Exploring the toxigenicity and genetic similarity of Detroit Reservoir’s recurring cyanobacterial bloom in 2017 and 2018

by
Tejas Godbole

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Microbiology
(Honors Scholar)

Presented May 30, 2019
Commencement June 2019

AN ABSTRACT OF THE THESIS OF

Tejas Godbole for the degree of Honors Baccalaureate of Science in Microbiology presented on May 30, 2019. Title: Exploring the toxigenicity and genetic similarity of Detroit Reservoir's recurring cyanobacterial bloom in 2017 and 2018.

Abstract approved:____

Theo Dreher

In order to understand the ability of Detroit Reservoir's recurring cyanobacterial bloom to produce toxins, and determine the genetic similarity of the bloom from year to year, environmental samples were taken from the Detroit Reservoir bloom biomass in the summers of 2017 and 2018. DNA from these samples was sequenced, assembled, binned, error corrected, and annotated. Both samples were aligned to each other to determine their genetic similarity between the years, and both samples were aligned individually to a group of known cylindrospermopsin-producing toxin genes from a well-studied *Aphanizomenon* sp. 10E6 genome to determine their toxigenicity. The 2018 sample genome was completed and sent to the NCBI for confirmation and annotation. After alignment, the 2017 and 2018 samples were found to be 99.99% identical organisms, and both were found to contain toxin-producing genes on plasmids within their larger genome structures.

Key Words: Cyanobacterial harmful algal blooms, Genome assembly, annotation, alignment

Corresponding e-mail address: godbolet@oregonstate.edu

Exploring the toxigenicity and genetic similarity of Detroit Reservoir's recurring cyanobacterial bloom in 2017 and 2018

by
Tejas Godbole

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Microbiology
(Honors Scholar)

Presented May 30, 2019
Commencement June 2019

Honors Baccalaureate of Science in Microbiology project of Tejas Godbole presented on May 30, 2019.

APPROVED:

____
Theo Dreher, Mentor, representing Department of Microbiology

____
Ryan Mueller, Committee Member, representing Department of Microbiology

____
Thomas Sharpton, Committee Member, representing Department of Microbiology

____
Toni Doolen, Dean, Oregon State University Honors College

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

____
Tejas Godbole, Author

Introduction:

Cyanobacteria are one of the most abundant and important phyla of bacteria that exist on the planet. Through their photosynthetic and nitrogen fixing abilities, they are key ecological players in the cycling of nutrients such as oxygen, carbon, and nitrogen, and they exist in a wide range of ecological niches. Able to grow in both fresh and salt water, along with under a number of other environmental conditions, cyanobacteria are one of the most studied bacteria being researched today (1). One feature of cyanobacteria that has sparked large amounts of interest is their ability to produce secondary metabolites. Secondary metabolites are organic compounds produced by bacteria, fungi, and plants, that are made outside the central pathways for the growth and development of the organism producing them. One common example of secondary metabolites produced by cyanobacterial organisms are cyanotoxins (2). Many cyanobacteria produce various types of toxins as a part of their normal life cycle for defense and signaling. Studying the genome of cyanobacteria provides insight into the types of toxins an organism can produce, and the phylogenetic relationship between cyanobacteria using homogeneity and variability between genomes.

Another aspect of many cyanobacteria that is critical to understand is their boom-bust cycle of exponential growth followed by a period of overwintering population drop-off, known more commonly as cyanobacterial bloom cycles. These bloom events often create a closely grouped mass of organisms during summer or fall that cover a patch of water during their boom time period, before disappearing nearly completely during their bust period, perhaps over just a few days or weeks (3). While the factors that drive bloom cycles have been studied and confirmed to be changes in temperature, nutrient availability, and access to sunlight among others, the specific conditions that create individual cyanobacterial booms and busts is variable based on each organism and the exact environmental factors that exist (4).

Due to the prevalence of cyanobacteria in fresh water, along with the toxins that many of these cyanobacteria produce as secondary metabolites, blooms can pose a threat to other organisms that use the water where the bloom occurs as a source of clean water. While the bloom is occurring, the toxins are held intracellularly and can therefore be avoided by simply not consuming the organism mass as it grows in the water. However, once the bloom goes through a collapse, and the cells lyse, they release their secondary metabolites into the water. This phenomenon has led to the recognition and study of what are commonly referred to as cyanobacterial harmful algal blooms, or CyanoHABs. CyanoHABs have been implicated in the sickness and sometimes even death of domestic animals such as dogs and cattle, and have at times been tied to the toxification of drinking water sources used by humans (4).

One source of drinking water that is known to suffer from CyanoHABs, and that has experienced annual toxic blooms in recent years, is Detroit Reservoir in Oregon. Detroit Reservoir, fed by the North Santiam River, is a popular summer destination for water recreation, and provides drinking water to the City of Salem, Oregon. In the early summer of 2018, the City of Salem underwent a minor public health crisis as levels of toxins in the finished drinking water were found to be at a level of concern for Salem residents who were infants, pregnant women, or elderly. Based on the Oregon Health Authority, recommended maximum cylindrospermopsin toxin levels are 3 ng/ml for adults and 0.7 ng/ml for children, the 0.7 ng/ml threshold being exceeded in the finished drinking water (5). A water advisory based on the threshold of toxins present in the water was issued on May 3rd and lasted until

July 3rd, with only a brief four-day gap at the end of June, when it was mistakenly lifted. The toxin water advisory caused almost all residents to require drinking water shipped in and dispersed at stations by the National Guard, even if they were not in the groups determined to be at risk. The main toxin detected in the water was 7-epi-cylindrospermopsin, an epimer of cylindrospermopsin toxin that is highly soluble, and absorbed through the gastrointestinal tract before spreading to the liver, kidneys, and spleen. Microcystin toxin was also detected between the recommended levels of 0.3 ng/ml for children and 1.6 ng/ml for adults (5, 6). Symptoms of 7-epi-cylindrospermospin consumption include fever, headache, vomiting, bloody diarrhea, and kidney damage with a loss of water, electrolytes, and proteins (7). The source of these toxins is known to be a CyanoHAB that occurs each year in Detroit Reservoir in the month of May. This bloom usually occurs around Memorial Day and lasts only a week or two before undergoing a die-off. This would be expected to release the intracellular toxins into the reservoir, and potentially enter Salem's finished drinking water if not treated properly. The Dreher lab has sampled this bloom from the years 2015 to 2018, and study of these samples will offer insight through genomic analysis of its ability to produce the specific toxins that triggered the Salem health advisory.

