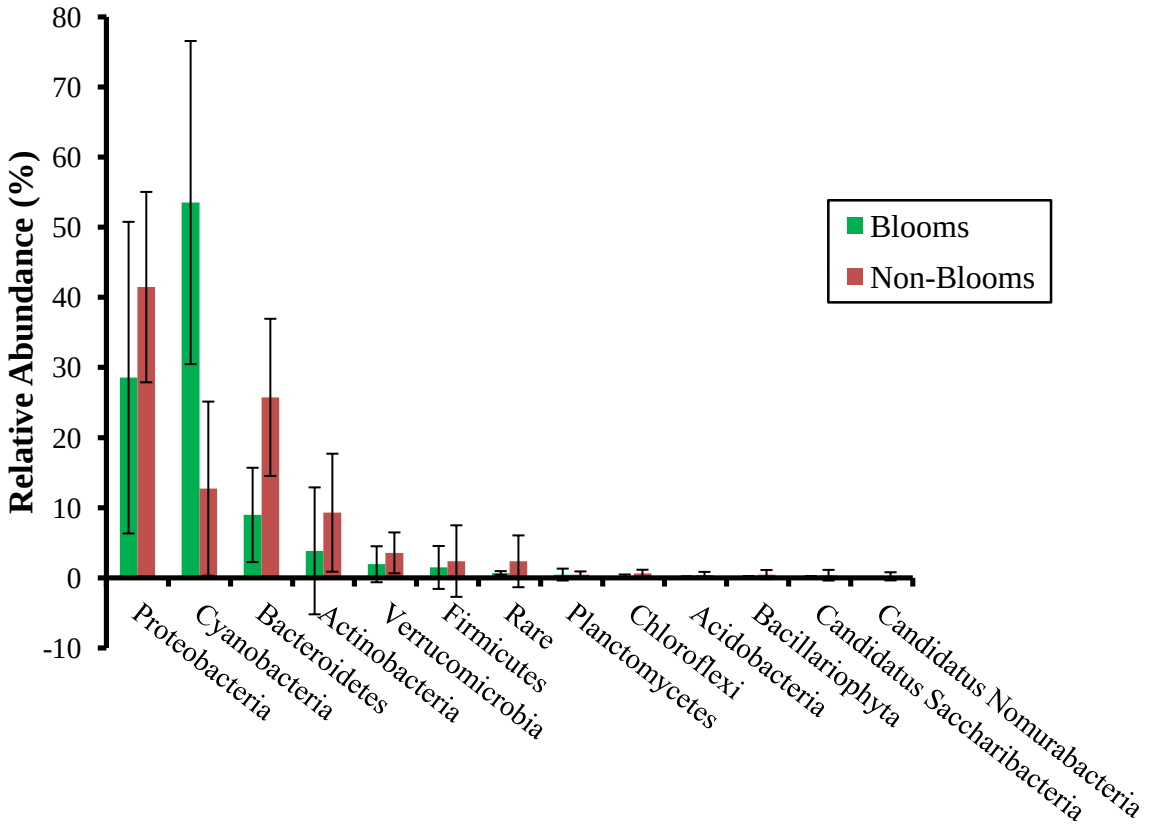


Figure 10. Phylum Composition Within Bloom and Non-Bloom Samples with Standard Deviation. The average for each taxonomic group was calculated within each environment and the standard deviation was determined.

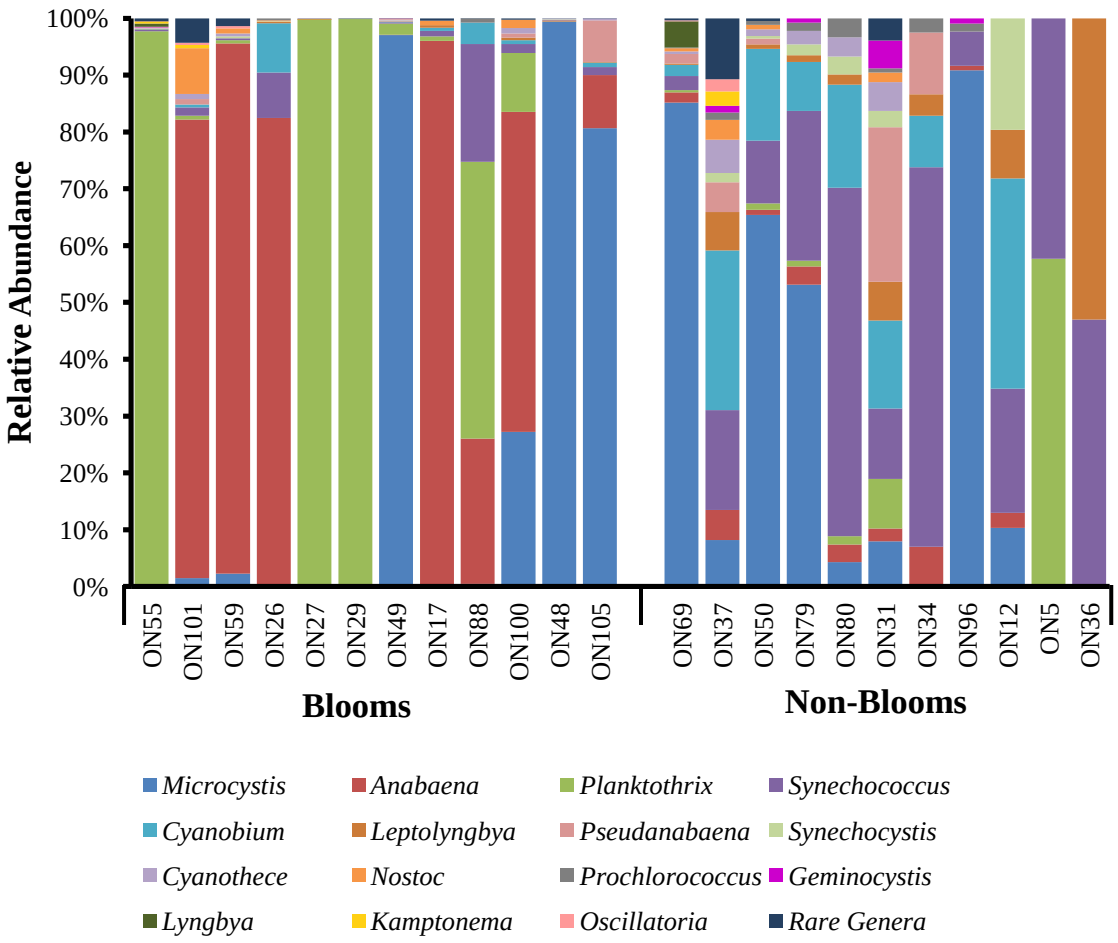


Figure 11. Relative Abundance of Cyanobacterial Genera at Bloom and Non-Bloom Sites. Reads were annotated at the cyanobacterial genera level at bloom and non-bloom sites and abundance was calculated relative to the cyanobacterial community. Cyanobacterial genera with <1% abundance in both groups were aggregated as Rare Genera.

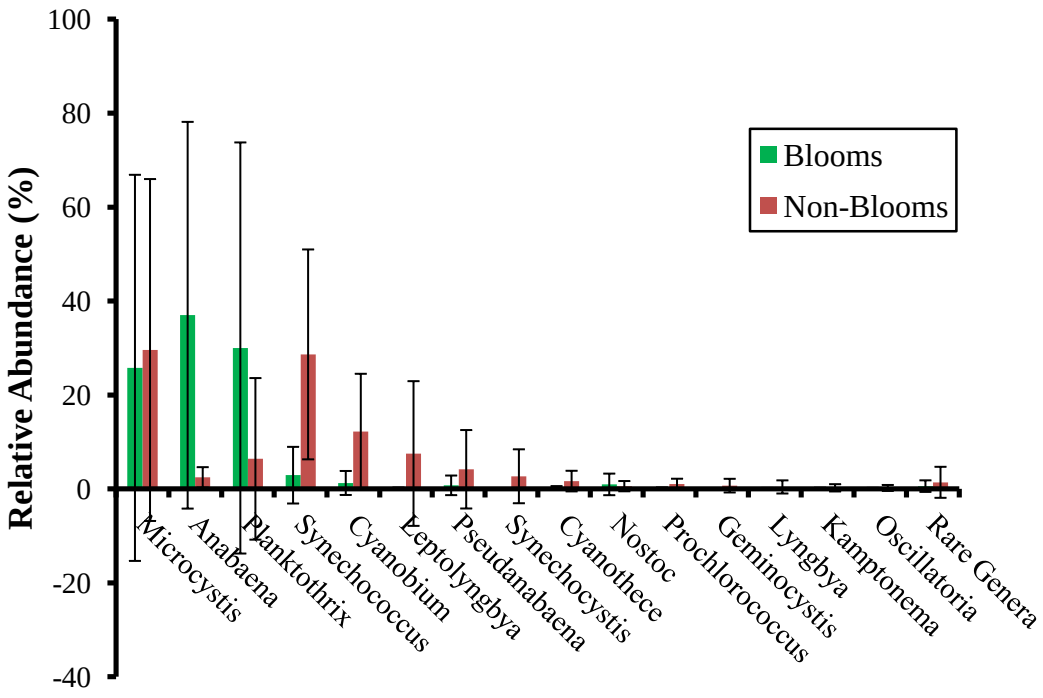


Figure 12. Cyanobacterial Genera Composition Within Bloom and Non-Bloom Samples with Standard Deviation. The average for each taxonomic group was calculated within each environment and the standard deviation was determined.

Functional Classification

Functional classification was first examined between the two environments using the EggNOG database **(Figure 13).** There was not much variation at the broad level functional capacity between bloom and non-bloom sites. Functions associated with amino acid transport and metabolism as well as DNA processes were most abundant with specialized functions such as extracellular, cytoskeleton and nuclear structures with a smaller abundance. This was further validated by examining the functional capacity using the SEED subsystems database at Level 1 **(Figure 14).** Again, there was a lack of variation at the broad level functional capacity with this database. Therefore, despite the high variation that was observed between the groups on the taxonomic level, the broad-scale functional composition was similar between and within the two groups.

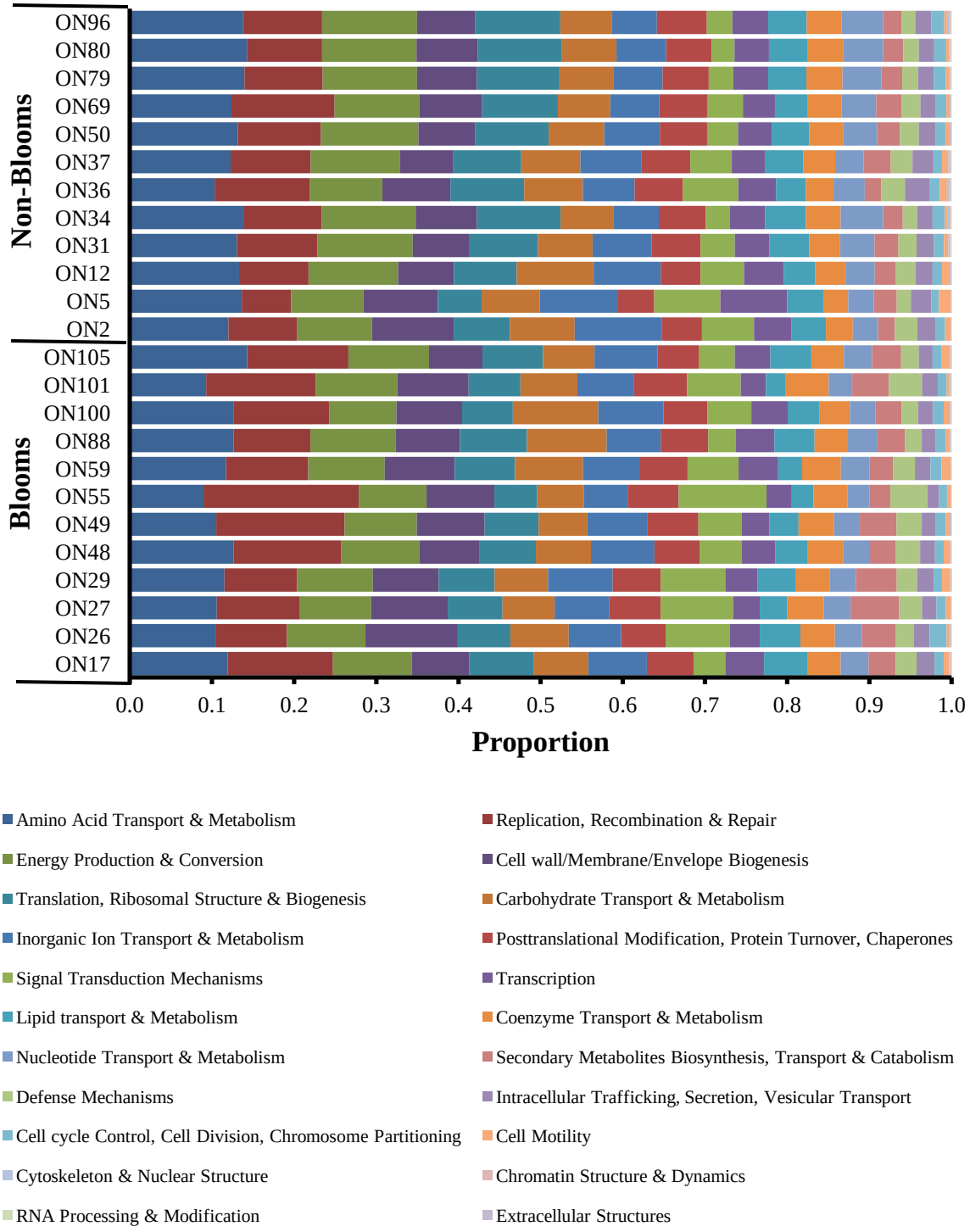


Figure 13. Broad-Scale Functional Classification Utilizing EggNOG Database. Reads were annotated at the functional level and the proportion of reads at each site was determined.

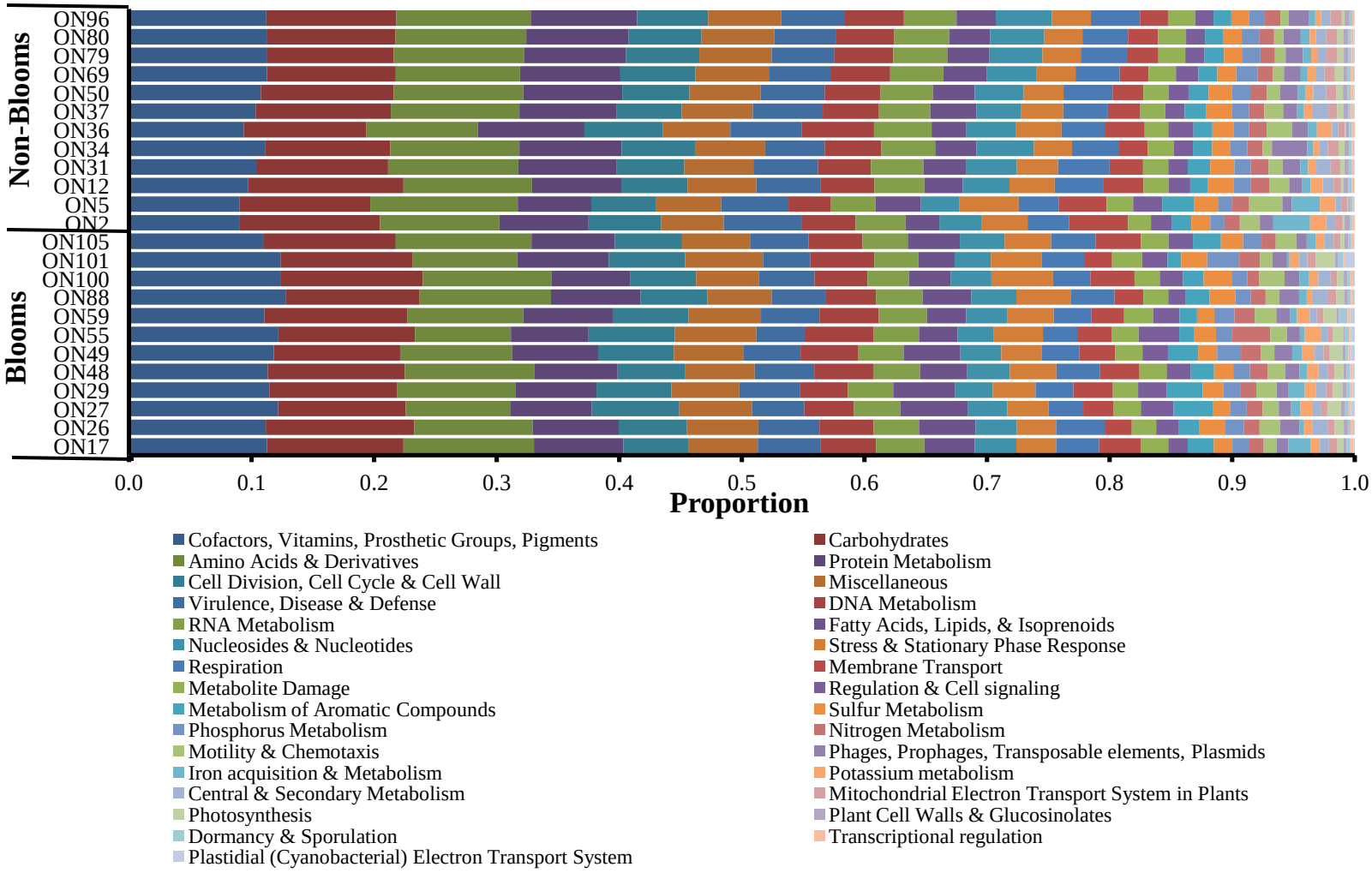


Figure 14. Large-Scale Functional Classification Utilizing the SEED Subsystems Database. Reads were annotated at the functional level and the proportion of reads at each site was determined.

