Genomic characterization of viruses infecting freshwater polar cyanobacteria

by

Caroline Chénard

B.Sc (Hon), Dalhousie University 2004

M.Sc., The University of British Columbia, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Oceanography)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2014

Abstract

There is wide recognition that cyanobacteria are major primary producers in polar freshwater regions. Filamentous cyanobacteria are commonly found in benthic mats and biofilms at the bottom of lakes, ponds and streams, while picocyanobacteria dominate the planktonic communities of many polar lakes. However, no representative viruses infecting this group of organisms have been characterized. This dissertation, which is a culmination of experiments and genomic and metagenomic analyses, presents the first characterization of viruses infecting freshwater polar cyanobacteria and the discovery of previously unknown groups of viruses. First, I isolated and genetically characterized a polar freshwater cyanophage (S-EIV1) that represents a new evolutionary lineage of bacteriophages that are globally widespread and abundant. Second, I described a new group of viruses (Cyanophage A-1 and Cyanophage N-1) infecting freshwater filamentous cyanobacteria that contain a distinct DNA polymerase. Third, during genomic analysis of Cyanophage N-1, I identified a DNA repeat region similar to a Clustered Regularly Interspaced Short Palindromic (CRISPR) array. The CRISPR array had direct repeats with high similarity to those commonly found in filamentous cyanobacteria. I showed that the viral-encoded CRISPR was transcribed and have the potential be viral-mediated transferred to its host. Finally, DNA-stable isotope probing (DNA-SIP) was used to recover and sequence viruses infecting primary producers in a polar cyanobacterial mat. Arctic freshwater systems are some of the most threatened environments because of rapid climate change, and viruses encompass the greatest genetic and biological diversity on Earth. This work presents previously unknown groups of viruses and a newly discovered virus-host system that provide new tools for investigating host-virus interactions and examining arctic viral diversity.

Preface

This statement certifies that the work presented in this thesis was conceived, conducted and written by Caroline Chénard. As my research advisor, Dr. Curtis A Suttle was involved in all aspects of this work, including conceptualization of the experiment and critical review of the thesis and manuscripts.

In Chapter 2, I performed the sample collection, isolated the cyanophage, carried out all the laboratory work, data analysis and wrote the manuscript. MiSeq Illumina sequencing was conducted at McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada. The *Synechoccocus* host was provided by Dr. Warwick Vincent (Laval University) and TEM picture was performed by Amy Chan (Research Scientist in the Suttle Lab) at the UBC Bioimaging Facility.

A version of Chapter 2 is currently under review:

> Chénard C, Chan AM, Vincent WF & Suttle CA. Polar freshwater cyanophage S-EIV1 represents a new evolutionary lineage of bacteriophages that is globally widespread and abundant. (*In review*)

In Chapter 3, I carried out the laboratory work, data analysis and wrote the manuscript. Library constructions, 454 sequencing and contigs assembly were conducted at McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada. The Nostoc PCC7210 and the cyanophages A-1 and N-1 were obtained from the American Type Culture Collection.

In Chapter 4, I carried out the laboratory work, data analysis and wrote the manuscript. Dr Jennifer Wirth (Post-Doctoral Fellow in the Suttle Lab) also helped with conceptualization of the experiment and gave critical advice.

In Chapter 5, I performed the sample collection, carried out all the laboratory work, data analysis and wrote the manuscript. MiSeq Illumina sequencing was conducted at McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada. Experimental design in the field was done in collaboration with Dr. Warwick Vincent (Laval University), and Dr. Anna D. Jungblut (Post-Doctoral Fellow in the Vincent Lab).

Table of contents

List of tables

List of figures

List of abbreviations and symbols

% ..... Percent

~ ..... Approximately

AT-rich..... Adenine- Thymine rich

bp ..... Base pair

BLAST..... Basic local alignment search tool

BLASTp..... Basic local alignment search tool protein search

CAMERA ...Community cyberinfrastruture for advanced microbial ecology research & analysis

CASCADE..... CRISPR-associated complex for antiviral defense

Cas genes ..... CRISPR-associated genes

CRISPR..... Clustered regularly interspaced short palindromic

crRNA..... CRISPR RNA

COG ..... Clusters of orthologous group

d..... Day

DCM ..... Deep chlorophyll maximum

DCTP deaminase ..... Deoxycytidine triphosphate deaminase

DGGE ..... Denaturing gradient gel electrophoresis

DNA..... Deoxyribonucleic acid

DNase..... Deoxyribonuclease

DNApolA..... DNA polymerase A

DNApolB..... DNA polymerase B

dNTP ..... Deoxyribonucleotide triphosphate

dTTP ..... Deoxythymidine triphosphate

dTMP ..... Thymidine monophosphate

DPS ..... DNA-binding proteins from starved cells

DR ..... Direct Repeat

DUF ..... Domain of unknown function

EDTA ..... Ethylenediaminetetraacetic acid

e-value ..... Expectation value

g ..... Gram

G+C ..... Guanine+cytosine

GC50 ..... Glass fiber filter grade GC50 (1.2-µm pore-size)

gp ..... Gene predicted

GOS ..... Global Ocean Survey

h ..... Hour

HGT ..... Horizontal gene transfer

kb ..... Kilobase

kDa-MW ..... Kilodalton- Molecular Weight

kV ..... Kilovolt

l ..... Liter

M ..... Molar

ml ..... Millitre

M.O.I ..... Multiplicity of infection

mM ..... Millimolar

ML.....Maximum likelihood

mg ..... Milligram

min .....Minute

Mt..... Megaton

MWCO .....Molecular weight cut off

nr ..... non-redundant

°C .....degree Celsius

ORF..... Open reading frame

PC.....Phycocyanin

PCCC ..... Polar Cyanobacteria Culture Collection

PCR..... Polymerase chain reaction

PE.....Phycoerythrin

PEG.....Polyethylene glycol

PFGE..... Pulse field gel electrophoresis

purM .....Phosphoribosylaminoimidazole synthetase

rpm..... Revolution per minute

RAMPs..... Repeat-Associated Mysterious Proteins

RT-PCR.....Reverse transcriptase- polymerase chain reaction

RAxML..... Randomized axelerated maximum likelihood

RNA ..... Ribonucleic acid

RNAse .....Ribonuclease

rRNA.....Ribosomal RNA

rlpA ..... Rare lipoprotein A

SDS-PAGE ..... Polyacrylamide gel electrophoresis

SIP ..... Stable Isotope Probing

TEM ..... Transmission electron microscopy

Tm ..... Melting temperature

TBE ..... Tris-borate-EDTA

terL ..... Terminase large subunit

tRNA ..... Transfer RNA

x g ..... Times gravity

UV ..... Ultraviolet

$\mu Em^{-2}s^{-1}$ ..... Micro Einsteins per square meter per second

VC ..... Virus concentrate

V ..... Volt

v/v ..... volume per volume

µL ..... Microlitre

µm ..... Micrometer

Acknowledgements

I would like to acknowledge my advisor, Dr. Curtis Suttle, who accepted me in his lab to work on this challenging and interesting project. I have learnt a lot from his mentorship and I am thankful for his guidance, expertise and support. I am also grateful to the members of my committee meeting, Dr. Steven Hallam and Dr Maite Maldonado as well as Dr. Warwick Vincent for their expertise and guidance.

I would also like to thank my wonderful labmates, past and present, especially Renat Aldesin, Amy Chan, Christina Charlesworth, Cheryl Chow, Jessie Clasen, Ricardo Cruz, Jan Finke, Matthias Fischer, Manuela Gimenes, Jessica Labonté, Julia Gustavsen, Tyler Nelson, Jérôme Payet, Emma Shelford, Alvin Tian, Marli Vlok, Danielle Winget, and Jennifer Wirth.
Many thanks to Julie Veillette, Dermot Antoniades, Anne Jungblut, Denis Sarrazin and Sophie Charvet for field assistance along with scientific and technical advice. I was very fortunate to have shared such unforgettable experience and memories of the north with them.

I would also like to acknowledge Parks Canada, the Polar Continental Shelf Project (PCSP), and the Northern Scientific Training Program (NSTP) for logistic and infrastructure support. I am also grateful to API-IPY program MERGE, the Fond Québécois de la Recherche et des Technologies (FQRNT), UBC University Graduate Fellowship, and BRITE for their financial support.

Last but not least, I am deeply grateful towards my friends and my paddling Ohana whom have been a source of distraction and understanding during the difficult times. Un dernier remerciement mais non le moindre à ma famille et Xavier pour m'avoir soutenu dans les moments difficiles et d'avoir été ma source de motivation. Merci.

À mes parents, pour leur amour et leur support

Chapter 1: Introduction

Inland waters of the Canadian High Arctic are habitats for aquatic life and important in biogeochemical processes. Northern freshwater systems include a spectrum of aquatic environments such as meltwaters, hypersaline ponds, ice-shelf pools, and oligotrophic, epishelf and meromictic lakes. They are subject to extreme conditions that include dramatic variations in light availability (from total darkness to continuous light) and temperature, a long freezing period, and a short growing season. Polar inland systems, once thought to have low biodiversity, are now believed to contain remarkable diversity which is mostly microbial(1–4). Consequently, they are of particular interest for examining microbial distribution and diversity, biogeochemical processes in microbial consortia, and microbial strategies for survival in cold environments (5, 6)

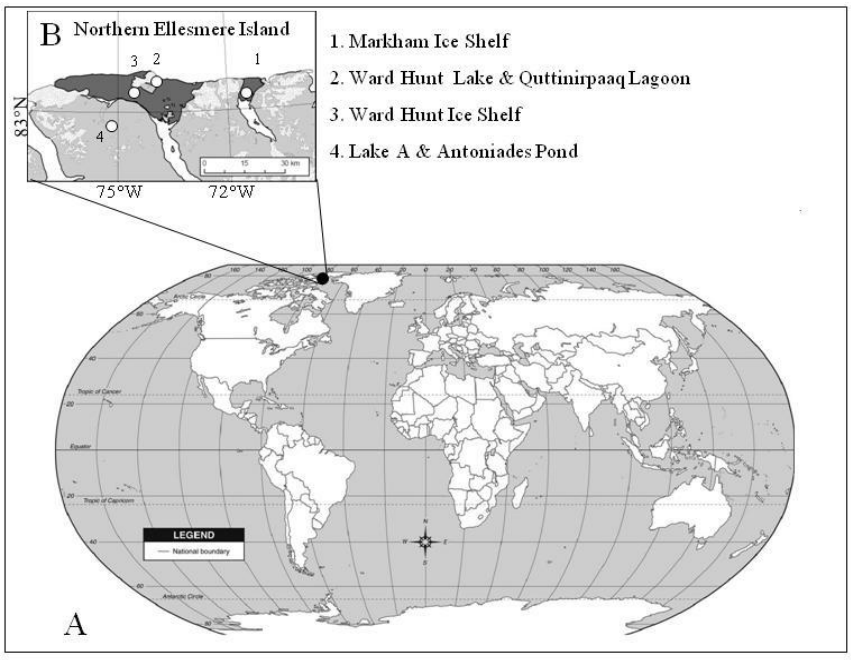


Figure 1-1. World map showing sampling sites.
(A) Location of study site in the Canadian High Arctic. (B) Detailed map showing the sampling locations along the northern coastline of Ellesmere Island, Canadian High Arctic (7)**.**

Unlike temperate inland waters, Arctic freshwaters are strongly dependent on the cryosphere (the region of the Earth where the surface is perennially frozen), and are very vulnerable to climate change. For example, the Ellesmere Ice Shelf on Northern Ellesmere Island (Nunavut) that was once a continuous shelf of thousands of square kilometres now consists of 4 main ice shelves (Ward Hunt, Ayles, Milne and Serson) with a total area of only a few hundred square kilometres (8) (Figure 1.1). Increase in global temperature is believed to be accelerating the fragmentation and loss of the remnant ice shelves and further changing polar environments. Break-up of the Ward Hunt Ice Shelf in 2003 led to the drainage of an epishelf lake which resulted in the loss of unique microbial communities (9). High temperatures in the summer of 2008 also caused the Markham Ice Shelf to break away and drift into the Arctic Ocean, destroying the uncharacterized associated microbial communities. Not only are polar aquatic ecosystems subject to habitat loss, but are also likely to be affected by sudden changes in light and temperature conditions as well as seasonal variation.

Cyanobacteria are common throughout polar regions, where they are major members of planktonic communities in lakes as well as forming benthic mats and biofilms at the bottom of lakes, ponds and streams. The diversity of the polar cyanobacteria is still an ongoing debate. While some studies suggest that several cyanobacteria are endemic to polar regions (10) others show that cyanobacterial isolates from both the Arctic and Antarctic are similar to those isolated from temperate regions (7, 11, 12). However, there is limited knowledge about the diversity of viruses infecting polar cyanobacteria. This dissertation presents the genetic characterization of viruses infecting polar cyanobacteria representatives. Hence, this introductory chapter provides a general description of cyanobacterial communities from polar freshwaters and briefly gives an

overview of what is known about viruses infecting cyanobacteria and states the objectives of this dissertation.

1.1 Cyanobacterial communities in polar inland waters

Cyanobacteria are a major component of microbial communities in polar inland aquatic systems. Given their ability to survive at low temperatures, they are often the dominant phototrophs in polar freshwaters. Generally, they are cold-tolerant microorganisms that grow optimally between 15 and 20°C (psychrotrophs) rather than cold-adapted microorganisms with optimal growth below 15°C (psychrophiles). Picocyanobacteria dominate the planktonic communities of many polar lakes, while filamentous cyanobacteria are commonly found in the pigmented microbial mats of glaciers, meltwater and ice-capped lakes (13).

1.1.1 Picocyanobacteria

Picocyanobacteria are the most abundant pelagic photosynthetic cells in polar inland waters. Picocyanobacteria are chroococcoid cells <3 µm in diameter. They are the most abundant members of the phytoplankton communities in lakes situated in the Vestfold Hills, Antarctica (14) and the northern coastline of the High Arctic (2). Picocyanobacterial concentrations have been found to be 10^4 cells ml^{-1} in a deep ice-covered lake on Ellesmere Island (15), and have even attained abundances of $8x10^6$ cells ml^{-1} in the saline lakes of the Vestfold Hills (16). Their high surface area:volume ratio allows high nutrient uptake efficiency and thus they are well suited to oligotrophic waters such as those found in high latitude lakes and rivers (13).

Polar picocyanobacteria usually belong to two genera, *Synechococcus* and *Cyanobium*. They are genetically diverse and *Synechococcus*-like in appearance (17). Phylogenetic analysis of 16S rRNA on samples from different Arctic lakes and fjords demonstrated a low diversity in Arctic picocyanobacterial communities (2). The 16S rRNA sequences from that study were

primarily from *Synechococcus* spp. and formed two closely related groups which were predominant in brackish and fresh waters. Polar picocyanobacteria are either red or green depending on whether cells have phycoerythrin (PE) or phycocyanin (PC) as their major light harvesting pigment (18). Pigmentation influences their distribution, with PE-rich cells dominating oligotrophic waters in which green and blue light deeply penetrate (19–21), and PC-rich cells dominant in more turbid water in which red light prevails (22). However, PC-rich and PE-rich cells can coexist in water of intermediate colouration including in coastal seas and many freshwater lakes (22, 23) In subarctic lakes, picocyanobacterial communities are a mixture of PC-rich and PE-rich strains with PC-rich strains being most abundant, whereas PE-rich strains usually dominate the ultra-oligotrophic High Arctic lakes (24).

1.1.2 Mat-forming species

The most abundant growth of polar cyanobacteria occurs as benthic mats and biofilms (i.e. microbial mats). Microbial mats from the High Arctic are dominated by cyanobacteria filamentous oscillatorian cyanobacteria such as *Phormidium*, *Leptolyngbya* and *Pseudanabaena* and nitrogen-fixing groups such as *Nostoc* (7). They secrete a mucilaginous matrix of mucopolysaccharides and proteins that bind with sediment particles to form cohesive mats. The formation of mats reduces physiological stresses caused by desiccation and freezing. The mats also include eukaryotic algae, in particular chlorophytes and diatoms, as well as a rich community of heterotrophic bacteria dominated by the phyla *Proteobacteria*, *Bacteroidetes* and *Actinobacteria* (3). In addition, rotifers, tardigrades, turbellarian worms, and viruses are found within these microbial mats (15).

Microbial mats are an important strategy for survival in polar freshwaters, where they are constantly exposed to low temperature, high fluctuations in salinity, and high exposure to UV

radiation (25). These mats can be found in shallow thermokarst (eroded permafrost) lakes (26, 27) in rivers and streams (28, 29), ice-covered lakes (30–32), glaciers (10) and ice-shelf ponds in the Antarctic and Arctic (5, 7, 25, 33, 34). Microbial mat communities were recently found to be the dominant biomass of the Ward Hunt and Markham ice shelves (25). Microbial mat communities usually mediate most of the primary production, organic and inorganic nutrient cycling, and chemical transformations in polar systems (35). Microbial mats comprised up to 80% of the total production in an Arctic lake (36), and microbial mat primary production was ten times higher than planktonic production in a permanently ice-covered Antarctic lake (32). Although these mat communities can have higher biomass than phytoplankton, their photosynthetic rates compared to their biomass are generally slower than those of planktonic communities (37). They usually retain a large over-wintering biomass and slowly grow during the brief season of liquid water availability (38). Microbial mats are less widespread in lower latitudes, where their relatively low growth rates make them vulnerable to grazing, and less competitive in contrast to other benthic species (35). However, they can be found in some wetland systems, especially in alkaline environments (13).

The cyanobacterial diversity present in polar microbial mats is still a subject of ongoing debate (38). A few studies suggest that several cyanobacterial strains are endemic to Antarctica (10, 39), while others suggest that these endemic cyanobacterial strains are globally distributed in the cold environment (7, 12). Jungblut *et al*. (7) demonstrated high 16S rRNA gene similarity (>99 %) from strains found in the Arctic, Antarctica and high mountain cyanobacterial mats. Strunecký *et al* (11) found several genera of *Phormidium* in polar regions and concluded that they were not endemic because of their similarity to temperate strains.

1.2 Cyanophages: viruses of Cyanobacteria

1.2.1 Viruses in the aquatic environment

Viruses have been known to be present in aquatic systems since the mid 20^{th} century, but it was decades later that transmission electron microscopy (TEM) revealed that they were the most abundant biological entity in aquatic environments (40, 41). They occur in both fresh and marine waters, from tropical to polar regions, and from surface waters to sediments. They have been found in cryoconite sediments from an Arctic glacier (42) to ice core samples from the bottom of the ice sheet over Lake Vostok (43).

Viruses range from approximately 10^6-10^8 ml^{-1} in marine and fresh waters, and differ by about an order of magnitude between coastal and open waters (44). In polar freshwater systems, viral abundances range from 0.05-94 x 10^7 ml^{-1} (43). Viruses are generally ~10-fold more abundant than bacteria (44, 45). Since viruses are obligate pathogens, their abundance often corresponds to that of the organisms they infect. Consequently, phages (viruses infecting bacteria) are the most abundant viruses in the biosphere. For example, there are about 10^{30} phage particles on Earth (46).

As significant agents of mortality, viruses are important players in biogeochemical and ecological processes (45, 47). Phages cause the lysis of a large proportion of both autotrophic and heterotrophic prokaryotes, shunting nutrients between particulate and dissolved phases and affecting community composition. As prophages, they confer immunity against infection and influence the transfer of genetic material among host organisms (48).

1.2.2 Cyanophage genomics

Evidence of viruses infecting marine cyanobacteria (i.e cyanophages) emerged from observations that a significant proportion of *Synechococcus* cells contains visible viral particles (41), and that viruses infecting *Synechococcus* spp. can be readily isolated from seawater (49–51). Cyanophages reach abundances in excess of 10^5 ml^{-1} (47), and their abundance fluctuates with temperature, salinity and host abundance (43, 49, 51). It is estimated that viral lysis removes from less than one to several percent of *Synechococcus* cells each day (41, 51).

The diversity of cyanophages has been examined in terms of morphology, host range and DNA sequence analysis. Based on morphology, cyanophages fall into three families, *Myoviridae*, *Podoviridae* and *Siphoviridae* (52). Myoviruses are viruses with a contractile tail and an isometric or prolate head. Podoviruses are viruses with a short non-contractile tail and an icosahedral head. Siphoviruses are viruses with a non-contractile flexible tail and an icosahedral head. Representatives of all three families have been isolated from seawater (50, 51, 53–59) and freshwater (60–66). Host-range studies have revealed that some cyanophages have broad host ranges and are able to infect strains that are distantly related (51, 53) or that even belong to different genera (54, 56).

Many genomes of cyanophages that infect marine *Synechococcus* (57–59, 67–69), *Prochlorococcus* (55, 68, 70, 71) and one (cyanomyovirus Syn9) that infects both *Synechococcus* and *Prochlorococcus* (56) have been sequenced. Most of these cyanophages are myoviruses and share similarity with the T4-like myoviruses. Their genome size ranges from 161 kb to 252 kb and they share core genes involved in virion structure, DNA replication, and host-derived genes(57–59, 67–69) . These host-derived genes include *psbA* and *psbD* which encode two important proteins (D1 and D2) of the photosystem II (55, 67, 72, 73). Other host-derived

genes involved in photosynthesis found in cyanophage genomes include *hoI (*haem oxygenase*)*, *pcyA* (phycocyanobilin:ferredoxin oxidoreductase), *pebS* (phycoerythrobilin synthase), *petE* (plastocyanin), *petF* (ferredoxin), *ptoX* (electron transfer to oxygen), *cpeT* (putative regulator of phycoerythrin biosyn- thesis), and *hli* (high-light inducible protein)(55, 57, 67, 68, 73). Cyanophage genomes also contain genes involved in carbon metabolism and nutrient acquisition (both for phosphorus and nitrogen). These genes include *talC* (transaldolase), *zwf* (glucose-6-phosphate dehydrogenase) and *gnd* (6-phosphogluconate dehydroge- nase).Acquisition of host-derived genes appears to be associated with a hyperplastic region within the conserved structural gene module. The marine cyanopodoviruses that have been sequenced so far are all part of the T7-like podoviruses supergroup (55, 69, 71). Their genomes are relatively small, ranging from 42 kb to 47 kb, and they all have similar genome architecture. The cyanopodoviruses also share core genes involved in virion structure, DNA replication and host-derived genes (55, 69, 71). The marine cyanosiphoviruses have genome sizes ranging from 30kb to 108 kb (55, 58). They are divergent from other siphoviruses but are mostly related to the lambda-like phage with which they share about 13 functional genes.

Comparatively, there are far fewer data for cyanophages infecting freshwater cyanobacteria. The available information reveals that they are not necessarily genetically related to marine cyanophages. To date, only five cyanophages that infect *Microcystis aeruginosa* (74), *Phormidium foveolarum* (63, 66), *Synechococcus* (65) and *Planktothrix agardhii* (75) have been sequenced (Table 1.1). The cyanomyovirus S-CRM01 that infects freshwater *Synechococcus* sp. is related to other cyanomyoviruses infecting marine *Synechococcus* and *Prochlorococcus* sp. (65). However, a region of about 58 kb in its genome contains a high proportion of genes that are unique (85%). In contrast, Ma-LMM01 that infects the freshwater toxic bloom-forming

cyanobacteria *Microcystis aeruginosa* has little similarity with previously sequenced myoviruses (74). Similarly, the podoviruses Pf-WMP3 and Pf-WMP4 are quite divergent from the other sequenced podoviruses at the nucleotide level but are similar in terms of genomic content and organization (63, 66).

Table 1-1. Freshwater cyanophages that have been sequenced to date.

Phage	Genome size	Family	Host strain	Accession	Reference
Pf-WMP3	43.2	Podoviridea	*Phormidium foveolarum*	EF537008	(63)
Pf-WMP4	40.8	Podoviridea	*P. foveolarum*	DQ875742	(66)
Ma-LMM01	162.1	Myoviridea	*Microcystis aeruginosa*	AB231700	(74)
S-CRM01	178.6	Myoviridea	*Cyanobium gracile*	NC015569	(65)
PaV-LD	95.2	NA	*Planktothrix agardhii*	HQ683709	(75)

1.2.3 Culture independent approaches to cyanophage diversity

Cyanophage isolates likely represent a small fraction of the diversity present in nature. The advent of culture–independent approaches such as targeting conserved genes and metagenomic approaches have been used to assess the genetic diversity of cyanophages in natural assemblages. These molecular techniques have demonstrated that we know little about cyanophage diversity in aquatic environments.

Although no gene is universally conserved in viruses, there are a number of genes that are found within specific groups; hence, the approach has been to develop PCR primers that target specific subsets of the viral assemblage. Due to their conserved nature, genes encoding structural proteins have frequently been targeted for PCR amplification, sequencing, and assessing the diversity of genes representative of specific viral groups (76–81). For example, a gene that is

homologous to g20, which encodes a portal vertex protein involved in capsid assembly in T4-like phages, has been used to assess cyanomyovirus diversity (76, 78, 81, 82). The results suggest that the diversity of myoviruses infecting cyanobacteria is very high. While sequences from cultured cyanomyoviruses fell within a well-supported monophyletic group, most environmental sequences fell into groups with no cultured representatives (76, 81). In addition, some environmental sequences recovered from samples collected in different environments such as the Southern Ocean, the Gulf of Mexico, Lake Constance in Germany, and a meltwater pond on the Ward Hunt ice shelf were highly similar or identical (76). This implies that some structural gene sequences were widely distributed in different environments. However, there are a number of problems with the use of gp20 to assess cyanomyovirus diversity, since T4-like phage other than cyanomyoviruses may also be targeted (76), and the primers target only a subset of cyanomyoviruses (80, 82). For instance, gp20 primers were not successful in amplifying a product from freshwater cyanophage isolates (80). Similar environmental studies targeting the DNA polymerase A of podoviruses have demonstrated high diversity as well as nearly identical gene sequences in fresh and marine waters as well as marine sediments (83, 84). Studies have also used a gene in cyanophages that is homologous to *psbA*, which encodes for a core photosynthetic protein (85–88). Unlike using structural gene sequences such as gp20 (76, 82), *psbA* targets more than one phage family, and distinguishes between phages that infect different marine genera, such as between *Prochlorococcus* and *Synechococcus* (85–87) However, *psbA* is not found in all cyanophages, host and phages sequences can group together, and the broad host range of some cyanophages can confound interpretations. As well, the three freshwater cyanophages sequenced to date lack the *psbA* gene (63, 66, 74).

High-throughput DNA sequencing and metagenomics have had a large impact on studying viral diversity and composition by providing more in-depth and less biased information of the genomic diversity of entire viral communities. Viral assemblages from different environments such as marine ecosystems (89–91), freshwater ecosystems (92, 93), and modern stromatolites (94), show that aquatic systems are dominated by dsDNA bacteriophages and are more diverse than previously thought. Indeed, metagenomic approaches have shown that most viral sequences have no homologues in current databases (89–91, 95, 96). Metagenomic analysis on samples from the Sargasso Sea, Gulf of Mexico, British Columbia, and the Arctic Ocean found that more than 90% of the 1.8 million sequences obtained did not have significant homologues to those in databases (89). The few sequences with significant homology were mostly related to cyanomyovirus (T4-like) sequences and were prevalent in all four samples (89). Metagenomic studies using fosmid libraries also demonstrated the prevalence of cyanophage in the environment (97–99). For example, a fosmid library from the Deep Chlorophyll Maximum (DCM) of the Mediterranean Sea reported that 34 out of 197 fosmid clones were attributed to cyanophages (97). Other studies have used read recruitments analysis to show the prevalence of different cyanophages on different metagenomic databases (58, 70, 71, 91). For example, a read recruitment performed with different cyanophages against the Global Ocean Survey (GOS) database demonstrated that cyanomyoviruses were the most abundant, following by cyanopodoviruses and then cyanosiphoviruses (58).

Cultured-dependent and independent approaches have demonstrated that we have only scratched the surface of cyanophage diversity present in aquatic environments and thus further work is needed to unveil more of this diversity.

1.3 Thesis objectives

Previous studies have demonstrated that polar freshwater systems harbour abundant, dynamic and diverse autotrophic and heterotrophic microbial assemblages. Cyanobacteria dominate these assemblages (38) but there is limited knowledge about the viruses infecting polar cyanobacteria. The overarching goal of this dissertation was to characterize the genomes of viruses infecting representative cyanobacteria found in polar regions. Viruses infecting polar cyanobacteria were characterized by culture-dependent (i.e. viral isolation) and culture-independent approaches. Most of the samples used in this dissertation were collected along the northern coastline of Ellesmere Island, Canadian High Arctic (Figure 1.1).

The dissertation had the following objectives:

1. Isolation and genomic characterisation of viruses infecting freshwater unicellular and filamentous cyanobacteria, Chapters 2, 3 and 4)
2. Development of a culture-independent approach to identify sequences associated with viruses infecting cyanobacteria in high-arctic microbial mats.

