AN ABSTRACT OF THE THESIS OF

ANDREW D. SCHWARTZ FOR THE DEGREE OF MASTER OF SCIENCE IN MICROBIOLOGY PRESENTED ON AUGUST 25, 2011
TITLE: BIOINFORMATICS APPLIED TO FRESHWATER CYANOBACTERIAL BLOOMS AND CYANOPHAGE

Abstract approved:

Theo W. Dreher                       Frederick S. Colwell

Freshwater cyanobacterial blooms are a nuisance and health threat in the Pacific Northwest. The accepted methods of characterizing these blooms by microscopic cell counts cannot differentiate between toxic and non-toxic strains of the cyanobacterium *Microcystis*. Also, there is limited understanding of freshwater cyanophage that may control bloom dynamics. In order to better understand the cyanobacterial and cyanophage populations of the Klamath River, two studies were conducted. Methods in the studies were largely composed of bioinformatic techniques.

First, using clone libraries to genetically track the *Microcystis* blooms, we found a distinct separation in bloom populations according to whether the toxin microcystin could be produced. The genotypic differences in two bloom cycles were parsed into a graphical representation, useful for describing many closely related organisms.

Further investigation of the bloom community led to the isolation of the novel, freshwater *Synechococcus*-infecting myophage, S-CRM01. The phage is very similar in gene content to marine cyanomyophage, but differs in synten and nucleotide composition. The bioinformatic techniques used that were either developed or customized to fit our data.

Our findings indicate that the use of bioinformatic techniques will greatly increase the resolution of population and genomic studies in freshwater environments.
This is apparent in the population differentiation seen throughout the bloom season in the Copco Reservoir as well as the comparative genomics of phage S-CRM01. The results of the research support the use of bioinformatics to acquire and interpret genetic data for cyanobacteria and cyanophage.
BIOINFORMATICS APPLIED TO FRESHWATER CYANOBACTERIAL BLOOMS AND CYANOPHAGE

by
Andrew D. Schwartz

A THESIS
Submitted to
Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 25, 2011
Commencement June 2012
Master of Science thesis of Andrew D. Schwartz presented on August 25, 2011

APPROVED:

______________________________
Co-Major Professor, representing Microbiology

______________________________
Co-Major Professor, representing Microbiology

______________________________
Chair of the Department of Microbiology

______________________________
Dean of the Graduate School

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

______________________________
Andrew D. Schwartz, Author
ACKNOWLEDGEMENTS

I would like to thank all the members of my committee for their time, effort and support of my Thesis.

My advisors: Dr. Theo Dreher has given me a priceless knowledge of science and research. Dr. Rick Colwell has always given me his time and knowledge.

The Dreher Lab has been a welcoming place for discussions.

Family, friends important for both discussions of science or otherwise and general support.

Especially Brenda Bennett, Tony Bertagnolli, Jonathan Halama and Matt Stinson.
CONTRIBUTION OF AUTHORS

CHAPTER 2: Dr. Connie Bozarth began the project before I arrived at OSU. She organized and designed the initial collection and production of the clone libraries. Her tireless work resulted in the numerous sequences that we were able to analyze and interpret. The project would not have been completed without her.

Jon Shepardson, an undergraduate, presented us with excellent data from the qPCR machine. His contribution allowed us to see the changing toxigenicity of the bloom through quantitative PCR of the mcy gene.

Dr. Frederick Colwell was involved in many aspects of discussion and interpretation of our results. His knowledge of ecology and determination of species in the environment was invaluable.

Dr. Theo Dreher directed the project and his ideas led to the development of new genotype maps used within the paper. Without his interpretation and perseverance this paper would not be nearly as interesting.

When I arrived at OSU many of the sequences had already been collected. I was assigned the task of organizing them on the computer. I created a pipeline for sequence analysis that allowed us to observe the numerous sequences and present them in a readable way. My use of bioinformatics was instrumental in creating a complete and interesting narrative from our research and in preparing a publishable product.

CHAPTER 3: Dr. Theo Dreher captained the expansive phage project. His direction and focus on key genetic aspects such as tRNAs produced a comprehensive view of the sequenced freshwater phage.

Nathan Brown, an undergraduate at the time began collecting water samples from the Klamath River and looking for phage. His initial discovery of a freshwater phage made the project possible.

Dr. Connie Bozarth sequenced the phage to completion. While initial 454 sequencing did not circularize the genome, Dr. Bozarth’s nested PCR approach was able
to finish S-CRM01. Connie’s initial idea and work with mass spectrometry led to a number of interesting and comprehensive results.

Erin Riscoe, an undergraduate, used PCR to determine other locations on the Klamath River where the phage occurs. Her results gave us a geographic depiction of freshwater cyanophage in the Klamath environment.

Dr. J. Cameron Thrash introduced us to the HAL phylogenomic pipeline. He also helped out with the interpretation of those results and contributed figures to the manuscript.

Dr. Samuel E. Bennett, Shin-Cheng Tzeng and Dr. Claudia Maier all worked with our lab to obtain excellent mass spectrometry data.

My contribution to this project included annotation of the phage genome through a multi step process and comparison of S-CRM01 genes to other freshwater phage. I was responsible for the phage sequence and confirming it was current. In this way, I was able to help others to discover tRNAs, perform mass spectrometry, HAL phylogenomics or answer various questions about the genetic makeup of the phage. I helped interpret the phylogenomics of the phage and made multiple comparisons to other cyanomyophage that had been sequenced, allowing this study to be published.
1. Introduction

Microcystis Genetics........................................................................................................... 4
Microcystin and toxigenicity.............................................................................................. 7
Phage of Cyanobacteria..................................................................................................... 8
Experimental Goals........................................................................................................... 13
References.......................................................................................................................... 14

2. Population Turnover in a *Microcystis* Bloom Results in Predominantly Non-toxigenic Variants Late in the Season.................................................................................................................. 18

Abstract................................................................................................................................ 19
Introduction............................................................................................................................ 20
Materials and Methods....................................................................................................... 21
Results and Discussion........................................................................................................ 24
Acknowledgments................................................................................................................ 39
References.............................................................................................................................. 40

3. A Freshwater Cyanophage Whose Genome Indicates Close Relationships to Photosynthetic Marine Cyanomyophage................................................................................................................. 48

Abstract................................................................................................................................ 49
Introduction............................................................................................................................ 50
Results and discussion.......................................................................................................... 51
Experimental procedures..................................................................................................... 78
Acknowledgements.............................................................................................................. 85
References.............................................................................................................................. 86

4. General Conclusion......................................................................................................... 94

Microcystis............................................................................................................................ 95
Cyanophage.......................................................................................................................... 97
References.............................................................................................................................. 99
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>The Klamath River system of Southern Oregon and Northern California.</td>
<td>3</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>The structure of microcystin-LR.</td>
<td>7</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Rarefaction curves estimating the population sampling of clone libraries.</td>
<td>24</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Progress of the <em>Microcystis</em> bloom in Copco Reservoir in 2007.</td>
<td>25</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Loss of the <em>mcyb</em> gene during the 2007 bloom season in Copco Reservoir.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Relationships between ITS genotypes collected from a 2007 <em>Microcystis</em> bloom in Copco Reservoir.</td>
<td>28</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Relationships between ITS genotypes at individual time points sampled from Copco Reservoir.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Changes in ITS and <em>cpeba</em> allele populations during the 2007 bloom season in Copco Reservoir.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>Relationships between <em>cpeba</em> genotypes at individual time points sampled from Copco Reservoir.</td>
<td>34</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>Relationships between <em>cpeba</em> genotypes collected from a 2007 <em>Microcystis</em> bloom in Copco Reservoir.</td>
<td>35</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td><em>Microcystis</em> genotypes in Copco Reservoir.</td>
<td>36</td>
</tr>
</tbody>
</table>
LIST OF FIGURES (Continued)

Figure 3.1 Distribution of S-CRM01 in the Klamath River valley, 2009....................... 51

Figure 3.2. Electron micrograph of S-CRM01 phage particles negatively stained with phosphotungstic acid.................................................................................. 52

Figure 3.3. Circular genetic map showing the gene organization of phage S-CRM01. .......................................................................................................................... 54

Figure 3.4 Extensive synteny between S-CRM01 and marine cyanomyophage.............. 60

Figure 3.5 Phylogenetic relationship of S-CRM01 to other myophage......................... 62

Figure 3.6. Late promoter consensus sequence ............................................................ 68

Figure 3.7. Phylogenetic tree showing relationship of S-CRM01 gp20 gene to Other gp20 sequences from freshwater or marine phage sources................................. 73

Figure 3.8. Phylogenetic tree showing relationship of S-CRM01 psba gene to Other psba sequences from freshwater or marine phage sources............................... 74

Figure 3.9. Klamath River watershed isolate LC16 is a member of the Cyanobium Gracile cluster of Synechococcus sp............................................................................. 80
LIST OF TABLES

Table 1.1 Fully sequenced genomes of the *Myoviridae* family infecting cyanobacteria (and T4) ................................................................. 12

Table 2.1 Diversity of clone libraries made from each sample ........................................ 30

Table 3.1. Predicted protein-coding genes of S-CRM01 ........................................... 55

Table 3.2. tRNA genes in the S-CRM01 genome ...................................................... 76
1. Introduction

Cyanobacteria are abundant in freshwater lakes. Photoautotrophs that are often diazotrophic, they are productive organisms during summer. In lakes, cyanobacteria can be somewhat dispersed but during the day, gas vacuoles provide buoyancy resulting in a highly visible surface scum. An increased prevalence of blooms has been linked to anthropogenic nutrient loading and a changing global climate (Paerl and Huisman, 2009).

Cyanobacterial blooms limit the use of recreational and drinking water and are capable of producing a number of noxious, toxic chemicals and anoxia in cases of rapid cell death. Large-scale events have raised awareness and caused alarm; a notable example is that of an *Anabaena* bloom in 1991: the Barwon-Darling River of Australia suffered a 1000 km anatoxin-a producing bloom that led to the deaths of numerous livestock (Bowling and Baker, 1996).

Water supplies contaminated by toxic cyanobacterial blooms are potentially deadly. In 1996 in Brazil, 47 deaths resulted after 117 people were intoxicated with microcystin-contaminated water used for dialysis (Jochimsen *et al*., 1998). Action was taken at a reservoir in Armidale, Australia in 1981 where a monitored recurring *Microcystis* bloom was present (Falconer *et al*., 1983). Addition of copper sulfate alleviated the reservoir of *Microcystis*, but released the toxin microcystin into the water supply, affecting liver function of those who consumed it. The release of toxins during the destruction of a bloom adds to the difficulty of bloom control.

In the Pacific Northwest, annual blooms of various cyanobacteria require lake managers to make decisions on posting lakes and reservoirs (Kann, 2006). The decision to post a health advisory is often based on morphological identification and cell counts. Although microscopy is the currently accepted measure, it lacks precision at species-level identification and because of the various morphotypes of cyanobacteria, toxigenicity cannot be determined based solely on morphology. Understanding the bloom-forming
population genetically would give insight into toxigenicity and possibly predictability of upcoming blooms. An integrated approach of microscopic methods with identification based on genetic characteristics would improve the understanding of freshwater cyanobacterial communities.

Genetic methods have been successful in obtaining reproducible phylogenies of cyanobacteria. *Anabaena* is morphologically distinct from *Aphanizomenon*, but they are genetically similar and do not segregate into clades based on several DNA targets (Lyra *et al.*, 2001). A study of *Anabaena* and *Aphanizomenon* in European lakes based on 16S, ITS and *rbcLX* (RubisCO) genes resulted in a monophyletic grouping of the two genera, with 97.5% 16S rDNA sequence similarity (Gugger *et al.*, 2002). Genetic differentiation has great possibility in identifying true relationships as evidenced by these studies.

My research analyzed *Microcystis* within a single freshwater system to thoroughly assess the genetic diversity as part of a larger study on *Microcystis* population dynamics. *Microcystis* has often been identified in the Pacific Northwest region as an annually recurring bloom and surface scum. It has been identified in a number of lakes and has consistently been found to produce toxin for at least a portion of the season. The choice of focus on *Microcystis* was based on its regional prevalence as well as it being a consistent, toxigenic health threat.

Viruses are the most numerous biological entities on the planet and greatly impact the population of bacteria in aquatic environments (Suttle, 2005). Viruses infecting cyanobacteria can be contributors to bloom decline and are almost constantly threatening infection and mortality. Viruses infect hosts at varying rates, affecting the microbial population based on bacterial growth rates and abundance (Suttle, 2007). Particular host specificities of phage mean an increased infection by a particular phage can impact only a portion of the population present. The effects of phage through mortality, host specificity and recombination can impact ecology, population structure and gene content.

The genetic reservoir of viruses should be studied to better understand their impact on aquatic environments and cyanobacterial populations. In most cyanobacterial blooms the role of phage and the degree to which they influence mortality remain
unknown. In the Klamath River system where cyanobacterial blooms occur annually cyanophage have not been studied. In general, freshwater phage have been overlooked and only one freshwater cyanophage has been fully sequenced (Microcystis phage Ma-LMM01)(Yoshida et al., 2007). The genetic diversity contained within phage is currently unaccounted for in ecological studies and could factor into the population dynamics of freshwater cyanobacteria.

*Synechococcus* is a widespread, well-studied cyanobacterium found in both marine and freshwaters; it has become a model host for cyanophage research. Of the marine myophage genomes sequenced, 11 of a total 17 infect *Synechococcus*. A freshwater *Synechococcus* infecting phage was isolated from the Klamath River to determine the genetic characteristics of freshwater cyanophage. Ultimately, the study focused on the comparison of a well-studied marine model cyanophage to its freshwater counterpart.

**Figure 1.1 The Klamath River system of Southern Oregon and Northern California.** The river runs from Upper Klamath Lake through Copco Reservoir and out at the Pacific Ocean. Various blooms have are reported annually, mostly in reservoirs on the river.

In order to more thoroughly characterize the dynamics of recurrent blooms the key identifying genes from *Microcystis* collected over a bloom season in the Copco Reservoir on the Klamath River were analyzed. Two genetic loci were surveyed and analysis found two separate, distinct population structures occurring at different times. In accompaniment, the genome sequence of a phage infecting *Synechococcus*, a model cyanobacterium was examined. The phage genome revealed a portion of the genetic diversity within freshwater cyanobacterial viruses that paralleled genomic sequences of
marine phage. Via these two studies we now have a better grasp on the freshwater cyanobacterial and phage population of the Klamath River.

**Microcystis Genetics**

Studies in Australia, Europe, Japan and the United States have focused on genetic identification and diversity of Microcystis in order to understand its toxigenic characteristics (Otsuka *et al.*, 2001; Tanabe *et al.*, 2007; van Apeldoorn *et al.*, 2007; Paerl and Huisman, 2009; Sabart *et al.*, 2009). *Microcystis* is known for its production of the hepatotoxin microcystin and research has often focused on the microcystin synthetase gene cluster (*mcy*) and the toxin. *Microcystis* is typically composed of diverse populations of multiple strains and morphologies some of which produce microcystin (*mcy*+).

**Genetic Techniques and Targets**

Sequencing of key genes is essential in identifying cyanobacteria in addition to determining phylogenetic relationships and even toxigenicity. Multiple genetic targets have been analyzed in *Microcystis* with the focus of finding an appropriate and informative genetic locus. The 16S rRNA gene sequence, conserved among bacteria, is unable to distinguish sub-populations of *Microcystis* (Neilan *et al.*, 1995). However, the internal transcribed spacer (ITS) of the 16S-23S rRNA operon and the phycocyanin intergenic spacer (*cpcBA*) both provide greater genetic resolution (Neilan *et al.*, 1995; Boyer *et al.*, 2001; Janse *et al.*, 2003). The use of these genetic loci has resulted in a greater understanding of *Microcystis* genetics.

Microscopy is unable to fully differentiate *Microcystis*, as organisms can be morphologically identical but genetically different (Baker *et al.*, 2002). To compound the problem, *Microcystis* strains that appear morphologically different are often not
genetically distinct (Otsuka et al., 2001). Furthermore different genetic targets produced conflicting phylogenetic trees when sequence composition was analyzed. Tillet et al. (2001) analysed 18 toxic and 17 non-toxic globally cultured strains in an effort to establish differentiated genetic markers. The comparison of two loci, 16S and 16S-23S ITS, resulted in erratically distributed toxic and non-toxic strain phylogenies. The sequence results were not phylogenetically separated by any measured parameter including geography, morphology and presence of mcyA (Tillett et al., 2001).

Later studies used more isolates and obtained more sequences through more advanced technologies. For example, denaturing gradient gel electrophoresis (DGGE), which separates amplified genetic targets based on size and sequence for subsequent sequencing, provided a large number of 100 bp or 200 bp ITS sequences (Sandaa et al., 2010). One DGGE study of 107 colonies from Europe and Morocco classified Microcystis sub-populations and placed toxic and non-toxic strains in different classes (Janse et al., 2004). Another technique, Real Time PCR (qPCR), amplifies a targeted genetic locus and reports the relative number of copies present. Other studies have used clone libraries to survey bacterial diversity. Clone libraries are created in a multi-step process; PCR amplification products are inserted into plasmids and transformed into E. coli cells, then sequenced from known priming sites in the plasmid. Clone libraries can produce an in depth survey of an environmental population and can be obtained with 700 – 900 bp lengths through Sanger sequencing. Use of clone libraries differentiated populations of freshwater and brackish Microcystis using the cpcBA locus (Moisander et al., 2009). The clone libraries produced 575bp sequences showing limited connectivity between the two populations.

A shift in focus from global to regional populations resulted in a greater understanding of population structures and dynamics in the environment. One observation is that of population shifts over the course of a bloom season. A team studying three Dutch lakes observed a few genotypes that dominate the Microcystis population were replaced by different genotypes later in the season, an observation of the succession of the Microcystis population (Kardinaal et al., 2007). Similar conclusions
were reached in a French reservoir through survey of the ITS genotype of *Microcystis* (Briand et al., 2008). Changes occurred in the population resulting in a shift in genotype as observed through sequencing of the ITS locus. Furthermore, this change was correlated with decreased proportion of *mcyB*+, pointing to a functional difference occurring through population dynamics.

Investigation of lake populations has led to an understanding of *Microcystis* population structures that must be studied in greater detail. To investigate the global diversity of *Microcystis*, Tanabe et al. (2007) sequenced multiple “house-keeping” genes from 164 isolates. The results revealed a high level of diversity and the “clonal” population structure of *Microcystis*. The “clonal” structure, meaning most populations arise from a single or predominant clone, was hypothesized. Later studies such as Kaardinaal et al. (2007) and Briand et al. (2008) observed lakes in which the most common genotypes could change throughout the season. Defining and assessing these population structures is an important tool in understanding the genetic makeup of *Microcystis* blooms and linking genotype to toxigenicity.