Differential Abundance Analysis

An important goal of this work was to detect differentially abundant taxa and functional capacities between bloom and non-bloom environments. Accordingly, the package DESeq2 was used to analyze differential abundance between the two groups. Although DESeq2 is a package that was initially designed to test differential expression analysis of RNA-Seq experiments, DESeq2 has proven to be an effective and powerful tool in analyzing differential abundance features in microbial communities (Sophie *et al.*, 2015, Jonsson *et al.*, 2016). In one study, the performance of 14 different methods for identifying differentially abundant features between two groups of metagenomes was conducted (Jonsson *et al.*, 2016). It was found that DESeq2 was one of the top 3 programs that had the best overall performance (Jonsson *et al.*, 2016). Additionally, DESeq2 has been proven to be most effective with sample group sizes less than 50 (Weiss *et al.*, 2015). This served as a guide for the utilization of DESeq2 in this current study to analyze differentially abundant features between the two groups. In DESeq2, un-normalized read counts for each sample are input into the program which then uses its own internal normalization process that corrects for the library size. The program then generates a table with Log2FoldChange, *p*-values and other statistical factors.

Differential abundance was first analyzed at the phylum level. There were significant differences in abundances between bloom and non-bloom sites **(Figure 15).** At bloom sites, only Cyanobacteria were the differentially abundant phyla. This supported the hypothesis that blooms were dominated by Cyanobacteria and that these were the causative agents of the blooms. In contrast, the non-bloom sites were enriched with Spirochaetes, Nitrospirae, Gemmatimonadetes and Bacteroidetes. It is important to make note that Spirochaetes, Nitrospirae and Gemmatimonadetes were under 1% relative abundance when all samples were normalized and were therefore aggregated as rare phyla in **Figure 9.**

The Nitrospirae are a phylum of bacteria that are important in nitrite oxidation in freshwater lake sediments (Wan *et al.*, 2017). Accordingly, they are usually found in high nitrogen conditions (Wan *et al.*, 2017). This is an interesting finding since most algal blooms arise from high phosphorus and low nitrogen conditions and therefore Nitrospirae would be expected to be in low abundance at bloom sites. The Gemmatimonadetes are soil-dwelling bacteria that are one of the top nine phyla found in soils and comprise approximately 2% of soil bacterial communities (DeBruyn *et al.*, 2011). It is a minor phylum associated with freshwater lake bacteria and very little reports are available on its occurrence with phytoplankton. However, Gemmatimonadetes is a dominant group associated with *Microcystis* colonies (Shia *et al.*, 2011). Therefore, its abundance at non-bloom sites suggests that not all blooms were *Microcystis spp.* dominated. For differential abundance of Bacteroidetes at non-bloom sites, this again was surprising since, as previously mentioned, Bacteroidetes abundance is positively correlated with algal bloom formation (Berry *et al.*, 2017, Zhu *et al.*, 2016, Te *et al.*, 2017). This differential

abundance analysis at the phylum level further illustrates the different microbial communities that arise in bloom and non-bloom environments.

Differential abundance analysis at the cyanobacterial genera level revealed that all of the significantly differentially abundant genera were enriched at the bloom sites and none were enriched at the non-bloom sites **(Figure 16).** The differentially abundant genera included *Planktothrix*, *Nostoc*, *Microcystis*, *Lyngbya*, *Kamptonema*, *Arthrospira* and *Anabaena*. The majority of these cyanobacterial genera produce cyanotoxins and these include *Planktothrix*, *Nostoc*, *Microcystis*, *Lynbgya*, and *Anabaena* (Blaha *et al.*, 2009). On the other hand, *Kamptonema* and *Arthrospira* are not known to have any species or strains to produce cyanotoxins (Mussagy *et al.*, 2006). However, it is important to not exclude these genera's toxicity since toxic species are likely to be discovered. This data reveals that bloom sites not only have a significant abundance in cyanobacterial genera, but that most of these cyanobacterial genera have the ability to produce toxins indicating the potential for toxicity at these sites.

More specifically, the *Nostoc* genera are widespread in blooms and have more than 200 taxa described. Most species of *Nostoc* produce heterocysts and are either found in terrestrial or freshwater environments. Several species of *Nostoc* have been reported across Canada and they have the potential to produce the hepatotoxin microcystin (Sheath & Steinman, 1981, Wehr *et al.*, 2015, Kurmayer, 2011). *Lyngbya* are a group of aggressive genera that form algal blooms in both freshwater and marine environments (Hudnell, 2008). They have the capacity to produce multiple different types of toxins including aplysiatoxin and lynbgyatoxin-a. Moreover, *Planktothrix*, *Microcystis* and *Anabaena* were the three most dominant cyanobacterial genera in the bloom samples in **Figure 11.** These three cyanobacterial genera are widespread in Ontario freshwater lakes and are the most predominant cyanobacterial genera (Rinta-Kanto *et al.*, 2005, Rinta-Kanto & Wilhelm, 2006, Hotto *et al.*, 2007). They can produce microcystins and their presence may indicate possible toxin production.

Next, the differential abundance of functional genes was examined to detect more specific functional differences between the two environments. This was done using the SEED subsystems database at level 2. Interestingly, there was a greater differential abundance of functions associated with bloom environments than with non-bloom environments **(Figure 17).** All functions associated with the following groups were differentially abundant at bloom sites: Cell Division and Cell cycle; Cell Wall and Cell Capsule; Cofactors, Vitamins, Prosthetic Groups, Pigments; DNA Metabolism; Dormancy and Sporulation; Heterocysts Formation in Cyanobacteria; Metabolite Damage and Repair on Mitigation; Phages, Prophages, Transposable Elements, Plasmids; Regulation and Cell Signaling; RNA Metabolism; and Plastidial Electron Transport System. Some of these functions are associated with Cyanobacteria such as heterocyst formation, cyanobacterial electron transport system, myxoxanthophyll biosynthesis and cyanobacterial circadian clock. This is consistent with the observed high abundance of Cyanobacteria at bloom sites.

In particular, functions associated with the carboxysome were differentially abundant in bloom samples. Carboxysomes are proteinaceous compartments within cells that envelope the fundamental CO_2-fixing enzyme (ribulose-1,5-bisphosphate carboxylase/oxygenase, RuBisCO) in Cyanobacteria and in some Proteobacteria (Rae *et al.*, 2013). They are important intracellular structures involved in carbon fixation. There are two types of carboxysomes: α-carboxysomes that are mainly found in oceanic Cyanobacteria and in some Proteobacteria, or β-carboxysomes that are predominantly found in freshwater Cyanobacteria (Rae *et al.*, 2013). There is a differential abundance of carboxysomes and more specifically, β-carboxysomes in bloom samples which further signifies the presence of Cyanobacteria at these sites. Additionally, functions associated with B12 and Cobalamin pathways were differentially abundant at bloom sites. A majority of microalgae require vitamin B12 for growth and it is synthesized by a small number of prokaryotic species including a majority of Cyanobacteria (Helliwell *et al.*, 2016). This further reveals the large presence of Cyanobacteria in blooms.

Additionally, differential abundance analysis was conducted on the nitrogen and phosphorus metabolism genes. **Table 8** reveals the functions Nitrogen Fixation and Phosphorus Metabolism to be differentially abundant at the bloom sites. Although other bacteria aside from Cyanobacteria can fix nitrogen, the functions associated with nitrogen fixation provide the initial support that cyanobacterial species that can form heterocysts are most likely present at bloom sites. It would be interesting if further analysis was conducted to see what genes are carrying out nitrogen fixation at these sites. Moreover, the differential abundance of phosphorus metabolism provides a possible indication of the presence of high phosphorus conditions at the bloom sites. As previously mentioned, high phosphorus conditions are one of the factors that lead to algal bloom formation (Schindler *et al.*, 2008).

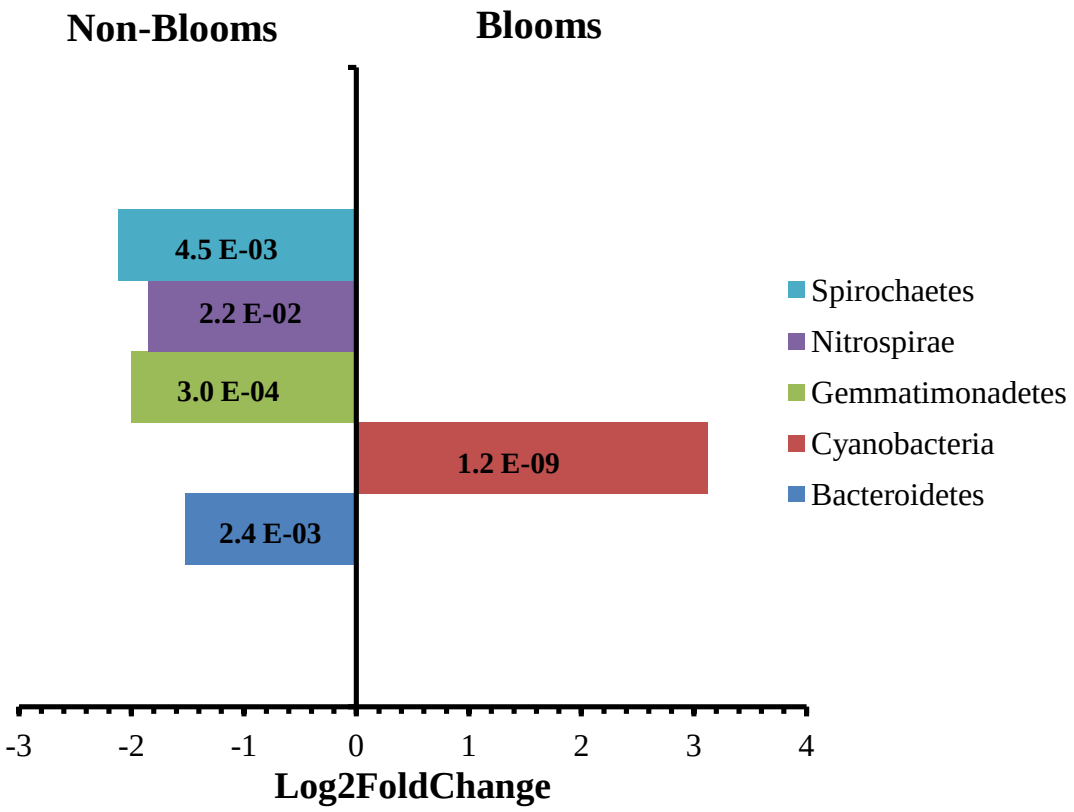


Figure 15. Differentially Abundant Phyla. The positive log2FoldChange indicates enrichment at bloom sites and the negative log2FoldChange indicates enrichment at non-bloom sites. The associated *p*-values are indicated for each phylum. The threshold adjusted *p*-value for significance is 0.05.

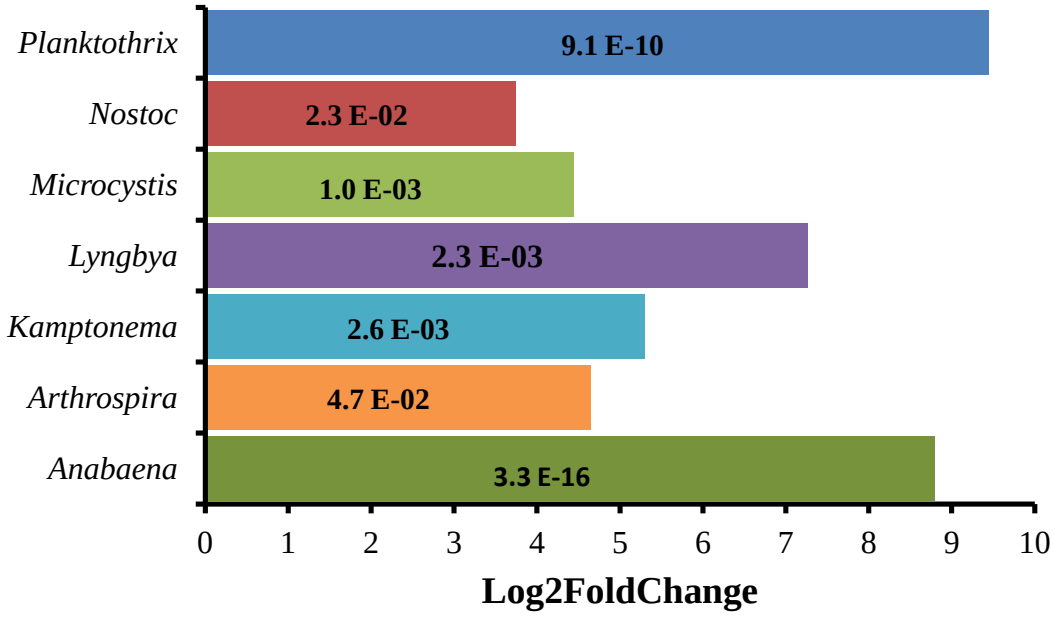


Figure 16. Differentially Abundant Cyanobacterial Genera. The positive log2FoldChange indicates enrichment of taxa at bloom sites. The associated *p*-values for each cyanobacterial genera are indicated. The threshold adjusted *p*-value for significance is 0.05.

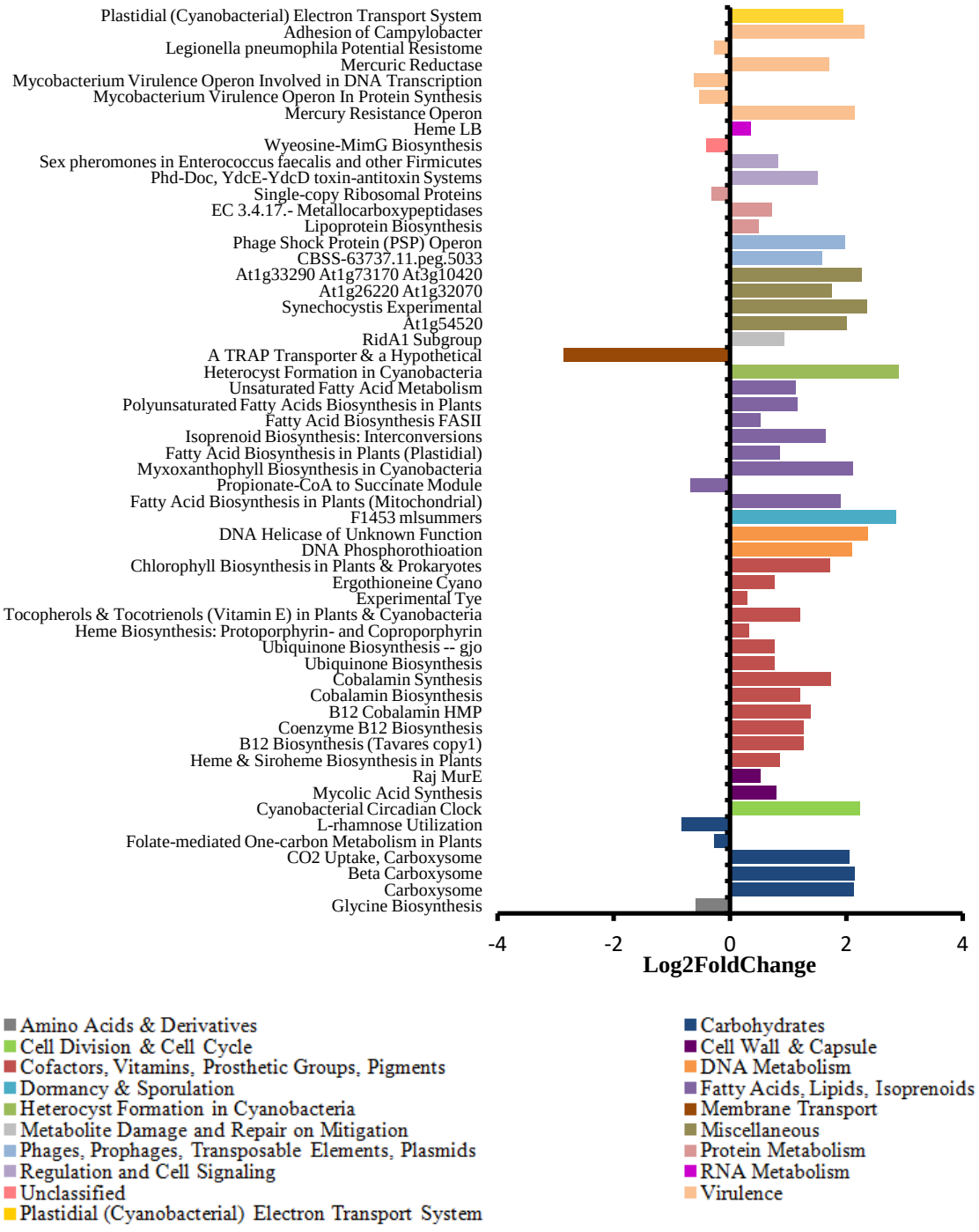


Figure 17. Differential Abundance of Functional Analysis. The SEED subsystems database was used at level 2 for differential abundance of functions. Positive log2foldchange indicates differential abundance at bloom sites and negative log2foldchange indicates differential abundance at non-bloom sites. The threshold of significance was an adjusted *p*-value of 0.0001.