In Chapter 2, genomic analysis of Cyanophage S-EIV1 isolated from a polar freshwater lake, revealed a new evolutionary lineage of bacteriophages that is globally widespread and abundant. While I isolated a virus infecting a polar unicellular cyanobacterium (Chapter 2), I could not isolate viruses infecting polar filamentous cyanobacteria, even though many were screened (Appendix A). Consequently, Cyanophages A-1 (L) and N-1, which infect filamentous cyanobacteria from the genus *Nostoc*, were sequenced. Even though the isolates are not of polar origin, *Nostoc* spp. are important taxa in high-arctic cyanobacterial mats, and a PCR-based analysis showed that cyanophages related to A-1 and N-1 were also present (Appendix B). Consequently, Cyanophages A-1 and N-1 were sequenced as representatives for viruses infecting

polar cyanobacteria, and genomic analysis revealed they represented a highly divergent group of cyanomyoviruses with distinct DNA polymerase and large subunit terminase genes. The sequencing of Cyanophage N-1 led to the discovery of a CRISPR array in its genome which was discussed in Chapter 4. Finally, in Chapter 5, DNA stable isotope probing (DNA-SIP) was used as a culture-independent approach to isolate DNA from replicating viruses infecting primary producers in high-arctic microbial mats.

Chapter 2: Polar freshwater cyanophage S-EIV1 represents a new evolutionary lineage of phages that are widespread and abundant

2.1 Synopsis

Cyanobacteria are often the dominant phototrophs in polar freshwater communities, yet the cyanophages that infect them remain unknown. Here, we present a genomic and morphological characterization of cyanophage S-EIV1 that was isolated from freshwaters on Ellesmere Island (Nunavut, High Arctic Canada), and which infects the polar *Synechococcus* sp., strain PCCC-A2c. S-EIV1 represents a newly discovered evolutionary lineage of bacteriophages whose representatives are widespread in aquatic systems. Among the 130 predicted open reading frames (ORFs) there is no recognizable similarity to genes that encode structural proteins other than the large terminase subunit and a distant viral morphogenesis protein, indicating that the genes encoding the structural proteins of S-EIV1 are distinct from other viruses. As well, only 19 predicted coding sequences on the 79,178bp circularly permuted genome have homology with genes encoding proteins of known function. Although S-EIV1 is divergent from other sequenced phage isolates, it shares synteny with phage genes captured on a fosmid from the deep-chlorophyll maximum in the Mediterranean Sea, as well as with an incision element in the genome of *Anabaena variabilis* (ATCC 29413). Sequence recruitment of metagenomic data indicates that S-EIV1-like viruses are cosmopolitan and abundant in a wide range of aquatic systems, suggesting they play an important ecological role.

2.2 Introduction

Cyanobacteria are often the dominant phototrophs in polar and subpolar lakes (38), and can account for more than 50% of the phytoplankton chlorophyll *a* in northern lakes (100). In meromictic lakes in the High Arctic (2) and Antarctica (16), planktonic cyanobacteria occur at abundances of up to 10^4 and 10^6 cells ml^{-1}, respectively. Most planktonic polar cyanobacteria are related to *Synechococcus* spp. and fall within two groups that contain brackish and freshwater representatives from different latitudes (2). Arctic and Antarctic cyanobacteria may be mostly cosmopolitan, generalist taxa rather than endemic specialists, but this will require ongoing genomic analysis to fully resolve (7).

Cyanophages have been well described in marine systems, where their abundances can be in excess of 10^5 ml^{-1} (47), varying with temperature, salinity and host abundance (49, 51). It is estimated that viral lysis removes up to a few percent of the *Synechococcus* population each day (49, 51). Based on morphology, cyanophages fall into the families *Myoviridae*, *Siphoviridae* and *Podoviridae* (52), although there is increasing evidence that morphology is a poor basis for classifying phage (101). Myoviruses have a contractile tail and an isometric to prolate head; siphoviruses have a non-contractile flexible tail and an icosahedral head; podoviruses have a short non-contractile tail and an icosahedral head. Representatives of all three families have been isolated from seawater (50, 51, 53–55) and freshwater (60–62, 102).

Despite the widespread distribution and ecological importance of polar cyanobacteria (38), little is known about viruses infecting these organisms. Here, we report on the isolation and genomic analysis of cyanophage S-EIV1, which along with its host, *Synechococcus* sp. strain PCCC-A2c, was isolated from polar freshwaters on northern Ellesmere Island, in the Canadian High Arctic. The virus S-EIV1 bears little resemblance to previously characterized cyanophages,

and along with its host, provides the first model system for studying cyanobacteria-virus interactions and examining viral diversity in aquatic polar ecosystems.

2.3 Material and methods

2.3.1 Host cells

The phycoerythrin-rich picocyanobacterium, *Synechococcus* sp. strain PCCC-A2c was isolated in July 2001 from the upper freshwater layer of Lake A, a meromictic lake at lat. 83º05'N, long. 75º30'W near the northern limit of the Canadian High Arctic (details in Van Hove et al. 2008). The strain was isolated from a water sample taken immediately under the ice at 2m depth, in the middle of the lake, by sequential dilution (2) in sterile BG-11 medium (103) at 10ºC under continuous light (50µmol photons m^{-2} s^{-1}), and then transferred to batch cultures for maintenance at 8ºC under continuous low irradiance (33µmol photons m^{-2} s^{-1}). The isolate is maintained in the CEN Polar Cyanobacteria Culture Collection at Laval University as strain PCCC Number A2c.

2.3.2 Cyanophage isolation

Cyanophage S-EIV1 was isolated from a composite of virus concentrates collected from the surface waters of lakes and ponds on Ellesmere Island, Nunavut, Canada (Table 2-1). Briefly, 20 to 40 l of water were filtered serially through 1.2-µm (GC50; Advantec MFS, Dublin CA) and 0.45-µm (HVLP; Millipore, Bedford MA) pore-size filters, and the remaining virus-size particles concentrated ca. 100- to 200-fold using a 30-kDa-MWcutoff ultrafiltration cartridge (Prep/Scale-TFF-2; Millipore) (104). Viral concentrates were stored at 4°C in the dark until processed. S-EIV1 was isolated by pooling several subsamples from the virus concentrates, adding the mix to an exponentially growing culture of *Synechococcus* sp. strain PCCC-A2c and incubating at 8°C under continuous irradiance of 33 µmol photons m^{-2} s^{-1} for 14 to 17 days. Culture lysis was

determined by a marked decrease in relative fluorescence (*in vivo* chlorophyll;Turner Designs TD-700 fluorometer, Sunnyvale CA, USA) compared to control cultures. The virus was then cloned by repetitive dilution to extinction in 96-well microtiter plates (53) containing exponentially growing *Synechococcus* sp. strain PCCC-A2c.

Table 2-1. Site information for virus concentrates collected from freshwater systems on Ellesmere Island, Nunavut (Canada).

Sample ID	Location	Latitude	Longitude	Date	Depth (m)
VC8	Ward Hunt Lake	83°05'N	74°10'W	14-Aug-08	0
VC9	Quttinirpaaq Lagoon	83°05'N	74°15'W	17-Aug-08	0
VC10	Ward Hunt Ice Shelf	83°01'N	71°30'W	19-Aug-08	0
VC12	Lake A-2m (oxic)	83°00'N	75°30'W	20-Aug-08	2
VC13	Lake A-12(oxicline)	83°00'N	75°30'W	20-Aug-08	12

2.3.3 Amplification and purification of S-EIV1

The cyanophage was amplified by adding 0.1% (v/v) of the virus isolate to five 35 ml cultures of *Synechococcus* sp. strain PCCC-A2c and incubated until lysis. The lysates were pooled and filtered through a 0.45 μm pore-size filter (HVLP; Millipore) to remove cellular debris. The virus was then concentrated (~50x) by ultrafiltration using Millipore Plus 70 Centricons. The concentrate was loaded onto a 20/30/40/50% (w/v in 50 mM Tris-HCl, pH 7.6) Optiprep^{TM} (Sigma-Aldrich, St. Louis, MO) step gradient, and ultracentrifuged for 8 h at 86,711 x g and 20˚C (SW40 rotor, Beckman Coulter, Indianapolis, IN). After centrifugation, the single visible band was extracted from the gradient by puncturing the side of the tube with a sterile 1-ml syringe and dialyzed overnight in a 20,000 MWCO dialysis cassette (3 ml Slide-A-Lyzer; Thermo Scientific-Pierce, Rockford, IL) against 500 ml of 200 mM Tris-HCl, pH 7.6, at 4°C.

2.3.4 Transmission electron microscopy

S-EIV1 lysate (70ml) was 0.45-μm filtered (HVLP; Millipore) and concentrated by ultracentrifugation for 6h at 119,577 *g* and 8°C in a 45Ti rotor (Beckman Coulter). The pelleted

viruses were resuspended in 1ml of supernatant. A portion of the virus suspension was fixed with glutaraldehyde (final 1% v/v) and adsorbed to the surface of formvar/carbon coated copper grids as previously described (53). The grids were briefly stained with 2% phosphotungstic acid (pH 7), viewed and photographed on a FEI Tecnai G2 200kV transmission electron microscope at the University of British Columbia Bioimaging Facility. Virus dimensions were estimated from electron micrographs of negatively stained particles.

2.3.5 Chloroform sensitivity

Sensitivity to chloroform was tested by adding 500 µl of 0.22-µm filtered lysate to an equal volume of chloroform and shaking by hand for 5 min. The chloroform was removed by centrifugation at 4,100 x g for 5 min at 10°C (Allegra X-30, F2402 rotor, Beckman Coulter). The aqueous phase was transferred to a microfuge tube and incubated for 6 h at room temperature to evaporate any remaining chloroform. As a control, 500 µl of chloroform was added to 500 µl of BG-11 medium. Chloroform-treated virus, chloroform-treated medium and non-treated viruses were added to exponentially growing *Synechoccoccus* sp. strain PCCC-A2c cultures and relative fluorescence measured for 2 weeks.

2.3.6 Host range

The host range of S-EIV1 was tested against 6 replicate cultures of 8 polar cyanobacterial strains (Table 2-2) grown as previously described. Susceptibility to infection was determined by a decline in relative fluorescence compared to control cultures to which no viruses were added.

Table 2-2. Cyanobacterial strains tested and their susceptibility to infection by Cyanophage S-EIV1

Strain ID	Location	Latitude	Longitude	Depth (m)	Susceptibility[1]
PCCC-A6.5a	Lake A	83º05'N	75 º30'W	6.5	-
PCCC-A2c	Lake A	83º05'N	75 º30'W	2	+
PCCC-A215	Lake A	83º05'N	75 º30'W	2	-
PCCC-A2b	Lake A	83º05'N	75 º30'W	2	-
PCCC-C112	Lake C	82º51'N	78 º12'W	12	-
PCCC-A2	Lake A	83º05'N	75 º30'W	2	-
PCCC-A2a	Lake A	83º05'N	75 º30'W	2	-
PCCC-PA	Antoniates pond	82º58'N	75 º24'W	0	-

2.3.7 Structural proteins

To identify the structural proteins, purified S-EIV1 was diluted in SDS buffer (4:1, v/v) and heated at 95ºC for 5 min. The sample was then resolved by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) using a Mini-PROTEAN Tetra Cell (Bio-Rad Laboratories; Hercules, CA). The 4 to 12% gel was run in a SDS running buffer (pH 8.3) at 100V for 2 h using a Novex Sharp Protein Standard (Invitrogen, Carlsbad CA) for size calibration. The gel was stained overnight with Coomassie Blue and de-stained for 2 d in a solution of 20% methanol and 10% acetic acid.

2.3.8 DNA extraction, sequencing and assembly

Synechococcus sp. strain PCCC-A2c was grown at 8ºC in 800 ml of BG-11 medium (103) in 1 l flasks and 33 µmol photons $m^{-2}s^{-1}$ continuous illumination. Exponentially growing cultures were infected with S-EIV1 and incubated as above for 14 to 17 d until lysis occurred. Sodium chloride (Sigma-Aldrich) was added to the lysate at a final concentration of 0.5 M at 4°C, which after 1 h was filtered through 1.2-µm pore-size glass-fiber (GC50; Advantec MFS) and 0.22-µm pore-size membranes (GVWP; Millipore) to remove cellular debris. The filtered

lysate was ultracentrifuged for 6h at 119,577 x g and 8˚C (Type 45Ti rotor,Beckman Coulter), the supernatant removed and the virus pellet resuspended in 200 µL of BG-11 medium.

The pellet was treated with DNase 1 and RNase A to remove free nucleic acids, and the nucleic acids extracted using a QiAamp MinElute Virus Spin Kit (Qiagen, Mississauga, ON). The DNA was sheared into ~300 bp fragments using a Covaris M220 ultrasonicator (Covaris, Woburn, MA), and purified using Agencourt AMPure XP beads (Beckham Coulter). The sequencing library was constructed using NxSeq® DNA Sample Prep Kits (Lucigen, Middleton, WI) and sequenced on an Illumina MiSeq at the Génome Québec Innovation Centre at McGill University (Montréal, QC). The adapters were trimmed from the reads using Trimmomatic-0.30 (http://www.usadellab.org/cms/index.php?page=trimmomatic), quality checked with Sickle (https://github.com/najoshi/sickle), and assembled using Ray with the default parameters, and 23 as the k value (105).

2.3.9 Genome annotation and identification of regulatory elements and motifs.

Open reading frames (ORFs) in S-EIV1 were predicted using GeneMark (106) and GLIMMER (107); where the predictions differed, the longer of the two was kept. The predicted ORFs were translated and assigned putative functions by using blastp to compare them with protein sequences in the GenBank (nr), Acclame and Procite databases. Sequences with e-values $<10^{-3}$ were considered to be homologues. PSI-BLAST and HHpred were used to predict more distant homologues. The genome was also analyzed for regulatory elements and motifs such as tRNA genes, promoter motifs and transcriptional terminators. tRNA genes were identified using tRNAscan-SE (108) and Aragorn v1.1(109). Putative promoter motifs were identified using PHIRE (110) with default parameters (20-mer sequences (S) with 4 base pair degeneracy (D=4)), and if they occurred in the 150 bp region immediately upstream of start codons of predicted

protein-coding genes. Rho-independent terminators were identified using Softberry's FINDTERM (http://linux1.softberry.com/berry.phtml) with the default energy threshold set to -16 kCal (56). The genomic map was constructed using GCview (111). The genome of S-EIV1 was compared to fosmid MEDDCM-OCT-S04-C348 (97) by using a tblastx analysis (cutoff e-value < 10^{-3}).

The prophage incision element, AvaD, in *Anabaena variabilis ATC1495* was confirmed by translating the ORFs and using blastp to compare them with sequences in the GenBank (nr), Acclame and Procite databases, as outlined above.

2.3.10 Phylogenetic analysis

DNA polymerase A (*DNApol A*) and the large terminase subunit (*terL*) were compared phylogenetically with those from other cyanophages by aligning the inferred amino-acid sequences with ClustalX for *DNApol A* and Promals for *terL* (112, 113) using default parameters followed by manually refining the alignments with Geneious v4.7 (114). Maximum likelihood trees were constructed with RAxML rapid bootstrapping and ML search (100 replicates) (115) assuming the James-Taylor Thornton model of substitution using empirical base frequencies and estimating the proportion of invariable sites from the data.

In an attempt to investigate the importance of S-EIV1 phages in the environment, phylogenetic analysis were also performed with DNA polymerase A from known phages, metagenomic sequences and amplicons from degenerate primers (83, 84, 116). Blast and the inferred amino acid sequence from S-EIV1 *DNApol A* was used to recover additional sequences from GenBank (Appendix D) that were used for phylogenetic analysis. The sequences were aligned and a maximum likelihood was constructed as desbribed above.

2.3.11 Recruitments of reads to metagenomic data

First, to interrogate other aquatic systems for S-EIV1-like phages, viral metagenomic databases (Table 2-9) in the CAMERA database (http://camera.cali2.net) were blasted using tblastn (e-value < 10^{-5}) against the protein sequences of S-EIV1 and fosmid MEDDCM-OCT-S04-C348 were blasted using. Reads with >20% amino-acid identity and >45nt in length were kept and the number of hits were recorded. Second, a database containing protein sequences from S-EIV1 and other aquatic phages (Table 2-3) was used to recruit reads from GOS database. GOS databases were blasted using tblastn (cut off e-value < 10^{-5}) against the constructed phage databases (Table 2-3). Reads with >20% amino-acid identity and >45 nt in length were kept. If a read was recruited for more than one phage, the read was associated with the phage that provided the lowest e-value. In order to reduce the effect of genome size, the read's counts were normalized by numbers of ORFs.

Table 2-3. Phages that were used in the recruitment of reads to metagenomic data

Phage	Genome size	Family	Host strain	Accession	Reference
T3	38,2	Podoviridae	*Escherichia coli*	AJ318471	(117)
N4	70.1	Podoviridae	*Escherichia coli*	NC008720	(118)
Pf-WMP3	43.2	Podoviridae	*Phormidium foveolarum*	EF537008	(63)
Pf-WMP4	40.8	Podoviridae	*Phormidium foveolarum*	DQ875742	(66)
S-CRM01	178.6	Myoviridae	*Cyanobium gracile*	NC015569	(65)
Syn5	46.2	Podoviridae	*Synechococcus*	NC009531	(56)
HTVC019P	42.1	Podoviridae	Pelagibacter	NC020483	(119)
P60	47.9	Podoviridae	*Synechococcus*	NC003390	(69)
SIO1	39.9	Podoviridae	Roseobacter	AF189021	(120)
SCBP3	47.3	Podoviridae	*Synechococcus*	EF535233	(121)
PSSP7	44.9	Podoviridae	*Prochlorococcus*	NC006882	(55)
HTVC011P	39.9	Podoviridae	Pelagibacter	NC020482	(119)
S-CBS2	72.3	Siphoviridae	*Synechococcus*	GU936714	(58)
PaV-LD	95.2	NA	*Planktothrix agardhii*	HQ683709	(75)
S-EIV1	79.1	NA	*Synechococcus*	KJ410740	This study
P-RSP2	42.2	Podoviridae	*Synechococcus*	HQ332139	(71)
S-TIM5	161.4	Myoviridae	*Synechococcus*	JQ245707	(59)
HTVC010P	34.9	Podoviridae	Pelagibacter	NC020481	(119)
S-SM2	190.8	Myoviridae	*Synechococcus*	NC015279	(68)
P-SSM2	252.4	Myoviridae	*Prochlorococcus*	NC006883	(55)
HTVC008M	147.2	Myoviridae	Pelagibacter	NC020484	(119)

2.4 Results and discussion

Cyanobacteria are major primary producers in freshwater ecosystems, and are often the most abundant phototrophs in polar lakes (38); however, representative cyanophages from these waters have not been previously described. In the present study, we isolated and characterized cyanophage, S-EIV1, from the Canadian High Arctic. This virus infects the freshwater polar cyanobacterium *Synechococcus* sp. strain PCCC-A2c. Based on morphology and genomic content cyanophage S-EIV1 represents a new evolutionary lineage of bacteriophage. The

circularly permuted genome of 79,178 bp has little similarity to other sequenced phages; yet, interrogation of metagenomic data suggests that viruses related to S-EIV1 are widely distributed in aquatic systems.

2.4.1 General features

Electron micrographs of negatively stained cyanophage S-EIV1 revealed icosahedral capsids with an estimated average diameter of ca. 95 nm (n=33). Evidence of short spiky extensions (Figure 2-1a, black open arrow) and long, fine tail fibers (Figure 2-1A, open white arrows) projecting from the base of the capsid were seen on both intact and empty capsids. A long tail-like structure was found to be associated only with empty capsids and extended up to ca. 125 nm in length (average = 109 nm, n=5; Figure 2-1A , closed white arrow). As well, infectivity of S-EIV1 is chloroform sensitive (Figure 2-1B), similar to some tailed phages, although this does not necessarily indicate the presence of lipids (122). S-EIV1 has a narrow host range and did not infect eight other cyanobacterial isolates from nearby freshwaters that were tested (Table 2-2).

2.4.2 Genomic analysis

The 79,178bp genome of S-EIV1-1 is circularly permuted (Figure 2-2, Appendix C), with a G+C content of 46.2%. Most ORFs in S-EIV1 do not share significant similarity with genes of known function. Of the 130 predicted ORFs encoded on both strands, 42 have homology to the database and of those only 15 shared homology with genes of known function (blastp, e-value cutoff < 10^{-3}), including genes associated with DNA metabolism, replication and cell lysis (Table 2-3, Appendix D); no sequences coding for tRNAs were found. PSI-BLAST and HHpred were used to ascribe function to additional ORFs, and resulted in the identification of putative coding sequences for a viral morphogenesis protein (ORF109), an exonuclease (ORF17), an o-

methyltransferase (ORF19) and a restriction endonuclease (ORF33) (Table 2-4). This gives a total of 19 ORFs that show homology with genes of known function. Three transcriptional terminators were predicted by Findterm (Table 2-5); two are downstream of ORFs with unknown function (ORF6 and ORF16), while one is downstream of a gene predicted to encode a peptidase (ORF106).

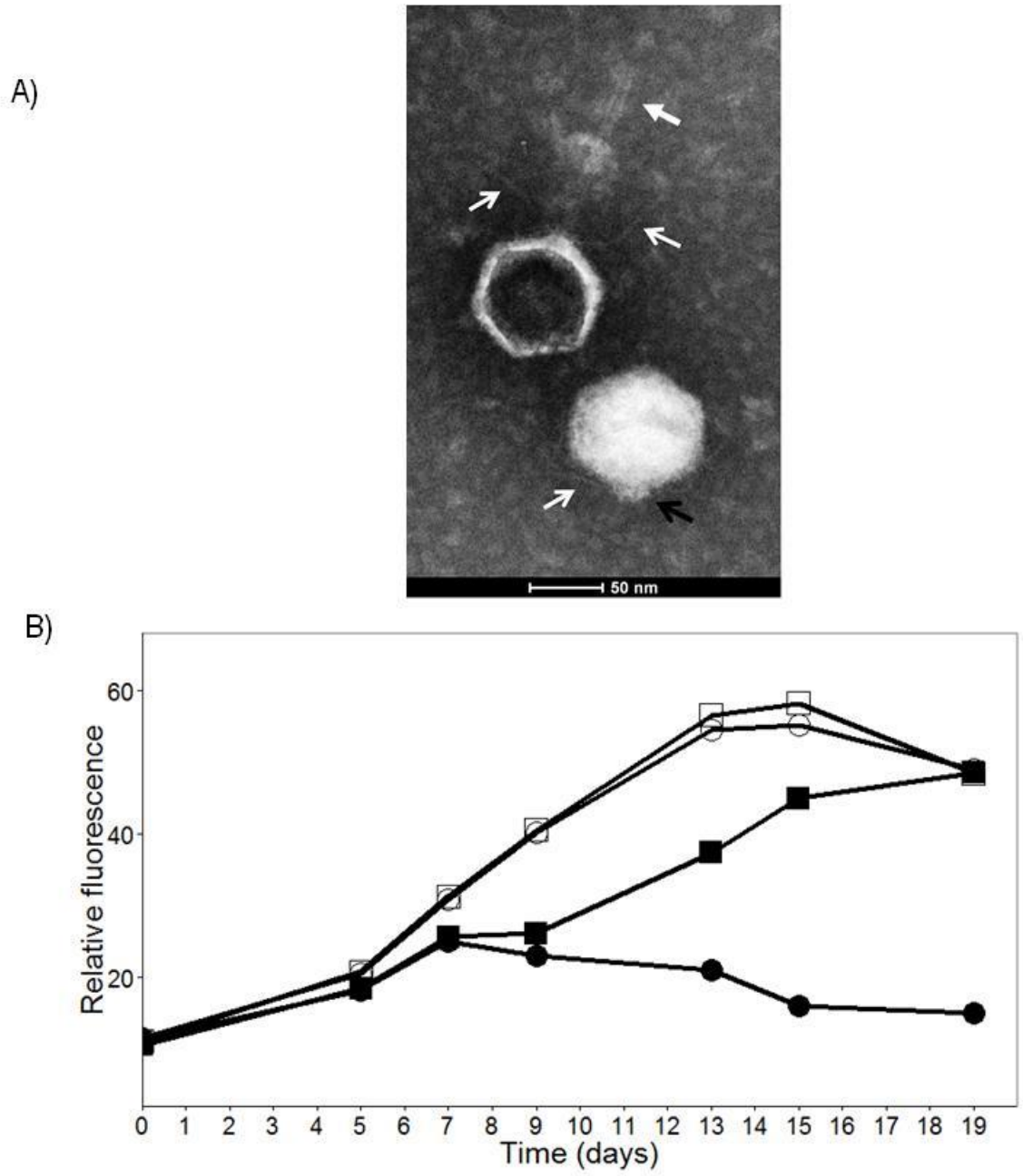


Figure 2-1. General features of cyanophage S-EIV1
A) Transmission electron micrograph of S-EIV1 negatively-stained with 2% phosphotungstic acid reveals large icosahedral capsids with a number of distinctive morphological features. Fine tail fibers (open white arrows) and short spiky extensions (black open arrow) were consistently seen in filled and empty capsids, while delicate tail-like structures were associated with empty capsids (closed white arrow). B) Effect of chloroform on the infectivity of S-EIV1. white circle = uninfected Synechococcus culture; black square = chloroform-treated viruses; white square = chloroform-treated medium, black circle = non-treated viruses. These are representative data from one of three independent experiments.

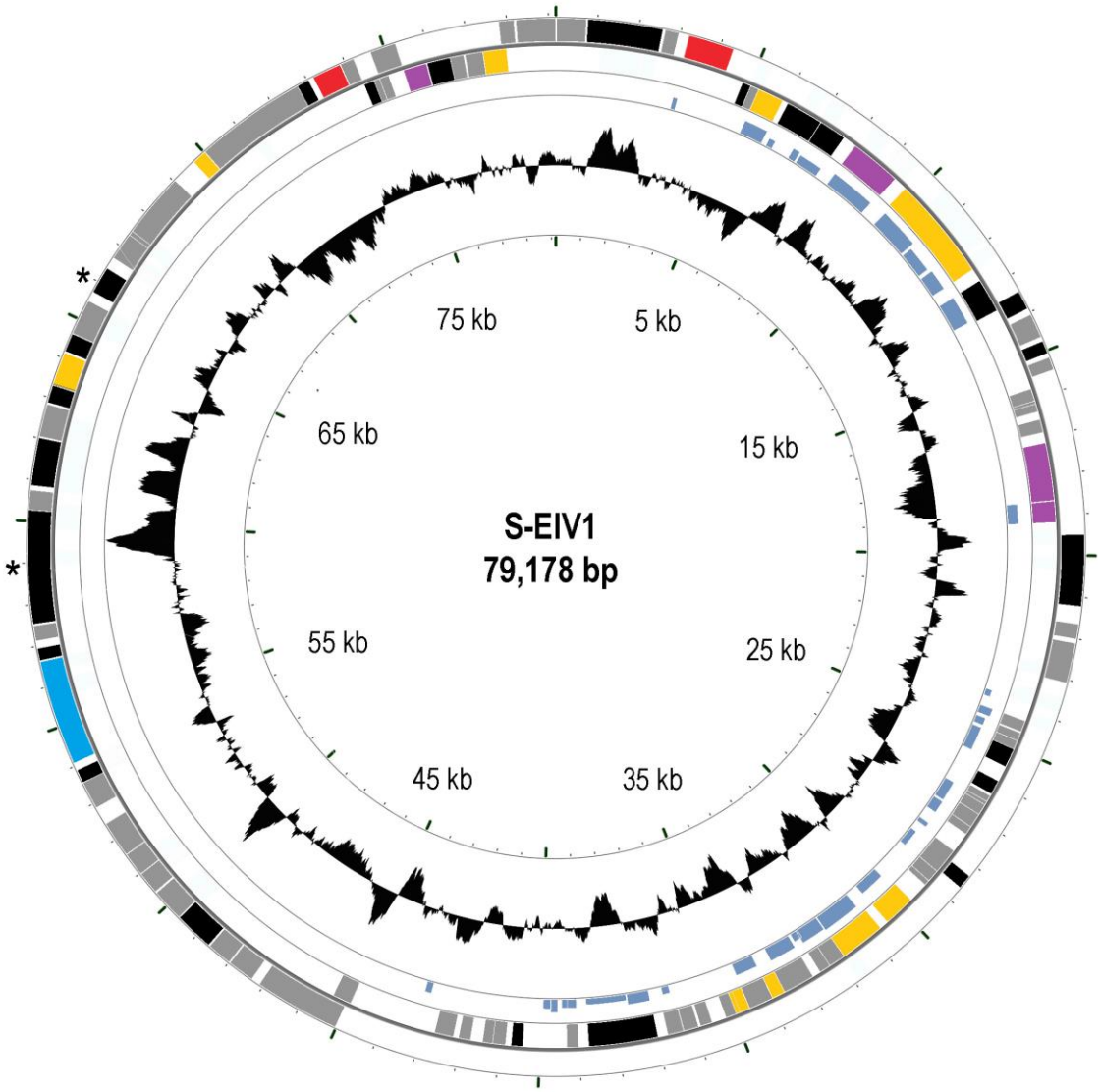


Figure 2-2. Genomic map of S-EIV1.
Circles from outmost to innermost correspond to: i) predicted ORFs (blastp, nr database, e-value < 10^{-3}) on forward strand and ii) reverse strand; iii) tblastx hits (e-value < 0.0001) against the fosmid MEDDCM-OCT-S04-C348 (the height of the bar is proportional to the e-value) and iv) GC content plotted relative to the genomic mean of 46.2 % G+C. Only ORFs greater than 200bp are shown and are colored as follows: red, lysis/lysogeny; grey, no homolog; black, hypothetical proteins; purple, host-derived genes; yellow, DNA metabolism and replication. * indicates structural genes that were identified by SDS-PAGE (see Figure 2-3).