The most prevalent and important toxin produced by *Microcystis* is the hepatotoxin microcystin. Mass spectrometry, HPLC, ELISA and protein phosphatase inhibition assay (PPIA) have all been used to determine the presence of the toxin in cultures or samples (Tillett et al., 2001; Baker et al., 2002; van Apeldoorn et al., 2007). However, these tests do not indicate the origin of the toxin. In order to further understand the toxigenicity in *Microcystis* research on the genetics of the microcystin operon is needed.

**Microcystin genes**

The *mcy* locus directs the synthesis of the cyclic heptapeptide hepatotoxin microcystin. The chromosomally located genes are 55 kbp in length (Fig. 1.1), and are found in multiple genera of freshwater cyanobacteria (Kaebernick et al., 2002; van Apeldoorn et al., 2007). Microcystin is expressed constitutively in most situations but is affected by light and nutrient availability (Kaebernick et al., 2000). The presence of the
*mcy* cluster is directly connected with toxigenic *Microcystis*, while the absence of the operon indicates non-toxigenicity (Nishizawa *et al.*, 2007).

![Figure 1.2 The structure of microcystin-LR.](image)

*Figure 1.2 The structure of microcystin-LR.* Where X and Z are variable L-amino acids. In most common variant (MC-LR) X is leucine (L) and Z is arginine (R). The synthetase gene cluster is shown above for *Microcystis*. Flanking regions *dnaN* (left) and *uma1* (right) are found in both toxigenic and non-toxigenic strains of *Microcystis* (Tooming-Klunderud *et al.*, 2008; Pearson *et al.*, 2010).

The *mcy* operon associated with *Microcystis* is not always present. It has been hypothesized that the *mcy* cluster is prone to a high level of recombination, evidenced by entire missing regions of the *mcy* cluster with flanking gene regions that are invariably present (Tanabe *et al.*, 2004; Tooming-Klunderud *et al.*, 2008). Non-toxic strains rarely possess *mcy* genes and in those rare cases only possess fragments of the *mcy* operon. *Mcy* genes have thus become measures and determinants of toxigenic ability.

**Microcystin and toxigenicity**

Freshwater studies of *Microcystis* have analyzed *mcy* frequently, particularly in quantification of toxigenic cells (Kurmayer and Kutzenberger, 2003). Quantifying DNA and relative number of gene copies is possible using real-time PCR (qPCR). This method of quantification has been used to determine the relative amount of toxin genes and the ratio of a toxin gene to housekeeping gene (Kurmayer and Kutzenberger, 2003).
It can be difficult to determine genotypes based on housekeeping genes to determine toxigenicity or other traits, but a few studies have been successful. Tanabe et al. (2007) sequenced multiple loci of seven housekeeping genes (ftsZ, glnA, gltX, gyrB, pgi, recA and tpi) from 164 different cultures, dividing the phylogeny of Microcystis into toxic and non-toxic classes. This study also found success differentiating cultures obtained from similar regions, inferring that Microcystis may exhibit geographical distribution. Using the ITS locus, Janse et al. (2004) found 107 colonies separated phylogenetically correlating with toxigenic ability. Further work is still required to elucidate the link between genotype and toxigenicity. A number of studies have established populations and their proportion of mcy+ genotypes or amount of toxin production (Kardinaal et al., 2007; Briand et al., 2008). However, to date only mcy+ can be used to verify a toxigenic organism. Any single mcy gene is not a good determinant of toxigenicity as it is a presence / absence assay and they have little variability in sequence (Tooming-Klunderud et al., 2008). It may not be possible to find the ITS or cpcBA sequences equating to toxigenicity, as recombination may have blurred the connection. A more nuanced view of Microcystis populations is possible through season-long sampling and observation of multiple genetic loci. Greater success likely lies in regional studies that attempt to discern toxigenicity in geographically co-located sub-populations.

Phage of Cyanobacteria

Viruses are the most abundant biological entities in aquatic environments. Those infecting cyanobacteria are referred to as cyanophage and are both abundant and widespread in marine and freshwaters. Three families of double-stranded DNA phage; Myoviridae, Siphoviridae, and Podoviridae are all known to have members that infect cyanobacteria. The Myoviridae are T4-like, Siphoviridae are λ-like and Podoviridae are T7-like (Suttle, 2002; Sullivan et al., 2009). The myo- and sipho- viruses tend to be lytic, while podoviruses are often lysogenic, with their genome integrated into the host genome.
Bioinformatic research has progressed our knowledge of cyanophage through genome sequencing and multiple comparison tools. Infection by phage often results in lysis or death. In the case of increasingly prevalent bloom-forming cyanobacteria it is unknown to what degree phage impact bloom decline through lysis (Suttle, 2002; Paerl and Huisman, 2009). Increasing cyanophage- and virus-like particle numbers during bloom decline implicate phage in bloom ecology (Manage et al., 1999; Manage et al., 2001; Deng and Hayes, 2008; Rodriguez-Brito et al., 2010). The decline of bacterial populations by phage killing is often described by the “Kill-the-winner” theory in which viruses rapidly kill and decrease the population of the most numerous bacteria (Suttle, 2007).

More complex population dynamics can occur in which the total bacterial species population does not fluctuate much; instead, genotypes not affected by a particular phage rise in abundance. A study of four aquatic environments of varying salinities found a stable level of species but a shuffling of the most successful members (Rodriguez-Brito et al., 2010). The shuffling of bacterial strains was observed through a metagenomic survey of the recA locus. The changing bacterial and viral components demonstrate the importance of host specificity in aquatic systems. Systems in which this model is observed would have a relatively constant number of bacteria or cyanobacteria but underlying population changes would be occurring constantly. In this way, phage could have a greater effect on the changing population structures of a cyanobacterial bloom than on the overall population decline.

Viral infection is a factor in bloom decline. Concentrations of cyanophage increased as a result of cyanobacterial cell lysis in a Japanese lake (Manage et al., 1999). Similarly, addition of virus-like particles caused a decrease of Microcystis cells in culture (Tucker and Pollard, 2005). Interestingly, cells re-cultured to log phase after initial infection were largely unaffected by addition of virus-like particles, indicating that natural resistance can develop in populations. Increased phage concentration does not necessarily impose bloom decline as host specificity of a phage could only affect a
portion of a cyanobacterial bloom, changing the abundance of bacterial strains (Suttle, 2007).

Phage induced mortality of bacteria affects biogeochemical cycles. The viral shunt hypothesis implicates phage as important players in the conversion of photosynthetically produced compounds to dissolved organic matter affecting utilization rates (Suttle, 2005). Through the shunt carbon bypasses higher trophic levels and is retained in photic zones increasing aerobic respiration (Suttle, 2005). The increased concentration of cyanobacteria during a bloom can lead to hyperactive phage infections, intensifying the ecological effects of the viral shunt (Suttle, 2007).

To appreciate the impact of phage on ecology and bacterial mortality it is essential to survey phage diversity. To define viral population diversity various genetic methods have been used. Unlike other organisms, all viruses do not have a single universal gene and therefore a variety of genes have been targeted for analysis (Suttle, 2005). In the study of aquatic viral diversity, a common choice is that of gp20, phage portal protein of the Myoviridae family. The limitation of this choice is that only myoviruses with sequence similarity will be observed in such a survey. In this case, gp20 has shown extensive diversity of myophage in freshwater and marine environments (Short and Suttle, 2005).

As whole genome sequencing becomes more efficient and less costly, the small genomes of phage have become ideal candidates for sequencing. The number of marine cyanomyophage with genomes available has been steadily increasing. For example, 12 newly isolated cyanophage were discussed in a single 2010 paper (Sullivan et al., 2010).

When the first cyanophage genomes were sequenced efforts were made to group and compare genes. In 2005, the genomes of three marine Prochlorococcus cyanophage; two T4-like myophage and one T7-like podophage were published (Sullivan et al., 2005). In the T4-like phage genes homologous to T4 were as follows: P-SSM2 and P-SSM4, 43 and 42 homologs of 75 T4 genes and the 15 out of 26 in the T7-like P-SSP7. The homologs of T7 and T4 genes are part of the ‘core’ group of universal genes. These genes
were compared and grouped with that of the first sequenced cyanophage, *Synechococcus* phage S-PM2, shotgun sequenced in 2005 (Mann *et al.*, 2005).

Genes without homology to T4 or T7 but homologous to other cyanophage or bacteria were denoted as "signature" and "host" genes (Sullivan *et al.*, 2005). The "signature" genes are proposed to reflect adaptations to photosynthetic hosts in low nutrient environments and are found in all cyanomyophage. Some "host" genes are also cyanophage “signature” genes but not all. The “signature” genes for marine cyanophage include photosystem gene *psbA*. Many of these genes are homologous to cyanobacterial genes and could provide a similar function to their bacterial counterparts (Rohwer and Thurber, 2009). Other genes found in cyanophage are most closely related to genes in cyanobacteria, including carbon metabolism genes and phosphate stress genes that could function to keep a host cell metabolically active during infection. Keeping the cell active is thought to increase the efficiency of infection and the time during which new viruses can be replicated.

Cyanobacterial or ‘host’ genes are some of the most interesting genetic discoveries within cyanophage. These genes encode proteins involved in host cellular processes. Genes such as *prnA, psbA, phoH, petE* and *talC* are phage homologs of bacterial genes (Millard *et al.*, 2009). The presence of these genes indicates a close genetic relationship between phage and host where horizontal gene transfer is common. Increased levels of phycoerythrin found in *Synechococcus* sp. WH7803 cells infected by phage S-PM2, may be a direct consequence of expression of the phage *cpeT* phycobilin lyase (Shan *et al.*, 2008).

Nineteen cyanomyophage genomes have been sequenced and deposited in Genbank, two of which infect freshwater organisms (Table 1.1). Typical of phage, a majority of their genome is dedicated to coding. The phage infecting *Microcystis* and T4 are the most different of the phage listed in Table 1.1. The other phage have somewhat large genomes, ~200kbp and around 300 genes and all contain photosystem gene *psbA*. As an indication of the momentum in genomic sequencing, 13 of the genomes in Table 1.1 were deposited in 2011.
Table 1.1 Fully sequenced genomes of the *Myoviridae* family infecting cyanobacteria (and T4). Ordered from the date deposited in Genbank, 13 sequences were deposited in 2011 more than all previously sequenced. Excepting T4, only two infect fresh water organisms, S-CRM01 and Ma-LMM01; all others infect marine organisms. *S-CRM01 is the freshwater *Synechococcus* phage described in Chapter 3 (Dreher et al., 2011).

<table>
<thead>
<tr>
<th>Host</th>
<th>Genome Size (bp)</th>
<th>ORFs (# of)</th>
<th>tRNAs (# of)</th>
<th>G+C%</th>
<th>Coding % of genome</th>
<th>Genbank accession #</th>
<th>Deposition Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>T4</td>
<td><em>E. coli</em></td>
<td>168903</td>
<td>278</td>
<td>10</td>
<td>35%</td>
<td>AF158101</td>
<td>7/5/99</td>
</tr>
<tr>
<td>S-PM2</td>
<td><em>Synechococcus</em></td>
<td>196280</td>
<td>236</td>
<td>1</td>
<td>37.80%</td>
<td>AJ630128</td>
<td>1/29/05</td>
</tr>
<tr>
<td>P-SSM2</td>
<td><em>Prochlorococcus</em></td>
<td>252401</td>
<td>327</td>
<td>1</td>
<td>35%</td>
<td>AY939844</td>
<td>3/25/05</td>
</tr>
<tr>
<td>P-SSM4</td>
<td><em>Prochlorococcus</em></td>
<td>178249</td>
<td>198</td>
<td>0</td>
<td>36%</td>
<td>AY940168</td>
<td>3/25/05</td>
</tr>
<tr>
<td>Syn9</td>
<td><em>Synechococcus</em></td>
<td>177300</td>
<td>226</td>
<td>6</td>
<td>40%</td>
<td>DQ149023</td>
<td>8/25/06</td>
</tr>
<tr>
<td>Ma-LMM01</td>
<td><em>Microcystis</em></td>
<td>162109</td>
<td>184</td>
<td>2</td>
<td>45%</td>
<td>AB231700</td>
<td>11/3/06</td>
</tr>
<tr>
<td>S-RSM4</td>
<td><em>Synechococcus</em></td>
<td>194454</td>
<td>237</td>
<td>12</td>
<td>41%</td>
<td>FM207411</td>
<td>8/10/09</td>
</tr>
<tr>
<td>S-SSM7</td>
<td><em>Synechococcus</em></td>
<td>232878</td>
<td>319</td>
<td>5</td>
<td>39%</td>
<td>GU071098</td>
<td>3/24/11</td>
</tr>
<tr>
<td>S-SM2</td>
<td><em>Synechococcus</em></td>
<td>190789</td>
<td>267</td>
<td>11</td>
<td>40%</td>
<td>GU071095</td>
<td>3/24/11</td>
</tr>
<tr>
<td>P-HM2</td>
<td><em>Prochlorococcus</em></td>
<td>183806</td>
<td>242</td>
<td>0</td>
<td>38%</td>
<td>GU075905</td>
<td>3/24/11</td>
</tr>
<tr>
<td>P-HM1</td>
<td><em>Prochlorococcus</em></td>
<td>181044</td>
<td>241</td>
<td>0</td>
<td>37%</td>
<td>GU071101</td>
<td>3/24/11</td>
</tr>
<tr>
<td>S-ShM2</td>
<td><em>Synechococcus</em></td>
<td>179563</td>
<td>230</td>
<td>1</td>
<td>41%</td>
<td>GU071096</td>
<td>3/24/11</td>
</tr>
<tr>
<td>S-SSM5</td>
<td><em>Synechococcus</em></td>
<td>176184</td>
<td>225</td>
<td>4</td>
<td>39%</td>
<td>GU071097</td>
<td>3/24/11</td>
</tr>
<tr>
<td>P-RSM4</td>
<td><em>Prochlorococcus</em></td>
<td>176428</td>
<td>239</td>
<td>3</td>
<td>37%</td>
<td>GU071099</td>
<td>3/24/11</td>
</tr>
<tr>
<td>S-SM1</td>
<td><em>Synechococcus</em></td>
<td>174079</td>
<td>234</td>
<td>6</td>
<td>41%</td>
<td>GU071094</td>
<td>3/24/11</td>
</tr>
<tr>
<td>Syn33</td>
<td><em>Synechococcus</em></td>
<td>174825</td>
<td>227</td>
<td>5</td>
<td>39%</td>
<td>GU071108</td>
<td>3/24/11</td>
</tr>
<tr>
<td>Syn19</td>
<td><em>Synechococcus</em></td>
<td>175230</td>
<td>215</td>
<td>6</td>
<td>41%</td>
<td>GU071106</td>
<td>3/24/11</td>
</tr>
<tr>
<td>P-SSM7</td>
<td><em>Prochlorococcus</em></td>
<td>182180</td>
<td>237</td>
<td>4</td>
<td>37%</td>
<td>GU071103</td>
<td>3/24/11</td>
</tr>
<tr>
<td>Syn1</td>
<td><em>Synechococcus</em></td>
<td>191195</td>
<td>234</td>
<td>6</td>
<td>40%</td>
<td>GU071105</td>
<td>3/24/11</td>
</tr>
<tr>
<td>S-CRM01*</td>
<td><em>Synechococcus</em></td>
<td>178563</td>
<td>294</td>
<td>33</td>
<td>39.70%</td>
<td>HQ615693</td>
<td>5/23/11</td>
</tr>
</tbody>
</table>

The ease with which these genomes can be sequenced has led to innovations in the cataloging of phage genomes. In cyanomyophage, groups of genes have been proposed to distinguish homologs from distinctly related phage genes (Millard et al., 2009; Henn et al., 2010; Sullivan et al., 2010). Recently, T4 Gene Clusters (‘T4-GCs’) have been used to define the orthologous genes found among all T4-like phage (Sullivan et al., 2010). The organization of genes and genomes has made it easier to compare selected aspects of entire phage genomes. The comparison of multiple characteristics will
allow for better identification of newly sequenced phage genomes and subsequently a
greater understanding of phage diversity and genetic content.

**Experimental Goals**

Blooms and the associated toxins continue to have a great effect on global and
regional water bodies. The genetic descriptions of blooms will give insight into the
complex communities that arise annually and often over a season. These descriptions are
made possible with new genetic techniques that enable better understanding of
cyanobacterial genetics and genetics of phage that have the potential to affect blooms. In
order to fully characterize a bloom our research set out to genetically define a bloom
season while also isolating and sequencing phage infecting freshwater *Synechococcus*.

The goal of the research on *Microcystis* blooms (Chapter 2) was to describe
regional blooms genetically for future diagnosis and prediction of potential bloom
toxigenicity. Multiple genetic loci, cell counts and quantitative PCR were used to produce
a representation of toxigenic and non-toxigenic *Microcystis* over a bloom season in the
Copco Reservoir of the Klamath River. Distilling the genotypic information into groups
with maximum parsimony networks simplified the representation of the results of a clone
library study.

To explore other components of blooms, we focused on phage infecting
cyanobacteria (Chapter 3). Our cultures, developed from environmental samples of
*Microcystis* blooms, were plated in order to isolate phage from the environment.
Although the samples were taken from a *Microcystis* bloom, and liquid cultures appeared
to be *Microcystis* monocultures, the plating of the culture selected for a freshwater
*Synechococcus* present in the culture. Our original intention was to obtain and sequence a
*Microcystis* phage but nonetheless, our isolation of a freshwater cyanophage was
interesting and novel. The phage isolated, S-CRM01, is the first freshwater
‘photosynthetic’ phage, meaning it contains photosystem gene *psbA*. Chapter 3 focuses
on the similarities and differences between freshwater and marine cyanophage whose
genomes have been fully sequenced.
References

Toxigenicity of a Cyanobacterial Bloom by Molecular Methods." Appl. Environ.
Microbiol. 68: 6070–76.

Res. 47: 643-57.

transcribed spacer region a good tool for use in molecular systematics and

"Spatiotemporal changes in the genetic diversity of a bloom-forming Microcystis
aeruginosa (cyanobacteria) population." The ISME Journal 3: 419-29.

Deng, L. and P. K. Hayes (2008). "Evidence for cyanophages active against bloom-

Dreher, T. W., N. Brown, C. S. Bozarth, A. D. Schwartz, E. Riscoe, C. Thrash, S. E.
Bennett, S.-C. Tzeng and C. S. Maier (2011). "A freshwater cyanophage whose
genome indicates close relationships to photosynthetic marine cyanomyophages." Environmental Microbiology 13: 1858-74.

Falconer, I., A. Beresford and M. Runnagar (1983). "Evidence of liver damage by toxin

"Phylogenetic Comparison of the Cyanobacterial Genera Anabaena and

Henn, M. R., M. B. Sullivan, N. Stange-Thomann, M. S. Osburne, A. M. Berlin, L. Kelly,
C. Yandava, C. Kodira, Q. Zeng, M. Weiand, T. Sparrow, S. Saif, G. Giannoukos,
S. K. Young, C. Nusbaum, B. W. Birren and S. W. Chisholm (2010). "Analysis of
high-throughput sequencing and annotation strategies for phage genomes." PLoS
ONE 5: e9083.