The purpose of this research experiment is to further understand the CyanoHABs that contribute to the toxicity of the water in Detroit Reservoir, which can in years of unusually high toxin production or when adequate treatment is lacking, prove dangerous for animals and humans (8). For the research reported here, samples taken from the May bloom for the years 2017 and 2018 were studied in order to gain insight as to just how similar the composition of the bloom's organisms are from year to year, and to look for the presence of toxin producing genes that match with the types of toxins that were detected in the crude and finished drinking water in Salem. After sequencing and studying the samples of the Detroit Reservoir bloom taken on the 31st of May in 2017 and on the 8th of May in 2018, I hypothesize that they will be found to contain toxin producing genes, and that they will be found to be the same species, or perhaps even the same strain for both years.

Materials and Methods:

Sequencing Methodology

Two of the most powerful high throughput sequencing methods available to researchers at this time are Illumina and PacBio sequencing (9). Illumina sequencing is used to create a high number of relatively short, usually 150 nucleotide-long, reads. These reads have a high level of accuracy, but can be difficult to assemble into scaffolds of long contiguous sequences needed to represent the full genome of an organism. Repeat sequences prove especially difficult when it comes to using Illumina reads in attempting to build complete genomes, as irreparable breaks in the assembly could occur if the repeat sequence is longer than any of the corresponding Illumina reads, and do not overlap with the rest of the sequence. Because the assembly program would be unable to decide to which repeated region of the larger genome structure the reads should be assigned, the section wouldn’t be completed. While some of the issues that arise when repeats occur can be fixed using paired-end Illumina sequencing, many large repeat sequences still prove troublesome. Illumina paired-end reads are DNA sequencing reads containing the ends of fragments of known average length, which allows scaffolds of contiguous sequences to be produced more accurately when sequence gaps occur (9). This can potentially resolve repeats if one of the paired ends is located on the non-repeat region at the correct distance from its mate on the repeat sequence, but that isn’t always possible and many long repeat sequences still cause irreparable contig breaks.

The other common sequencing method available for use is PacBio sequencing. PacBio sequencing has a higher error rate in raw reads than Illumina sequencing, but also creates much longer reads, averaging around 10 to 15 kilobase pairs long. PacBio sequencing does this by using single molecule real time, or SMRT cell sequencing (10). A circular SMRTbell template is created with the DNA being sequenced and ligated to two primers in a circular conformation, allowing the DNA polymerase to repeatedly sequence the same DNA insert. This allows for the production of numerous polymerase subreads that can be aligned to create a long error-corrected circular consensus sequence (CCS), demonstrated in Figure 2.

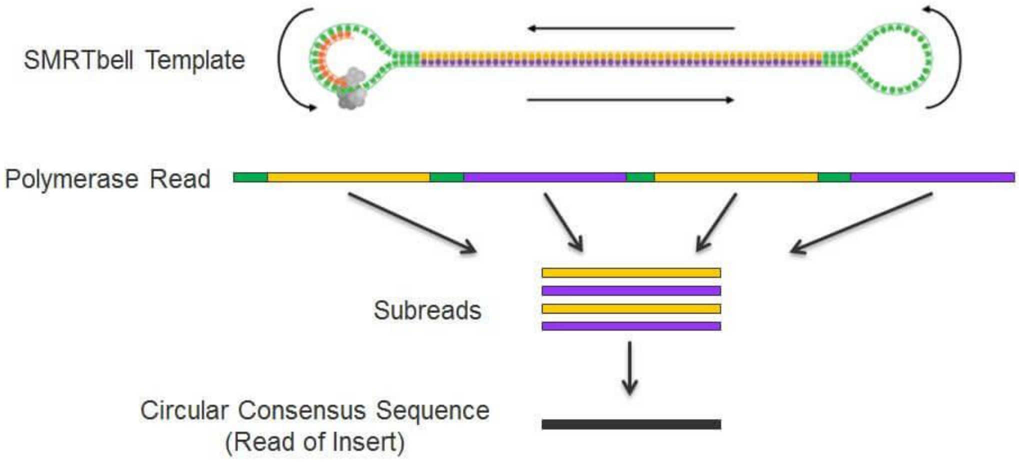


Figure 2. Derivation of Circular Consensus Sequences by PacBio sequencing using the sample reads and the added primers to circularize the path of the DNA polymerase.

These long sequence reads are better suited for assembling the entire genome of an organism because they avoid many of the problems brought up by repeat sequences. The increased length makes it more likely for a sequence to stretch across the repeat, or to at least provide an overlap that can be used to place the sequence in the correct repeat region in the genome. The higher error rate of PacBio raw sequencing reads as compared to Illumina sequencing is corrected by aligning multiple runs of the reads because by adding more CCS reads together, the likelihood of the same random error occurring in a majority of the reads is very unlikely (11). After sequencing is finished, those reads are assembled into larger contiguous sequences using programs such as PacBio HGAP (12). Based on what a researcher is looking to do and on the resources available to them, different projects call for the use of one, or the other, or both of these sequencing methods.

The HGAP program is specifically designed for assembling PacBio sequences into *de novo* assemblies. This occurs by first going through a preassembly process of generating long and highly accurate sequences that are mapped as consensus sequences with trimming based on quality to improve the accuracy of the sequence overlaps being created. Assembly is then used to create full genome closure by finding overlap consensus sequences and generating ultra-long trimmed preassembly reads, that anchor the longest repeat regions to the correct areas of the assembly. Finally, consensus polishing is used to reduce the presence of insertions, deletions, and base substitution errors in the draft assembly using a quality-aware consensus algorithm (12). After HGAP is finished, the goal is to have highly accurate consensus sequences in contigs representing a final assembly that will provide the full representation of the genome or genomes sampled and sequenced. This full representation is achieved by using binning and assessment programs to create a draft genome, and error-correcting that draft genome by aligning the PacBio reads to insure it is the best representation of the genome of the organism of interest.

The program MetaBAT bins contigs by aligning the reads of each sample separately to the assembled metagenome, and then calculating the distances between reads based on a number of factors including tetranucleotide frequency, abundance, and GC content (13). The metagenome takes into account all of the genomes present in an environmental sample, which is why sorting and binning is used to isolate the individual genome of the organism of interest (14). The MetaBAT binning technique has proven to be highly accurate in empirically deriving a distance value that is in turn used to potentially separate contigs into different genome bins, creating draft genomes. These draft genomes made up of contigs sorted by MetaBAT were analyzed by CheckM to assess the identity and quality of the draft genome as cyanobacterial. CheckM assesses the quality of genome bins by calculating the genome completeness and contamination using sets of ubiquitous and single-copy genes that are found within a phylogenetic lineage (15). When Illumina reads from the same sample are available, the Pilon program is used to further refine draft genomes through read alignment analysis to identify inconsistencies between the input genome and the evidence in the reads (16). This is a further step taken to correct errors that may have arisen in the creation of the draft genome.