Table 2-4. Predicted ORFs of cyanophage S-EIV1 with similarity to genes of known function.

ORF	Length (bp)	Strand	Significant hit	Organism	e*	%id (shared aa)
5	1107	+	Lysozyme	*Synechococcus* phage S-CBP3	e^{-115}	55% (205)
8	888	-	PurM	uncultured phage MedDCM-OCT-S04-C348	$9e^{-78}$	60% (135)
13	1266	-	Glycosyl transferase group 1	uncultured phage MedDCM-OCT-S04-C348	e^{-121}	53% (209)
14	156	-	S-adenosylmethionine decarboxylase proenzyme (DUF206)	*Prochlorococcus marinus* str. AS9601	$2e^{-05}$	44%(22)
15	2640	-	P4 phage primase	uncultured phage MedDCM-OCT-S04-C348	e^{-161}	70% (265)
28	552	-	DNA-binding ferritin-like protein	*Opitutaceae* bacterium TAV1	$4e^{-04}$	30% (39)
50	807	-	ssDNA-binding protein	*Thermo* uncultured phage MedDCM-OCT-S04-C348	$8e^{-67}$	45%(237)
52	1869	-	DNA polymerase family A	uncultured phage MedDCM-OCT-S04-C348	e^{-148}	70% (245)
55	627	-	FAD dependent thymidylate synthase	uncultured phage MedDCM-OCT-S04-C348	$3e^{-57}$	53% (116)
56	378	-	Endodeoxyribonuclease	*Pseudomonas* sp. HPB0071]	$3e^{-05}$	47% (26)
95	2538	+	Putative terminase large subunit Ava_D0014	*Anabaena variabilis* ATCC 29413	$1e^{-67}$	43%(137)
106	744	+	Peptidase, M23 family	*Acinetobacter* sp. WC-743	$3e^{-07}$	33%(44)
115	360	+	HNH nuclease	*Synechococcus* sp. CC9902	$4e^{-16}$	42% (40)
119	663	+	Lysozyme	*Acinetobacter* sp. RUH2624	$6e^{-18}$	37% (57)
121	237	-	Protein of unknown function DUF3310	*Clostridium* sp.	$6e^{-06}$	47% (34)
124	585	-	deoxycytidine triphosphate deaminase	*Synechococcus* sp. WH 7803	$2e^{-53}$	53% (104)

***e-value**

Table 2-5. Identification of distant homologs of S-EIV1 ORFs using HHpred analysis.

ORF	Length (bp)	Strand	Predicted function	PfamA ID	e-value
17	849	-	Exonuclease	Pf12684	$4.2e^{-17}$
19	588	+	O-methyltransferase	PF01596	$1.6e^{-20}$
33	621	+	Restriction endonuclease	NA	$3.4e^{-11}$
109	327	-	Morphogenesis protein	NA	$1.2e^{-39}$

Sequences encoding phage structural proteins (i.e. capsid, tail tube, portal or tail fiber) were not found with the exception of the terminase large subunit and a viral morphogenesis

protein classified with HHpred, providing further evidence that S-EIV1 represents a previously unknown phage lineage. However, SDS-PAGE analysis resolved 6 visible structural proteins of about 23, 32, 35, 39, 42 and 85 kDa (Figure 2-3). Of these, the viral morphogenesis protein (ORF109), corresponding to the 23 kDa band, and ORF99 being the only putative coding sequence long enough to encode a 85 kDa protein, were the only structural proteins that could be matched with specific ORFs. The detection of only 6 structural proteins for a phage of this size is an underestimation of the structural proteins for cyanophage S-EIV1. For example, the SDS-PAGE analysis for cyanophage PaV-LD which is about 80nm in diameter resolved 13 structural proteins (75). Similar findings were also reported for the cyanophage syn5 which contain 14 structural proteins (123). A mass spectrometry-based proteomics analysis would be needed to further document structural genes given that it more sensitive than SDS-PAGE.

Table 2-6. Transcriptional terminators of S-EIV1 as assigned by FINDTERM

Terminator	Start	End	Length (bp)	Strand	Energy (kCal)	Upstream ORF	Dist to ORF (bp)	Status
Term1	4335	4386	52	-	-22.9	ORF6	1	Putative
Term2	1249	12549	54	-	-22.7	ORF16	10	Putative
Term3	6336	63321	44	+	-19.5	Peptidase	15	Putative

Host-derived genes that have been found in other cyanophages, such as those encoding proteins involved in photosynthesis, carbon metabolism and phosphorus-related functions (Lindell et al. 2004,Millard et al. 2004; Mann et al. 2005; Sullivan et al. 2005; Weigele et al. 2007; Labrie et al. 2013;), were not found in S-EIV1, however, genes for proteins involved in nucleotide metabolism and stress response were identified. First, S-EIV1 encodes a homolog of S-adenosylmethionine decarboxylase (*SpeD*), a key enzyme in the biosynthesis of spermidine and spermine, polyamines that are important in photoadaptation and photoinhibition in oxygenic

phototrophs (125). For example, a mutant of *Synechocystis* sp. PCC6803 with reduced spermidine content exhibits reduced *psbA2* transcript stability (126). *SpeD* gene is commonly found in marine T4-like cyanophage genome (127, 128). Another sequence (ORF28) has distant homology (blastp, e-value< 10^{-3}) to genes encoding DNA binding protein from starved cells (DPS). These intracellular iron-binding proteins in the bacterioferritin/ferritin superfamily (129) act in iron storage, DNA binding and oxidative stress prevention. Prokaryotes have highly regulated enzymatic systems to protect DNA from oxidative damage due to reactive oxygen species such as hydroxyl radicals, superoxide, and H_2O_2 (130). During starvation, when the ability to cope with environmental stress is compromised by the lack of nutrients, DPS efficiently and rapidly responds to oxidative and nutritional stresses by making the cells more resistant to reactive oxygen (130–132). A gene encoding DPS might help cyanophages in polar lakes where oxygen tensions are high but DNA-repair rates are low because cold temperatures, and nutrient availability constrain phytoplankton production (13).

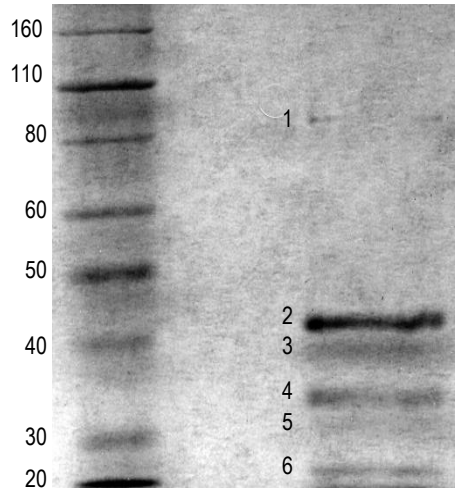


Figure 2-3. Migration of the cyanovirus S-EIP1 on a 12% SDS-PAGE gel following by Coomassie blue staining.
Numbers of the left indicate the molecular masses of the ladder (Novex Sharp Protein Standard, Invitrogen). Numbers of right (1 to 6) indicate bands that were visible on the gel. These are representative data from one of two independent experiments.

Other genes identified in the genome of S-EIV1 include a phosphoribosylaminoimidazole synthetase (purM) homolog that encodes an enzyme involved in purine ribonucleotide biosynthesis and a deoxycytidine triphosphate (*dctp*) homolog that encodes an enzyme required for pyrimidine metabolism. S-EIV1 also encodes a thymidylate synthase homolog which may be involved in scavenging host nucleotides, and which may assist viruses with the synthesis of thymidylate from uridylate after host transcription has stopped (133).

2.4.3 S-EIV1: a new evolutionary lineage of cyanophage

S-EIV1 represents a new evolutionary lineage of cyanophages based on genome content and organization. Although S-EIV1 shows some morphological similarity with members of the *Podoviridae* (*i.e.* intact capsid), there is no evident homology between genes encoding the core structural proteins of podoviruses and genes in S-EIV1, it indicates that S-EIV1 does not belong within this family. S-EIV1 has core genes found in cyanopodoviruses including ssDNA-binding protein (ORF50), endonuclease (ORF 56), primase (ORF15), terminase large subunit (*terL*) (ORF95) and DNA polymerase family A (*DNApol*) (ORF52), but many others are missing including core genes involved in DNA metabolism, assembly and capsid structure (71). As well, cyanopodoviruses generally have a genomic architecture similar to coliphage T7, which encodes genes on a single strand arranged as follows: 1) transcription, 2) RNA polymerase, DNA metabolism and replication and 3) phage assembly and DNA maturation (71). In contrast, S-EIV1 codes from both strands and similar to Roseophage SIO1 (120) does not encode RNA polymerase, implying that host transcription machinery is used during infection, as has been suggested for the siphovirus P-SS2 that infects *Prochlorococcus* sp. (MIT9313) (70).

Phylogenetic analysis of *DNApol* and *terL* shows that S-EIV1 is evolutionarily distinct from other cyanophages. Although *DNApol* is similar to those found in podoviruses, it groups

more closely with viruses infecting *Pelagibacter ubique* (HTCV-like) (119) and *Roseobacter* sp. (SIO-like) (120) than it does with the P60 group of podoviruses infecting other cyanobacteria (Figure 2-4). Further evidence of the evolutionary divergence of S-EIV1 from other viruses is provided by *terL* which encodes a protein involved in DNA packaging. S-EIV1 clusters in a well-supported clade with terminases found in prophage elements in the filamentous cyanobacteria *Anabaena variabilis* (AvaD,) and *Nodosilinea nodulosa* (134) (Figure 2-5). The phylogenetic divergence in *terL* between S-EIV1 and cyanopodoviruses is not surprising given that the genome of S-EIV1 is circularly permuted while in cyanopodoviruses it is linear with direct repeats, which likely involves different DNA packaging processes (135). The clade containing S-EIV1 *terL* was actually more closely related to the cyanosiphoviruses (ie. S-CBS2, P-SS2) than the cyanopodoviruses.

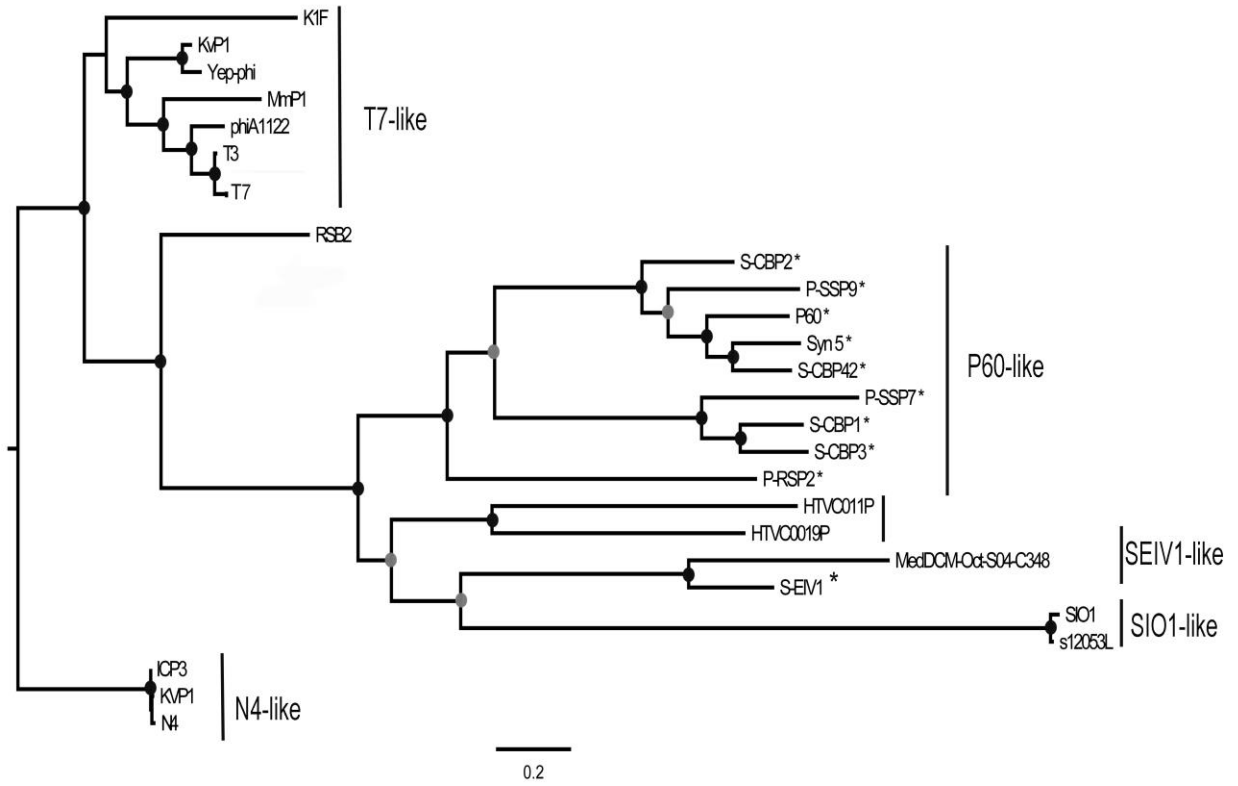


Figure 2-4. Unrooted maximum likelihood amino-tree of DNA polymerase A.
Bootstrap values are indicated as black (90 to 100%) or grey (75 to 89%) dots at the nodes. The groups are labeled as follows: P-60-like, marine cyanopodoviruses; HTVC-like, *Pelagibacter* viruses; N4-like, *Vibrio* viruses; SIO1-like, *Roseobacter* viruses. Cyanophages are labelled with *. S-EIV1 is shown by a black arrow. Scale bar represents amino acid substitutions per site.

The isolation of S-EIV1 suggests that new evolutionary lineages of viruses are likely to be discovered if different host strains are screened. This has been clearly shown recently with previously unknown groups of viruses isolated on *Pelagibacter ubique* (119, 136) and *Cellulophaga baltica* (137). Similarly, most cyanophages have been isolated using a few strains of *Synechococcus* spp.; however, cyanophage S-TIM5, which was from a previously unknown lineage of myoviruses, was isolated from the RedSea using a different *Synechococcus* strain (59). Clearly, there is enormous potential to isolate representatives of previously unknown groups of viruses by screening untested taxa of host organisms.

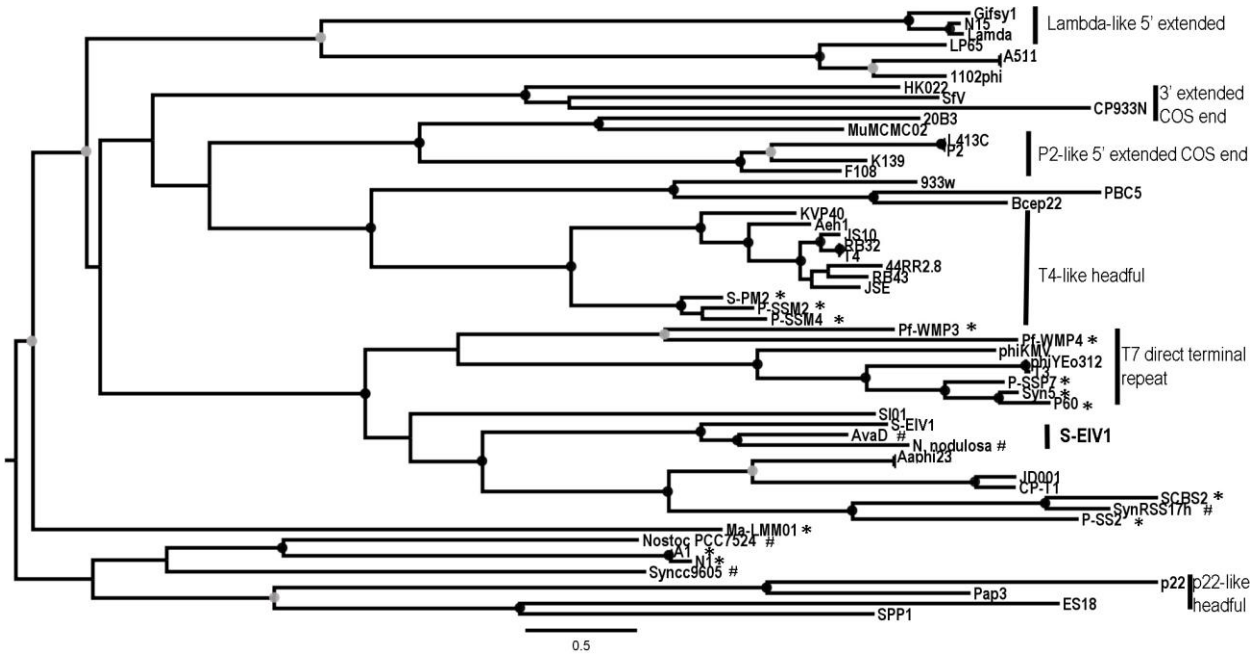


Figure 2-5. Maximum likelihood amino-acid tree of the viral terminase large subunit (*terL*). Bootstrap values are indicated as black (90 to 100%) or grey (75 to 89%) dots at the nodes. Cyanophage genomes are denoted by * and cyanobacterial host genomes by #. The clade containing S-EIV1 (S-EIV1-like) is in bold. Scale bar represents amino acid substitutions per site.

2.4.4 S-EIV1-like viruses in nature

Although S-EIV1 represents a previously unknown phage lineage, it shares synteny with a sequence from an uncultured phage and an incision element in a filamentous cyanobacterium. A blastx analysis (e-value < 10^{-3}) of S-EIV1 genome against the sequence of MEDDCM-OCT-S04-C348, captured in a fosmid library from the deep-chlorophyll maximum in the Mediterranean Sea (97) demonstrate synteny between a region of 40kb from S-EIV1 and the fosmid (Figure 2-2). A total of 26 ORFs are shared between S-EIV1 and MEDDCM-OCT-S04-C348. Of these 19 ORFs encode for hypothetical proteins and seven encode putative proteins with known functions including lysozyme (ORF5), phosphoribosylaminoimidazole synthetase (ORF8), glycosyl transferase (ORF13), primase (ORF15), DNA-binding ferritin-like protein (ORF28), ssDNA-binding protein (ORF50), and *DNApol* (ORF52) (Table 2-6). The gene content

similarity between S-EIV1 and MEDDCM-OCT-S04-C348, as well as the phylogenetic affiliation of *DNApol* indicates that MEDDCM-OCT-S04-C348 is from a relative of S-EIV1.

Despite huge differences in temperature, salinity and wide geographic separation, High Arctic lakes and the Mediterranean Sea are oligotrophic regions where *Synechococcus* is a major primary producer. At the deep chlorophyll maximum in the Mediterranean, *Synechococcus* abundances range from 1.75 to 4 x 10^6 cells ml^{-1} (138), and in Lake A, picocyanobacterial populations reach up to 6 x 10^4 cells ml^{-1} (2). Moreover, *Synechococcus* sp. strain PCCC-A2c has high 16S rDNA gene sequence similarity to cyanobacteria isolated from freshwater, brackish and marine systems (Table 2-7). This may also reflect the range of salinity conditions in meromictic Lake A, from freshwater at the surface where the strain was isolated, to saline conditions at depth. Consequently, strains similar to *Synechococcus* sp. strain PCCC-A2c may occur in the Mediterranean Sea, as do closely related cyanophages such as MEDDCM-OCT-S04-C348.

Table 2-7. Predicted ORFs of Cyanophage S-EIV1 with similarity to predicted ORFs in the uncultured phage sequence MedDCM-Oct-S04-C348.

ORF	Length (bp)	Strand	Significant hit	e-value	%- identity (shared aa)
8	888	-	PurM	$9e^{-78}$	60% (135)
9	936	-	hypothetical protein	$2e^{-10}$	50% (35)
10	438	-	hypothetical protein	$6e^{-19}$	37% (54)
13	1266	-	lycosyl transferase group 1	e^{-121}	53% (209)
15	2640	-	P4 phage primase	e^{-161}	70% (265)
17	849	-	hypothetical protein	e^{-110}	71% (190)
37	261	-	hypothetical protein	$6e^{-19}$	58% (43)
38	219	-	hypothetical protein	$1e^{-08}$	43%(30)
39	717	-	hypothetical protein	$2e^{-26}$	35% (75)
44	963	-	hypothetical protein	$5e^{-36}$	44% (84)
46	216	+	hypothetical protein	$5.3e^{-5}$	48.9% (44)
50	807	-	ssDNA-binding protein	$8.7e^{-67}$	45%(237)
52	1869	-	DNA polymerase family A	e^{-148}	70% (245)
53	759	-	hypothetical protein	$3e^{-108}$	56% (71)
55	627	-	FAD dependent thymidylate synthase	$3e^{-57}$	53% (116)
64	1755	-	hypothetical protein	$1e^{-42}$	46% (98)
67	228	-	hypothetical protein	$7e^{-10}$	48 %(25)
68	282	-	hypothetical protein	$3e^{-09}$	48% (34)

A second 10 kb module on the positive strand of S-EIV1 shares synteny with, and has 5 ORFs with high similarity to a 37 kb incision element (AvaD) in the filamentous cyanobacterium *Anabaena variabilis* ATCC29413. Shared ORFs include putative genes encoding *terL* (Figure 2-6) and a structural protein, suggesting an evolutionary relationship between S-EIV1 and AvaD. Annotation of AvaD revealed more phage-like genes including two integrases (AvaD0049 and AvaD0026), hnh nuclease (AvaD0046), endonuclease (AvaD0044), primase (AvaD0041), DNA

polymerase (AvaD0033), RNA polymerase (AvaD0022) and DNA binding protein (AvaD0015), providing evidence that AvaD is a viral element.

Table 2-8. BLAST summary of 16S rDNA gene sequences with high similarity to *Synechococcus* PCCC-A2c c using the Green Genes database(http://greengenes.lbl.gov).

Sequences producing significant alignments	Accession Number	Location	Score (bits)	e*	%
Cyanobium sp. str. Bright	AY172837	Marine, Red Sea	569	e^{-161}	98.6
Synechococcus sp T7cc1. str	AF448061	Estuary , Tokunoshima Island, Japan	569	e^{-161}	98.6
Synechococcus sp. str. PS838	AF448068	Estuary, Japan	569	e^{-161}	98.6
Synechococcus sp. str. MBIC10089	AB058226	Estuary, Japan	565	e^{-161}	98.4
Synechococcus sp. str. HOS	AF448064	Brackish Marshland, Japan	565	e^{-160}	98.4
Cyanobium sp. str. PCC 7001	AM709626	Intertidal water, Long Island, USA	562	e^{-159}	98.1
Synechococcus sp. str. TAGS	AF448067	Estuary, Japan	562	e^{-159}	98.1
Synechococcus sp. str. PCC7001	AB015058	Intertidal mud, City Island, New York, USA	562	e^{-159}	98.1
Synechococcus sp. str. . MA0607	KFJ763779	Mazurian Lake, Poland	556	e^{-157}	97.5
Cyanobium sp. str. JJ21RS4 21-RS4	AM710355	Freshwater reservoir, Czech Republic	556	e^{-157}	97.5

*e-value

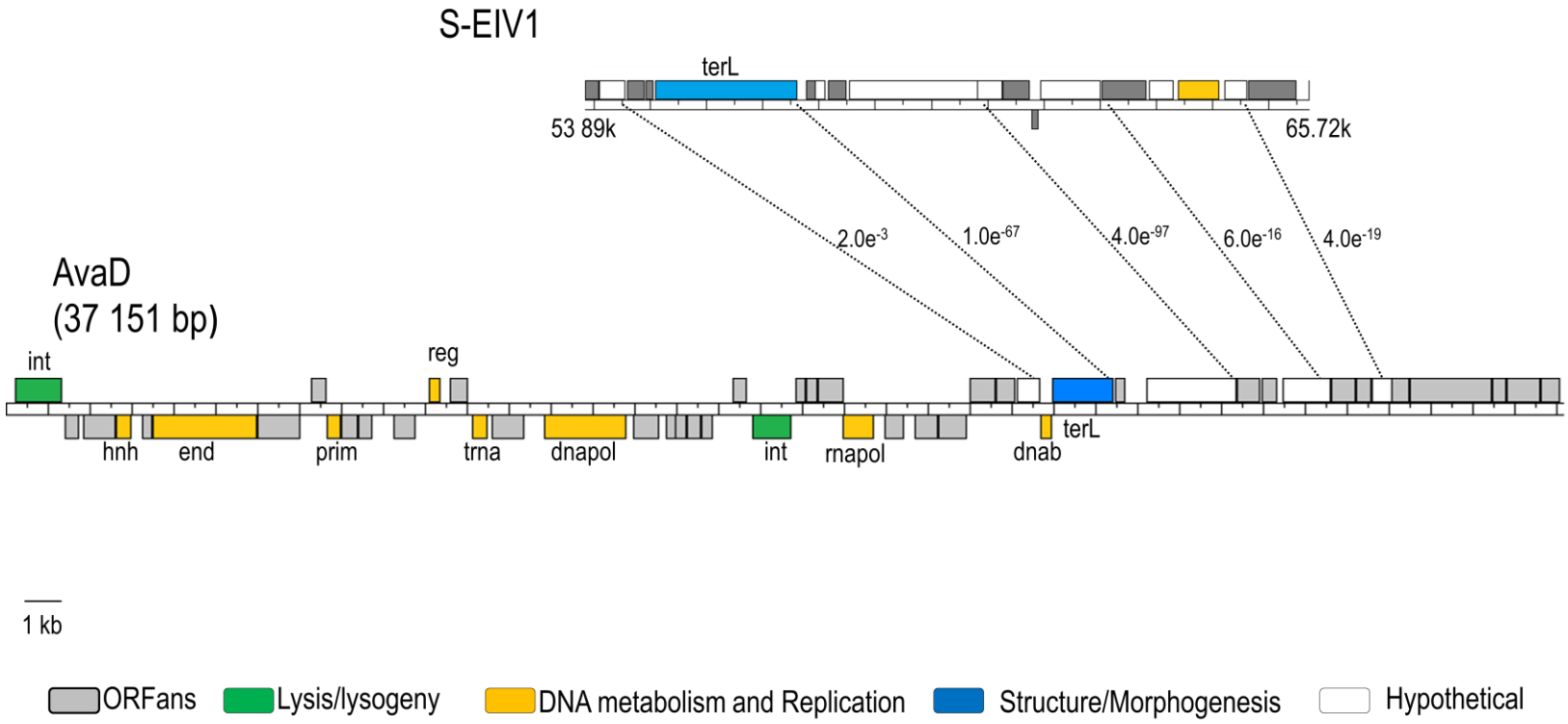


Figure 2-6. Genomic map of the incision element AvaD in *Anabeana variabilis* ATCC29413.
Gene abbreviations and putative functions are as follows: int = integrase, hnh = HNH nuclease, end=endonuclease, prim = primase, reg = transcription regulator, trna = tRNA, dnapol = DNA polymerase; rnapol = RNA polymerase, dnab = DNA binding protein, terL = terminase large sub-unit.

Phylogenetic analysis of *DNApolA* with known phages, metagenomic sequences and amplicons (83, 84, 116) reveals that S-EIV1 along with enviromental sequences form an unrecognized *DNApolA* group (Figure 2-7, named ENV5). Further evidence for S-EIV1-like phages in aquatic systems was obtained from a database of protein sequences from S-EIV1, MEDDCM-OCT-S04-C348 and other aquatic phages. This database was used to retrieve sequences with high similarity from the Global Ocean Survey (GOS) and viral metagenomic data available on CAMERA. Sequences with similarity to S-EIV1 and MEDDCM-OCT-S04-C348 were recruited from all the metagenomic databases (Table 2-8) with most recruited sequences being similar to ORFs coding proteins involved in DNA replication or viral structure. For example, ORF99, which is believed to encode a structural protein, shows similarity to many reads from viral metagenomic data. In addition, recruited reads from the GOS database for different phages indicates that S-EIV1-like phages are as prevalent as phages similar to Vibriophage N4, Roseophage SIO1, and Cyanophages P60, syn5, SCBP3 and PSSP7 (Figure 2-7). This comparison overestimates the prevalence of S-EIV1 relative to other cyanophages, because each recruited read was only assigned to the phage with the most similar sequence. Hence, for phages with very similar sequences, such as the P60-like marine cyanophages (cyanophages P60, S-CBP3, P-SSP7 and P-RSP2), which share many core genes, only one phage genome would recruit the read. As there are several representatives of marine cyanopodoviruses (P-60 like viruses), but only one S-EIV1-like phage, the overall effect is to dilute the number of reads assigned to each P60-like cyanophage. Regardless, the data indicate that S-EIV1-like phages are widespread and abundant in aquatic systems.