"Toxic and nontoxic Microcystis colonies in natural populations can be

differentiation of Cyanobacteria by using rRNA-internal transcribed spacer

Holmes, M. B. Antunes, D. A. de Melo Filho, T. M. Lyra and V. S. T. Barreto


2. Population Turnover in a *Microcystis* Bloom Results in Predominantly Non-toxigenic Variants Late in the Season

Connie S. Bozarth, Andrew Schwartz, Jon She pardson, Frederick Colwell and Theo W. Dreher
Abstract

Surface samples of the 2007 *Microcystis* bloom occurring in Copco Reservoir on the Klamath River in Northern California were analyzed genetically by sequencing clone libraries made with amplicons at three loci: ITS, *cpcBA* and *mcyA*. Samples were taken between June and October, during which time two cell-count peaks occurred, in mid-July and early September. The ITS and *cpcBA* loci could be classified into 4 or 5 allele groups, which provided a convenient means for describing the *Microcystis* population and its changes over time. Each group was numerically dominated by a single, highly represented sequence. Other members of each group varied by changes at 1 to 3 nucleotide positions, while groups were separated by up to 30 nucleotide differences. As deduced by partial sampling of the clone libraries, there were marked population turnovers during the season, indicated by changes in allele composition at both the ITS and *cpcBA* loci. Different ITS and *cpcBA* genotypes appeared to be dominant at the two population peaks. Toxicity (microcystin per cell and *mcyB* copy number) was lower during the second peak, and *mcyB* copy number fell further as the bloom declined.
Introduction

Toxic freshwater cyanobacterial blooms, commonly caused by *Microcystis*, are of current concern in many parts of the world because of their effects on drinking water, water-based recreation and watershed ecology (Dittmann and Wiegand, 2006). *Microcystis* cells are able to produce microcystin, a non-ribosomally synthesized cyclic heptapeptide hepatotoxin with potent inhibitory activity against mammalian protein phosphatases (Vareli, 2009) whose synthesis is directed by the 55 kb *mcy* gene cluster (Tillett, 2000). The *Microcystis* genus exhibits world-wide occurrence, although the extent of genetic differentiation between or within geographical regions is currently uncertain due to a relatively sparse database, in spite of a growing number of studies (Allender, 2009; Bittencourt-Oliveira, 2001; Humbert, 2003; Janse, 2004; Tillett, 2001; Vareli, 2009; Ye, 2008).

Only a few studies to date have used gene-specific tools to investigate the changes in *Microcystis* population structure throughout the development of a bloom season. In some instances, there has been little indication of major population changes. Thus, the proportion of toxigenic (*mcyB*+) *Microcystis* was stable over the course of two consecutive bloom seasons in Lake Wannsee (Berlin, Germany) (Kurmayer, 2003). The ITS genotype (internal transcribed spacer of the ribosomal RNA operon), as assessed by DGGE and sequencing, was also stable in Lake Volkerak (The Netherlands) during 2001 (Kardinaal, 2007). By contrast, studies in other lakes have observed changes in the *Microcystis* genotypes and in the proportion of potentially toxigenic cells during a bloom season (Briand, 2009; Kardinaal, 2007; Rinta-Kanto, 2009; Yoshida, 2007). A better understanding of the population changes that occur during the development of toxic blooms is important in understanding their ecology and in assessing whether it might be feasible to manage *Microcystis* blooms in order to minimize toxicity.

Copco Reservoir is a lake formed by a hydroelectric dam on the Klamath River in northern California. Beginning in 2004, highly toxic blooms dominated by *Microcystis* have developed between June and November (Jacoby, 2007, Kann, 2007). Most studies of *Microcystis* blooms have been conducted in lakes with low in- and out-flows. Copco
Reservoir sits on a major river with normal through-flows during bloom season of 1000-3000 cfs (Kann, 2007). The consequences of toxic blooms in the reservoir may be carried to downstream reaches of the river, since elevated Microcystis levels have been present downstream of Copco Reservoir (Kann, 2009). We report here the results of a survey of the genotypic structure of the Microcystis population in Copco Reservoir during the 2007 bloom season. Major population shifts evident at the ITS and cpcBA loci coincided with the replacement of toxigenic with non-toxigenic strains.

**Materials and Methods**

**Sample Collection and DNA extraction.** Water samples (250 ml) were collected as surface grabs including floating scum by the Department of Natural Resources of the Karuk Tribe of Northern California from site CR01 in Copco Reservoir. CR01 is situated at a water depth of about 25 m near the dam at 41.982 latitude, -122.328 longitude, river mile 198.6. At one date (June 13), the sample was taken from the nearby shore (CRCC, Copco Cove ramp).

One sample collected from the top 3.5 m of the water column at site MDT (42.385 latitude, -121.927 longitude) of Upper Klamath Lake (UKL)(Lindenberg, 2006), 55 miles upstream of Copco Reservoir, (provided by the USGS), was included in the analysis. UKL is a shallow, well-mixed lake (90% of the lake is less than 4 m deep) that experiences persistent summer blooms of non-toxic Aphanizomenon flos-aquae; low levels of toxic Microcystis are commonly present.

Samples from both lakes were provided to us as splits from ongoing water quality monitoring programs in Copco Reservoir and Upper Klamath Lake. Water quality data from both sites and all sampling dates are available at Karuk Tribe [http://karuk.us/dnr/documentation.php](http://karuk.us/dnr/documentation.php) and USGS websites [http://or.water.usgs.gov/projs_dir/klamath_ltmn/](http://or.water.usgs.gov/projs_dir/klamath_ltmn/). Samples were shipped to the laboratory on ice and processed within 4 days of collection. Five to 25 ml of sample were filtered onto glass fiber disks (Whatman GF/C) stacked onto 0.2 µm Supor membrane filters (Pall) and frozen until extraction. The glass fiber filters were used in all
subsequent DNA extractions and the 0.2 μm filters were archived. No green coloration was detected on the 0.2 μm filters for any of the samples analyzed during this study. DNA was extracted as previously described (Neilan, 2001). Briefly, frozen glass fiber filters were ground, exposed to lysozyme (1 mg/ml), SDS (0.4%) and proteinase K (0.2 mg/ml) before double extraction with phenol:chloroform:isoamyl alcohol.

DNA analysis. The ITS, cpcBA and mcyA loci were PCR amplified with primers designed for cyanobacteria-wide specificity, with products cloned for clone library analysis. The primers were: ITS (481-491 bp amplicons), CS1F (GYCACGCCGAAGTCRTTAC) and ULR (CCTCTGTGTCCTAGGTATC) (Janse, 2003); cpcBA (623 bp amplicons), PCBF (GGCTGCTTGTTTACGGCACA) and PCAR (CCAGACCACGCAACTAA) (Neilan, 2001); mcyA (246-252 bp amplicons), mcyA-Cd1F (AAAATTAAAAGCCGTATCAAA) and mcyA-Cd1R (AAAAGTGTATTAGCGGCTCAT) (Hisbergues, 2003). All PCR reactions were in 25 µl with 1-10 ng DNA template, Platinum Taq buffer, 2.5 mM MgCl₂, 200 µM dNTPs, and 200 nM each of the forward and reverse primers, and 1.25 units Platinum Taq polymerase (Invitrogen). Reactions were conducted for 30 cycles with 56C annealing and 72C extension steps for 30 sec each, with a final 5 min extension at 72C. PCR products were detected by ethidium bromide staining after electrophoresis on 1.5% agarose gels.

Quantitative PCR targeting Microcystis-specific cpcBA and mcyB genes was conducted using primers and Taqman probes as described previously (Kurmayer, 2003). Reactions (25 µl) consisted of 12.5 µl Maxima Probe qPCR Master Mix (Fermentas), 900 nM each primer and 250 nM probe, and 5 ng of genomic DNA as template. Cycle conditions were 50C for 2 min, 95C for 10 min, followed by 40 cycles of 95C for 15 sec and 60C for 1min, in an ABI 7500 Real Time PCR machine. For each gene, a plasmid containing the target sequences amplified from Copco Reservoir Microcystis DNA was constructed. Dilutions of these plasmids and of DNA extracted from reference M. aeruginosa cultures UTEX 2385 (microcystin-producing) and UTEX 2386 (microcystin non-producing) (UTEX Culture Collection of Algae) were used as standards. Dilutions of
UTEX 2385 DNA were used to generate standard curves that were included in each run. All reactions were conducted in triplicate and in the linear response range.

The *cpcBA*, ITS and *mcyA* amplicons were cloned using a pGEM kit from Promega and chemically competent Top10 One Shot *E. coli* cells from Invitrogen. Selected colonies were screened by colony PCR using either the *cpcBA* or ITS primer pairs. Colonies that gave positive PCR products were grown overnight in 2X YT + carbenicillin to generate plasmid preparations (Purelink Quick Miniprep Kit, Invitrogen). Insert DNA sequences were confirmed by two-way sequencing and submitted to GenBank (accession numbers GU249152-GU249302).

Population diversity parameters were estimated with Arlequin v3.11 (nucleotide diversity)(Excoffier, 2005) and Fastgroup II (Shannon index)(Yu, 2006). Rarefaction curves (Fig. 2.1) showed that sampling was incomplete to varying degrees, as is common for clone library analysis of multiple samples (Sabart, 2009). Maximum parsimony network analysis was performed using the statistical parsimony program TCS v1.21 (Clement, 2000).
Figure 2.1 Rarefaction curves estimating the population sampling of clone libraries. Curves were constructed by Fastgroup II for each ITS and cpcBA clone library with sequence similarity set to 100%.

Results and Discussion

Transition from high to low toxicity during development of the Microcystis bloom during 2007. Surface samples were collected near the dam of Copco Reservoir throughout the 2007 bloom season, at CR01 from July through October, and at CRCC (Copco Cove boat ramp) on 27 June. Microscopic phycological analysis and microcystin
toxin analyses (ELISA) were obtained by the Karuk Tribe (results summarized in (Kann, 2009)). The cyanobacterial bloom was dominated by *Microcystis* at all times, with small appearances of *Aphanizomenon* and *Anabaena*. There were two peaks in *Microcystis* accumulation at CR01, an early peak measured at July 11 and a larger peak centered on the first week of September (Fig. 2.2).

**Figure 2.2 Progress of the *Microcystis* bloom in Copco Reservoir in 2007.** *Microcystis* cell counts (♦) and microcystin content (□) in surface samples taken at site CR01 at the indicated dates were assessed by microscopic phycological analysis and enumeration. Microcystin (extracellular plus intracellular) was determined by ELISA. Cell counts and toxin determinations were conducted with splits of the same samples, some of which were the same samples genetically analyzed in this study. Data were derived from reference (Kann, 2007).

The reported amount of microcystin present per *Microcystis* cell fell over 100-fold at CR01 during the course of the bloom, with the greatest decline occurring in the period spanning the last week of August and the first half of September (Fig. 2.2). This coincided with the peak of the bloom. The decline in toxin per cell coincided with a
decrease in the toxigenic potential of the *Microcystis* bloom as measured by the ratio of *mcyB* to *cpcBA* gene copy number (Fig. 2.3).

![Figure 2.3 Loss of the mcyB gene during the 2007 bloom season in Copco Reservoir.](image)

Quantitative PCR was used to determine the relative copy numbers of the *Microcystis*-specific *mcyB* and *cpcBA* loci (average from three experiments). Samples were taken from Copco Reservoir at the indicated dates and from the MDT8 site in Upper Klamath Lake (UKL) on 21 August 2007.

**Large-scale population turnover in Copco Reservoir indicated by analysis of the ITS alleles present.** Clone libraries representing cyanobacterial ITS amplicons were constructed for several time points to obtain a picture of the genetic makeup of the *Microcystis* population. 155 full-length sequences of the 481-491 bp *Microcystis* ITS sequence were obtained, among which there were 61 different sequences. There were no matches to these sequences in the GenBank database.

Initial phylogenetic analysis of the ITS sequences suggested classification into four sequence groups. This was confirmed by construction of a maximum parsimony network (Fig. 2.4), which also revealed the makeup of each sequence group and their relationships to each other. Groups 1-4 are dominated by abundant single sequences (termed *most abundant sequences*, MAS) with an array of rare closely related sequences.
differing from the MAS by 1, or in a few cases 2 or 3 nucleotide differences (Fig. 2.4). Groups 1-4 are well separated genetically, with between 9 and 30 nucleotide differences separating the MAS of each group. Only three sequences did not fit into the four groups. Classification of sequences into groups thus provides a convenient framework for analyzing the population changes occurring throughout the season.
Figure 2.4 Relationships between ITS genotypes collected from a 2007 *Microcystis* bloom in Copco Reservoir. The maximum parsimony network created with the TCS version 1.21 program (Clement, 2000) from all sequences generated from clone libraries made from the June through October samples separates the ITS sequences into four allele groups. The larger circles represent the MAS genotypes of each group, which are identified by name. The areas of circles are roughly proportional to the number of times a sequence was found, which is also indicated by the numbers entered within the larger circles. The number of nucleotide differences between two genotypes is the sum of steps on the shortest connecting path, summing crosshatches, intervening genotypes, and junction nodes (small circles). Sampling times are color coded as indicated in the legend. MAS GenBank accession numbers are as follows: ITS-1, GU249289; ITS-2, GU249253; ITS-3, GU249225; ITS-4, GU249216.
Genotype diversity statistics estimated from ITS clone libraries made from the different time-point samples (Table 2.1) indicate overall low diversity with a diversity peak for the August samples. Maximum parsimony networks for clone libraries made from each sample (Fig. 2.5) confirm a simple predicted population structure, with more diversity present during August. Changes in the ITS alleles across the bloom season are shown in Fig. 4, which plots the relative contributions of ITS alleles from Groups 1-4 for each time point. Group 2 sequences predominated early, during June and July, with Group 2 MAS accounting for up to 92% of the Group 2 sequences analyzed. By contrast, Group 1 sequences predominated late in the season, during September and October (Fig. 2.6), and a single sequence was again a large part of that population (up to 67%). The shift from predominantly Group 2 sequences coincided with the decline of the early (June/July) bloom (Fig. 2.2), while the second bloom peak seemed to correspond to the ascendancy of Group 1 sequences, which persisted beyond the bloom decline in mid-September. High levels of toxicity (Fig. 2.2) and the presence of the mcyB gene (Fig. 2.3) were associated with the early, but not the late, peak. A mix of sequences was recovered in the August clone libraries, during the period between Group 2 and Group 1 dominance.
Table 2.1 Diversity of clone libraries made from each sample

<table>
<thead>
<tr>
<th>Clone library and sample date</th>
<th>No. of clones</th>
<th>No. of genotypes</th>
<th>No. of clones of most-abundant genotype (MAS)</th>
<th>Nucleotide diversity</th>
<th>Shannon index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ITS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copco</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 June</td>
<td>24</td>
<td>10</td>
<td>15 (ITS-2)</td>
<td>0.005 ± 0.003</td>
<td>1.27</td>
</tr>
<tr>
<td>11 July</td>
<td>26</td>
<td>5</td>
<td>22 (ITS-2)</td>
<td>0.004 ± 0.003</td>
<td>0.48</td>
</tr>
<tr>
<td>8 August</td>
<td>26</td>
<td>13</td>
<td>5 (ITS-2)</td>
<td>0.029 ± 0.015</td>
<td>2.22</td>
</tr>
<tr>
<td>23 August</td>
<td>32</td>
<td>21</td>
<td>11 (ITS-1)</td>
<td>0.031 ± 0.016</td>
<td>1.93</td>
</tr>
<tr>
<td>19 September</td>
<td>23</td>
<td>14</td>
<td>10 (ITS-1)</td>
<td>0.008 ± 0.002</td>
<td>0.70</td>
</tr>
<tr>
<td>17 October</td>
<td>24</td>
<td>7</td>
<td>16 (ITS-1)</td>
<td>0.008 ± 0.003</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>UKL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 August</td>
<td>17</td>
<td>9</td>
<td>9 (ITS-2)</td>
<td>0.053 ± 0.027</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>epcBA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copco</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 June</td>
<td>14</td>
<td>7</td>
<td>7 (CPC-A)</td>
<td>0.014 ± 0.007</td>
<td>0.90</td>
</tr>
<tr>
<td>11 July</td>
<td>21</td>
<td>9</td>
<td>13 (CPC-A)</td>
<td>0.002 ± 0.001</td>
<td>1.29</td>
</tr>
<tr>
<td>8 August</td>
<td>16</td>
<td>9</td>
<td>8 (CPC-A)</td>
<td>0.007 ± 0.004</td>
<td>1.12</td>
</tr>
<tr>
<td>23 August</td>
<td>17</td>
<td>6</td>
<td>12 (CPC-A)</td>
<td>0.031 ± 0.001</td>
<td>0.44</td>
</tr>
<tr>
<td>19 September</td>
<td>17</td>
<td>15</td>
<td>4 (CPC-B2)</td>
<td>0.014 ± 0.007</td>
<td>1.87</td>
</tr>
<tr>
<td>17 October</td>
<td>23</td>
<td>6</td>
<td>12 (CPC-B1)</td>
<td>0.022 ± 0.011</td>
<td>1.19</td>
</tr>
<tr>
<td><strong>UKL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 August</td>
<td>19</td>
<td>4</td>
<td>16 (CPC-A)</td>
<td>0.013 ± 0.007</td>
<td>0.61</td>
</tr>
</tbody>
</table>

* For GenBank accession numbers of the listed most-abundant sequences (MAS), see Fig. 3 and 5.
* Nucleotide diversity obtained using Arlequin version 3.11.
* Shannon index obtained using FastGroup II (set at 100% sequence identity with gaps).
Figure 2.5 Relationships between ITS genotypes at individual time points sampled from Copco Reservoir. Maximum parsimony networks created from sequences generated from clone libraries for the individual time samples are shown. Details as for Fig. 2.4.
Figure 2.6 Changes in ITS and cpcBA allele populations during the 2007 bloom season in Copco Reservoir. (A) The relative abundances of the members of ITS sequence groups 1 to 4 are plotted across the bloom season; poorly clustered sequences GU249218 and GU249219 are represented as "other." The June sample was from site CRCC, and all others were from CR01. The dashed lines report the relative abundances of the MAS genotypes belonging to groups 1 and 2 (ITS-1 and ITS-2, respectively). (B) The relative abundances of the members of cpcBA sequence groups A to D are plotted across the bloom season. The dashed lines report the relative abundances of the MAS
genotypes belonging to groups A, B1, and C (CPC-A, CPC-B1, and CPC-C, respectively).

Different timing of population shifts suggested by analysis at the cpcBA locus. A similar clone library analysis was conducted for cpcBA amplicons. 132 full-length sequences of the 623-bp Microcystis cpcBA amplicon were recovered, representing 46 different sequences. Several of these matched sequences recently reported from Copco Reservoir or its sister impoundment Iron Gate Reservoir (Moisander, 2009), and identical sequences were also present in GenBank from worldwide sources such as Spain, Japan, Israel and Switzerland.