Experimental Methods

For the 2017 bloom, an environmental sample was taken on the 31st of May 2017 from Heater Creek, an inlet of Detroit Reservoir on its Western end. This sample will from here on out be referred to as DET68. For the 2018 bloom, the sample was taken on the 8th of May 2018 from Blowout Creek, another inlet of Detroit Reservoir, on its Southern end. This 2018 sample will be referred to as DET69. The locations of the samples can be seen on the map and are indicated with labels and by the red arrow for DET68's Heater Creak sample location and the black arrow for DET69's Blowout Creek sample location in Figure 1.

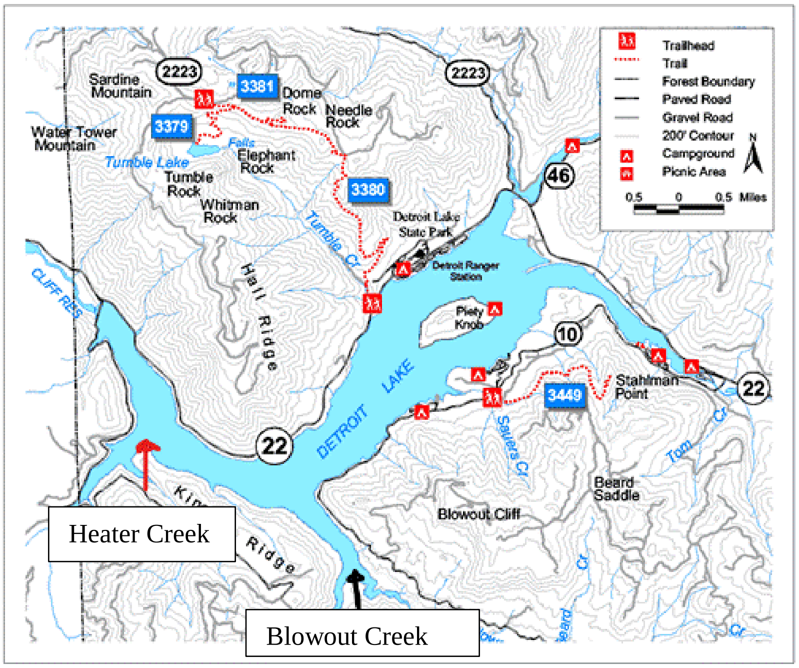


Figure 1. A map of Detroit Reservoir with the Heater Creek inlet sample location for DET68 indicated by a red arrow and the Blowout Creek inlet sample location for DET69 indicated by a black arrow.

Both DET68 and DET69 were environmental metagenomic samples taken from the blooms present in their respective years. DNA preparations were made from both of these samples, with the DET68 sample processed by filtration through a 10 µm mesh and yielding 27 ng/µl of DNA and the DET69 sample being filtered onto GF/C filters 1.2 µm in size and yielding 31 ng/µl of DNA. The samples were filtered to remove small bacterial cells. The DNA preparations were made by extracting with lysozyme, SDS, proteinase K and phenol/chloroform for lysis, followed by ethanol for precipitation (17). Final purification used the Qiagen Power Biofilm Kit #24000-50, which was used to selectively recover the desired DNA from the ethanol precipitate while avoiding the large amounts of mucilage polysaccharide that is also found in the precipitate. The DNA preparations were quantified using a Nanodrop spectrophotometer before submitting the preps for DNA sequencing. For this project PacBio Sequel chemistry was used as the main sequencing technique for its increased ability to yield a sequence that would be able to be assembled into a complete or nearly complete genome, but Illumina paired-end sequencing was also done as an additional resource to help order the PacBio contigs into a full genome where errors or sequence breaks occurred (9).

For the DET68 sample, assembly was run using PacBio Hierarchical Genome Assembly Process (HGAP), which has many parameter settings in order to best find the seeds of a contiguous sequence and to elongate those contigs to their maximum length with high levels of accuracy. Parameter seed coverage settings of c30, c60, c90, c120, and c150, with an expected genome size of 6 mega base pairs were each run. The most complete and effective results came from the run with the c60 setting (aggressive option off). The DET69 sample PacBio assembly was also run multiple times using different HGAP settings, and again the c60 setting was found to create the best combination of long contiguous sequences that represented the genome. After assembly was finished, CCS and Illumina reads were used to order the contiguous sequences into a full genome, MetaBAT and CheckM were used to bin the contigs, Pilon was used to error correct the contig bins, PROKKA and PGAP were used to annotate the genome, and Mauve was used to compare the DET68 sequences to the DET69 sequences.

To manually determine whether contigs could be further joined or circularized. The first step taken in determining the correct order of contigs was to map the CCS reads to the HGAP contigs using the Minimap2 program in order to find areas where overlaps existed between the borders of what were seen as contig breaks. Even if HGAP did not assemble the contigs together, perhaps because of an apparent sequencing break, finding areas where individual CCSs stretched between contigs provided direction for looking at the sequences on a closer level and manually ordering them. In addition, Illumina paired-end sequencing data was used to order the contigs because the presence of one pair on one contig and another the correct distance away on another contig would provide strong evidence that those two contigs were adjacent to each other in the genome. The manual circularization consisted of looking at the CCSs and Illumina reads and mapping them onto the known contigs using the program Minimap2, a versatile pairwise aligner for genomic and spliced nucleotide sequences (18). Using Minimap2 to see if the beginning and end of the longest contig sequences overlapped at a level to allow for the circularization of that genome, potential circular genomes or plasmids were detected. Both DET68 and DET69 underwent this close manual scrutiny and mapping with the help of CCS reads, Illumina reads, with sequence visualization done using the Geneious program.

After assembly of the DET68 and DET69 cyanobacterial genomes was completed, gene annotation was conducted. Annotation is a critical step in looking into the functions that various aspects of a genome might code for. Understanding the sequencing and finding the homology between sequences is valuable, but when hoping to answer questions of whether or not an organism has the genomic capability to produce a toxin, annotation is required. For both DET68 and DET69 genomes, Prokka was the program used for annotation. Prokka annotation identifies features of interest in a set of genomic DNA sequences from bacterial, archaeal, or viral genomes, and labels them with comparisons to known genes that can be examined by the researcher (19). In this case, Prokka was used to find the proteins or hypothetical proteins that the genome sequence encodes. In addition to Prokka annotation, the DET69 genome sequence was submitted to the National Center for Biotechnological Information (NCBI) and annotated using the Prokaryotic Genome Annotation Pipeline (PGAP). PGAP annotates bacterial chromosomes and plasmids through a multi-level process including the prediction of protein-coding genes and other functional genome units. It works by combining *ab initio* gene prediction algorithms, protein profile hidden Markov models, and complex domain architectures for functional annotation of proteins (20).