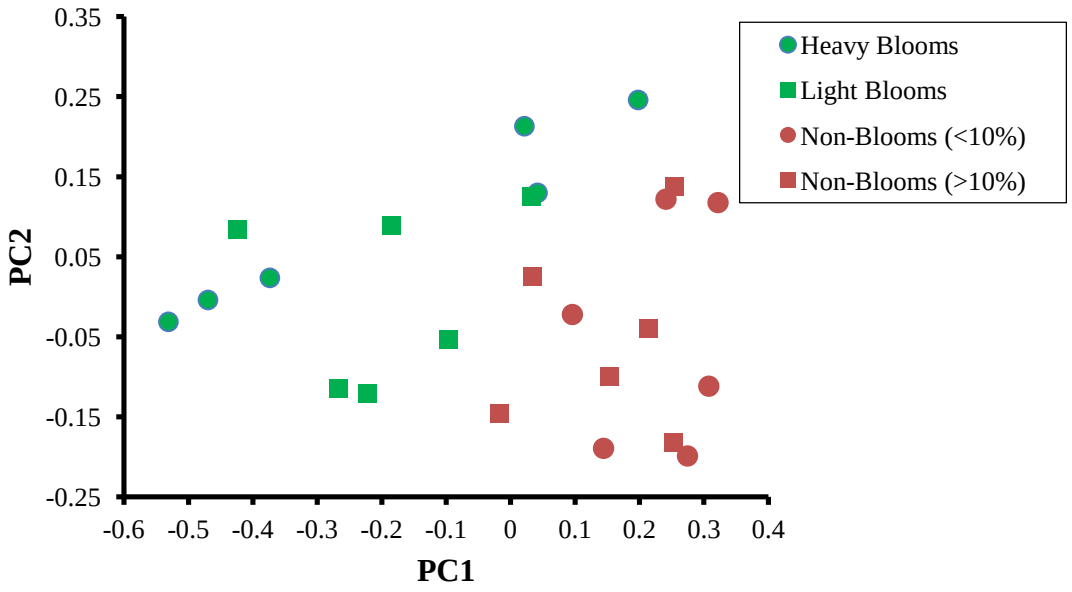
Table 8. Differential Abundance of Nitrogen and Phosphorus-associated Functions. The adjusted *p*-value cut off is 0.05. Positive L2FC indicates enrichment at bloom sites.

Function	baseMean	Log2FoldChange	lfcSE	stat
Nitrogen fixation	412.0183	0.927	0.492	1.883
Phosphorus Metabolism	10783.05	0.257	0.112	2.304

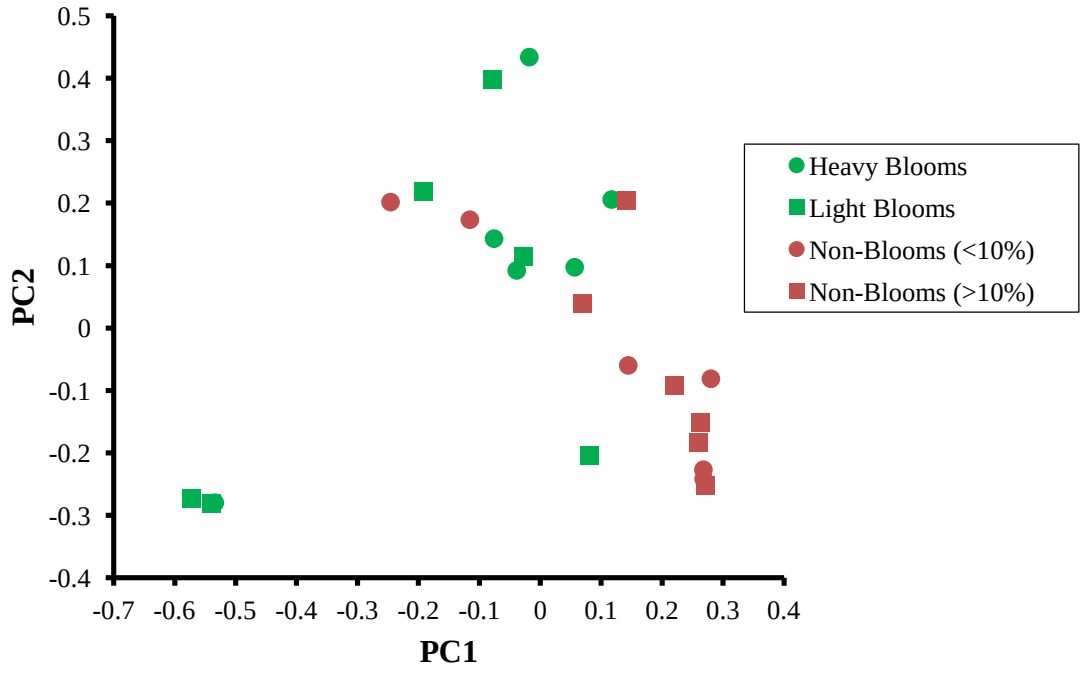
Beta-Diversity Analysis

PCoA plots were constructed with a different distance metric than that used in the 16S rRNA sequencing analysis. Here, the Bray-Curtis dissimilarity measure was utilized for PCoA analysis (Bray & Curtis, 1957). Bray-Curtis is widely applied in community ecological data for multivariate analysis. It is a non-phylogenetically based measure that takes abundance data into account and its measure has a range between 0 and 1 (Bray & Curtis, 1957). If a value of 0 is obtained then both sites being compared share all the same species whereas a value of 1 indicates the two sites do not share any of the species (Bray & Curtis, 1957). As such, Bray-Curtis dissimilarity measure ignores cases where the species are absent in both sites and is dominated by the abundant species. PCoA plots were calculated using the Bray-Curtis dissimilarity measure for taxonomic and functional levels. The two bloom and non-bloom environments were divided into two further groups. Within the bloom samples, there were the heavy bloom sites that had high cyanobacterial density and the light bloom sites where that had low cyanobacterial density. Within the non-bloom samples, half the samples had <10% cyanobacterial abundance according to 16S rRNA sequencing analysis and half had >10% cyanobacterial abundance. These subdivisions were utilized to see if different cyanobacterial densities within the two groups would have an effect on diversity.

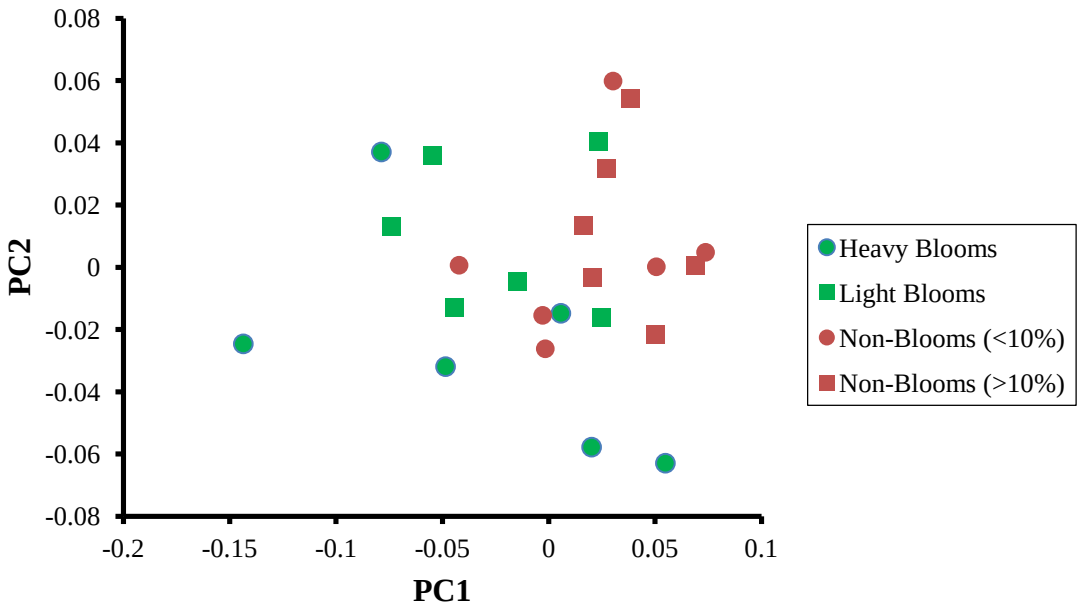
Between bloom and non-bloom sites there were differences at the taxonomic phylum level that were revealed **(Figure 18A).** However, there is no difference that is noted within the four subdivisions. Therefore, one can conclude that there is a difference in the two communities based on whether it is a bloom or non-bloom site and that there is no variation between heavy and light blooms as well as having a high or low cyanobacterial abundance at non-bloom sites. Differences at the genus level was examined next **(Figure 18B).** There was clustering for most of the non-bloom samples but otherwise there were no differences revealed. In terms of differences based on the subdivisions, there were no differences that were observed. The taxonomic variation at the phylum level was not correlated with functional variation **(Figure 18C).** The EggNOG database was used for this analysis and it was determined that large-scale functionalities illustrated no variation between the two groups and their subdivisions.



(A) Phylum Level



(B) Genus Level.



(C) Functional Level

Figure 18. PCoA Plots Utilizing Bray-Curtis Dissimilarity Measure. PCoA plots were calculated using the Bray-Curtis dissimilarity measure at the **(A)** Phylum Level **(B)** Genus Level and **(C)** Functional Level of the EggNOG database.

2.5 Discussion

Cyanobacterial blooms are a global concern that over the past two decades have been increasing in occurrence, intensity and geographic distribution. It is important that algal blooms be studied extensively due to its potential danger to humans, animals and associated ecosystems. Cyanobacteria have the capacity to produce cyanotoxins that have adverse health effects on humans and animals affecting the liver, skin, nervous system and other organs. In this work, metagenomic sequencing was highlighted as an important method of monitoring and studying algal blooms to gain a better understanding of taxonomic compositions and functional capacities. This method circumvents the problem of culturing Cyanobacteria and its associated microbial community so that insight into algal blooms can be gained that traditional methods and its counterparts cannot achieve.

In collaboration with the Ministry of the Environment and Climate Change, 108 samples were collected across Ontario bloom and non-bloom regions. The goal of this work was to analyze blooms and to compare them to non-bloom sites with regards to their microbial communities and functional capacities. To our current knowledge, there has been no other work that has completed such an extensive comparative metagenomic sequencing approach to assess over 100 samples of Ontario bloom and non-bloom sites. Despite these samples only being single samples, our work is able to provide a snapshot of the microbial communities at these sites.

With the utilization of amplicon sequencing, differences between bloom and non-bloom microbial communities were revealed. Although there was a lot of microbial community variation across the samples and a lack of metadata to facilitate explanation of the results, it is still important to draw preliminary conclusions from the data and use this as a basis for future work. At the phylum level, Planctomycetes and Verrucomicrobia were associated with bloom samples in greater abundance than with non-bloom samples, supporting previous findings (Haukka *et al.*, 2006, Kolmonen *et al.*, 2004, Louati *et al.*, 2015, Pizzetti *et al.*, 2011, Eiler & Bertilsson, 2004). Additionally, Cyanobacteria were found to be at a greater overall abundance in bloom samples when compared to non-bloom samples. Not all bloom samples had a high abundance of Cyanobacteria and this can be attributed to the fact that not all samples collected by the MOECC were cyanobacterial in origin. Similarly, there were some non-bloom samples that had a high abundance of Cyanobacteria and this could be due to the close proximity of sampling to an algal bloom dominated by Cyanobacteria. When dissecting the cyanobacterial communities, it was found that the three cyanobacterial orders Oscillatoriales, Chroococcales, and Synechococcales were the most abundant in both groups with Oscillatoriales having the greatest proportion across bloom sites and Synechococcales having the greatest proportion across non-bloom sites. Diversity analysis utilizing the Weighted Unifrac measure revealed no difference between the two groups as well as on a monthly basis. This was possibly due to UniFrac utilizing a phylogenetic distance measure, thereby indicating that the two groups showed no phylogenetic difference.

Temporal variation analysis of bloom sites at the taxonomic phylum level revealed Cyanobacteria being the most abundant in July and August, a finding that is supported in literature (van Apeldoorn *et al.*, 2007). However, a Weighted UniFrac PCoA plot of only bloom sites on a monthly basis depicted no variation. This again can be due to the lack of phylogenetic differences between the samples since the weighted UniFrac measure was used.

Toxin PCR analysis of the *mcyE* gene of the microcystin operon revealed that only 4 samples of the 108 had the *mcyE* gene present and all 4 samples were bloom sites. These 4 positive samples were also sampled in July and August, indicating a temporal trend. The low number of positive samples does not imply that toxin production was not present at the sites that tested negative. This is because the PCR was developed for detecting *mcyE* present in *Microcystis spp.* and toxin producing potential of *Planktothrix spp.* and *Anabaena spp.* would not be detected. Also, some toxic strains of Cyanobacteria might lack the *mcyE* gene of the microcystin operon (Vaitomaa *et al.*, 2003) or the *mcyE* gene was present in concentrations too low to be detected.

Next, a more in-depth analysis was employed with the utilization of shotgun metagenomic sequencing on a subset of the 108 samples. Here, 24 samples were chosen where half were bloom samples and half were non-bloom samples. The bloom samples chosen were determined to be caused by Cyanobacteria by the MOECC through microscopic analysis. Analysis into the taxonomic composition between the two groups revealed Proteobacteria, Cyanobacteria, and Bacteroidetes to be the predominant phyla. Additionally, differential abundance analysis revealed that Cyanobacteria were the only differentially abundant phylum at bloom sites whereas Spirochaetes, Nitrospirae, Gemmatimonadetes and Bacteroidetes were the differentially abundant phyla at the non-bloom sites. The differential abundance of Bacteroidetes in non-bloom samples was an interesting finding since it has been previously shown that Bacteroidetes is positively correlated with algal bloom sites (Berry *et al.*, 2017, Zhu *et al.*, 2016, Te *et al.*, 2017). At the cyanobacterial genera level, it was determined that *Anabaena spp.*, *Planktothrix spp.*, and *Microcystis spp.*, were the three most abundant Cyanobacteria at bloom sites. Differential abundance analysis at the cyanobacterial genera level revealed no differentially abundant cyanobacterial genera at the non-bloom sites and that the bloom sites had *Planktothrix*, *Nostoc*, *Microcystis*, *Lyngbya*, *Kamptonema*, *Arthrospira*, and *Anabaena* as the differentially abundant cyanobacterial genera. Most of these genera have been cited to release toxins and this was an indication of possible toxicity at the bloom sites (Blaha *et al.*, 2009).

Broad-level functional analysis indicated no differences between and within the two environments. However, differential abundance analysis of specific functions revealed functions that were significantly different between the two groups. It was shown that the majority of functions were differentially abundant in bloom samples. Additionally, several functions were associated with Cyanobacteria. These functions included heterocyst formation, cyanobacterial circadian clock, cyanobacterial electron transport

system and myxoxanthophyll biosynthesis. Moreover, beta-diversity analysis utilizing the Bray-Curtis dissimilarity measure illustrated differences at the taxonomic phylum level between bloom and non-bloom environments. At the genus level, there was clustering for most non-bloom samples but otherwise there were no differences revealed. The PCoA plot revealed no large-scale functional variation between the two groups.

Overall, this study was able to describe taxonomic and functional differences between 108 geographically distinct bloom and non-bloom sites utilizing metagenomic sequencing approaches. It has provided an in-depth analysis of the causative agents of algal blooms in Ontario, their associated microbial communities and functional capacities. An extension of this work should focus on identifying the presence of nitrogen assimilation genes such as those associated with nitrogen fixation (*nifD),* urea assimilation (*ureA-G*), nitrate reduction (*nar/nir)* and ammonium utilization (*glnA)* as previously described (Steffen *et al.*, 2012). This could provide a basis in the potential role of bloom formation and the major contributors. More importantly, future work in this field should focus on collecting more metadata on the sampling site. Information such as the temperature of water, sampling depth, time of day, and nutrient levels are all important factors that can have an impact on the microbial community. These factors would provide a greater insight into the taxonomic and functional capacities of the communities and will allow for a greater distinction when comparing bloom and non-bloom sites.