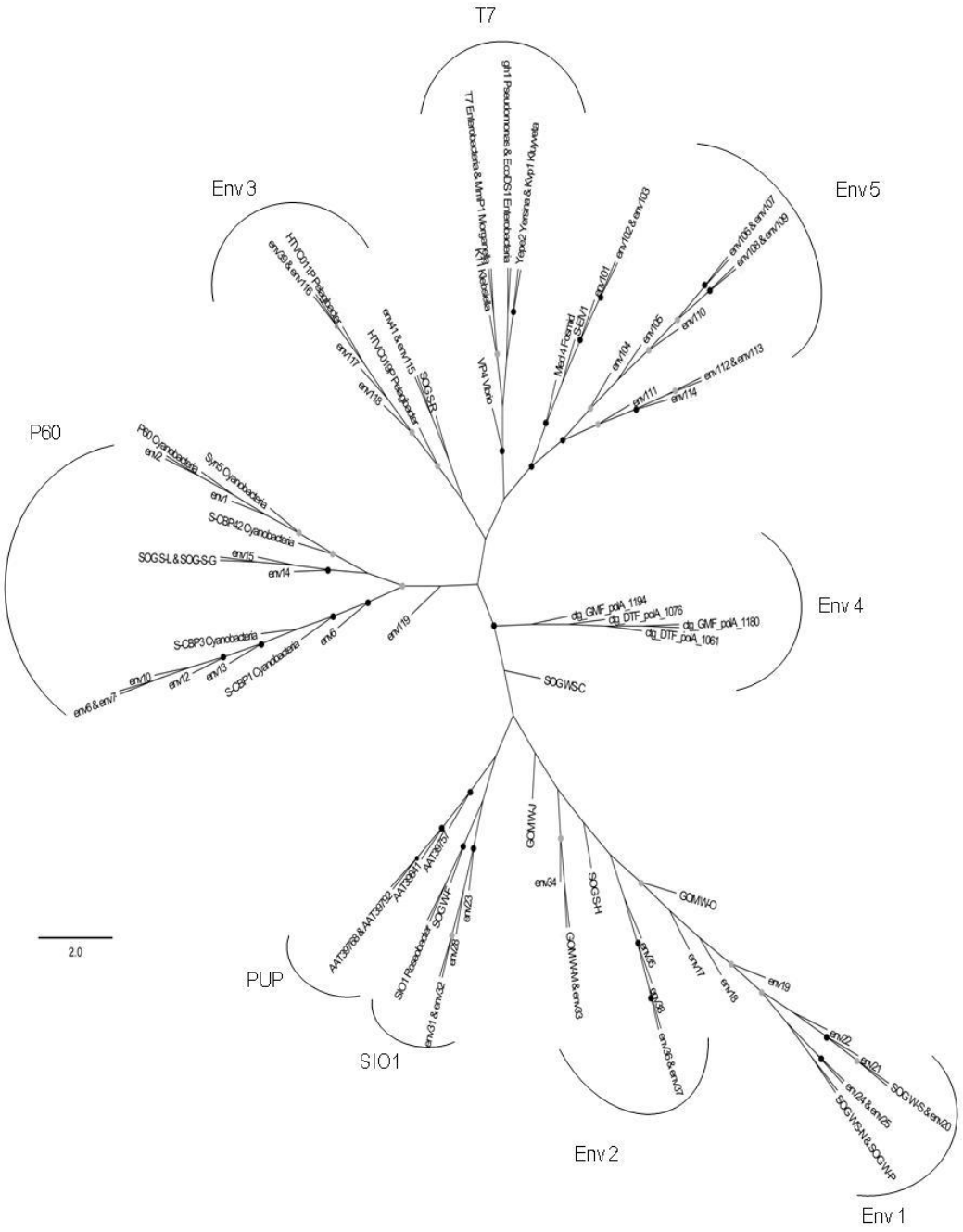


Figure 2-7.Unrooted maximum likelihood amino-tree of DNA polymerase A with known phages, metagenomic sequences and amplicons from degenerate primers from Labonté et al. (83) **and Breitbart et al.** (84)**.**

Bootstrap values are indicated as black (90 to 100%) or grey (70 to 89%) dots at the nodes. The groups are labeled as follows: P-60, marine cyanopodoviruses; T7, Enterobacter podoviruses; SIO1-like, *Roseobacter* viruses. PUP clade was identified by Breitbart et al. (105), Env clade 1-3 were identified by Labonté et al. (104). Env-4 was identified by Schmidt et al. (116)**. Env 5 is a new environmental clade including cyanophage S-EIV1. Scale bar represents amino acid substitutions per site.**

Table 2-9. Number of reads recruited from each metagenomic database for Cyanophage S-EIV1 and the uncultured phage sequence, MedDCM-Oct-S04-C348.

Metagenome Name	Location	S-EIV1 # reads	Med # reads	Reference
Global Ocean Survey	Various Locations	5552	2766	(139)
Tampa Bay Induced phages	Tampa Bay	286	75	(91)
Marine Virome	Sargasso Sea, Bristish Columbia, Arctic Oceans and Gulf of Mexico	400	201	(89)
Reclaimed Water Virus	South West Florida, USA	51	19	(92)
Virus Stromatolites	Mexico	19	10	(94)
Lake Limnopolar	Lake Limnopolar, Antarctica	95	61	(140)
Virus spring	Octopus and Bear Paw Hot Spring, Yellowstone National Park, USA	9	5	(141)

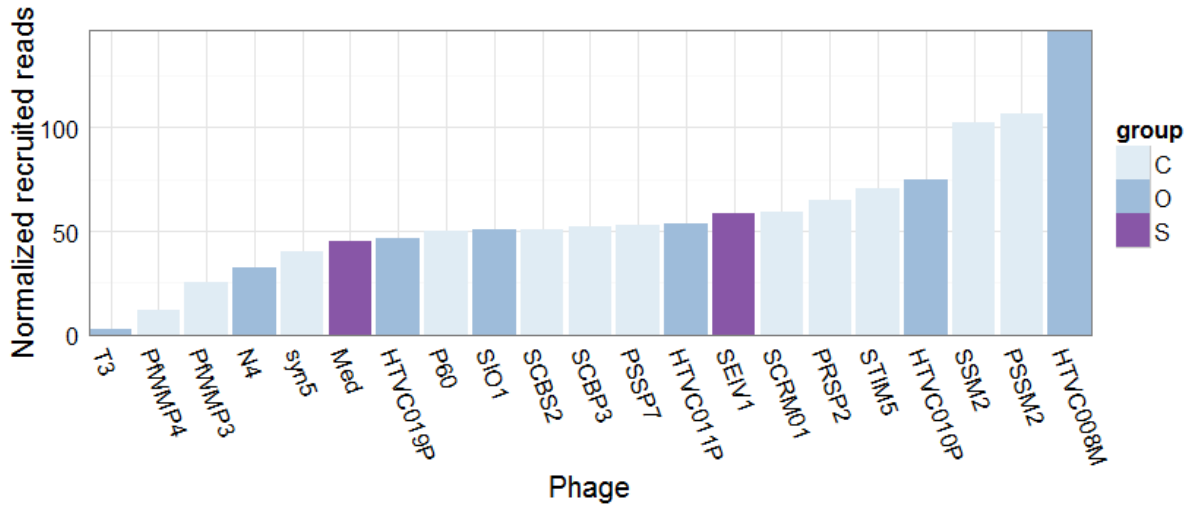


Figure 2-8. Abundance of Cyanophage S-EIV1 relative to other phages (including cyanophages, pelagiphages, roseophage, vibriophage and enterobacteriophage) using the Global Ocean Survey database

The number of reads was normalized to the number of ORFs in each genome. The bars are coloured as follows: light blue, cyanophages (C); dark blue, other phages (O); purple, S-EIV1-like phages (S).

2.5 Concluding remarks

S-EIV1 infects a polar isolate of *Synechococcus*, and represents a previously unknown lineage of cyanophages. Metagenomic data indicate that related viruses are relatively abundant and widespread in aquatic systems. Given the importance of picocyanobacteria for primary production in marine and fresh waters, this new viral lineage may play a major ecological role. With an icosahedral head and short non-contractile tail, S-EIV1 morphologically resembles viruses belonging to the family *Podoviridae* but its genome bears little resemblance. S-EIV1 lacks a number of core genes found in other podoviruses infecting cyanobacteria, including those encoding RNA polymerase, and most of the conserved structural proteins. The genomic organisation is also different, and unlike podoviruses, S-EIV1 encodes genes on both DNA strands. These features, along with phylogenetic analyses of DNA polymerase A and the terminase large subunit indicate that S-EIV1 represents a previously unknown evolutionary group of bacteriophage.

Chapter 3: A new lineage of viruses infecting freshwater filamentous cyanobacteria contain a distinct DNA polymerase

3.1 Synopsis

Nostoc spp. are ecologically important cyanobacteria that are widespread in freshwaters, and here, we present the first genome sequences for cyanophages infecting this genus. The viruses infect *Nostoc* sp, PC7210; their 68,304 bp and 64,960 bp genomes have G+C contents of 38.3 and 35.3%, respectively. Their genomes are not similar to those of other sequenced viruses; 80 and 77% of the predicted ORFs for A-1(L) and N-1, respectively, have no recognizable similarity to other sequences in databases. However, many of the coding sequences for proteins with known functions are highly similar to those found in filamentous cyanobacteria, showing a long evolutionary relationship with their host. Both phages contain a distinct DNA polymerase B that is closely related to those found in plasmids of the cyanobacteria *Cyanothece* sp. PCC7424 (plasmid pP742402), *Nostoc* sp. PCC7120 (plasmid pPCC7120beta) and *Anabaena variabilis* ATCC29413 (plasmid C). Together, these polymerase sequences form a distinct group that is more related to proteobacterial DNA polymerases than those found in other viruses, suggesting it was acquired from a proteobacterium by a virus and then transferred to the cyanobacterial plasmid. As well, many ORFs are similar to a prophage-like element identified in the genome of Nostoc PCC7524. The sequencing of the *Nostoc* phages revealed the history of numerous gene transfers between these viruses and their hosts, which helped forge the evolutionary trajectory of this previously unrecognized group of phages.

3.2 Introduction

Bacteriophages are the most abundant biological entities, with typical abundances of 10 to 100 million per mL in marine and fresh waters; they usually outnumber their potential hosts by an order of magnitude (Suttle 2007; Suttle 2005, Weinbauer 2004). Phages affect biogeochemical and ecological processes by facilitating nutrient cycling, maintain bacterial biodiversity, mediate microbial mortality and genetic transfer, and represent a vast source of uncharacterized genetic diversity (45, 47, 142).

Tailed bacteriophages with icosahedral heads and containing double-stranded DNA (dsDNA) are commonly isolated from aquatic environments. In particular, myoviruses, characterized by having a contractile tail, and which infect marine unicellular cyanobacteria from the genera *Synechococcus* and *Prochlorococcus*, have been relatively well studied. Cyanophages infecting *Synechococcus* spp. can occur at abundances $>10^5$ ml^{-1} in coastal seawater (Suttle and Chan, 1994; Waterbury and Valois, 1993; Sullivan et al., 2003). They are morphologically similar to, and share about 40 genes with T4-like phages infecting coliform bacteria (68, 143). They also possess many open reading frames (ORFs) that have only been found in cyanophages, as well as a number of host-derived genes, including those coding for proteins involved in photosynthesis, the pentose-phosphate pathway and phosphate acquisition(55, 73, 124, 133) .

There are few data for cyanoviruses infecting freshwater cyanobacteria. The sparse information suggests that freshwater myoviruses are not necessarily closely related to marine cyanomyoviruses. For example, S-CRM01 infects a freshwater *Synechococcus* sp., and is closely related to myoviruses infecting marine *Synechococcus* spp. (65); whereas Ma-LMM01 infects the freshwater toxic bloom-forming cyanobacteria *Microcystis aeruginosa* and has little similarity with previously sequenced myoviruses (74). Similarly, primers designed to amplify

sequences encoding the major capsid protein of T4-like phages, including marine cyanophages (Filée et al. 2005), do not amplify DNA from myoviruses infecting filamentous cyanobacteria assigned to *Nostoc spp*. and *Anabaena spp.* (80).

Filamentous cyanobacteria of the genera *Nostoc* and *Anabaena* are abundant and active members of freshwater and terrestrial microbial communities. They are found in various habitats including ice-covered polar lakes (7, 144) hypertrophic coastal lagoons (145), rice paddy soils (146) and rock-pool communities in Karst regions (147). They can fix nitrogen and form symbiotic associations with a wide range of plants and fungi (147).Despite the widespread distribution and ecological importance of *Nostoc* spp. and *Anabaena* spp. (147), a genomic analysis for cyanophages infecting these genera has not been reported.

In this chapter, I analyse the genome sequences of myoviruses infecting the freshwater filamentous cyanobacterium *Nostoc* sp. PCC 7120. The results demonstrate that cyanophages A-1(L) (148), and N-1 (Adolph, 1971; Adolph and Haselkorn, 1972), are distantly related to other phages that have been genetically characterized. Few of the predicted coding genes were similar to those found in other cyanophages. Predicted coding genes involved in DNA metabolism were similar to those found in filamentous cyanobacteria.

3.3 Material and methods

3.3.1 Cyanophage isolation, purification, DNA preparation and genome sequencing

Cyanophages A-1(L) and N-1, as well as their host, *Nostoc* sp. strain PCC 7120, were obtained from the American Type Culture Collection. Cyanophages were amplified and purified, as follows: 800 ml batch cultures of *Nostoc* sp. strain PCC 7120 was grown in BG-11 medium (103) in 1 L Erlenmeyer flasks under constant illumination (33 µmol photons $m^{-2}s^{-1}$

photosynthetically active radiation) at 26ºC with constant shaking at 75 rpm. Exponentially growing cultures were infected with either A-1(L) or N-1 left for 4 to 7 d until transparent, indicating lysis. To prevent phage binding to the filter, sodium chloride was added to the lysate at a final concentration of 0.5 M and incubated at 4°C for 1 h before filtration once through a 1.2 μm pore-size GC50 glass-fiber filter (Advantec MFS, Dublin, CA), and then twice through GVWP 0.22 μm pore-size polyvinylidine low-protein binding filters (Millipore, Bedford, MA). Subsequently, the viral particles were concentrated using polyethylene glycol (PEG) precipitation (150). Briefly, the filtered lysate was centrifuged at 10,000 xg for 10 min in a Sorvall RC-5C centrifuge (GSA rotor, 4°C*)* to remove cellular debris. PEG 6000 was added to the supernatant to a concentration of 10 % solution and incubated overnight at 4°C with constant shaking. The PEG solution was centrifuged at 16,000 xg for 20 min in a Sorvall RC-5C centrifuge (GSA rotor, 4°C*)*, the supernatant removed and the pelleted viruses re-suspended in 200 μL of BG-11 medium.

DNA was extracted by treating the re-suspended pellet with DNase 1 and RNase A to remove free nucleic acids, and using the QiAamp MinElute Virus Spin Kit (Qiagen, Mississauga, ON) according to the manufacturer's instructions. The DNA was sequenced using 454 GS FLX Titanium pyrosequencing at the Génome Québec, Innovation Centre, McGill University (Montréal, Canada). For each phage, more than 36,000 reads with an average length of ~350 bp were assembled into three contiguous sequences (contigs) using the GS De Novo Assembler (Roche), and closed into a single circular contig by PCR. The sequencing coverage was approximately 179 fold for A-1 and 250 fold for N-1.

3.3.2 Genome annotation

Open reading frames (ORFs) were predicted using GeneMark (106) and GLIMMER (107). To create the final predictions, the ORF calls from the two programs were combined. For ORFs predicted by both programs that differed in size, the longer of the two was kept. The final set of predicted ORFs was translated and assigned putative functions by comparing them with known protein sequences found in the GenBank (nr database), Acclame and Procite databases using the BLASTp program. The ORFs were considered to be homologous to a protein-encoding gene if the e-value was less than 10^{-3}. Identification of tRNA genes was performed using tRNAscan-SE (108). Putative promoter motifs were identified using PHIRE (110) and the default parameters of 20-mer DNA sequences (S) with 4 base pair degeneracy (D=4). The motif was considered as a putative promoter if it was found in the 150-bp region immediately upstream of the start codon of a predicted protein-coding gene. Sequence logos of the motifs were created with Weblogo using the alignment of their sequence (151). Identification of rho- independent terminators was performed with FINDTERM (Softberry, Inc). The default energy threshold was set to -16 kCal for the analysis as previously described (56).

3.3.3 Phylogenetic analysis

The open reading frames with similarity to genes coding for deoxycytidine triphosphate deaminase **(**DCTP deaminase), DNA polymerase B, and the large terminase subunit were used for phylogenetic analysis. Inferred amino-acid sequences for the DCTP deaminase and DNA polymerase B were aligned in ClustalX using default parameters, while the large terminase gene (terL) was aligned using the Promals web server (112, 113) with default parameters. Geneious v4.7 (114) was used to manually refine the alignment and construct neighbour-joining trees. Maximum likelihood trees (ML) were constructed with the RAxML Web-Server rapid

bootstrapping and ML search (100 replicates) (115) assuming the James-Taylor Thornton model of substitution using empirical base frequencies and estimating the proportion of invariable sites from the data.

3.3.4 Gene comparison with other cyanomyoviruses

Predicted ORFs were compared to a database for T4-like phages infecting heterotrophic bacteria (10), and marine (16) (68) and freshwater (1) cyanobacteria (65). An ORF was considered as shared if the e-value was $< 10^{-3}$.

3.3.5 Metagenomic analysis

The genomes of A-1(L) and N-1 were used to recruit similar fragments from different metagenomic data sets available from CAMERA (http://camera.calit2.net). BLASTn ($e > 10^{-3}$) was used to compare the genomes against the viral data in the CAMERA and NCBI environmental databases.

3.4 Results and discussion

Although, Cyanophages A-1 and N-1 are morphologically similar to T4-like cyanophages (62, 148, 149), the genomic sequences of Cyanophages A-1 and N-1 that infect *Nostoc* sp. strain PCC 7120 indicate they represent a previously unknown lineage of viruses. Few of the predicted coding genes were similar to those found in other phages, while the DNA polymerase B sequences were similar to those found in a host plasmid.

3.4.1 Genome features

The dsDNA genomes of cyanophages A-1 and N-1 were 68,304 and 64,960 bp, with G+C content of 38.3 and 35.4 %, respectively. Their genome lengths are about half those of described cyanomyoviruses. Members of the *Myoviridae* generally have larger genomes than other phage families. For example, the coliphage T4 has a genome of 168 kb (152), while cyanomyoviruses

typically have genomes ranging from 161 to 231 kb in length (55–57, 65, 67, 74). Myoviruses with much smaller genomes have been less studied, although 'dwarf' myoviruses with genome sizes of less 50 kb that infect a diverse range of proteobacteria (i.e. *Aeromonas salmonicida*, *Vibrio chlolerae*, *Bdellovibrio* spp. and *Pectobacterium caratovorum*) were recently characterized (153). However, these viruses do not have sequence similarity to A-1 (L) and N-1.

Bioinformatic analysis of A-1 and N-1 revealed that only about a quarter of translated ORFs had a similarity to protein sequences in current databases (40 of 153 for A-1(L), and 33 of 141 for N-1, e-value < 10^{-3}). In contrast, often about 80% of ORFs in virus isolates are reported to have recognizable homologues (154), while in marine cyanophages the percentage is typically >60%(67, 68). However, for other freshwater cyanophages the percentage of predicted proteins has been similarly low, indicating the lack of representative genomes. For example, about 61% and 76% of translated ORFs for the cyanomyoviruses S-CRM01 (65) and Ma-LMM01(74), respectively, and 65% for the cyanopodoviruses Pf-WMP4 and Pf-WMP3 ORFs (63, 66) did not have similarity to amino-acid sequences in databases.

Translated ORFs with significant hits to proteins of known functions were associated with DNA replication, DNA metabolism and repair, and structural components (Figure 3-1, Appendix F and Appendix G). Cyanophage A-1 (L) also contained putative genes coding for a transposase and two phage anti-repressors that are associated with a lysogenic life style.

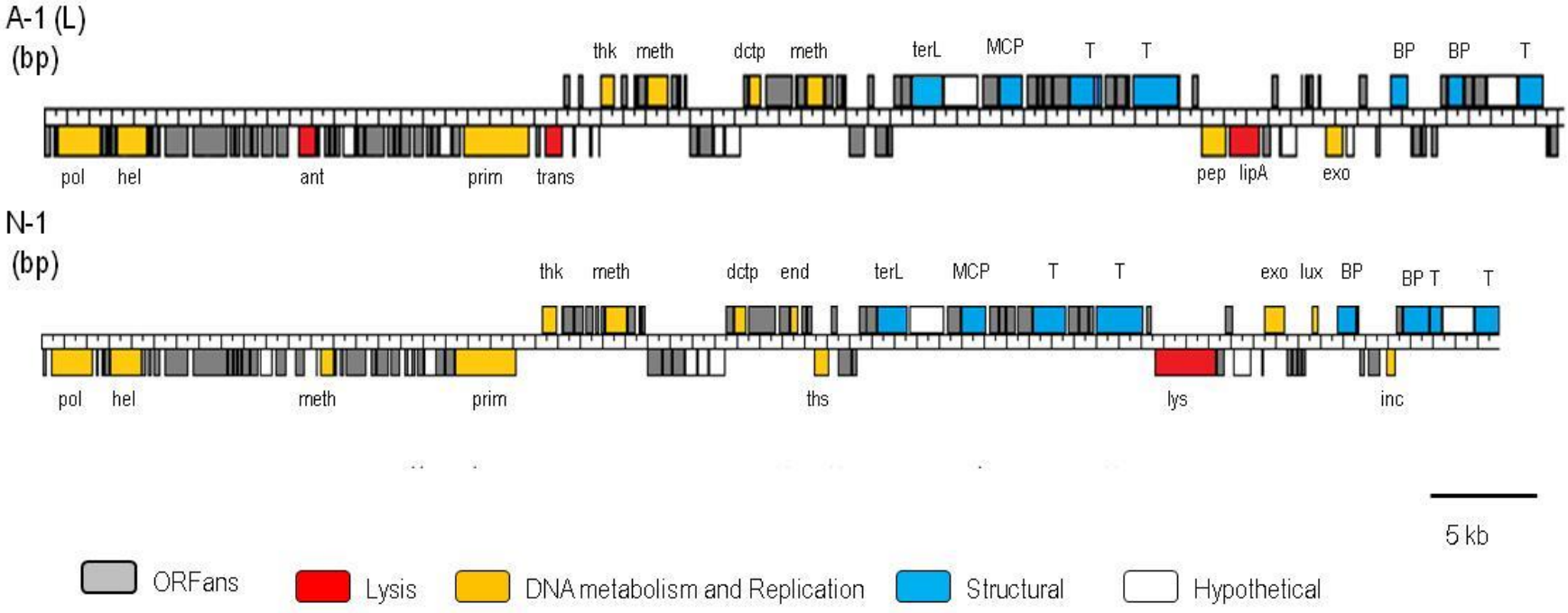


Figure 3-1. Comparative genomics of the two Nostoc cyanomyoviruses.
The genomes are presented as linear molecules for better comparison: A) Cyanomyovirus A-1(L) and B) Cyanomyovirus N-1. Gene abbreviations and functions are as follows: ant, anti-repressor; BP, baseplate; dctp, dCTP deaminase ; exo, exonuclease; hel, helicase; incl, viral A inclusion factor; lipA, lipoprotein A, luX= LuxR transcription factor, lyz, lysozyme; MCP, major capsid protein; meth, methylase; pol, DNA polymerase; prim, primase; T, tail; terL, terminase large sub-unit; thk, thymidylate kinase; ths, thymidylate synthase; trans, transposase.

Although there was high similarity between the *Nostoc* cyanophages (Appendix H), they had little similarity to cyanophages infecting other cyanobacteria. Genomic studies on 26 T4-like phages, including 16 marine cyanophages, identified 38 T4-like core genes coding primarily for structural and DNA replication proteins (68). In contrast, only six T4-like core genes have significant similarity to ORFs in the *Nostoc* cyanophages (Table 3-1), and those core genes are associated with replication and DNA modification. In addition, none of the 25 genes found exclusively within marine cyanomyoviruses were present in the *Nostoc* cyanomyoviruses.

Table 3-1. Predicted ORFs in cyanophage A-1 and N-1 with similarity to T4-like genes.

	CyanophageA-1 (L)		**Cyanophage N-1**	
T4-core genes	**e***	**Phage hit**	**e***	**Phage hit**
gp5 baseplate hub+tail lysozyme	$4.0e^{-5}$	Phage RB49	$4.0e^{-5}$	Cyanophage S-PM2
gp17 terminase DNA packaging enzyme large subunit	$4.0e^{-2}$	Cyanophage S-PM2	$4.0e^{-2}$	Cyanophage S-PM2
gp41 DNA primase-helicase	$2.0e^{-13}$	Cyanophage S-PM2	$1.0e^{-11}$	Cyanophage S-RSM4
gp43 DNA polymerase B	$4.0e^{-5}$	Phage Aeh1	$9.0e^{-6}$	Phage Aeh1
thymidylate synthase	$1.0e^{-35}$	Cyanophage S-CRM01	$9.0e^{-32}$	Cyanophage P-SSM4

*e-value

3.4.2 Regulatory elements and motifs

Both genomes were analyzed for regulatory elements and motifs such as tRNA genes, promoter motifs and transcriptional terminators. Unlike in other cyanomyoviruses, but not unusual for myoviruses with genomes smaller than 70 kb, tRNAs genes were not identified in cyanophages A-1 (L) and N-1. A recent study found that 7 of 10 small myoviruses did not

contain tRNAS (153). In addition, the presence or absence of tRNAs might influence host range, as cyanomyoviruses that only infect *Prochlorococcus* spp. tend to code for few, if any, tRNAs; whereas, those infecting both *Prochlorococcus* spp. and *Synechococcus* spp. or just *Synechococcus* spp., code for more tRNAs (155). In contrast, early studies described that cyanophage A-1(L) and N-1 displayed a broad host range and could infect members of both *Anabaena* spp. and *Nostoc* spp. (156). However, the reported broad host range might reflect errors in taxonomy (52, 157).

Putative promoter motifs in the genomes for the *Nostoc* cyanomyoviruses were identified using PHIRE (110), and were different from each other. For A-1(L), 43 motifs were identified that contained a consensus sequence for a putative promoter motif consisting of two highly conserved regions separated by a base pair (Figure 3-2). For N-1, 21 motifs were identified that included a highly conserved region of 11 bases pairs in the putative motif.

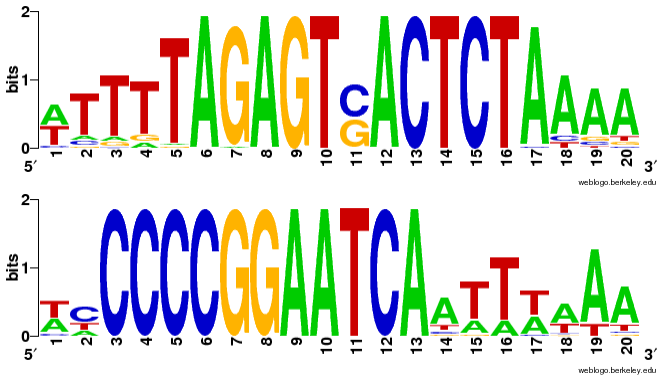


Figure 3-2. Sequence logo of the predicted promoter motifs predicted from alignments of the 5' upstream regions.
A) Putative promoter for cyanomyovirus A-1 (L) showing a sequence logo created from an alignment of 43 sequences. B) Putative promoter for cyanomyovirus N1 showing a sequence logo created from an alignment of 21 sequences. The height of each letter is proportional to the level of sequence conservation of the nucleotides at the respective positions.

3.4.3 Presence of a distinct DNA polymerase B

DNA polymerase type B (*pol* B) sequences were found in A-1 (L) and N-1. DNA polymerase catalyzes the polymerisation of deoxyribonucleotides into a DNA strand, and because there are a many viral and cellular homologues, it is a useful phylogenetic marker. The closest relatives to the *pol* B sequences in A-1(L) and N-1 are plasmids in the cyanobacteria *Cyanothece* sp. PCC 7424 (plasmid pP742402), *Nostoc* sp. PCC 7120 (plasmid pCC7120beta) and *Anabaena variabilis* ATCC 29413 (plasmid C) (Figure 3-3). These viral and plasmid *pol* B sequences form a distinct group that is related to proteobacterial and archaeal *pol* B clades. Moreover, while *pol* B is common in proteobacteria, the only representatives known in cyanobacteria are from these three plasmid sequences.