Analysis at the cpcBA locus generally paralleled the results from the ITS analysis, indicating an allele population similarly low in diversity (Table 2.1) and dominated by four sequence groups, A, B1, B2 and C, with a minor group, D (Fig. 2.8, Fig. 2.7). Only two sequences did not fit into these five groups. Identical cpcBA sequences recently reported from Copco Reservoir (Moisander, 2009) mostly belonged to Groups A, B2 and C. As deduced from the ITS analysis, large population shifts are also indicated from cpcBA analysis, but the timing is different (Fig. 2.9). Group A sequences predominated in clone libraries made from June through 23 August, after which time they were replaced by a mixture of sequences from the other groups. Because the mcyB copy number was very low in the September and October samples (Fig. 2.3), most of the Group B1, B2, C and D cpcBA alleles would be expected to derive from non-toxigenic Microcystis strains.
Figure 2.7 Relationships between cpcB.4 genotypes at individual time points sampled from Copco Reservoir. Maximum parsimony networks created from sequences generated from clone libraries for the individual time samples are shown. Details as for Fig. 2.4.
Figure 2.8 Relationships between cpcBA genotypes collected from a 2007 Microcystis bloom in Copco Reservoir. The maximum parsimony network created from all sequences generated from clone libraries made from June through October separates the cpcBA genotypes into five allele groups. Details are as in Fig. 2.4. MAS GenBank accession numbers are as follows: CPC-A, GU249203; CPC-B1, GU249162; CPC-B2, GU249171; CPC-C GU249154; CPC-D, GU249176.

The major population switch indicated by cpcBA alleles thus occurred in between 23 August and 19 September, and the period of greatest cpcBA diversity occurred during September and October (Fig. 2.9). This contrasts with major shifts in ITS alleles between 11 July and late August, with apparent ITS diversity highest during August (Fig. 2.6). Analysis at the two loci thus reveal population transitions that are more complex than indicated by either locus alone. Because of the dominance of certain allele groups in most samples, and incorporating our analyses of relative mcyB levels (Fig. 2.3), we can
infer the predominant 3-locus genotypes (ITS/cpcBA/mcy) of Microcystis present at different sampling times in Copco Reservoir (Fig. 2.9). Most importantly, because the transition from the dominant early genotype occurred at different times for the ITS and cpcBA loci, there seem to be strains present in the Copco system that reflect recombination between these genes. Thus, cpcBA Group A strains with either Group 2 or Group 1 ITS alleles seem to be present, and these are both probably mcy+. Similarly, it appears that ITS Group 1 strains can have cpcBA Group A, C or B1 alleles; these may all be non-toxic, except perhaps the Group A members (Fig. 2.9). Such complexities have not previously been revealed in Microcystis populations, which have usually been studied at a single locus outside the mcy gene cluster.

**Figure 2.9 Microcystis genotypes in Copco Reservoir.** The predicted genotypes at the three analyzed loci for most of the population at the 2007 sampling times are indicated at left. Mixed populations were present on 8 August and 23 August. The determined genotypes of three isolated colonies taken from Copco Reservoir on 15 July 2008 (colonies 8 and 15) or 5 August 2008 (colony 24) are given at right. The amplicon sequences derived from these colonies matched exactly the MAS of the indicated sequence groups.

Gene linkages such as those hypothesized in Fig. 2.9 must be verified by direct means, such as genetic analysis of single cells or clonal colonies (Janse, 2004). No colonies were isolated from the 2007 Copco Reservoir bloom, but three colonies from the 2008 bloom were shown to have the genotypes cpcBA-A/ITS-2/mcy+ and cpcBA-C/ITS-
These correspond to postulated genotypes that were prevalent during 2007. This type of colony analysis coupled with population analysis at more than one locus (such as cpcBA, ITS) is needed to obtain a full picture of the population dynamics of a bloom.

Transitions in Microcystis blooms from initial high toxicity to low toxicity have been observed previously, involving a variety of lake types (Briand, 2009; Kardinaal, 2007; Yoshida, 2007). Among these, the study by Briand et al. (2009) in Grangent Reservoir, a major dam on a large river, is the case most closely resembling Copco Reservoir. On the other hand, bloom seasons without major changes in toxicity have been reported from other lakes (Kardinaal, 2007; Kurmayer, 2003). More studies are necessary to determine the lake-specific and seasonal influences that govern such population changes.

Temporal population differentiation at the mcya locus. Clone libraries of mcya amplicons were analyzed for the Copco Reservoir samples taken at three time points, 11 July, 23 August and 19 September. About 10 amplicons were fully sequenced from each clone library, revealing seven different sequences. Sequence GU249297 was the most prevalent, found 10 times for 11 July and 5 times for 23 August. This sequence (and 4 other cloned sequences) encodes an McyA protein with Thr-Phe appearing between amino acids Lys-1610 and Ser-1611 of McyA encoded by M. aeruginosa PCC 7806 (Genbank AM778952). mcya genes with this Lys-Ser dipeptide were first reported from the Great Lakes (Rinta-Kanto, 2006); see also (Allender, 2009).

A single sequence found in the 23 August library encoded a Thr-Trp dipeptide at the above location of the McyA protein. Similar sequences were recently reported from Copco Reservoir (Moisander, 2009) and from Florida (Yilmaz, 2009).

A final distinct mcya sequence lacked the additional dipeptide (as in M. aeruginosa PCC 7806, AF183408); this sequence was the most prevalent clone (found 8 times) in the 19 September library. Four closely related sequences were also found mostly during September 2007 in another study of Copco reservoir (Moisander, 2009).
The Upper Klamath Lake Microcystis population is closely related to the early season Copco population. ITS, *cpcBA* and *mcyA* amplicon clone libraries were analyzed for a single sample from Upper Klamath Lake (UKL) collected on 21 August, 2007. Sequences identical to those found in Cocpo Reservoir dominated each library. Thus, the ITS Group 2 MAS was found in 9/17 clones (53%), Group 2 was represented in 13/17 clones (76%), with Group 1 and 3 sequences accounting for 2/19 clones (12%) each (Fig. 2.6A). In the *cpcBA* library, the Group A MAS was found in 12/19 clones (63%), with Group A sequences accounting for 16/19 clones (84%) (Fig. 2.6B). The remaining sequenced clones were one from Group B2 and two clones related to Group B1. Seven ITS and 5 *cpcBA* sequences not found in the Copco Reservoir libraries were recovered from UKL; all 8 *mcyA* clones sequenced were identical to the most prevalent sequence found in the Copco Reservoir clone libraries (GU249297).

The sequence groups represented in the UKL sample suggested the presence of a bloom similar to the early bloom present in Copco Reservoir during June (Fig. 2.6). This is the period when toxicity was high in Copco (Fig. 2.2), and a high *mcyB* to *cpcBA* gene copy ratio was indeed measured in the UKL sample (Fig. 2.3). The coexistence of several genotypes in UKL and Copco Reservoir, separated by about 80 river km, may not be surprising. However, the two lakes and their bloom characteristics are very different, a situation that a priori may be expected to favor different genotypes. UKL is a relatively high altitude (1260 m), large shallow lake that is dominated by annual non-toxic *Aphanizomenon* blooms, while Copco Reservoir is an impoundment on a major river situated at lower altitude (800 m) and in a canyon, thereby exposing it to higher summer temperatures. A recent study of *Microcystis* blooms on a major river in France showed that blooms in different impoundments on the same river do not necessarily equilibrate (Sabart, 2009). A major variable in whether blooms are transported to inoculate sites downstream of a reservoir may be the depth at which water is released relative to the location of the bloom in the water column (Sabart, 2009).
Potential for genetic group-specific monitoring and population studies.

Description of the *Microcystis* population in terms of the small number of sequence groups that dominate the network maps of Figs. 2.4 and 2.8 has facilitated the recognition of the inferred population changes that occurred during 2007 (Figs. 2.6). It also offers the potential to develop nucleic acid probes for group-specific detection to aid in population monitoring. For pragmatic reasons, our study did not involve a complete sampling of clone libraries, but inclusion of sequences of the same cpcBA and mcyA amplicons from the 2007 Copco Reservoir bloom water samples collected in another recent study (Moisander, 2009) supported our classification into discrete allele groups (not shown). Only a few sequences do not fall into the groups we have described in this study.

Previous authors have classified alleles present in *Microcystis* blooms into genotype groupings (Briand, 2009; Kardinaal, 2007; Tanabe, 2009; Yoshida, 2007). However, this is the first indication from two loci (ITS and cpcBA) that a *Microcystis* bloom is comprised of well-separated allele groups that are dominated by single genotypes. ITS analyses of *Microcystis* blooms on the Loire River (France) have indicated a similar population structure (Briand, 2009; Sabart, 2009), suggesting that a simplified group-based description may be more generally applicable. Group-specific PCR or hybridization assays may provide a convenient and economical approach to simplified monitoring of *Microcystis* populations that provides valuable insight into population structure. Such an approach is widely used in other fields for single nucleotide polymorphism (SNP) genotyping (Kim, 2007). This approach would likely need an initial inventory of the major allele groups present in a given study area, and would also need to incorporate controls for detecting the emergence of novel sequences that fall outside the known allele groups.

Acknowledgments

We gratefully acknowledge the assistance of Susan Corum of the Karuk Tribe of California in collecting Copco Reservoir samples, and of Mary Lindenberg and Tamara Wood of the US Geological Survey in collecting the UKL sample. We thank Andrea
McHugh for assisting in the laboratory, and Joe Eilers and Richard Raymond for discussions. This work was supported by Oregon Sea Grant award NA060AR4170010.

References


3. A Freshwater Cyanophage Whose Genome Indicates Close Relationships to Photosynthetic Marine Cyanomyophage.

Theo W. Dreher, Nathan Brown, Connie S. Bozarth, Andrew D. Schwartz, Erin Riscoe, Cameron Thrash, Samuel E. Bennett, Shin-Cheng Tzeng and Claudia S. Maier
Abstract

Bacteriophage S-CRM01 has been isolated from a freshwater strain of *Synechococcus* and shown to be present in the upper Klamath River valley in northern California and Oregon. The genome of this lytic T4-like phage has a 178,563 bp circular genetic map with 297 predicted protein-coding genes and 33 tRNA genes that represent all 20 amino acid specificities. Analyses based on gene sequence and gene content indicate a close phylogenetic relationship to the “photosynthetic” marine cyanomyophage infecting *Synechococcus* and *Prochlorococcus*. Such relatedness suggests that freshwater and marine phage can draw on a common gene pool. The genome can be considered as being comprised of three regions. Region 1 is populated predominantly with structural genes, recognized as such by homology to other T4-like phage and by identification in a proteomic analysis of purified virions. Region 2 contains most of the genes with roles in replication, recombination, nucleotide metabolism and regulation of gene expression, as well as 5 of the 6 signature genes of the photosynthetic cyanomyophage (*hli03, hsp20, mazG, phoH* and *psbA; cobS* is present in Region 3). Much of Regions 1 and 2 are syntenous with marine cyanomyophage genomes, except that a segment encompassing Region 2 is inverted. Region 3 contains a high proportion (85%) of genes that are unique to S-CRM01, as well as most of the tRNA genes. Regions 1 and 2 contain many predicted late promoters, with a combination of CTAAATA and ATAAATA core sequences. Two predicted genes that are unusual in phage genomes are homologs of cellular *spoT* and *nusG*. 
Introduction

Research over recent years has brought into focus the important contribution of bacteriophage to the ecology of microbial populations and to biochemical and geochemical cycles in the environment; bacteriophage are both enormously abundant and diverse (Rohwer, 2003; Weinbauer, 2004; Breitbart and Rohwer, 2005; Rohwer et al., 2009). Marine environments have attracted the most attention (Fuhrman, 1999; Wommack and Colwell, 2000; Breitbart et al., 2002; Mann, 2003; Suttle, 2005; 2007), with relatively few studies dedicated to investigating bacteriophage present in freshwater environments (Middelboe et al., 2008; Wilhelm and Matteson, 2008). We have been interested in studying the microbial diversity and population dynamics of toxic cyanobacterial blooms (Bozarth et al., 2010), which are an increasingly common ecological dysfunction seen in a wide variety of lakes and reservoirs (Paerl and Huisman, 2009).

The only currently available genome sequence from a phage infecting freshwater cyanobacteria is that of the myophage Ma-LMM01 infecting Microcystis aeruginosa from a Japanese lake (Yoshida et al., 2008). We describe here the genome sequence and other properties of a cyanomyophage isolated from Copco Reservoir on the Klamath River in Northern California in September 2008. The phage was associated with a toxic Microcystis aeruginosa bloom, but an endemic Synechococcus lineage is the host for this phage. The genome sequence revealed close relationships to a well-studied group of ‘photosynthetic’ exoT4-even cyanomyophage infectious to marine Synechococcus and Prochlorococcus (Mann et al., 2005; Sullivan et al., 2005; 2010; Weigele et al., 2007; Millard et al., 2009). This similarity between freshwater and marine cyanomyophage supports indications from metagenomic (Rodriguez-Brito et al., 2010) and amplicon studies with psbA and gp20 (T4 portal protein) that related phage can be found in freshwater and marine environments (Dorigo et al., 2004; Short and Suttle, 2005; Wilhelm et al., 2006; Chénard and Suttle, 2008; Sullivan et al., 2008), although there is also evidence for distinctively freshwater cyanophage lineages (Deng and Hayes, 2008; Yoshida et al., 2008; Wang et al., 2010). Our study is the first involving whole
genome characterization to address the relationship between freshwater and marine cyanophages, and supports the possibility that the gene complement of water-borne phage has been shaped by gene pools in both freshwater and marine environments (Sano et al., 2004).

Results and discussion
Isolation and physical characteristics

Phage S-CRM01 was isolated from a surface sample taken in September 2008 from a Microcystis-dominated bloom in Copco Reservoir on the Klamath River in Northern California. At 319.8 river km from the coast, this site is far removed from salt water habitats. S-CRM01 infects and lyases LC16, a culture belonging to the Cyanobium gracile cluster of mostly freshwater Synechococcus (see Experimental procedures). We have observed lytic infection of no other hosts, including the freshwater cyanobacteria Microcystis aeruginosa, Synechococcus elongatus PCC 7942 and Synechocystis sp. PCC6803. During 2009, a survey was conducted to assess the distribution of S-CRM01 between Upper Klamath Lake and the Klamath River estuary. On the basis of PCR identification of plaques on LC16 plates, S-CRM01 was present across a length of about 250 km along the Klamath River valley, from the Williamson River Delta at the northern end of Upper Klamath Lake as far downstream as Seiad Valley (Fig. 3.1).

Figure 3.1 Distribution of S-CRM01 in the Klamath River valley, 2009. Positive identifications were made in samples from the Williamson River delta area of Upper Klamath Lake (WRD; 15 October), Copco Reservoir (near dam; 13 October), Iron Gate Reservoir (IG, near dam, 18 August & 15 September), and at two sites on the lower Klamath River: Brown Bear sampling site at Horse Creek (BB, 6 August & 15 September) and Seiad Valley (SV, 15 September).
Negative staining electron microscopy revealed a T4-like morphology, with isometric heads about 85–100 nm in diameter and rigid contractile tails 15–20 nm in diameter and 140–170 nm long (Fig. 3.2). Contracted tails reveal a sheath and core 20–30 nm and about 10 nm in diameter respectively. A double-ringed baseplate and narrow neck are visible.

![Electron micrograph of S-CRM01 phage particles negatively stained with phosphotungstic acid. An intact particle is shown at left and a contracted particle at right. Note the icosahedral head, prominent neck, two-ringed baseplate and injection tube.](image)

The genome migrated as a 180 kbp band in PFGE. Nuclease BAL-31 digestion followed by cleavage with BamHI resulted in simultaneous loss of material from all bands, indicative of circularly permuted linear DNA (not shown). This is consistent with the presence of a gene homologous to the *gp17* large terminase subunit of phage T4, which determines a circularly permuted packaging strategy (Casjens and Gilcrease, 2008).

**S-CRM01 is a ‘photosynthetic’ cyanomyophage most closely related to marine phage of *Synechococcus* and *Prochlorococcus***