Chapter 3. Assessing the Validity of Conserved Signature Proteins in Identifying Cyanobacteria in Environmental Samples

3.1 Introduction

The taxonomic classification of Cyanobacteria has been largely debated for the past few decades. Cyanobacteria are a difficult group to classify due to their long and complex evolutionary history that makes it challenging to discern classification based on morphology. Prior to the 1970s, the classification of Cyanobacteria was governed by the International Code of Botanical Nomenclature and under this code they were considered to be "Cyanophytes" and accordingly were treated as microscopic plant organisms in the field of botany (Stanier *et al.*, 1978). This was due to the morphological and size similarities Cyanobacteria shared with eukaryotic algae. However shortly after, cytological and biochemical studies performed on axenic culture strains led to the proposal that they were prokaryotic in nature and should accordingly be classified as "Cyanobacteria" (Stanier *et al.*, 1978, Gibbons & Murray, 1978). Therefore, in the late 1970s Cyanobacteria were classified under the International Code of Nomenclature of Bacteria (Stanier *et al.*, 1978, Rippka, 1988).

Although Cyanobacteria are now under the Bacteriological Code, most cyanobacterial names are still governed by the Botanical Code and very few taxa are ruled by the Bacteriological Code (Hoffmann *et al.*, 2005, Parte, 2014). This coexistence of two separate nomenclature codes for the same group of organisms creates immense problems (Pinevich, 2015, Oren & Garrity, 2014). The names of Cyanobacteria previously elucidated and validly published under the International Code of Botanical Nomenclature have no standing with the bacterial nomenclature and must be redescribed under the Bacteriological Code (Pinevich, 2015, Oren & Garrity, 2014). The Botanical Code rules are different from the Bacteriological Code, making reconciliation between the two nomenclature systems problematic. For instance, the Bacteriological Code uses axenic cultures as the basic taxonomic unit to establish valid names. On the other hand, the Botanical Code uses preserved-type specimens as the taxonomic reference. Overall, classification of Cyanobacteria is a difficult task and there have been several published systems of classifying Cyanobacteria since their introduction to the Bacteriological Code (Hoffmann, 2005, Cavalier-Smith, 2002, Euzeby, 1997, Nelissen *et al.*, 1994). This difficult with classifying Cyanobacteria provides the initial support for the use of novel genetic markers to dissect the problems with cyanobacterial classification and identification.

16S rRNA gene analysis is important for understanding phylogenetic relationships among prokaryotes including Cyanobacteria, and is useful in identifying and

distinguishing certain taxa (Naushad *et al.*, 2014). However, 16S rRNA gene analysis is used as the only marker to identify most prokaryotic taxa and no other unique biochemical or molecular characteristic is used as a supplement to distinguish certain taxonomic groups from others (Naushad *et al.*, 2014). Therefore, it is imperative that accurate and reliable genetic markers be identified and used to supplement our current understanding of taxonomic identification and evolutionary relationships (Gupta, 2009, Gupta & Griffiths, 2002). Conserved Signature Proteins (CSPs) are one such genetic marker that can be utilized for taxonomic identification and classification (Gao & Gupta, 2012, Gupta, 2009). CSPs are proteins that are present in bacterial groups that share common ancestors and that can be used to delineate the organism of interest (Howard-Azzeh *et al.*, 2014). One limitation to the utilization of CSPs is that the vast majority of cyanobacterial CSPs have unknown functions. This can lead to questions regarding whether the identified CSPs are in fact proteins that are important for delineating the organism of interest or are some other protein byproduct found in the cells.

CSPs are uniquely found in several taxa including the cyanobacterial phylum and thus have the ability to assist with the difficulties of classifying Cyanobacteria (Gupta & Mathews, 2010, Gupta *et al.*, 2003). Therefore, to utilize these molecular markers to supplement current methods of classifying Cyanobacteria, Dr. Radhey Gupta in the Department of Biochemistry at McMaster University developed a database that contains CSPs specific to several taxa. The aim of this work was to test the validity of utilizing CSPs to identify Cyanobacteria in bloom and non-bloom environmental samples. This was achieved by comparing taxonomic identification at the phylum level using CSPs, shotgun metagenomic sequencing (NCBI-nr database) and amplicon sequencing (Greengenes database) annotation methods.

3.2 Methods

The sample protocols for DNA Extraction, DNA Library Preparation and Sequencing, and Bioinformatic Analysis are the same as in Chapter 2. There were 24 samples that were chosen for signature protein analysis of which 12 consisted of bloom samples and 12 of non-bloom samples. These are the same 24 samples that were utilized in shotgun metagenomic sequencing analysis in Chapter 2. For reader convenience purposes, the methods outlined in Chapter 2 are restated below. It is also important to mention that throughout this Chapter, the term "signature proteins" refers to "CSPs" and the two terms are interchangeable.

3.2.1 DNA Extraction

Approximately 5.0mL of each water sample was centrifuged at 10,000 rpm for 15 minutes. The supernatant was then discarded and the pellet was used for DNA extraction. DNA extractions were performed using the Norgen Biotek Corp Soil DNA Isolation Kit (Norgen Biotek Corporation, Canada). Manufacturer's protocol was followed, with the addition of 10µg/mL of lysozyme and 200mM of β-mercaptoethanol to maximize lysis of cyanobacterial cells (Mehta *et al.*, 2015). Approximately 150µL of DNA was extracted for each sample. Extracted DNA was stored at -20°C until further use for library preparation, sequencing and polymerase-chain reaction.

3.2.2 Preparation of Libraries and Sequencing

DNA was diluted to 0.2ng/µL and prepared for sequencing using the Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA). The samples were sent to the Farncombe Metagenomics Facility (McMaster University, Hamilton, Ontario, Canada) for 250bp paired-end sequencing on the Illumina HiSeq 1500 platform.

3.2.3 Bioinformatic and CSP Analysis

Raw reads obtained from the Illumina HiSeq platform were quality screened using FastQC (v. 0.11.3)(Andrews, 2010) to detect adapter sequences and low-quality reads. Trimmomatic (v.0.36) (Bolger *et al.*, 2014) was used for adapter removal and quality trimming. Trimmomatic's *Nextera-PE* template was used for adapter clipping and a quality score of 25 was selected as the quality score threshold. Overlapping paired-ended reads were then merged using Fast Length Adjustment of Short Reads (FLASh v.1.2.11) (Magoc & Salzberg, 2011). Default parameters were used for merging with the exception

of the minimum overlap set to 15bp. Processed reads were again quality tested using FastQC and reads that passed were retained for further analysis.

Once reads were trimmed and merged, each sample file was individually run against Dr. Radhey Gupta's CSP database by his lab. Dr. Radhey Gupta has developed an extensive internal CSP database that contains CSP sequences unique to several taxa including Cyanobacteria. The input files were normalized to a file size of 100 MB.

3.3 Results

To assess the use of CSPs in identifying Cyanobacteria, the output from the CSP database annotations were divided into bloom and non-bloom samples in **Table 9** and **Table 10**, respectively. The numbers of hits shown were normalized to a file size of 100MB. Only the top three phyla hits are illustrated in the tables. In all bloom samples, Cyanobacteria were one of the top three phyla that were detected via signature proteins. Moreover, 10 out of the twelve bloom samples had Cyanobacteria with the largest number of CSP hits. This revealed the primary importance and validity of CSP work in identifying Cyanobacteria since all twelve bloom samples were previously shown in Chapter 2 to have a high abundance of Cyanobacteria. On the other hand, only seven non-bloom samples had Cyanobacteria in the top three phyla. This again further validated the use of signature proteins in identifying Cyanobacteria in environmental samples since Chapter 2 illustrated a relatively high abundance of Cyanobacteria in these non-bloom samples.

In order to compare the results of using CSPs to annotate taxa as opposed to the utilization of amplicon sequencing and shotgun metagenomic sequencing annotation methods, **Table 11** and **Table 12** were created for bloom and non-bloom samples, respectively. These tables compare the top three phyla within each sample and annotation scheme, allowing us to gain a more comprehensive understanding of how powerful CSP work is. Upon examination into bloom samples, nine had Cyanobacteria in the top three phyla in all three annotation methods. Additionally, in nine of the bloom samples, Cyanobacteria was in the top three phyla using CSP and amplicon sequencing whereas in all twelve bloom samples Cyanobacteria was in the top three phyla in CSP and shotgun metagenomic sequencing annotation techniques. This reveals that signature proteins and shotgun metagenomic sequencing annotation methods were the most similar to each other for cyanobacterial abundance. Moreover, this analysis of bloom samples provides the initial basis that CSP work in identifying Cyanobacteria in environmental samples is an important tool and is one that has initial validity. However, it should be noted that more work should be completed on CSP annotation and that further research should extend into looking past the phyla level and into taxonomic levels below phylum. One caveat to this work is that because each classification technique utilizes a different database, it is possible to get a different number of hits for each phylum across the different methods. In some databases some organisms may be overrepresented and in other databases they may be underrepresented and this may lead to some disparity and biases when comparing the top phyla.

Non-bloom samples had similar results for cyanobacterial abundance. In 7 of the samples, cyanobacterial abundance was either in the top three phyla in all annotation methods or not in the top three for all annotation methods. There was only one sample, ON5, where amplicon sequencing and shotgun metagenomic sequencing methods detected Cyanobacteria as the top three phyla but was not present as the top three in

signature protein classification. There were three non-bloom samples that did not have Cyanobacteria as the top three phyla in all annotation methods. Additionally, in two of the non-bloom samples, Cyanobacteria was identified as the top three phyla using CSPs and shotgun metagenomic sequencing annotation as opposed to the one non-bloom sample that identified Cyanobacteria as the top three phyla using CSPs and amplicon sequencing. This analysis of non-bloom samples further reveals that CSP work is an important tool for supplementing current methods of identifying cyanobacteria.

Conversely, when other phyla are examined there are some contrasting results. For instance, in some bloom samples, Firmicutes is listed as a top three phylum using signature proteins but is not listed as a top three phylum using either amplicon sequencing or shotgun metagenomic sequencing annotation in any of those samples. This trend is also seen in the non-bloom samples whereby Planctomycetes is in the top three phyla in some samples for CSP annotation but in none of those samples using amplicon and shotgun metagenomic sequencing annotation methods. This may be explained due to the CSP database having a more extensive collection of CSPs towards these two groups of organisms and a less extensive collection of other phyla's signature proteins. This would then lead to an overrepresentation of certain groups of organisms and an underrepresentation of other groups.

Table 9. Signature Protein Hits of Bloom Samples

Sample	Phyla	Number of Normalized Hits
ON17	Actinobacteria	827
	Cyanobacteria	527
	Alphaproteobacteria	281
ON26	Cyanobacteria	844
	Planctomycetes	413
	Firmicutes	51
ON27	Cyanobacteria	1692
	Firmicutes	65
	Bacteroidetes	48
ON29	Cyanobacteria	294
	Alphaproteobacteria	66
	Firmicutes	14
ON48	Cyanobacteria	795
	Alphaproteobacteria	124
	Bacteroidetes	86
ON49	Cyanobacteria	1079
	Planctomycetes	57
	Alphaproteobacteria	46
ON55	Cyanobacteria	1418
	Actinobacteria	98
	Firmicutes	39
ON59	Cyanobacteria	1860
	Gammaproteobacteria	552
	Proteobacteria	120
ON88	Cyanobacteria	971
	Alphaproteobacteria	56
	Firmicutes	53
ON100	Cyanobacteria	904
	Alphaproteobacteria	148
	Bacteroidetes	129
ON101	Cyanobacteria	2638
	Actinobacteria	85
	Firmicutes	59
ON105	Alphaproteobacteria	325
	Cyanobacteria	233
	Actinobacteria	198

Table 10. Signature Protein Hits of Non-Bloom Samples

Sample	Phyla	Number of Normalized Hits
ON2	Bacteroidetes	677
	Alphaproteobacteria	91
	Firmicutes	47
ON5	Proteobacteria	90
	Gammaproteobacteria	65
	Bacteroidetes	44
ON12	Alphaproteobacteria	131
	Planctomycetes	55
	Firmicutes	47
ON31	Cyanobacteria	90
	Planctomycetes	58
	Alphaproteobacteria	39
ON34	Cyanobacteria	172
	Actinobacteria	132
	Firmicutes	82
ON36	Actinobacteria	77
	Alphaproteobacteria	22
	Firmicutes	21
ON37	Planctomycetes	231
	Cyanobacteria	148
	Firmicutes	27
ON50	Cyanobacteria	380
	Planctomycetes	89
	Firmicutes	58
ON69	Cyanobacteria	630
	Firmicutes	68
	Planctomycetes	60
ON79	Cyanobacteria	306
	Firmicutes	52
	Planctomycetes	38
ON80	Cyanobacteria	348
	Actinobacteria	87
	Firmicutes	72
ON96	Gammaproteobacteria	224
	Actinobacteria	61
	Firmicutes	37

Table 11. Comparison of Annotation Methods for Bloom Samples

Sample	Conserved Signature Proteins	Amplicon Sequencing (Greengenes Database)	Shotgun Metagenomic Sequencing (NCBI-nr Database)
ON17	Actinobacteria Cyanobacteria Alphaproteobacteria	Proteobacteria Bacteroidetes Actinobacteria	Proteobacteria Cyanobacteria Bacteroidetes
ON26	Cyanobacteria Planctomycetes Firmicutes	Cyanobacteria Planctomycetes Proteobacteria	Cyanobacteria Proteobacteria Verrucomicrobia
ON27	Cyanobacteria Firmicutes Bacteroidetes	Cyanobacteria Proteobacteria Bacteroidetes	Cyanobacteria Bacteroidetes Proteobacteria
ON29	Cyanobacteria Alphaproteobacteria Firmicutes	Cyanobacteria Proteobacteria Bacteroidetes	Cyanobacteria Proteobacteria Bacteroidetes
ON48	Cyanobacteria Alphaproteobacteria Bacteroidetes	Proteobacteria Cyanobacteria Actinobacteria	Proteobacteria Cyanobacteria Bacteroidetes
ON49	Cyanobacteria Planctomycetes Alphaproteobacteria	Proteobacteria Bacteroidetes Cyanobacteria	Cyanobacteria Bacteroidetes Proteobacteria
ON55	Cyanobacteria Actinobacteria Firmicutes	Proteobacteria Cyanobacteria Verrucomicrobia	Cyanobacteria Proteobacteria Bacteroidetes
ON59	Cyanobacteria Gammaproteobacteria Proteobacteria	Proteobacteria Actinobacteria Cyanobacteria	Cyanobacteria Proteobacteria Bacteroidetes
ON88	Cyanobacteria Alphaproteobacteria Firmicutes	Proteobacteria Cyanobacteria Actinobacteria	Cyanobacteria Actinobacteria Proteobacteria
ON100	Cyanobacteria Alphaproteobacteria Bacteroidetes	Proteobacteria Actinobacteria Bacteroidetes	Proteobacteria Cyanobacteria Bacteroidetes
ON101	Cyanobacteria Actinobacteria Firmicutes	Proteobacteria Actinobacteria Bacteroidetes	Cyanobacteria Verrucomicrobia Proteobacteria
ON105	Alphaproteobacteria Cyanobacteria Actinobacteria	Proteobacteria Cyanobacteria Bacteroidetes	Proteobacteria Cyanobacteria Bacteroidetes

