Summer 2021

Investigation into Bacterial Impairment of Greene and Polk County Water Systems

John C. Kincaid
Missouri State University, Kincaid2015@live.missouristate.edu

INVESTIGATION INTO BACTERIAL IMPAIRMENT OF GREENE AND POLK COUNTY WATER SYSTEMS

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Biology

By

John Kincaid

July 2021

INVESTIGATION INTO BACTERIAL IMPAIRMENT OF GREENE AND POLK COUNTY WATER SYSTEMS

Biology

Missouri State University, July 2021

Master of Science, Biology

John Kincaid

ABSTRACT

Bacterial impairment of water systems is a major issue facing mankind. Bacteria that are introduced into a system have the potential to cause harmful diseases to wildlife and humans. In Greene and Polk counties, many water systems have become bacterially impaired over the years. Despite this, little is known about the contamination of known harmful bacteria in this region. To address this issue, I investigated the presence of known human pathogens across water systems in these two counties, many of which have displayed high levels of *E. coli* and fecal indicator microorganisms over many years. I used a high-throughput sequencing approach to determine what pathogens were present at these locations. Overall, it was found that pathogens such as *Legionella*, *Shigella/E. coli*, and *Yersinia* were found in high abundance in one site in the Little Sac watershed and one site in the Pearson Creek watershed. Additionally, these pathogens were found to be abundant at Sequiota Park's spring at different time points across a year. qPCR was also used to determine the presence of four commonly reported cyanotoxin genes: cylindrospermopsin, anatoxin-a, microcystin, and saxitoxin. Cyanotoxins are produced by Cyanobacteria and are known to cause various diseases. Of the four tested, only the anatoxin-a gene was detected. It was found that there was an increase in anatoxin-a containing cyanobacteria at site LS_171. Additionally, this site displayed increased abundance in *Planktothrix*-related sequences, suggesting that the increase in the anatoxin-a gene could be due to the increase in *Planktothrix* at this site. Conclusively, utilizing these methods may be able to help prevent diseases outbreaks associated with these pathogens.

KEYWORDS: bacterial impairment, bacterial pathogens, cyanobacteria, cyanotoxins, anatoxin-a, high-throughput sequencing, qPCR

INVESTIGATION INTO BACTERIAL IMPAIRMENT OF GREENE AND POLK COUNTY WATER SYSTEMS

By

John Kincaid

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master's of Science, Biology

July 2021

Approved:

Babur S. Mirza, Ph.D., Thesis Committee Chair

Christopher R. Lupfer, Ph.D., Committee Member

Thomas E. Tomasi, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to first and foremost thank Dr. Babur Mirza for all the time and effort he has put into helping me become the best scientist that I can be. I could not have gotten this far without all of your guidance. I could never thank you enough for everything you have done for me. Also, I would like to thank my committee members, Dr. Tom Tomasi and Dr. Chris Lupfer, for their support throughout my grad school career. I would also like to thank every person who has come through the Mirza lab that has helped me along the way. I especially want to thank Paris Mayhood and Scott McElveen. We all started around the same time in the lab, and I am grateful to have had you two for support when learning the things we did. Also, thank you to everyone at MSU who has been there for me throughout my time here. Your support has helped me get to where I am now.

I also would like to thank my father, Brian Kincaid, who has pushed me to do my best no matter how hard it is. I have you to thank for the man that I have become. Thank you to all of my family for all the love and support you have given me throughout my life.

I dedicate this thesis to my mother, June Kincaid, whom I wish could have been here with us to this day. I love and miss you bunches and bunches.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

OVERVIEW

Harmful Aspects of Bacteria

Waterborne diseases are major health concerns worldwide. Approximately 1.5 million deaths occur worldwide due to waterborne illnesses (WHO, 2019). Although most deaths originate from developing countries, instances of waterborne disease still occur in developed countries like the United States. Fecal contamination due to urbanization and agricultural developments are major causes of water contamination. Frequent disease outbreaks due to waterborne pathogens highlights the significance and need of close monitoring of relative distribution of bacterial pathogens in impaired watersheds.

In general, most bacteria play an important role in survival and wellbeing of humans, such as their role in biogeochemical cycles and the human microbiome. However, some bacteria are pathogenic and can cause serious diseases. Many of these diseases are caused by waterborne pathogens originating from drinking and/or recreational use of water that is contaminated with fecal material. Most of these gastrointestinal bacterial pathogens are transmitted through the fecal-oral route (Browne *et al*., 2017). Waterborne pathogens such as *Salmonella*, *Shigella* and *Yersinia* spp., can cause gastrointestinal illnesses, in particular diarrhea and dysentery. Detection of various waterborne pathogens is critical information for the water resource managers to identify the health risks associated with the potential use of a water resource. On the other hand, as water sources are becoming scarce, it may not be a pragmatic approach to prohibit the recreational use of water resources based on the presence of indicator microorganisms (*E. coli* and fecal coliforms). Hence, studying the relative distribution of various pathogens in a water environment is critical for accurate assessment of health risk associated with the potential use of

a water resource. The detection of waterborne pathogens is even more critical in the karst environments where groundwater can easily be contaminated by old leaky septic tanks and broken sewer lines (Sercu *et al*., 2011). Additionally, some pathogens are known to be carried in the feces of animals and can be transmissible to humans if consumed (Renter *et al*., 2006; Rouffaer *et al*., 2017; Skov *et al*., 2008).

Not all waterborne pathogens cause gastrointestinal infections. Some waterborne pathogens such as *Legionella* spp. can cause respiratory diseases. *Legionella* spp. can persist in natural water environments and cause lung diseases when aerosolized. Inside the human lungs, *Legionella* spp. can infect the resident alveolar macrophages. This will then lead to the development of pneumonia (Newton *et al*., 2010). The health risk associated with these bacteria has increased due to their ability to persist in water pipes through biofilm association, especially those that are used for facility cooling systems. This allows for the bacterium to become easily aerosolized, allowing for entry into the lungs (Fields, Benson, and Besser, 2002).

Some bacterial pathogens can also cause diseases through indirect interaction with the human body by producing toxins that may be ingested, inhaled, or come in direct skin contact. A few examples of these toxin-producing bacteria are cholera-toxin (produced by *Vibrio Cholera*), shiga-toxins (produced by *Shigella/E. coli* that produce Shiga-toxin) and cyanotoxins (produced by *Cyanobacteria* spp.). These toxins are commonly produced by the bacteria as part of their metabolism and can cause human health problems through direct or indirect interactions. Diseases like these mostly occur due to contaminated sources of drinking water or contaminated food.

Description of Study Sites Within Greene and Polk Counties

In the state of Missouri, many streams and lakes have been declared as impaired because of abundance of *E. coli* and fecal indicator microorganisms. These bacterially impaired water systems have been placed on the Missouri's 303(d) list for microbial impairment. This requires local governments to devise a plan to reduce the overall extent of this contamination. This plan is called a Total Maximum Daily Load (TMDL). Once a water system has a TMDL made, it is taken off the 303(d) list. Within Greene and Polk counties, many water systems are on or have been on this list, for example, the Little Sac watershed (LSW) was placed on this list in 1998 (WCO, 2016). It had a TMDL created to reduce contaminants in 2006 (Baffaut, 2006). Several locations within LSW consistently demonstrated high levels of indicator microorganisms (Figure 1).

Similarly, the Pearson Creek watershed is another water system that was placed on this list due to high abundance of *E. coli* in 2006, and this watershed is slated to have a TMDL developed by 2031 (MDNR, 2020; Figure 2). Pearson Creek is a tributary of the James River, which is one of the sources of drinking water for the city of Springfield. Contamination of these water systems can have a direct impact on human health, causing waterborne disease outbreaks if the water is not treated properly.

In addition to above mentioned surface waters, underground water bodies such as Sequiota Spring cave have also shown a high abundance of indicator microorganisms. Sequiota Spring (Figure 3), which is part of a tributary to Lake Springfield, is currently not placed on the 303(d) list despite a high abundance of indicator microorganisms in the past (Bullard *et al*., 2001; WWE, 2001). This can be a potential area of concern due to the recreational use of waters that are connected to this spring, which flows into Galloway creek. These areas are used for

many recreational purposes, so monitoring of these waters is crucial for the safety of the people who use them.

In addition to impaired water systems, constant monitoring and testing of drinking water sources is also imperative for the wellbeing of humans. The city of Springfield draws 80% of its drinking water supply from surface waters such as Fellows, McDaniel, and Stockton lakes. These lakes are connected through the Little Sac River, which raises concerns of contamination of harmful bacteria and high nutrient loadings.

Assessment of Bacterial Communities Using DNA Sequencing

Considering that the city of Springfield's drinking water supply comes mainly from surface waters, it is important that we are consistently monitoring these systems for the potential threat of known pathogens. Prior methods of testing in this area have focused mainly on fecal indicator bacteria, which has many limitations when determining the health risk of a water system. Thus, in this thesis, I investigated the presence of bacteria that are known to cause disease in various water systems in Greene and Polk Counties utilizing a high-throughput DNA sequencing approach.

In order to assess the bacterial communities, genomic DNA from bacteria in the environment must be isolated and then prepared for DNA sequencing (Figure 4). To this end, I collected water samples in 5-gallon sterile containers from the priorly mentioned water systems. This water was then filtered through 0.2 μm filters to isolate the bacteria. The genomic DNA was then extracted from the bacteria using a DNA extraction kit. The DNA was then prepared for sequencing using a library preparation method (Figure 5). This method involves amplifying the

16S rRNA gene and then adding indices to the amplicons to tag what site they were from. The 16S rRNA gene was used due to being highly conserved across all bacteria yet has many variable regions that can decern between species. Once the DNA was prepared, it was sent to be sequenced. Sequences that I retrieved were then screened for quality, and then used for bacterial community analysis.

This thesis focuses on two major parts: i) detection of commonly reported waterborne pathogens (i.e. *Legionella*, *Salmonella*, etc.) in various streams along Green and Polk counties, and ii) investigating the diversity of *Cyanobacteria* genera and cyanotoxin genes within recreational and drinking water systems of the city of Springfield. The aim of these studies is to assess the potential health risks associated with the water use of these resources for drinking and recreational purposes.

Figure 1: Sites of high bacterial contamination throughout the Little Sac watershed.

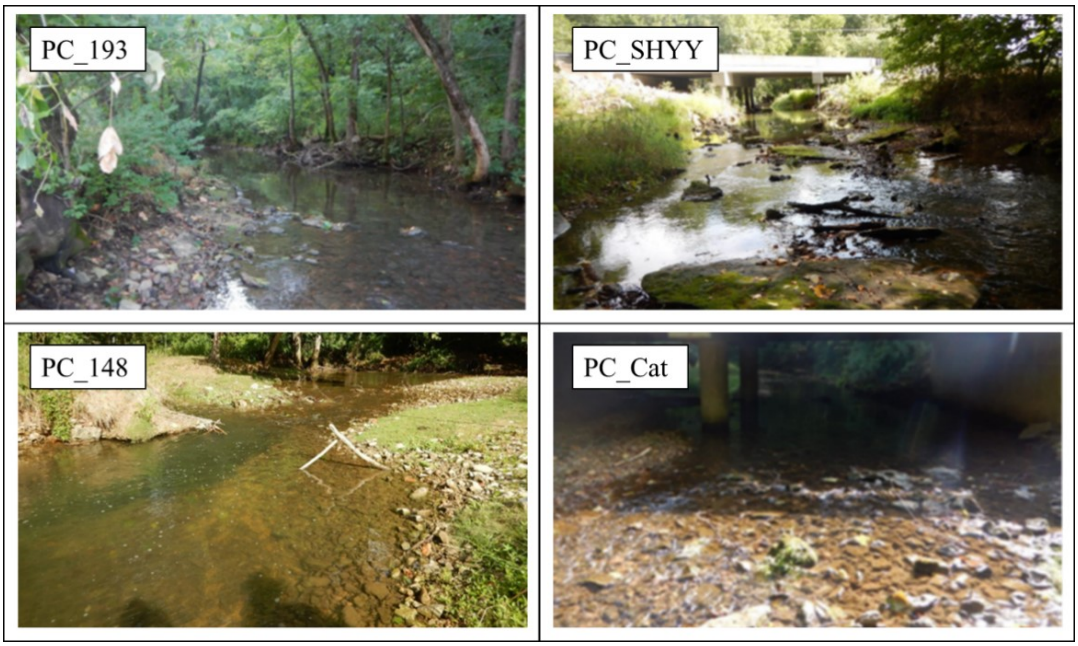


Figure 2: Sites of high bacterial contamination along the Pearson Creek watershed.

Figure 3: Sequiota Park's recharge spring.

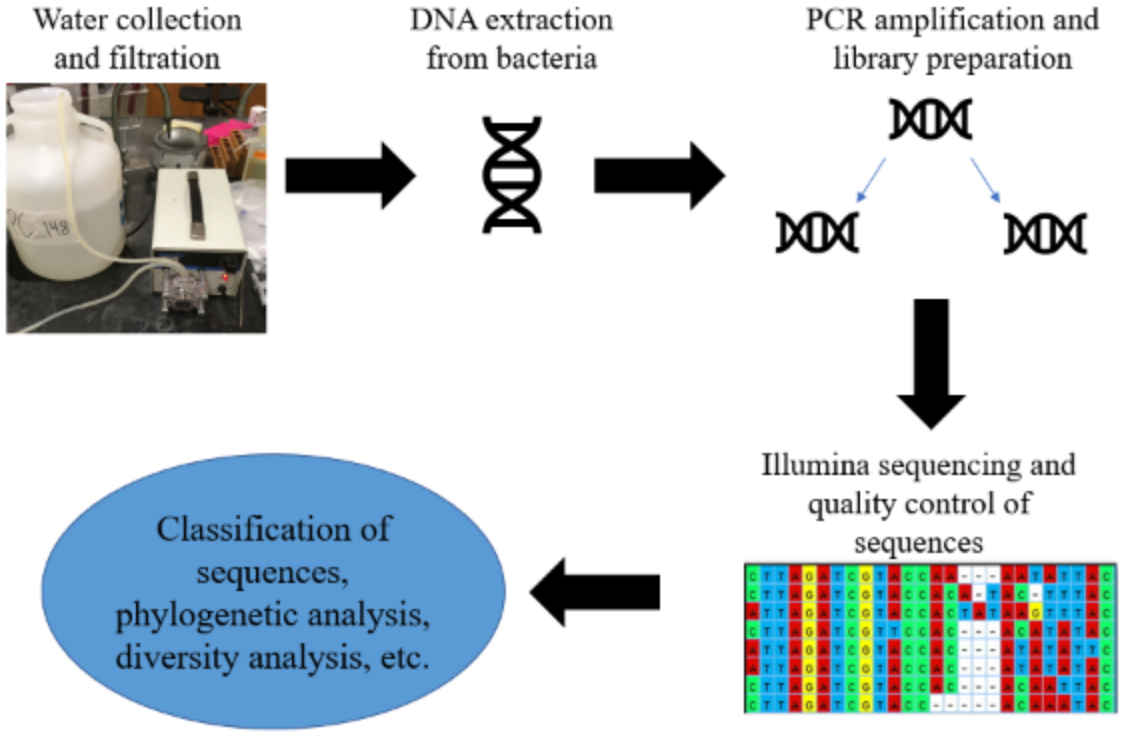


Figure 4: Workflow of how bacterial communities were assessed in this thesis.

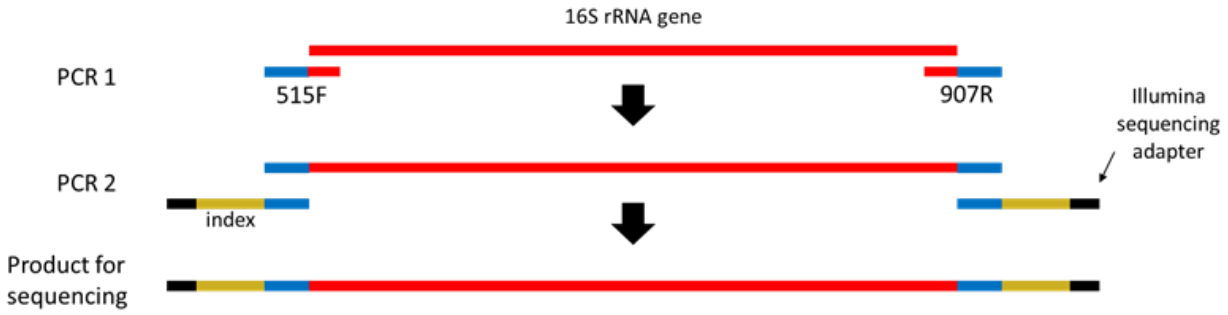


Figure 5: Library preparation of DNA for Illumina sequencing using universal 16S rRNA primers 515F and 907R.

PATHOGEN DETECTION WITHIN BACTERIALLY IMPAIRED WATERSHEDS IN KARST REGIONS

Abstract

Bacterial impairment of water systems has been a common problem worldwide. However, studies on the relative distribution of bacterial pathogens within impaired water systems have been limited. Mostly, the impaired aquatic systems have been classified based on the detection of indicator microbes or characterized through microbial source tracking. In the current study, I assessed general bacterial diversity, as well as presence of potential human pathogens within three bacterially impaired aquatic systems. I used Illumina paired-end DNA sequencing of bacterial 16S rRNA gene amplicons from the three water systems which consistently showed high abundance of *E. coli* and fecal coliforms for the past few years. I observed both spatial and temporal variations in the relative distribution of DNA sequences related to different potential pathogens such as *Legionella*, *Salmonella*, *Yersinia*, and *Plesiomonas* spp. across different water systems. I identified a few potential hotspots within the impaired water systems where the relative abundance of different pathogen-related sequences was high. This study demonstrates the effectiveness of high-throughput DNA sequencing for an initial screening of potential pathogens, which can be followed by a targeted approach for monitoring the specific pathogens in the impaired water environments. This method of testing could allow for a more cost-effective way of monitoring the true health risk our waters pose to those who use them.

Keywords: bacterial impairment; bacterial pathogens; high-throughput sequencing

Introduction

Bacterial contamination of drinking and recreational water is a major environmental health concern (Pandey *et al*., 2014). According to the World Health Organization (WHO 2019), every year waterborne diseases cause more than 1.5 million human deaths worldwide. Although most of these deaths are reported in developing countries, there are still many incidences of waterborne diseases coming from developed countries (CDC, 2016). For example, within the United States, every year 75,000 deaths and hospitalizations have been attributed to waterborne bacterial pathogens (Adam *et al*., 2017). Bacterial pathogens such as *Salmonella*, *Shigella* and pathogenic *E. coli* from human or animal fecal material can enter the drinking and recreational water resources (Byappanahilli *et al*., 2012). The fecal material from infected individuals can carry high loading of bacterial pathogens, such as up to 10^{10} *Salmonella* cells and 10^{9} *Shigella* cells per gram of human feces (Pepper *et al*., 2014). In addition to fecal pathogens, many pathogens can persist in water and cause disease under optimum growth conditions. This highlights the need for continuous monitoring of the water systems for potential pathogens.

Frequently, the health risks associated with water contamination are tested through the presence of fecal indicator bacteria (FIB) using quick and cost-effective methods, such as IDEXX Colilert and membrane filtration-based testing (Kinzelman *et al*., 2005; Ramirez-Castillo *et al*., 2015). The presence of FIB (like *E. coli,* enterococci and fecal coliforms) suggest fecal contamination of water systems. Detection of FIB has been used as the gold standard for determining bacterial impairment of watersheds. However, FIB testing does not provide information on the source of fecal contamination (Unno *et al*., 2018; Jang *et al*., 2017). The microbial source tracking (MST) approach was developed to assess the specific source of fecal contamination, i.e., fecal material originating from human or other animal sources. The specific

bacterial markers unique to different animals can be quantified using host-specific quantitative PCR (qPCR) primers (Green *et al*., 2014; Layton *et al*., 2006). The FIB and MST approaches are commonly used to identify fecal contamination of water environments and design measures to reduce the extent of contamination (Dickerson *et al*., 2007).

Both FIB and MST approaches are used as exploratory tools for the detection and identification of fecal contamination sources in a watershed. However, they do not provide information on the relative distribution of different waterborne human pathogens. The information on the relative distribution of different bacterial pathogens in a water source is important for water resource managers and can be used to reduce the chances of disease outbreaks. Quantifying the abundance of a specific human pathogen can be done through qPCR using target-specific primers. However, testing hundreds of different pathogens such as *Salmonella*, *Campylobacter*, *Legionella*, pathogenic strains of *E. coli*, and *Vibrio vulnificus* in a water sample without any prior knowledge can be very costly and time-consuming. Next-Gen DNA sequencing can be used for the initial screening of the presence of bacterial pathogens in an impaired watershed. This can help to identify the potential targets for qPCR for the temporal monitoring of these pathogens on a consistent basis within the impaired watershed to determine the specific health risks associated with the potential use of a watershed.