Consistent with a circularly permuted organization and its migration in PFGE, the S-CRM01 genome assembled into a circular map 178 563 bp long (Fig. 3.3). The G+C content of the genome is 39.7%, and 297 protein-coding genes and 33 tRNA genes are predicted. Protein-coding genes have been annotated with the following conventions
(Table 3.1): S-CRM01 genes are designated as g1, etc.; in addition, genes with homology to numbered phage T4 genes are designated as gp1, etc.; genes lacking direct homology with T4 but with similarity to other cyanomyophage genes that themselves have significant homology to T4 are designated as gp-like; other genes are designated with conventional gene names, e.g. nrdA.
Figure 3.3. Circular genetic map showing the gene organization of phage S-CRM01. The genome can be considered as comprised of Regions 1–3 (outer ring) that are dominated by structural genes, replication-related genes and unique genes respectively. Subsequent rings represent: positive strand ORFs, negative strand ORFs, tRNAs (maroon), and positive mass spectrometry identification (mauve). The inner rings show GC content around a 50% mid-line (black), and GC skew (G+ = cyan, C+ = blue). Putative ORF function is described by colour as listed below the circular map.
Table 3.1. Predicted protein-coding genes of S-CRM01. The genome contains 34 genes with homology to T4, and a total of 86 ORFs (29% of total) with homology to ORFs found in at least one of a group of 17 closely related myophages that lytically infect marine *Synechococcus* or *Prochlorococcus* (Mann et al., 2005; Sullivan et al., 2005; 2010; Weigele et al., 2007; Millard et al., 2009) (Tables 3.1). Phage S-PM2 from this group shares the largest number of genes with S-CRM01 (75), while the others have between 60 and 73 genes in common with S-CRM01. In contrast, the S-CRM01 genome shares a mere four genes with the only other known genome from a freshwater cyanomyophage, that of Ma-LMM01, which infects *Microcystis aeruginosa* (Yoshida et al., 2008).
<table>
<thead>
<tr>
<th>ORF No.</th>
<th>Strand</th>
<th>Size (aa)</th>
<th>T4 similarity*</th>
<th>Other genes</th>
<th>Mass spectrometry ID</th>
<th>E-value</th>
<th>Annotation from best hit</th>
<th>Accession number of best hit</th>
<th>T4 gene clusters‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 +</td>
<td>143</td>
<td>gp35</td>
<td></td>
<td></td>
<td></td>
<td>8.0E-20</td>
<td>Base plate wedge - P-SSM2</td>
<td>YP_214341</td>
<td>105</td>
</tr>
<tr>
<td>4 +</td>
<td>621</td>
<td>gp8</td>
<td>X</td>
<td></td>
<td></td>
<td>3.0E-10</td>
<td>Base plate wedge subunit - S-PM2</td>
<td>YP_196114</td>
<td>106</td>
</tr>
<tr>
<td>5 +</td>
<td>1455</td>
<td>gp7-like</td>
<td>X</td>
<td></td>
<td></td>
<td>1.0E-31</td>
<td>Base plate wedge initiator - S-RSM4</td>
<td>YP_00097385</td>
<td>121</td>
</tr>
<tr>
<td>7 +</td>
<td>505</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0E+00</td>
<td>Hypothetical protein SSMp2116 - P-SSM2</td>
<td>YP_214348</td>
<td>112</td>
</tr>
<tr>
<td>8 +</td>
<td>455</td>
<td>gp8-like</td>
<td>X</td>
<td></td>
<td></td>
<td>2.0E-25</td>
<td>Tail collar domain protein similar to Desulfovibrio salernensis DSM 2638</td>
<td>YP_00099129</td>
<td>167/2199</td>
</tr>
<tr>
<td>9 +</td>
<td>2485</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>5.0E-18</td>
<td>Putative structural protein SRSmp1494 - S-RSM4</td>
<td>YP_196118</td>
<td>828</td>
</tr>
<tr>
<td>10 +</td>
<td>69</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>5.0E-06</td>
<td>Hypothetical protein SSMp2116 - P-SSM2</td>
<td>YP_214348</td>
<td>112</td>
</tr>
<tr>
<td>11 +</td>
<td>437</td>
<td>gp37</td>
<td>gp12-like</td>
<td>X</td>
<td></td>
<td>7.0E-25</td>
<td>Tail collar domain protein similar to Desulfovibrio salernensis DSM 2638</td>
<td>YP_00099129</td>
<td>167/2199</td>
</tr>
<tr>
<td>12 +</td>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.0E-10</td>
<td>Collagen triple helix repeat domain protein Burkholderia phymatum STM91</td>
<td>YP_001859603</td>
<td></td>
</tr>
<tr>
<td>13 +</td>
<td>237</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0E-09</td>
<td>Phage phiL001 g88</td>
<td>YP_224012</td>
<td></td>
</tr>
<tr>
<td>14 +</td>
<td>250</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.0E-10</td>
<td>Hypothetical protein SSMp2116 - P-SSM2</td>
<td>YP_214348</td>
<td>112</td>
</tr>
<tr>
<td>16 +</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-41</td>
<td>Phage tail collar domain protein Ralstonia phage RSL1</td>
<td>YP_00099074</td>
<td>127</td>
</tr>
<tr>
<td>17 +</td>
<td>174</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-12</td>
<td>3-deoxy-D-arabino-heptulosonate 7-phosphate synthase from Campylobacter jejuni RM3289</td>
<td>YP_05426792</td>
<td></td>
</tr>
<tr>
<td>18 +</td>
<td>326</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-17</td>
<td>Hypothetical protein SSMp2116 - P-SSM2</td>
<td>YP_214348</td>
<td>112</td>
</tr>
<tr>
<td>19 +</td>
<td>133</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-09</td>
<td>Hypothetical protein SSMp2116 - P-SSM2</td>
<td>YP_214348</td>
<td>112</td>
</tr>
<tr>
<td>20 +</td>
<td>579</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.0E-26</td>
<td>DNA terminase - S-RSM4</td>
<td>YP_00097385</td>
<td>121</td>
</tr>
<tr>
<td>22 +</td>
<td>512</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-09</td>
<td>Hypothetical protein from Micrococcus phage MalMM01p164</td>
<td>YP_851178</td>
<td>1348</td>
</tr>
<tr>
<td>24 +</td>
<td>272</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0E-10</td>
<td>Tail sheath stabilizer - Syn9</td>
<td>YP_717778</td>
<td>117</td>
</tr>
<tr>
<td>26 +</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.0E-09</td>
<td>Hypothetical protein from Micrococcus phage MalMM01p164</td>
<td>YP_851178</td>
<td>1348</td>
</tr>
<tr>
<td>29 +</td>
<td>382</td>
<td>gp13</td>
<td></td>
<td></td>
<td></td>
<td>6.0E-31</td>
<td>Neck protein - S-PM2</td>
<td>YP_196127</td>
<td>116</td>
</tr>
<tr>
<td>30 +</td>
<td>352</td>
<td>gp14</td>
<td></td>
<td></td>
<td></td>
<td>5.0E-34</td>
<td>Neck protein - Syn9</td>
<td>YP_717778</td>
<td>117</td>
</tr>
<tr>
<td>31 +</td>
<td>283</td>
<td>gp15</td>
<td>gp16-like</td>
<td>X</td>
<td></td>
<td>1.0E-57</td>
<td>Tail sheath stabilizer - Syn9</td>
<td>YP_717778</td>
<td>117</td>
</tr>
<tr>
<td>32 +</td>
<td>141</td>
<td></td>
<td>gp16-like</td>
<td>X</td>
<td></td>
<td>9.0E-26</td>
<td>DNA terminase - S-RSM4</td>
<td>YP_00097385</td>
<td>121</td>
</tr>
<tr>
<td>33 +</td>
<td>219</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.0E-09</td>
<td>Hypothetical protein from Micrococcus phage MalMM01p164</td>
<td>YP_851178</td>
<td>1348</td>
</tr>
<tr>
<td>34 +</td>
<td>517</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.0E-13</td>
<td>Tail sheath stabilizer - P-SSM4</td>
<td>YP_214693</td>
<td>125</td>
</tr>
<tr>
<td>35 +</td>
<td>568</td>
<td>gp17</td>
<td></td>
<td></td>
<td></td>
<td>6.0E-13</td>
<td>Tail sheath stabilizer - P-SSM4</td>
<td>YP_214693</td>
<td>125</td>
</tr>
<tr>
<td>36 +</td>
<td>805</td>
<td>gp18</td>
<td></td>
<td></td>
<td></td>
<td>6.0E-43</td>
<td>Tail tube protein - P-SSM2</td>
<td>YP_214362</td>
<td>126</td>
</tr>
<tr>
<td>37 +</td>
<td>200</td>
<td>gp19</td>
<td></td>
<td></td>
<td></td>
<td>6.0E-13</td>
<td>Tail tube protein - P-SSM2</td>
<td>YP_214362</td>
<td>126</td>
</tr>
<tr>
<td>38 +</td>
<td>583</td>
<td>gp20</td>
<td></td>
<td></td>
<td></td>
<td>4.0E-17</td>
<td>Portal vertex head protein - S-RSM4</td>
<td>YP_00359734</td>
<td>127</td>
</tr>
<tr>
<td>39 +</td>
<td>415</td>
<td>gp21</td>
<td>gp12-like</td>
<td>X</td>
<td></td>
<td>1.0E-79</td>
<td>Prolactin core scaffold protein and protease - S-PM2</td>
<td>YP_196150</td>
<td>129</td>
</tr>
<tr>
<td>40 +</td>
<td>637</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.0E-91</td>
<td>Pre-peptidase C-terminal domain-containing protein Acaryochloris</td>
<td>YP_00151712</td>
<td></td>
</tr>
<tr>
<td>41 +</td>
<td>361</td>
<td>gp22</td>
<td>gp12-like</td>
<td>X</td>
<td></td>
<td>4.0E-54</td>
<td>Prolactin core scaffold protein and protease - P-SSM2</td>
<td>YP_214366</td>
<td>130</td>
</tr>
<tr>
<td>42 +</td>
<td>440</td>
<td>gp23</td>
<td>gp12-like</td>
<td>X</td>
<td></td>
<td>7.0E-17</td>
<td>Major capsid protein - P-SSM2</td>
<td>YP_214367</td>
<td>131</td>
</tr>
<tr>
<td>43 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0E-28</td>
<td>Tail completion and sheath stabilizer - P-SSM2</td>
<td>YP_214369</td>
<td>133</td>
</tr>
<tr>
<td>44 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.0E-29</td>
<td>UvoW - P-SSM2</td>
<td>YP_214370</td>
<td>134</td>
</tr>
<tr>
<td>45 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.0E-13</td>
<td>UvoW - P-SSM2</td>
<td>YP_214374</td>
<td>138</td>
</tr>
<tr>
<td>46 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.0E-12</td>
<td>Hypothetical Protein S-PMP113 - S-PM2</td>
<td>YP_195147</td>
<td>139</td>
</tr>
<tr>
<td>47 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0E-45</td>
<td>Late transcription sigma factor - P-SSM2</td>
<td>YP_717814</td>
<td>140</td>
</tr>
<tr>
<td>48 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.0E-82</td>
<td>Endotopic peptidase - P-SSM2</td>
<td>YP_214377</td>
<td>141</td>
</tr>
<tr>
<td>49 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.0E-10</td>
<td>Endotopic peptidase - P-SSM2</td>
<td>YP_214377</td>
<td>141</td>
</tr>
<tr>
<td>50 +</td>
<td>470</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0E-12</td>
<td>Endotopic peptidase - P-SSM2</td>
<td>YP_214377</td>
<td>141</td>
</tr>
</tbody>
</table>

Table 1. Predicted protein-coding genes of S-CRM01.
<table>
<thead>
<tr>
<th>Region 1C</th>
<th>Region 2</th>
<th>Region 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>56 – 98</td>
<td>9.6-12</td>
<td>PAA repeat-containing protein Aminobacterium colombianense DSM 12261</td>
</tr>
<tr>
<td>58 – 302</td>
<td>gp5</td>
<td>2.0-21</td>
</tr>
<tr>
<td>59 – 736</td>
<td>424</td>
<td>1.0-22</td>
</tr>
<tr>
<td>61 – 865</td>
<td>gp61-like</td>
<td>4.0-10</td>
</tr>
<tr>
<td>62 – 61</td>
<td>2.0-09</td>
<td>Similar to base plate hub assembly catalytic gp51 S-PM2p200 – S-PM2</td>
</tr>
<tr>
<td>63 – 235</td>
<td>gp4-like</td>
<td>1.0-43</td>
</tr>
<tr>
<td>64 – 140</td>
<td>gp4</td>
<td>2.0-21</td>
</tr>
<tr>
<td>66 + 114</td>
<td>2.0-20</td>
<td>8.0-24</td>
</tr>
<tr>
<td>67 + 283</td>
<td>gp4-like</td>
<td>1.0-16</td>
</tr>
<tr>
<td>68 + 219</td>
<td>gp53</td>
<td>2.0-18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region 2</th>
<th>Region 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>69 – 302</td>
<td>gp32</td>
</tr>
<tr>
<td>74 – 204</td>
<td>gp59</td>
</tr>
<tr>
<td>75 – 95</td>
<td>gp33-like</td>
</tr>
<tr>
<td>76 – 102</td>
<td>phnH</td>
</tr>
<tr>
<td>77 – 252</td>
<td>3.0-06</td>
</tr>
<tr>
<td>78 – 82</td>
<td>1.0-12</td>
</tr>
<tr>
<td>79 – 61</td>
<td>1.0-09</td>
</tr>
<tr>
<td>80 – 229</td>
<td>2.0-22</td>
</tr>
<tr>
<td>81 – 301</td>
<td>4.0-05</td>
</tr>
<tr>
<td>83 – 281</td>
<td>mnh</td>
</tr>
<tr>
<td>84 – 85</td>
<td>4.0-07</td>
</tr>
<tr>
<td>86 – 81</td>
<td>nirC</td>
</tr>
<tr>
<td>87 – 64</td>
<td>hht03</td>
</tr>
<tr>
<td>88 – 30</td>
<td>2.0-30</td>
</tr>
<tr>
<td>89 – 10</td>
<td>1.0-20</td>
</tr>
<tr>
<td>90 – 364</td>
<td>pabA</td>
</tr>
<tr>
<td>97 – 256</td>
<td>2.0-24</td>
</tr>
<tr>
<td>98 – 124</td>
<td>2.0-10</td>
</tr>
<tr>
<td>101 – 55</td>
<td>2.0-07</td>
</tr>
<tr>
<td>104 – 194</td>
<td>nusG</td>
</tr>
<tr>
<td>105 – 68</td>
<td>1.0-06</td>
</tr>
<tr>
<td>107 – 87</td>
<td>3.0-06</td>
</tr>
<tr>
<td>110 – 126</td>
<td>3.0-03</td>
</tr>
<tr>
<td>111 – 74</td>
<td>1.0-05</td>
</tr>
<tr>
<td>112 – 251</td>
<td>3.0-56</td>
</tr>
<tr>
<td>113 – 186</td>
<td>4.0-24</td>
</tr>
<tr>
<td>114 – 171</td>
<td>4.0-18</td>
</tr>
<tr>
<td>115 – 59</td>
<td>4.0-12</td>
</tr>
<tr>
<td>116 – 362</td>
<td>nirB</td>
</tr>
<tr>
<td>118 – 132</td>
<td>4.0-10</td>
</tr>
<tr>
<td>119 – 771</td>
<td>nirA</td>
</tr>
<tr>
<td>120 – 399</td>
<td>gp51</td>
</tr>
<tr>
<td>122 – 214</td>
<td>2.0-36</td>
</tr>
<tr>
<td>123 – 220</td>
<td>6.0-30</td>
</tr>
<tr>
<td>124 – 129</td>
<td>2.0-51</td>
</tr>
<tr>
<td>125 – 242</td>
<td></td>
</tr>
<tr>
<td>127 – 289</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. cont.

<table>
<thead>
<tr>
<th>ORF No.</th>
<th>Strand</th>
<th>Size (aa)</th>
<th>T4 similarity*</th>
<th>Other genes</th>
<th>Mass spectrometry ID</th>
<th>E-value</th>
<th>Annotation from best hit</th>
<th>Accession number of best hit</th>
<th>T4 gene clusters(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>132</td>
<td>-</td>
<td>398</td>
<td></td>
<td></td>
<td>1.0E-62</td>
<td></td>
<td>Hypothetical protein P700755_P27646 Psychrophilus torquis ATCC 700755</td>
<td>YP_0030097298</td>
<td>71</td>
</tr>
<tr>
<td>136</td>
<td>-</td>
<td>137</td>
<td>mazG</td>
<td></td>
<td>1.0E-46</td>
<td></td>
<td>Helicase - P-SSM2</td>
<td>YP_214420</td>
<td>184</td>
</tr>
<tr>
<td>138</td>
<td>-</td>
<td>459</td>
<td>gp41</td>
<td></td>
<td>6.0E-151</td>
<td></td>
<td>DNA Polymerase - P-SSM2</td>
<td>ACY76597</td>
<td>182</td>
</tr>
<tr>
<td>139</td>
<td>-</td>
<td>193</td>
<td></td>
<td></td>
<td>3.2E-22</td>
<td></td>
<td>2OG-Fe(lII) oxygenase superfamily dioxygenase - P-SSM4</td>
<td>YP_717846</td>
<td>104</td>
</tr>
<tr>
<td>141</td>
<td>-</td>
<td>338</td>
<td>ursX</td>
<td></td>
<td>3.0E-142</td>
<td></td>
<td>DNA -&gt; RNA-like recombination protein - P-SSM2</td>
<td>YP_214417</td>
<td>181</td>
</tr>
<tr>
<td>142</td>
<td>-</td>
<td>838</td>
<td></td>
<td></td>
<td>6.0E-00</td>
<td></td>
<td>DNA Polymerase - P-SSM2</td>
<td>YP_214414</td>
<td>178</td>
</tr>
<tr>
<td>145</td>
<td>-</td>
<td>136</td>
<td>hsp20</td>
<td></td>
<td>6.0E-30</td>
<td></td>
<td>Small heat shock protein - P-SSM2</td>
<td>YP_195165</td>
<td>170</td>
</tr>
<tr>
<td>146</td>
<td>-</td>
<td>135</td>
<td>regA</td>
<td></td>
<td>8.0E-53</td>
<td></td>
<td>RegS translation repressor - P-SSM4</td>
<td>YP_214992</td>
<td>168</td>
</tr>
<tr>
<td>148</td>
<td>-</td>
<td>382</td>
<td></td>
<td></td>
<td>2.0E-70</td>
<td></td>
<td>DNA-cytosine methyltransferase Octobacterium anthropi</td>
<td>YP_001370055</td>
<td>2562</td>
</tr>
<tr>
<td>149</td>
<td>-</td>
<td>232</td>
<td></td>
<td></td>
<td>1.0E-32</td>
<td></td>
<td>Methyltransferase domain-containing protein Bacteroides sp. 2p2p4</td>
<td>ZP_0453186</td>
<td>167</td>
</tr>
<tr>
<td>153</td>
<td>-</td>
<td>124</td>
<td>gp62-like</td>
<td></td>
<td>4.0E-21</td>
<td></td>
<td>Clamp loader subunit - P-SSM2</td>
<td>YP_00397312</td>
<td>167</td>
</tr>
<tr>
<td>155</td>
<td>-</td>
<td>311</td>
<td></td>
<td></td>
<td>1.0E-92</td>
<td></td>
<td>Clasp clamp - P-SSM2</td>
<td>YP_214993</td>
<td>157</td>
</tr>
<tr>
<td>156</td>
<td>-</td>
<td>223</td>
<td>gp45</td>
<td></td>
<td>7.0E-59</td>
<td></td>
<td>Sliding clamp - P-SSM2</td>
<td>YP_214993</td>
<td>153</td>
</tr>
<tr>
<td>161</td>
<td>-</td>
<td>27</td>
<td></td>
<td></td>
<td>6.0E-21</td>
<td></td>
<td>GOSL family Isopse Serratia proteamaculans 568</td>
<td>YP_001478498</td>
<td>Region 3</td>
</tr>
<tr>
<td>166</td>
<td>+</td>
<td>714</td>
<td>cobS</td>
<td></td>
<td>9.0E-97</td>
<td></td>
<td>Peptidase S-PM20119 - S-PM2</td>
<td>YP_195154</td>
<td>146</td>
</tr>
<tr>
<td>167</td>
<td>+</td>
<td>422</td>
<td>cobS</td>
<td></td>
<td>6.0E-27</td>
<td></td>
<td>Hypothetical protein Swi62139 Sphingomonas wittich RW1</td>
<td>YP_001262262</td>
<td>150</td>
</tr>
<tr>
<td>171</td>
<td>+</td>
<td>157</td>
<td></td>
<td></td>
<td>3.2E-12</td>
<td></td>
<td>Hypothetical protein - P-SSM2</td>
<td>YP_195072</td>
<td>150</td>
</tr>
<tr>
<td>172</td>
<td>+</td>
<td>175</td>
<td></td>
<td></td>
<td>7.0E-07</td>
<td></td>
<td>Acetyltransferase Acrvohochotus manna MBIC1017</td>
<td>YP_001520933</td>
<td>150</td>
</tr>
<tr>
<td>176</td>
<td>+</td>
<td>58</td>
<td></td>
<td></td>
<td>1.0E-63</td>
<td></td>
<td>DNA ligase Coeliariviridae</td>
<td>YP_001370060</td>
<td>1627</td>
</tr>
<tr>
<td>185</td>
<td>+</td>
<td>132</td>
<td></td>
<td></td>
<td>3.0E-18</td>
<td></td>
<td>Conserved hypothetical protein Streptomyces clavuligerus ATCC 27064</td>
<td>ZP_0505046</td>
<td>1627</td>
</tr>
<tr>
<td>194</td>
<td>+</td>
<td>288</td>
<td>lig</td>
<td></td>
<td>4.0E-26</td>
<td></td>
<td>Hypothetical protein - P-SSM2</td>
<td>YP_195072</td>
<td>150</td>
</tr>
<tr>
<td>201</td>
<td>+</td>
<td>183</td>
<td></td>
<td></td>
<td>2.0E-35</td>
<td></td>
<td>NHC endonuclease - Cyanothece sp. PCC 7425</td>
<td>YP_002482562</td>
<td>1465</td>
</tr>
<tr>
<td>205</td>
<td>+</td>
<td>157</td>
<td></td>
<td></td>
<td>1.0E-15</td>
<td></td>
<td>gp39.1 hypothetical protein Enterobacteria phage JS10</td>
<td>YP_000923535</td>
<td>395</td>
</tr>
<tr>
<td>206</td>
<td>+</td>
<td>45</td>
<td>gp39.1</td>
<td></td>
<td>9.0E-12</td>
<td></td>
<td>Rodoviridae - P-SSM2</td>
<td>YP_00151301</td>
<td>395</td>
</tr>
<tr>
<td>208</td>
<td>+</td>
<td>162</td>
<td></td>
<td></td>
<td>3.0E-19</td>
<td></td>
<td>S-adenosylmethionine decarboxylase S-PM2</td>
<td>YP_753372</td>
<td>411</td>
</tr>
<tr>
<td>209</td>
<td>+</td>
<td>116</td>
<td>spoD</td>
<td></td>
<td>1.0E-12</td>
<td></td>
<td>Hypothetical protein - P-SSM2</td>
<td>YP_195072</td>
<td>150</td>
</tr>
<tr>
<td>215</td>
<td>+</td>
<td>297</td>
<td>spoT</td>
<td></td>
<td>9.0E-37</td>
<td></td>
<td>Paralys phosphate guanino-3'-pyrophosphohydrolase (spO7)</td>
<td>ACX99173</td>
<td>1803</td>
</tr>
<tr>
<td>220</td>
<td>+</td>
<td>261</td>
<td></td>
<td></td>
<td>8.0E-37</td>
<td></td>
<td>Metallophosphatase - delta proteobacterium LM6-1</td>
<td>ZP_01288066</td>
<td>2146</td>
</tr>
<tr>
<td>230</td>
<td>+</td>
<td>115</td>
<td></td>
<td></td>
<td>2.0E-09</td>
<td></td>
<td>Hypothetical protein S-PM20066 - S-PM2</td>
<td>YP_195090</td>
<td>68</td>
</tr>
<tr>
<td>234</td>
<td>+</td>
<td>184</td>
<td></td>
<td></td>
<td>2.0E-55</td>
<td></td>
<td>Hypothetical protein S-PM20151 - S-PM2</td>
<td>YP_214383</td>
<td>147</td>
</tr>
<tr>
<td>249</td>
<td>+</td>
<td>110</td>
<td></td>
<td></td>
<td>5.0E-22</td>
<td></td>
<td>Hypothetical protein S-PM2016 - S-PM2</td>
<td>YP_195050</td>
<td>66</td>
</tr>
<tr>
<td>255</td>
<td>+</td>
<td>79</td>
<td></td>
<td></td>
<td>1.0E-05</td>
<td></td>
<td>Hypothetical protein S-PM2043 - S-PM2</td>
<td>YP_195077</td>
<td>66</td>
</tr>
<tr>
<td>281</td>
<td>+</td>
<td>54</td>
<td></td>
<td></td>
<td>1.0E-09</td>
<td></td>
<td>Hypothetical protein - P-SSM2</td>
<td>YP_004250202</td>
<td>432</td>
</tr>
<tr>
<td>291</td>
<td>+</td>
<td>201</td>
<td></td>
<td></td>
<td>7.0E-06</td>
<td></td>
<td>Hypothetical protein S-PM2016 - S-PM2</td>
<td>YP_195050</td>
<td>66</td>
</tr>
<tr>
<td>296</td>
<td>+</td>
<td>79</td>
<td></td>
<td></td>
<td>4.0E-07</td>
<td></td>
<td>Hypothetical protein syn39.12 - Syn9</td>
<td>YP_717765</td>
<td>916</td>
</tr>
</tbody>
</table>