Table 12. Comparison of Annotation Methods for Non-Bloom Samples

Sample	Conserved Signature Proteins	Amplicon Sequencing (Greengenes Database)	Shotgun Metagenomic Sequencing (NCBI-nr Database)
ON2	Bacteroidetes	Proteobacteria	Bacteroidetes
	Alphaproteobacteria	Bacteroidetes	Proteobacteria
	Firmicutes	Verrucomicrobia	Verrucomicrobia
ON5	Proteobacteria	Proteobacteria	Proteobacteria
	Gammaproteobacteria	Bacteroidetes	Bacteroidetes
	Bacteroidetes	Cyanobacteria	Cyanobacteria
ON12	Alphaproteobacteria	Proteobacteria	Proteobacteria
	Planctomycetes	Cyanobacteria	Firmicutes
	Firmicutes	Bacteroidetes	Bacteroidetes
ON31	Cyanobacteria	Proteobacteria	Proteobacteria
	Planctomycetes	Cyanobacteria	Bacteroidetes
	Alphaproteobacteria	Actinobacteria	Cyanobacteria
ON34	Cyanobacteria	Proteobacteria	Proteobacteria
	Actinobacteria	Actinobacteria	Actinobacteria
	Firmicutes	Cyanobacteria	Bacteroidetes
ON36	Actinobacteria	Proteobacteria	Proteobacteria
	Alphaproteobacteria	Firmicutes	Bacteroidetes
	Firmicutes	Bacteroidetes	Verrucomicrobia
ON37	Planctomycetes	Proteobacteria	Proteobacteria
	Cyanobacteria	Cyanobacteria	Cyanobacteria
	Firmicutes	Actinobacteria	Bacteroidetes
ON50	Cyanobacteria	Proteobacteria	Proteobacteria
	Planctomycetes	Bacteroidetes	Cyanobacteria
	Firmicutes	Verrucomicrobia	Bacteroidetes
ON69	Cyanobacteria	Proteobacteria	Cyanobacteria
	Firmicutes	Actinobacteria	Bacteroidetes
	Planctomycetes	Cyanobacteria	Proteobacteria
ON79	Cyanobacteria	Proteobacteria	Proteobacteria
	Firmicutes	Actinobacteria	Bacteroidetes
	Planctomycetes	Cyanobacteria	Cyanobacteria
ON80	Cyanobacteria	Proteobacteria	Bacteroidetes
	Actinobacteria	Actinobacteria	Proteobacteria
	Firmicutes	Bacteroidetes	Cyanobacteria
ON96	Gammaproteobacteria	Proteobacteria	Proteobacteria
	Actinobacteria	Actinobacteria	Bacteroidetes
	Firmicutes	Bacteroidetes	Actinobacteria

3.4 Discussion

This work has provided an initial confirmation that Conserved Signature Proteins are a powerful molecular marker that can aid in the identification and classification of Cyanobacteria. It is preliminary work that compared identification of Cyanobacteria using signature proteins with amplicon sequencing and shotgun metagenomic sequencing annotation methods. When compared to these taxonomic identification techniques, signature proteins did exceptionally well in identifying the cyanobacterial phylum in environmental samples that consisted of bloom and non-bloom sites. The use of Dr. Radhey Gupta's extensive CSP database for identification and classification of Cyanobacteria is warranted for future environmental research and this work reveals a promising future for this field. However, other phyla such as Firmicutes and Planctomycetes would appear in the top three phyla in an environmental sample using signature proteins, but would not be present in shotgun metagenomic sequencing and amplicon sequencing annotation methods. This can be attributed to the CSP database having a more extensive list of CSPs annotated to these phyla and less CSPs identified for other phyla. Consequently, this would lead to an overrepresentation of Firmicutes and Planctomycetes being detected with signature proteins. When the signature proteins database hits were compared to shotgun metagenomic sequencing and amplicon sequencing database hits, it was found that CSP annotation had more similar results to shotgun metagenomic sequencing than amplicon sequencing. This can be attributed to the fact that the databases used within each sequencing method differ and what might be well characterized in one database may not be well characterized in the other. Nonetheless, cyanobacterial CSP work has proven to be an efficient method for supplementing the identification and classification of Cyanobacteria.

Future work on CSPs should utilize Dr. Radhey Gupta's CSP database to characterize what cyanobacterial organisms are present at the family, order, and genus level. From there, research could focus on comparing the use of signature proteins at these taxa levels to other annotation methods to further validate CSPs as a supplement for identification of Cyanobacteria. Additionally, future work should also look at identifying the functions of these signature proteins since most have unknown functions. Identifying the functions of the signature proteins will help in recognizing and answering questions regarding the importance of these conserved proteins and will aid in overcoming the limitations of CSP work.

Appendix A. Chapter 2 Supplement

DNA Extraction SOP (Modified from Norgen Biotek's Soil DNA Isolation Kit):

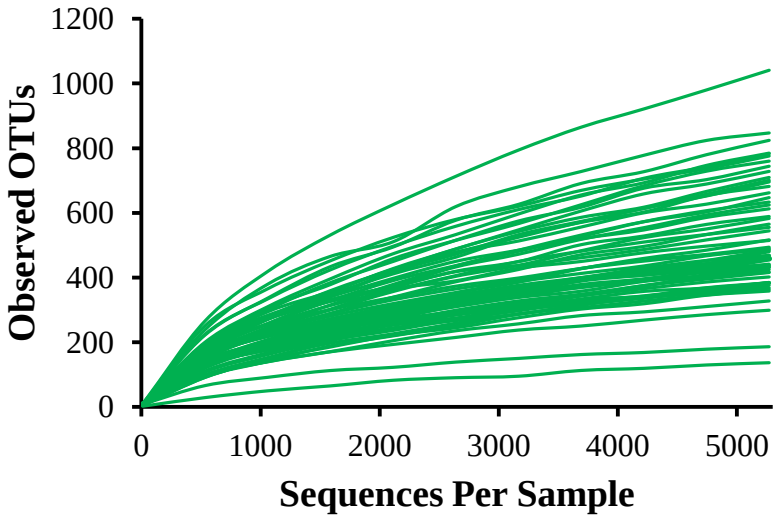
1. **Lysate Preparation**
   a. Centrifuge 5.0mL of each water sample at 10,000 rpm for 15 minutes
   b. Remove the supernatant and resuspend pellet in 750µL of Lysis Buffer G
   c. Transfer the sample to a Bead Tube
   d. Add 10µg/mL of Lysozyme and 200mM β-mercaptoethanol and mix
   e. Add 100µL of Lysis Additive A and vortex for 5 seconds
   f. Secure the bead tube onto the bead beater and vortex for 4 minutes
   g. Centrifuge the tube at 14,000g for 2 minutes
   h. In a microcentrifuge tube, transfer approximately 450µL of supernatant
   i. Add 100µL of Binding Buffer I, inverting the tube a few times and then incubating on ice for 5 minutes
   j. Centrifuge the tube for 2 minutes at 14,000 g to pellet any proteins and particles
   k. Transfer 450µL of supernatant to another microcentrifuge tube and add 230µL of 96-100% ethanol

2. **Binding to Column**
a. Mix the lysate and ethanol with a pipette and transfer approximately 630µL of the lysate with ethanol onto a Spin Column with its collection tube
b. Centrifuge for 1 minute at 8,000 g. Discard the flow through and reassemble the spin column with the collection tube

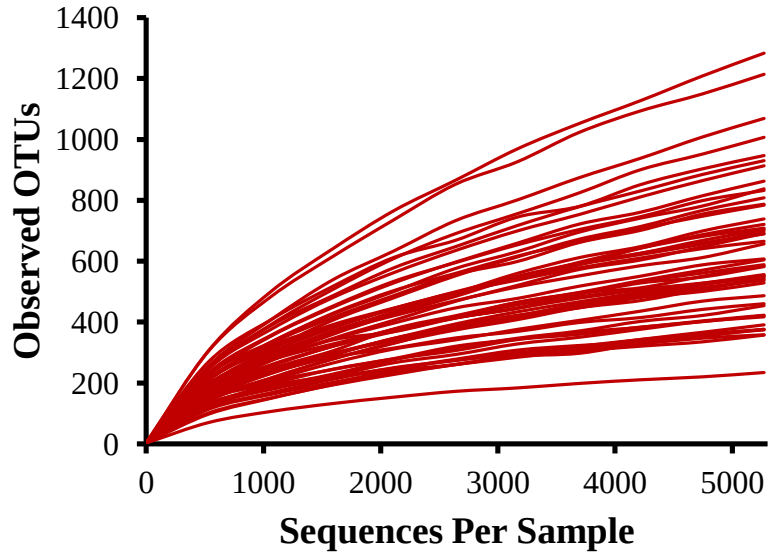
3. **Column Wash**
a. Add 500µL of Buffer SK to the Spin Column and centrifuge for 1 minute at 8,000g
b. Discard the flow through and reassemble the spin column with the collection tube
c. Apply 500µL of Wash Solution A and centrifuge for 1 minute at 8,000g
d. Discard the flow through and reassemble the spin column with its collection tube
e. Spin the column for 2 minutes at 14,000rpm. Discard the collection tube

4. **DNA Elution**
a. Place the Spin Column into a 1.7mL Elution tube
b. Add 100µL of Elution Buffer B to the column and incubate for 5 minutes at room temperature
c. Centrifuge for 1 minute at 8,000g
d. Another elution is performed by repeating Steps 4b,4c and using 50µL of Elution Buffer B in a different elution tube. This increases the yield by approximately 20-30%. Therefore the total amount of DNA extracted is 150µL
e. Store the DNA at -20°C until needed for further analysis

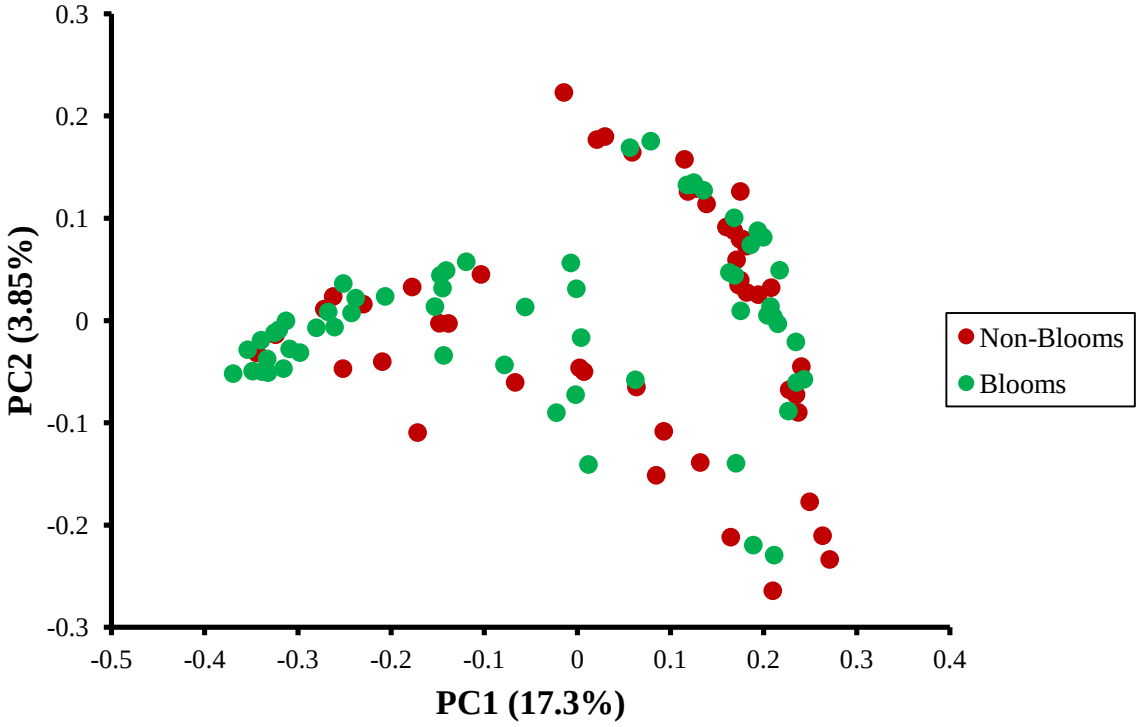
A.



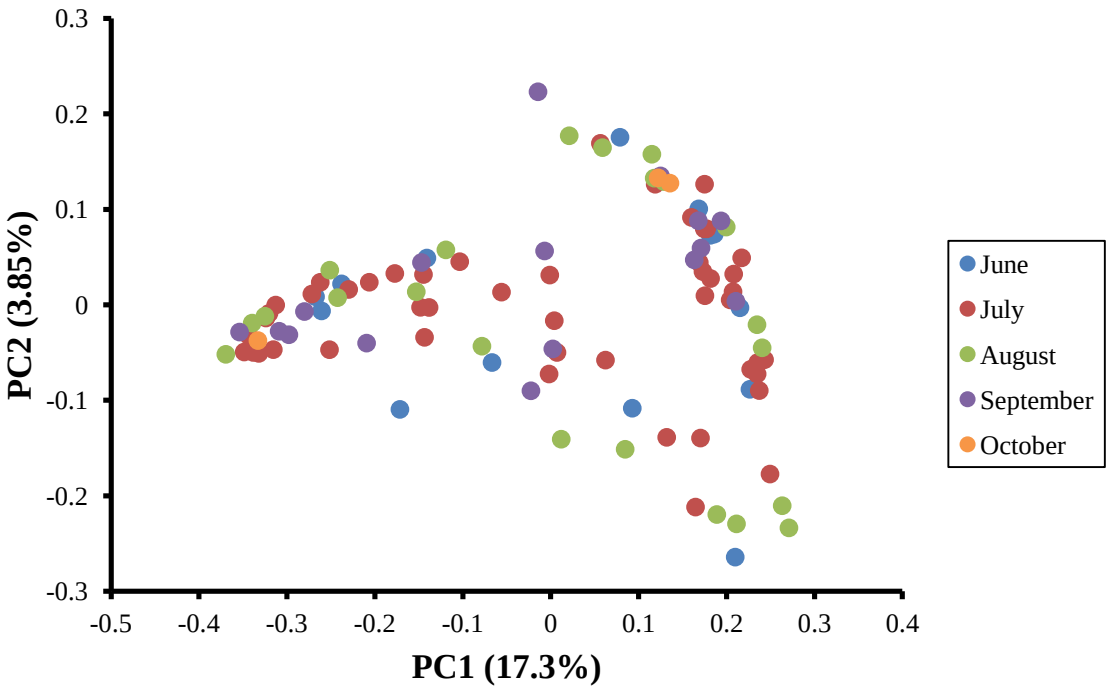
B.



Supplementary Figure 19. Rarefaction Curves of Amplicon Sequencing Data. Rarefaction curve of **(A)** Bloom Samples and **(B)** Non-bloom Samples.

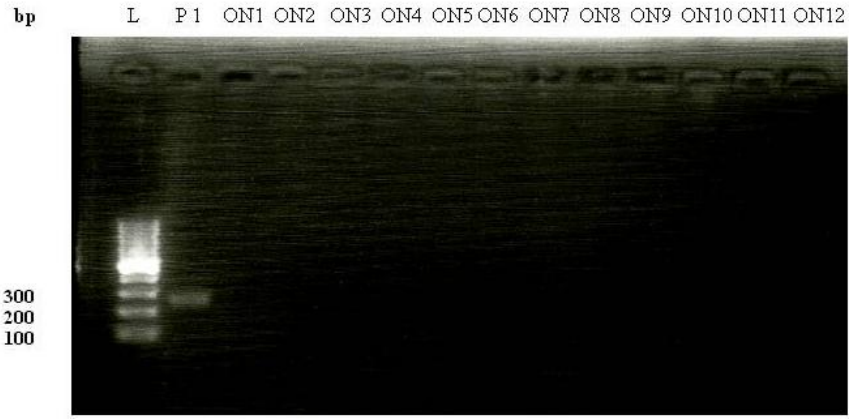


Supplementary Figure 20. PCoA Plot Utilizing Unweighted UniFrac Measure for Amplicon Sequencing Data based on Bloom and Non-Bloom Sites. PCoA Plot was made using the Unweighted UniFrac Measure for all 108 samples and samples were divided into their respective group of blooms and non-blooms.

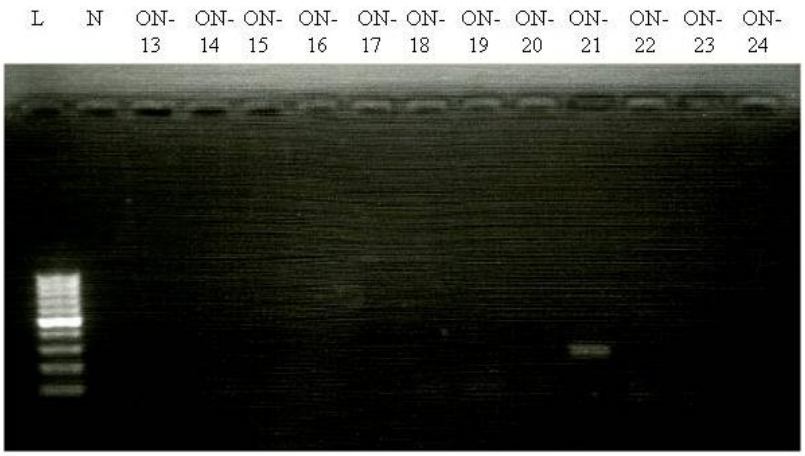


Supplementary Figure 21. PCoA Plot Utilizing Unweighted UniFrac Plot for Amplicon Sequencing Data on a Monthly Basis. PCoA plot was made using the Unweighted UniFrac Measure for all 108 samples and samples were divided based on their respective months of sampling.