In the current study, I assessed the diversity and relative distribution of potential bacterial pathogens in three bacterially impaired water systems. Two surface water systems, Little Sac watershed (LSW) and Pearson creek (PC), were sampled spatially to determine if there were differences at various sites along these impaired systems. The third water system was a recharge spring located at Sequiota Park's cave. This site was sampled temporally to determine if variations in pathogen diversity change during summer and winter season. The three water

systems chosen for this study previously showed a consistently high abundance of FIB determined using the IDEXX Colilert testing approach (Bullard *et al*., 2001; WWE, 2001; Baffaut, 2006; Owen and Pavlowsky, 2014; MDNR, 2020). So far, the testing of these watersheds has been mainly focused on FIB alone, which provides little information on presence of actual pathogens and the potential health risks that are associated with these bacterially impaired water systems. This study aims to investigate potential pathogens that may be associated with these three systems.

Methods

Water Collection and processing. Water samples were collected at five sites along the Little Sac Watershed (LSW) (Figure S1), four sites along Pearson Creek (PC) (Figure S2) and a site located at the Sequiota Park's cave recharge spring. The samples from Sequiota Park's cave spring were collected from a single location but multiple samples were collected overtime from this location. The detailed description of sampling locations, sampling dates and volume of water filtered are reported in supplemental table (Table S1). Each sample was collected in 5-liter sterilized polypropylene carboys and transported back to the lab on ice.

Isolation of bacterial cells and DNA extraction. Water samples were filtered through 0.22 μm Sterivex filters (Millipore Corporation, MA) using a peristaltic pump (Masterflex, Cole–Pamer Co, Vernon Hills, IL, USA). Approximately, 0.75-1.75 liters of water, depending on the concentration of suspended material, were filtered for each sample (Table S1). The filters were stored at -20^{o}C until further processing. Then filters were cut into small fragments using sterile scissors and placed into 50 mL tubes. Sterile water (25 mL) was added to the tubes containing fragments of filter and vortexed for five minutes to detach the bacterial cells from the

filter. Cells were harvested by centrifugation at 10,000 rpm for five minutes and used for the DNA extraction using Qiagen's DNeasy PowerLyzer PowerSoil kits (Mo Bio, Carlsbad, CA). DNA was eluted with 25 µL sterile water and stored at -20°C until further processing.

DNA sequencing. Bacterial communities from each sample were assessed using Illumina MiSeq paired-end DNA sequencing. A two-step PCR approach was used as described previously (Mayhood & Mirza, 2021). Briefly, the first PCR amplified the V3-V5 region of the bacterial 16S rRNA gene using primers F515 (5′-GTGCCAGCMGCCGCGG-3′) and R907 (5′-CCGTCAATTCMTTTRAGTTT-3′). These universal bacterial primers also contained Illumina sequencing primers that were used for the 2nd PCR amplification. Each 25 µL PCR reaction contained 1X buffer, 0.2 µM of each primer, 2.0 mM $MgSO_4$, 0.2 µM of each deoxynucleoside triphosphates (dNTPs), 1.0 µL of template DNA, and 5 units of High-Fidelity Platinum Taq polymerase (Invitrogen, USA). Conditions for the first PCR were: an initial denaturing step for 5 min at 95 °C, followed by 30 cycles of 95 °C for 45 sec, an annealing step at 56 °C for 45 sec, an extension at 72 °C for 45 sec, and a final extension at 72 °C for 7 minutes. As a standard PCR procedure, I ran positive and negative controls along with each set of PCR reactions. The successful amplification of PCR was evaluated by gel electrophoresis and staining with ethidium bromide. Amplified PCR products were cleaned using ExoSap-IT PCR Cleanup System (Invitrogen, USA) as per manufacture's protocol.

Cleaned PCR products of the first PCR reaction were used as the templates for the second PCR step. PCR reagents and their concentrations were the same as described above, with an exception of the PCR primers. The primers used in the second PCR contained the Illumina sequencing adapters A and B along with the standard unique multiplex identifier sequences. The PCR conditions for the second PCR were: initial denaturation for 3 minutes at 95°C, followed by

10 cycles of denaturation at 94°C for 30 s, annealing at 60°C for 30 s, and extension at 72°C for 30 s, with a final extension at 72°C for 7 minutes. Products from the 2nd PCR were quantified using Nanodrop 2000 spectrophotometer (ThermoFisher Scientific) and pooled together in equimolar concentrations. Pooled PCR amplicons from different samples were purified using the Agencourt AMPure beads (Beckman Coulter, Brea, CA). The purified PCR products were sequenced using Illumina MiSeq paired-end DNA sequencing.

Sequence processing and phylogenetic analysis. Sequences were processed for initial quality control parameters such as read length, ambiguous bases, removal of chimeric sequences, etc. using Ribosomal Database Project II (RDP) (Fish *et al*., 2013) (http://rdp.cme.msu.edu). High-quality filtered DNA sequences were classified using the RDP classifier. Bacterial sequences related to *Legionellaceae, Enterobacteriaceae* and *Bacteroidaceae* families were extracted using Mothur (Schloss *et al*., 2009). These sequences were further assessed for additional quality control parameters such as presence of ambiguous bases ('N') and trimmed to uniform sequence length using Sequencher 5.4.6 (Gene Codes Corp., Ann Arbor, MI, USA). Cleaned DNA sequences along with the reference sequences of these families from GenBank were aligned and clustered into operational taxonomic units (OTUs) using RDP. The Neighbor-joining based phylogenetic trees were constructed using MEGA version 10.1.8. (Kumar *et al*., 2018). Pie charts were created to visualize the relative distribution of sequences originating from each watershed. The size of the pie chart depicts the relative abundance of sequences within a specified cluster. The distribution of sequence abundances among different clusters can be directly compared within the same watershed, but not across different watersheds.

Statistical analysis. Overall, the significance of differences among bacterial communities across different sites was assessed by analysis of similarity (ANOSIM) and

visualized with nonmetric multidimensional scaling (NMDS) plots generated in R version 3.6.0 (R Core Team, 2019). The Bray-Curtis similarity data were square-root transformed prior to ANOSIM analysis.

Results and Discussion

Next-Gen DNA sequencing. Next-Gen DNA sequencing resulted in 2,815,170 good quality sequences across all sites. Table S1 breaks down the number of sequences per site from the three watersheds. Sequences classified as relating to three families of bacteria, *Legionellaceae, Enterobacteriaceae,* and *Bacteroidaceae,* were extracted for phylogenetic analysis to investigate the presence of known human pathogens.

Distribution of *Legionellaceae*-related sequences. I retrieved 9,849 *Legionellaceae* related sequences across all sampling locations. These sequences were affiliated with both pathogenic and non-pathogenic *Legionella* spp. from GenBank. Representative sequences from each OTUs containing 50 or more sequences were randomly selected to construct a phylogenetic tree. I used >50 sequences as a cutoff because an overwhelming number of OTUs contained >3 sequences. Overall, the tree was comprised of 4,339 sequences from OTUs containing >50 sequences along with 23 *Legionella* references from GenBank (Figure 1). Of these sequences, 97 sequences were originated from Little Sac sites, 2,629 from Pearson Creek and 1,613 from Sequiota Park cave. The phylogenetic tree suggests the presence of 9 distinct phylogenetic clusters (I to IX). The sequences within clusters I, IV, VII, and VIII were closely related pathogenic species of *Legionella* such as *L. dumoffi*, *L. anisa*, *L. pneumophila* and several others (Figure 1). The sequences in clusters I, IV, VII, VIII, and IX were related to non-pathogenic

species of *Legionella*. Whereas, clusters III, V, and VI are unique without any references. These sequences did not show close relatedness to the *Legionella* references used in this study.

Pie charts next to clusters are depicting the relative distribution of sequences from each site or sampling season that are reported for each cluster. The size of the pie chart is proportional to the number of sequences in that cluster. It is important to mention that the size of the pie chart can only be compared to other pie charts from the same watershed.

Within the Little Sac watershed, most of the sequences among different clusters belonged to site PR_102 (50 to 80%). Cluster I and VI contained the most sequences and were closely related to three pathogenic species (*L. dumoffii, L. anisa,* and *L. bozemanii*), and two non-pathogenic species (*L. drancourtii* and *L. worsleiensis*). Site PR_102 resides within the city of Springfield. The increase in *Legionella*-related sequences could be due to contamination coming from the old sanitary sewer systems present in the northern part of the city. *Legionella* spp. can survive or grow in corroded pipes by forming biofilms that sustain higher abundances of these bacteria (EPA, 2016). These pathogens have been known to make up a small percent of Legionnaire's disease that mainly occurs in immunocompromised patients (Muder & Yu, 2002).

Similarly, within the Pearson Creek watershed at site PC_Cat (Figure 1), I retrieved a high abundance of *Legionella*-related sequences (35 to 100%). The distribution of *Legionella*-related sequences at the two downstream sites (PC_193 and PC_148) was relatively low as compared to the PC_Cat site. I observed low water levels at the time of sampling at these sites which may have led to less downstream movement of *Legionella* spp. Clusters I and VIII have a higher abundance of sequences, with Cluster VIII containing more non-pneumophila pathogenic species like *L. maceachemii*, *L. israelensis*, *L. jordanis* and others (Figure 1). These pathogens mostly cause pneumonia, but also have been reported to cause Pontiac fever (Muder & Yu,

2002). Plumbing systems can harbor these bacteria in biofilms and allow them to grow in number (EPA, 2016). Since this stream runs through residential areas, it is possible that leaking plumbing systems or leaky domestic sewer systems may be contributing to this influx.

In the Sequiota Spring's samples, the clusters I, III, and IX made up the majority of sequences, with cluster IX being the major cluster with *L. geestiana* belonging to this cluster. *L. geestiana* is not known to cause disease in humans. Of these three major clusters, the abundance of *Legionella*-related sequences was increased in the winter of 2019 (26 to 92%) compared to the summer sampling events. Normally, incidences of Legionnaire's disease/Legionellosis are increased during the summer and fall seasons due to the preferred warmer temperatures (CDC, 2018b). Although disease cases increase more in the summer/fall, *Legionella* species are able to persist in the environment in relatively the same abundance year-round (Parthuisot *et al*., 2010). Since temperature does not vary much in groundwater systems, it is possible that increases in humidity and availability of protozoan hosts may be influencing the increase in *Legionella*-related sequences in the winter at this recharge spring (Newton *et al*., 2010; Simmering *et al*., 2017). Changes in flow rate may also influence the detection of *Legionella*, given that faster water flow can disrupt biofilms that they are contained in (Cassell *et al*.,

In general, *Legionella* spp. are respiratory pathogens that can infect alveolar macrophages. The most well know of these pathogenic species is *L. pneumophilia*, which is the main causative agent of Legionnaire's disease. *Legionella* spp. are frequently found in natural environments such as lakes and streams, but their numbers are relatively low due to improper growth conditions like low numbers of their protozoan hosts (Atlas, 1999). Factors such as the presence of biofilms and protozoa they can parasitize help them replicate and survive nutrient-poor environments (Newton *et al*., 2010). The frequency of *Legionella*-associated diseases can

increase due to high concentrations of suspended or aerosol *Legionella* cells in an environment (Fields, Benson, and Besser, 2002). A major concern for this is if water containing *Legionella* is used for showering or other household purposes, in which the bacteria can easily become aerosolized. In the United States, around 10,000 cases of Legionnaires disease were reported in 2018 with a fatality rate of 10% (CDC, 2018a). Considering the health risks associated with *Legionella*, it is important to monitor this potential pathogen to reduce the disease incidence.

It is important to mention that partial 16S rRNA sequencing may not be able to accurately assign the phylogeny at the species and strain levels. Further testing of specific pathogenic species through qPCR using species-specific marker genes could help to quantify the specific species in water environments. Overall, I observed the assignment of partial reference sequence corroborated well with the phylogenetic placement of *Legionella* species from other studies based on the complete 16S rRNA gene sequences (Burstein *et al*., 2016).

Distribution of *Enterobacteriaceae*-related sequences. Phylogenetic analyses of 6,760 *Enterobacteriaceae* related sequences retrieved from the three different water systems could be divided into three distinct clusters (I to III). OTUs containing >50 sequences (3,213 sequences) along with the 11 references sequences of known human pathogens from GenBank were used for the construction of a phylogenetic tree (Figure 2). The distribution of these sequences within different water systems were 204 from Little Sac, 2,466 from Pearson Creek, and 543 sequences from Sequiota park's cave. Most of the sequences were classified into two clusters (I&III) (Figure 2). At the site PR_102 and PC_Cat, cluster I related sequences ranged 79% to 87% of the *Enterobacteriaceae* sequences. Similarly, these two sites made up 59% and 85% of sequences in cluster III (Figure 2). Like *Legionella* sequences, I detected a similar trend in the distribution of *Enterobacteriaceae*-related sequences at the PR_102 and PC Cat sites.

Cluster I contained references from pathogenic *Salmonella* spp., *Shigella* spp. and *E. coli*. *Salmonella* species are intracellular parasites that usually invade the gastrointestinal tract and cause salmonellosis. Some species of *Salmonella* such as *S. enterica* cause typhoid fever. Mostly, typhoid fever is associated with drinking water contaminated with fecal material. Each year approximately 200 million to 1.3 billion infections and about 3 million deaths world-wide each year are caused by non-typhoid species of Salmonella (Coburn *et al*., 2007). The two serovars *S. typhi* and *S. paratyphi* cause 21.7 million infections and 217,000 deaths each year (Ashurt *et al*., 2018), mostly in developing counties (Ao *et al*., 2015). Similarly, *Shigella* spp. and *E. coli* cause gastrointestinal diseases like diarrhea and dysentery. Both organisms are reported to annually cause 491,828,692 disease incidences worldwide (Khalil *et al*., 2018). It is important to mention that the 16S rRNA gene sequences from both genera (*Escherichia* and *Shigella*) are highly similar and cannot be reliably distinguish. (Ragupathi *et al*., 2017).

Sequences in clusters II and III were related to references of pathogenic *Yersinia* spp. and *Plesiomonas shigelloides*, respectively. *P. shigelloides* causes gastroenteritis infections in human and is frequently reported to originate from freshwater environments and through the consumption of aquatic foods like shellfish and catfish (Janda *et al*., 2016). Likewise, *Yersinia* spp. (*Y. enterocolitica* and *Y. pseudotuberculosis*) also cause gastrointestinal infections and are reported to originate from the untreated water that is contaminated with fecal material (Tacket *et al*., 1985; Thompson & Gravel, 1986). *Yersinia* spp. related infections can also be caused by the consumption of contaminated meat. Previously, Galindo *et al*. (2011) reported the spreading of *Y. enterocolitica* through the fecal-oral route that can significantly increase human-human transmission of this pathogen.

Previously, MST-based study of this site suggested the presence of human fecal contamination (Mirza *et al*., 2018). It is possible that the influx of these *Enterobacteriaceae*-related sequences came from human fecal contamination due to the old sanitary sewer systems in the area. Although, previous studies (Renter *et al*., 2006; Rouffaer *et al*., 2017; Skov *et al*., 2008) have also suggested the presence of *Enterobacteriaceae*-related bacteria such as *Salmonella* and *Yersinia* spp. within fecal material of deer, birds, rodents, etc. These other sources could contribute to the extent of contamination occurring at these sites.

Most of the *Enterobacteriaceae*-related sequences from Sequiota Park's cave samples were related to the phylogenetic cluster I and a few sequences were related to cluster III. During summer 2020, I observed an increase in the number of *Enterobacteriaceae* sequences (56%) as compared to the summer of 2019 (21%) and the winter of 2019 (23%) within cluster I (Figure 2). Sequiota Spring is located in southeast Springfield and this watershed is part of a karst landscape that has several sinkholes, losing streams, and springs in the area (Thomson, 1986). The Sequiota Spring recharge area is approximately 4.8 mi^2 and drains significant portions of urbanized southeast Springfield. Previous studies have suggested the presence of human fecal contamination of this cave, with an increase in contamination in the summer of 2020 (Owen *et al*., 2021). The City of Springfield has been working to remediate aging sanitary sewer infrastructure in the recharge area to help improve water quality and reduce human health risk in local streams. The observed seasonal variation in the *Enterobacteriaceae* sequences could be due to repair work on the sanitary sewer lines. Since restoration work on these sewer systems mainly started in the summer of 2020, it could be possible that the increase seen in this study is influenced by accidental contamination from these repairs. Previous studies reported increases in

Enterobacteriaceae abundance during warmer months of the year, as well as increases due to sewage contamination (Arnade, 1999; Gallart *et al*., 2005; Buckalew *et al*., 2006).

Human fecal marker detected among *Bacteroidaceae*-related sequences. I also evaluated the variation in the relative abundance of *Bacteroidaceae*-related sequences because species with this family are commonly used as a marker for the identification of sources of fecal contamination (Field & Samadpour, 2007). Previous studies (Mirza *et al*., 2018; Owen *et al*., 2019; and Owen *et al*., 2021), evaluated these sites for various sources of fecal contamination. The PR_102 site with LSW, the PC_SHYY site with Pearson Creek, and Sequiota cave have shown the presence of human fecal contamination. I retrieved 4,351 *Bacteroides* classified sequences from the different water samples in this study (Figure 3). The phylogenetic analysis suggested the presence of two major phylogenetic clusters within *Bacteroidaceae* sequences. Cluster I contained the reference sequence *Bacteroides dorei* which is commonly used as a marker gene for human fecal contamination. The sequences within OTU1 in cluster I were highly similar to *B*. *dorei*. Cluster I made up most of the sequences that came from Little Sac and Sequiota Park. In contrast, cluster II contained twice as many sequences as cluster I of the Pearson Creek samples.

Overall, I observed a similar trend in the distribution of *Bacteroidaceae* sequences within cluster I as was seen for *Enterobacteriaceae*-related sequences. Site PR_102 within LSW, PC_Cat within Pearson creek and summer 2020 samples from Sequiota cave contained the most sequences in this cluster. These results corroborated well with the previous MST report-based studies on these sites (Mirza *et al*., 2018). Conversely, PC_Cat site within Pearson Creek did not previously display human fecal contamination (Owens *et al*., 2019). For Sequiota Spring, summer 2020 made up 84% of sequences in cluster I. This correlates well with the

Enterobacteriaceae tree. Similarly, to the prior two watersheds, Sequiota spring was studied using MST methods (Owen *et al*., 2021). That study showed that the relative abundance of the human fecal marker was increased in the summer of 2020 as compared to the other two sampling periods. This corresponds well with my results, suggesting that the pathogens possibly present in this system could be coming from the accidental leakage from the renovation of the sewer lines and human fecal contamination.

Distribution of other genera containing pathogenic species. In addition to the above three bacterial families, I also evaluated the distribution of other potential pathogenic bacterial genera (Table 1). I observed the presence of *Pseudomonas*, *Acinetobacter*, *Clostridium* and *Aeromonas* related sequences. Although some species within these genera are pathogenic, most of the species that can persist in aquatic environments are non-pathogenic. Further testing would be needed to determine if pathogenic species within these genera are present at these sampling sites. DNA sequences related to other bacterial genera that contain known pathogens such as *Campylobacter*, *Peptoclostridium* and *Helicobacter* were not detected at these sites. Using this method of sequencing and screening for abundant pathogen-containing genera could help quickly identify bacteria of concern that could then be tested and monitored using quantitative methods like real-time PCR.

Phylum level distribution of sequences from surface and recharging streams. To determine if there were differences in community structure, the abundance of classified phyla was used. Overall, 2,035,589 good quality sequences from LSW and PC sites were classified at the phylum level (Figure S3). This accounts for sequences that were unclassified at the family/genus level. Most of the sequences (47 to 85% of sequences per site) were related to phylum *Proteobacteria*. *Bacteroidetes* was the second most abundant phylum (5 to 40%

sequences per site), followed by *Actinobacteria* (2 to 14%) and *Cyanobacteria* (1 to 4%) (Figure S3). This trend is similar to previous findings in bacterial stream diversity (Zeglin, 2015). I observed variation in the distribution of phylum related sequences at different sites. For example, site PR_102 from LSW displayed an elevated abundance of *Proteobacteria* as compared to other locations (25% increase).

For Sequiota Spring's samples, a similar trend was observed. Overall, 779,581 sequences were classified at the phylum level (Figure S4). *Proteobacteria* was the most abundant (38-67%), followed by *Actinobacteria*, *Cyanobacteria*, and *Bacteroidetes* at varying percentages. There were some variations depending on the time of year. In the winter of 2019, *Cyanobacteria/Chloroplast* sequences increased by 20%, and *Actinobacteria* decreased by 14-30% compared to the summer sampling events. During the summer of 2019, *Proteobacteria* decreased in abundance by 25%. These changes in phyla distributions suggest a variation in the bacterial community structure over the course of the sampling year.