Finally, in order to test the relationship between the DET68 and DET69 samples taken approximately one year apart from Detroit Reservoir, the two genomes were aligned against each other to explore and visualize their similarities and differences. Whole genome alignment is used to inspect the relationships between genomes at the nucleotide level, which can then be interpreted to determine differences in the gene structure and protein coding capabilities of the genome. It is commonly used to determine phylogenetic relationships and to complete genomic annotations from one well-annotated and well-studied genome to another genome based on homologous regions. The program used here for the alignment between DET68 and DET69 was Mauve. Mauve is useful for aligning homologous regions amongst two genome sequences and tracking changes such as recombination, rearrangement, gene loss, duplication, and horizontal transfer that may have occurred between the two sequences. It does this by creating locally colinear blocks, which are regions where the genome is free from genome rearrangements (21). Just because two aligned genomes aren’t represented as one large colinear block does not mean that those two organisms can’t be closely related and have highly conserved functionality. Ideally, organisms that are almost identical and phylogenetically close will be represented by one colinear block representative of the entire length of the genome, but if not Mauve can also be used to closely examine why a block may be broken into multiple pieces.

Results and Discussion:

From the 2018 Blowout Creek DET69 sample, 31.0 ng/µl of DNA was extracted and sent away to the Center for Genomic Research and Biocomputing (CGRB) lab at Oregon State University for PacBio sequencing. The results of the run were assembled with HGAP resulting in 45 contiguous sequences accounting for a total of 6,634,259 base pairs. MetaBAT binning created a genome bin of four contiguous sequences, all with a closely related GC content of 37.8%. The total length of this suspected cyanobacterial genome was 6,085,511 base pairs long. When CheckM was run on this genome, it was found to be cyanobacterial with 99.78 percent completeness and only 0.22 percent contamination. Based on this strong evidence for the bin being representative of the cyanobacterial organism of interest in the DET69 sample, it was examined more closely in order to test the genome's ability to be circularized. Using CCS overlaps, Illumina paired-end reads, and by manually looking for overlaps between ends of contigs, the DET69 sample was circularized into a full genome of 5,839,262 base pairs with a plasmid of 200,795 base pairs. These complete assembled chromosome and plasmid are represented in Figures 3 and Figure 4, respectively. Two other cyanobacterial contigs of length 27,923 and 17,531 did not appear to circularize or align with the larger DET69 genome structure.

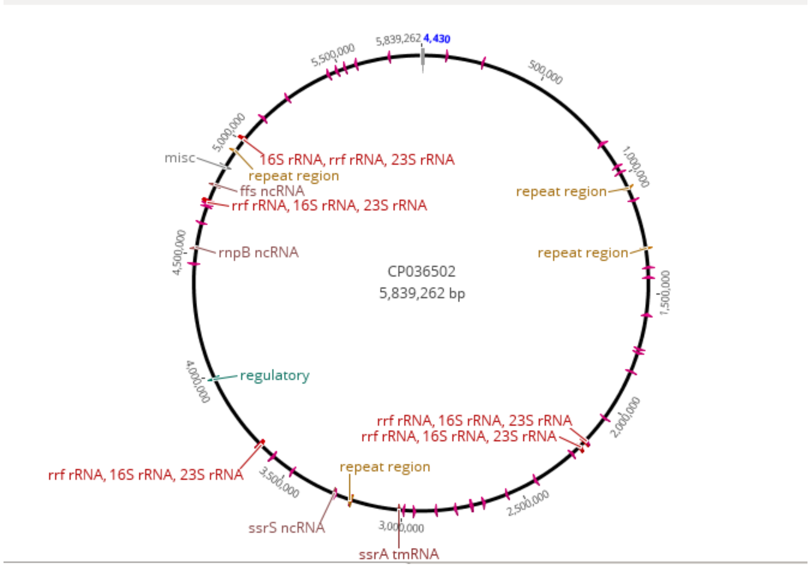


Figure 3. Full DET69 chromosome as annotated by NCBI PGAP

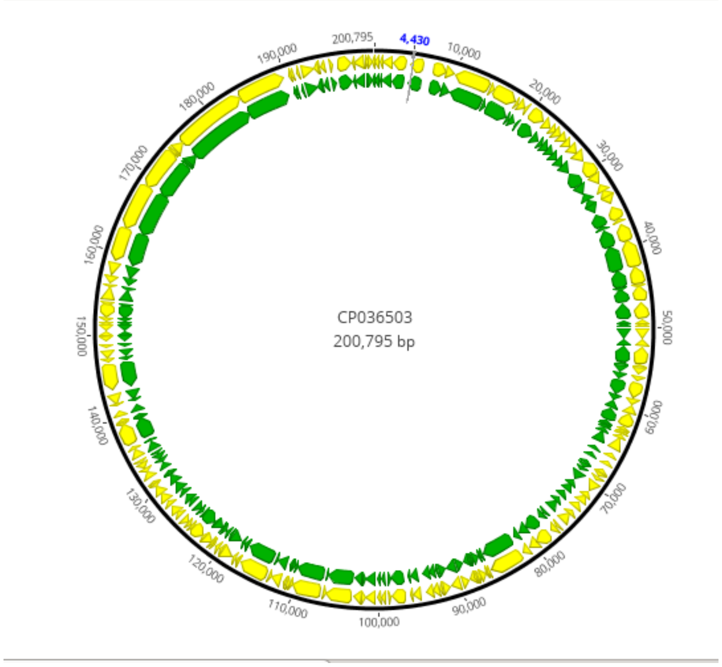


Figure 4. DET69 plasmid as annotated by NCBI PGAP

Based on the fact that DET69 could be assembled into a complete genome with a circular plasmid, with its MetaBAT and CheckM results implicating it as the cyanobacterial genome of interest, it was submitted to the NCBI for annotation through the PGAP algorithm.

The assembly of DET69 into a full genome was a critical step towards reaching the research goal of this project providing a complete reference to compare other genomes to. While this project only took the DET68 sample and aligned it to the DET69 reference genome, the more large-scale goal of completing a genome and having it annotated in a national database like the NCBI means that it is now available for public use past this single project. The fact that DET69 is from the same year as Salem's water crisis means the genome can provide insight into the toxin producing genes that led to the toxification of the drinking water.