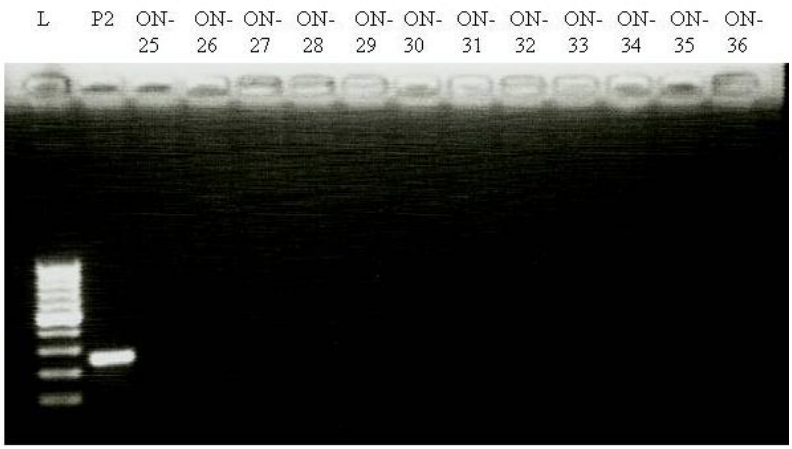
A.


bp
L
P 1
ON1
ON2
ON3
ON4
ON5
ON6
ON7
ON8
ON9
ON10
ON11
ON12
300
200
100

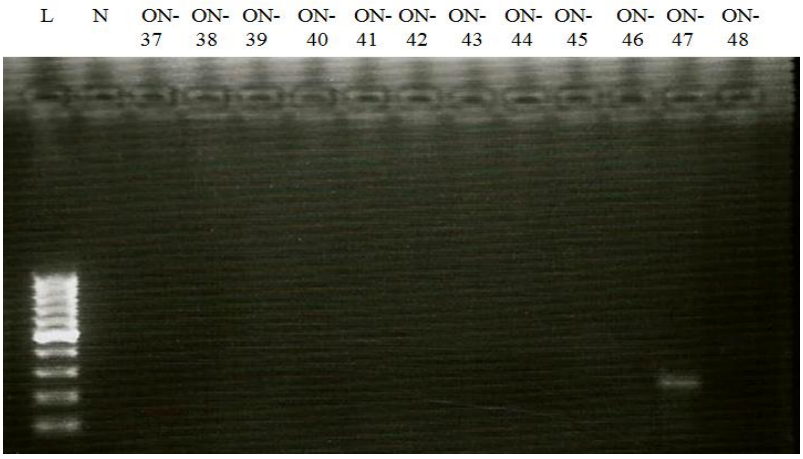
B.


L
N
ON-13
ON-14
ON-15
ON-16
ON-17
ON-18
ON-19
ON-20
ON-21
ON-22
ON-23
ON-24

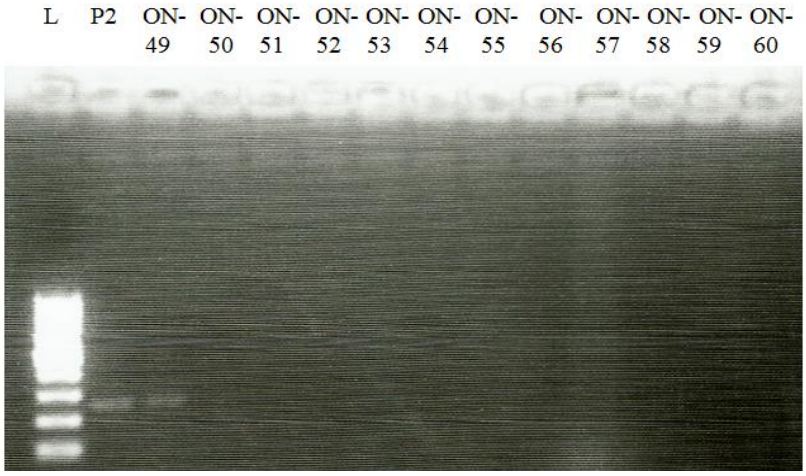
C.


L
P2
ON-25
ON-26
ON-27
ON-28
ON-29
ON-30
ON-31
ON-32
ON-33
ON-34
ON-35
ON-36

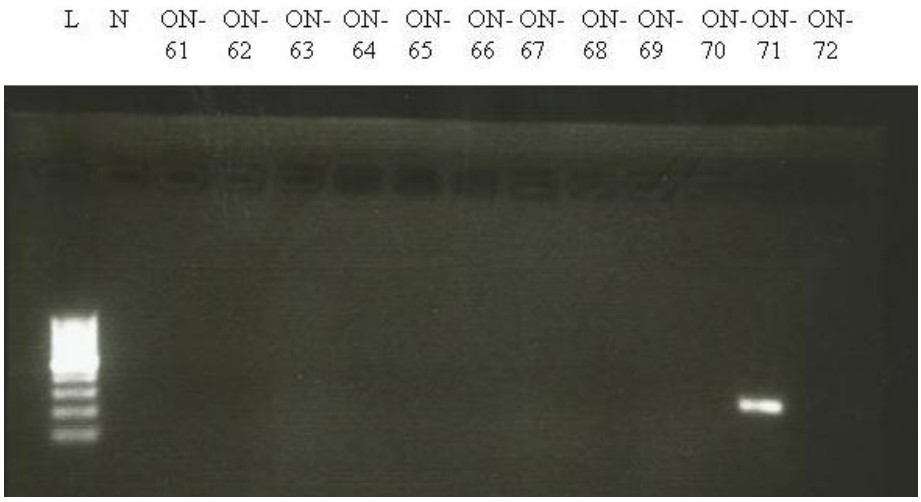
D.


L
N
ON-37
ON-38
ON-39
ON-40
ON-41
ON-42
ON-43
ON-44
ON-45
ON-46
ON-47
ON-48

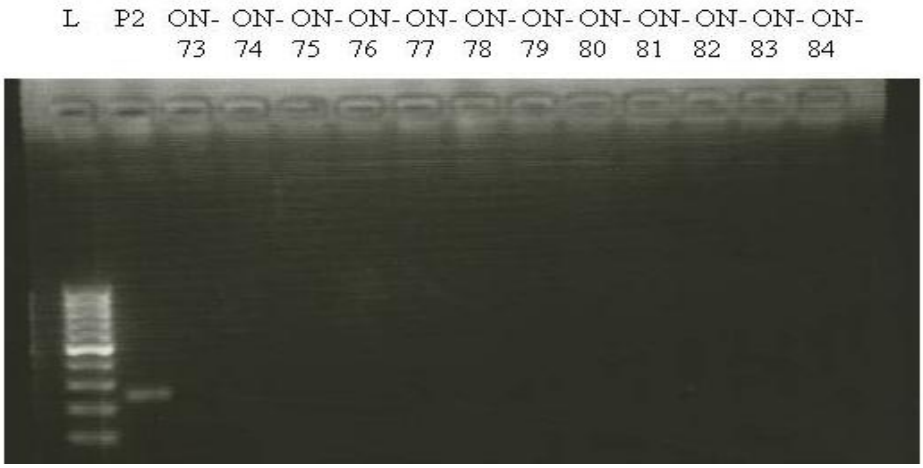
E.


L
P2
ON-49
ON-50
ON-51
ON-52
ON-53
ON-54
ON-55
ON-56
ON-57
ON-58
ON-59
ON-60

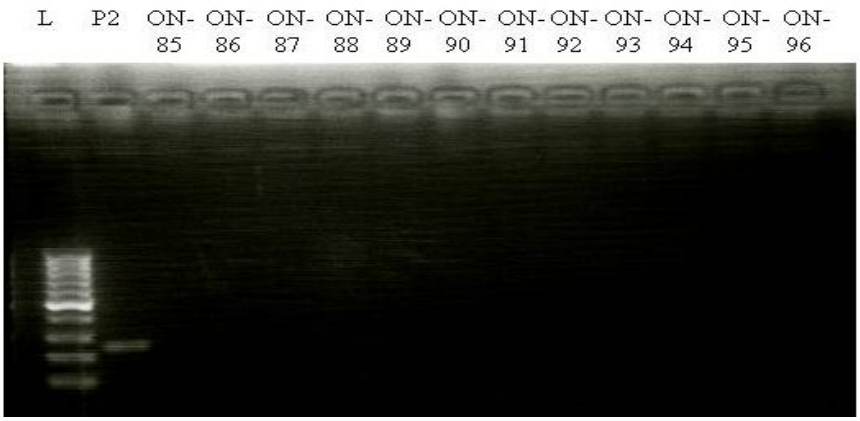
F.


L N ON-61 ON-62 ON-63 ON-64 ON-65 ON-66 ON-67 ON-68 ON-69 ON-70 ON-71 ON-72

G.


L P2 ON-73 ON-74 ON-75 ON-76 ON-77 ON-78 ON-79 ON-80 ON-81 ON-82 ON-83 ON-84

H.

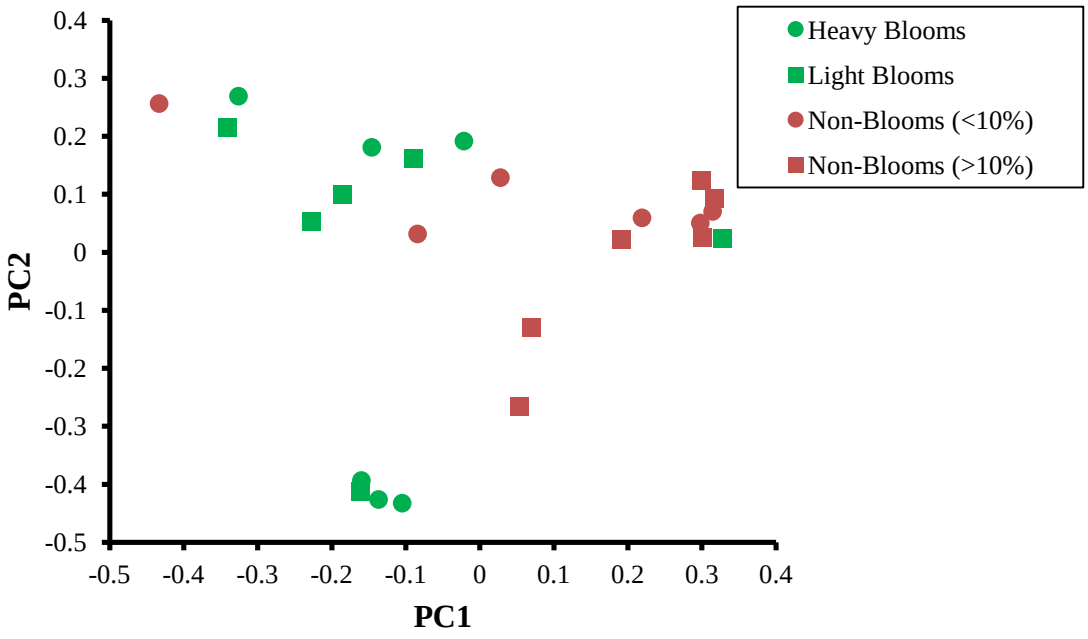


I.

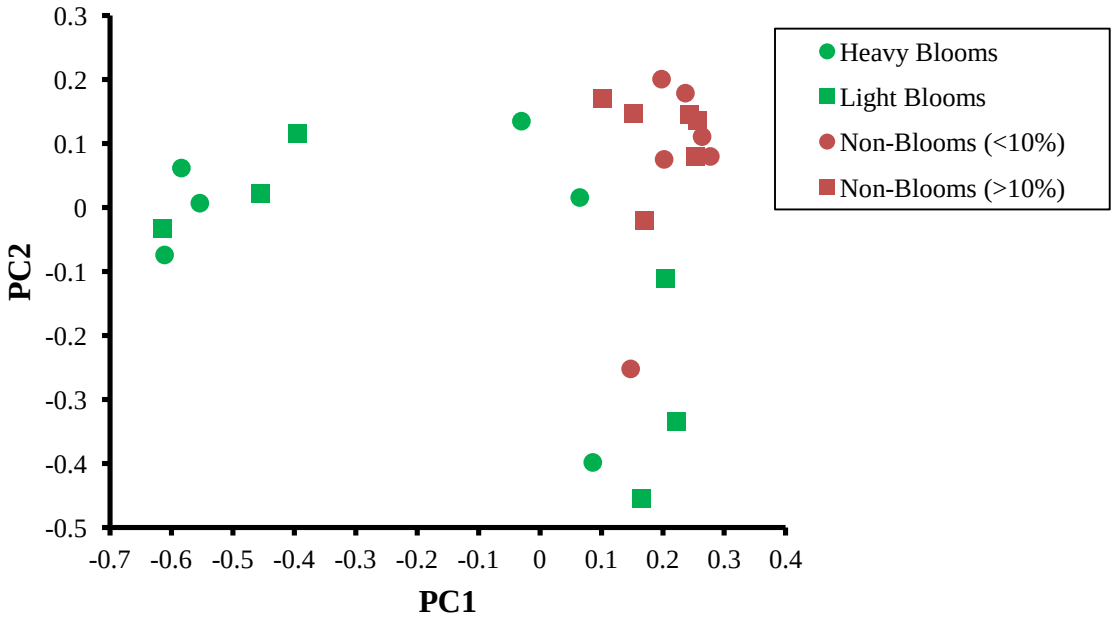
Supplementary Figure 22. Conventional PCR Reaction of the *mcyE* toxin gene with DNA isolated from all 108 samples. Figures **A-I** represent all 108 samples. 100bp ladder was used. P1 is positive control CPCC 300 *Microcystis aeruginosa*, P2 is positive control CPCC 299 *Microcystis aeruginosa* and N is negative control *E.coli* MG1655.

Supplementary Table 13. Concentrations of Shotgun Metagenomic Sequencing Samples

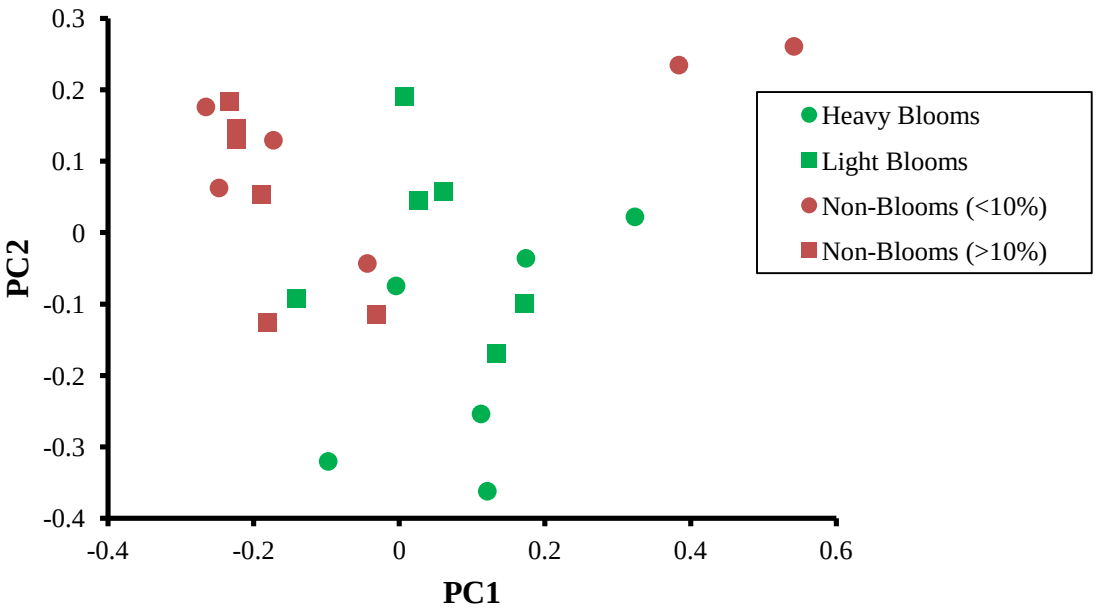
Bloom Samples	Concentration (µg/mL)	Non-Bloom Samples	Concentration (µg/mL)
ON17	0.81	**ON2**	0.2
ON26	0.20	**ON5**	0.86
ON27	0.53	**ON12**	0.53
ON29	0.31	**ON31**	0.20
ON48	0.20	**ON35**	0.20
ON49	0.48	**ON36**	0.20
ON55	0.20	**ON37**	0.20
ON59	1.58	**ON50**	1.42
ON88	0.20	**ON69**	0.21
ON100	0.20	**ON79**	0.20
ON101	0.20	**ON80**	0.21
ON105	0.20	**ON96**	0.20



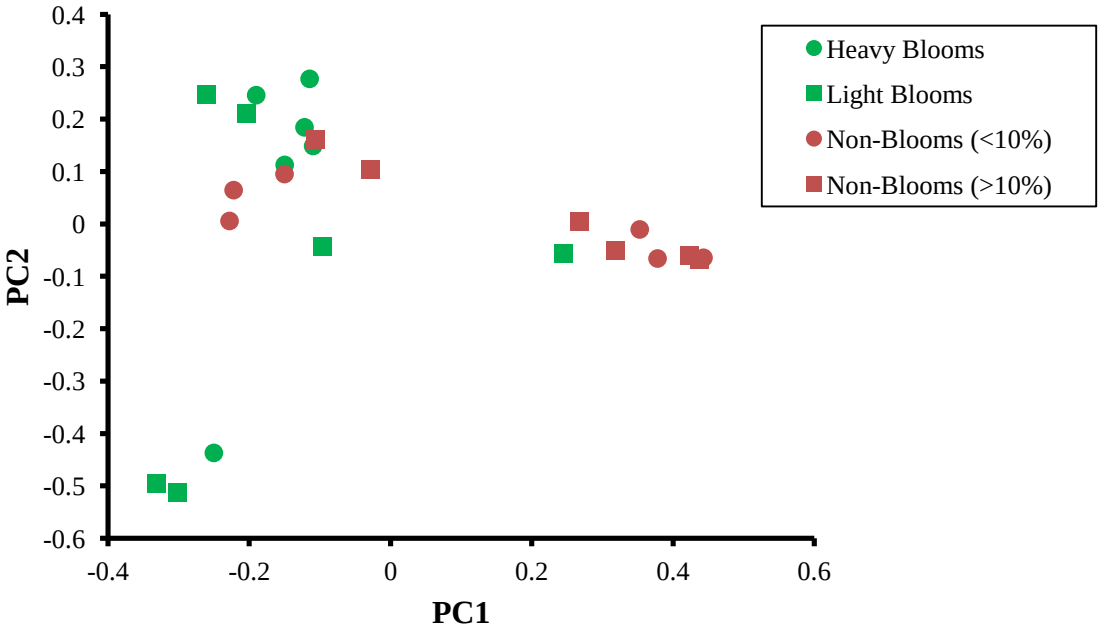
(A) Class



(B) Order



(C) Family



(D) Species

Supplementary Figure 23. PCoA Plots Utilizing Bray-Curtis Dissimilarity Measure at Each of the Respective Taxa Levels.