The close phylogenetic relationship between the *pol* B sequences in the phages and plasmids suggests that the gene was transferred laterally. Moreover, DNA polymerase B in the *Cyanothece* plasmid pP742402 is adjacent to a CRISPR, a region where recombination can occur and thus be promiscuous to gene exchange. DNA *pol* in A-1 (L), N-1 and cyanobacterial plasmids share a common ancestor with proteobacteria implying that the transfer of DNA to the cyanobacterial plasmids may have been phage mediated.

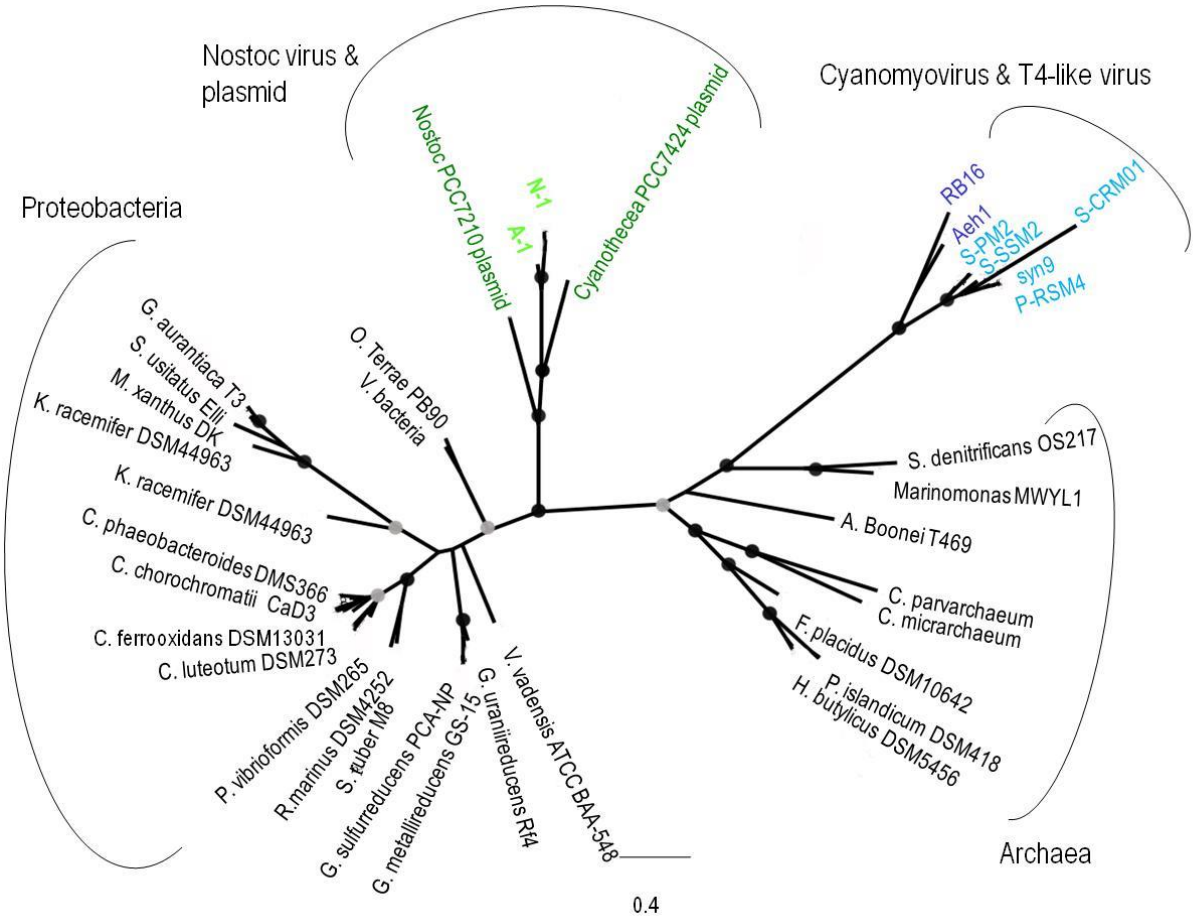


Figure 3-3. Unrooted maximum likelihood phylogenetic tree of DNA polymerase B protein sequences found in viruses, bacteria and archaea.
Bootstrap values corresponding to between 90 and 100 % (black circles) and 75 and 89 % (grey circles) are shown at the nodes. The sequence names are coloured as follows: black (Bacteria and Archaea), light blue (marine cyanomyoviruses), dark blue (T4-like myoviruses), light green (Nostoc phages) and dark green (cyanobacterial plasmid). Scale bar represents amino acid substitutions per site.

3.4.4 Phylogeny of the terminase large subunit

Further evidence of the evolutionary divergence of A-1 and N-1 from other cyanophages is provided by the gene (*terL*) encoding the terminase large subunit, a protein involved in DNA packaging in dsDNA phages. Phylogenetic analysis reveals that the translated *terL* sequences from A-1 and N-1 cluster separately from those in other viruses (Figure 3-4), and branch most

closely with those in the freshwater heterocystous cyanobacterium *Nostoc* sp. PCC7524, and the marine tropical and subtropical unicellular N-fixer, *Crocosphaera watsonii*, with the marine *Synechococcus* sp. PCC9605 being somewhat more distant. It has been argued that terminases are good phylogenetic markers for phage evolution and that sequences found in cyanobacteria may be remnant prophage (70). Indeed, the gene for the terminase large subunit in *Crocosphaera watsonii* occurs near genes encoding for a phage tail collar (EAM53192), a transposase (EAM53191), and a hypothetical protein (EAM53190) that also show similarity to putative genes in the *Nostoc* cyanophages. The terminase large subunit in *Nostoc* PCC7524 is also part of a prophage-like element (see below). The A-1 and N-1 terminase sequences have revealed a new evolutionary group of phage terminases with similarity to prophage elements in several genera of divergent cyanobacteria, suggesting that relatives of A-1 and N-1 infect a broad range of marine and freshwater hosts.

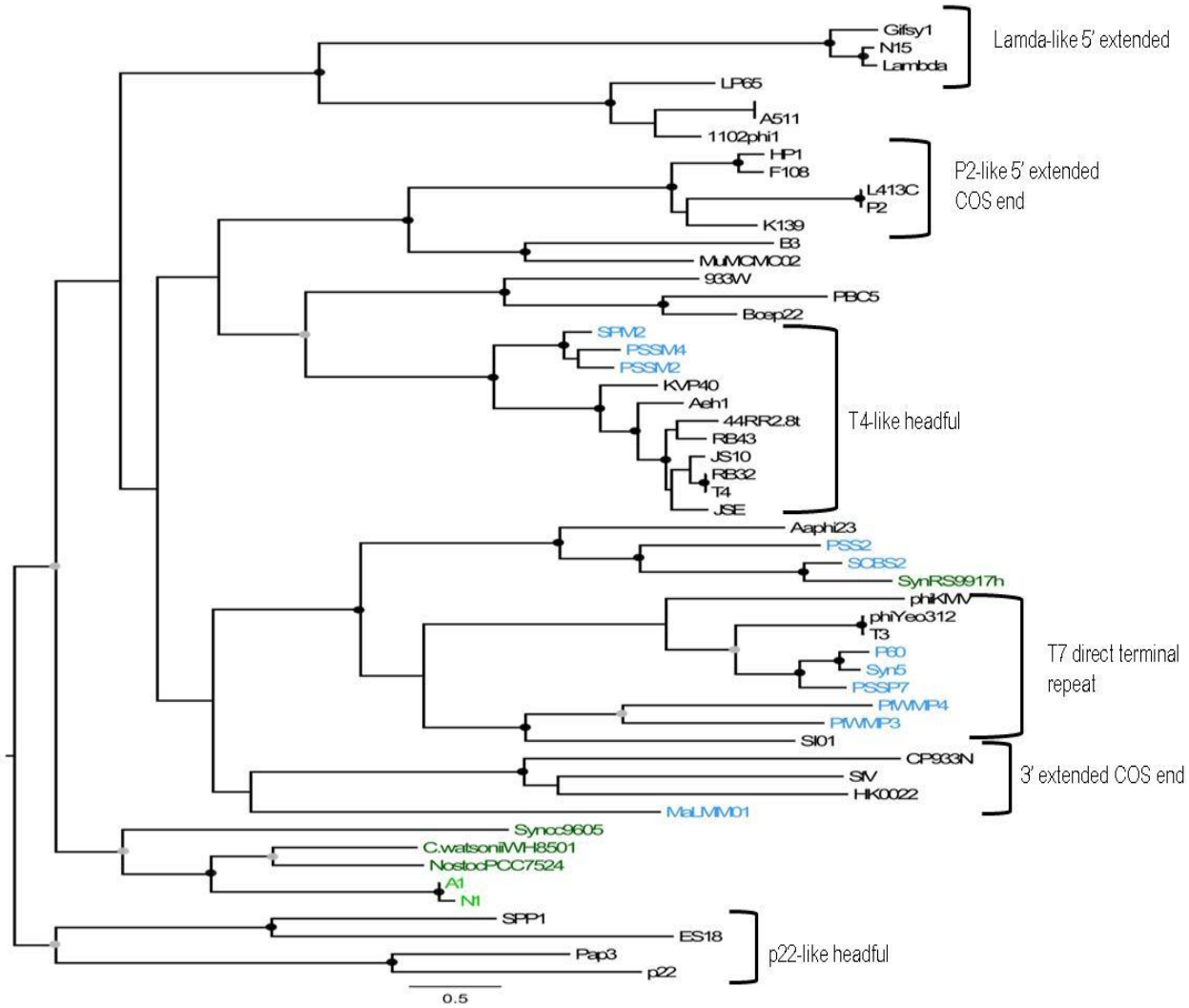


Figure 3-4. Phylogenetic relationship of terminase large subunit (terL) protein sequences found in phages.
A maximum likelihood tree is shown and the bootstrap values are on the nodes of the branches. Bootstrap values corresponding to between 90 and 100 % (black circles) and 75 and 89 % (grey circles) are shown at the nodes. The sequence names are coloured as followed: (other), light green (Nostoc cyanophages), dark green (cyanobacteria), light blue (other cyanophages), black (other viruses). Scale bar represents amino acid substitutions per site.

3.4.5 Genetic exchange between filamentous cyanobacteria and Nostoc cyanophages

Some ORFs in A-1 and N-1 exhibit high similarity to genes in cyanobacteria that code for proteins with known functions (Table 2), including purine and pyrimidine metabolism. One example is deoxycytidine triphosphate (dCTP) deaminase, an enzyme involved in the production

of dUMP, the immediate precursor of thymidine nucleotides. Phylogenetic analysis demonstrates that dCTP deaminase sequences from the *Nostoc* cyanophages are more similar to those found in cyanobacteria; whereas, the dCTP deaminase homologue in the marine cyanophage S-PM2 is more closely related to other virus sequences (Figure 3-5). Thymidylate synthase and thymidylate kinase genes in A-1 and N-1 were also similar to host genes. The proteins encoded by these genes likely catalyze two subsequent steps in deoxythymidine triphosphate (dTTP) synthesis. Thymidylate synthase is involved in the production of thymidine monophosphate (dTMP), while thymidylate kinase phosphorylates dTMP to dTDP. This reaction is crucial to both the *de novo* synthetic and the salvage pathways for pyrimidine deoxyribonucleotides. The gene encoding thymidylate kinase is commonly found in eukaryotes and their DNA viruses, and has been reported from some myoviruses with genome sizes > 200 kb (158), however, to our knowledge this is the first time a homologue of this gene in A-1(L) and N-1 has been reported in a cyanophage, or in phages with genomes < 70kb.

Table 3-2. Predicted ORFs with high similarity to cyanobacterial genes for cyanophage A-1

ORFs	Length (bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
5	2016	R	DNA polymerase B	*Cyanothece* PCC7425	$7.00e^{-115}$	40.4% (237)
24	939	R	putative ant AntA/AntB antirepressor	*Leptolyngbya* sp. PCC 7375	$1.90e^{-24}$	45%(51)
28	618	R	DNA N-6-adenine-methyltransferase	*Synechocystis* sp. PCC 7509	$1.00e^{-06}$	27.3%(44)
51	1209	F	Transposase	*Nostoc* sp. PCC7120	0	388(100%)
53	285	F	Hypothetical protein	*Anabeana variabilis* ATCC 29413	$5.00e^{-31}$	
59	417	F	Hypothetical protein	*Calothrix* sp. PCC 7103	$2.59e^{-48}$	58% (98)
62	1044	F	DNA-cytosine methyltransferase	*Thrichodesmium erythraeum* IMS101	$3.0e^{-62}$	37.8%(140)
76	600	F	dCTP deaminase/dUTPase superfamily	*Cyanothece* PCC7425	$2.00e^{-76}$	67.3%(134)
81	888	F	DNA methylase N-4/N-6 domain protein	*Arthrospira maxima*	$1.00e^{-65}$	49.4%(133)
83	432	F	endodeoxyribonuclease RusA	*Cyanothece* PCC7425	$8.00e^{-04}$	28.4%(29)
93	1359	F	Terminase large subunit	*Nostoc* sp. PCC 7524 AFY48994.1	$2.28e^{-54}$	30% (129)
98	1605	F	Hypothetical protein	*Nostoc* sp. PCC 7524 AFY48995.1	$7.16e^{-17}$	21%(113)
107	1521	F	Tail sheath protein	*Nostoc* sp. PCC 7524 AFY49006.1	$1.14e^{-69}$	38% (137)
115	1416	R	Lysozyme-like domain,Rare lipoprotein A	*Anabeana variabilis* ATCC 29413	$2.3e^{-19}$	57.7%(60)
121	738	R	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49010.1)	$9.39e^{-11}$	29%(67)
124	891	R	Exonuclease RNase T and DNA polymerase	*Cyanobacterium aponinum* PCC10605	$1.63e^{-08}$	27% (48)
129	813	F	Phage-related baseplate assembly protein	*Nostoc* sp PCC7524 (AFY49015.1)	$1.46e^{-32}$	34%(91)
136	738	F	BaseplateJ phage tail	*Nostoc* sp PCC7524 (AFY49018.1)	$3.76e^{-33}$	41% (94)
139	576	F	Phage tail protein (tail_P2_I)	*Nostoc* sp PCC7524 (AFY49020.1)	$5.63e^{-40}$	59%(77)
140	1389	F	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49021.1)	$4.83e^{-57}$	37%(157)
142	1140	F	Tail collar protein	*Nostoc* sp PCC7524 (AFY49022.1)	$2.1e^{-52}$	50% (149)

Table 3-3. Predicted ORFs with high similarity to cyanobacterial genes for cyanophage N-1

ORFs	Length (bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
2	1896	R	DNA polymerase B	*Cyanothece* sp. PCC 7424	7.0e-^{118}	41.5%(243)
23	600	R	Hypothetical protein	*Nostoc punctiforme* PCC 73102	1.10e^{-06}	21.4%(40)
30	633	R	C-5 cytosine-specific DNA methylase	*Nostoc punctiforme* PCC 73102	1.81e^{-05}	36.2%(34)
51	627	F	Thymidylate kinase	*Lyngbya* PCC8106	2.45e^{-23}	36%(74)
61	1047	F	DNA-cytosine methyltransferase	*Anabaena variabilis* ATCC 29413	1.02e^{-56}	33.2(127)
70	579	R	Hypothetical protein	*Thermoanaerobacter italicus* Ab9	8.67e^{-03}	32.7%(33)
72	507	R	Hypothetical protein	*Calothrix* sp. PCC 7103	1.73e-^{47}	56%(94)
74	726	R	Hypothetical protein	*Calothrix* sp. PCC 7103	4.94e^{-06}	30.4%(78)
76	585	F	dCTP deaminase	*Synechococcus* sp. PCC 7335	1.45e^{-69}	65.5%(131)
84	741	R	Thymidylate synthase complementing protein/FAD-dependent thymidylate synthase	*Chlorobaculum parvum* NCIB 8327	8.35e^{-33}	39.1%(86)
92	1359	F	phage terminase, large subunit	*Nostoc* PCC7524	2.32e^{-55}	31%(134)
95	1602	F	Hypothetical protein	*Nostoc* PCC7524 (AFY48995)	3.55e^{-18}	22%(106)
103	330	F	Hypothetical protein	*Nostoc* PCC7524 (AFY49001)	9.62e^{-06}	30%(30)
106	1521	F	Tail sheath protein	*Nostoc* PCC7524 (AFY49006)	3.74e^{-74}	39%(138)
116	2694	R	Lysozyme-like domain, rare lipoprotein A (RlpA)-like double psi beta barrel	*Nostoc* PCC7524 (AFY49014)	1.48e^{-37}	28 % (137)
121	849	R	Hypothetical protein	*Nostoc* PCC7524 (AFY49010)	9.82e^{-10}	27.9%(51)
125	918	F	Exonuclease RNase T and DNA polymerase III	*Thauera* sp MZIT	4.14e^{-04}	28.4%(38)
134	813	F	gp5 baseplate hub subunit and tail lysozyme	Acinetobacter phage Ac42	4.24e^{-07}	28.6%(30)
144	330	R	Lysosyme	*Nostoc* PCC7524 (AFY49017)	2.51e^{-03}	24%(32)
145	1167	F	Baseplate J phage tail protein	*Nostoc* PCC7524 (AFY49018)	4.18 e^{-68}	43%(144)
146	576	F	Phage tail protein	*Nostoc* PCC 7524 (AFY49020)	2.32e^{+00}	26.3% (45)
147	1395	F	Phage tail fiber protein	*Nostoc* PCC 7524 (AFY49021)	2.97e^{-64}	37%(172)
148	1155	F	Tail collar protein	*Nostoc* . PCC 7524 (AFY49022)	1.62e^{-47}	33%(137)

Sequences with similarity to putative genes encoding DNA adenine methyltransferase (dams) and DNA cytosine methyltransferases (dcm) were also found in the *Nostoc* phages. In general, DNA methyltransferases mediate post-replicative methylation at a specific recognition site, and protect bacterial DNA against digestion by specific restriction endonucleases; whereas, unmethylated infective DNA, such as in phages, is cleaved. However, DNA methyltransferases occur in some phages, and modify the viral DNA to be resistant to bacterial restriction systems. In general, phage DNA methyltransferases are similar to those of their hosts. The presence of DNA methyltransferase genes supports the observation that that A-1(L) tolerates dam-like and dcm-like methylation, and shares similar restriction enzyme cleavage resistance as for its host (159).

Putative coding sequences in A-1 and N-1, as well as in the freshwater cyanophage S-CRM01 (65) and the marine cyanophage S-PM2 (57), are similar to host-like genes encoding the rare lipoprotein A (*rlpA*) (Table 3-2, Table 3-3). Although its function is not known, *rlpA* was strongly induced during hyperosmotic stress in *Synechocystis* sp. PCC 6803 (160) and is upregulated as part of the general stress response of *Synechococcus* sp. WH8102 grown under low phosphate (161).

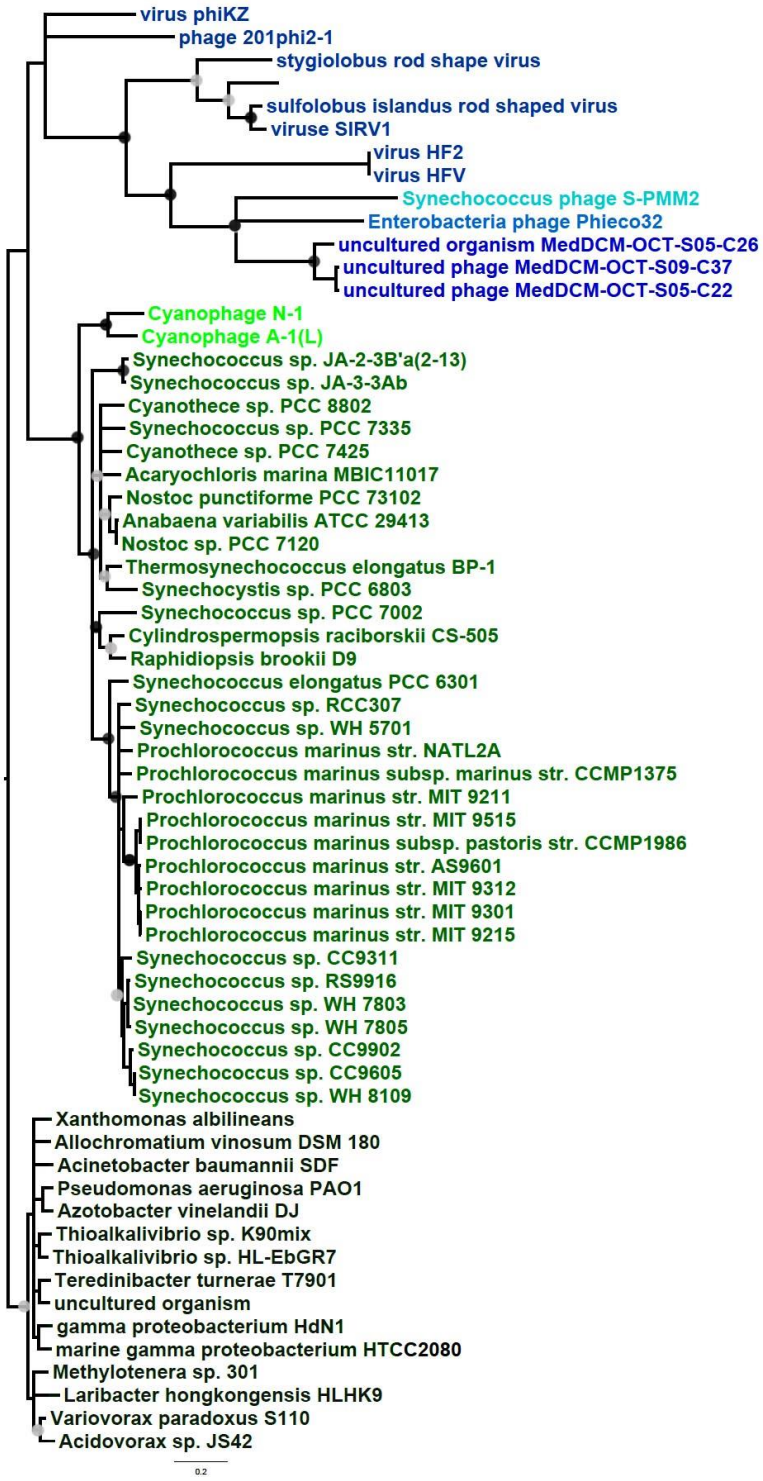


Figure 3-5. A maximum likelihood phylogenetic tree of dCTP deaminase protein sequences from viruses and bacteria.
Bootstrap values corresponding to between 90 and 100 % (black circles) and 75 and 89 % (grey circles) are shown at the nodes. Scale bar represents amino acid substitutions per site.

3.4.6 Nostoc cyanophage-related genes were also found in the genome of Nostoc PCC7524

Fourteen ORFs in cyanophages A-1 and N-1 that putatively encode structural proteins, terminase, lysozyme and peptidase had high similarity to ORFs in *Nostoc* sp. PCC7524 (NCBI Reference Sequence: NC019684.1), leading to the re-annotation of about 30 kb of sequence and the identification of a prophage-like element (Figure 3-6). Although lysogeny has been reported in natural *Synechococcus* communities (162, 163), few prophage-like elements have been detected in cyanobacterial genomes; for example, none were found in a dozen marine picocyanobacterial genomes (164, 165). However, evidence of lysogeny was recently found in the genomes of *Synechococcus elongatus* strains PCC 6301 and PCC 7942 (58). The lifestyles of cyanophage A-1(L) and N-1 have been reported as being lytic (62, 156), but similar prophage-like elements in *Nostoc* PCC7524 suggest that related phage have the potential for lysogeny. Moreover, the presence of ORFs in the host with high similarity to sequences in A-1 and N-1 indicates that genetic exchange occurs possibly via prophage integration or homologous recombination.

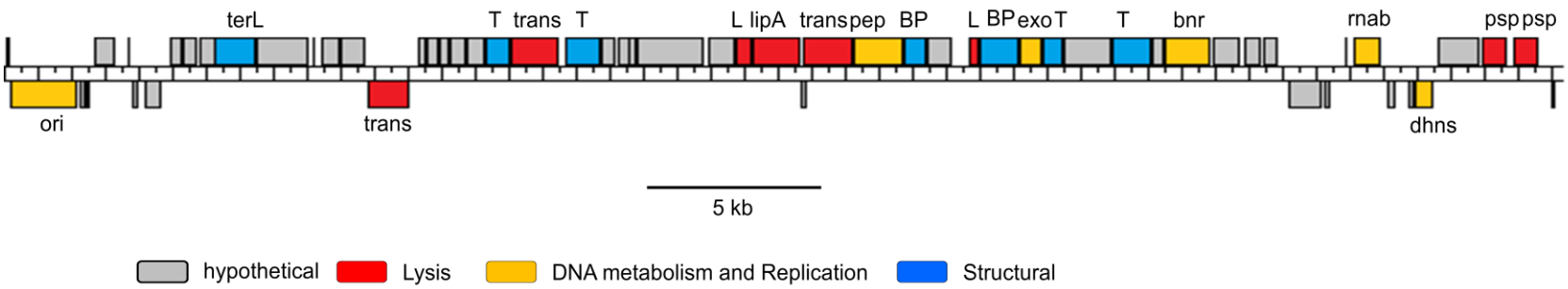


Figure 3-6. Genomic map of the prophage-like element in the genome of Nostoc PCC7524 (NC019684.1).
Gene abbreviations and functions are as follows: BP, baseplate; bnr,bnr repeat protein; dhns, dihydroxynaphtoate synthase; exo, exonuclease; H, hypothetical protein L, lysozyme; lipA, lipoprotein A; ori, origin of replication; pep, peptidase; psp, phage shock protein; rnab, RNA binding protein; T, tail; terL, terminase large sub-unit; trans, transposase.

3.4.7 Prevalence of Nostoc cyanophage in aquatic systems

No sequences homologues to the A-1(L) and N-1 genomes were found in viral metagenomic data on CAMERA (e.g. Marine Virome, Chesapeake Bay Virome, Reclaimed Water virus, Tampa Bay induced phages); however, in the NCBI environmental database nine matches to A-1(L) were found in a microbial mat metagenome, and three homologues to N-1 occurred in a stromatolite metagenome. Although there were few hits to the A-1 and N-1 genomes, environments in which *Nostoc* is an important member of the community are not well represented in environmental metagenomic datasets.

3.5 Concluding remarks

The cyanophages A-1(L) and N-1 that infect *Nostoc* sp. strain PCC 7120 belong to a previously unrecognized evolutionary lineage of tailed phages. Most of their predicted protein-coding genes have no obvious similarity to sequences in databases, and those that do are generally most similar to genes found in filamentous cyanobacteria. These findings indicate that lateral gene transfer between similar phages and their hosts have played an important role in forging the evolutionary trajectory of this previously unrecognized evolutionary lineage of phages.

Chapter 4: Cyanophage N-1 contains a functional CRISPR array

4.1 Synopsis

Clustered regularly interspaced short palindromic repeats (CRISPRs) and CRISPR-associated (*cas*) genes are part of an adaptive immune system that protects *Bacteria* and *Archaea* against foreign nucleic acids (166). The CRISPR array is a series of non-contiguous 20 to 50 bp direct repeats interspaced by non-repetitive spacers of 25 to 75 bp that are sequences usually derived from previously encountered viruses or plasmids (166–168). The CRISPR-Cas system provides immunity by recognising and cleaving incoming foreign genetic material that has sequence similarity to the spacers (166, 168–170). The patchy distribution of CRISPR-Cas system types has raised questions on their acquisition and transfer (171). Here, I report on a functional CRISPR array from a cyanophage that infects filamentous cyanobacteria from the ecologically important genera *Nostoc* and *Anabaena*. The CRISPR array has direct repeats with high similarity to the DR5 family that is commonly found in filamentous cyanobacteria. I show that the viral-encoded CRISPR is transcribed and transferred to its host. These findings indicate that not only can viruses serve as vectors for moving CRISPRs among cells, but that viruses likely carry CRISPRs to confer host-resistance to infection by competing phages and plasmids, thereby conferring a selective advantage to both the host and CRISPR-encoding phage.

4.2 Introduction

Bacteria and viruses have a shared evolutionary history stretching for billions of years that has led to a myriad of adaptations for cells to avoid infection and counter measures for phage to avoid them (172). One of these defense mechanisms is the clustered regularly interspaced short palindromic repeats (CRISPRs) and the CRISPR associated (*cas*) genes (166,

168) (i.e. CRISPR-Cas system). Evidence of this adaptive immune system is found in almost all archaeal genomes and in about 40% of bacterial genomes(173). The CRISPR array consists of a series of non-contiguous direct repeats (DR), 20 to 50 bp in length, that are separated by variable sequences (spacers) usually derived from viruses and plasmids (166–168). A leader region, which is an AT sequence of up to 550 bp, is directly adjoining the leader region. In some cases, CRISPR associated (*cas*) genes are found upstream or downstream of CRISPR arrays.