General bacterial community structure. Differences at the genera level were investigated to determine if the bacterial communities at each site were unique. The differences in bacterial community structure of the surface streams Little Sac and Pearson Creek were compared using multivariate analysis. Overall, 771,696 good quality sequences classified into 596 genera were used for non-metric multidimensional scaling analyses using the Bray-Curtis similarity matrix (Figure 4). Only genera containing 5 or more sequences were used. Overall difference in bacterial community structure at two water systems was non- significant ($p=0.178$). I also assessed the variation within each watershed. For Pearson Creek sites only, there was a significant difference in community composition ($p<0.01$) and site PC_Cat was grouped separately (Figure 4). These differences in overall bacterial community could be due to

differences in water characteristics or due to differences in the sources of bacterial contamination. Likewise, within LSW there was a significant difference in bacterial community structure ($p<0.05$). Site PR_102 was separately clustered as compared to other sites in the LSW, but rather was grouped with PC sites. Since PR_102 and PC sites are located in urbanized areas, it is possible that this is playing a role in their bacterial communities.

Similarly, non-metric multidimensional scaling analyses was done on Sequiota Park's recharge spring with genera containing 5 or more sequences. A total of 368,033 good quality sequences were used, which were classified into 656 genera. A significant difference was observed between these three sampling dates (p-value < 0.001). Each sampling event was grouped separately from one another, indicating that the bacterial composition varied with each season (Figure 5). It could be possible that the renovating of the sewage systems could be impacting these differences. Other factors such as flow rate, inhabitance of animals (e.g. bats), and temperature may be influencing these difference that are seen.

References

Adam, E. A., Collier, S. A., Fullerton, K. E., Gargano J. W. & Beach, M. J. 2017 Prevalence and direct costs of emergency department visits and hospitalizations for selected diseases that can be transmitted by water, United States. *Journal of Water and Health* **15**(5), 673-683.

Ao, T. T., Feasey, N. A., Gordon, Keddy, K. H., Angulo, F. J. & Crump, J. A. 2015 Global Burden of invasive Nontyphoidal *Salmonella* Disease, 2010. *Emerging Infectious Diseases* **21**(6), 941-949.

Arnade, L. J. 1999 Seasonal correlation of well contamination and septic tank distance. *Groundwater* **37**(6), 920-923.

Ashurst, J. V., Truong, J. & Woodbury, B. 2018 Salmonella typhi.

Atlas, R. M. 1999 Legionella: from environmental habitats to disease pathology, detection and control. *Environmental Microbiology* **1**(4), 283-293.

Baffaut, C. 2006 Total Maximum Daily Load for Little Sac River Watershed. FAPRI-UMC Report #07-05. Food and Agricultural Policy Research Institute (FAPRI), University of Missouri. Missouri Department of Natural Resources Water Protection Program.

Buckalew, D. W., Hartman, L. J., Grimsley, G. A., Martin, A. E. & Register, K. M. 2006 A long-term study comparing membrane filtration with Colilert® defined substrates in detecting fecal coliforms and Escherichia coli in natural waters. *Journal of Environmental Management* **80**(3), 191-197.

Bullard, L., Thomson K.C., and Vandike J.E. 2001 The Springs of Greene County Missouri. Missouri Department of Natural Resources Geological Survey and Resource Assessment Division. Water Resources Report No. 68.

Burstein, D., Amaro, T. Z., Lifshitz, Z., Cohen, O., Gilbert, J. A., Pupko, T., Shuman, H. A. & Segal, G. 2016 Genomic analusis of 38 Legionella species identifies large and diverse effector repertoires. *Nature Genetics* **48**, 167-175. https://doi.org/10.1038/ng.3481

Byappanahilli, M. N., Nevers, M. B., Korajkic, A., Staley Z. R. & Harwood, V. J. 2012 Enterococci in the environment. *Microbiology and Molecular Biology Reviews* **76**(4), 685-706.

Cassell, K., Gacek, P., Warren, J. L., Raymond, P. A., Cartter, M. & Weinberger, D. M. 2018 Association between sporadic legionellosis and river systems in Connecticut. *The Journal of Infectious Diseases* **217**(2), 179-187.

Centers for Disease Control and Prevention (CDC). 2016 Magnitude & Burden of Waterborne Disease in the U.S. Retrieved from https://www.cdc.gov/healthywater/burden/index.html

CDC. 2018a *Legionella* (Legionnaires' disease and Pontiac fever). https://www.cdc.gov/legionella/fastfacts.html

CDC. 2018b National notifiable infectious diseases and conditions: United States. https://wonder.cdc.gov/nndss/static/2016/annual/2016-table2h.html

Coburn, B., Grassl, G. A. & Finlay, B. B. 2007 Salmonella, the host and disease: a brief review. *Immunology and Cell Biology* **85**(2), 112-118.

Ragupathi, N. D., Sethuvel, D. M., Inbanathan, F. Y. & Veeraraghavan, B. 2018 Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. *New Microbes and New Infections* **21**, 58-62.

Dickerson, J. W., Hagedorn, C. & Hassall, A. 2007 Detection and remediation of human-origin pollution at two public beaches in Virginia using multiple source tracking methods. *Water Research* **41**(16), 3758-3770. https://doi.org/10.1016/j.watres.2007.02.055

EPA. 2016 Technologies for *Legionella* Control in Premise Plumbing Systems: Scientific Literature Review. Retrieved from: https://www.epa.gov/sites/production/files/2016-09/documents/legionella_document_master_september_2016_final.pdf

Field, K. G. & Samadpour, M. 2007 Fecal source tracking, the indicator paradigm, and managing water quality. *Water Research* **41**, 3517–3538. https://doi.org/10.1016/j.watres.2007.06.056

Fields, B. S., Benson, R. S. & Besser, R. E. 2002 *Legionella* and Legionnaires's disease: 25 years of investigation. *Clinical Microbiology Reviews* **15**, 506-526

Fish J.A., Chai B., Wang Q., Sun Y., Brown C.T., Tiedje J.M. & Cole J.R. 2013 FunGene: the functional gene pipeline and repository. *Frontiers in Microbiology* **4**, 291. https://doi.org/10.3389/fmicb.2013.00291.

Galindo, C. L., Rosenweig, J. A., Kirtley, M. L. & Chopra A. K. 2011 Pathogenesis of *Y. enterocolitica* and *Y. pseudotuberculosis* in Human Yersiniosis. *Journal of Pathogens* 2011 doi:10.4061/2011/182051

Gallert, C., Fund, K. & Winter, J. (2005). Antibiotic resistance of bacteria in raw and biologically treated sewage and in groundwater below leaking sewers. *Applied Microbiology and Biotechnology* **69**(1), 106-112.

Green, H. C., Haugland, R. A., Varma, M., Millen, H. T., Borchardt, M. A., Field, K. G., Walters, W. A., Knight, R., Kelty, C. A. & Shanks, O. C. 2014 Improved HF183 quantitative real-time PCR assay for characterization of human fecal pollution in ambient surface waters samples. *Applied and Environmental Microbiology* **80**, 3086 –3094.

Janda, J. M., Abbott, S. L. & Mclver, C. J. 2016 *Plesiomonas shigelloides* Revisited. *Clinical Microbiology Reviews* **29**(2), 349–374.

Jang, J., Hur, H. G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T. & Ishii, S. 2017 Environmental Escherichia coli: ecology and public health implications-a review. *Journal of Applied Microbiology* **123**(3), 570–581. https://doi.org/10.1111/jam.13468

Khalil, I. A., Troeger, C., Blacker, B. F., Rao, P. C., Brown, A., Atherly, D. E., Brewer, T. G., Engmann, C. M., Houpt, E. R., Kang, G., Kotloff, K. L., Levine, M .M., Luby, S. P., MacLennan, C. A., Pan, W. K., Pavlinac, P. B., Platts-Mills, J. A., Qadri, F., Riddle, M. S., Ryan, E. T., Shoultz, D. A., Steele, A. D., Walson, J. L. Sanders, J. W., Mokdad, A. H., Murray, C. J. L., Hay, S. I. & Reiner Jr, R. C. 2018 Morbidity and mortality due to shigellaand enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990–2016. *The Lancet Infectious Diseases* **18**(11), 1229-1240. https://doi.org/10.1016/S1473-3099(18)30475-4

Kinzelman J. L., Singh A., Clem N. G., Pond K. R., Bagley R. C. & Gradus S. 2005 Use of IDEXX Colilert-18® and Quanti-Tray/2000 as a rapid and simple enumeration method

for the implementation of recreational water monitoring and notification programs, *Lake and Reservoir Management* **21**, 73-77. DOI: 10.1080/07438140509354414

Kumar S., Stecher G., Li M., Knyaz C. & Tamura K. 2018 MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**(6), 1547-1549.

Layton, A., Mckay, L., Williams, D., Garrett, V., Gentry, R., Sayler, G. 2006 Development of Bacteroides 16S rRNA geneTaqMan-based real-time PCR assays for estimation of total, human, and bovine fecal pollution in water. *Applied and Environmental Microbiology* **72**, 4214–4224.

Mayhood, P. & Mirza, B. S. 2021 Soybean Root Nodule and Rhizosphere Microbiome: Distribution of Rhizobial and Nonrhizobial Endophytes. *Applied and Environmental Microbiology* **87**(10).

Missouri Department of Natural Resources (MDNR) 2020 Approved Section 303(d) Listed Waters. https://dnr.mo.gov/env/wpp/waterquality/303d/docs/2020-303d-list-cwc-approved-2020-04-02.pdf

Muder, R. R. & Yu, V. L. 2002 Infection due to *Legionella* species other that *L. pneumophila*. *Clinical Infectious Diseases* **35**(8), 990-998. https://doi.org/10.1086/342884

Mirza, B. S., Owen, M. R., Kincaid, J. C. & Pavlowsky, R. T. 2018 Bacteria source tracking to support watershed planning, Little Sac River, southwest Missouri. Missouri State University. Retrieved from: https://oewri.missouristate.edu/assets/OEWRI/LSW_SourceTracking_6-7-18.pdf

Newton, H. J., Ang, D. K. Y., Driel, I. R. & Hartland, E. L. 2010 Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clinical Microbiology Reviews* **23**(2), 274-298.

Owen, M.R. and R.T. Pavlowsky (2014). Water Quality Assessment and Load Reductions for Pearson Creek, Springfield, Missouri. Final Report to the James River Basin Partnership. Ozarks Environmental and Water Resources Institute EDR-14-001.

Owen, M. R., Mirza, B. S., Kincaid, J. C., Roman, G. F. & Pavlowsky, R. T. 2019 Bacteria source tracking to support watershed planning, Pearson Creek, Greene county, Missouri. Missouri State University. Retrieved from: https://oewri.missouristate.edu/assets/OEWRI/PCBacteriaReportFINAL02082019.pdf

Owen, M. R., Mirza, B. S., Pursley, T. J. & Pavlowsky, R. T. 2021 Bacteria source tracking assessment of Sequiota Spring (June 2019-July 2020). Springfield Missouri. Missouri State University.

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S. & Singh, V. P. 2014 Contamination of water resources by pathogenic bacteria. *AMB Express* **4**, 51. https://doi.org/10.1186/s13568-014-0051-x

Pepper, I. L., Gerba, C. P. & Gentry, T. J. 2014 Environmentally Transmitted Pathogens. In Gentry, T. J. (Eds.), *Environmental Microbiology (*3rd ed., pp. 509-550) Elsevier.

Parthuisot, N., West, N.J., Lebaron, P. & Baudart, J. 2010 High divesity and abundance of *Legionella* spp. in a pristine river and impact of seasonal and anthropogenic effects. *Applied and Environmental Microbiology* **76**(24), 8201-8210. doi: 10.1128/AEM.00188-10

R Core Team 2019 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

Ramirez-Castillo, F. Y., Loera-Muro, A., Jacques, M., Garneau, P., Avelar-gonzález, F. J., Harel, J., Guerrero-barrera, A. L. 2015 Waterborne Pathogens: Detection Methods and Challenges. *Pathogens* **4**, 307–334. https://doi.org/10.3390/pathogens4020307

Renter, D. G., Gnad, D. P., Sargeant, J. M. & Hygnstrom, S. E. 2006 Prevalence and Serovars of *Salmonella* in the feces of free-ranging white-tailed deer (*Odocoileus virginianus*) in Nebraska. *Journal of Wildlife Diseases* **42**(3), 699-703. https://doi.org/10.7589/0090-3558-42.3.699

Rouffaer, L. O., Baert, K., Van den Abeele, A. M., Cox, I., Vanantwerpen, G., De Zutter, L., Strubbe, D., Vranckx, K., Lens, L., Haesebrouck, F., Delmée, M., Pasmans, F. & Martel, A. 2017 Low prevalence of human enteropathogenic Yersinia spp. in brown rats (Rattus norvegicus) in Flanders. *PloS one* **12**(4), e0175648. https://doi.org/10.1371/journal.pone.0175648

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J. & Weber C.F. 2009 Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537–7541. http://dx.doi.org/10.1128/AEM.01541-09

Simmering, J. E., Polgreen, L. A., Hornick, D. B., Sewell, D. K. & Polgreen, P. M. 2017 Weather-dependent risk for Legionnaires' disease, United States. *Emerging Infectious Diseases* **23**(11), 1843.

Skov, M. N., Madsen, J. J., Rahbek, C., Lodak, J., Jespersen, J. B., Jørgensen, J. C., Dietz, H. H., Chriél, M. & Baggesen, D. L. 2008 Transmission of *Salmonella* between wildlife and meat-production animals in Denmark. *Journal of Applied Microbiology* **105**(5), 1558-1568. doi: 10.1111/j.1365-2672.2008.03914.x.

Tacket, C. O., Harris, N., Allard, J., Nolan, C., Nissinen, A., Quan, T. & Cohen, M. L. 1985 An outbreak of Yersinia enterocolitica infections caused by contaminated tofu (soybean curd). *American Journal of Epidemiology* **121**(5), 705-711.

Thomson, K.C. 1986 Geology of Greene County, Missouri. Watershed Management Coordinating Committee.

Thompson, J. S. & Gravel, M. J. 1986 Family outbreak of gastroenteritis due to Yersinia enterocolitica serotype 0: 3 from well water. *Canadian Journal of Microbiology* **32**(8), 700-701.

Wright Water Engineers (WWE). 2001 Southern Hills Lakes Preliminary Evaluation and Management Plan: Summary Report. Prepared for the City of Springfield, April 2001.

Unno, T., Staley, C., Brown, C. M., Han. D., Sadowsky, M. J, & Hur, H. G. 2018 Fecal pollution: new trend and challenges in microbial source tracking using next-generation sequencing. *Environmental Microbiology* **20**(9), 3132-3140.

World Health Organization (WHO). 2019 Water sanitation hygiene. Retrieved from: https://www.who.int/water_sanitation_health/diseases-risks/en/

Wright Water Engineers (WWE) 2001 Southern Hills Lakes Preliminary Evaluation and Management Plan: Summary Report. Prepared for the City of Springfield, April 2001.

Zeglin, L. H. 2015 Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Frontiers in Microbiology* **6**, 454. doi:10.3389/fmicb.2015.00454

Table 1. Distribution of sequences related to pathogen-containing genera across the three watersheds.

Genus	Little Sac	Pearson Creek	Sequiota Park
Pseudomonas	403	65094	1506
Acinetobacter	1999	12992	939
Clostridium	228	989	1433
Aeromonas	366	1443	656
Staphylococcus	12	1470	126
Streptococcus	6	839	631
Bacillus	99	645	449
Alistipes	5	251	281
Rickettsia	39	346	39
Microcystis/Synechocystis	2	317	77
Enterococcus	14	246	43
Vibrio	3	244	1
Mycobacterium	5	120	76
Treponema	4	65	48
Klebsiella	0	3	48
Leptospira	0	32	12
Neisseria	0	32	0
Haemophilus	0	24	1
Proteus	0	21	3
Campylobacter	0	1	0
Peptoclostridium	0	0	0
Helicobacter	0	0	0

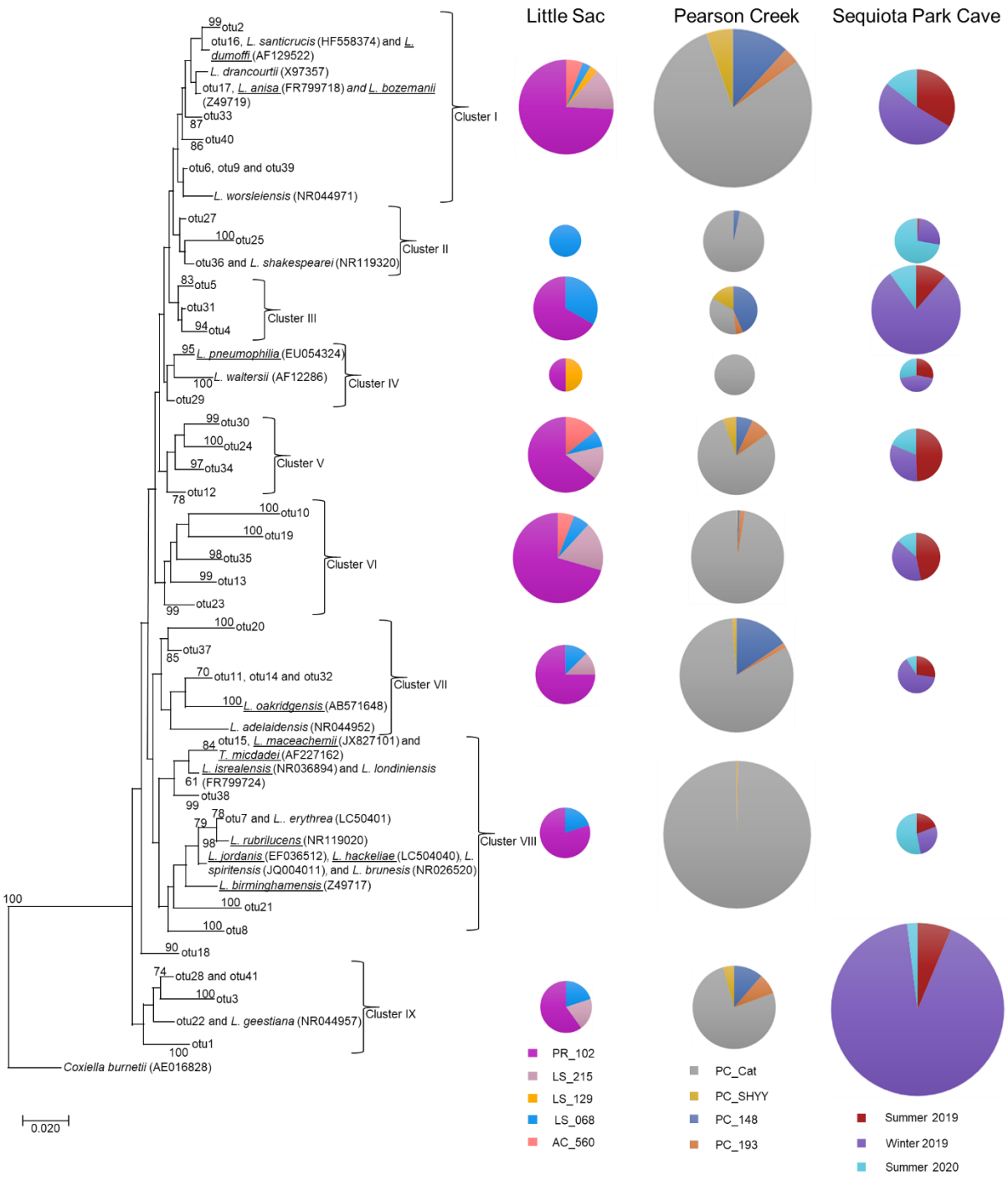


Figure 1. Neighbor-joining tree of *Legionellaceae* OTUs (95% similarity), along with pathogenic and non-pathogenic reference species from GenBank (accession numbers following reference's name). A total of 4,339 sequence are represented from all watersheds. The resulting tree suggests 9 distinct clusters. Distribution of sequences from each site in each watershed are depicted by pie-charts relating to what cluster they are in. The size of the pie-chart is relating to the amount of sequences are contained in that cluster, but size of pie charts can only be compared within the same watershed.