ORF number, gene orientation (strand) and protein size (number of amino acids) are noted for predicted protein-coding genes that either have a database match or were detected by mass spectrometry. ORFs were identified with BLASTP run against the NCBI nr database with a threshold E-value of \( \leq 10^{-5} \). The best-hit E-values, homologous gene, and NCBI accession numbers are noted.

a. T4 similarity identifies genes (by T4 name) with BLASTP hits to the phage T4 genome; T4-like genes are homologous to cyanophage genes annotated with T4 gene names.

b. Gene annotation based on sequence homology and colour coded as in Fig. 3: structural genes (including assembly catalyst), yellow; replication, recombination, nucleotide metabolism and gene expression control genes, blue; (marine) cyanophage signature genes (Millard et al., 2009), green; Novel S-CRM01 genes, pink; genes of proteins identified by mass spectrometry, purple.

c. T4 Gene Clusters (T4-GC) numbers are based on sequence homology with a threshold E-value \( \leq 10^{-6} \) from Sullivan and colleagues (2010).
The S-CRM01 genome encodes all six genes proposed as (marine) cyanophage signature genes (Millard et al., 2009): cobS, hli03, hsp20, mazG, phoH and psbA. The psbA gene, encoding the D1 protein of photosystem II, and hli03 (high light inducible) gene are characteristic ‘photosynthetic phage’ genes that are thought to function in augmenting the photosynthetic capacity of infected cells or in protecting against oxidative stress resulting from high light intensities (Lindell et al., 2005; Mann et al., 2005; Clokie et al., 2006). The S-CRM01 genome shows extensive synteny with the marine cyanomyophage genomes, but a 64 kbp part of the syntenic region is in an inverted orientation (Fig. 3.4) with most genes between g56 and g165 expressed from the negative strand. Together with the GC-skew profile (Fig 3.3, inner-most ring; the leading strand in prokaryotic genomes is enriched in G; Lobry, 1996), this is suggestive of bidirectional replication with an origin near the gene inversion boundary (nt 123 K). No such pronounced GC-skew patterns are evident in the marine cyanomyophage genomes (not shown).
Figure 3.4 Extensive synteny between S-CRM01 and marine cyanomyophage. A. Dot plots comparing gene order between S-CRM01 and phage T4 and indicated cyanomyophage genomes. The S-CRM01 genome regions 1–3 (see Fig. 3.3) are indicated by coloured shading and are labelled at right. Syntenic segments I–IV are gp13-gp46 (SCRM01 g29-g55), gp5-td(g59-g80), nrdB-gp45 (g116-g156) and g166-167 (cobS) respectively. B. Diagram indicating the location of the inversion between the S-CRM01 and S-PM2 genomes. Note that the P-SSM2 and S-PM2 genomes have insertions between syntenic segments II and III, and that P-SSM2 and Syn9 genome numbering convention places the regions syntenic with Region 1C at the start of the genomes.

The S-CRM01 genome possesses an unusually large number (182) of ORFs with no significant homologues in the GenBank database (as of 30 August, 2010), representing 61% of predicted ORFs (genes shown in black in Fig. 3.3). This is far higher than the number of unique genes in the marine cyanomyophage genomes of similar size (about 60–100) (Millard et al., 2009; Sullivan et al., 2010).

Based on both gene content and sequence relatedness, S-CRM01 is a member of a discrete clade that encompasses all 17 currently sequenced marine cyanomyophage, but it is the most divergent member of this group (Fig. 3.5). The phylogenetic relationships
among the cyanomyophage were explored using the Hal phylogenomics pipeline (Robbertse et al., 2006), which produces whole genome phylogenies using single-copy protein coding genes. The pipeline can be configured for inclusion of different numbers of orthologous clusters to allow the analysis to be expanded by including genes missing from a few genomes (e.g. Fig. 3.5D; see Experimental procedures). As shown in the consensus tree in Fig. 3.5A, the cyanomyophage (including S-CRM01) form a monophyletic group supported by high bootstrap values. However, the group is quite diverse, with S-CRM01 being the most divergent member and the marine phage partitioning into at least four clades. To examine the possibility of long branch attraction artifacts, the same analysis was run without the T4, Aeh1 and KVP40 genomes, with the overall topology matching that of the consensus tree in Fig. 3.5A. Phylogenetic analysis based on gene content (using a subset of the genomes analysed here) showed similar overall relationships (Millard et al., 2009).
Figure 3.5 Phylogenetic relationship of S-CRM01 to other myophage. A. Consensus tree for the most highly supported topology across all tested values of included non-universal genes (missing data). Values at nodes indicate the average bootstrap value for all trees with that node/the number of trees for which that node occurred out of a total of 27 trees (see Experimental procedures). The marine cyanomyoviruses have been grouped into the clades MC1-4 as indicated. B and C. Consensus trees showing alternative topologies among clades MC1-3 when including 0–10% and 15–40% included non-universal genes respectively. D. The number of orthologous clusters analysed for each phage as a function of % included non-universal genes are indicated and colour-matched.

With different numbers of genes (orthologous clusters) included in the analysis, two additional alternative tree topologies with differences in the relative positions of MC1, MC2 and MC3 were observed (Fig. 5B and C). Analysis of individual gene phylogenies produced a similar range of topologies, although the S-CRM01 branch (typically the longest) was at times placed among the MC1-3 clades. No pattern between tree topology and gene type or location in the genome could be discerned. The variable relationships indicated above suggest high levels of horizontal gene exchange among the cyanomyophage, analogous to that proposed among the *Synechococcus* or *Prochlorococcus* hosts of these phage (Zhaxybayeva et al., 2009).
Evidence for gene exchange at individual loci among the marine cyanophage has been reported by Zeidner and colleagues (2005), Sullivan and colleagues (2006) and Bryan and colleagues (2008). Note that the analysis of Fig. 3.5 does not support the classification of phage based on host (Synechococcus or Prochlorococcus), since both hosts are represented in clades MC2 and MC4 (cf. Sullivan et al., 2010).

A structural gene cluster mostly on the plus strand

The bioinformatically identifiable structural genes are all present in a 72 kbp segment of the genome (Region 1), comprised of two clusters: g1 through g45, covering nts 1–51380 (genome Region 1A), and g59 through g68, covering nts 62104–72062 (Region 1C). Region 1A includes strong synteny with the genomes of marine exoT-even cyanomyophage and phage T4 (Fig. 3.4). This region encodes most of the recognizable structural genes, with all but one gene expressed from the plus strand. Expression is predicted to be dominated by the activity of late promoters, as appropriate for structural protein genes.

Both transcriptional directions are represented in Region 1C, with transcription again predicted to be dominated by late promoters. This region is also syntenic, although in inverted orientation, to the marine cyanomyovirus genomes (Fig. 3.4), but is not syntenic to T4.

As has been observed for marine cyanomyophage, the conserved genes that define synteny are variably interspersed with additional genes. Twenty-three ORFs with no homologues in the GenBank database are located within Region 1; other genes are most similar to phage or bacterial proteins (Table 3.1). Millard and colleagues (2009) have described a hyperplastic region between gp15 and gp18 in S-RSM4 and other cyanomyophage. In S-CRM01 this segment contains only one non-conserved gene, but several genes of varied apparent origins are located upstream of gp13.

Virion proteomics

Mass spectrometry was used to identify phage proteins present in a preparation of S-CRM01 virions purified through a CsCl density gradient. Forty-three proteins were identified with high certainty (Fig. 3.3, Table 3.1). All of these, except two proteins
encoded in Region 3 of the genome, are encoded by genes that are closely associated with late promoters. Most of the identified proteins are encoded in Regions 1A and 1C, emphasizing the clustering of structural protein genes in these parts of the genome. Thirteen of these proteins correspond to structural proteins of T4 (gp or gp-like genes), 13 correspond to genes with BLAST hits to other phage or bacterial genes, while 9 are encoded by genes unique to S-CRM01. These results indicate that the S-CRM01 virion is composed of proteins with a variety of origins: homologues of T4 proteins, homologues of proteins from the related marine cyanomyophage, proteins with closest BLAST hits in other phage or in bacterial genomes, and proteins that have no currently known homologues.

An additional eight proteins encoded by genes not clustered with structural genes were identified by our proteomic study. These genes are located in Regions 1B, 2 and 3 (Fig. 3.3, Table 3.1). Electron microscopy suggested that host material attached to phage baseplates may have been present in the phage preparation made for proteomic analysis. Consequently, it is uncertain whether these proteins are truly virion-associated (structural) proteins or phage-encoded proteins that have been inserted into host structural components. Their identification does prove the expression of the respective genes during viral infection, indicating that these proteins (four of which are unique to S-CRM01) are functionally relevant.

As expected, abundant mass spectrometry signals were registered for gp23 major capsid protein (g44), gp18 contractile tail sheath protein (g36) and gp19 tail tube protein (g37), which are present in multiple copy number in the T4 capsid (Miller et al., 2003a). Abundant signals were also observed for peptides from g9, at 271 kDa the largest protein encoded by the S-CRM01 genome. This huge protein has a predicted strong \( \beta \)-strand character and no identifiable sequence motifs, but has BLAST matches to a wide range of phage proteins (including putative tail protein) and to glycosyl hydrolase bacterial neuraminidase repeat (BNR) proteins. An internal dot plot analysis shows numerous different internal repeat elements 10–20 residues long, mostly repeated only once. These
properties suggest that g9 is a tail fibre gene (Weigele et al., 2007), although we have not observed tail fibres under the electron microscope.

Three proteins with collagen-like triple helix repeat domains were detected by mass spectrometry: g13 with 36 GXY repeats, g20 with 85 GXQ repeats, and g23 with 40 GXQ repeats. Collagen-like proteins have been reported from some phage, participating in spike formation in phage PRD1 (Caldentey et al., 2000) and suggested to be found in tail fibres (Smith et al., 1998; Sullivan et al., 2005). More careful virion characterization will be needed to determine whether S-CRM01 possesses the tail fibres or whiskers that are predicted by the presence of these proteins.

**T4-like non-structural genes predominantly on the minus strand**

A second major region of synteny with the exoT-even marine cyanomyophage genomes is the 51 kbp segment between genes g69 and g165 (Region 2, nts 72 094–123 515) (Figs 3.3 and 3.4; Table 3.1) (cf. Sullivan et al., 2010). These genes, all of which are on the minus strand, include core T4-like genes with roles in DNA replication, recombination and repair (e.g. gp43 DNA polymerase, gp61 primase, gp45 sliding clamp, uvsX), nucleotide metabolism (nrdA, nrdB, nrdC, td thymidylate synthase) and 5 of the 6 marine cyanophage signature genes (phoH, hli03, psbA, mazG, hsp20). Other significant genes also present in Region 2 are: a member of the 2OG-FeII oxygenase superfamily that is present in nearly all marine cyanomyophage, a nusG transcription anti-termination factor homologue (not found in T4-related phage), regA regulator of early gene translation (common to all marine cyanomyo and T4-like phage), and RNase H (rnh, common to T4-like phage but found in only S-PM2 and Syn19 among the marine cyanomyophage).

This segment of the genome is also predicted to be expressed predominantly via late transcription, though some putative early promoters have been identified, notably for the expression of gp45, gp44, gp62 and gp33, all of which are involved in establishing and maintaining T4 late gene transcription (Miller et al., 2003a). The other key gene needed for late transcription, gp55 alternative sigma factor, is located in Region 1B between the two structural gene segments and on the plus strand, and it also appears to be
expressed from an early promoter. The recombination
genes uvsY, uvsW, gp47 and gp46 are also located in Region 1B, all on the plus strand in
another region of synteny with marine cyanomyophage genomes (Figs 3.3 and 3.4).

As in the structural gene region, but even more so, the conserved genes in Region
2 are interspersed with additional genes at multiple sites. Forty-eight (49%) of ORFs in
Region 2 are unique to S-CRM01.

49 kbp (27%) of the genome contains 132 ORFs, 85% of which are unique genes

The remainder of the genome (Region 3, genes g166-g297) possesses no overall
similarity to the genomes of other phage apart from g166 and g167 (Fig. 3.4) and contains
only 20 genes with homology to previously identified ORFs. Fifteen of these have
homologues in at least one T4-like or cyanomyophage. This part of the genome also
contains all but two of the 33 tRNA genes. The significant genes present in this segment
are cobS (g167), DNA ligase (g194), speD (g109), spoT (g215) and HNH endonuclease
(g205). cobS, involved in cobalamin biosynthesis, may support the synthesis of
deoxyribonucleotides (Sullivan et al., 2005) and is one of the marine cyanophage signature
genes (Millard et al., 2009). DNA ligase is common in T4-like phage genomes, though
not found in the marine cyanomyophage; the S-CRM01 DNA ligase gene is most closely
related to Chorella virus and PB1-like myophage ligases. speD encodes S-
adenosylmethionine decarboxylase, which is involved in polyamine synthesis and found
in a few marine cyanomyophage genomes (Mann et al., 2005). spoT, which encodes
ppGpp synthetase and hydrolase, has not previously been recognized in myophage
genomes but is present in the Aeromonas myophage Aeh1, 44RR, PHG25 and PHG31.
HNH endonuclease is possibly a member of a family of homing endonucleases, although
no introns have been detected in the S-CRM01 genome.

All but one of the genes (g186) in Region 3, including the tRNA genes, reside on
the plus strand. Region 3 genes are atypical in a number of ways. Identifiable promoters
are sparse; the G+C content is generally high (Fig. 3.3) and the ORFs are more likely to
have a higher G+C% in the third codon position; the ORFs are shorter and less tightly
spaced. A similarly extensive array of unique genes is also present in the Aeh1 genome.
(Comeau et al., 2007) and to a lesser extent in other myophage, including the marine cyanomyophage (e.g. S-PM2; Mann et al., 2005).

**Genome transcription**

The predicted transcription control signals – start sites of 34 early promoters and 81 late promoters, and coordinates of 17 terminator hairpins. As for T4-like phage such as RB49 (Desplats et al., 2002) and marine cyanomyophage such as S-PM2 (Mann et al., 2005), S-CRM01 gene expression lacks the middle phase of transcription found in T4. Either early or late (in a few instances both) promoters are predicted for genes that are middle-transcribed in T4. Thus, early promoters are predicted upstream of the genes involved in establishing late transcription: gp55 (sigma factor), gp44, gp45, gp46 and gp33 (sliding clamp and clamp loader proteins). DNA replication and repair genes that are either middle or late expressed in T4, such as gp43 (DNA polymerase), gp32 (ssDNA binding protein), gp41 (helicase), uvsX, uvsW, and gp46 and gp47 endonucleases, are associated with putative late promoters. This is also true of nucleotide metabolism genes td, nrdA and nrdC, as well as each of the marine cyanomyophage signature genes except cobS. A few of these genes – rnh, td, phoH, hsp20 – are associated with both predicted early and late promoters. This is also true of some, though not all, tRNA genes.