The CRISPR-Cas system provides immunity to the host cell by recognising and cleaving incoming foreign genetic material with sequence similarity to the spacers (166, 168, 169, 173). During an encounter of the host with a virus or plasmid, a Cas complex recognizes the foreign DNA and integrates a novel repeat-spacer at the leader end of the CRISPR. The newly acquired spacer matches the sequence of the virus (proto-spacer). Studies have shown that spacers are incorporated in a polar manner at the leader end of the array (174, 175). First, a comparison of CRISPR arrays for two related *Sulfolobus* strains showed conserved spacers at the downstream end and highly variable spacers at the upstream end (176). In addition, new spacers were incorporated at the leader end of the array in the surviving cells of *S. thermophilus* (175). The CRISPR array, therefore, represents a record of previous infection with the most recent encounter present at the upstream end and the more ancient infections further downstream.

The CRISPR array is constantly transcribed from the leader region by RNA polymerase. The CRISPR transcript, often described as the pre-CRISPR RNA, is further cleaved by CAS proteins at the base of the hairpin into a smaller CRISPR RNA (crRNA) that contains a single spacer and a partial repeat by Cas proteins. In the well-studied model CRISPR in *E. coli*, a set of five Cas proteins (Cas1-4 and Cas5e), more commonly defined as the CASCADE (CRISPR ASsociated Complex for Anti-viral DEfense) complex, is believed to be involved with the

processing of crRNA. The crRNA, together with specific Cas proteins, then form a CRISPR complex that interferes with invading nucleic acids (both DNA and RNA, depending on the system). Any incoming phage carrying a sequence identical to a spacer is inactivated and the phage infection process blocked (166, 168, 169, 173).

Many types of CRISPRs have been recognized, based on the sequence similarity of the repeats (177). The distribution of closely related CRISPR-Cas systems in phylogenetically distant organisms suggests exchange by horizontal gene transfer (171). The CRISPR-Cas systems in cyanobacteria have not been widely investigated; however, there is strong evidence for widespread distribution of this adaptive immune system within the group. A recent study found that 86 out of 126 sequenced cyanobacterial genomes contained CRISPR-Cas systems (178), although they were absent from a marine subclade that includes representatives from the genera *Synechococcus* and *Prochlorococcu*s. CRISPR-Cas systems are rare in plasmids and prophages. To date, only one viral-encoded CRISPR-Cas system has been described (179). In this case, the ICP1 phage CRISPR-Cas system is used to counteract a phage inhibitory chromosomal island of the bacterial host. In addition, it has been suggested that they could mediate the exchange of CRISPRs among organisms (180–182).

In this chapter, I describe a functional CRISPR array found in the genome of Cyanophage N-1. The CRISPR array has direct repeats with high similarity to the DR5 family that is commonly found in filamentous cyanobacteria. The viral-encoded CRISPR is transcribed and transferred to its host. These findings indicate that not only can viruses serve as vectors for moving CRISPRs among cells, but that viruses likely carry CRISPRs to confer host-resistance to infection by competing phages and plasmids, thereby conferring a selective advantage to both the host and CRISPR-encoding phage.

4.3 Material and methods

4.3.1 Identification and analysis of the CRISPR array

During the genomic analysis of N-1, a repeat DNA region was found and further classified as a CRISPR array using CRISPR finder (183). To show that this region was not a sequencing and/or assembly error, the CRISPR region was checked with PCR and sequencing. Two µL from a 0.22 µm filtrate N-1 lysate was added to a 48 µL PCR mixture containing Platinum *Taq* DNA polymerase assay buffer (50 mM KCl, 20 mM Tris-HCl, pH 8.4), 10 mM $MgCl_2$, 200 µM deoxyribonucleoside triphosphate, the primers sCRF and sCRR (0.25µM each sCRF=CAATTGGCAAAAGATTTAGCAGC and CR3R=GGGGAGAGGTTTGGAGAGGGGT) and 2.0 U of Platinum Taq DNA polymerase (Invitrogen, Carlsbad, Ca). Negative controls contained all of the reagents, but sterile water was used as the template. PCR was carried out as follows: denaturation at 94°C for 5 min, followed by 35 cycles of denaturation at 94°C for 30s , annealing at 57°C for 45s, extension at 72°C for 1 min, and a final extension at 72°C for 10 min (184). The amplification products were subjected to electrophoresis using 1.5% agarose–0.5xTris-borate-EDT buffer (45 mM Tris-borate, 1 mM EDTA [pH 8.0]) at 100 V for 60 min. Gels were stained with GelGreen (Invitrogen) and visualized under UV illumination.

Cyanophage N-1 direct repeats were compared with those from the CRISPR database using the follow parameters: BLASTN, e-value < 10^{-5}. A sequence logo of the aligned direct repeats from N-1 was created with Weblogo(151). The genome of N-1 and its host were screened for *cas* genes using the nr database. A Neighbor-Joining tree (Juke-Cantor Model) was created with the consensus direct repeats from the N-1 CRISPR and other cyanobacterial CRISPRs using Geneious (114) and edited with FigTree v1.3.1 (http://tree.bio.ed.ac.uk /software/figtree).

4.3.2 RNA isolation and Reverse Transcriptase –PCR (RT-PCR)

Total RNA was extracted from host cells infected with N-1. Total RNA from an uninfected control culture was also extracted. Two flasks each containing 100 ml of an exponentially growing *Nostoc sp PC7210* culture were infected with N-1 lysate and 15 mL were collected at day 5. *Nostoc* cells were pelleted by centrifugation and resuspended in BG-11 media. RNA was extracted using TRIzol® Reagent following the manufacturer's protocol (Life technologies). Briefly, 0.75 mL of TRIzol® Reagent was added to the resuspended pellets. The cells were lysed by pipetting up and down several times. The samples were centrifuged at 12,000 g for 10 minutes at 4°C (Beckham Coulter, Allegra X-22R). The supernatant was then transferred to a new microcentrifuge tube and incubated for 5 minutes at room temperature to permit complete dissociation of the nucleoprotein complex. Afterward, 0.2 mL of chloroform was added to the tube and incubated for 3 minutes at room temperature. The sample was then centrifuged at 12,000 xg for 15 minutes at 4°C (Beckham Coulter, Allegra X-22R) and the resulting aqueous phase was collected and placed into a new microcentrifuge tube. 0.5 mL of 100% isopropanol was added to the aqueous phage and incubate room temperature for 10 minutes. The sample was centrifuged at 12,000xg for 10 minutes at 4ºC (Beckham Coulter, Allegra X-22R) . The supernatant was removed from the tube and the pellet was washed with 1 mL of 75% ethanol. Afterward, the RNA pellet was resuspended in 50 uL of RNase- free water.

To confirm the transcription of the N1 CRISPR loci, the presence of precursor CRISPR RNA (pre-crRNA) was analysed using Reverse Transcriptase-PCR (RT-PCR). RT-PCR targeted the sequence between spacer 1 and spacer 4 (~ 150 bp). First, an aliquot of the extracted RNA was treated with DNase I (Invitrogen) to remove DNA. The cDNA was generated using Superscript III reverse transcriptase (Invitrogen) with random hexamers (50 ng/µl).

Amplification was carried out in 25 µl PCR reactions containing 10 ng cDNA template, 1 µM of each primer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, and 0.5 units of Platinum Taq DNA polymerase (Invitrogen). PCR cycle parameters consisted of a single denaturation step for 5 min at 95°C, followed by 35 cycles for 30 sec at 95°C, 1 min at 57°C, and 3 min at 72°C, and a final extension step for 10 min at 72°C. Resulting PCR products were sequenced.

4.3.3 Culture of surviving *Nostoc* cells

To investigate the potential lateral transfer between the cyanophage and its host, a *Nostoc* culture was infected as previously described. This time, the culture was incubated until surviving *Nostoc* cells start growing again. These *Nostoc* cells were collected and inoculated on BG-11 plates. Inoculation was repeated 4 times for each *Nostoc* colonies to remove any leftover viral particles. The absence of virus was confirmed using primers targeting N-1 major capsid protein (80). *Nostoc* genomic DNA was extracted from using a phenol chloroform extraction method (150). The DNA were then screened with primers targeting the CRISPR region (see 4.2.1 for PCR info). The *Nostoc* cells with positive amplification for the presence of CRISPR array were transferred into a liquid culture (BG-11) and further use for analysis.

4.3.4 Pulse field gel electrophoresis

Pulse field gel electrophoresis (PFGE) was used to separate the *Nostoc* genomic DNA and its plasmid. A subsample (100 µL) from a culture of *Nostoc* PCC7210 was mixed with an equal volume of molten low-melt agarose solution (1% agarose), and dispensed into a plug mold. After solidification, the plugs were immersed in a Proteinase K digestion buffer (250 mM EDTA, 1% sodium dodecyl sulfate, 1 mg of proteinase K ml^{-1}) overnight at room temperature. After incubation, the buffer was decanted, and the plug was washed by submerging it in 10 mM Tris mM EDTA, pH 8.0 for 30 min. The washing step was repeat 3 times.

One agarose plug for each culture (*Nostoc* control & *Nostoc* surviving cells) was used to target the plasmid DNA. The agarose plug was loaded onto a 1% agarose PGFE gel (0.5x TBE buffer-45 mM Tris-borate, 1 mM EDTA pH 8.0). PFGE was performed under the following conditions: 0.5× TBE tank buffer (45 mM Tris-borate, 1 mM EDTA pH 8.0), 1 to 15 s pulse ramp, 120° included angle, 6.0 V cm^{-1} 14°C, and 22 h. After electrophoresis, the gel was stained in 0.1x SYBR Green solution (Invitrogen) for 60 min, then visualized and photographed with an Alpha Imager 3400 system.

4.3.5 PCR amplification of the CRISPR array

To survey plasmid DNA from the surviving cells, the following PCR was used. Two μL from a culture was added to a 48 μL PCR mixture containing Platinum *Taq* DNA polymerase assay buffer (50 mM KCl, 20 mM Tris-HCl, pH 8.4), 10 mM $MgCl_2$, 200 μM deoxyribonucleoside triphosphate, the primers sCRF and sCRR (0.25μM each (sCRF= CAATTGGCAAAAGATTTAGCAGC and CR3R= GGGGAGAGGTTTGGAGAGGGGT.) and 2.0 U of Platinum Taq DNA polymerase (Invitrogen) Negative controls contained all of the reagents, but sterile water was used as the template. PCR was carried out as follows: denaturation at 94°C for 5 min, followed by 35 cycles of denaturation at 94°C for 30s , annealing at 57°C for 45s, extension at 72°C for 1 min, and a final extension at 72°C for 10 min (184)The amplification products were subjected to electrophoresis using 1.5% agarose–0.5x TBE buffer at 100 V for 60 min. Gels were stained with GelGreen and visualized under conditions of UV illumination.

4.4 Results

4.4.1 CRISPR array in the genome of Cyanophage N-1

During genomic analysis of Cyanophage N-1 (Chapter 3), a DNA repeat region of about 400 bp was identified (Figure 4-1A). The array comprises four spacers and five 37 bp long direct repeats that are similar in structure to the DR5 family of CRISPRs commonly found in cyanobacteria. (Table 4-1, Figure 4-1B), Spacers in the N-1 CRISPR vary in length from 29 to 37 bp and from 24 to 54% in GC-content (Table 4-2), but did not have significant matches to other sequences in the NCBI non-redundant nucleotide (nr/nt) database. . An AT-rich sequence region (~25.6% G+C content) of approximately 120 bp upstream of the CRISPR array was considered as the leader region. Neighbor-Joining analysis revealed that the direct repeats in Cyanophage N-1 clustered among those in filamentous cyanobacteria (Figure 4-2), and were most similar to three sets of consensus direct repeats from CRISPR arrays found in the genome of *Calothrix* sp. PCC7507. Overall, DRs found within a cyanobacterial genome were not necessarily most closely related to each other. For example, some CRISPRs in *Nostoc* PCC7210 and *A. variabilis* ATCC29413 have repeats that are more similar to those in N-1 than to other repeats within their own genomes (Figure 4-2).

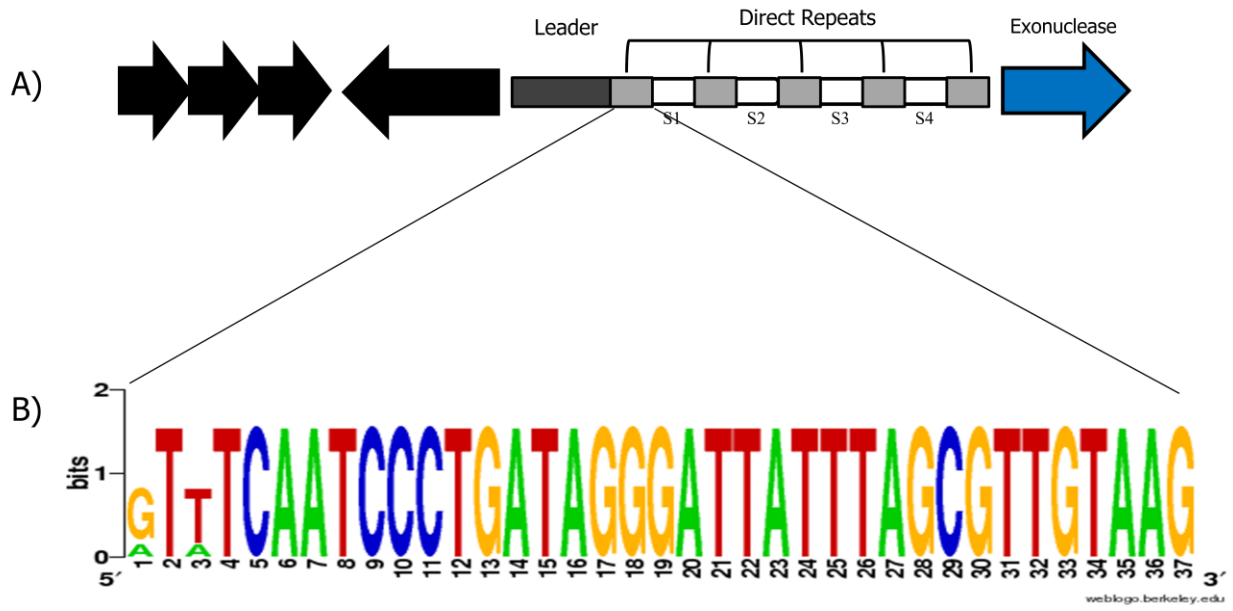


Figure 4-1.Characterization of the CRISPR in Cyanophage N-1.
A) The CRISPR in Cyanophage N-1 consists of five direct repeats (DRs) (grey boxes) and four spacers (white boxes), and a leader sequence (dark-grey box). The CRISPR array is surrounded by ORFs which putatively encode for an exonuclease (blue arrow), or which are hypothetical (black arrow). B) A consensus alignment shows little sequence variation across direct repeats in the N-1 CRISPR.

Table 4-1. BLAST results for the direct repeats from the CRISPR array present in the genome of Cyanophage N-1.

Cyanobacterial stain (CRISPR ID)	e-value	Start	End	DR concensus sequence
Microcoleus sp. PCC 7113 (CRISPR6)	$3.0e^{-08}$	2087079	2087333	GTCTGAATTCCATATAATCCCTATCAGGGATTGAAAC
Microcoleus sp. PCC 7113 (CRISPR11)	$1.0e^{-07}$	3131505	3132053	GTTTAAATTCCACTTAATCCCTATCAGGGATTGAAAC
Microcoleus sp. PCC 7113 (CRISPR15)	$1.0e^{-07}$	3859458	3860565	GTTTCAATCCCTGATAGGGATTAAGTGGAATTTAAAC
Microcoleus sp. PCC 7113 (CRISPR9)	$1.0e^{-07}$	2916270	2917173	GTTTAAATTCCACTTAATCCCTATCAGGGATTGAAAC
Nostoc sp. PCC 7107 (CRISPR27)	$1e^{-07}$	5817107	5818597	GTTGCAATTTCTATTAATCCCTATCAGGGATTGAAAC
A. variabilis ATCC 29413 (CRISPR3)	$8.0e^{-07}$	1234556	1237182	GTTTTAATTAACAAAAATCCCTATCAGGGATTGAAAC
Nostoc PC7210 (CRISPR13)	$8.0e^{-07}$	3516819	3517367	GTTTCAATCCCTGATAGGGATTTTTGTTAGTTAAAAC
Nostoc PC7210 (CRISPR14)	$8.0e^{-07}$	3517542	3518084	GTTTCAATCCCTGATAGGGATTTTTGTTAGTTAAAAC
A. cylindrica PCC 7122 (CRISPR21)	$4e^{-07}$	5001036	5003639	GTTTCAATCCCTAATAGGGATTATTTGAAATTTCAAC
Cylindrospermum stagnale (CRISPR5)	$4e^{-07}$	1101502	1103134	GTTACAATTCACCCAAATCCCTATCAGGGATTGAAAC
Calothrix sp. PCC 6303 (CRISPR6)	$4e^{-07}$	2021864	2022765	GTTCCTATAAACTAAAATCCCTATCAGGGATTGAAAC
Calothrix sp. PCC 6303 (CRISPR8)	$4e^{-07}$	2038085	2039287	GTTCCTATAAACTAAAATCCCTATCAGGGATTGAAAC
Calothrix sp. PCC 7507 (CRISPR24)	$4e^{-07}$	5067421	5070798	GTTTCAATCCCTGATAGGGATTTAAGTTAATTGGAAC
Calothrix sp. PCC 7507 (CRISPR14)	$4e^{-07}$	3375025	3376306	GTTTCAATCCCTGATAGGGATTTAAGTTAATTGGAAC
N. punctiforme PCC73102 (CRISPR16)	$3.0e^{-07}$	3338172	3341197	GTTTCAATCCCTGATAGGGATTTTGATGAATTGCAAT
Nostoc PCC7210(CRISPRS8)	$3.0e^{-07}$	1836813	1837723	GTTTCTATTAACACAAATCCCTATCAGGGATTGAAAC
Nostoc PCC7210 (CRISPRS3)	$3.0e^{-07}$	807452	807558	ATTGCAATTAACTAAAATCCCTATCAGGGATTGAAAC
A. variabilis ATCC 29413 (CRISPR19)	$3.0e^{-07}$	5764010	5766428	ATTGCAATTAACTAAAATCCCTATCAGGGATTGAAAC
Nostoc PCC7210 (CRISPRS20)	$3.0e^{-07}$	5654133	5654384	GTTAAAACCCTCTAAAATCCCTATCAGGGATTGAAAC
A. variabilis ATCC 29413 (CRISPR9)	$3.0e^{-07}$	2395670	2398703	GTTTCAATCCCTGATAGGGATTTTAGAGGGTTTTAAC
A. variabilis ATCC 29413 (CRISPR18)	$1.0e^{-06}$	5227213	5229282	GTTTCTATTAACACAAATCCCTATCAGGGATTGAAAG
Cyanothece sp. PCC 7424 (CRISPR26)	$4.0e_{-06}$	3575124	3575742	GTTACAATTAAAATGAATCCCTATTAGGGATTGAAAC
A. variabilis ATCC 29413 (CRISPR16)	$4.0e^{-06}$	4821250	4823752	GTTGCAACACCACATAATCCCTATTAGGGATTGAAAC

Table 4-2. Spacer information for the CRISPR array.

Spacer	Length	%GC content	Sequence
1	34	32.4	CAATTGGCAAAAGATTTAGCAGCTTTTTTGATC
2	29	24.1	TGTAAAGTACTCTTCACAAATTCAAAACAAAAATAC
3	33	54.5	CCAAAGTACCATCGGCACTTCTTGTCCACCGGA
4	37	35.1	TCTCATAAAAGATTTTCGTCGCAATGCAACAAAAGCT

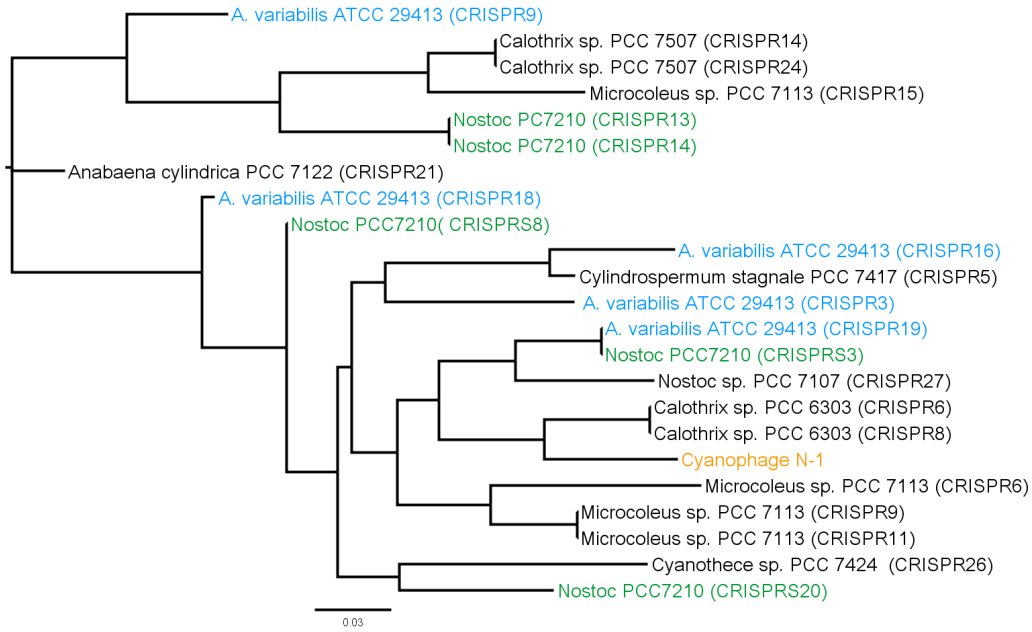


Figure 4-2Phylogeny of Direct Repeats for the Cyanophage N-1 and cyanobacteria.
An unrooted neighbour-joining tree shows that the consensus direct repeat in Cyanophage N-1 (orange) is not most closely related to those found in its known hosts, *Nostoc* PCC7210 (green) and *A. variabilis* ATCC 29413 (light blue). Consensus directs repeat from other cyanobacteria, including those which are most closely related in *Calothrix* sp. PCC 6303, are shown in black. The scale bar represents 0.003 nucleotide changes.

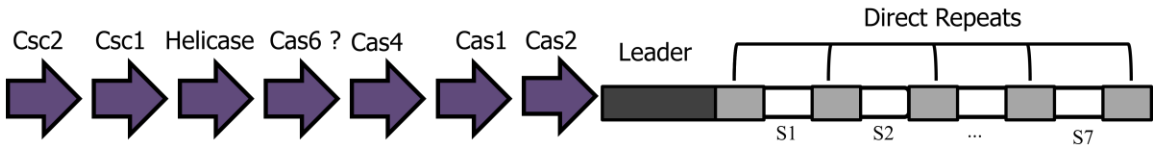


Figure 4-3.The CRISPR8 found in *Nostoc* PCC7210 which is adjacent to a *cas* operon (purples arrows).
The CRISPR8 consists of seven direct repeats (DRs) (grey boxes) and six spacers (white boxes), and a leader sequence (dark-grey box).

4.4.2 Transcription of N-1 CRISPR

No *cas* genes were identified in the genome of N-1, while *Nostoc* PCC7210 CRISPR8 was the only *Nostoc* CRISPR with direct repeat similarity to N-1 CRISPR to have a complete *cas* operon (Figure 4-3). The transcription of the N1 CRISPR array (pre-crRNA) was demonstrated by reverse transcriptase PCR (RT-PCR) on two independent *Nostoc PCC7210* cultures infected with cyanophage N-1 (Figure 4-4).

4.4.3 Identification of surviving cells containing N-1 CRISPR

In order to investigate the potential lateral transfer between the cyanophage and its host, surviving cells of infected *Nostoc* cultures were collected and screened for potential integration of the CRISPR array in their genomes. Four out of eight surviving cells showed positive amplification of CRISPR arrays (Figure 4-5). Using pulsed-field-gel electrophoresis (PFGE) to separate genomic and plasmid DNA, and PCR for screening, the N-1 CRISPR was further localized to the 182.2 kb pPCC7120beta plasmid (Figure 4-6).

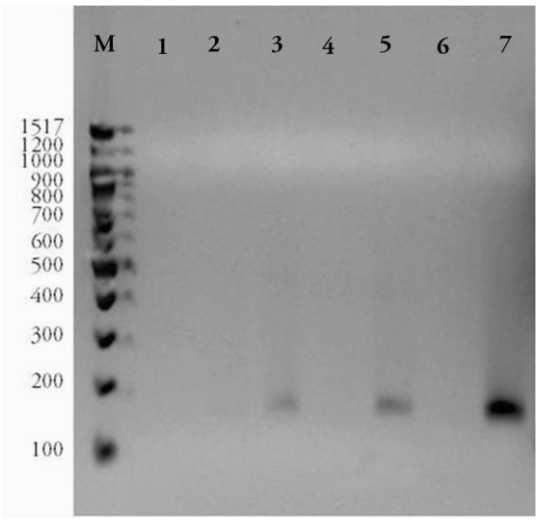


Figure 4-4 Transcription of the N-1 CRISPR into pre-crRNA.

The N1-CRISPR RT-PCR products were separated on a 1.5 % agarose gel and stained with GelGreen. The lanes on the gel are labeled as follows: M) Invitrogen 100 bp ladder, 1) PCR negative control; 2) Culture A- RNA control; 3) Culture A- cDNA 4) Culture B- RNA control 5) culture B cDNA, 6) RT negative control; 7) PCR positive control.

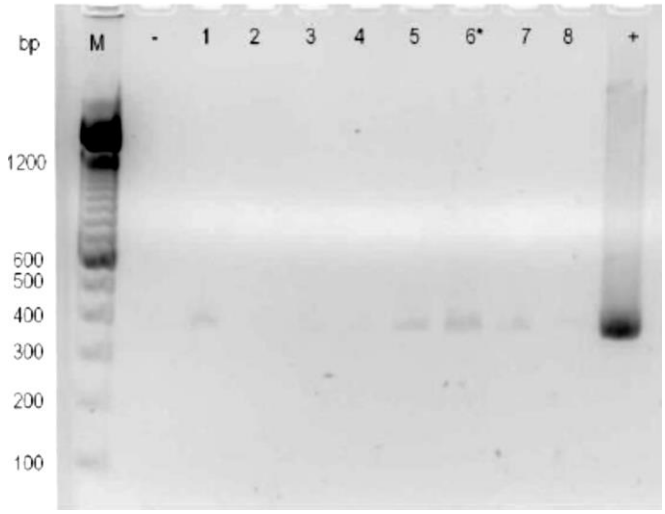


Figure 4-5. Identification of surviving *Nostoc* cells containing N-1 CRISPR array . The N1-CRISPR PCR products were separated on a 1.5 % agarose gel and stained with GelGreen. The lanes on the gel are labeled as follows: M) Invitrogen 100 bp ladder; -) PCR negative control; 1to 8) Surviving *Nostoc* cells and +) PCR positive control. * Surviving *Nostoc* cells were used for the PFGE analysis.

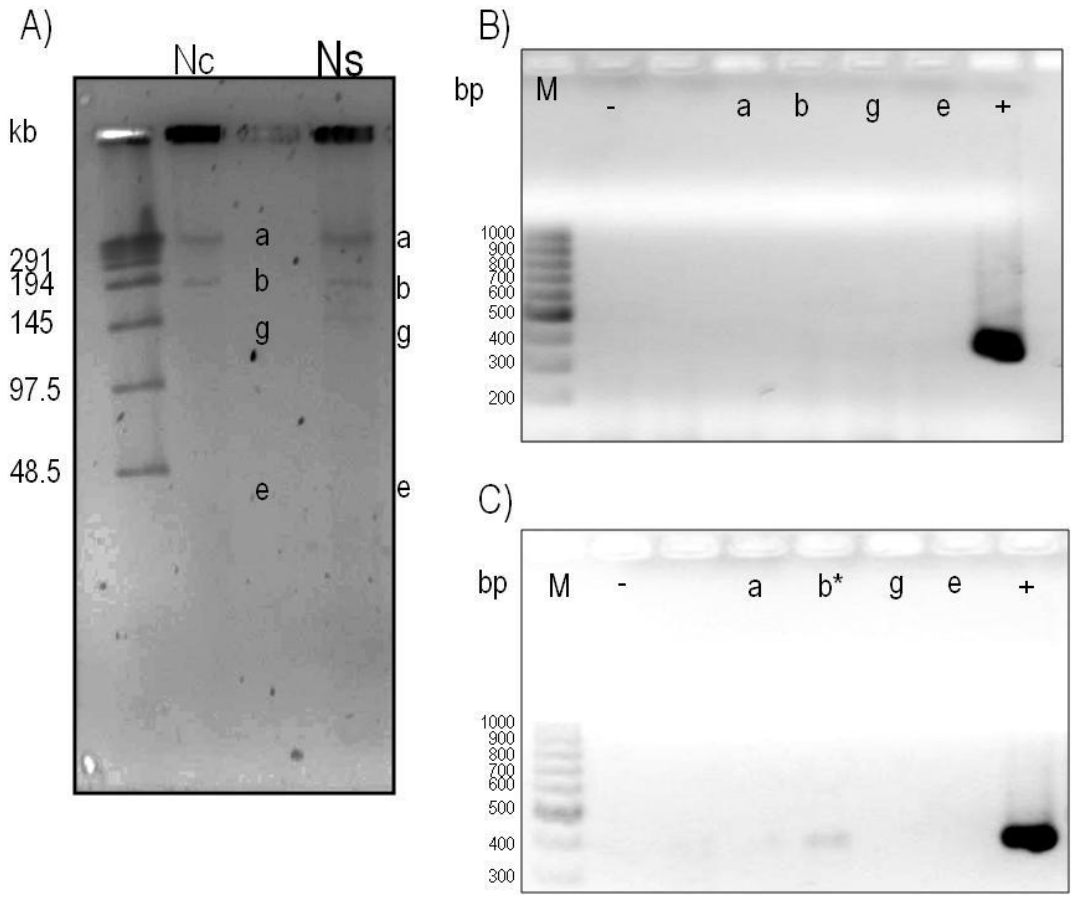


Figure 4-6. Identification of N-1 CRISPR recombination site.
A) 0.8% agarose PGFE gel on the *Nostoc* cells to extract the plasmid DNA (M: yeast chromosome ladder, Nc: *Nostoc* control, Ns: Nostoc surviving cell 6. B) CRISPR PCR on the plugs from Nc control (M: 100 bp NEB ladder, -: negative control, a: alpha plasmid, b: beta plasmid, g: gamma plasmid, e: epsilon plasmid, +: positive control. C) CRISPR PCR on the plugs from Ns (M: 100 bp NEB ladder, -: negative control, a: alpha plasmid, b: beta plasmid, g: gamma plasmid, e: epsilon plasmid, +: positive control. *asterisk designate plugs with amplification.