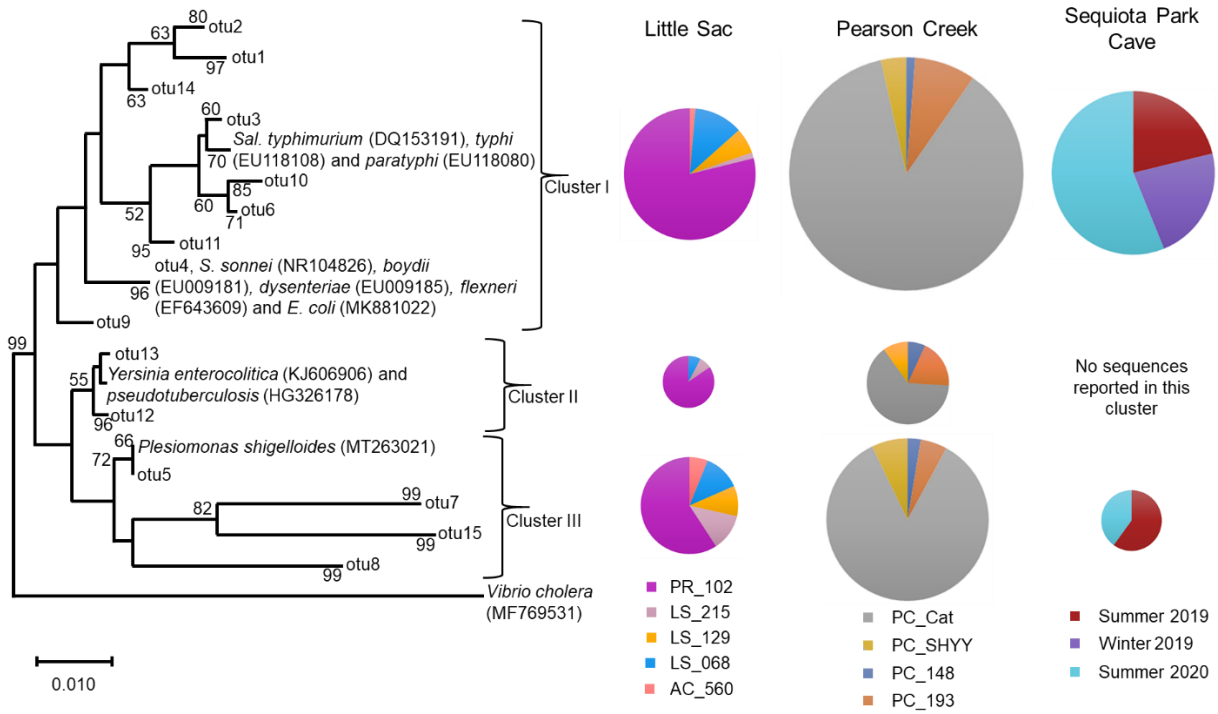


Figure 2. Neighbor-joining tree of *Enterobacteriaceae* OTUs (95% similarity) with 50 or more sequences in them (3,213 sequences represented), and known human pathogens obtained from GenBank (accession numbers following reference's name). Resulting tree suggests 3 clusters. Distribution of sequences from each site in the three watersheds are depicted by pie-charts relating to what cluster they are in. The size of the pie-chart is relating to the amount of sequences are contained in that cluster, but size of pie charts can only be compared within the same watershed.

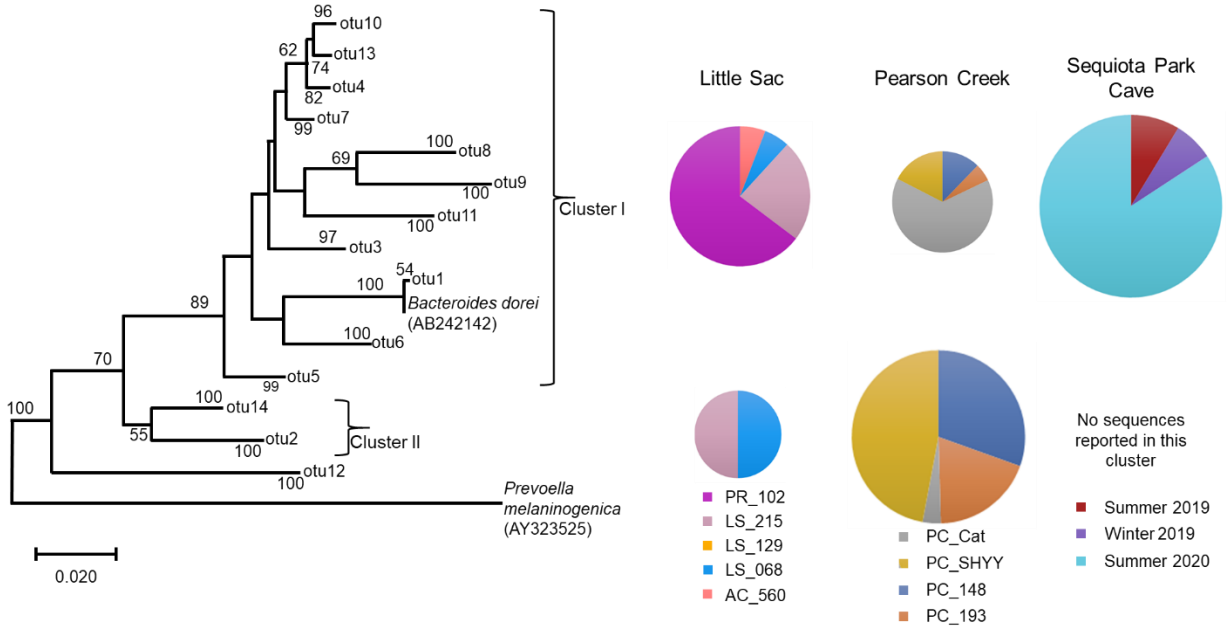


Figure 3. Neighbor-joining tree of *Bacteroidaceae*-related sequences along with the human fecal marker, *Bacteroides dorei*. OTUs (95% similarity) with 50 or more sequences (2,999 sequences represented) were used for construction of tree. Two distinct clusters were produced from the tree. Distribution of sequences from each site in each watershed are depicted by pie-charts relating to what cluster they are in. The size of the pie-chart is relating to the amount of sequences are contained in that cluster, but size of pie charts can only be compared within the same watershed.

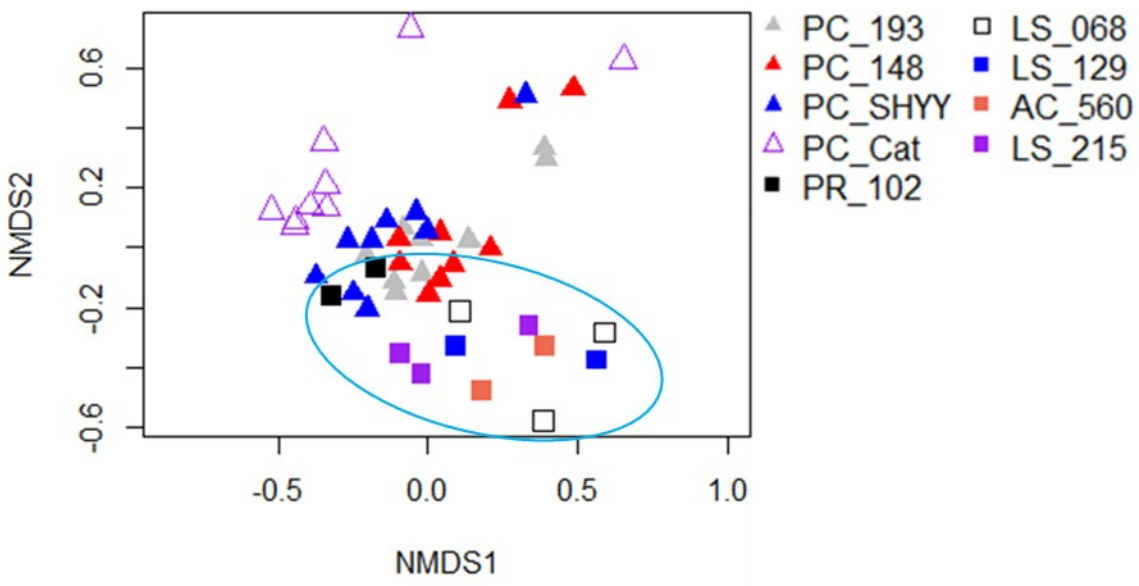


Figure 4. Non-metric multidimensional scaling plot of Pearson creek and Little Sac sites based on Bray-Curtis similarity index. Genera with 5 or more sequences across all sites were used for analysis, totaling 771,696 sequences across 569 genera. The blue circle outlines the Little Sac sites. Stress=0.173. P-value = 0.184

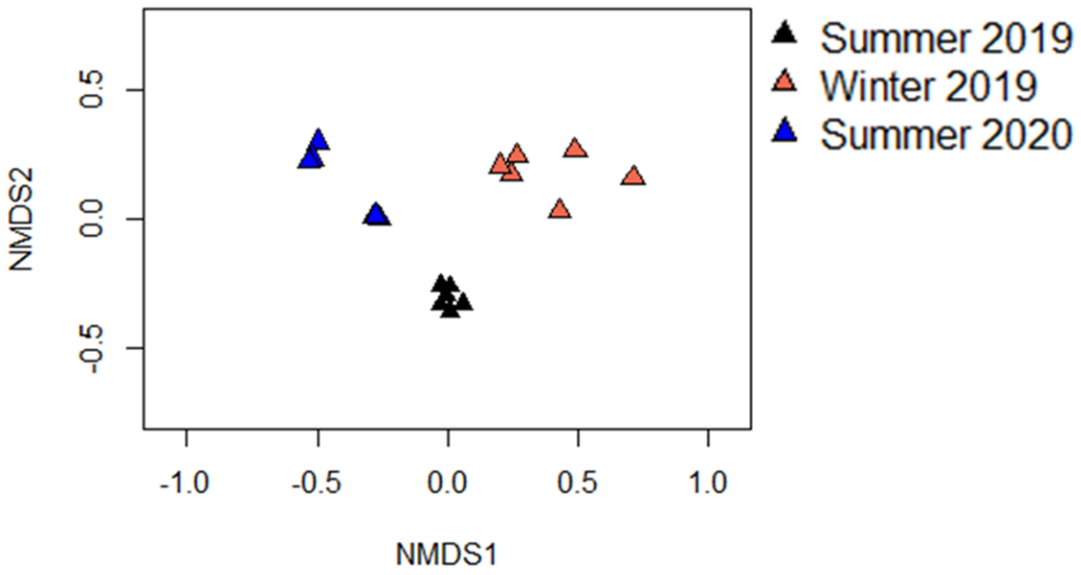
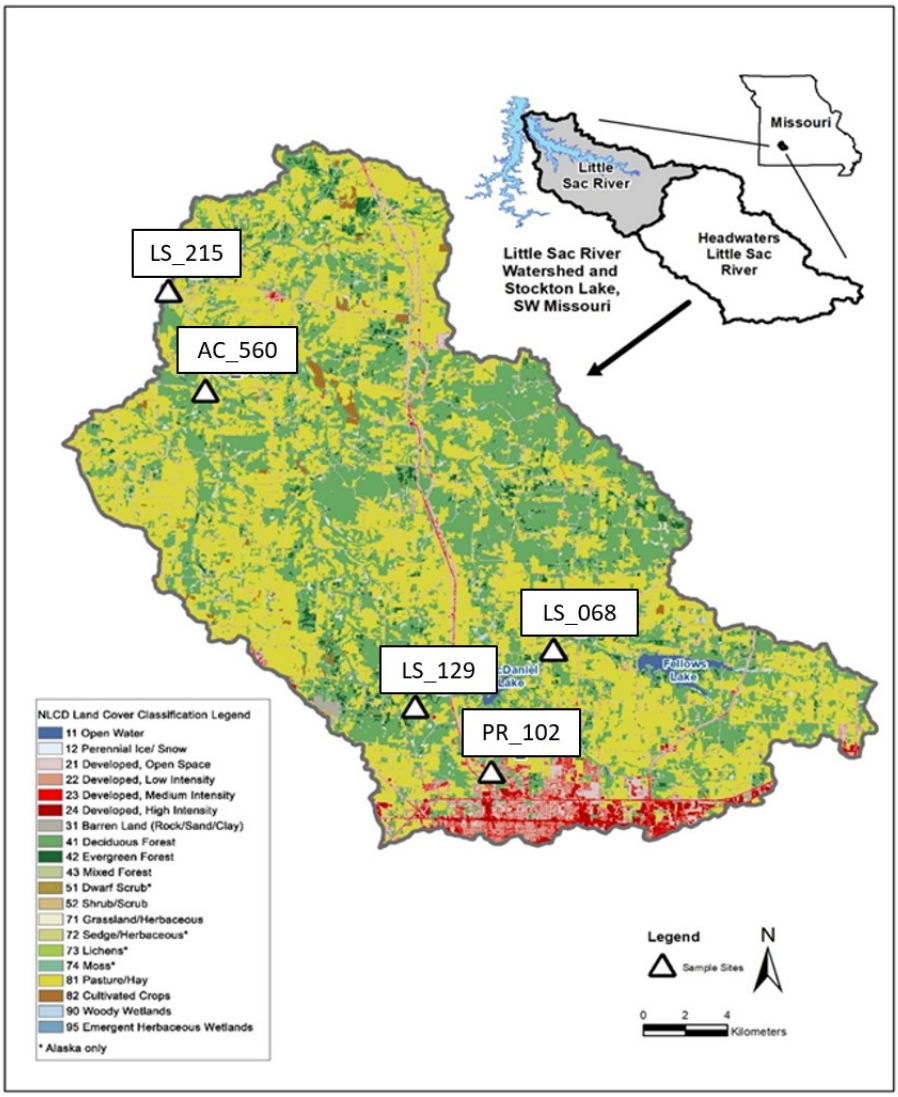


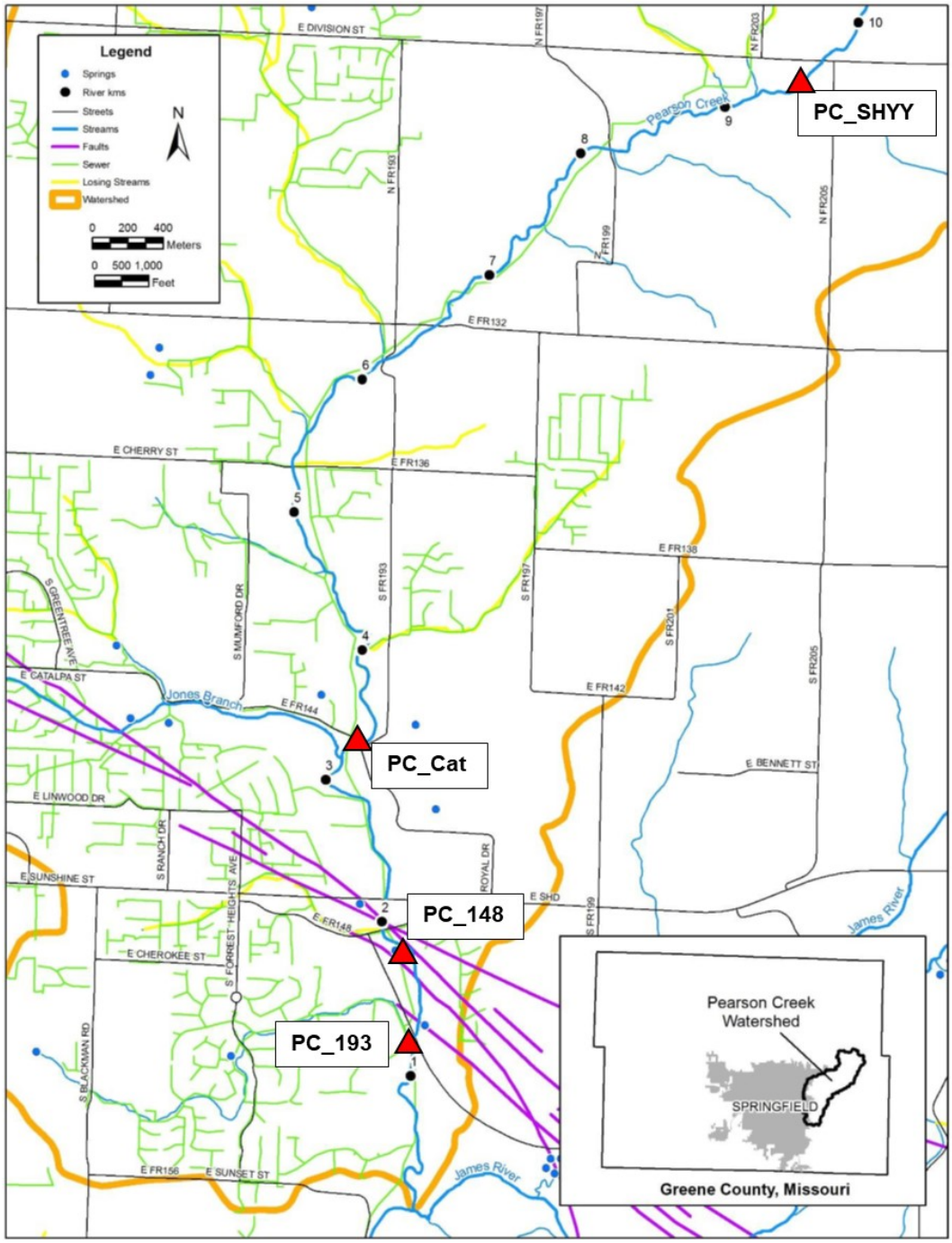
Figure 5. Non-metric multidimensional scaling plot of Sequiota Park's temporal sampling based on Bray-Curtis similarity index. Genera with 5 or more sequences across all sites were used for analysis, totaling 368,033 sequences across 656 genera. Stress = 0.061. P-value < 0.001

Supplemental table 1. Breakdown of sites from the three watersheds tested. Coordinates, months sampled, water filtered in liters (L), and total sequence after quality check are provided for each site.

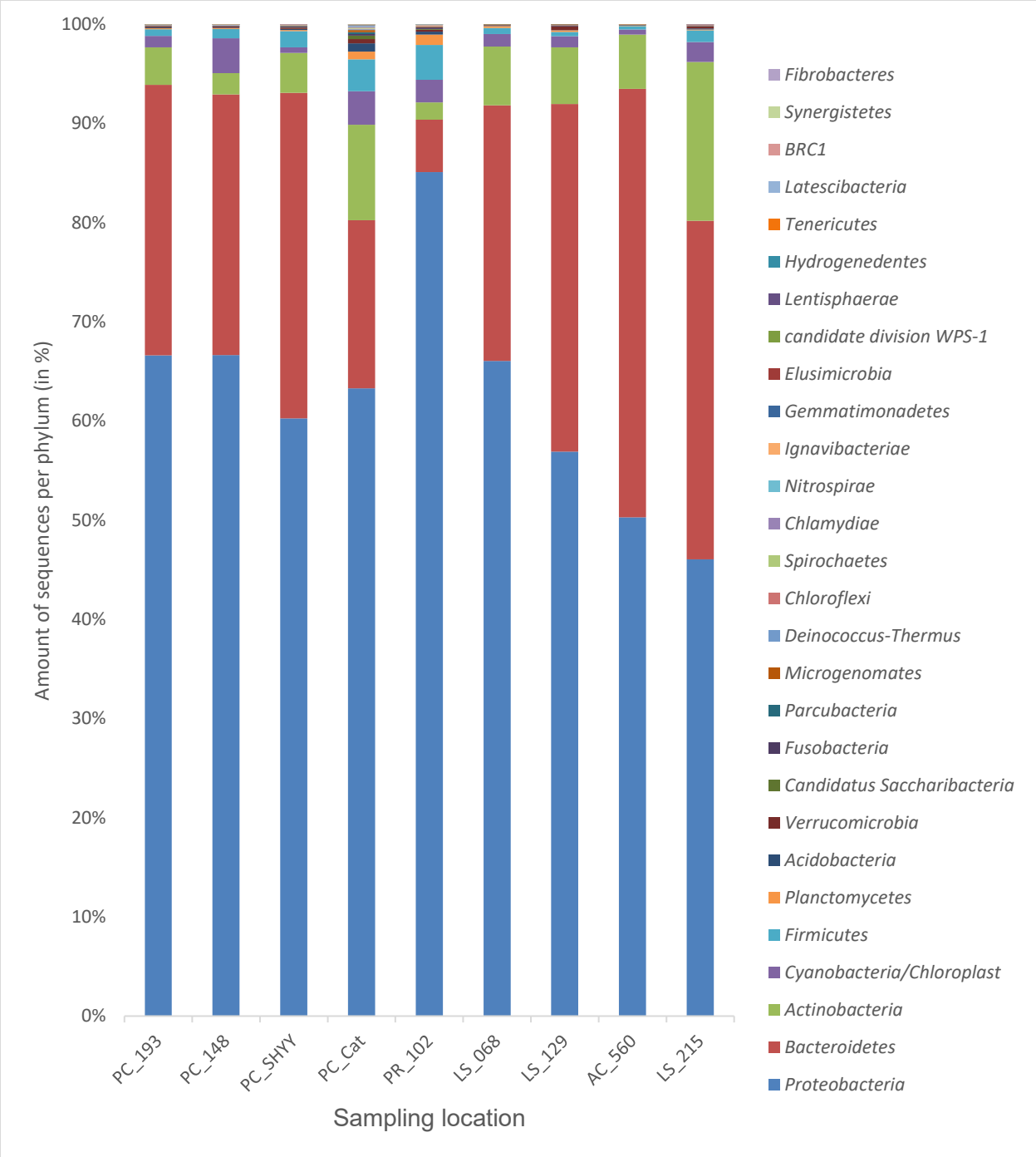
Watershed	Site ID	Coordinates	Month sampled in	Water filtered	Total Sequences
Little Sac	AC_560	37.437105 -93.465077	September/October of 2017	1 L	36,575
	LS_068	37.319043, -93.27665		0.4-1 L	47,543
	LS_129	37.292627, -93.350902		0.5 L	29,992
	LS_215	37.482787, -93.485443		1 L	53,280
	PR_102	37.263116, -93.309926		1-1.5 L	43,281
Pearson Creek	PC_148	37.17796, -93.198469	August of 2018	0.5-1.25 L	491,195
	PC_193	37.172568, -93.196443		0.75-1.5 L	583,940
	PC_Cat	37.187289, -93.199984		1.25-1.5 L	394,835
	PC_SHYY	37.17796, -93.198469		1-1.25 L	354,948
Sequiota Park's Spring	Summer 2019	37.147763, -93.236811	June of 2019	1.75-2 L	213,540
	Winter 2019		November/December of 2019	1.5-1.75 L	297,221
	Summer 2020		July of 2020	1-1.5 L	268,820