All identifiable structural genes are associated with putative late promoters, although in several cases (gp8, gp13 & gp14, gp17, gp18 &gp19, gp22, gp5) early promoters have unexpectedly also been predicted with good prediction scores. Transcription patterns will need to be experimentally assessed to verify these predictions.

Late promoters are predicted to fall into two categories, 46 with a CTAAATA core sequence and 35 with an ATAAATA core sequence (Fig. 3.6). ACTAAATA is the most frequent promoter sequence. ATAAATA promoters are most common with T4 (Miller et al., 2003a) and the marine cyanomyophage (Mann et al., 2005; Weigele et al., 2007; Sullivan et al., 2010). The CTAAATA sequence is largely absent from the latter genomes, indicating that these promoters have not merely been overlooked in previous annotations. CTAAATA promoters are unusual among T4-like phage, previously
reported for RB49 (Desplats et al., 2002) and Aeh1 (Nolan et al., 2006). Like RB49, S-CRM01 late promoters are a combination of CTAAATA and ATAAATA types, with no evident specialization within gene groups that might suggest separate regulation. We have also only detected a single late transcription gp55 sigma factor, which may thus have unusually broad promoter recognition. Interestingly, S-CRM01 gp55 is clearly more closely related to marine cyanomyophage gp55 genes (which recognize ATAAATA promoters) than to RB49 or Aeh1 gp55.

![Figure 3.6. Late promoter consensus sequence. Weblogo representation of the sequences of 81 predicted late promoters, indicating the prevalence of both CTAAATA and ATAAATA core sequences. Nucleotide 0 marks the predicted start of transcription based on T4 late promoters (Miller et al., 2003).](image)

**A full set of tRNAs**

The 33 tRNA genes in the S-CRM01 genome (Table 3.2 below) is more than found in some bacteria and includes all of the 20-amino-acid specificities, although there is no initiator tRNA^{16} as found in some phage genomes. KVP40 also has all 20 specificities represented, though some are thought to be pseudogenes that may not be functional (Miller et al., 2003b). The S-PM2 genome contains 24 tRNA genes representing all but the cysteine and phenylalanine specificities, but the other marine cyanomyophage genomes have much smaller subsets of these. tRNA^{16} is the only S-CRM01 tRNA gene with an encoded 3′ CCA terminus, while in S-PM2 five tRNA genes (including tRNA^{16}) include the 3′ CCA. Three tRNA genes with a CAU anticodon are present. Based on features in the anticodon loop (C32 and A38) that are recognized by the TilS enzyme, which modifies C34 in the anticodon to lysidine (Nakanishi et al.,
2009), the \( t2 \) and \( t31 \) genes are proposed to be \( tRNA^{Ile}_2 \) genes, leaving one \( tRNA^{Ile}_3 \) gene \((t22)\). Gene \( t2 \) has multiple additional nucleotides that are proposed to be characteristic of lysidine-containing \( tRNA^{Ile}_2 \) in cyanobacteria (Freyhult et al., 2007).

It cannot be determined without experimentation whether all predicted tRNAs are functional, although key identity elements and conserved features (Giegé et al., 1998) are generally present. Some of the predicted tRNAs do have highly unusual features (refer to tRNA database; Juhling et al., 2009) that may compromise function: e.g. U73 and long variable loop in tRNA\(^{Glu}_{t12}\), A1 in tRNA\(^{Glu}_{t13}\), U73 in tRNA\(^{Ile}_{t31}\), U10:U25 in tRNA\(^{His}_{t21}\) (probable cause for identification as possible pseudogene by tRNAScan-SE).

The \( tRNA^{Ile}_2 \) gene \( t23 \), like other phage \( tRNA^{Ile}_2 \) occurrences (e.g. S-PM2), lacks the additional 5′ residue G-1 that is in most systems critical for histidine identity (Giegé et al., 1998). G-1 would be either replaced with U-1 (adjacent nucleotide in the genome) or lacking in the mature S-CRM01 tRNA\(^{Ile}_2\), a situation that in \( E. \ coli \) allows only partial translational function (Yan and Francklyn, 1994). Another possibility is that G-1 could be added post-transcriptionally by a Thg1-like activity (Heinemann et al., 2010) or that a variant histidyl–tRNA synthetase that does not rely on G-1 is present in host cells (Wang et al., 2007). No Thg1 or histidyl–tRNA synthetase is evident among the S-CRM01 genes. Establishing the functionality of phage tRNA genes is important in determining their role during infection and the selective forces that act on them.

A connection between the suite of S-CRM01 tRNA genes and codon usage in phage ORFs or in those ORFs expected to be expressed late could not be discerned. It is thought that phage tRNAs serve to optimize expression of proteins by allowing more efficient decoding of codons in phage mRNAs that are under-represented in host mRNAs (Bailly-Bechet et al., 2007). In the absence of the host genome sequence, this hypothesis cannot be tested for S-CRM01. However, there must be a different reason for the presence of tRNAs for the two specificities represented by single codons (methionine and tryptophan) or of tRNA genes with unusual features that could be expected to decrease decoding efficacy; perhaps the viral tRNAs provide an advantage by being more flexible in their accommodation to the interactions on the ribosome in a stressed (infected) cell.
In several cyanomyophage and T4-like genomes, some tRNA genes are located in the general region downstream of the *nrdA* and *nrdB* genes. Only two of the 33 tRNA genes in S-CRM01 are found in this location, the remainder being loosely spaced across a 22 kbp segment of Region 3 among mostly genes that are unique to S-CRM01 (Fig. 3.3). The S-CRM01 genome includes homologues of 22 of the 24 tRNA genes from S-PM2, with nucleotide identity ranging from 42% to 81%, with an average of 68%. tRNAs among the MC1 clade (S-PM2, Syn1 and S-RSM4) share average identities of 76–80%, consistent with their close relationship on the basis of protein-coding genes. This would suggest vertical inheritance of tRNAs, but the lack of extensive synteny, wide range in nucleotide identity and wide variety of tRNA gene content present a complex picture of tRNA gene evolution in the cyanomyophage; the MC1 clade members possess 24, 12 and 6 tRNA genes, and the MC2 clade members possess 10, 5 and 1 tRNA genes.

**Noteworthy genes encoded by S-CRM01**

Two predicted genes, *spoT* and *nusG* (indicated in pink in Fig. 3.3), are unusual genes in T4-like phage. SpoT is a pyrophosphohydrolase that in *E. coli* removes the alarmone (p)ppGpp, which accumulates as a result of RelA action in cells when the stringent response is triggered by amino acid starvation (Srivatsan and Wang, 2008). SpoT is also capable of synthesizing (p)ppGpp under certain conditions, such as fatty acid starvation. Through a network of interactions with other proteins, it seems to sense the physiological status of the cell and modulate (p)ppGpp levels (Potrykus and Cashel, 2008; Srivatsan and Wang, 2008). The effect of elevated (p)ppGpp is a downregulation of normal macromolecular synthesis and a switch to gene expression governed by alternative sigma factors, turning gene expression to pathways such as amino acid biosynthesis. The introduction into a cell of a replicating phage can be expected to deplete cellular nutrients, with the risk of inducing elevated (p)ppGpp and shutting off ribosome synthesis, general transcription and replication. Indeed, one may speculate that host cells utilize (p)ppGpp to establish an antiviral state that inhibits viral replication and amplification, perhaps until nucleases can attack the viral nucleic acid. Specific effects on viral gene expression are also possible, as illustrated by the ability of (p)ppGpp to shut
down some lambda phage promoters (Potrykus and Cashel, 2008). It could thus be advantageous for an infecting phage to counter this type of innate defence, and a virally expressed SpoT enzyme could do that by hydrolysing (p)ppGpp or otherwise altering the regulation of (p)ppGpp levels.

The predicted S-CRM01 SpoT possesses most of the key amino acids needed for (p)ppGpp hydrolysis (Hogg et al., 2004), but it is a very small protein, suggesting activity as a (p)ppGpp hydrolase with little or no regulatory control. Phage-encoded spoT has not been previously recognized, although homologous genes are present in the *Aeromonas* phage Aehl, 44RR, PHG25 and PHG31, which were assigned to T4 gene cluster (T4-GC) 1803 by Sullivan and colleagues (2010). All of the marine cyanomyophage genomes possess a mazG gene (as does S-CRM01), which has been postulated to hydrolyse (p)ppGpp and avoid stationary phase conditions in the cell (Bryan et al., 2008; Clokie et al., 2010). However, MazG proteins have varied nucleotide substrate specificities (Galperin et al., 2006) and it has been cautioned that they may not function in (p)ppGpp hydrolysis (Sullivan et al., 2010). In fact, it may be that the broad substrate specificity of MazG (Zhang and Inouye, 2002; Zhang et al., 2003; Lu et al., 2010) involves this protein both in the removal of mutagenic nucleotides that are produced as a result of oxidative stress (Lu et al., 2010) associated with photosynthesis as well as suppression of (p)ppGpp levels.

S-CRM01 may be the first phage genome to contain a putative nusG gene. The NusG protein associates with elongating RNA polymerase to modulate transcription in various ways. Based mainly on *E. coli* studies, NusG is considered a transcription elongation factor because it increases transcription elongation rates (Squires and Zaporojets, 2000; Yakhnin et al., 2008), although in some bacteria the opposite is true (Sevostyanova and Artsimovitch, 2010). In addition, NusG can promote or suppress pausing at different sites (and thereby facilitate attenuation control) (Yakhnin et al., 2008; Sevostyanova and Artsimovitch, 2010), promote transcript release at termination sites (Chalissery et al., 2007), and may exert effects on translation (Squires and Zaporojets, 2000). S-CRM01 may benefit by influencing the elongation and termination
phases of transcription of the phage, or even host, genome through the action of viral NusG. Additionally, Cardinale and colleagues (2008) have shown that NusG functions in concert with Rho to decrease doubling time and prevent expression of the cryptic rac prophage in *E. coli* MG1655. Overexpression of NusG by S-CRM01 may be able to increase the metabolic activity of the infected host and prevent interference or competition for transcriptional machinery from prophage endemic to the host. If such benefits do exist, the absence of identifiable *nusG* from other phage genomes might suggest that unrecognized transcriptional regulators are encoded by other phage.

**Relationship between freshwater and marine phage**

There has not been enough data to derive a clear picture of the genetic relationships between phage in freshwater and marine environments. Some early observations with podophage (Breitbart et al., 2004) and myophage (Short and Suttle, 2005) emphasized the discovery of very similar sequences in the viral populations of the two aquatic environments. For myophage, additional phylogenetic studies with the *gp20, psbA* and *psbD* genes have verified this observation, but have also shown that most (though not all) freshwater sequences map at high resolution to clades that do separate from marine isolates (Wilhelm et al., 2006; Chénard and Suttle, 2008; Sullivan et al., 2008; Wang et al., 2010). S-CRM01 seems to represent both of these scenarios, with the closest currently known *gp20* sequence found in the Atlantic Ocean isolate P-ShM1 (Sullivan et al., 2008) (Fig. 3.7). On the other hand, the *psbA* gene locates to a unique branch between freshwater and marine clades (Fig. 3.8).
Figure 3.7. Phylogenetic tree showing relationship of S-CRM01 gp20 gene to other gp20 sequences from freshwater or marine phage sources. See Fig. 3.9 for tree construction details. Cyanophage whose genomes have been sequenced are labeled by name and with a red hexagon. The group designations are taken from Sullivan et al. (2008). EU715812, closest to S-CRM01, represents phage P-ShM1 (Sullivan et al. (2008).
Figure 3.8. Phylogenetic tree showing relationship of S-CRM01 psbA gene to other psbA sequences from freshwater or marine phage sources. See Fig. 3.9 for tree construction details. Cyanophage whose genomes have been sequenced are labeled by name and with a red hexagon. The group designations are taken from Chénard and Suttle (2008). S-CRM01 sits between two a freshwater clade (EU258991-6 and EU404136) and a marine-dominated clade containing phage S-SSM7.

The complete S-CRM01 genome sequence allows more meaningful consideration of the relationship between related marine and freshwater phage. The data in Fig. 5 show the relationships to be complex. On the basis of synteny (Fig. 3.4A) and protein-coding (Fig. 3.5D) and tRNA gene content, S-CRM01 is most similar to S-PM2, while on the basis of some phylogenomic comparisons, P-SSM2, S-SM2 and S-SSM7 can be considered more similar (Fig. 3.5D). It thus seems likely that there have been multiple
gene exchanges between the S-CRM01 and marine cyanomyophage lineages during evolution. Brackish estuarine waters are likely connections between freshwater and marine environments, where genetic exchanges could occur between phage that principally exist in one or the other aquatic habitat, especially for phage of bacteria such as *Synechococcus*, which have broad distributions. Highly unique phage do seem to exist in freshwater, however, such as the *Microcystis*-infecting Ma-LMM01 (Yoshida et al., 2008), and it will be interesting to learn in the future whether this is the pattern for phage infecting hosts that do not exist in the oceans.
Table 3.2. tRNA genes in the S-CRM01 genome.

<table>
<thead>
<tr>
<th>tRNA gene</th>
<th>Identity</th>
<th>Anticodon</th>
<th>Nt positions</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>Leu</td>
<td>UAA</td>
<td>88617-544</td>
<td>-</td>
</tr>
<tr>
<td>t2</td>
<td>Ile2</td>
<td>CAU</td>
<td>88643-715</td>
<td>-</td>
</tr>
<tr>
<td>t3</td>
<td>Gly</td>
<td>UCC</td>
<td>127547-618</td>
<td>+</td>
</tr>
<tr>
<td>t4</td>
<td>Arg</td>
<td>UCU</td>
<td>127623-697</td>
<td>+</td>
</tr>
<tr>
<td>t5</td>
<td>Lys</td>
<td>UUU</td>
<td>128953-9025</td>
<td>+</td>
</tr>
<tr>
<td>t6</td>
<td>Pro</td>
<td>UGG</td>
<td>129165-236</td>
<td>+</td>
</tr>
<tr>
<td>t7</td>
<td>Tyr</td>
<td>GUA</td>
<td>129535-619</td>
<td>+</td>
</tr>
<tr>
<td>t8</td>
<td>Gly</td>
<td>GCC</td>
<td>130519-590</td>
<td>+</td>
</tr>
<tr>
<td>t9</td>
<td>Asp</td>
<td>GUC</td>
<td>130857-930</td>
<td>+</td>
</tr>
<tr>
<td>t10</td>
<td>Trp</td>
<td>CCA</td>
<td>130985-1055</td>
<td>+</td>
</tr>
<tr>
<td>t11</td>
<td>Asn</td>
<td>GUU</td>
<td>131716-787</td>
<td>+</td>
</tr>
<tr>
<td>t12</td>
<td>Glu</td>
<td>UUC</td>
<td>131866-941</td>
<td>+</td>
</tr>
<tr>
<td>t13</td>
<td>Glu</td>
<td>CUC</td>
<td>132043-115</td>
<td>+</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>------------</td>
<td>---</td>
</tr>
<tr>
<td>t14</td>
<td>Pro</td>
<td>UGG</td>
<td>134674-748</td>
<td>+</td>
</tr>
<tr>
<td>t15</td>
<td>Ile</td>
<td>GAU</td>
<td>134944-5014</td>
<td>+</td>
</tr>
<tr>
<td>t16</td>
<td>Ile</td>
<td>GAU</td>
<td>135055-128</td>
<td>+</td>
</tr>
<tr>
<td>t17</td>
<td>Phe</td>
<td>GAA</td>
<td>135201-274</td>
<td>+</td>
</tr>
<tr>
<td>t18</td>
<td>Ala</td>
<td>UGC</td>
<td>135313-385</td>
<td>+</td>
</tr>
<tr>
<td>t19</td>
<td>Val</td>
<td>GAC</td>
<td>135762-833</td>
<td>+</td>
</tr>
<tr>
<td>t20</td>
<td>Ser</td>
<td>GCU</td>
<td>136579-663</td>
<td>+</td>
</tr>
<tr>
<td>t21</td>
<td>Ser</td>
<td>GGA</td>
<td>136668-759</td>
<td>+</td>
</tr>
<tr>
<td>t22</td>
<td>Met</td>
<td>CAU</td>
<td>139590-660</td>
<td>+</td>
</tr>
<tr>
<td>t23</td>
<td>His</td>
<td>GUG</td>
<td>139662-738²</td>
<td>+</td>
</tr>
<tr>
<td>t24</td>
<td>Thr</td>
<td>GGU</td>
<td>138739-810</td>
<td>+</td>
</tr>
<tr>
<td>t25</td>
<td>Thr</td>
<td>UGU</td>
<td>141268-341</td>
<td>+</td>
</tr>
<tr>
<td>t26</td>
<td>Gln</td>
<td>UUG</td>
<td>141517-587</td>
<td>+</td>
</tr>
</tbody>
</table>
Table 2. (continued)

<table>
<thead>
<tr>
<th>t27</th>
<th>Lys</th>
<th>CUU</th>
<th>141839-912</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>t28</td>
<td>Cys</td>
<td>GCA</td>
<td>142050-123</td>
<td>+</td>
</tr>
<tr>
<td>t29</td>
<td>Leu</td>
<td>CAA</td>
<td>145010-084</td>
<td>+</td>
</tr>
<tr>
<td>t30</td>
<td>Arg</td>
<td>ACG</td>
<td>149226-300</td>
<td>+</td>
</tr>
<tr>
<td>t31</td>
<td>Ile2</td>
<td>CAU</td>
<td>149448-521</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(LysAU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t32</td>
<td>Leu</td>
<td>GAG</td>
<td>149554-627</td>
<td>+</td>
</tr>
<tr>
<td>t33</td>
<td>Val</td>
<td>UAC</td>
<td>149712-783</td>
<td>+</td>
</tr>
</tbody>
</table>

1 Probable modification to lysidine-A-U; 2 Includes 3′ CCA

Multiple genes: Arg (2x): t4, t30; Glu (2x): t12, t13; Gly (2x): t3, t8; Ile (2x): t15, t16; Ile2 (2x): t2, t31; Leu (3x): t1, t29, t32; Lys (2x): t5, t27; Pro (2x): t6, t14; Ser (2x): t20, t21; Thr (2x): t24, t25; Val (2x): t19, t33.

Tightly juxtaposed genes are t3/t4, t20/t21, t22/t23/t24.

Experimental procedures

Sample collection and cyanophage enrichment

Water was collected from the top 0.5 m of Copco Reservoir (mid-channel near the dam wall at latitude 41.979°N and longitude 122.333°W) during a *Microcystis* bloom on the Klamath River in Northern California on 10 September 2008, and transferred to the laboratory in the dark on ice. In order to enrich for cyanophage present in the sample, a 200 ml aliquot of 0.2 µm filtered water was supplemented with 4 ml of 50× BG-11 medium (Sigma-Aldrich, St. Louis, MO, USA) and 20 ml of *Microcystis*-dominated cultures derived from the Klamath River system (August 2007). The enriched water
sample was incubated at 24°C under fluorescent lamps at 10 µE m⁻² s⁻¹ for 10 days. Water quality data relevant to the collected sample are available at http://www.pacificorp.com/es/hydro/hl/kr.html#.