4.5 Discussion

4.5.1 CRISPR in Cyanophage N-1

The CRISPR array found in Cyanophage N-1 is similar to the DR5 family of CRISPRs commonly found in cyanobacteria, and is predicted to have the same characteristic hairpin structure (177) found in DR5 (Type I-D) CRISPR repeats (data not shown) (178, 185). The

CRISPR-Cas system is widespread among cyanobacteria, with 86 out of 126 sequenced genomes containing CRISPR-Cas systems, and with multiple CRISPR arrays in many genomes (178). This includes *Nostoc* PCC7210 and *A. variabilis* ATCC29413, which include 13 and 11 CRISPR arrays, and 106 and 183 spacers, respectively (186). The sequence similarity of CRISPR repeats between cyanobacteria and N-1 suggests that the N-1 CRISPR was transferred from a cyanobacterium to an ancestor of N-1 during an infection, confirming that the CRISPR-Cas system has been exchanged by lateral gene transfer among microorganisms (171, 187).

Although no *cas* genes were identified in the N-1 genome, the N-1 CRISPR array was transcribed during infection (Figure 4-4). Possibly, N-1 contains unidentified genes encoding Cas proteins or host Cas proteins may be used for initiation. CRISPR loci can function without proximate *cas* genes (177, 187), and different CRISPR loci with similar repeats in the same genome can use the same set of Cas proteins. In *Nostoc* PCC7210, CRISPR8 is adjacent to a *cas* operon and is similar to the repeats in the N-1 CRISPR (Figure 4-3). Nothing is known about expression of the CRISPR-Cas system in *Nostoc*, but in another Type I CRISPR-Cas system in *Escherichia coli*, Cas proteins are continuously transcribed (188). This provides a mechanism for expression of the N-1 CRISPR array, even in the absence of viral encoded Cas proteins.

4.5.2 Recombination of N-1 CRISPR with the host

The N-1 CRISPR array was amplified from the plasmid PCC7120beta in cells that survived infection by N-1, indicating transfer of the N-1 CRISPR to the host plasmid (Figure 4-6). The mechanism by which the N1-CRISPR was transferred to its host is enigmatic. The N-1 CRISPR was detected in cells that grew up subsequent to culture lysis by Cyanophage N-1; yet, there was no evidence that N-1 is temperate; PCR screens for the MCP were negative (data not

shown), while those for the N-1 CRISPR were positive. Resistance to infection, however, was not conferred by the CRISPR, which had no matches to the N1 genome other than the CRISPR.

Interestingly, plasmid PCC7120beta, conferring resistance to competing phages. The recent isolation of phages infecting *V. cholerae* that encode a CRISPR-Cas system demonstrated that the system defeats an inhibitory chromosomal island of the bacterial host (179). In the case of N-1 CRISPR, the origin of the spacers is unknown which make it impossible to identify the target of the array. However, I suggest that the CRISPR array in cyanophage N-1 acts as a mechanism against co-infection and offers a fitness advantage to both host and virus by preventing lysis by competing phage.

The ability of viruses to contain and likely transfer CRISPRs benefits both the host and the carrier virus by introducing new spacers that protect the host from a wider spectrum of viruses that are potential competitors of the carrier virus. Even in cases where a viral CRISPR is not incorporated into the host, its expression may be of selective advantage. In filamentous cyanobacteria, for example, molecules are believed to be exchanged between cells through non-specific junctions that connect the cytoplasm of adjacent cells (189). Viral particles likely cannot pass through these channels, while, small molecules such as crRNA probably can. This would allow cells adjacent to those infected by N-1 to acquire the CRISPR sequences, and thereby immunity to infection by other competing viruses.

The success of this defense mechanism against co-infection implies that the spacers in the N-1 CRISPR are derived from current competitors, and that the CRISPR is able to incorporate new spacers from competing viruses. Consequently, the acquisition of new spacers into the viral CRISPR array is crucial. Co-infection of N-1 with other viruses could be a way to experimentally test the acquisition of new spacers in the N-1 CRISPR array.

4.6 Concluding remarks

CRISPR-Cas systems in bacteria and archaea are powerful defenses against the invasion of specific foreign nucleic acids. The sporadic distribution of CRISPRs within and among prokaryotic genomes, as well as phylogenetic analysis of direct-repeat sequences, indicates that horizontal transfer has played a prominent role in their occurrence. Here, I demonstrate that a CRISPR encoded by a phage is not only expressed when the host is infected, but it can be transferred to the host. When transferred to the host, phage-encoded CRISPRs will increase both the fitness of the host and the virus by making the host resistant to a broader suite of viruses while remaining susceptible to infection by the CRISPR-encoding virus.

Chapter 5: Use of stable isotope probing to characterize viruses infecting primary producers in high-arctic cyanobacterial mats

5.1 Synopsis

Cyanobacteria and their viruses are important members of polar microbial communities, yet analysis of their diversity and host interactions are hampered by the lack of appropriate model cyanophage-host systems and methods to study infection in the environment. Here, I used a DNA stable-isotope-probing technique (DNA-SIP) along with ^{13}C-labeled sodium bicarbonate as an approach to characterize cyanophages from Arctic microbial mat communities. Initially, we demonstrated in a laboratory setting that cyanophage nucleic acids incorporate ^{13}C from labeled cyanobacterial cells during replication. Subsequently DNA-SIP was applied to identify viruses infecting cyanobacteria in microbial mats in Ward Hunt Lake (Canadian High Arctic). Assembly of more than 9 million metagenomic reads from the virus assemblage revealed cyanophage-like contigs, including a 9.8 kb contig that contained open reading frames (ORFs) with similarity to phage, prophage and cyanobacterial hypothetical genes. Cyanobacterial sequences were associated with the orders *Chroococcales* (unicellular), *Oscillatoriales* (filamentous) and *Nostocales* (nitrogen-fixing). In particular, contigs derived from cyanophages infecting representatives of *Leptolyngbya* spp. or *Pseudanabaena* spp. were common in the high-arctic cyanobacterial assemblage. The results suggest that cyanophage assemblages are diverse and play an active role in the carbon cycling in microbial mats from the High Arctic.

5.2 Introduction

Cyanophages are present in many ecosystems including coastal, open-ocean and polar inland waters (76). They can reach abundance in excess of 10^5 mL^{-1} in coastal regions (47) and

are believed to lyse from <1% to >5 % of *Synechococcus* cells each day (41, 49, 51). Cyanophages can have great influence on community structure and genetic diversity of cyanobacterial assemblages and also contribute to carbon cycling (190).

Cyanobacteria are among the dominant phototrophs in polar regions, where they form benthic mats at the bottom of ponds and lakes (38). Microbial mats from the High Arctic are dominated by filamentous oscillatorian cyanobacteria from genera such as *Phormidium*, *Leptolyngbya* and *Pseudanabaena* and nitrogen-fixing groups such as *Nostoc* (7). Viruses are likely an integral part of the High Arctic microbial mats microbial mat assemblages (191); yet are uncharacterized to date. Moreover, viral sequences within metagenomic data from microbial mats, often has little similarity to other sequences, is in low coverage, and provides limited information on viral and host diversity. This is especially true for freshwater cyanophages in high-arctic aquatic ecosystems.

A major challenge in isolating and sequencing new cyanophages is that their hosts are not in culture (192); therefore, non-culture based approaches need to be developed to identify and characterise freshwater cyanophages. Metagenomic approaches coupled with techniques such as stable isotope probing (SIP) can be used to identify microorganisms in environmental samples that have particular metabolic functions (193, 194). For example, this technique was used with different substrates and organisms to identify the grazers of marine picocyanobacteria (195). DNA-SIP relies on the incorporation of a particular growth substrate that is highly enriched with a stable isotope, such as ^{13}C (Figure 5-1). This method potentially provides an ideal tool for identifying viruses infecting primary producers by selectively recovering and analyzing isotopically enriched viral DNA.

In this chapter, I used DNA-SIP combined with metagenomic analysis to identify viruses infecting cyanobacteria, the dominate primary producer in high-arctic microbial mats. I showed in a laboratory setting that cyanophage incorporate ^{13}C into their nucleic acids from labeled filamentous cyanobacteria. Afterwards, I used DNA-SIP to identify viruses infecting cyanobacteria in a microbial mat from Ward Hunt Lake, in the Canadian High Arctic (83°N). Assembly of sequences showed the presence of cyanophage-like contigs, including a 9.8 kb contig containing ORFs that are similar to sequences found in phage, prophage and oscillatorian, unicellular and nitrogen-fixing cyanobacteria.

5.3 Materials and methods

5.3.1 *Nostoc* sp. strain PCC 7120 and Cyanophage A-1: A model system for DNA-SIP

5.3.1.1 Culture growth

Liquid 800 ml batch cultures of *Nostoc* sp. strain PCC 7120 (obtained from the American Type Culture Collection) were grown in 1L flasks in BG-11 medium (103) under constant illumination (33 $\mu Em^{-2}s^{-1}$) at 26ºC with continuous shaking at 75 rpm. One culture was incubated in BG-11 containing ^{13}C sodium bicarbonate (Cambridge Isotope Laboratories, Andover, Ma), while another control culture was grown in unlabeled medium. The cultures were incubated for 5 d. *Nostoc* PCC7210 cells were concentrated using centrifugation for 15min at 4,000xg (Beckham Coulter, Allegra X-22R). The supernatant was removed and the pelleted cells resuspended in 200 µL of BG-11 medium. *Nostoc* PCC7210 genomic DNA was extracted from the resuspend cell pellets using a phenol chloroform extraction method (150).

5.3.1.2 Amplification and purification of the cyanophage

Liquid cultures of *Nostoc* sp. strain PCC7120 were grown as described in 5.2.1.1. Exponentially growing cultures were infected with Cyanophage A-1(L) and left for 4 to 7 d until

the cultures were transparent, indicating lysis. To separate virus particle from cell debris, sodium chloride was added to the lysate to a final concentration of 0.5 M and incubated at 4°C for 1 h before vacuum filtration through a 1.2-µm pore-size glass-fiber filter (GC50;Advantec MFS, Dublin, CA) and twice through a 0.22-µm pore-size membrane filter (GVWP; Millipore, Bedford). The viral particles were then concentrated using ultracentrifugation for 6 h at 119,577xg in a Beckman Coulter ultracentrifuge (45Ti rotor, 8°C). The supernatant was removed and the pelleted viruses resuspended in 200 µL of BG-11 medium. The resuspended viral pellet was treated with DNase 1 and RNase A to remove free nucleic acids, and then extracted using the QiAamp MinElute Virus Spin Kit (Qiagen, Mississauga, Ontario) according to the manufacturer's instructions.

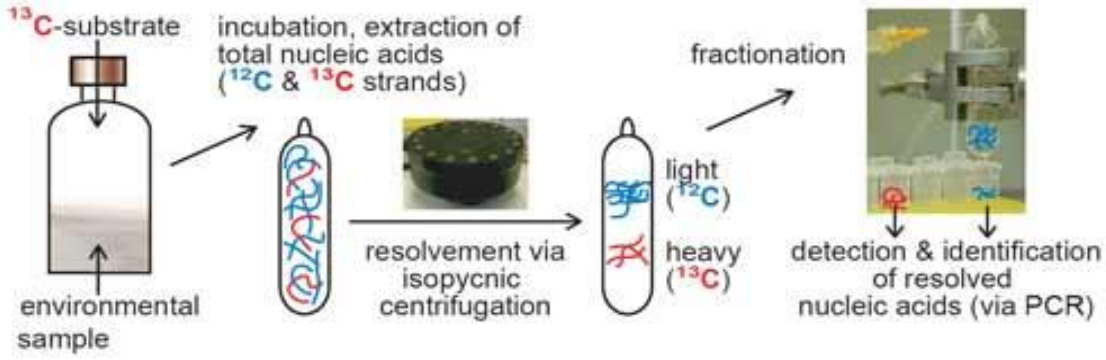


Figure 5-1. Schematic diagram of DNA-based stable isotope probing (SIP). (Adapted from http://www.helmholtz-muenchen.de)

5.3.1.3 DNA extraction, fractionation and quantification

DNA centrifugation and fractionation were performed as described by Neufeld et al (196)(2007). Briefly, ~1 µg of each DNA extract was combined with CsCl and gradient buffer into a 6.5 ml polyallomer Quick-Seal centrifuge tube (Beckman, Fullerton) for a final density of

~1.7205 as measured by refractometry (r^2 mini, Reichert, Depew, NY, USA). For *Nostoc* cells, one gradient contained unlabeled DNA and one gradient contained ^{13}C-labeled DNA. For Cyanophage A-1, one gradient contained unlabeled DNA and one gradient contained half ^{13}C-labeled DNA and half unlabeled DNA.

The gradients were centrifuged for 40 h at 148,361g in a Beckman Coulter ultracentrifuge (NVT65 rotor, 20ºC). DNA was retrieved by gradient fractionation, resulting in 20 fractions of approximately 250 µl each, where fraction 1 was the heaviest and fraction 20 was the lightest. Density was measured for each fraction using refractometry. DNA was precipitated from the CsCl by adding 20 µg of linear polyacrylamide and two volumes of PEG solution (30% PEG, 1.6 M NaCl) to each fraction. The samples were incubated at room temperature for 2 h and centrifuged at 13,000xg for 30 min (Beckham Coulter, Allegra X-22R). The pellets were washed with 70% ethanol and eluted in 50 µl of Tris-EDTA buffer.

For the DNA-SIP with the *Nostoc* culture, DNA for each fraction was quantified by running 5µL on a 1.5% agarose gel. For the DNA-SIP with Cyanophage A-1, DNA for each fraction was quantified using PicoGreen (Molecular Probes, Inc., Eugene, OR) and a NanoDrop 3300 Fluorospectrometer (Thermo Scientific, Delaware, US).

5.3.2 DNA-SIP on environmental samples

5.3.2.1 Sample description and incubation

The cyanobacterial mat samples used in the study were collected from the shoreline of Ward Hunt Lake (83°04′47″N, 074°08′17″W, Figure 5-2) on August 8, 2008. The cyanobacterial mats were approximately 10 cm deep, orange-brown in colour and flakey in appearance. For each treatment, approximately 20 g wet weight of mat was placed in a 100ml polypropylene bag containing $H^{13}C_3O$ solution (Cambridge Isotope Laboratories, Andover, MA) at a final

concentration of 25 mg l^{-1}. Cyanobacterial mats were incubated *in situ* for 3, 6 and 11 d at the water's edge of Ward Hunt Lake. A control sample of cyanobacterial mat that were placed in a whirlpak bag without ^{13}C sodium bicarbonate solution was also collected at 11 d.

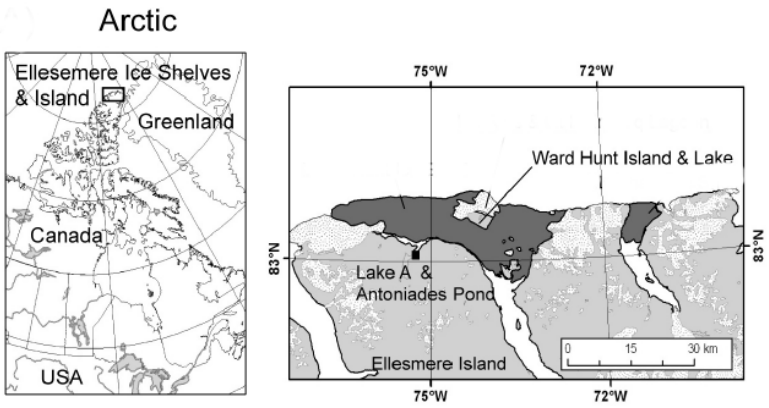


Figure 5-2- Location of Ward Hunt Lake. Adapted from (197)

5.3.2.2 Extraction of viral particles from cyanobacterial mats

Viral particles were extracted from the cyanobacterial mats using potassium citrate buffer as previously described by Williamson *et al* (2005). Briefly, the cyanobacterial mats were freeze-dried using a Flexi-Dry^{TM} MP (FTS, Stone Ridge, NY) and added to 70 ml of 1% potassium citrate buffer (10 g potassium citrate, 1.92 g $Na_2HPO_4 \cdot 12H_2O$, and 0.24 g KH_2PO_4 l^{-1}, pH 7). The samples were vortexed and sonicated (Branson 3200) on ice for 3 min, with 30s of manual shaking and sonication after each min, and then centrifuged at 10,000 x g for 10 min in a Sorvall RC-5C centrifuge (GSA rotor, 4°C). The supernatant was carefully transferred to another bottle. The mat pellets were resuspended in fresh buffer and the extraction procedure was repeated twice. The supernatant was then pooled and filtered through a 1.2-µm pore-size glass-fiber filter (GC50; Advantec MFS, Dublin, CA) and through a 0.22-µm pore-size membrane filter (GVWP; Millipore, Bedford). The viral particles from the filtered lysate were then concentrated by

ultracentrifugation for 6h at 119,577xg in a Beckman Coulter ultracentrifuge (45Ti rotor, 8˚C). The supernatant was removed and the pelleted viruses resuspended in 500 µL of potassium citrate buffer.

5.3.2.3 DNA extraction, fractionation and quantification

DNA was extracted by first treating the resuspended pellet with DNase I and RNase A to remove free nucleic acids, and then using the QiAamp Ultrasens Virus Kit (Qiagen, Mississauga, ON) according to the manufacturer's instructions. DNA centrifugation and fractionation were performed as described in section 5.2.1.3. However, only ~500 ng of each DNA extract were combined with the CsCl gradient. The DNA quantification for each fraction was performed using a PicoGreen dsDNA quantification protocol. PicoGreen dye was diluted 1:200 with TE, pH 8. Each reaction contained 15 µl of dye solution, 1µL of fraction sample and 15 µl of TE. Standard curves were constructed by serial dilution of lamda DNA (5µg), based on quantitation by the commercial provider. Ninety-six well plates were read on the Bio-Rad iQ5 (Biorad Laboratories, Hercules, CA).

5.3.2.4 Sequencing and assembly

The nucleic acids from the heavy fractions of the gradient for samples from 6 d and 11 d were pooled, purified and further used for sequencing. The sequencing library was constructed using a Nextera XT DNA Sample Preparation Kit (Illumina) and sequenced on an Illumina MiSeq at the Génome Québec Innovation Centre at McGill University (Montréal, QC). The adapters were trimmed using Trimmomatic-0.30 (http://www.usadellab.org/cms/index.php?page=trimmomatic) and quality checked with Sickle (https://github.com/najoshi/sickle). Transposon contamination from the Nextera XT kit was

checked and removed using a custom perl script, and the reads assembled using Ray with the default parameters, and 33 as the k value (105).

5.3.2.5 Sequence analysis

Contigs > 250 bp were searched against the NCBI non-redundant sequence (nr) database as of November 23, 2013. BLASTx (e-value < 10^{-3}) was performed on the contigs and classified based on the top hit. ORFs in contigs larger than 2000 bp were predicted using GeneMark (106). The predicted ORFs were translated and assigned putative functions by using BLASTp to compare them with protein sequences in the GenBank (nr), Acclame and Procite databases. Sequences with e-values <10^{-3} were considered to be homologues. Contigs and their annotations were illustrated using DNA master (J.G. Lawrence) (http://cobamide2.bio.pitt.edu). The GC content of contigs was calculated using the R package sequin (199) and plotted against contig length using the R package ggplot2 (200).

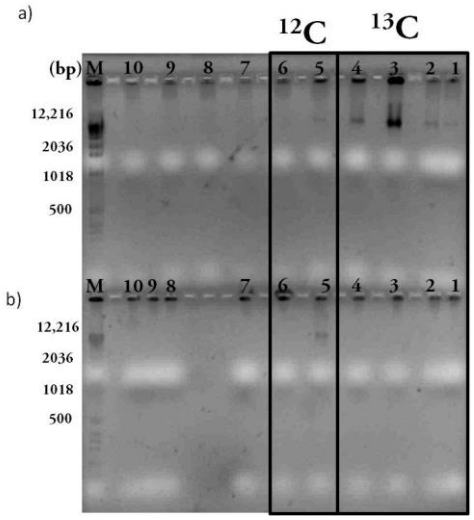
5.4 Results and discussion

The work presented here has shown that DNA sequences from uncultured viruses infecting primary producers can be recovered and sequenced using DNA-SIP. Assembly of sequences revealed the presence of cyanophage-like contigs, including a 9.8 kb contig containing open reading frames (ORFs) that are similar to phage or prophage genes and hypothetical proteins in filamentous cyanobacteria. These data and their interpretation are discussed below.

5.4.1 *Nostoc* sp. PCC7210 and Cyanophage A-1: Development of viral DNA-SIP

In order to assess whether ^{13}C-labeled DNA-SIP can be used to isolate DNA from viruses infecting photoautotrophs, Cyanophage A-1 and *Nostoc* sp. PCC7210 was used as a model system, because of the importance of filamentous cyanobacteria in high-arctic microbial mats (7). ^{13}C was incorporated into both *Nostoc* and cyanophage DNA, and could be separated from

unlabeled DNA using density-gradient centrifugation; the higher the proportion of ^{13}C in the DNA, the greater its density (193). The ^{13}C-labeled culture of *Nostoc* PCC7210 contained DNA with a range of buoyant densities from 1.7165 to 1.6926 g ml^{-1} (Fractions 1 to 5, Figure 5-3 a), while DNA from the unlabeled culture was only recovered from the 1.6937 and 1.6926 g ml^{-1} fractions (Fraction 5 & 6, Figure5.3b). In general, unlabeled DNA from bacterial genomic material ranges from 1.69 to 1.72 g ml^{-1} depending on G+C-content (150). Based on its % GC content, *Nostoc* PCC7210 genomic DNA should have a buoyant density of 1.69 to 1.70 g ml^{-1} (201). The presence of heavier DNA in the ^{13}C-labeled *Nostoc* culture demonstrates that ^{13}C was incorporated, although the range of densities is higher because only a portion of the ^{12}C is labeled during the incubations.



c)

	^{13}C		^{12}C	
Fraction	***R.I**	**Density**	***R.I**	**Density**
1	**1.4009**	**1.7165**	1.3994	1.7002
2	**1.3994**	**1.7002**	1.3993	1.6991
3	**1.3991**	**1.6969**	1.3992	1.6980
4	**1.3989**	**1.6947**	1.399	1.6958
5	**1.3987**	**1.6926**	**1.3988**	**1.6937**
6	1.3986	1.6915	**1.3987**	**1.6926**
7	1.3983	1.6882	1.3984	1.6893
8	1.3982	1.6871	1.3982	1.6871
9	1.398	1.6850	1.398	1.6850
10	1.3978	1.6828	1.3978	1.6828

***R.I= refraction Index**

Figure 5-3. DNA collected from different fractions of the density gradients
a) ^{13}C-labeled and b) unlabeled DNA from *Nostoc* sp. PCC7210 was run on a 1% agarose gel. Each lane represents a fraction collected from the density gradient, where Fraction 1 is the heaviest and Fraction 10 is the lightest. Lanes 1 to 4 are from the heaviest fractions and contain the labeled ^{13}C DNA, while Lanes 5 to 6 are from lighter fractions and contain unlabeled DNA. M= 1 kb ladder (Invitrogen). c) Refractive index and corresponding density are shown for labeled and unlabeled DNA from *Nostoc* sp. PCC7210, with the fractions containing measurable DNA shown in bold.

^{13}C-labeled DNA was also produced when Cyanophage A-1 was amplified on a ^{13}C-labeled *Nostoc* culture (Figure 5-4). Density-gradient fractionation showed that when Cyanophage A-1 was amplified on an unlabeled *Nostoc* culture it has a buoyant density of ~1.72 g ml^{-1} (Figure 5-4). The gradient containing half unlabeled and half ^{13}C-labeled DNA from-Cyanophage A-1 had a buoyant density range of between 1.727 and 1.741 g ml^{-1} (Figure 5-4). The heavier DNA in the cyanophages amplified on the ^{13}C-labeled cultures versus the control is consistent with labeled DNA being incorporated into the DNA of the replicating viruses. Based on the protocol, the unlabeled DNA should have densities of ~1.705 to 1.720 g ml^{-1} while the labeled DNA should be in the range of ~1.720 to 1.735 g ml^{-1} (196). However, labeled nucleic acids were detected in fractions up to 1.743 g ml^{-1}. As shown in a study investigating the heterotrophic growth of cyanobacteria in soil, ^{13}C-labeled viral DNA was detected in fractions as heavy as 1.754 g ml^{-1} (Taylor et al., 2013). This indicates that the cyanophage DNA was largely being synthesized from newly incorporated C, rather than recycling of cellular C.

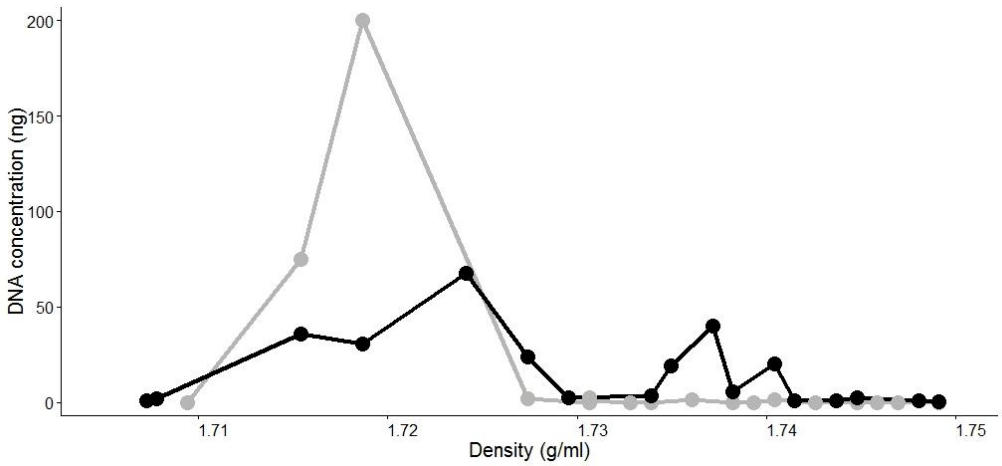


Figure 5-4.Total DNA per fraction (ng) versus density (g ml^{-1}) for gradients containing DNA from Cyanophage A-1(L) from either unlabeled (grey), or half unlabeled and half ^{13}C-labeled (black) samples .

5.4.2 Separation of active viruses infecting primary producers in cyanobacterial mats.

To identify viruses replicating on High-Arctic primary producers, cyanobacterial mats from Ward Hunt Lake were incubated with ^{13}C sodium bicarbonate. DNA quantification of density-gradient fractions demonstrated that ^{13}C-incubated samples contained nucleic acids in heavier fractions than those incubated without ^{13}C (Figure 5-5). Density gradients for samples collected on Days 6 and 11 showed small amounts of DNA in the heavier fractions (1.7425, 1.7371 and 1.7415 g ml^{-1}) that was pooled and sequenced.