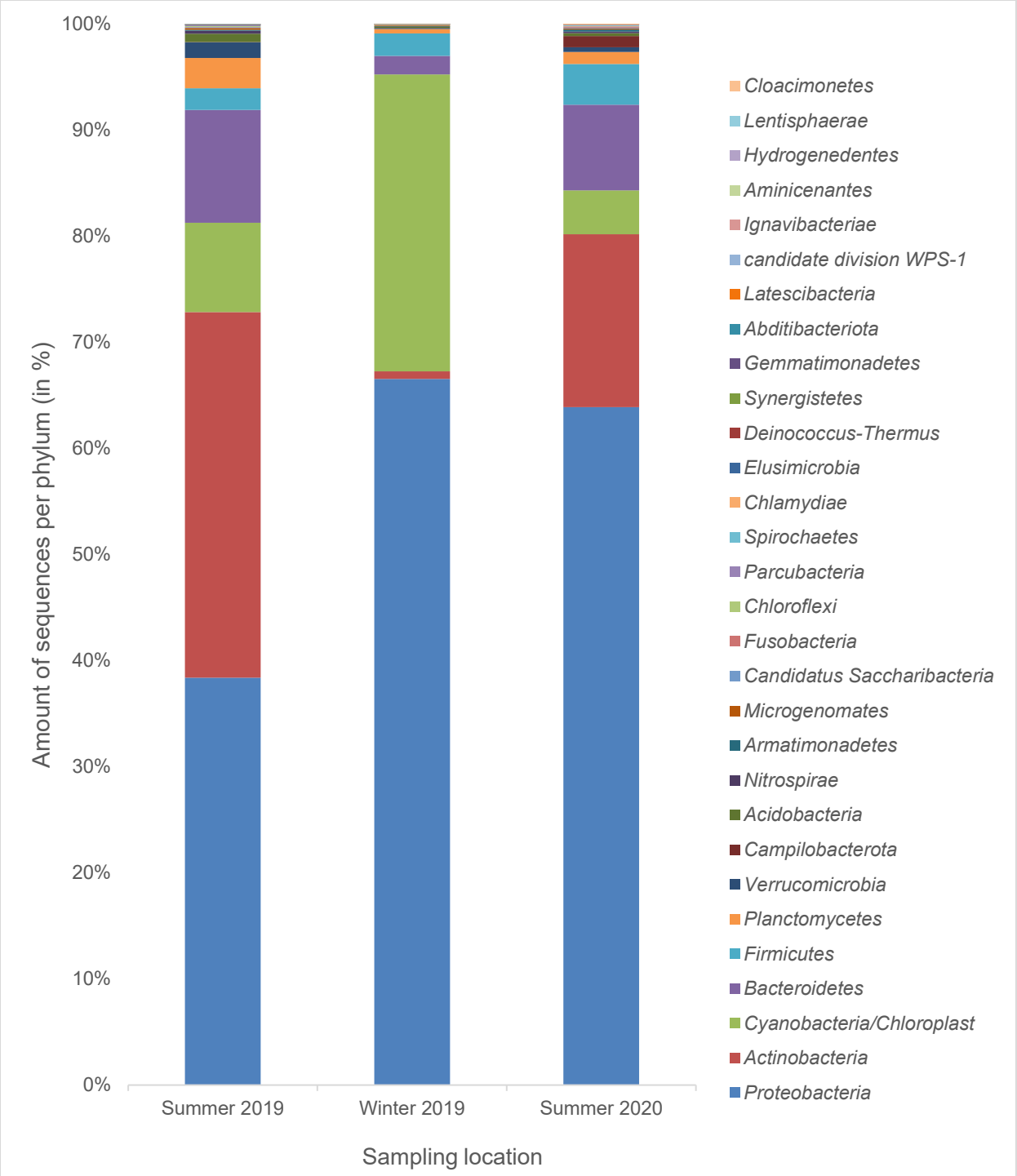
Supplemental Figure 1. Five sites along the Little Sac watershed that were chosen for sequence analysis are displayed with white triangles. The watershed flows from the southern sites (PR_102) toward the northern sites (LS_215). Figure modified from Mirza *et al*., 2018 with permission from OWERI, Missouri State University.



Supplemental Figure 2. Map of Pearson Creek. Numbered dots represent the distance going upstream in km. The four sites displayed (in red triangles) were chosen for sequencing analysis. The stream flows from north to south. Figure modified from Owen *et al*., 2019 with permission from OWERI, Missouri State University.



Supplemental Figure 3. Phylum level distribution of sequnces (2,035,589) as compared across PC and LSW sites. Sequnces classified as phylum-related were used in construction of of bargraph.



Supplemental Figure 4**.** Phylum level distribution of sequnces (779,581) as compared across temporal sampling of Sequiota Park's recharge spring. Sequnces classified as phylum-related were used in construction of of bargraph.

DIVERSITY OF CYANOBACTERIA AND ABUNDANCE OF CYANOTOXIN ASSOCIATED GENES WITHIN SPRINGFIELD WATER RESOURCES

Abstract

Cyanobacteria-dominated harmful algal blooms (CyanoHABs) are increasing in occurrence worldwide. Many of these blooms comprise naturally occurring species of cyanobacteria that reach nuisance levels under appropriate growth conditions. CyanoHABs can produce toxic secondary metabolites or cyanotoxins, with potentially detrimental impacts on ecosystem function as well as human and animal health. Despite the potential harmful effects on human health associated with the use of water containing toxic algal blooms, there are no formal federal regulation in place to address the presence or abundance of toxigenic species of cyanobacteria within water systems we use. So far, a limited number of studies have been conducted within the state of Missouri to investigate the diversity of cyanobacterial species and relative abundance of toxin associated genes. This study investigated cyanobacterial diversity and potential for cyanotoxin production in major water systems within the city of Springfield, Missouri. Samples were collected in the summer of 2020. Bacterial DNA was extracted, and samples were prepared for the Illumina-paired end DNA sequencing using both cyanobacterial-specific and universal 16S rRNA gene primers. I also investigated the relative abundance of various cyanotoxins gene such as genes associated with the production of anatoxin-a, microcystin, cylindrospermopsin and saxitoxin toxins using quantitative PCR. These results suggest the consistent presence of anatoxin-a gene containing cyanobacteria species within almost all samples. The genes for other cyanotoxins such as microcystin, cylindrospermopsin and saxitoxin were not detectable at all sampling sites. Overall, I detected high diversity of

cyanobacteria within different water samples. At the site LS_171 I observed high numbers of *Planktothrix* related sequences, and I also observed high abundance of anatoxin-a gene at this site. This suggests that an increase in anatoxin-a gene abundance is most likely due to the increase in *Planktothrix*-related species. Overall, these methods of testing can be useful for predicting and prevention of health risks associated with the cyanotoxin within waters of Springfield.

Keywords: Cyanobacteria; cyanotoxins; DNA sequencing; qPCR

Introduction

Cyanobacteria is a diverse phylum of naturally occurring photosynthetic bacteria in both freshwater and marine environments. Cyanobacteria are naturally present in water in low abundance, but under favorable growth conditions, their numbers can increase significantly and develop into Cyanobacteria-dominated harmful algal blooms (CyanoHABs). The development of CyanoHABs is reported to have a detrimental impact on aquatic ecosystems (Mur *et al*. 1999, Jonasson *et al*. 2010), and related toxins have been known to cause various waterborne illnesses in humans (Burch 2008) and wildlife (Hilborn and Beasley, 2015). Broadly, these cyanotoxins are classified into two main groups: hepatotoxins and neurotoxins. Within hepatotoxins group three frequently detected toxins are microcystins, nodularins, and cylindrospermosins which targets liver cells. Within neurotoxins group, which primarily targets nerve cells, three frequently reported toxins are saxitoxins, anatoxin-a and homoanatoxin-a (Quiblier *et al*., 2013). These toxins, along with other known cyanotoxins, can be produced by diverse cyanobacterial genera (Blaha, Babica, & Marsalek, 2009, Table S1) that are present in different aquatic environments.

To assess the toxigenic potential of a water system, it important to assess the diversity of overall cyanobacterial species as well as abundance of toxin associated genes within a water system. The production of toxins can be suddenly triggered by specific growth conditions (mainly unknown so far), and can seriously impact human health. In general, the presence of cyanobacteria that produces these toxins does not necessarily mean that the toxin is being produced. However, it indicates the toxigenic potential risk associated with the use of this water body. Hence, frequent testing and monitoring the relative abundance of various toxin-associated genes can indicate an early warning signal for the potential cyanotoxin in the water systems. Since different toxins can only be detected in a water sample when they are at high concentrations, if these toxins are detected within different water systems, then it is already too late to use that water for either drinking or recreational purposes.

Over the years, the frequency of CyanoHABs has substantially increased due to factors like climate change and increased nutrient availability (Lopez *et al*. 2008; Pearl *et al*. 2011). The worldwide increase in occurrence of CyanoHABs and associated toxins in diverse aquatic environments is a major health concern. Although cyanotoxins are widely acknowledged as a public health risk, there are no formal federal regulations in place in the United States that mandate CyanoHAB monitoring or management of impaired water ecosystems. Some states have created monitoring regulations for certain toxins based on the EPA's (2015) recommendations for drinking water. In the state of Missouri, the "EPA's 2007 National Lakes Assessment" reported that 13 of 28 lakes assessed within the state of Missouri showed moderate to high health risk due to a high abundance of cyanobacteria (> 20,000 cells/mL), a threshold set by the World Health Organization in its *Guidelines for Safe Recreational Water Environments*

(2003). Despite this, studies on cyanobacterial species diversity and associated cyanotoxins have been very limited.

For the city of Springfield, MO, 80% of the drinking water comes from surface waters (City Utilities, 2019). Two major bodies of water that are used for this are Fellows lake and McDaniel lake, which are both interconnected by the Little Sac River. Since little research has investigated the cyanobacterial diversity within these systems, this study aims to investigate the different cyanobacterial species present along these waters using Illumina pair-end DNA sequencing. In addition to this water system, a pond located in Sequiota Park and Lake Springfield were investigated. The pond is fed by a recharge spring, which then flows into Galloway creek. This creek empties into Lake Springfield, which is a system highly used for recreational activities such as boating, fishing, and swimming. Along with the diversity of Cyanobacteria, the presence of various cyanotoxin-coding genes was analyzed to determine the potential for cyanotoxin over-production during CyanoHABs. This study aims to gauge the potential health risks that could be posed by cyanobacterial blooms under favorable conditions, and to establish a methodology for future water monitoring.

Materials and Methods

Sample collection and filtration. Water samples were collected from 14 sites, five sites within Fellows lake, three sites of McDaniel lake, and six sites across the Little Sac watershed (Figure 1). Additionally, two locations, Sequiota Park's pond and Lake Springfield, were sampled (Table S2). From each location, in triplicate, water samples were collected in sterilized polypropylene carboy containers. These samplings were done on sunny days between the hours of 10:30am to 4pm to ensure the Cyanobacteria had risen to the top of the water. The water

samples were placed on ice and immediately transported to the laboratory where these water samples were immediately filtered through 0.22 μm Sterivex filters (Millipore Corporation, MA) using a peristaltic pump (Masterflex, Cole–Pamer Co, Vernon Hills, IL, USA). Approximately, 0.15 to 3.5 liters of water was filtered for each sample (Table S2). The volume of water filtered per location varied based upon concentration suspended particles which saturated filters and no more water could be filtered. These filters containing bacterial cells were frozen at -20^{o}C until further processing.

DNA extraction. For the DNA extraction, filters were cut into small fragments and placed into 50 mL tubes. Sterile water (25 mL) was added and vortexed for five minutes to detach bacterial cells from filters. Cells were harvested by centrifugation in an ultracentrifuge at 20,000 rpm for 15 minutes. Cell pellet was resuspended with 750 μL of PowerBead Solution for DNA extraction. Total genomic DNA was extracted using Qiagen's DNeasy PowerLyzer PowerSoil kits (Mo Bio, Carlsbad, CA) as per the manufacture's protocol. DNA was eluted with 35 μL sterile water and stored in -20^{o}C until further processing.

Real-time PCR-based quantification of cyanobacterial genes. Cyanobacterial-specific genes were quantified using real-time PCR. Reactions were performed using TaqMan or SYBR green assays, depending on the availability of probe (Table S3). Briefly, PCR reactions were performed in 20 μL, with each reaction containing 10 μL master mix (TaqMan or SYBR green), 50 nM of each primer, 50 nM of probe (if used) and 1 μL of template DNA. The PCR conditions were 95 °C for 5 minutes, followed by 40 cycles of 95 °C for 30 sec, and annealing/extension at 56 °C to 60 °C (depending on the gene) for 30 sec (Table S3). For SYBR green assays, dissociation curves was also analyzed for the specific gene amplification. For all the assays, a standard curve was created using plasmids containing the gene of interest. The plasmid was

serially diluted (10^{-1} to 10^{-9}) to create the standard curve. R^2 values were determined from these standards and ranged from 0.96-0.997 from the various assays. Copy numbers were standardized to copies per 100 mL of filtered water to ensure that the total amount of water filtered did not influence trends.

DNA sequencing. Bacterial communities from each site were assessed using Illumina MiSeq paired-end DNA sequencing. I amplified and sequenced cyanobacteria-specific 16S rRNA using cyan-specific PCR primers (Table S3). Also, the bacterial 16S rRNA gene was amplified and sequenced to assess the overall bacterial community structure using universal primers (Table S3). Both primer pairs were attached with adapter sequences and unique sample-based indices. These PCR reactions were in 25 µL reactions each reaction containing 12.5 µL 1X buffer, 0.2 µM of each primer, 2.0 mM $MgSO_4$, 0.2 µM of each deoxynucleoside triphosphates (dNTPs), 1.0 µL of template, and 5 units of High-Fidelity Platinum Taq polymerase high-fidelity PCR system enzyme (Invitrogen, USA). I used a two-step PCR approach, in the first PCR conditions were: an initial denaturing at 95ºC for 5 min, followed by 35 cycles of 95 °C for 30 sec, an annealing step at 56 °C for 30 sec, an extension at 72 °C for 30 sec, and a final extension for 7 minutes. The PCR products were tested for amplification on a 1% agarose gel and stained with ethidium bromide. The same PCR condition as described above were used the amplification of Cyanobacterial-specific 16S rRNA gene within the exception of specific primers and a 60ºC annealing temperature. Amplified PCR products were cleaned using ExoSap-IT PCR Cleanup System (USB, Cleveland, OH) as per manufacture's protocol.

Cleaned PCR products were used as the templates for a second PCR step. Reaction mixtures were the same as stated above, with the exception of the PCR primers. Details were the same as described previously (Hakim *et al*., 2018). The PCR conditions for the 2nd PCR were:

initial denaturation at 95 °C for 3 minutes, followed by 15 cycles of denaturation at 94 °C for 30 s, annealing at 60°C for 30 s, and extension at 72 °C for 30 s, with a final extension at 72°C for 7 minutes. PCR products from the 2nd PCR were pooled together in equimolar concentrations. Pooled DNA samples were purified using the Agencourt AMPure magnetic bead system (Beckman Coulter, Brea, CA). The purified PCR products were sequenced using Illumina MiSeq paired-end DNA sequencing platform. DNA sequencing was performed at the Center for Integrated Biosystems (CIB), Utah State University, USA.

Sequence processing and analysis. Sequences were processed for initial quality control parameters such as read length, ambiguous bases, removal of chimeric sequences, etc. using Ribosomal Database Project II (RDP) (Fish *et al*., 2013) (http://rdp.cme.msu.edu). High-quality filtered DNA sequences were classified using the RDP classifier. Sequences classified as GpI-XIII were extracted and dereplicated using Mothur (Schloss *et al*., 2009). These were each aligned and clustered using RDP at 97% similarity. Re-replication and abundance count tables were done using Python (version 2.7.2). Representative sequences from the most abundant operational taxonomic units (OTUs) were extracted and BLAST searched to determine the genus make-up of each group. Analysis and creation of box-and-whisker plots denoting the "Gp" distribution between sites was done using R version 4.0.2 (R Core Team, 2019). Shannon-Chao diversity was analyzed using RDP. Analysis of sequences was also done using Qiime2. Briefly, sequences were denoised and chimeric sequences were removed using the dada2 plugin. Sequences were clustered into amplicon sequence variances (ASVs) at 99% similarity, and then classified using the Greengenes 16S dataset.

Statistical analysis. Once sequences were processed and families/genera of interest were extracted, they were normalized to ensure total sequences retrieved from a location didn't

influence the trends. This was done by calculating the percent of amount of a specific classified family/genus to the total amount of sequences retrieved from a location. This was then multiplied by 100,000. This simulates as if every location contained only 100,000 sequences, in order to properly compare between sites.

A one-way ANOVA and post-hoc Tukey test was done on the results from the qPCR experiments to determine if there were significant differences between locations, as well as which locations were statistically similar to one another.

Results and Discussion

Sequences obtained after quality control. Next-Gen DNA sequencing using the universal 16S rRNA primers resulted in 4,883,733 good quality sequences after initial screening and removal of chimeric sequences. Approximately 9% of these sequences (444,542) were classified as Cyanobacteria-related genera. Illumina sequencing using Cyanobacteria-specific PCR primers resulted in 5,001,381 sequences obtained. Approximately 60% of these sequences (3,084,964) were classified as Cyanobacteria-related genera. Distribution of these sequences from each site retrieved by both primers are summarized in Table S4.

Cyanobacterial diversity using cyanobacterial-specific primers. Sequences classified at the family level under the class *Cyanobacteria* were investigated for differences between locations (2,120,380 total sequences). Due to similarities in Cyanobacterial 16S rRNA genes and lack of consistent formal nomenclature, most sequences were defined at the family level as Families I-XIII using RDP's classification methods. These families were determined based on phylogenetic analysis reported in the second edition of the Bergey's manual (Castenholz *et al*., 2001), which classifies Cyanobacterial families into 13 clusters (I-XIII). Overall, a majority of

all sequences retrieved were classified as Family II (79.6%). This was followed by Family IV (5.3%), Family VIII (4.7%), Family VI (4.2%), Family XIII (2.7%), and Family I (2.6%). The remaining families were less than 1% of the total sequences classified, and thus were not investigated further. When the distribution of these sequences was explored, a few differences were observed between the sites. Most of sequences retrieved from a majority of the sites were related to the Family II (70-96%) (Figure 2). However, at site LS_171, Family XIII (50%) was the most abundant family detected. Likewise, at sites PR_102 and AC_560 the most dominant sequences were closely related to Family VIII (53% and 62%, respectively), and site SP_Pond was primarily made up of Family VI (63%).

Within each of the 13 families using RDP's classifier, there are a group of genera defined as "Gp" (GpI-XIII, respectively per family), along with other genera that have normal nomenclature. Classified genera were investigated to determine which were the most abundant across all the sampling locations. Briefly, GpII from Family II made most of the sequences classified at the genera level (82.8%), followed by GpIV (5.8%), GpVI (4.6%), *Planktothrix* (2.5%, genus in Family XIII), GpVIII (2,1%), and GpI (1%). Genera less than 1% of all sequences were not investigated further. Since these Gp groups are related to multiple genera according to the Bergey's manual groupings, I investigated what genera were contributing to these sequences (See Table S5).

The genus *Synechococcus* was identified using qiime2 analysis to be the main contributor of GpII. This genus is commonly reported to be one of the most abundant freshwater cyanobacteria (Cabello-Yeves *et al*., 2017). Using python and BLAST, GpIV was determined to be mainly identified as *Plectomena* (84%). For GpVI, this group mostly appeared to be *Psuedoanabaena* (78% of sequences). GpVIII was mostly *Pleurocapsa* (98%). Lastly, group GpI

was mainly *Dolichospermum* (61%), with part of the sequences being related to *Chamaesiphon* (11%) and *Aphanizomenon* (3%). Of these genera detected, *Dolichospermum, Aphanizomenon,* and *Planktothrix* are known to produce the cyanotoxins saxitoxins and anatoxin-a(s) (Otten & Paerl, 2015). In addition, *Dolichospermum and Aphanizomenon* have been reported to produce cylindrospermosins, while *Dolichospermum* and *Planktothrix* also have the ability make microcystins in some species (Otten & Paerl, 2015).