From the Heater Creek sample of DET68, 27 ng/µl of DNA were extracted from the mesh filtered slurry. This PacBio sequencing and HGAP assembly yielded 31 contigs with a total length of 6,383,534 base pairs. These contigs were then sorted and binned using MetaBAT, which created a bin of five cyanobacterial contigs with a total length of 6,091,194 base pairs with a GC content of 37.8 percent. After running CheckM on this group of 5 contiguous sequences, the phylogenetic identity was found to be cyanobacterial with 99.67% completeness and only 1.78% contamination. The total length and the individual contig lengths in this particular bin of contigs matched well with the previously run DET69 sample, as seen in Table 1.

Sample	**DET69**	**DET68**
Longest chromosome length (bp)	5,839,262	5,839,290
Plasmid chromosome length (bp)	200,795	200,799
GC Percent	37.8	37.8

Table 1. Similarities in length between the longest circularized DET69 contig and the longest DET68 chromosome, and between the circular plasmid chromosome of DET69 and the second longest DET68 contig.

Now using the DET69 sample annotated by the NCBI and accepted as a complete genome with its additional plasmid genome, the DET68 sample was aligned to DET69 in order to determine the closeness of the relationship between the two. Figure 5 shows the result of using Mauve to align the longest contiguous sequence of DET68 to the circularized complete genome of DET69, Figure 6 shows the sequence view of the alignment, and Figure 7 shows a closer view at the similarity stats and number of sequence differences shown between the longest contig of DET68 and the main genome of DET69.

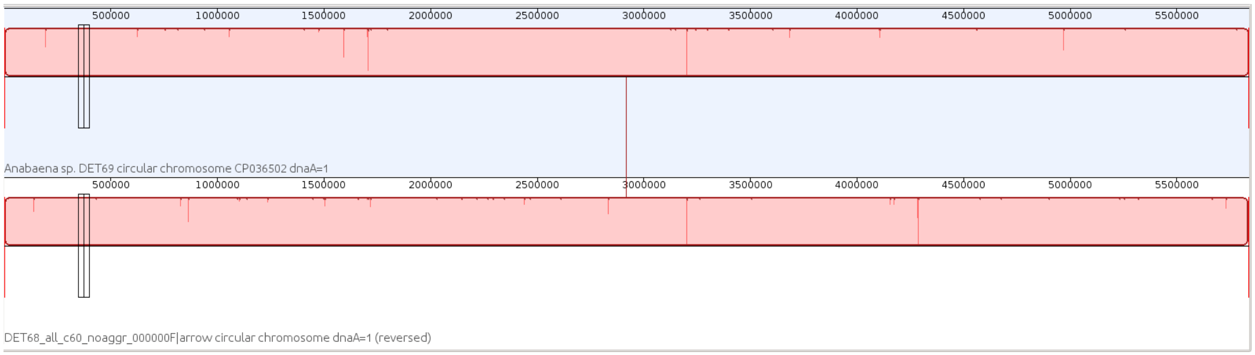
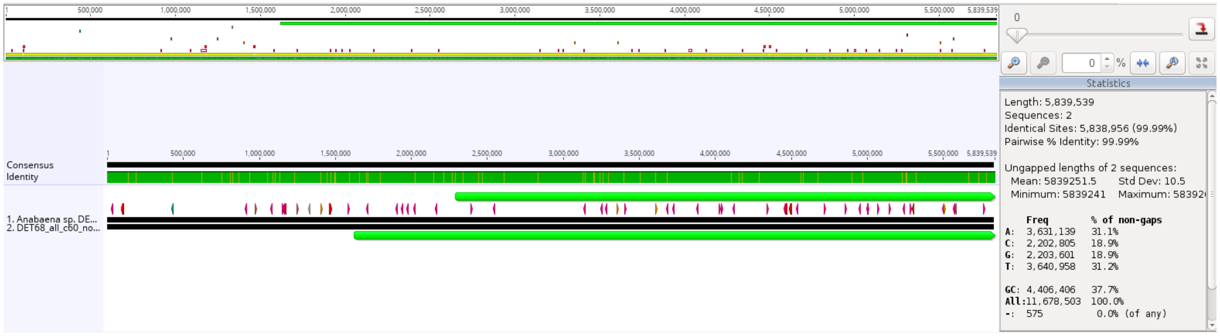
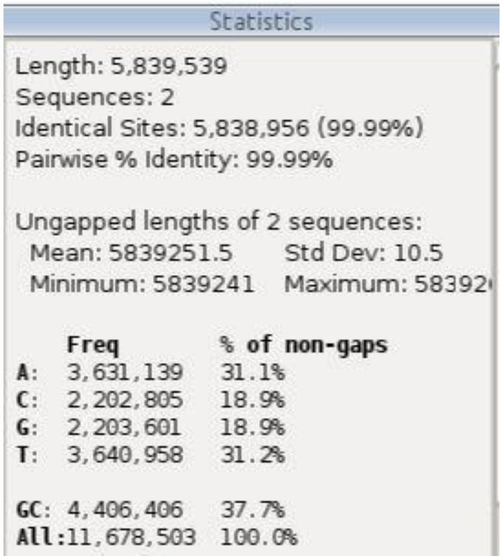

500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
5000000
5500000
Anabaena sp. DET69 circular chromosome CP036502 dnaA=1
DET68_all_c60_noaggr_000000F|arrow circular chromosome dnaA=1 (reversed)

Figure 5. Alignment of DET68 contig0 to DET69 contig CO036502 circular genome



```
Length: 5,839,539
Sequences: 2
Identical Sites: 5,838,956 (99.99%)
Pairwise % Identity: 99.99%

Ungapped lengths of 2 sequences:
  Mean: 5839251.5       Std Dev: 10.5
  Minimum: 5839241      Maximum: 58392

    Freq        % of non-gaps
A:  3,631,139   31.1%
C:  2,202,805   18.9%
G:  2,203,601   18.9%
T:  3,640,958   31.2%

GC: 4,406,406   37.7%
All:11,678,503  100.0%
-:  575           0.0% (of any)
```

Figure 6. The Mauve sequence view of the DET68 longest contig aligned to DET69 reference main genome.



```
Statistics
Length: 5,839,539
Sequences: 2
Identical Sites: 5,838,956 (99.99%)
Pairwise % Identity: 99.99%

Ungapped lengths of 2 sequences:
  Mean: 5839251.5      Std Dev: 10.5
  Minimum: 5839241     Maximum: 58392

    Freq        % of non-gaps
A:  3,631,139   31.1%
C:  2,202,805   18.9%
G:  2,203,601   18.9%
T:  3,640,958   31.2%

GC: 4,406,406   37.7%
All:11,678,503  100.0%
```

Figure 7. The statistics calculated for the alignment of the DET68 longest contig to the DET69 reference main genome using Mauve.