Appendix B. Chapter 3 Supplement

Supplementary Table 14. Number of Signature Protein Hits in Bloom Samples

Sample	Phyla	Number of Hits	Number of Normalized Hits
ON17	Actinobacteria	1127	827
	Cyanobacteria	719	527
	Alphaproteobacteria	383	281
ON26	Cyanobacteria	3025	844
	Planctomycetes	1480	413
	Firmicutes	181	51
ON27	Cyanobacteria	4951	1692
	Firmicutes	191	65
	Bacteroidetes	141	48
ON29	Cyanobacteria	1248	294
	Alphaproteobacteria	281	66
	Firmicutes	60	14
ON48	Cyanobacteria	2726	795
	Alphaproteobacteria	427	124
	Bacteroidetes	296	86
ON49	Cyanobacteria	4639	1079
	Alphaproteobacteria	197	46
	Planctomycetes	243	57
ON55	Cyanobacteria	3078	1418
	Actinobacteria	212	98
	Firmicutes	85	39
ON59	Cyanobacteria	3199	1860
	Gammaproteobacteria	949	552
	Proteobacteria	207	120
ON88	Cyanobacteria	2739	971
	Alphaproteobacteria	157	56
	Firmicutes	149	53
ON100	Cyanobacteria	2832	904
	Alphaproteobacteria	463	148
	Bacteroidetes	405	129
ON101	Cyanobacteria	8279	2638
	Actinobacteria	266	85
	Firmicutes	185	59
ON105	Alphaproteobacteria	1022	325
	Cyanobacteria	733	233
	Actinobacteria	625	198

Supplementary Table 15. Number of Signature Protein Hits in Non-Bloom Samples

Sample	Phyla	Number of Hits	Number of Normalized Hits
ON2	Bacteroidetes	778	677
	Alphaproteobacteria	104	91
	Firmicutes	54	47
ON5	Proteobacteria	171	90
	Gammaproteobacteria	123	65
	Bacteroidetes	84	44
ON12	Alphaproteobacteria	363	131
	Planctomycetes	153	55
	Firmicutes	129	47
ON31	Cyanobacteria	208	90
	Planctomycetes	134	58
	Alphaproteobacteria	90	39
ON34	Cyanobacteria	414	172
	Actinobacteria	317	132
	Firmicutes	197	82
ON36	Actinobacteria	313	77
	Alphaproteobacteria	89	22
	Firmicutes	86	21
ON37	Planctomycetes	456	231
	Cyanobacteria	292	148
	Firmicutes	53	27
ON50	Cyanobacteria	987	380
	Planctomycetes	232	89
	Firmicutes	151	58
ON69	Cyanobacteria	1851	630
	Firmicutes	201	68
	Planctomycetes	175	60
ON79	Cyanobacteria	925	306
	Firmicutes	156	52
	Planctomycetes	115	38
ON80	Cyanobacteria	1382	348
	Actinobacteria	346	87
	Firmicutes	285	72
ON96	Gammaproteobacteria	276	224
	Actinobacteria	75	61
	Firmicutes	46	37

References

Andrews, S., (2010) FastQC: a quality control tool for high throughput sequence data.

Bao, E., T. Jiang, I. Kaloshian & T. Girke, (2011) SEED: efficient clustering of next-generation sequences. *Bioinformatics* **27**: 2502-2509.

Baron-Sola, A., Y. Ouahid & F.F. del Campo, (2012) Detection of potentially producing cylindrospermopsin and microcystin strains in mixed populations of cyanobacteria by simultaneous amplification of cylindrospermopsin and microcystin gene regions. *Ecotoxicology and Environmental Safety* **75**: 102-108.

Bautista, A.C., C.E. Moore, Y. Lin, M.G. Cline, N. Benitah & B. Puschner, (2015) Hepatopathy following consumption of a commercially available blue-green algae dietary supplement in a dog. *BMC Veterinary Research* **11**: 136.

Berry, M.A., T.W. Davis, R.M. Cory, M.B. Duhaime, T.H. Johengen, G.W. Kling, J.A. Marino, P.A. Den Uyl, D. Gossiaux, G.J. Dick & V.J. Denef, (2017) Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities. *Environmental Microbiology* **19**: 1149-1162.

Blaha, L., P. Babica & B. Marsalek, (2009) Toxins produced in cyanobacterial water blooms - toxicity and risks. *Interdisciplinary Toxicology* **2**: 36-41.

Bolger, A.M., M. Lohse & B. Usadel, (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Botts, S., (2016) Exploring microbial community dynamics: Positive selection for gain of RpoS function in Escherichia coli & microbial profiling of the Niagara Region. In: Department of Biology. McMaster University.

Bray, J.R. & J.T. Curtis, (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27**: 325-349.

Bryant, D.A. & N.U. Frigaard, (2006) Prokaryotic photosynthesis and phototrophy illuminated. *Trends in Microbiology* **14**: 488-496.

Buchfink, B., C. Xie & D.H. Huson, (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59-60.

Cai, H.Y., Z.S. Yan, A.J. Wang, L.R. Krumholz & H.L. Jiang, (2013) Analysis of the attached microbial community on mucilaginous cyanobacterial aggregates in the eutrophic Lake Taihu reveals the importance of Planctomycetes. *Microbial Ecology* **66**: 73-83.

Cai, Y. & C.P. Wolk, (1997) Nitrogen deprivation of Anabaena sp. strain PCC 7120 elicits rapid activation of a gene cluster that is essential for uptake and utilization of nitrate. *Journal of Bacteriology* **179**: 258-266.

Campos, A. & V. Vasconcelos, (2010) Molecular mechanisms of microcystin toxicity in animal cells. *International Journal of Molecular Sciences* **11**: 268-287.

Caporaso, J.G., J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Pena, J.K. Goodrich, J.I. Gordon, G.A. Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C.A. Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W.A. Walters, J. Widmann, T.

Yatsunenko, J. Zaneveld & R. Knight, (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336.

Carmichael, W.W., S.M. Azevedo, J.S. An, R.J. Molica, E.M. Jochimsen, S. Lau, K.L. Rinehart, G.R. Shaw & G.K. Eaglesham, (2001) Human fatalities from cyanobacteria: chemical and biological evidence for cyanotoxins. *Environmental Health Perspectives* **109**: 663-668.

Cavalier-Smith, T., (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *International Journal of Systematic and Evolutionary Microbiology* **52**: 7-76.

Chakravorty, S., D. Helb, M. Burday, N. Connell & D. Alland, (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods* **69**: 330-339.

Chen, K. & L. Pachter, (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology* **1**: 106-112.

Chislock, M.F., K.L. Sharp & A.E. Wilson, (2014) Cylindrospermopsis racibarskii dominates under very low and high nitrogen-to-phosphorus ratios. *Water Research* **49**: 207-214.

Chorus, I. & J. Bartram, (1999) Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management. *WHO*.

Coutinho, F., D.A. Tschoeke, F. Thompson & C. Thompson, (2016) Comparative genomics of Synechococcus and proposal of the new genus Parasynechococcus. *PeerJ* **4**: e1522.

Cumino, A.C., C. Marcozzi, R. Barreiro & G.L. Salerno, (2007) Carbon cycling in Anabaena sp. PCC 7120. Sucrose synthesis in the heterocysts and possible role in nitrogen fixation. *Plant Physiology* **143**: 1385-1397.

DeBruyn, J.M., L.T. Nixon, M.N. Fawaz, A.M. Johnson & M. Radosevich, (2011) Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Applied and Environmental Microbiology* **77**: 6295-6300.

DeSantis, T.Z., P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu & G.L. Andersen, (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**: 5069-5072.

Dixon, P., (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**: 927-930.

Dolman, A.M., J. Rucker, F.R. Pick, J. Fastner, T. Rohrlack, U. Mischke & C. Wiedner, (2012) Cyanobacteria and Cyanotoxins: The Influence of Nitrogen versus Phosphorus. *PloS One* **7**.

Eiler, A. & S. Bertilsson, (2004) Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environmental Microbiology* **6**: 1228-1243.

Elliott, J.A., I.D. Jones & S.J. Thackeray, (2006) Testing the sensitivity of phytoplankton communities to changes in water temperature and nutrient load, in a temperate lake. *Hydrobiologia* **559**: 401-411.

Eriksson, J.E., D. Toivola, J.A.O. Meriluoto, H. Karaki, Y.G. Han & D. Hartshorne, (1990) Hepatocyte Deformation Induced by Cyanobacterial Toxins Reflects Inhibition of Protein Phosphatases. *Biochemical and Biophysical Research Communications* **173**: 1347-1353.

Euzeby, J.P., (1997) List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *International Journal of Systematic Bacteriology* **47**: 590-592.

Fogg, G.E., (1969) The physiology of an algal nuisance. *Proceedings of the Royal Society B: Biological Sciences* **173**: 175-189.

Gao, B. & R.S. Gupta, (2012) Microbial systematics in the post-genomics era. *Antonie van Leeuwenhoek* **101**: 45-54.

Genitsaris, S., K.A. Kormas & M. Moustaka-Gouni, (2011) Airborne algae and cyanobacteria: occurrence and related health effects. *Frontiers in Bioscience* **3**: 772-787.

Gerdtzen, Z.P., J.C. Salgado, A. Osses, J.A. Asenjo, I. Rapaport & B.A. Andrews, (2009) Modeling heterocyst pattern formation in cyanobacteria. *BMC Bioinformatics* **10 Suppl 6**: S16.

Ghai, R., C.M. Mizuno, A. Picazo, A. Camacho & F. Rodriguez-Valera, (2014) Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* **23**: 6073-6090.

Gibbons, N.E. & R.G.E. Murray, (1978) Validation of Cyanobacteriales Stanier in Gibbons and Murray 1978 as a New Order of the Kingdom Procaryote Murray 1968, and the use of Neuter Plural Endings for Photobacteria and Scotobacteria classes nov. Gibbons and Murray 1978. *International Journal of Systematic Bacteriology* **28**: 332-333.

Graham, J.L., G.M. Foster, T.J. Williams, A.R. Kramer & T.D. Harris, (2017) Occurence of cyanobacteria, microcystin, and taste-and-odor compounds in Cheney Reservoir, Kansas, 2001-16. *Scientific Investigations Report 2017-5016.*

Graham, J.L., K.A. Loftin, M.T. Meyer & A.C. Ziegler, (2010) Cyanotoxin mixtures and taste-and-odor compounds in cyanobacterial blooms from the Midwestern United States. *Environmental Science & Technology* **44**: 7361-7368.

Gupta, R.S., (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. *International Journal of Systematic and Evolutionary Microbiology* **59**: 2510-2526.

Gupta, R.S. & E. Griffiths, (2002) Critical issues in bacterial phylogeny. *Theoretical Population Biology* **61**: 423-434.

Gupta, R.S., P.I. Mark, C. Chandrasekera & V. Johari, (2003) Molecular signatures in protein sequences that are characteristic of cyanobacteria and plastid homologues. *International Journal of Systematic and Evolutionary Microbiology* **53**: 1833-1842.

Gupta, R.S. & D.W. Mathews, (2010) Signature proteins for the major clades of Cyanobacteria. *BMC Evolutionary Biology* **10**: 24.

Haukka, K., E. Kolmonen, R. Hyder, J. Hietala, K. Vakkilainen, T. Kairesalo, H. Haario & K. Sivonen, (2006) Effect of nutrient loading on bacterioplankton community composition in lake mesocosms. *Microbial Ecology* **51**: 137-146.

Havens, K.E., (2008) Cyanobacteria blooms: effects on aquatic ecosystems. *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* **619**: 733-747.

Hawkins, P.R., M.T. Runnegar, A.R. Jackson & I.R. Falconer, (1985) Severe hepatotoxicity caused by the tropical cyanobacterium (blue-green alga) Cylindrospermopsis raciborskii (Woloszynska) Seenaya and Subba Raju isolated from a domestic water supply reservoir. *Applied and Environmental Microbiology* **50**: 1292-1295.

Heinz, S., P. Liauw, J. Nickelsen & M. Nowaczyk, (2016) Analysis of photosystem II biogenesis in cyanobacteria. *Biochimica et biophysica acta* **1857**: 274-287.

Helliwell, K., A. Lawrence, A. Holzer, U. Kudahl, S. Sasso, B. Krautler, D. Scanlan, M. Warren & A. Smith, (2016) Cyanobacteria and eukaryotic algae use different chemical variants of vitamin B12. *Current Biology* **26**: 999-1008.

Hisbergues, M., G. Christiansen, L. Rouhiainen, K. Sivonen & T. Borner, (2003) PCR-based identification of microcystin-producing genotypes of different cyanobacterial genera. *Archives of Microbiology* **180**: 402-410.

Hoffmann, L., (2005) Nomenclature of Cyanophyta/Cyanobacteria: roundtable on the unification of the nomenclature under the Botanical and Bacteriological Codes. *Algological Studies* **117**: 13-29.

Hoffmann, L., J. Komarek & J. Kastovsky, (2005) System of cyanoprokaryotes (cyanobacteria) state in 2004. *Algological Studies* **117**: 95-115.

Hoiczyk, E. & A. Hansel, (2000) Cyanobacterial cell walls: news from an unusual prokaryotic envelope. *Journal of Bacteriology* **182**: 1191-1199.

Hotto, A.M., M.F. Satchwell & G.L. Boyer, (2007) Molecular characterization of potential microcystin-producing cyanobacteria in Lake Ontario embayments and nearshore waters. *Applied and Environmental Microbiology* **73**: 4570-4578.

Howard-Azzeh, M., L. Shamseer, H.E. Schellhorn & R.S. Gupta, (2014) Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. *Photosynthesis Research* **122**: 171-185.

Hudnell, H.K., (2008) *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs*. Springer.

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

Huson, D.H., A.F. Auch, J. Qi & S.C. Schuster, (2007) MEGAN analysis of metagenomic data. *Genome Research* **17**: 377-386.

Ionescu, D., M. Hindiyeh, H. Malkawi & A. Oren, (2010) Biogeography of thermophilic cyanobacteria: insights from the Zerka Ma'in hot springs (Jordan). *FEMS Microbiology Ecology* **72**: 103-113.

Jensen, L.J., P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks & P. Bork, (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36**: D250-254.

Johnk, K.D., J. Huisman, J. Sharples, B. Sommeijer, P.M. Visser & J.M. Stroom, (2008) Summer heatwaves promote blooms of harmful cyanobacteria. *Global Change Biology* **14**: 495-512.

Jonsson, V., T. Osterlund, O. Nerman & E. Kristiansson, (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* **17**: 78.

Kahru, M., J.M. Leppanen & O. Rud, (1993) Cyanobacterial Blooms Cause Heating of the Sea-Surface. *Marine Ecology Progress Series* **101**: 1-7.

Kauff, F. & B. Budel, (2010) Phylogeny of Cyanobacteria: An Overview. In: Progress in Botany. Springer-Verlag Berlin Hedelberg, pp. 209-224.

Kellmann, R., T.K. Mihali & B.A. Neilan, (2009) Identification of a Saxitoxin Biosynthesis Gene with a History of Frequent Horizontal Gene Transfers (vol 67, pg 526, 2008). *Journal of Molecular Evolution* **68**: 292-292.