The distribution of these GpII was most abundant in Fellows lake and Lake Springfield compared to the other locations (Figure 3a). Similarly, to the family level, LS_171, PR_102, AC_560, and SP_Pond displayed very low numbers of sequences related to this genus. GpIV was isolated to Fellows lake sites (Figure 3b). GpVI was only detected at site SP_Pond, while *Planktothrix* was detected only at LS_171 (Figure 3c and 3d, respectively). The two furthest downstream sites, PR_102 and AC_560, indicated the presence of sequences related to GpVIII (Figure 3e). Lastly, sites LS_171 and ML-2 contributed the bulk of the sequence classified as GpI (Figure 3f). Overall, similar trends that were observed at the family level were seen through these genera. In addition, *Synechococcus* appears to be dominate in lake systems, as seen in the distribution of Family II and GpII, which is comparable to other studies (Becker *et al*., 2007; Ruber *et al*., 2016).

Additionally, the distribution of the family and genus level sequences retrieved using the universal 16S primers were investigated to determine if a primer bias influenced what I observed. Similar methods were used for these sequences. Overall, abundances using 16S universal primers yielded similar results among most sites to the cyanobacteria-specific primers at the family and genus level (Figure S1 & S2a-e). One difference that was observed at site SP_Pond was the reduction of Family VI using the universal primers and an increase in Family I

when comparing to the Cyanobacterial-specific sequences from this site. Family VI was the most abundant family at this location using the Cyanobacterial-specific primer sequences, whereas with the 16S-primer sequences this family was hardly detected. Across all the sites, only 365 sequences were retrieved from this family using the universal primers. This suggests that the 16S primers may not be able to detect Family VI effectively in the environment as compared to the Cyanobacterial-specific primers. Additionally, at the genus level, there was a decrease in GpVIII at site AC_560 looking at the universal 16S-primer sequences.

Cyanobacterial diversity. Overall diversity of cyanobacterial sequences was investigated using Shannon-Chao estimators. Sequences obtained using the cyanobacterial specific primers showed a decrease in diversity across the Fellows lake locations with Shannon indices from 2.75-3.55, while all other locations had a Shannon index between 4.19-5.17, with the exception of site ML-3 (Table S6). The decrease in cyanobacterial diversity could be due to the high abundance of *Synechococcus* sequences in Fellows lake samples.

In addition to this, diversity across all of the bacterial community was investigated at phylum level (Figure S3) using sequences from the universal 16S rRNA primers. I observed four main bacterial phyla: *Proteobacteria, Bacteroidetes, Actinobacteria,* and *Cyanobacteria* (total 93-98%). This was consistent with what is normally reported within other aquatic environments (Zeglin, 2015). Shannon diversity indices were slight decreased at sites FL_3 and FL_5 (4.53 and 3.56, respectively), however, no consistent trend was observed between the rest of the sites which had diversity indices between 5.16-7.03 (Table S7). Although slight increases in cyanobacterial abundance were observed in Fellows lake, it does not appear that there were major shifts in the total bacterial communities.

Larger and stagnant waters display higher levels of *Cyanobacteria*. Total Cyanobacteria were analyzed using real-time PCR. Fellows lake displayed a higher relative abundance of Cyanobacteria as compared to the rest of the sites within the Little Sac watershed (Figure 4). Cyanobacteria decrease in abundance when leaving Fellows lake into LS_175 and LS_171, and then increase again in McDaniel lake. Additionally, both Sequiota park's pond and Lake Springfield displayed elevated levels of cyanobacteria. These data suggest that the stagnant environments of the lakes and pond sites is conducive for cyanobacterial growth. This observation is consistent with earlier studies as it is easier for cyanobacteria to grow on the shorelines of stagnant waters like lakes and ponds rather than flowing systems like rivers and streams (Crayton, 1993).

Anatoxin-a detected across sampling locations. The potential for cyanotoxin production was assessed using qPCR. The presence of four commonly reported cyanotoxins, saxitoxin, microcystin, cylindrospermopsin, and anatoxin-a, was determined using primers specific to each gene (Table S3). Of these four, only anatoxin-a was detected (Figure 5). Detection of this toxin gene was similar to the total amount of cyanobacteria detected and was detected at every sampling location, but there was not a significant difference between any of the sampling locations ($p = 0.062$). Furthermore, I investigated the proportion of the cyanobacteria that had the potential to produce this toxin by comparing the total copy numbers per liter of anatoxin-a gene detected to that of total cyanobacteria detected. Interestingly, LS_171 had a larger percentage of anaC containing cyanobacteria compared to the other sites (Table 1). Comparing this to the sequencing results, it could be suggested that the increase in this toxin gene stems from the increase in *Planktothrix* abundance at this site.

Health risk posed by anatoxin-a producing *Planktothrix*. I detected the presence of anatoxin-a gene at all sampling sites which suggests a potential health risk associated with anatoxin-a within the waters of Springfield. The anatoxin-a causes muscle contractions by simulating the effects of acetylcholine on muscle cells. However, anatoxin-a is unable to be degraded, leaving it bound to the receptor. This does not let the muscle go back to a resting state, leaving the muscle continually contracted. (Ilieva *et al*., 2019). Anatoxin-a is known to cause death in domestic animals and livestock within a few hours of its ingestion (Carmichael *et al*., 1978; Puschner *et al*., 2008). These data suggest that *Planktothrix* is the main contributor of the anatoxin-a gene at LS_171. In prior studies of *Planktothrix* that can produce microcystins, it has been shown that they produce more toxins per gram of biomass than other toxic Cyanobacteria (Fastner *et al*., 1999). In addition, blooms dominated by *Planktothrix* have been shown to occur in nitrogen-limited and phosphorus-rich environments (Davis *et al*., 2015). Contamination of this site through runoff rains could lead to increases in abundance of this bacterium, and possibly a largescale production of anatoxin-a. This could lead to hazardous conditions for wildlife and residents within this area.

Conclusion

Monitoring of the waters we use is important for assessing health risks and preventing many diseases. Since there are few regulations on the presence of cyanotoxins in water systems, it is imperative that studies like this one are done to determine what risks could be present in times of CyanoHABs. The current study suggests that anatoxin-a could be a health concern for the wildlife and domestic animals within the Springfield area during bloom forming conditions. Animals that use these waters for drinking may ingest this toxin, which could be fatal to the

animal within a few hours. In particular, site LS_171 appears to be of major concern due to the increase in *Planktothrix* spp. that are possibly able to produce anatoxin-a. It is important to note that this study was conducted during non-bloom forming conditions, which means it is unlikely that this toxin is produced in high abundance. To further this study in the future, the presence of this toxin should be analyzed to determine if it is being produced. This can be achieved most accurately through ELISA-based methods or testing for the mRNA of the toxin using qPCR. In addition, water quality parameters like total nitrogen, phosphorus, iron, etc. should be investigated to determine if the increases in *Planktothrix* is driven by these factors. To ensure that my sequencing results match the true environmental presence of *Planktothrix*, qPCR should be done using a genus-specific assay to determine if this Cyanobacteria is truly increased at site LS_171. Knowing this information would help to better assess the health risk of *Planktothrix* and anatoxin-a across Greene county water systems.

References

Al-Tebrineh, J., Gehringer, M. M., Akcaalan, R. & Neilan, B.A. 2011a A new quantitative PCR assay for the detection of hepatotoxigenic cyanobacteria. *Toxicon* **57**, 546-554

Al-Tebrineh, J., Pearson, L.A., Yasar, S.A. & Neilan, B.A. 2011b A multiplex qPCR targeting hepato- and neurotoxigenic cyanobacteria of global significance. *Harmful Algae* **15**, 19-25.

Becker, S., Richl, P. & Ernst, A. 2007 Seasonal and habitat-related distribution pattern of Synechococcus genotypes in Lake Constance. *FEMS Microbiology Ecology* **62**(1), 64-77.

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

Burch, M.D. 2008 Effective doses, guidelines & regulations. *Proceedings of Interagency, International Symposium on Cyanobacterial Harmful Algal Blooms (ISOC-HAB)*, p.831-853. In Hudnell HK (ed), State of science and research needs, Springer, New York.

Cabello-Yeves, P. J., Haro-Moreno, J. M., Martin-Cuadrado, A. B., Ghai, R., Picazo, A., Camacho, A. & Rodriguez-Valera, F. 2017 Novel Synechococcus genomes reconstructed from freshwater reservoirs. *Frontiers in Microbiology* **8**, 1151.

Carmichael, W. W. & Gorham, P. R. 1978 Anatoxins from clones of Anabaena flos-aquae isolated from lakes of western Canada: With 3 figures and 2 tables in the text. *Internationale Vereinigung für Theoretische und Angewandte Limnologie: Mitteilungen* **21**(1), 285-295.

Castenholz, R. W., Wilmotte, A., Herdman, M., Rippka, R., Waterbury, J. B., Iteman, I. & Hoffmann, L. 2001 Phylum BX. cyanobacteria. In *Bergey's Manual® of Systematic Bacteriology* (pp. 473-599). Springer, New York, NY.

City Utilities. 2019 Water quality report 2019. Retrieved from: https://www.cityutilities.net/wp-content/uploads/water-qualityreport.pdf

Crayton, M. A. 1993 Toxic cyanobacteria blooms: a field/laboratory guide. *Olympia: Office of Environmental Health Assessments, Washington State Department of Health*.

Davis, T. W., Bullerjahn, G. S., Tuttle, T., McKay, R. M. & Watson, S. B. 2015 Effects of increasing nitrogen and phosphorus concentrations on phytoplankton community growth and toxicity during Planktothrix blooms in Sandusky Bay, Lake Erie. *Environmental Science & Technology* **49**(12), 7197-7207.

EPA. 2015 Recommendations for public water systems to manage cyanotoxins in drinking water. Retrieved from: https://www.epa.gov/sites/production/files/2017-06/documents/cyanotoxin-management-drinking-water.pdf

Fastner, J., Neumann, U., Wirsing, B., Weckesser, J., Wiedner, C., Nixdorf, B. & Chorus, I. 1999 Microcystins (hepatotoxic heptapeptides) in German fresh water bodies. *Environmental Toxicology: An International Journal* **14**(1), 13-22.

Fish J.A., Chai B., Wang Q., Sun Y., Brown C.T., Tiedje J.M. & Cole J.R. 2013 FunGene: the functional gene pipeline and repository. *Frontiers in Microbiology* **4**, 291. https://doi.org/10.3389/fmicb.2013.00291.

Hakim, S., Mirza, B.S., Zaheer, A., Mclean, J.E. & Mirza, M.S. 2018 Retrieved 16S rRNA and nifH sequences reveal co-dominance of Bradyrhizobium and Ensifer (Sinorhizobium) strains in field-collected root nodules of the promiscuous host Vigna radiata (L.) R. Wilczek. *Applied Microbiology & Biotechnology* **102**, 485-497.

Hilborn, E. D. & Beasley, V. R. 2015 One health and cyanobacteria in freshwater systems: animal illnesses and deaths are sentinel events for human health risks. *Toxins* **7**(4), 1374–1395. https://doi.org/10.3390/toxins7041374

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. & Visser, P. M. 2018 Cyanobacterial blooms. *Nature Reviews Microbiology* **16**(8), 471-483.

Ilieva, V., Kondeva-Burdina, M., Georgieva, T. & Pavlova, V. 2019 Toxicity of cyanobacteria. Organotropy of cyanotoxins and toxicodynamics of cyanotoxins by species. *Pharmacia* **66**, 91.

Jonasson, S., Eriksson, J., Berntzon, L., Spáčil, Z., Ilag, L.L., Ronnevi, L.O., Rasmussen, U. & Bergman, B. 2010 Transfer of a cyanobacterial neurotoxin within a temperate aquatic ecosystem suggests pathways for human exposure. *Proceedings of the National Academy of Sciences* **107***,* 9252-9257.

Lopez, C.B., Jewett, E.B. & Dortch, Q., Walton, B.T., Hudnell, H.K. 2008 Scientific Assessment of Freshwater Harmful Algal Blooms. Interagency Working Group on Harmful Algal Blooms, Hypoxia, and Human Health of the Joint Subcommittee on Ocean Science and Technology. Washington, DC.

Mirza, B.S., Muruganandam, S., Meng, X., Sorensen, D.L., Dupont, R.R. & McLean, J.E. 2014 Arsenic(V) reduction in relation to iron (III) transformation and molecular characterization of the structural and functional microbial community in sediments of a basin-fill aquifer in Northern Utah. *Applied and Environmental Microbiology* **80**(10), 3198-3208

Mur, R., Skulberg, O.M. & Utkilen, H. 1999 Cyanobacteria in the environment. Toxic Cyanobacteria in Water: A Guide to their Public Health Consequences, p. 25-54. In *Chorus* I, Bartram J (eds), E & FN Spon, London

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

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

Pearl, H.W., Hall, N.S. & Calandrino, E.S. 2011 Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climate-induced change. *Science of the Total Environment* **409**, 1739-1745.

Puschner, B., Hoff, B. & Tor, E. R. 2008 Diagnosis of anatoxin-a poisoning in dogs from North America. *Journal of Veterinary Diagnostic Investigation* **20**(1), 89-92.

Quiblier, C., Wood, S., Echernique-Subiabre, I., Heath M, Villeneuve A, & Humbery J. F. 2013 A review of current knowledge on toxic benthic freshwater cyanobacteria – Ecology, toxin production and risk management. *Water Research* **47**, 5464-5479 http://dx.doi.org/10.1016/j.watres.2013.06.042

R Core Team 2019 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

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

Ruber, J., Bauer, F. R., Millard, A. D., Raeder, U., Geist, J. & Zwirglmaier, K. 2016 Synechococcus diversity along a trophic gradient in the Osterseen Lake District, Bavaria. *Microbiology* **162**(12), 2053-2063.

Sabart, M., Crenn, K., Perriere, F., Abila, A., Leremboure, M., Colombet, J., Jousse, C. & Latour, D. 2015 Co-occurrence of microcystin and anatoxin-a in the freshwater lake Aydat (France): Analytical and molecular approaches during a three-year survey. *Harmful Algae* **48**, 12-20

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J. & Weber C.F. 2009 Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537–7541. http://dx.doi.org/10.1128/AEM.01541-09

World Health Organization. 2003 *Guidelines for safe recreational water environments: Coastal and Fresh Waters* (Vol. 1). World Health Organization.

Zanchett, G. & Oliveira-Filho, E. C. 2013 Cyanobacteria and cyanotoxins: from impacts on aquatic ecosystems and human health to anticarcinogenic effects. *Toxins* **5**(10), 1896-1917.

Zeglin, L. H. 2015 Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Frontiers in Microbiology* **6**, 454. doi:10.3389/fmicb.2015.00454

Table 1. Gene copies of all *Cyanobacteria* and anatoxin-a gene (anaC) per liter of water.

Sampling location	All *Cyanobacteria* copies (L^{-1} water)	AnaC gene copies (L^{-1} water)	Percent of *Cyanobacteria* that contain AnaC
FL_1	2.39E+10	1.51E+06	0.01%
FL_2	7.72E+09	2.14E+06	0.03%
FL_3	3.56E+09	4.08E+05	0.01%
FL_4	3.5E+09	6.08E+05	0.02%
FL_5	1.06E+10	1.62E+06	0.02%
LS_175	4.17E+08	1.18E+05	0.03%
LS_171	2.13E+08	4.81E+05	0.23%
LS_159	2.59E+09	3.52E+05	0.01%
ML-1	2.25E+09	1.32E+06	0.06%
ML-2	8.55E+08	4.95E+05	0.06%
ML-3	1.13E+09	6.09E+05	0.05%
LS_SHO	3.58E+08	1.94E+05	0.05%
PR_102	76381665	3.42E+04	0.04%
AC_560	1.18E+08	1.16E+05	0.10%
SP_Pond	8.11E+09	1.68E+05	0.00%
SL	3.32E+10	1.96E+06	0.01%

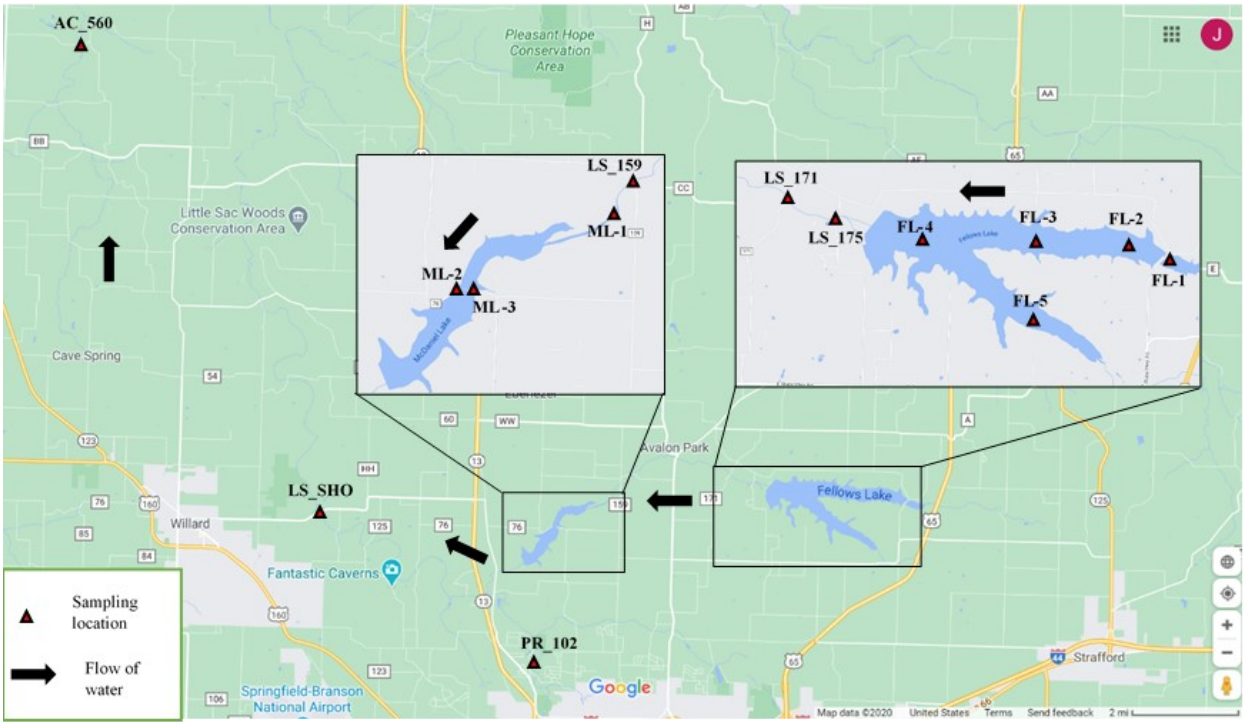


Figure 1. Map of Fellows/McDaniel lake watershed. Sampling locations are marked with red triangles. Black arrows denote the direction of water flow throughout the system. Map was modified from google maps.