As can be seen, the entire genome of length of DET68 aligns to the annotated and confirmed NCBI genome of DET69 in one singular colinear block. In addition, there is a 99.99% representation and pairwise identity of the DET68 longest contig in the confirmed reference DET69 main genome.

The second longest contiguous sequence of DET68 was also aligned to the plasmid genome in DET69, as can be visualized in Figures 8, 9, and 10 through the annotated alignment, the sequence view, and the readable statistics respectively. All these images and statistics were produced using the Mauve program run as a plug-in through Geneious.

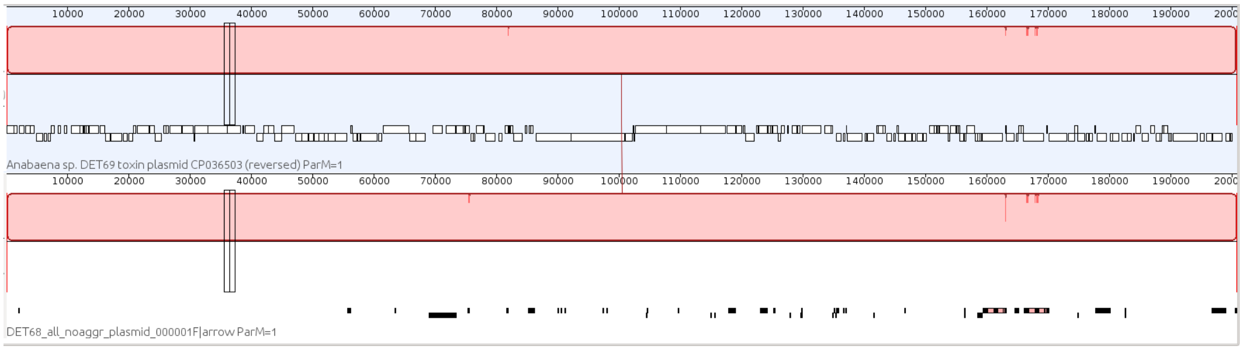
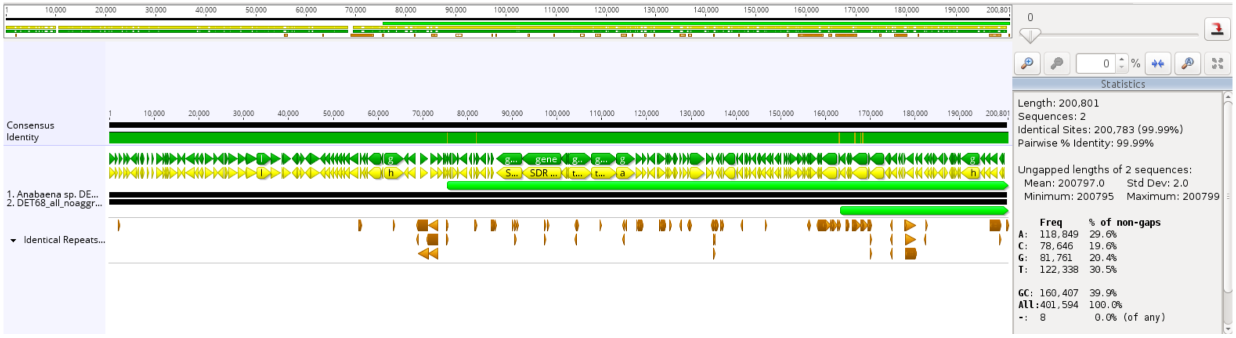
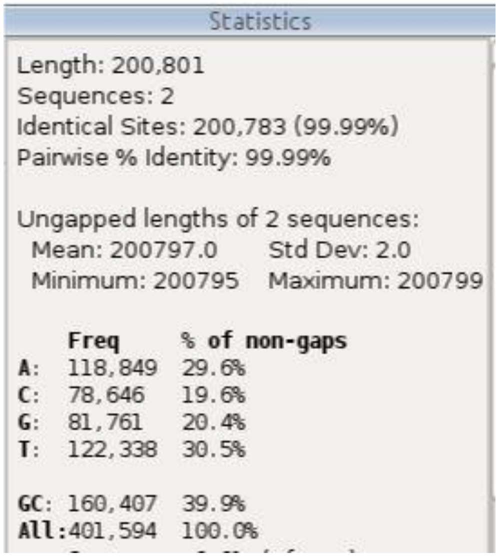


Figure 8. Alignment of DET68 contig1 to DET69 contig CO036503 circular plasmid genome



```
Length: 200,801
Sequences: 2
Identical Sites: 200,783 (99.99%)
Pairwise % Identity: 99.99%

Ungapped lengths of 2 sequences:
  Mean: 200797.0      Std Dev: 2.0
  Minimum: 200795     Maximum: 200799

     Freq     % of non-gaps
A:   118,849  29.6%
C:   78,646   19.6%
G:   81,761   20.4%
T:   122,338  30.5%

GC:  160,407  39.9%
All:401,594   100.0%
-:   8          0.0% (of any)
```

Figure 9. Mauve sequence alignment view of the DET68 plasmid contig aligned to DET69 plasmid genome.



Statistics

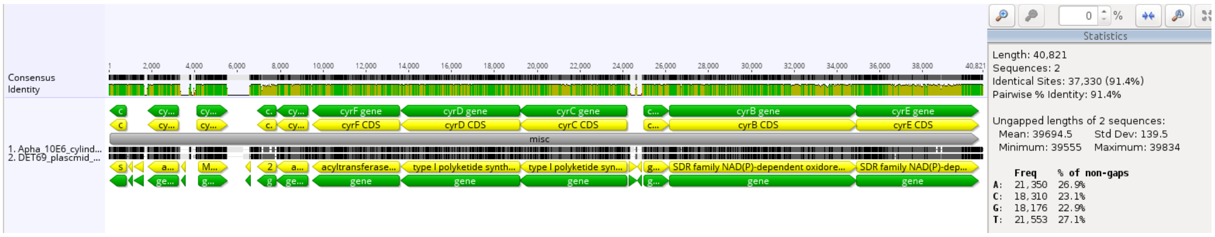
```
Length: 200,801
Sequences: 2
Identical Sites: 200,783 (99.99%)
Pairwise % Identity: 99.99%

Ungapped lengths of 2 sequences:
  Mean: 200797.0      Std Dev: 2.0
  Minimum: 200795     Maximum: 200799

     Freq     % of non-gaps
A:   118,849  29.6%
C:   78,646   19.6%
G:   81,761   20.4%
T:   122,338  30.5%

GC:  160,407  39.9%
All:401,594   100.0%
```

Figure 10. The statistics calculated for the alignment of the DET68 second longest contig to the DET69 reference plasmid genome using Mauve.