Kleinteich, J., S.A. Wood, F.C. Kupper, A. Camacho, A. Quesada, T. Frickey & D.R. Dietrich, (2012) Temperature-related changes in polar cyanobacterial mat diversity and toxin production. *Nature Climate Change* **2**: 356-360.

Kolmonen, E., K. Sivonen, J. Rapala & K. Haukka, (2004) Diversity of cyanobacteria and heterotrophic bacteria in cyanobacterial blooms in Lake Joutikas, Finland. *Aquatic Microbiology* **36**: 201-211.

Kozdeba, M., J. Borowczyk, E. Zimolag, M. Wasylewski, D. Dziga, Z. Madeja & J. Drukala, (2014) Microcystin-LR affects properties of human epidermal skin cells crucial for regenerative processes. *Toxicon : Official Journal of the International Society on Toxinology* **80**: 38-46.

Krishnamurthy, T., W.W. Carmichael & E.W. Sarver, (1986) Toxic peptides from freshwater cyanobacteria (blue-green algae). I. Isolation, purification and characterization of peptides from Microcystis aeruginosa and Anabaena flos-aquae. *Toxicon : Official Journal of the International Society on Toxinology* **24**: 865-873.

Kurmayer, R., (2011) The Toxic Cyanobacterium Nostoc Sp. Strain 152 Produces Highest Amounts of Microcystin and Nostophycin under Stress Conditions. *Journal of Phycology* **47**: 200-207.

Louati, I., N. Pascault, D. Debroas, C. Bernard, J.F. Humbert & J. Leloup, (2015) Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PloS One* **10**: e0140614.

Love, M.I., W. Huber & S. Anders, (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Lozupone, C. & R. Knight, (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**: 8228-8235.

Lozupone, C.A., M. Hamady, S.T. Kelley & R. Knight, (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that

structure microbial communities. *Applied and Environmental Microbiology* **73**: 1576-1585.

Magoc, T. & S.L. Salzberg, (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Mehta, K.K., N.H. Evitt & J.R. Swartz, (2015) Chemical lysis of cyanobacteria. *Journal of Biological Engineering* **9**: 10.

Moffitt, M.C. & B.A. Neilan, (2001) On the presence of peptide synthetase and polyketide synthase genes in the cyanobacterial genus Nodularia. *FEMS Microbiological Letters* **196**: 207-214.

Mohamed, Z.A. & A.M. Al Shehri, (2011) Occurence and germination of dinoflagellate cysts in surface sediments from the Red Sea off the coasts of Saudi Arabia. *Oceanologia* **53**: 121-136.

Muhseen, Z.T., Q. Xiong, Z. Chen & F. Ge, (2015) Proteomics studies on stress responses in diatoms. *Proteomics* **15**: 3943-3953.

Mussagy, A., H. Annadotter & G. Cronberg, (2006) An experimental study of toxin production in Arthrospira fusiformis (Cyanophyceae) isolated from African waters. *Toxicon : Official Journal of the International Society on Toxinology* **48**: 1027-1034.

Muyzer, G., E.C. de Waal & A.G. Uitterlinden, (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* **59**: 695-700.

Naushad, H.S., B. Lee & R.S. Gupta, (2014) Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera Dickeya, Pectobacterium and Brenneria. *International Journal of Systematic and Evolutionary Microbiology* **64**: 366-383.

Nelissen, B., A. Wilmotte, J.M. Neefs & R. Dewachter, (1994) Phylogenetic-Relationships among Filamentous Helical Cyanobacteria Investigated on the Basis of 16s Ribosomal-Rna Gene Sequence-Analysis. *Systematic and Applied Microbiology* **17**: 206-210.

Nubel, U., F. Garcia-Pichel & G. Muyzer, (1997) PCR primers to amplify 16S rRNA genes from cyanobacteria. *Applied and Environmental Microbiology* **63**: 3327-3332.

Oren, A. & G.M. Garrity, (2014) Proposal to change General Consideration 5 and Principle 2 of the International Code of Nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* **64**: 309-310.

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

Paerl, H.W. & J.F. Ustach, (1982) Blue-Green-Algal Scums - an Explanation for Their Occurrence during Fresh-Water Blooms. *Limnology and Oceanogr*aphy**27**: 212-217.

Pandey, K.D., S.P. Shukla, P.N. Shukla, D.D. Giri, J.S. Singh, P. Singh & A.K. Kashyap, (2004) Cyanobacteria in Antarctica: ecology, physiology and cold adaptation. *Cellular and Molecular Biology* **50**: 575-584.

Parte, A.C., (2014) LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Research* **42**: D613-616.

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

Pinevich, A.V., (2015) Proposal to consistently apply the International Code of Nomenclature of Prokaryotes (ICNP) to names of the oxygenic photosynthetic bacteria (cyanobacteria), including those validly published under the International Code of Botanical Nomenclature (ICBN)/International Code of Nomenclature for algae, fungi and plants (ICN), and proposal to change Principle 2 of the ICNP. *International Journal of Systematic and Evolutionary Microbiology* **65**: 1070-1074.

Pizzetti, I., B.M. Fuchs, G. Gerdts, A. Wichels, K.H. Wiltshire & R. Amann, (2011) Temporal variability of coastal Planctomycetes clades at Kabeltonne station, North Sea. *Applied and Environmental Microbiology* **77**: 5009-5017.

Qin, B., G. Zhu, G. Gao, Y. Zhang, W. Li, H.W. Paerl & W.W. Carmichael, (2010) A drinking water crisis in Lake Taihu, China: linkage to climatic variability and lake management. *Environmental Management* **45**: 105-112.

Quince, C., A. Lanzen, T.P. Curtis, R.J. Davenport, N. Hall, I.M. Head, L.F. Read & W.T. Sloan, (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6**: 639-641.

Rae, B., B. Long, L.F. Whitehead, B. Forster, M. Badger & G. Price, (2013) Cyanobacterial carboxysomes: microcompartments that facilitate CO2 fixation. *Journal of Molecular Microbiology and Biotechnology* **23**: 300-307.

Rai, A.N., B. Bergman & B. Rasmussen, (2002) *Cyanobacteria in Symbioses*, p. 335. Kluwer Publications, Boca Raton, FL, USA.

Ranjan, R., A. Rani, A. Metwally, H.S. McGee & D.L. Perkins, (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications* ***469***: 967-977.

Rantala-Ylinen, A., S. Kana, H. Wang, L. Rouhiainen, M. Wahlsten, E. Rizzi, K. Berg, M. Gugger & K. Sivonen, (2011) Anatoxin-a synthetase gene cluster of the cyanobacterium Anabaena sp. strain 37 and molecular methods to detect potential producers. *Applied and Environmental Microbiology* **77**: 7271-7278.

Rantala, A., P. Rajaniemi-Wacklin, C. Lyra, L. Lepisto, J. Rintala, J. Mankiewiez-Boczek & K. Sivonen, (2006) Detection of microcystin-producing cyanobacteria in Finnish lakes with genus-specific microcystin synthetase gene E (mcyE) PCR and associations with environmental factors. *Applied and Environmental Microbiology* **72**: 6101-6110.

Rasmussen, B., I.R. Fletcher, J.J. Brocks & M.R. Kilburn, (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**: 1101-1104.

Rastogi, R.P., D. Madamwar & A. Incharoensakdi, (2015) Bloom Dynamics of Cyanobacteria and Their Toxins: Environmental Health Impacts and Mitigation Strategies. *Frontiers in Microbiology* **6**: 1254.

Rellan, S., J. Osswald, M. Saker, A. Gago-Martinez & V. Vasconcelos, (2009) First detection of anatoxin-a in human and animal dietary supplements containing cyanobacteria. *Food and Chemical Toxicology : an International Journal Published for the British Industrial Biological Research Association* **47**: 2189-2195.

Reynolds, C.S., (1987) Cyanobacterial Water-Blooms. *Adv Bot Res* **13**: 67-143.

Rinta-Kanto, J.M., A.J. Ouellette, G.L. Boyer, M.R. Twiss, T.B. Bridgeman & S.W. Wilhelm, (2005) Quantification of toxic Microcystis spp. during the 2003 and 2004 blooms in western Lake Erie using quantitative real-time PCR. *Environmental Science & Technology* **39**: 4198-4205.

Rinta-Kanto, J.M. & S.W. Wilhelm, (2006) Diversity of microcystin-producing cyanobacteria in spatially isolated regions of Lake Erie. *Applied and Environmental Microbiology* **72**: 5083-5085.

Rippka, R., (1988) Recognition and Identification of Cyanobacteria. *Method Enzymology* **167**.

Rippka, R., J. Deruelles, J.B. Waterbury, M. Herdman & R.Y. Stanier, (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Journal of General Microbiol*ogy**111**: 1-61.

Robarts, R.D. & T. Zohary, (1987) Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *New Zealand Journal of Marine and Freshwater Research* **21**: 391-399.

Rodriguez, R.L. & K.T. Konstantinidis, (2014) Estimating coverage in metagenomic data sets and why it matters. *Isme Journal* **8**: 2349-2351.

Runnegar, M., N. Berndt & N. Kaplowitz, (1995) Microcystin Uptake and Inhibition of Protein Phosphatases - Effects of Chemoprotectants and Self-Inhibition in Relation to Known Hepatic Transporters. *Toxicology and Applied Pharmacology* **134**: 264-272.

Runnegar, M.T., I.R. Falconer & J. Silver, (1981) Deformation of isolated rat hepatocytes by a peptide hepatotoxin from the blue-green alga microcystis aeruginosa. *Naunyn-Schmiedeberg's Archives of Pharmacology* **317**: 268-272.

Saker, M.L., A.D. Thomas & J.H. Norton, (1999) Cattle mortality attributed to the toxic cyanobacterium Cylindrospermopsis raciborskii in an outback region of north Queensland. *Environmental Toxicol*ogy **14**: 179-182.

Sanschagrin, S. & E. Yergeau, (2014) Next-generation sequencing of 16S ribosomal RNA gene amplicons. *Journal of visualized experiments : JoVE*.

Sanseverino, I., D. Conduto, L. Pozzoli, S. Dobricic & T. Lettieri, (2016) Algal bloom and its economic Impact. In. Italy: Joint Research Centre, the European Commission's in-house science service.

Sayers, E.W., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, V. Miller, I.

Mizrachi, J. Ostell, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, E. Yaschenko & J. Ye, (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37**: D5-15.

Schindler, D.W., R.E. Hecky, D.L. Findlay, M.P. Stainton, B.R. Parker, M.J. Paterson, K.G. Beaty, M. Lyng & S.E.M. Kasian, (2008) Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 11254-11258.

Sheath, R.G. & A.D. Steinman, (1981) A checklist of freshwater algae of the Northwest Territories, Canada. *NRC Research Press*.

Shia, L., Y. Cai, X. Wang, P. Li, Y. Yu & F.X. Kong, (2011) Community structure of bacteria associated with Microcystis colonies from cyanobacterial blooms. *Journal of Freshwater Ecology* **25**: 193-203.

Singh, S., P.K. Rai, R. Chau, A.K. Ravi, B.A. Neilan & R.K. Asthana, (2015) Temporal variations in microcystin-producing cells and microcystin concentrations in two fresh water ponds. *Water Research* **69**: 131-142.

Sonntag, S. & I. Hense, (2011) Phytoplankton behavior affects ocean mixed layer dynamics through biological-physical feedback mechanisms. *Geophysical Research Lett*ers **38**.

Sophie, J.W., ZHenijiang X., A. Amir, S. Peddada, K. Bittinger, A. Gonzalez, C. Lozupone, J.R. Zaneveld, Y. Vasquez-Baeza, A. Birmingham & R. Knight, (2015) Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints*.

Stanier, R.Y., W.R. Sistrom, T.A. Hansen, B.A. Whitton, R.W. Castenholz, N. Pfennig, V.N. Gorlenko, E.N. Kondroatieva, K.E. Eimhjellen, R. Whittenbury, R.L. Gherna & H.G. Truper, (1978) Proposal to Place the Nomenclature of the Cyanobacteria (Blue-Green Algae) under the Rules of the International Code of Nomenclature of Bacteria. *International Journal of Systematic Bacteriology* **28**: 335-336.

Steffen, M.M., Z. Li, T.C. Effler, L.J. Hauser, G.L. Boyer & S.W. Wilhelm, (2012) Comparative metagenomics of toxic freshwater cyanobacteria bloom communities on two continents. *PloS One* **7**: e44002.

Stewart, I., A.A. Seawright & G.R. Shaw, (2008) Cyanobacterial poisoning in livestock, wild mammals and birds - an overview. *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* **619**: 613-637.

Tang, K., W.L. Ding, A. Hoppner, C. Zhao, L. Zhang, Y. Hontani, J.T. Kennis, W. Gartner, H. Scheer, M. Zhou & K.H. Zhao, (2015) The terminal phycobilisome emitter, LCM: A light-harvesting pigment with a phytochrome chromophore. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 15880-15885.

Te, S.H., B.F. Tan, J.R. Thompson & K.Y. Gin, (2017) Relationship of Microbiota and Cyanobacterial Secondary Metabolites in Planktothricoides-Dominated Bloom. *Environmental Science & Technology* **51**: 4199-4209.

Tencalla, F.G., D.R. Dietrich & C. Schlatter, (1994) Toxicity of Microcystis-Aeruginosa Peptide Toxin to Yearling Rainbow-Trout (Oncorhynchus-Mykiss). *Aquatic Toxicology* **30**: 215-224.

Thomas, J., (1970) Absence of the pigments of photosystem II of photosynthesis in heterocysts of a blue-green alga. *Nature* **228**: 181-183.

Tripathi, S.N., I.K. Chung & J.A. Lee, (2007) Diversity and characteristics of terrestrial cyanobacteria near Gimhae city, Korea. *J Plant Biol* **50**: 50-59.

Vaitomaa, J., A. Rantala, K. Halinen, L. Rouhiainen, P. Tallberg, L. Mokelke & K. Sivonen, (2003) Quantitative real-time PCR for determination of microcystin synthetase e copy numbers for microcystis and anabaena in lakes. *Applied and Environmental Microbiology* **69**: 7289-7297.

van Apeldoorn, M.E., H.P. van Egmond, G.J. Speijers & G.J. Bakker, (2007) Toxins of cyanobacteria. *Molecular nutrition & food research* **51**: 7-60.

Walsby, A.E., (1994) Gas Vesicles. *Microbiological Reviews* **58**: 94-144.

Wan, Y., X. Ruan, Y. Zhang & R. Li, (2017) Illumina sequencing-based analysis of sediment bacteria community in different trophic status freshwater lakes. *MicrobiologyOpen*.

Warren-Rhodes, K.A., K.L. Rhodes, S.B. Pointing, S.A. Ewing, D.C. Lacap, B. Gomez-Silva, R. Amundson, E.I. Friedmann & C.P. McKay, (2006) Hypolithic cyanobacteria, dry limit of photosynthesis, and microbial ecology in the hyperarid Atacama Desert. *Microbial Ecology* **52**: 389-398.

Watanabe, M.F., S. Oishi, K. Harda, K. Matsuura, H. Kawai & M. Suzuki, (1988) Toxins contained in Microcystis species of cyanobacteria (blue-green algae). *Toxicon : Official Journal of the International Society on Toxinology* **26**: 1017-1025.

Wehr, J.D., R.G. Sheath & J.P. Kociolek, (2015) *Freshwater Algae of North America*. Elsevier.

Weiss, S.J., Z. Xu, A. Amir, S. Peddada, K. Bittinger, A. Gonzalez, C. Lozupone, J. Zaneveld, Y. Vazquez-Baeza, A. Birmingham & R. Knight, (2015) Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints*.

Wolk, C.P., J. Thomas, P.W. Shaffer, S.M. Austin & A. Galonsky, (1976) Pathway of nitrogen metabolism after fixation of 13N-labeled nitrogen gas by the cyanobacterium, Anabaena cylindrica. *The Journal of Biological Chemistry* **251**: 5027-5034.

Zanchett, G. & E.C. Oliveira-Filho, (2013) Cyanobacteria and cyanotoxins: from impacts on aquatic ecosystems and human health to anticarcinogenic effects. *Toxins* **5**: 1896-1917.

Zegura, B., A. Straser & M. Filipic, (2011) Genotoxicity and potential carcinogenicity of cyanobacterial toxins - a review. *Mutation Research* **727**: 16-41.

Zhu, L., A. Zancarini, I. Louati, S. De Cesare, C. Duval, K. Tambosco, C. Bernard, D. Debroas, L. Song, J. Leloup & J.F. Humbert, (2016) Bacterial Communities Associated with Four Cyanobacterial Genera Display Structural and Functional Differences: Evidence from an Experimental Approach. *Frontiers in Microbiology* **7**: 1662.

Zwart, G., R. Huismans, M. Agterveld, Y. Peer, P. Rijk, H. Eenhoorn, G. Muyzer, E. Hannen, H. Gons & H. Laanbroek, (1998) Divergent members of the bacterial division Verrucomicrobiales in a temperate freshwater *Microbial Ecology* **25**: 159-169.