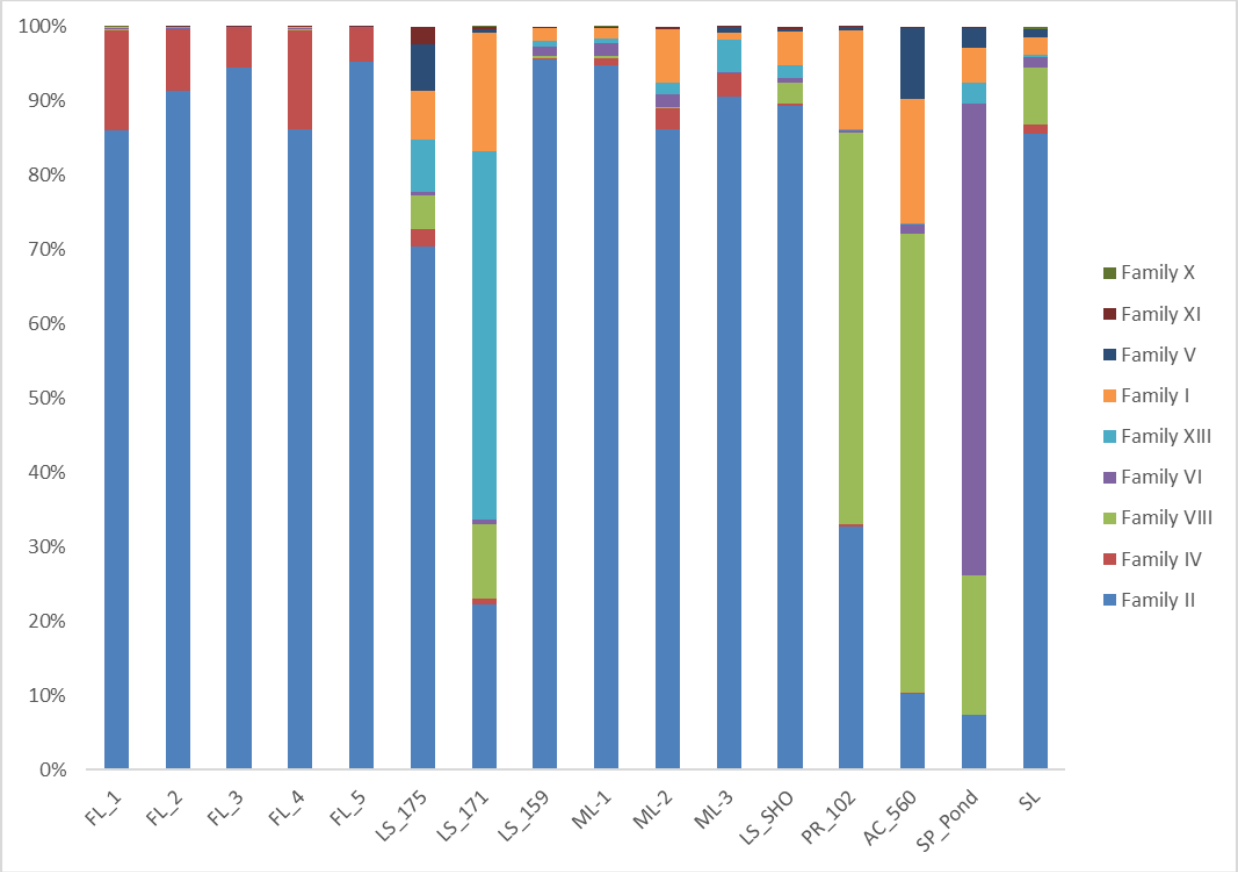


Figure 2. Distribution of family-level sequences among the 16 sampling sites using the Cyanobacteria-specific primers. The data expressed are normalized to prevent the influence of the total amount of sequences retrieved from each location (2,120,380 total sequences prior to normalization).

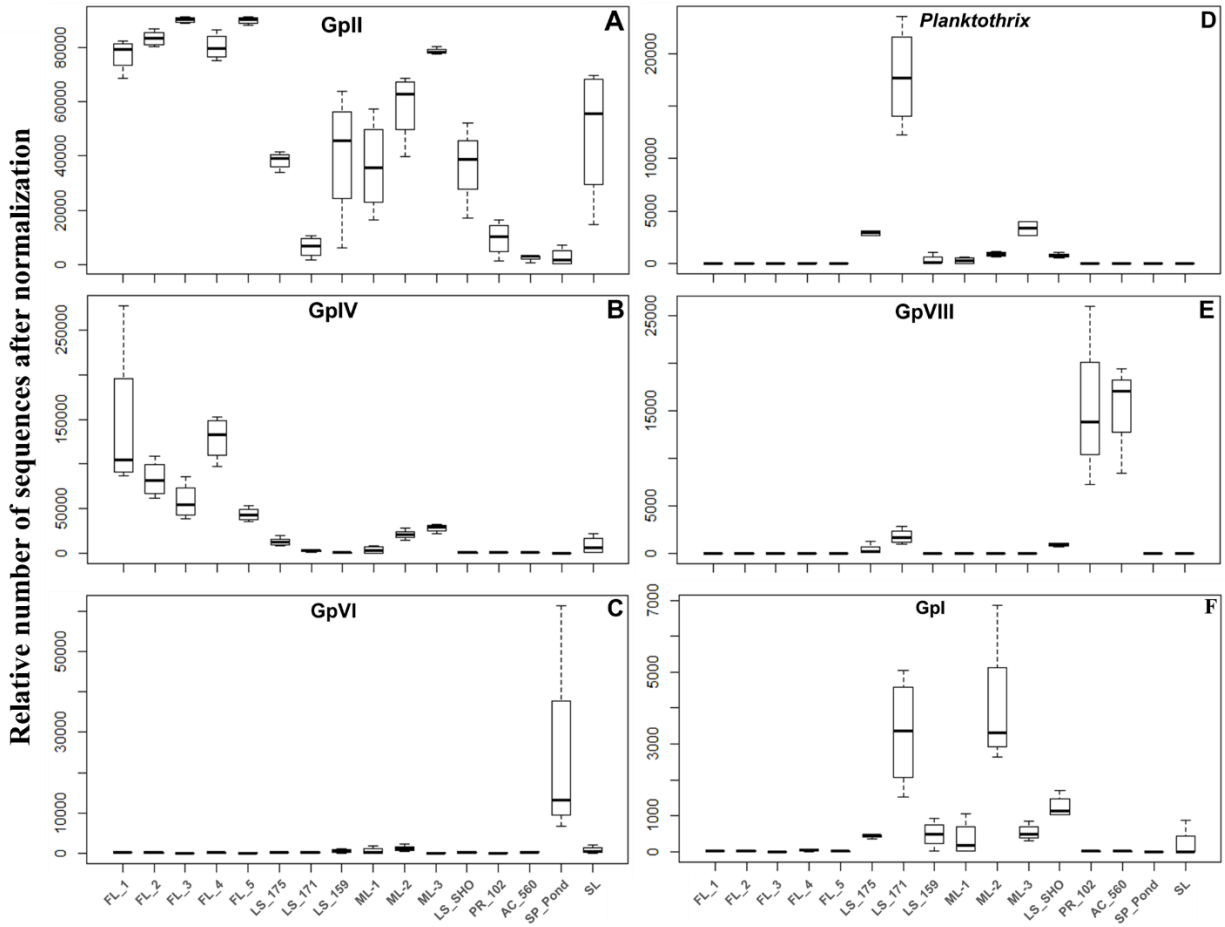


Figure 3. Distribution of sequences classified under "Gp" groups using Cyanobacteria-specific primers. Genera numbers were normalized as if each site contained 100,000 total sequences. A) Distribution of GpII-related sequences across sampling locations. B) GpIV. C) GpVI. D) *Planktothrix*. E) GpVIII. F) GpI.

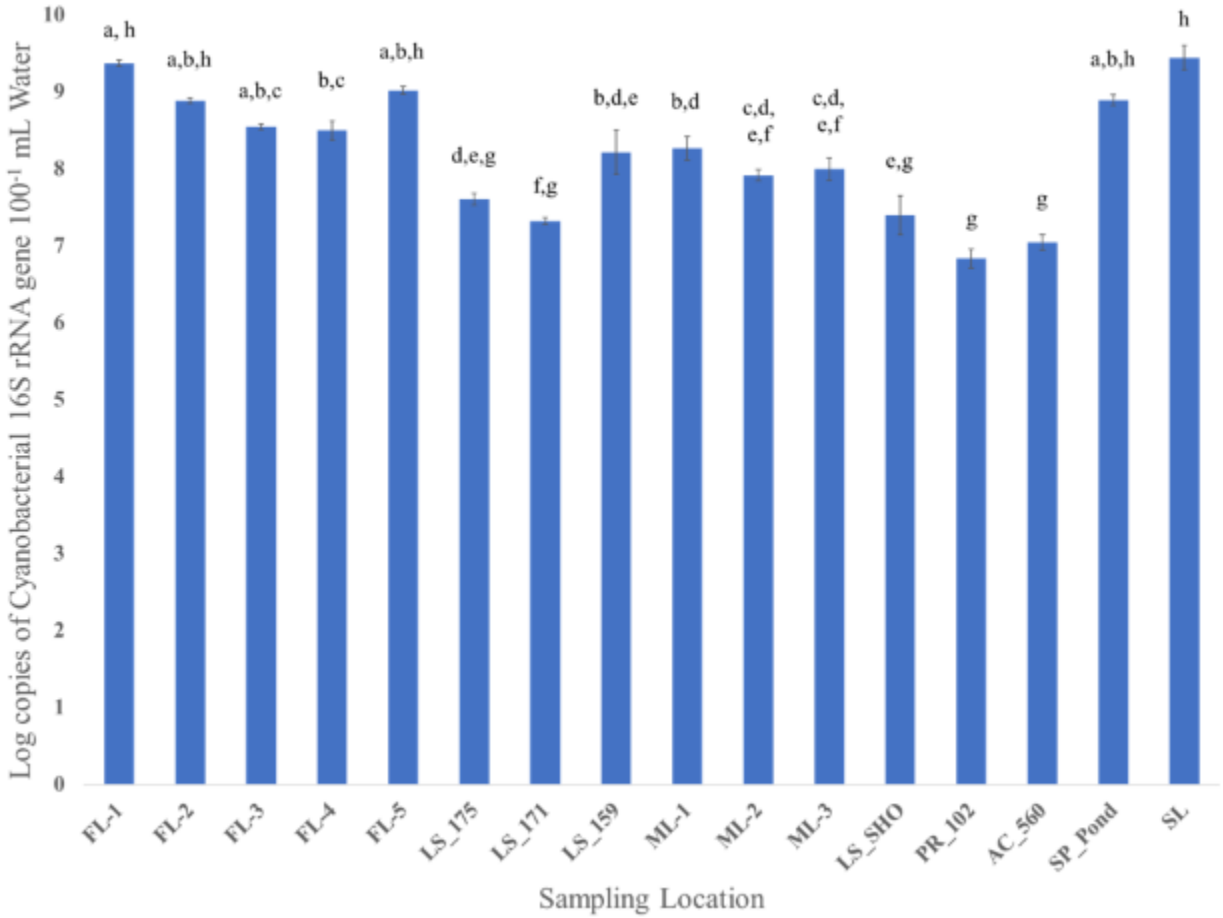


Figure 4. Log gene copies of all *Cyanobacteria* 16S gene per 100 mL of water. Each bar is an average of the three replicates and the error bars signify the standard error. A one-way ANOVA found that there was a significant difference between sampling locations ($p < 0.05$), and a post-hoc Tukey test was done to determine statistically different groups. Statistically similar groups were denoted with a letter above each bar.

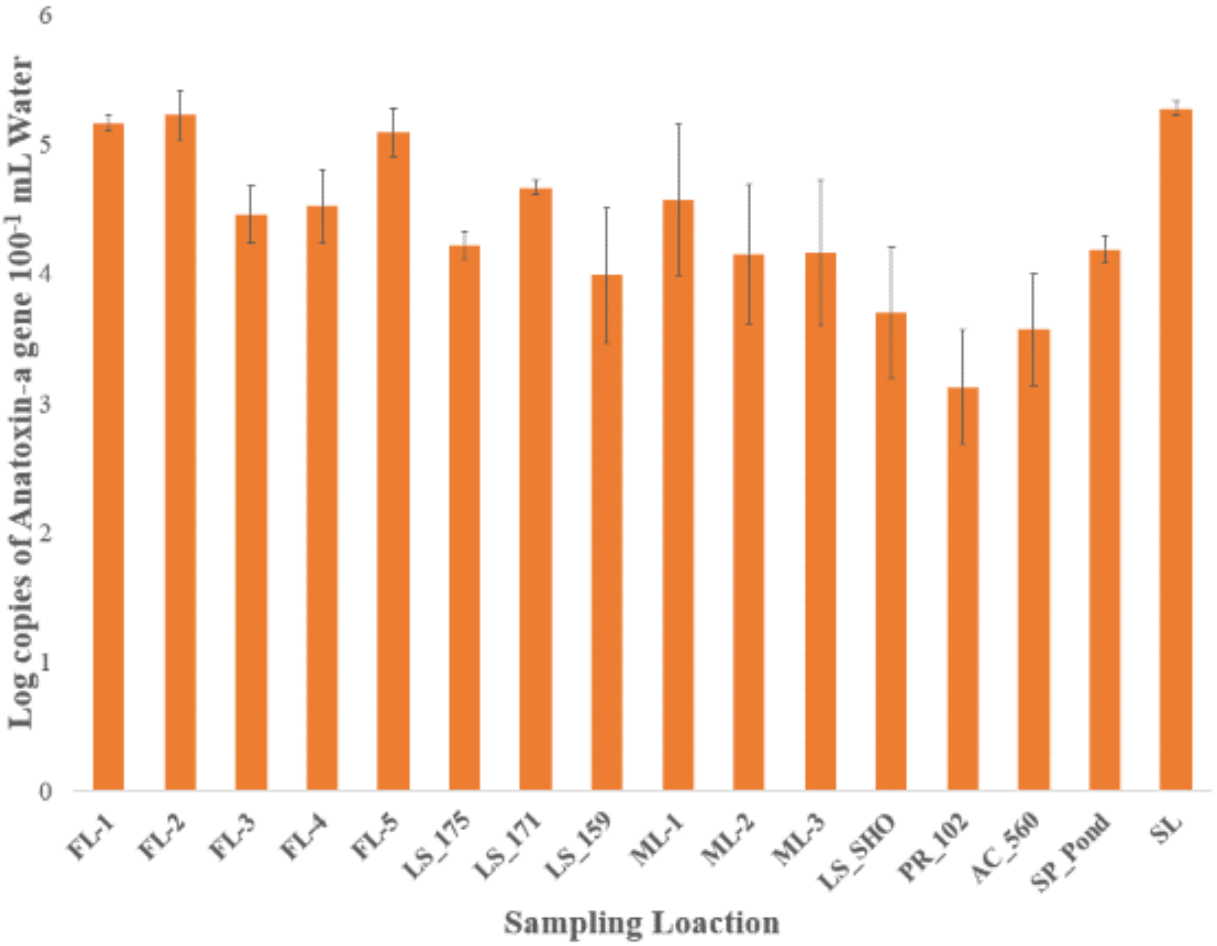


Figure 5. Log gene copies of the anatoxin-a gene (anaC) per 100 mL of water. Each bar is an average of the three replicates and the error bars signify the standard error. A one-way ANOVA found that there was no significant difference between the groups ($p = 0.062$).

Supplemental Table 1. Genera that contain species that are commonly reported to produce cyanotoxins, and which cyanotoxins are associated with the genera. Information gathered from literature (Zanchett & Oliveira-Filho, 2013; Otten & Paerl, 2015; Huisman *et al*., 2018; & Ilieva *et al*, 2019).

Genus	ANA	APLY	CYLIND	LYNG	MIC	NOD	SAX
Anabaenopsis					✓		
Aphanizomenon	✓		✓				✓
Chrysosporum			✓				
Cuspidothrix	✓						✓
Cylindrospermum	✓		✓				✓
Dolichospermum	✓		✓		✓		✓
Fischerella					✓		
Gloeotrichia					✓		
Hapalosiphon					✓		
Lyngbya		✓		✓			✓
Microcystis					✓		
Nodularia						✓	
Nostoc					✓		
Oscillatoria	✓	✓		✓	✓		
Phormidium	✓				✓		
Planktothrix	✓	✓			✓		✓
Raphidiopsis	✓		✓				✓
Schizothrix		✓					

ANA Anatoxin-a, **APLY** Aplysiatoxin, **CYLIND** Cylindrospermopsin, **LYNG** Lyngbyatoxin-a, **MIC** Microcystin, **NOD** Nodularin, **SAX** Saxitoxin

Supplemental Table 2. Sampling location details including coordinates, dates sampled, and water filtered. Replicates from each location were pooled into a range for water filtered.

Site ID	Coordinates	Date sampled	Water filtered
FL_1	37.310999, -93.182910	7/8/2020	0.35 L
FL_2	37.312036, -93.185893	7/8/2020	0.5 L
FL_3	37.314186, -93.204984	7/8/2020	0.4-0.5
FL_4	37.313876, -93.222157	7/8/2020	0.5 L
FL_5	37.301665, -93.200064	7/8/2020	0.45-0.5
LS_175	37.316432, -93.235795	6/19/2020	1.25-2 L
LS_171	37.319034, -93.243507	6/19/2020	0.75-2 L
LS_159	37.316183, - 93.280486	6/16/2020	0.4-0.5 L
ML-1	37.315949, -93.281259	6/16/2020	0.35-0.4 L
ML-2	37.304336, -93.304525	6/16/2020	0.5 L
ML-3	37.304002, -93.302202	6/16/2020	0.5-0.6 L
LS_SHO	37.308388, -93.383731	6/17/2020	1.4-2.5 L
PR_102	37.263116, -93.309926	6/19/2020	2.5-3.5 L
AC_560	37.437105, -93.465077	6/17/2020	1.25-2 L
SP_Pond	37.146647, -93.237862	7/21/2020	0.5 L
SL	37.116927, -93.249851	7/23/2020	0.15-0.25 L

Supplemental Table 3. Primers used for sequencing and qPCR. Primers/probe sequences are read from the 5' end to the 3' end. For qPCR sets, Taqman assays were used if a probe was available, and if no probe is denoted, then a SYBR Green assay was used.

Target and primer/probe sets	Primer/probe sequences	Annealing Temp	Reference
16S rRNA (sequencing)		56 ºC	Mirza *et al*. 2014
F515	GTGCCAGCMGCCGCGG		
R907	CCGTCAATTCMTTTRAGTTT		
Cyanobacteria-specific 16S rRNA (sequencing)		60 ºC	Nübel *et al*., 1997
CYAN359F	GGGGAATYTTCCGCAATGGG		
CYAN781R	GACTACTGGGGTATCTAATCCYTT		
All Cyanobacteria		58 ºC	Rinta-Kanto *et al*., 2005
CYAN108F	ACGGGTGAGTAACRCGTR		
CYAN377R	CCATGGCGGAAAATTCCCC		
General microcystin gene (mycE)		60 ºC	Al-Tebrineh *et al*., 2011a
DQmcyF	TTTAGAACSGGVGATTTAGG		
DQmcyR	CGRBTVADTTGRTATTCAATTTCT		
Saxitoxin gene (sxtA)		60 ºC	Al-Tebrineh *et al*., 2011b
sxtF	GGAGTGGATTTCAACACCAGAA		
sxtR	GTTTCCCAGACTCGTTTCAGG		
sxtP	FAM-TGCCGATTTAGAAGAAAGTATCCTCTCAG-Tamra		
Anatoxin-a gene (anaC)		56 ºC	Sarbart *et al*., 2015
anaC-gen-F2	TCTGGTATTCAGTMCCCTCYAT		
anaC-gen-R2	CCCAATARCCTGTCATCAA		
Cylindrospermopsin gene (cyrA)		60 ºC	Al-Tebrineh *et al*., 2011b
cyrF	GTCTGCCCACGTGATGTTATGAT		
cyrR	CGTGACCGCCGTGACA		
cyrP	Fam-CCTTTGGGAACGAAATTCTCGAAGCAACT-Tamra		

Supplementary Table 4. Amount of sequences retrieved from each location the three replicates. Amounts are shown as a range.

Location	Range for Cyanobacterial-specific sequences	Range for 16S sequences
FL1	49,362 - 78,128	46,250 - 76,382
FL2	69,689 - 107,373	55,532 - 84,983
FL3	28,409 - 55,124	40,135 - 118,195
FL4	102,181 - 124,698	67,135 - 111,498
FL5	131,047 - 143,156	134,434 - 161,256
LS_175	48,964 - 81,462	61,470 - 94,794
LS_171	43,091 - 133,954	67,532 - 106,609
LS_159	46,107 - 62,586	37,079 - 70,698
ML1	30,462 - 75,901	40,394 - 84,444
ML2	36,892 - 70,781	54,067 - 83982
ML3	42,425 - 81,830	44,885 - 76,749
LS_SHO	40,809 - 100,366	67,408 - 103,148
PR_102	23,771 - 70,076	62,365 - 79,994
AC_560	30,587 - 68,840	66,523 - 130,760
SP_Pond	112,226 - 147,216	53,962 - 166,569
SL	110,636 - 147,745	48,747 - 53,512

Supplemental Table 5. RDP classification of cyanobacterial sequences. Most abundant families that were detected are shown, along with the genera that RDP reports are in the families. Genera that contain species known to produce cyanotoxins are denoted, as well as the genera that my classified sequences showed were present.