Again, the entirety of the known DET68 contiguous sequence aligns to the plasmid genome of DET69 as one colinear block 99.99% sequence alignment representation and pairwise identity.

The closeness of the alignment between the two genomes indicates the closeness of the relationship between these two sampled organisms. Based on this evidence, it is reasonable to conclude that the May bloom that occurred in Detroit Reservoir in the years 2017 and 2018, is in fact a single organism. While the alignment of samples for just two years cannot be used to extrapolate the relationship between the blooms that occur every year with much confidence, it is an indicator that if other years were sampled and compared to the DET69 reference genome, they could likely be the same organism blooming again and again. Although the surrounding organisms may change with the environmental conditions, different years can yield different amounts of bloom growth, and the expression of certain genes may change slightly based on how the organism reacts to that year's challenges, the genome of the organism from 2017 to 2018 appears to remain almost identical. While the genome appears identical, the level of toxins that enters the water depends at least in part on other factors that could contribute to the finished drinking water coming from Detroit Reservoir in the year 2018 being higher than that for the year 2017.

One of the strongest motivating factors of this project besides determining the similarity of the bloom year to year was the proposed ability of the Detroit Reservoir May blooms to produce toxins as secondary metabolites that could toxify the body of water. To this end, both the DET68 and DET69 genomes were searched for genes that were known to be linked to the production of toxins, starting with the known reference DET69 genome. While no such genes were found in the main genome assembled from the DET69 sample, toxin producing genes were found to be present in the circular plasmid sequence that is also included in the larger DET69 genome structure. This plasmid sequence was aligned to a database of well-studied *Aphanizomenon* cylindrospermopsin biosynthesis genes (22), as shown in Figure 11 and Figure 12.

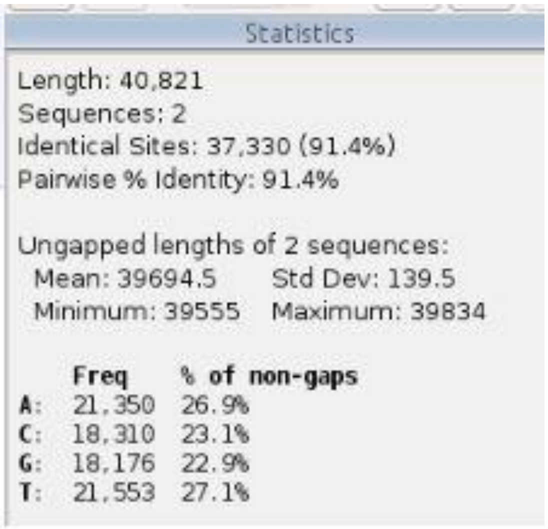


Length: 40,821
Sequences: 2
Identical Sites: 37,330 (91.4%)
Pairwise % Identity: 91.4%

Ungapped lengths of 2 sequences:
Mean: 39694.5 Std Dev: 139.5
Minimum: 39555 Maximum: 39834

	Freq	% of non-gaps
A:	21,350	26.9%
C:	18,310	23.1%
G:	18,176	22.9%
T:	21,553	27.1%

Figure 11. Sequence alignment view of the DET69 plasmid sequence to the well-studied cylindrospermopsin toxin producing genes from a sample of *Aphanizomenon* 10E6 using Mauve.


Statistics

Length: 40,821
Sequences: 2
Identical Sites: 37,330 (91.4%)
Pairwise % Identity: 91.4%

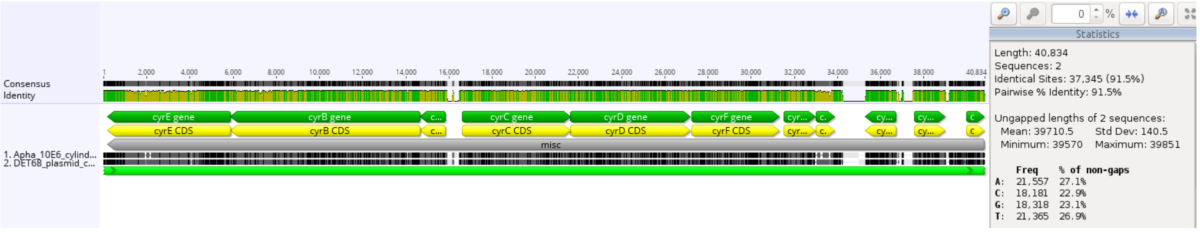
Ungapped lengths of 2 sequences:
Mean: 39694.5 Std Dev: 139.5
Minimum: 39555 Maximum: 39834

	Freq	% of non-gaps
A:	21,350	26.9%
C:	18,310	23.1%
G:	18,176	22.9%
T:	21,553	27.1%

Figure 12. Statistics from the sequence alignment view of the DET69 plasmid and the toxin-producing *Aphanizomenon* genes using Mauve.

As can be seen, the known *Aphanizomenon* 10E6 toxin-producing genes were found in the DET69 plasmid sequence based on the homology and pairwise identity score. While the 91.4% average pairwise identity is not nearly as good as the 99.99% similarity between DET69 and DET68, that level of similarity is not expected because the two samples that were aligned came from different organisms. What is able to be seen from this comparison is that the DET69 plasmid contains the Cyr genes necessary to biosynthesize cylindrospermopsin toxins. Represented in the DET69 genome are the genes CyrJ, CyrH, CyrK, CyrI, CyrG, CyrF, CyrD, CyrC, CyrA, CyrB, and CyrE. From the study of *Aphanizomenon*, it is known that each of these genes play a role in the production of a cylindrospermopsin toxin (22).

Since the ability of the DET69 plasmid to produce the toxins of concern was confirmed, and the similarity of the DET68 contig to the DET69 plasmid was also confirmed, an alignment was run between the DET68 contig and the same known toxin-producing *Aphanizomenon* genes. The results of this alignment are shown in Figure 13 and Figure 14.