**Cyanophage isolation and amplification**

The enriched culture was treated with chloroform, and cellular material was removed by centrifugation. The 0.2-µm-filtered supernatant (Supor-200; Pall Life Sciences) was ultracentrifuged using a Ti60 rotor (Beckman) at 4°C and 177 520 g, for 90 min. Pellets were resuspended in SM (50 mM Tris-HCl, 100 mM NaCl, 10 mM MgSO₄·7H₂O, pH 7.5) and stored at 4°C. Plaque assays were conducted using BG-11 top agar layered onto BG-11 agar plates that were incubated under the growth conditions described above until plaques were visible (~1 week). Three serial plaque isolations were performed using the Klamath River system culture, which was also used for phage amplification in liquid culture. Phage particles were collected by ultracentrifugation. During the course of phage isolation, it was noticed that the culture characteristics had changed. The identity of the resultant culture (LC16) was examined by PCR amplification of genomic DNA using the cyanobacteria-wide primers CS1F and ULR directed at the internal transcribed spacer of the rRNA operon (ITS), followed by DNA sequencing of the PCR product as described (Bozarth et al., 2010). A single derived sequence indicated culture purity. Comparison of the sequence with the GenBank database identified LC16 as a member of the *Cyanobium gracile* cluster that includes freshwater *Synechococcus* isolates (Ernst et al., 2003; Chen et al., 2006) (Fig. 3.9). The closest known relative has been isolated from Lake Balaton (Hungary), with other related isolates from freshwater sources in Germany and Wisconsin (USA) and brackish or saline sources in California, Baltic Sea (Denmark) and White Sea (Russia).
Figure 3.9. Klamath River watershed isolate LC16 is a member of the *Cyanobium gracile* cluster of *Synechococcus* sp. Consensus phylogenetic tree depicting relationships between ITS sequences from *Synechococcus* isolates, including selected marine *Synechococcus* and *Prochlorococcus* isolates that serve as hosts for characterized cyanomyophage (indicated next to host names). The indicated *Cyanobium gracile* cluster of mostly freshwater isolates follows the definition of Ernst et al. (2003). The tree was constructed by maximum likelihood analysis with a TN93 (Tamura-Nei) substitution model using the PhyML package (Guindon and Gascuel, 2003). Bootstrap values are indicated.
Genome structure analysis

Genome size was estimated using pulsed field gel electrophoresis (PFGE) after genome preparation as described (Lingohr et al., 2008), using a 1.4% agarose gel in a CHEF II PFGE unit (Bio-Rad) set to run at 6 V cm⁻¹ for 18 h with a 0.1 s switch time.

To determine whether the genome was linear, circular or circularly permuted, BAL-31 nuclease digestion of the phage genome was performed as described (Yoshida et al., 2008). Purified phage DNA (200 ng) was incubated with 0.1 U µl⁻¹ BAL-31 nuclease (NEBiolabs) at 30°C for 0, 10, 20, 40 and 60 min. The DNA was then extracted with phenol/chloroform, ethanol precipitated and digested overnight at 37°C with BamHI endonuclease. The restriction products were separated on 0.8% agarose gel and visualized with ethidium bromide. DNA ladder (1 kb) (GeneRuler, Fermentas) and a plasmid (pGEM, Promega) were included as linear and supercoiled circular controls.

Genome sequencing

The genome (500 ng) was sequenced by Roche (Branford, CT, USA) using GS FLX Titanium Sequencing, accumulating 232 kbp at an average coverage of 350X. A first draft of the genome sequence was assembled using Newbler (gsAssembler) software (Roche). Five contigs larger than 5000 bp in length were assembled, amounting to a total of 173.5 kbp. To determine the order of contigs and to fill the gaps, a multiplex PCR approach was used (Tettelin et al., 1999). Briefly, pairs of outward-oriented primers positioned about 100 bp from the two ends of each contig were used in all combinations to direct PCR amplifications using genomic DNA as template. Products visible on gels after 30 cycles were confirmed by PCR using individual primer pairs, and were extracted from agarose gels and submitted for direct Sanger sequencing (Genewiz). Each gap was sequenced in both directions, with additional primers designed when needed.

Contigs and junction sequences were arranged and assembled in Geneious (Biomatters, http://www.geneious.com). The gap sequences could be confirmed using smaller contigs and unassembled reads from the 454 data. The average sequencing
coverage for contigs ranged between 360 and 500 with an average Phred Equivalent (Roche) quality score of 64.0.

**Genome annotation and analysis**

Regions of coding sequence were predicted using Glimmer 3 (Delcher et al., 1999) and Genemark S (Besemer et al., 2001); tRNAs were predicted using tRNAscan-SE (Lowe and Eddy, 1997). Annotated protein coding sequences were determined using a BLASTx search against the NCBI nr database. Annotations were made based on an E-value cut-off < 1e-5. Phage-associated ORFs were compared with a custom database of T4-like cyanophage using a BLASTx search. Genome annotation was curated in both Geneious and Artemis (Rutherford et al., 2000). Early promoters were predicted in regions upstream of ORFs by similarity to putative S-PM2 early promoters (Mann et al., 2005) and using BPROM (LDF>5; Softberry, Mount Kisco, NY, USA). Putative terminator sequences were identified by TransTerm (http://uther.otago.ac.nz). The gene-annotated S-CRM01 sequence is available under GenBank accession HQ615693.

Phylogenomic analysis was carried out using the Hal pipeline (http://aftol.org/pages/Halweb3.htm; http://sourceforge.net/projects/bio-hal/), which consists of a set of Perl scripts that automates a series of phylogenomic analyses using existing software and sequence analysis programs (Robbertse et al., 2006); analysis was executed on a 64-bit Linux cluster operating Red Hat Linux 3.2.3, Linux version 2.4.21. The proteins encoded by 21 phage genomes were analysed at 9 levels of ‘missing data’ (inclusion in the analysis of genes/orthologous clusters not present in all phage), in each case using 3 gap removal methodologies (complete gap removal and liberal and conservative gap-removal with GBlocks), resulting in a total of 27 phylogenetic trees. Inclusion of genes missing from a few taxa can improve phylogenetic analyses by increasing the number of genes analysed (Wiens, 2006).

**Transmission electron microscopy**

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

**Mass spectrometry-based proteomics**

Phage was purified using a CsCl density gradient. Phage-associated proteins were prepared for SDS-PAGE by boiling for 2 min in SDS-PAGE protein sample loading buffer and separated on a 8–16% gradient gel (Bio-Rad). The sample lane was cut into 11 sections, dehydrated with 50% acetonitrile in 50 mM NH$_4$HCO$_3$, and dried in a SpeedVac. Reduction and alkylation with DTT and iodoacetamide, and in-gel trypsin digestion were performed as described in the Protease-Max (Promega, Madison, WI, USA) manual. A blank section of the gel was processed and included in all subsequent analyses.

LC-MS/MS analyses of the extracted peptides were performed on a LTQ-FT Ultra mass spectrometer (Thermo) with an IonMax ion source. The mass spectrometer was coupled to a nanoAcquity Ultra performance LC system (Waters) equipped with a Michrom Peptide CapTrap and a C$_{18}$ column (Agilent Zorbax 300SB-C18, 250 × 0.3 mm, 5 µm). A binary gradient system was used consisting of solvent A, 0.1% aqueous formic acid and solvent B, acetonitrile containing 0.1% formic acid. Peptides were trapped and washed with 1% solvent B for 3 min. Peptide separation was achieved using a linear gradient from 10% B to 30% B at a flow rate of 4 µl min$^{-1}$ over 35 min.

For the LC-MS/MS analysis, the LTQ-FT mass spectrometer was operated in a data-dependent mode. A full FT-MS scan (m/z 350–2000) was alternated with collision-induced dissociation (CID) MS/MS scans of the 5 most abundant doubly or triply charged precursor ions. As the survey scan was acquired in the ICR cell, the CID experiments were performed in the linear ion trap where precursor ions were isolated and subjected to CID in parallel with the completion of the full FT-MS scan. CID was performed with helium gas at a normalized collision energy 35% and activation time of 30 ms. Automated gain control (AGC) was used to accumulate sufficient precursor ions (target value, 5 × 10$^5$/micro scan; maximum fill time 0.2 s). Dynamic exclusion was used with a repeat count of 1 and exclusion duration of 60 s. Data acquisition was controlled by Xcalibur (version 2.0.5) software (Thermo).
For the sequence database search, Thermo RAW data files were processed with Proteome Discoverer v1.0 using default parameters except for a S/N threshold setting of 10. A Mascot (v2.2.04) search of a phage-encoded protein database (299 sequences; 54263 residues) was launched from Proteome Discoverer with the following parameters: the digestion enzyme was set to Trypsin/P and two missed cleavage sites were allowed. The precursor ion mass tolerance was set to 10 p.p.m., while fragment ion tolerance of 0.8 Da was used. Dynamic modifications included carboxymethyl (+57.0214 Da) for Cys and oxidation (+15.9994 Da) for Met.

**PCR to detect S-CRM01 in the Klamath system**

Enriched phage fractions were prepared (see above) from samples collected from the Klamath River system during August–October 2009. Multiple plaques isolated on LC16 top agar plates from each sample were tested for the presence of S-CRM01 by PCR analysis using two primer pairs directed at g44 (gp23) and g34 (Ma-LMM01-like hypothetical protein). In all cases, reactions were either positive or negative for both primer pairs.

The primer pair for detecting g34 was CRM01-g34(F) 5′-GTCAAATAGAATCCAGGATGAATTA and CRM01-g34(R) 5′-TACCATAGTCTCCACCCTTTC. The primer pair detecting g44 was CRM01-g44(F) 5′-GACGTATGTGGCGTTCAGCCAATGA and CRM01-g44(R) 5′-CGGTTGATTCTGCAAGGATTTC. PCR reactions used High Fidelity Platinum Taq polymerase (Invitrogen) in the provided buffer with 0.2 µM of each primer, 0.2 mM dNTPs and 2.5 mM MgSO₄; after initial denaturation, 35 cycles of 0.5 min at 94°C, 0.5 min at 52°C, and 1 min at 68°C were run. These primers were designed to be specific for S-CRM01, avoiding amplification from known related cyanomyophage genomes; detection of S-CRM01 was scored as positive only when products of the expected size were amplified with both primer pairs.
Acknowledgements

We gratefully acknowledge the assistance of Susan Corum of the Karuk Tribe of California, Richard Raymond and PacifiCorp Inc., and Mary Lindenberg and Tamara Wood of the US Geological Survey for sample collections from the Klamath valley. We thank Elizabeth Kutter for advice on initial experiments, Josh Powell and Theresa Sawyer for assistance with electron microscopy, Andrea McHugh for assisting in the laboratory, Michael Schwalbach for assistance with PFGE, Joey Spatafora for discussions, Susan Golden and Roger Ely for providing cultures, and Roger Hendrix for comments on the manuscript. The OSU mass spectrometry facility and services core is supported in part by NIH/NIEHS grant P30 ES000210. This work was supported by Oregon Sea Grant award NA060AR4170010 and by Oregon State University Agricultural Experiment Station.
References


Lingohr, E., Frost, S., and Johnson, R. (2008) Determination of bacteriophage genome size by Pulsed-Field Gel Electrophoresis In *Bacteriophage: Methods and*


4. General Conclusion

Cyanobacterial blooms are an increasingly prevalent phenomenon. Considering the changing global climate and our increasing freshwater usage, blooms will only become more common and more problematic (Paerl and Huisman, 2009). As they arise, it will be important to understand the genetic composition of a bloom system in order to track and predict the toxigenic populations. Using current bioinformatic techniques part of our research focused on trends in genotypes of toxigenic bloom-former *Microcystis aeruginosa* in the Copco Reservoir of the Klamath River.

Viruses are an abundant and influential part of aquatic environments. To further characterize the population of freshwater bloom systems, our research focused on a phage infecting *Synechococcus*, a model cyanobacterium, isolated from the Klamath River. The research compared all of the fully sequenced, well-researched marine cyanomyophage to S-CRM01 – the phage isolated from freshwater *Synechococcus*. In its entirety, our research derived a clear conceptual understanding of a phage that infects freshwater cyanobacteria and its relation to marine cyanomyophage.

Bioinformatics was integral to the research presented in this thesis. The use of software, scripts and visualization techniques allowed us to compare the nuances of genetic structures from cyanobacteria and phage. In Chapter 2, the sequences obtained were analyzed through Perl scripts, visualization software and statistics packages. In Chapter 3 a draft genome was created using these tools to help in the final construction of the genome. Once sequenced, Perl scripts helped annotate and describe the genomic sequence. To compare the multiple genomes available, the HAL phylogenomic pipeline (see chapter 3) was implemented. HAL produced a concatenated comparison of multiple genes from all cyanomyophage genomes (Robbertse *et al.*, 2011).
**Microcystis**

*Microcystis* blooms are widespread toxigenic events (Yoshida *et al.*, 2008; Moisander *et al.*, 2009; Rinta-Kanto *et al.*, 2009; Wang *et al.*, 2010). The organism *Microcystis* is well studied and one aspect of current research is the population dynamics of bloom events. The research described in Chapter 2 was aimed at an annually occurring cyanobacterial bloom event in the Copco Reservoir on the Klamath River system. Focusing on the sampled population of various surface scums, current bioinformatic methods were used to sort and represent genotypic data. These data were used to depict the population dynamics occurring within a *Microcystis* bloom.

To characterize the genetics of *Microcystis*, the focus was on two DNA targets and obtaining quantitative data on the presence of the *mcy* operon that is transcribed to express multiple microcystin synthetase proteins. This approach has allowed us to analyze the dominant population of the bloom and see changes in relative toxigenicity. Using clone libraries to obtain 600-800 bp sequences from the ITS and the *cpcBA* regions single base pair changes were compared. In both gene targets the population was dominated by abundant sequences as seen in previous freshwater research projects. The most abundant sequences at different time points were distinct enough to be separated into genotype groups. This was useful in presenting the data and recognizing the inferred population shift. The genotype groups have great potential for the future of monitoring or researching a particular bloom. As the studied population is shown to only have a few groups in both the *cpcBA* and ITS loci, it is possible to target research of these groups with specific PCR or genotyping methods (Bozarth *et al.*, 2010). This approach could be used in monitoring toxigenic populations or emerging trends over a season. Genotype groups were depicted in a unique way to help visualize the seasonal trend seen in our study.

The presentation of the sequences in dendrograms can often be confusing and uninformative. This study was aimed at describing single base pair relationships between
sequences and interpreting those relationships. To describe both the relationship between sequences and the most abundant sequence, maximum parsimony network diagrams were constructed. The diagrams showed the most abundant sequence of any group as the center and sequences with nucleotide differences extending outward. As the population shifted, the most abundant sequence groups became significantly different as represented by length in the network diagrams. Presentation of data in this unique form allows for easy recognition of population shifts that occur in *Microcystis* blooms and could be applied to other microbial communities.

The use of clone libraries as the method of population surveying was adequate for our study but is not without drawbacks. The quality and length of sequence works well for population surveys and allows for in depth and even functional analysis not always possible with other methods such as DGGE and T-RFLP. The drawbacks come from the time and work involved with making clone libraries ultimately leading to fewer amplicons. Fewer amplicons can mean insufficient sampling of a population, a problem that obscures the actual diversity in the environment. The recent advent of pyrosequencing, sequencing by ligation, etc. can provide high quality and abundant sequence data.

High throughput technologies such as pyrosequencing yield numerous reads and have been used to determine bacterial diversity in natural populations (Acosta-Martinez *et al.*, 2008; McKenna *et al.*, 2008; Lauber *et al.*, 2009). Population studies focusing on cyanobacterial blooms can take advantage of these advancements to obtain large sequence inventories of bloom events elucidating most if not all of the genotypes present. Research such as this would benefit from large data sets as they take into account a greater portion of the population. The costs of high throughput sequencing technology are lower per sequence, although the upfront cost of the technology is quite expensive. High throughput technologies have not been used in *Microcystis* studies to determine toxigenicity and non-toxigenicity but high throughput methods could work similarly to clone library studies but with smaller sequence length.
Cyanophage

Viruses that infect cyanobacteria can have a profound effect on the genetics of the host. Accompanying the viral infection of a cyanobacterium and subsequent decline of a bloom, cyanophage have been observed to carry genes used in host metabolism and functions such as photosynthesis. Phage are a catalyst, if not cause, of bloom fluctuation and cyanobacterial recombination. Studies focusing only on the cyanobacterial population are missing potential reservoirs of genetic information by not including phage.

Chapter 3 describes research that aimed to characterize phage infecting freshwater cyanobacterial blooms. The first step in this study was to isolate phage from a freshwater cyanobacterial bloom. After isolation of a phage (designated as S-CRM01) obtained from the culture of LC-16 (Synechococcus), the phage genome was sequenced and found to be similar to that of myophage infecting marine cyanobacteria. Specifically, S-CRM01 shares synteny and gene content with marine myophage whose genomes had been sequenced. For example, the majority of structural genes of S-CRM01 and other marine myophage were found to be in the same order. Furthermore, S-CRM01 and 17 marine cyanomyophage share key genes including T4-like structural genes and cyanophage signature genes such as psbA. Based on these genetic details S-CRM01 has been identified as a ‘photosynthetic’ cyanomyophage and the first freshwater ‘photosynthetic’ cyanomyophage to be fully sequenced (Dreher et al., 2011).

In order to better understand the relationship between marine and freshwater cyanophage, the entire gene content of all sequenced cyanomyophage was analyzed and compared. Using the HAL phylogenomic pipeline, single-copy protein coding genes were compared to produce whole genome phylogenies of the cyanomyophage (Robbertse et al., 2011). S-CRM01, the freshwater phage, was found to be genetically distinct from the marine myophage (Dreher et al., 2011). Whole genome analysis is a novel tool for cyanomyophage and provides a superior depiction of genetic relationships by comparing all available genetic information for individual phage.
The use of phylogenomics will be essential in future studies of cyanomyophage. Only recently have whole genome analyses become accessible through the use of high throughput genome sequencing. This provides a more detailed account of the sampled population due to the increased amount of genetic information. Comparisons of entire genomes can be performed using gene content, gene order and nucleotide similarity. Accessing the genome of phage leads to a better understanding of their ability to infect host cells, and discovery of novel genes that affect the infection cycle e.g. *psbA* and *phoH*.

Future research on cyanophage would be approached similar to the methods discussed in Chapter 3. The phage population of freshwater blooms is largely unidentified but could potentially be a source of genetic diversity. Host genes found in S-CRM01 and marine myophage indicate a close relationship between phage and host. Further examination of this relationship will illuminate phage-host interaction and events leading to bloom decline.
References