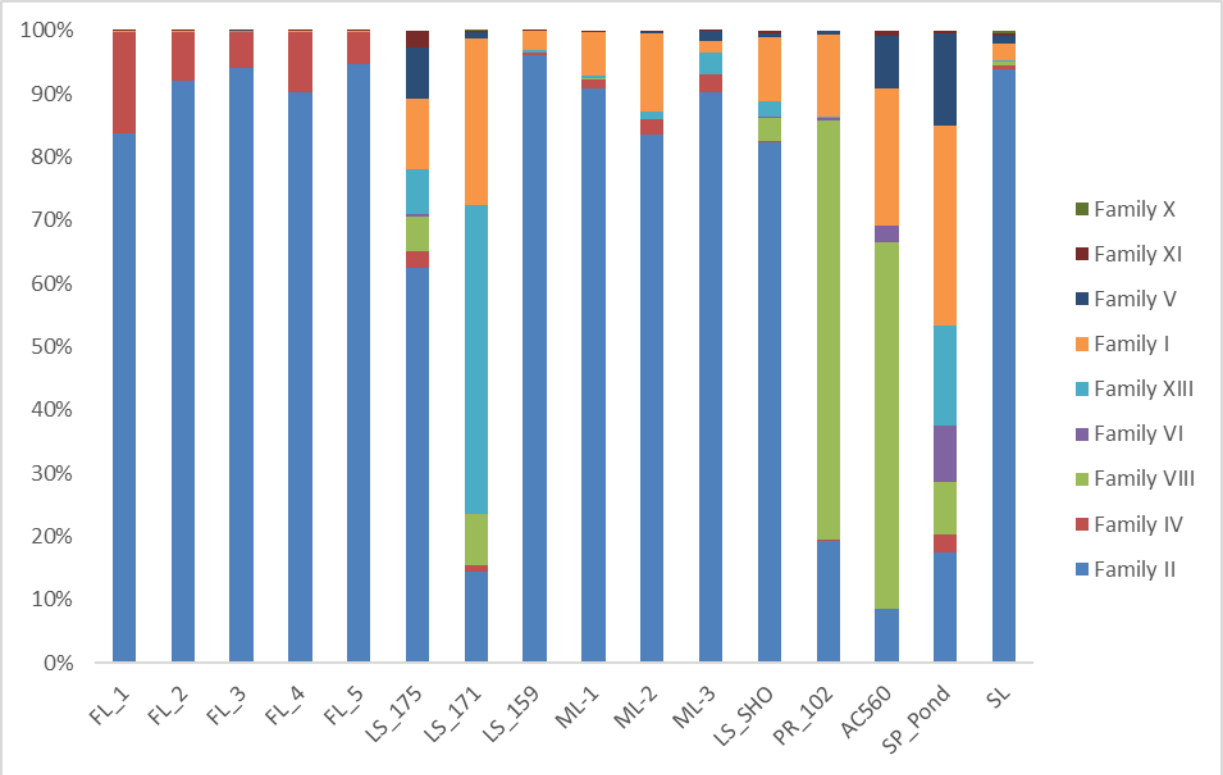
Most abundant Families	**Genera in family**	**Genera commonly reported to produce cyanotoxins**	**What I detected in high abundance**
Family I (2.6%)	GpI (*Nostoc*, *Calothrix*, *Cylindrospermum*, *Tolypothrix*, *Nodularia*, *Dolichospermum*, *Aphanizomenon*, *Fisherella*, *Raphidiopsis* & *Chamaesiphon*)	*Nostoc*, *Cylindrospermum*, *Nodularia*, *Dolichospermum*, *Aphanizomenon*, *Fisherella* & *Raphidiopsis*	*Dolichospermum*, *Aphanizomenon* & *Chamaesiphon*
Family II (79.6%)	GpII (*Synechococcus*, *Cyanobium* & *Microcystis*), *Prochlorococcus*	*Microcystis*	*Synechococcus*
Family IV (5.3%)	GpIV (*Plectomena*, *Leptolyngbya* & *Synechococcus*)	None	*Plectomena*
Family VI (4.2%)	GpVI (*Pseudanabaena*, *Limnothrix*, *Oscillatoria* & *Phormidium*)	*Oscillatoria* & *Phormidium*	*Pseudanabaena*
Family VIII (4.7%)	GpVIII (*Myxosarcina* & *Pleurocapsa*)	None	*Pleurocapsa*
Family XIII (2.7%)	GpXIII (*Arthrospira* & *Lyngbya*), *Planktothrix*, *Cephalothrix* & *Planktothricoides*	*Planktothrix*	*Planktothrix*

Supplemental Table 6. Shannon-Chao diversity of Cyanobacteria-specific sequences at each sampling location. Replicates were averaged and standard error is shown to the right of the average.

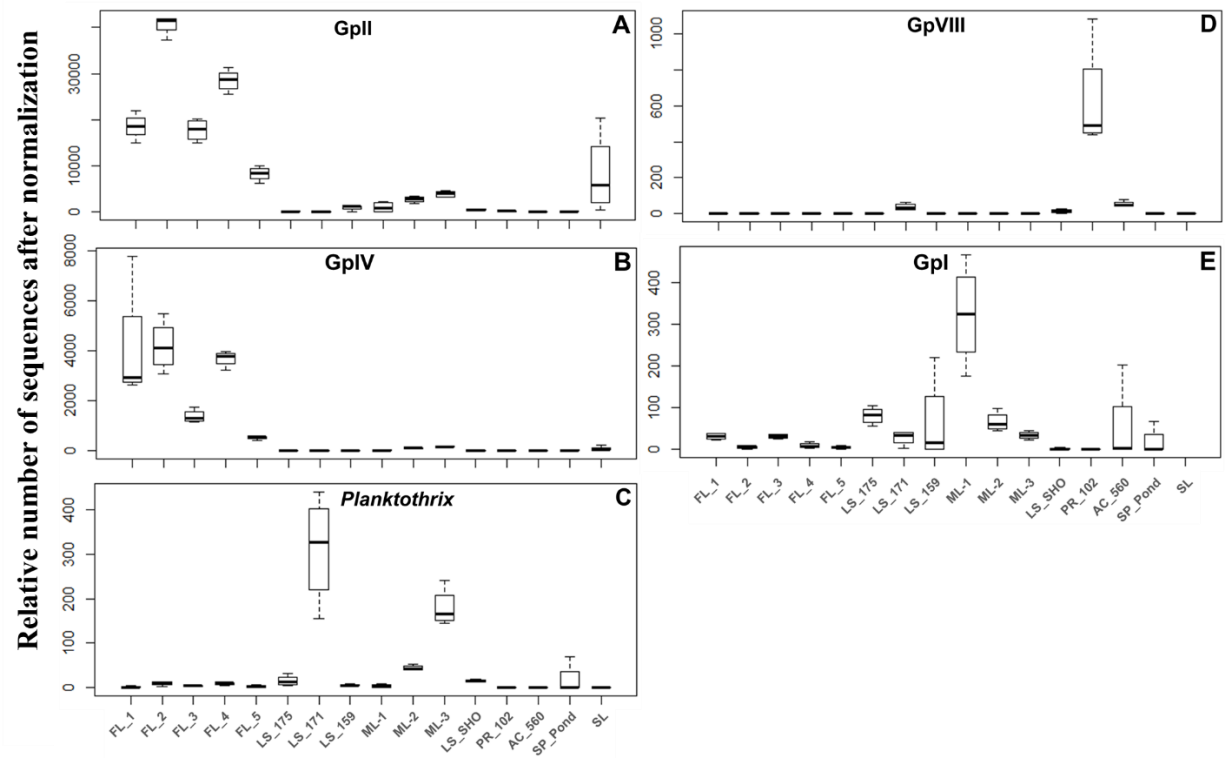
Sample ID	Average number of sequences	Chao 1 estimator	Shannon Diversity
FL_1	70233 ± 6965	1735 ± 359	3.29 ± 0.03
FL_2	85837 ± 8387	1840 ± 212	3.55 ± 0.003
FL_3	43727 ± 6192	926 ± 147	2.76 ± 0.01
FL_4	114788 ± 5231	2044 ± 229	2.86 ± 0.02
FL_5	137238 ± 2959	2565 ± 137	2.99 ± 0.01
LS_175	69353 ± 8315	6892 ± 373	5.16 ± 0.01
LS_171	75032 ± 20240	4927 ± 496	4.59 ± 0.02
LS_159	56068 ± 3927	3553 ± 426	4.31 ± 0.02
ML-1	56081 ± 9912	2927 ± 636	4.21 ± 0.01
ML-2	56946 ± 7653	2933 ± 527	4.21 ± 0.02
ML-3	64337 ± 8485	2411 ± 251	3.26 ± 0.01
LS_SHO	66859 ± 12361	3718 ± 727	4.36 ± 0.01
PR_102	51189 ± 10004	3415 ± 874	4.45 ± 0.03
AC_560	53587 ± 8164	5264 ± 1110	5.24 ± 0.01
SP_Pond	122593 ± 6662	4722 ± 425	4.32 ± 0.04
SL	131121 ± 9595	9789 ± 1153	5.08 ± 0.03

Supplemental Table 7. Shannon-Chao diversity of universal 16S rRNA gene sequences at each sampling location. Replicates were averaged and standard error is shown to the right of the averages.

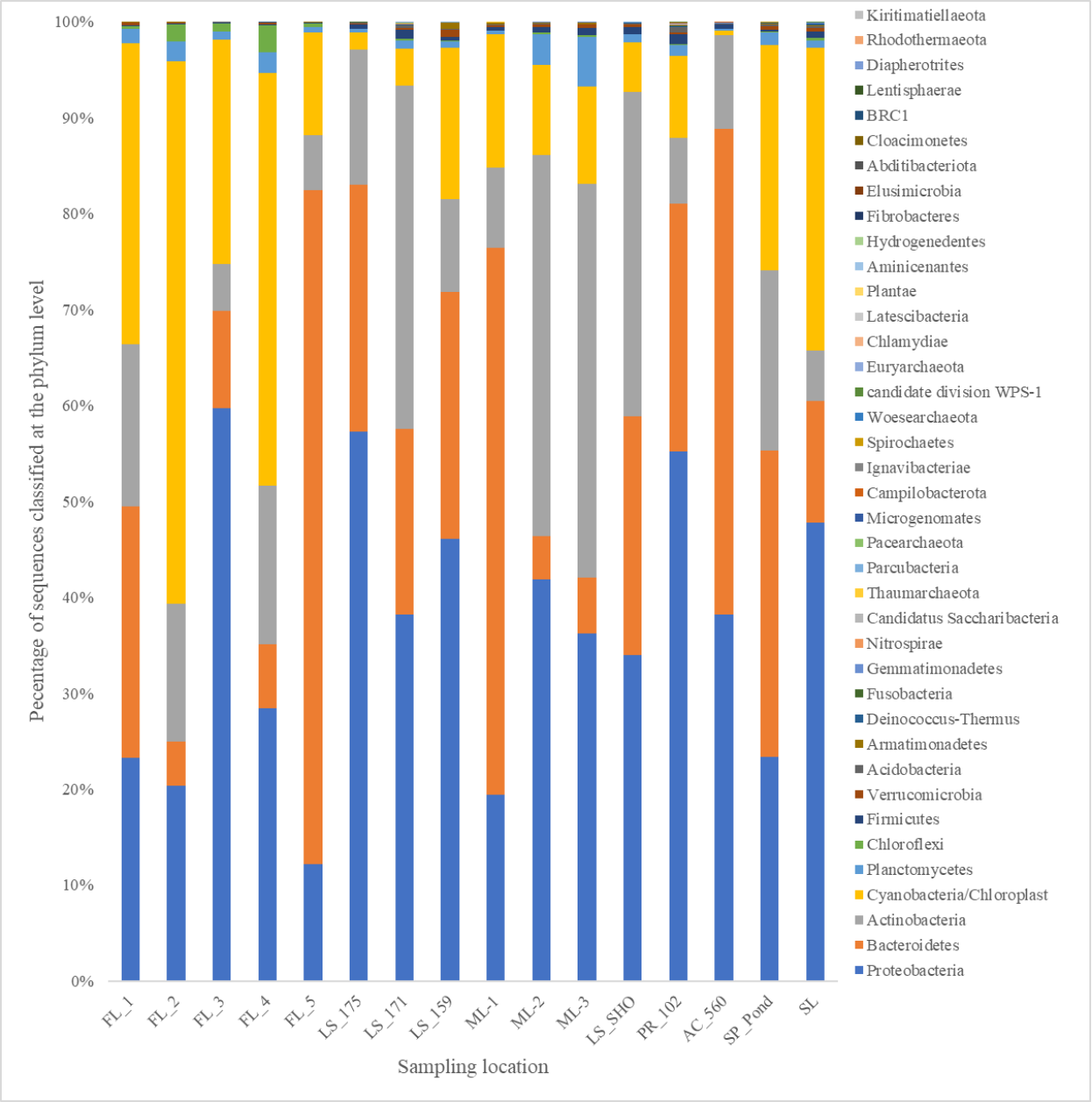
Sample ID	Average number of sequences	Chao 1 estimator	Shannon Diversity
FL_1	55338 ± 7054	14450 ± 1299	6.24 ± 0.03
FL_2	67289 ± 6311	13894 ± 1095	5.76 ± 0.04
FL_3	76945 ± 16124	9794 ± 1047	4.53 ± 0.12
FL_4	75328 ± 7079	18301 ± 1312	6.15 ± 0.05
FL_5	143216 ± 3206	10902 ± 496	3.56 ± 0.15
LS_175	74389 ± 8041	14177 ± 724	5.80 ± 0.04
LS_171	84690 ± 10000	26512 ± 3211	6.23 ± 0.18
LS_159	55569 ± 7120	15395 ± 2138	6.62 ± 0.14
ML-1	59809 ± 9469	9232 ± 3105	5.40 ± 0.59
ML-2	65263 ± 6737	13812 ± 1775	5.84 ± 0.19
ML-3	65154 ± 6974	14529 ± 1392	6.10 ± 0.11
LS_SHO	81102 ± 8145	20607 ± 1436	6.52 ± 0.03
PR_102	73394 ± 3988	28857 ± 3084	6.98 ± 0.19
AC_560	102612 ± 12019	16754 ± 595	5.16 ± 0.11
SP_Pond	81074 ± 22698	22830 ± 3185	6.28 ± 0.22
SL	50852 ± 1134	24421 ± 2707	7.03 ± 0.41



Supplemental Figure 1. Distribution of family-level sequences among the 16 sampling sites using universal 16S rRNA gene primers. The data expressed are normalized to prevent the influence of the total amount of sequences retrieved from each location (545,707 total sequences prior to normalization).



Supplemental Figure 2. Distribution of sequences classified under "Gp" groups using universal 16S rRNA gene primers. Genera numbers were normalized as if each site contained 100,000 total sequences. A) Distribution of GpII-related sequences across sampling locations. B) GpIV. C) *Planktothrix*. D) GpVIII. E) GpI.



Supplemental Figure 3. Bacterial composition at the phylum level between sampling locations using universal 16S rRNA primers.

SUMMARY

Rapid urbanization and agricultural development have caused the deterioration of water quality. Consistent monitoring of the water systems for waterborne pathogens is important to prevent disease outbreaks. Previous studies in the region heavily relied on the testing of fecal indicator bacteria (FIB) to assess health risks associated with a watershed. These methods helped to identify health concerns associated with the potential use of bacterially impaired watershed. However, the presence of FIB does not always correlate well with the presence of actual human pathogens (Levy *et al*., 2012).

In this thesis, I used a high-throughput DNA sequencing approach to assess the diversity and abundance of bacterial pathogens in bacterially impaired water systems within Greene and Polk counties. I classified overall bacterial communities, as well as determined the relative distribution of pathogenic genera. This initial screening approach could be useful in identifying pathogens that need to be closely monitored. This method can also avoid wasted time and resources in the testing for pathogens that are not present. In this study, I detected the presence of DNA sequences related to genera that are known to cause human diseases such as *Legionella*, *Yersinia*, and *Shigella/E. coli*. Knowing the relative distribution of these different pathogens could help to assess the health risks directly associated with water use and prevent disease outbreaks.

I also evaluated the diversity of Cyanobacteria species within the waters of Springfield along with the relative abundance of different cyanotoxin genes. I detected the presence of the anatoxin-a gene in waters used for drinking and recreational purposes in Springfield. Although detection of a gene responsible for toxins is an indirect method of testing for the toxin, it can be

used to infer that if cyanobacterial blooms occur, this toxin is most likely to be present in higher concentrations in the water (Ribeiro *et al*., 2020). The presence of the anatoxin-a gene and the cyanobacterial species containing this gene suggests this toxin can be a major concern for the use of this water for drinking, recreation, and for livestock and pets. Although death in humans is not reported, this toxin is known to be very deadly to many animals (Hilborn and Beasley, 2015). In addition to detecting this cyanotoxin, I used a similar approach with high-throughput DNA sequencing to determine if genera of known anatoxin-a producing Cyanobacteria are present in our water systems. Overall, there was an increase in *Planktothrix* at one site, LS_171, which may be driving the increase in anatoxin-a producing cyanobacteria at this location. Knowing what cyanotoxins could be produced by certain *Cyanobacteria* during blooms can allow for preventative measures to be taken and inform the public of the potential risks to themselves and their domestic animals. Possible variables that could be a trigger for cyanotoxins could be nutrient limitation, quorum sensing, or temperature variations.

Overall, knowing the true health risk of the waters we use is an important factor to prevent more diseases and deaths from occurring each year. In this thesis, I analyzed many of the water systems that are used in Greene and Polk counties for the presence of known human pathogens and cyanotoxins.

REFERENCES

Browne, H. P., Neville, B. A., Forster, S. C. & Lawley, T. D. 2017 Transmission of the gut microbiota: spreading of health. *Nature Reviews Microbiology* **15**(9), 531-543.

Baffaut, C. 2006 Total maximum daily load (TMDL) for Little Sac river watershed. *FAPRI-UMC Report*, 07-05.

Bullard, L., Thomson K. C. & Vandike J. E. 2001 The Springs of Greene County Missouri. Missouri Department of Natural Resources Geological Survey and Resource Assessment Division. Water Resources Report No. 68.

Fields, S. 2004 Global nitrogen: cycling out of control. *Environmental Health Perspectives* **112**, A556-A563. https://doi.org/10.1289/ehp.112-a556

Fields, B. S., Benson, R. S. & Besser, R. E. 2002 *Legionella* and Legionnaires's disease: 25 years of investigation. *Clinical Microbiology Reviews* **15**, 506-526.

Hilborn, E. D. & Beasley, V. R. 2015 One health and cyanobacteria in freshwater systems: animal illnesses and deaths are sentinel events for human health risks. *Toxins* **7**(4), 1374-1395. https://doi.org/10.3390/toxins7041374

Levy, K., Nelson, K. L., Hubbard, A. & Eisenberg, J. N. 2012 Rethinking indicators of microbial drinking water quality for health studies in tropical developing countries: case study in northern coastal Ecuador. *The American Journal of Tropical Medicine and Hygiene* **86**(3), 499-507. https://doi.org/10.4269/ajtmh.2012.11-0263

Missouri Department of Natural Resources (MDNR) 2020 Approved Section 303(d) Listed Waters. https://dnr.mo.gov/env/wpp/waterquality/303d/docs/2020-303d-list-cwc-approved-2020-04-02.pdf

Newton, H. J., Ang, D. K. Y., Driel, I. R. & Hartland, E. L. 2010 Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clinical Microbiology Reviews* **23**(2), 274-298.

Renter, D. G., Gnad, D. P., Sargeant, J. M. & Hygnstrom, S. E. 2006 Prevalence and Serovars of *Salmonella* in the feces of free-ranging white-tailed deer (*Odocoileus virginianus*) in Nebraska. *Journal of Wildlife Diseases* **42**(3), 699-703. https://doi.org/10.7589/0090-3558-42.3.699

Ribeiro, M. S., Tucci, A., Matarazzo, M. P., Viana-Niero, C. & Nordi, C. S. 2020 Detection of Cyanotoxin-Producing Genes in a Eutrophic Reservoir (Billings Reservoir, São Paulo, Brazil). *Water* **12**(3), 903.

Rouffaer, L. O., Baert, K., Van den Abeele, A. M., Cox, I., Vanantwerpen, G., De Zutter, L., Strubbe, D., Vranckx, K., Lens, L., Haesebrouck, F., Delmée, M., Pasmans, F. & Martel, A. 2017 Low prevalence of human enteropathogenic Yersinia spp. in brown rats (Rattus norvegicus) in Flanders. *PloS one* **12**(4), e0175648. https://doi.org/10.1371/journal.pone.0175648

Sercu, B., Werfhorst, L. C. Van De, Murray, J. L. S. & Holden, P. A. 2011 Sewage Exfiltration as a Source of Storm Drain Contamination during Dry Weather in Urban Watersheds. *Environmental Science & Technology* **45**, 7151–7157. https://doi.org/10.1021/es200981k

Skov, M. N., Madsen, J. J., Rahbek, C., Lodak, J., Jespersen, J. B., Jørgensen, J. C., Dietz, H. H., Chriél, M. & Baggesen, D. L. 2008 Transmission of *Salmonella* between wildlife and meat-production animals in Denmark. *Journal of Applied Microbiology* **105**(5), 1558-1568. doi: 10.1111/j.1365-2672.2008.03914.x.

Watershed Committee of the Ozarks (WCO). 2016 Sac River Healthy Watershed Plan. October 2016.

World Health Organization (WHO). 2019 Water sanitation hygiene. Retrieved from: https://www.who.int/water_sanitation_health/diseases-risks/en/

Wright Water Engineers (WWE) (2001) Southern Hills Lakes Preliminary Evaluation and Management Plan: Summary Report. Prepared for the City of Springfield, April 2001.