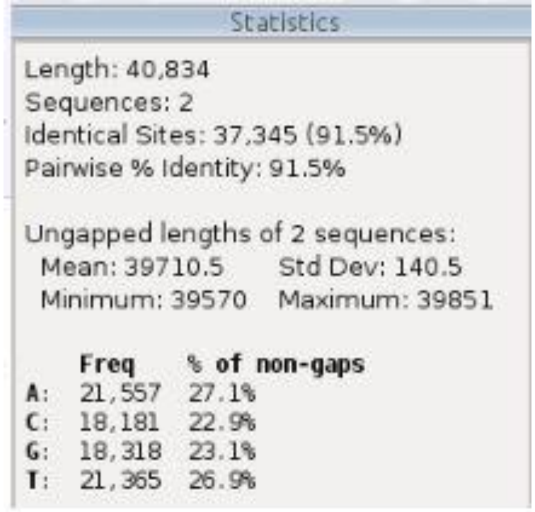


Length: 40,834
Sequences: 2
Identical Sites: 37,345 (91.5%)
Pairwise % Identity: 91.5%

Ungapped lengths of 2 sequences:
Mean: 39710.5 Std Dev: 140.5
Minimum: 39570 Maximum: 39851

```
    Freq    % of non-gaps
A:  21,557  27.1%
C:  18,181  22.9%
G:  18,318  23.1%
T:  21,365  26.9%
```

Figure 13. Sequence alignment view of the DET68 theorized plasmid sequence to the well-studied cylindro toxin producing genes from a sample of *Aphanizomenon* using Mauve.



Length: 40,834
Sequences: 2
Identical Sites: 37,345 (91.5%)
Pairwise % Identity: 91.5%

Ungapped lengths of 2 sequences:
Mean: 39710.5 Std Dev: 140.5
Minimum: 39570 Maximum: 39851

```
    Freq    % of non-gaps
A:  21,557  27.1%
C:  18,181  22.9%
G:  18,318  23.1%
T:  21,365  26.9%
```

Figure 14. Statistics from the sequence alignment view of the DET68 theorized plasmid and the toxin-producing *Aphanizomenon* genes using Mauve.

As expected based on how similar we now know the DET69 plasmid sequence and the proposed DET68 plasmid sequence to be, there is representation of the known toxin genes in the DET68 sequence. The same genes are present in both the DET68 and DET69 plasmid sequences, and both seem to be capable of producing cylindrospermopsin toxins, consistent with the 7-epi-cylindrospermopsin toxin that was detected as breaking the warning threshold and causing the issues with the Salem drinking water.

Conclusions:

The two-fold purpose of this study to gain an understanding of how the cyanobacterial bloom that occurs in May each summer in Detroit Reservoir plays a role in the toxification of the water and the potential toxification of the drinking water of Salem residence amongst others, and to understand just how similar the primary cyanobacterial organism was that made up the bloom in the years 2017 and 2018, were both successfully accomplished. While there is ample room for further study, much progress was made towards the understanding of the questions asked in this study through the genomic analysis conducted. The presence of toxin-producing genes that are known to produce the type of toxin that was detected in the water indicates that the organisms studied do in fact play a critical role in the potential toxification of drinking water that must be considered in the future. Additionally, the completed genome and plasmid of the DET69 sample can be used as a reference in years to come to test for the similarity of future blooms and infer their ability to produce toxins. While the confirmation of a future bloom's ability to produce harmful secondary metabolites and toxins may not indicate that they will in fact trigger such a degree of toxification that certain members of the public will no longer be safe in drinking the water, the similarity of a future bloom to DET69 could at least provide a warning for water treatment specialists that they could have a potential toxification event on their hands.

Conflict of Interest:

There are no potential conflicts of interest exist in this project.

Acknowledgements:

All research was done at Oregon State University in Dr. Theo Dreher's laboratory. Sequencing work was done by the Center for Genomic Research and Biocomputing at Oregon State University.

References:

1. Stanier RY, Bazire GC. 1977. Phototrophic prokaryotes: the cyanobacteria. Microbiology 31:225-74.
2. Carmichael WW. 1991. Cyanobacteria secondary metabolites – the cyanotoxins. Journal of Applied Bacteriology 72:445-459.
3. Oliver RL, Ganf GG. 2000. Freshwater Blooms. The Ecology of Cyanobacteria 149-194.
4. Paerl HW, Hall NS, Calandrino ES. 2011. Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change. Science of the total environment 409:1739-1745.
5. Oregon Health Authority. 2018. OHA files temporary rules for cyanotoxin testing by water suppliers. Oregon.gov. https://www.oregon.gov/oha/ERD/Pages/OHAFilesTemporaryCyanotoxinTestingRules.aspx.
6. Statesmen Journal. 2018. Updated July 3: Full test results of Salem's drinking water. https://www.statesmanjournal.com/story/news/2018/06/03/salem-drinking-water-advisory-test-results/667330002/
7. Poniedzialek B, Rzymski P, Kokocinski M. 2012. Cylindrospermopsin: Water-linked potential threat to human health in Europe. Environmental Toxicology and Pharmacology 34:651-660.
8. Carmichael WW. 2012. Health effects of toxin-producing cyanobacteria: "The CyanoHABs". Human and Ecological Risk Assessment 7:1393-1407.
9. Nie SJ, Liu YQ, Wang CC, Gao SW, Xu TT, Liu Q, Chang HL, Chen YB, Yan PC, Peng W, Zheng TQ, Xu JL, Li ZK. 2017. Assembly of an early-matured *japonica* (*Geng*) rice genome, suijing18, based on PacBio and Illumina sequencing. Scientific Data 4:170195.

10. Zautner AE, Goldschmidt AM, Thurmer A, Shuldes J, Bader O, Lugert R. 2015. SMRT sequencing of the Campylobacter coli BfR-CA-9557 genome sequence reveals unique methylation motifs. MBC Genomics 16:1087.
11. Fichot EB, Norman RS. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome 1:10.
12. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner, SW, Jonas K. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods 10:563-569.
13. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:1165.
14. Ascher J, Ceccherini MT, Pantani OL, Agneli A, Borgogni F, GuirriG, Ninnipieri P, Pietramellara G. 2009. Sequential extraction and genetic fingerprinting of a forest soil metagenome. Applied soil ecology 42:176-181.
15. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2014. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25: 1043-1055.
16. Walker BJ, Abeel T, Shea T, Priest M, Abouelleil A, Sakthikumar S, Coumo C, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 10:1371.
17. Atshan SS, Shamsudin MN, Lung LT, Ling KH, Sekawi Z, Pei CP, Rad EG. 2012. Improved method for the isolation of RNA from bacteria refactory to disruption, including *S. aureus* producing biofilm. Gene 494:219-224.
18. Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094-3100.
19. Seemann T. 2014. *Prokka: rapid prokaryotic genome annotation.* Bioinformatics 30:2068-2069.
20. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research 44:6614-6624.
21. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genomic Research 14:1394-1403.
22. Stuken A, Jakobsenn KS. 2010. The cylindrospermopsin gene cluster of *Aphanizomenon* sp. Straing 10E6: organization and recombination. Microbiology 156: 2438-2451.