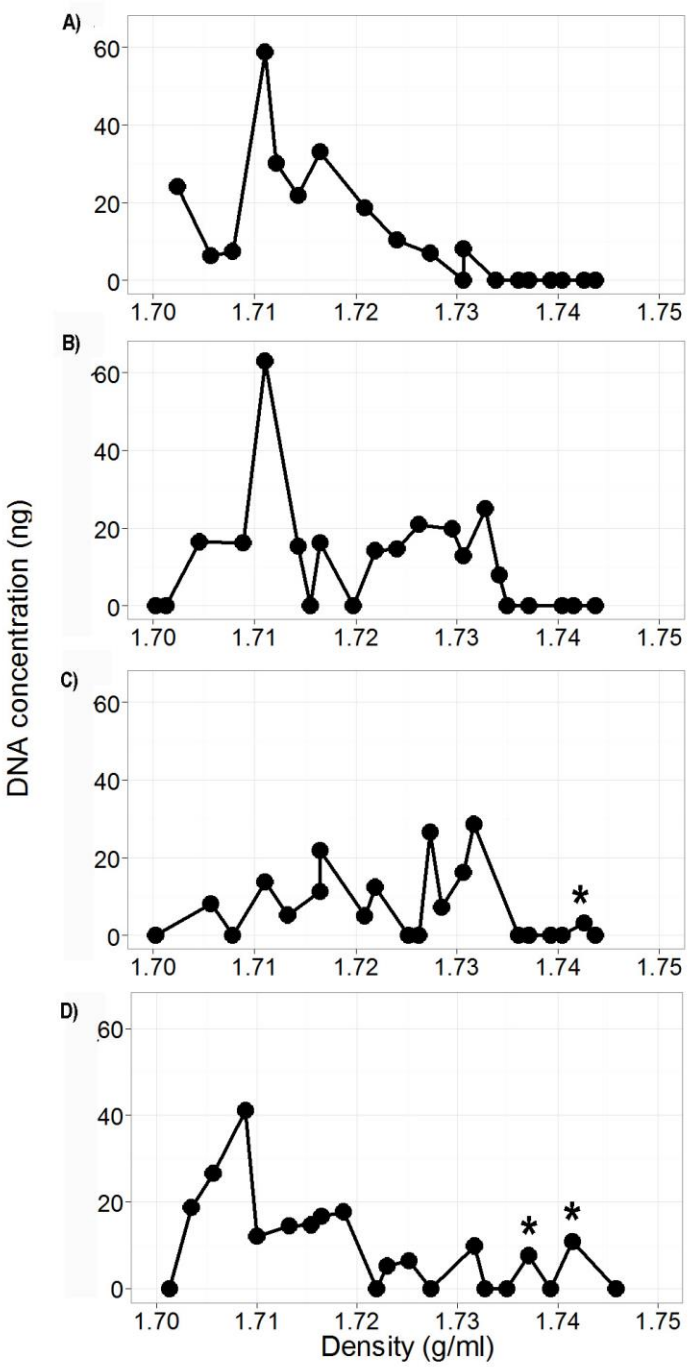


Figure 5-5.Total DNA per fraction (ng) versus density (g ml^{-1}) for gradients containing DNA from the viral fraction of ^{13}C-incubated cyanobacterial mats (black lines) incubated for a) control, b) 3 d, c) 6 d and d)11 days. The asterix (*) represent the fractions that was pooled and sequenced.

5.4.3 Sequencing analysis of the ^{13}C-labeled nucleic acids.

Assembly of the sequences (9,273,402 reads) resulted in 1,209 contigs with a length higher than 250 bp. A BLAStx analysis using the nr database revealed that 51.6 % of the contigs had no significant similarity (e-value $<10^{-3}$) to other deposited sequences. For the contigs with similarity, 30.7% were similar to sequences from Bacteroidetes (Genus *Flavobacterium*), 6.9 % were similar to sequences from cyanobacteria, 4.4% were similar to sequences from other bacteria, 2.3% were similar to sequences from cyanophages, and 4.4% were similar to sequences from other phages. Contigs with homology to *Flavobacteria* were considered contaminants and removed from further analyses. The contamination of the ^{13}C-labeled fraction with DNA from *Flavobacterium* spp. may be from dissolved-organic ^{13}C leaking from the primary producers, either directly or via viral lysis. Although filtration should have removed the bacteria, some bacteria from the CFB group can pass through 0.2 μm pore-size filters (Suttle, pers comm). Other contigs attributed to cyanobacteria and other bacteria could also be contaminants, unidentified prophages in bacterial genomes (203) or host-derived phage genes (72, 133, 204). Contigs with similarity to cyanobacterial genes were closest to sequences derived from hypothetical proteins in filamentous oscillatorian cyanobacteria (e.g. *Nostoc*, *Pseudanabaena*, *Leptolyngbya*, and *Oscillatoria*), nitrogen-fixing filamenouts *Nostoc* and unicellular cyanobacterial taxa (*Synechococcus* and *Microcystis*), suggesting that some could be unidentified prophage. The remaining contigs had similarity with sequences derived from genes involved in recombination and repair, replication and transcription. This agrees with 16S rRNA gene and morphological surveys of cyanobacteria in Ward Hunt Lake microbial mats where these oscillatorian and nitrogen-fixing genera were determined to be a dominant part of the cyanobacterial assemblage (7, 205). 16S rRNA gene sequences with highest match to

Synechococcus where also identified, however at lower abundance (7). For the virus-like contigs (including prophages), 48% were similar to sequences derived from siphoviruses, 22% to sequences from podoviruses, 8% to sequences from myoviruses and 11% to sequences from prophages. Matches to virus-derived genes included structural proteins, replication proteins, prophage-like proteins (*i.e* intergrases, transposases, phage anti-repressors), but nearly half (49%) matched hypothetical genes. The presence of sequences similar to those associated with lysogeny, and the high proportion of siphovirus-like sequences is consistent with many of the phages being temperate. Temperate phages can switch between a lytic and lysogenic lifestyles (206), with lysogeny being favored when host abundance and productivity is low (207–210).

5.4.4 Large contigs: Identification of cyanophage-like contigs

Identification of putative genes in contigs larger than 2 kb revealed cyanophage-like contigs. At 9.8 kb, Contig 6322 (43.6 % G+C content) is the largest and contains 16 ORFs (Figure 5-6, Table 5-1). These include three with homology to phage or prophage sequences, including DNA primase, an unknown structural protein and a tail fiber protein (Table 5-1). Six shared homology with hypothetical proteins in filamentous cyanobacteria, one with a hypothetical protein in proteobacteria and six did not have recognizable similarity to other sequences. Prophage sequences in cyanobacterial genomes often go unrecognized and are annotated as hypothetical proteins, and may be more common than previously recognized (58, 70) (Chapter 2 and Chapter 3). As most ORFs found in contig 6322 were similar to hypothetical proteins in cyanobacteria from the *Pseudanabaenaceae*, the contig might derive from temperate phages infecting members of the genera *Leptolyngbya* or *Pseudanabaena*, which are important members of high-arctic cyanobacterial mats (7).

Contigs 0 (43.7% G+C content, 2.2 kb), 6452 (46.7% G+C content, 2.3 kb) and 3569 (48.8% G+C content, 2.4 kb) also shared similarity to putative genes in cyanophages and cyanobacteria (Figure 5-6), including putative genes in the cyanopodovirus S-CBP2 and two ORFs from a prophage element in *Leptolyngbya* PCC7376 (134). Contig 3002 (60.9% G+C content, 2.1 kb) and Contig 7007 (65% G+C content, 2.7 kb) also appear to be of phage origin, but did not have similarity with sequences from cyanophages or cyanobacteria (Figure 5-6).

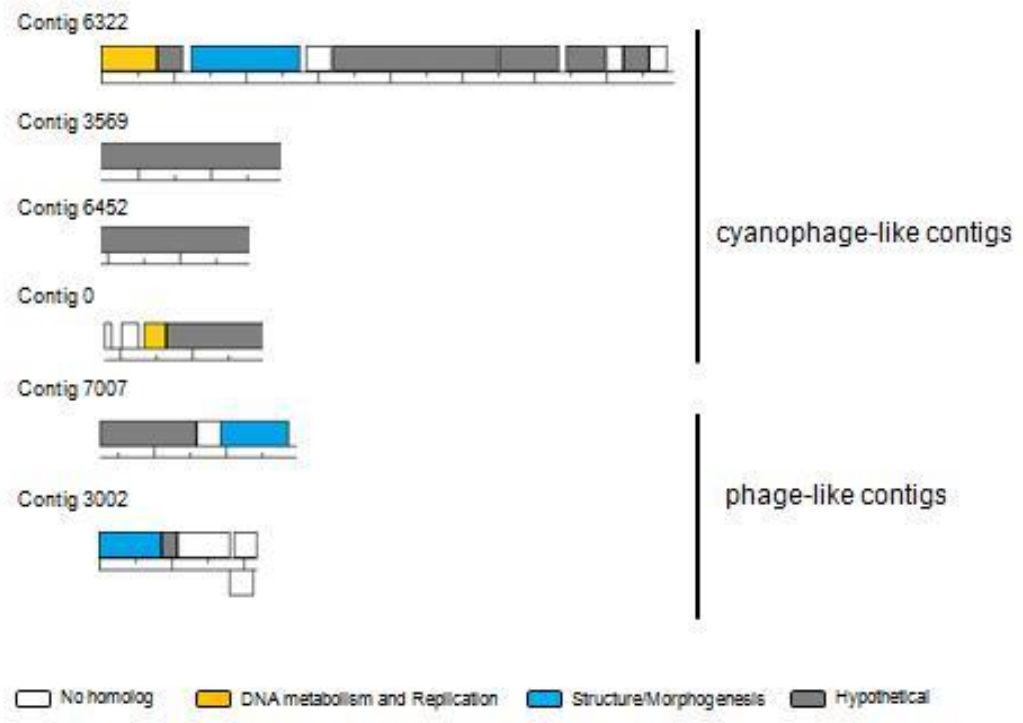


Figure 5-6.Annotation of the functional grouping for ORFs in assembled contigs larger than 2 kb.

Table 5-1. Annotation of the predicted ORFs for the contig 6322

ORF	Length (bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
1	780	+	Phage DNA primase	*Pseudanabaena sp.* PCC6802	$2e^{-32}$	37% (87)
2	348	+	hypothetical protein	*Pseudanabaena biceps*	$1e^{-17}$	69% (82)
3	1521	+	Structural protein (gp16)	Mycobacterium phage RidgeCB	$4e^{-4}$	53% (17)
4	357	+	-	-	-	-
5	2298	+	Structural protein (tail fiber)	*Geitlerinema* sp. PCC7407	$4e^{-39}$	38% (92)
6	819	+	hypothetical protein	*Leptolyngbya boryana*	$7e^{-72}$	43% (120)
7	555	+	hypothetical protein	*Geitlerinema* sp. PCC7407	$2e^{-34}$	42% (72)
8	252	+	-	-	-	-
9	378	+	hypothetical protein	*Leptolyngbya boryana*	$3e^{-5}$	28% (35)
10	252	+	-	-	-	-
11	198	+	-	-	-	-
12	273	+	hypothetical protein	*Pseudanabaena biceps*	$1e^{-7}$	37% (34)
13	372	+	hypothetical protein	*Pseudanabaena biceps*	$2e^{-40}$	80% (66)
14	192	+	hypothetical protein	*Sphingobium xenophagum*	$3e^{-3}$	43% (18)
15	234	+	-	-	-	-
16	252	+	-	-	-	-

5.4.5 G+C content reveals three "viral-like groups"

Although the G+C content ranges from 32 to 65%, the ranges fall into three distinct groups that may represent three groups of viruses (Figure 5-7). Contigs with similarity to cyanobacteria and cyanophages were mostly in the groups with low (35 to 38%) and mid (41 to 50%) ranges in G+C, while contigs with similarity to phages were primarily in the high G+C range (51 to 65 %). While there is no evidence that these high GC range contigs might derived from cyanophages, their putative functions were mainly associated in phage structural proteins which are usually highly conserved and found within different groups. This is the case for the marine cyanomyoviruses which share core genes involved in virion structure (e.g. major capsid proteins, tail fiber) with other T4-like phages (127). Therefore, these sequences might from an unrecognized group of cyanophages.

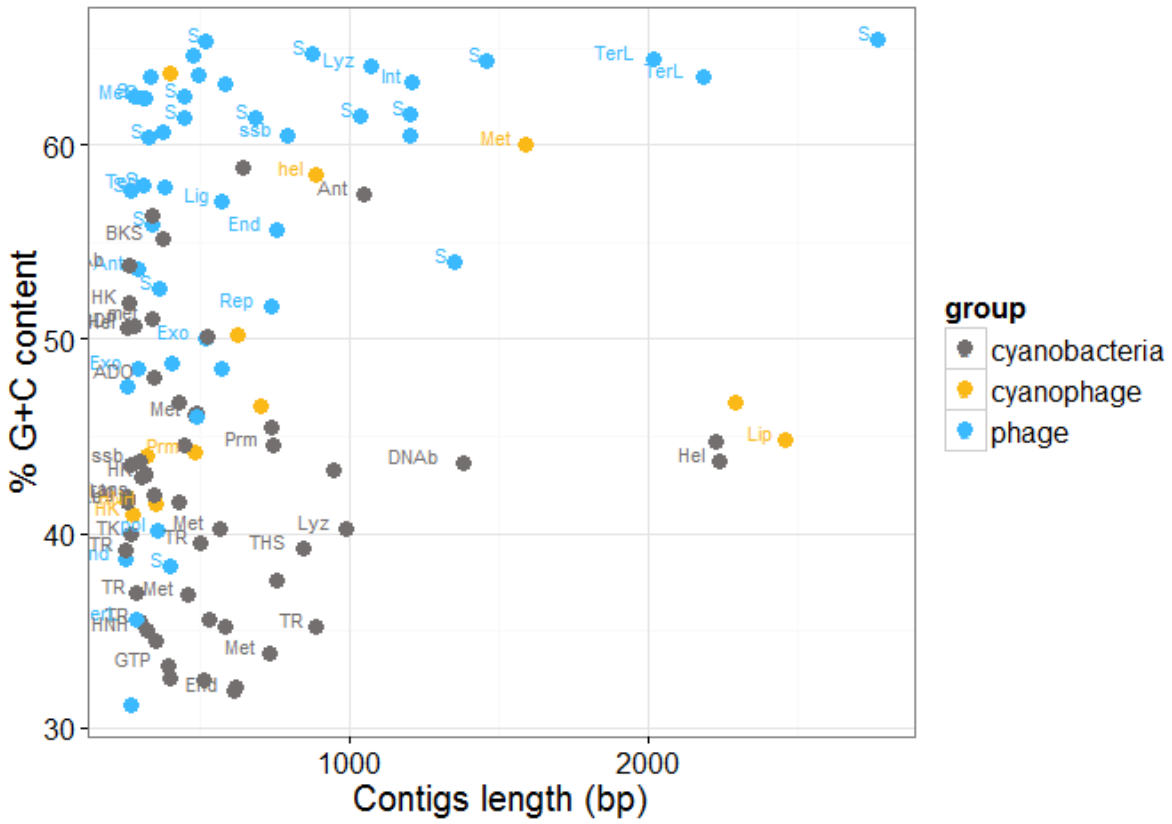


Figure 5-7. G+C content (%) versus length (bp) for contigs with homology with cyanobacteria, cyanophages, phages and other bacteria.
Contigs with known functions are labeled as follow: ADO, S-Adenosylmethionine decarboxylase; ant, anti-depressor gene; BKS, beta-ketoacyl synthase; DNAb, DNA binding protein; end, endonuclease; exo, exonuclease; GTP, Guanosine triphosphate synthesis ; Hel, helicase; HK, histodine kinase; HNH, homing endonuclease; int; integrase; Lig, ligase; Lip, lipoprotein; terL, Lyz, lysozyme; terminase large sub-unit; trans, transposase; met, methylase; Lig, ligase; Prm, primase; rep, replication protein; S, structural protein; ssb, single strand binding protein; THS, trehalose synthase/maltokinase; TK, tymidine kinase; TR, transcription regulator. Contigs with no labels are hypothetical proteins. Contig 6322 (9.8 kb) was not included in this analysis.

5.4.6 Methodological considerations

There are technical challenges to using DNA-SIP to identify replicating environmental viruses. DNA-SIP requires substrate concentrations and incubation times that reflect *in situ* conditions. First, a substrate added in high concentration may change the kinetics and cause biases in the community being examined. This is particularly true if the added substrate may

stimulate growth. In the case of bicarbonate additions to polar cyanobacterial mats this is not of concern because the cold temperature and slow growth rates ensures that CO_2 is not limiting to growth. Second, a long incubation period increases the risk of cross-feeding of the ^{13}C from the primary producers to the rest of the microbial community (193) through leakage of dissolved organic C by exudation, viral lysis or sloppy feeding by grazers. This was seen with the contamination of the ^{13}C-labeled fraction with DNA from *Flavobacterium* spp. Third, bottle effects may affect the relative abundances of subsets of the microbial community (211, 212). Fourth, SIP requires sufficient label to be incorporated so that the nucleic acids can be separated and detected. As previously mentioned, at least 20% of DNA must be ^{13}C-labeled in order to separate unlabeled and ^{13}C-labeled DNA (193). This implies that DNA from some active cyanophages might not been detected during our experiment. Fifth, contamination of unlabeled DNA within the ^{13}C-labeled DNA fraction can be generated during the ultracentrifugation and fractionation. Lastly, recovery of sufficiently-intact, high-quality ^{13}C-labeled DNA from a density gradient can be difficult. This is especially true for the small genomes of viruses that require more viral particles to be purified, in order to load at least 1 µg of DNA onto the gradient. In these experiments, ~0.5 µg of DNA was used and only ~1 % of the DNA was retrieved from the labeled fractions, which made sequencing challenging.

5.4.7 Future perspectives

DNA-SIP has great potential for identifying the viruses which are infecting specific groups of host organisms without culturing. Although metagenomic analysis of viral assemblages provides insights into the composition and diversity, it cannot be used to identify the hosts that specific viruses infect. Since DNA-SIP can be used with complex carbon and nitrogen substrates, it can be used to label specific subsets of microbes, and subsequently the

viruses infecting them. Hence, DNA-SIP is potentially a powerful approach to identify the viruses infecting specific functional groups of microbes.

5.5 Concluding remarks

In this chapter, I used ^{13}C-labeled sodium bicarbonate and DNA-SIP to assess whether this method can be used to identify viruses infecting cyanobacteria. I used *Nostoc* sp. PCC7210 and Cyanophage A-1 to show that cyanophages produced from ^{13}C-labeled host cells incorporate ^{13}C into their DNA that can be separated and purified by density-gradient fractionation. I applied this approach to high-arctic cyanobacterial mat samples, and purified the ^{13}C-labeled viral fraction, which contained phage and cyanophage-like contigs. Likely, these contigs originate from cyanophages that infect representatives of the most abundant taxa in these mats, namely the genera *Nostoc*, *Pseudanabaena*, *Leptolyngbya*, and *Oscillatoria* (7). Hence, DNA-SIP has the potential to resolve previously unknown viruses infecting specific microbial functional groups.

Chapter 6: Concluding chapter

6.1 Recapitulation of the work

There is wide recognition that cyanobacteria are major primary producers in polar freshwater regions. Filamentous cyanobacteria are commonly found in benthic mats and biofilms at the bottom of lakes, ponds and streams (38), while picocyanobacteria dominate the planktonic communities of many polar lakes. However, no representative viruses infecting this group of organisms have been characterized. This dissertation, which is a culmination of experiments and genomic and metagenomic analyses, presents the first characterization of viruses infecting freshwater polar cyanobacteria and the discovery of previously unknown groups of viruses.

Chapter 2 details the isolation and genomic characterisation of cyanophage S-EIV1 that infects the polar *Synechococcus* sp. strain PCCC-A2c. This cyanophage represents a new evolutionary lineage of phages that are globally widespread and abundant. Among its 130 ORFs, there is no recognizable similarity to genes that encode known structural proteins other than the large terminase subunit and a distant viral morphogenesis protein, indicating that the genes encoding structural proteins of S-EIV1 are distinct from other known bacteriophages. Only 20 predicted coding sequences are similar to genes encoding proteins with known functions, and most do not bear resemblance to genes found in other cyanophages. Metagenomic data indicate that related viruses are abundant in a wide range of aquatic systems suggesting that they play an important ecological role.

In Chapter 3, I found that Cyanophage A-1(L) and Cyanophage N-1 which infect filamentous cyanobacteria from the genera of *Nostoc* and *Anabaena* also demonstrate little similarity with other sequenced cyanophages. Although they are morphologically similar to myoviruses, their genome sizes are half those of other cyanomyoviruses. Many of the coding

sequences for proteins with known functions are highly similar to those found in filamentous cyanobacteria. Both phages contain a distinct DNA polymerase B that is closely related to those found in plasmids of cyanobacteria. Together these polymerase sequences form a distinct group that is more closely related to proteobacterial DNA polymerases than those found in other viruses, suggesting it was acquired from a proteobacterium by a virus and then transferred to the cyanobacterial plasmid. As well, many ORFs showed similarity to a prophage-like element identified in the genome of *Nostoc* PCC7524. The sequencing of the *Nostoc* phages reveals that numerous gene transfers occurred between these viruses and their hosts that helped to forge the evolutionary trajectory of this previously unrecognized group of phage.

Another example of gene transfer between virus and host was described in Chapter 4. The genomic analysis of the cyanophage N-1 reveals the presence of a CRISPR array with direct repeats highly similar to those associated with CRISPRs commonly found in a filamentous cyanobacterial genome (DR-5). Based on the phylogeny of the direct repeats and a survey of the surviving *Nostoc* cells, I show evidence of viral-mediated transfer of CRISPR array between the Cyanophage N-1 and its host. In addition, the evidence of transcription of N-1 CRISPR was also shown. These findings suggest that the cyanophage N-1 might use the CRISPR as a defense mechanism against co-infection.

Finally, in Chapter 5, I used ^{13}C-labeled sodium bicarbonate and DNA-SIP to assess whether this method can be used to identify viruses infecting cyanobacteria. I used *Nostoc* sp. PCC7210 and Cyanophage A-1 to show that cyanophages produced from ^{13}C-labeled host cells incorporate ^{13}C into their DNA that can be separated and purified by density-gradient fractionation. I applied this approach to high-arctic cyanobacterial mat samples, and purified the ^{13}C-labeled viral fraction, which contained phage and cyanophage-like contigs. Likely, these

contigs originate from cyanophages that infect representatives of the most abundant taxa in these mats, namely the genera. *Nostoc*, *Pseudanabaena*, *Leptolyngbya*, and *Oscillatoria* (7). Hence, DNA-SIP has the potential to resolve previously unknown viruses infecting specific microbial groups.

The little similarity for the cyanophages studied in this dissertation to previously characterized viruses clearly indicates the lack of knowledge of polar viruses and the importance of further describing these viral communities. In addition, it specifies the depth of genetic diversity present in polar cyanophage populations of which we are only beginning to scratch the surface. CRISPR-Cas immune systems found in cyanophage N-1 suggest that bacterial immune systems have the potential to be transferred by viruses. The DNA-SIP technique developed in this dissertation will help us to further understand and characterise viral communities and the phototrophic hosts they infect, providing us with exciting new avenues of exploration.

6.2 Limitations

Major challenges about the work presented in this dissertation were focused on the methodological limitations of virus isolation. Virus isolation favours viruses with broad host ranges, high affinities to the targeted host and/or high burst size. However, viruses with these specifications are not necessarily a representation of the *in situ* viral assemblage.

While I successfully isolated a virus infecting a polar picocyanobacterium (Chapter 2), I was not as fortunate in the isolation of viruses infecting polar filamentous cyanobacteria. Many polar filamentous cyanobacteria (Appendix A) were screened for this purpose but with no success. Consequently, the cyanophages A-1 and N-1 were used as representatives of viruses infecting filamentous cyanobacteria. As representative cyanophages that infect polar cyanobacteria, their genetic characterization can be used in future studies to examine viral

diversity in polar regions (Chapter 3). Although they are not of polar origin, members of the *Nostoc* genera are important taxa in High Arctic cyanobacterial mats.

To tackle the methodological constraints of a culture-dependent approach, I used DNA-SIP along with sequencing to characterize active viruses infecting cyanobacteria (Chapter 5). The DNA-SIP technique has some limitations and they were discussed in detail in section 5.3.6. Other methodological limitations presented in Chapter 5 included procedures to generate the sequencing of the uncultured viruses infecting the primary producers. To separate and concentrate the virus particles, filtration along with ultracentrifugation were used; however, not all bacteria cells were removed and there was cellular contamination. Adding a cesium chloride gradient step to the protocol could have helped with the removal of the remaining cellular contamination but it could also have led to viral loss (213). Filtration through a 0.2 μm pore-size membrance can cause viral loss. Although I did not use multiple displacement amplification (MDA), which causes sequencing bias, I used the Nextera protocol which can introduce bias for low %GC content (214).

6.3 Significance of the work

This dissertation presents new and significant information for a site and groups of viruses that were previously uncharacterized. This work represents the first study characterizing polar cyanophages, and reveals genomic information that be used to examine viral diversity in polar as well as other regions. The isolation of cyanophage S-EIV1 using the polar isolate *Synechococcus* sp. strain PCCC-A2c, also provide the first cyanophage-host system from a polar environment.

From a global perspective, the new lineages of cyanophages sequenced in this dissertation increase the database of known viruses, and helps close the sequence space in natural

viral assemblages (Figure 6-1). The sequencing of cyanophages S-EIV1, A-1, and N-1 increases the sequences in metagenomic databases that can be classified. As well, the discovery of a functional CRISPR array in the genome of a cyanophage represents a significant contribution to the field of viral ecology, as it is the first report of a CRISPR in a cyanophage. These findings indicate that not only can viruses serve as vectors for moving CRISPRs among cells, but suggest that viruses carry CRISPRs to confer host-resistance to infection by competing phages; thereby, conferring a selective advantage to both the host and CRISPR-encoding phage.

Finally, the development and use of DNA-SIP to identify DNA sequences associated with viruses infecting cyanobacteria presents a new and exciting technique. Indeed, there are only few methods available to study uncultured active viruses.

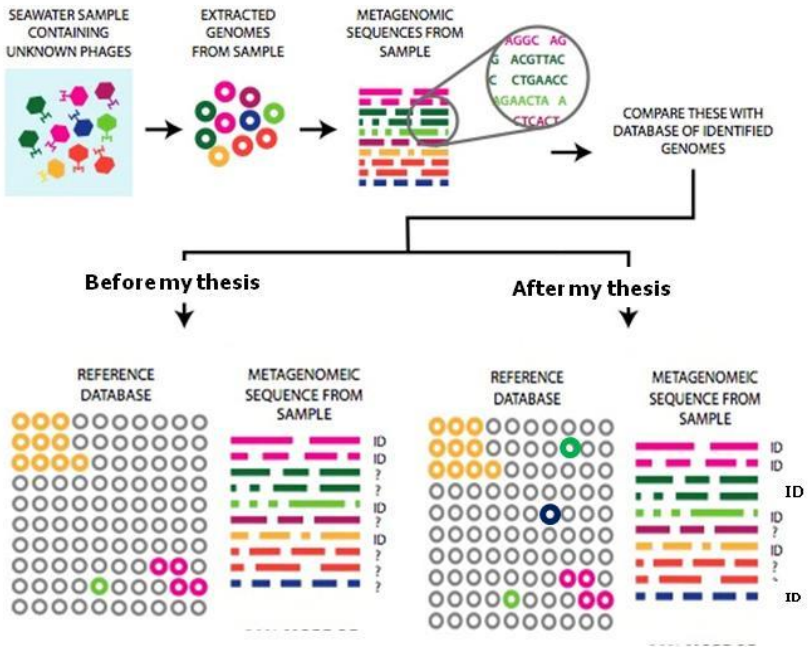


Figure 6-1. A simplified model that shows how the addition of the cyanophage sequences from this dissertation increase the database of known viruses and help in the identification of sequences in metagenomic databases.
To identify the viruses in an aquatic sample, the viral genomes are extracted and metagenomic data are generated through random sequencing, and the sequences are compared to a reference library of known viruses. The addition of cyanophages S-EIV1, A-1, and N-1 results in the identification of more sequences from metagenomic databases. In the reference database, the orange, light green and pink circles represent the known phages, while the dark green and blue circles represent the cyanophages S-EIV1(Chapter 2) and the *Nostoc* cyanophages (Chapter 3), respectively. A question mark next to a row of sequences indicates the sequences remain unidentified, and an ID indicates the sequences are known. Adapted from Culley (215).

6.4 Future perspectives

The work presented in this dissertation advocates the importance of screening different host strains for discovery of new viruses, as demonstrated recently with previously unknown groups of viruses isolated on *Pelagibacter ubique* (119) and *Cellulophaga baltica* (137). Most isolated and sequenced phages infect only four of 45 known bacterial phyla (Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria of the class Gammaproteobacteria) (137). Clearly,

there is enormous potential to isolate representatives of previously unknown groups of viruses by screening untested taxa of host organisms.

In addition, this dissertation underlines the importance of focusing on the biology of new viruses. First, there is a need to identify the function of unknown genetic sequences. Cyanophage genomes in this dissertation showed a high proportion of genetic sequences of unknown function. For example, the S-EIV1 ORFs did not share recognizable similarity to genes that encode structural proteins other than the large terminase subunit and a distant viral morphogenesis protein. Experimental phage proteomics is essential to identify more structural proteins. Second, there is a need to isolate new model systems to further study host-virus interactions. For example, the cyanophage S-EIV1 along with its host is a new model system for studying host-virus interactions and examining viral diversity in polar regions. Lytic cycle experiments for the cyanophage S-EIV1 were not performed for this dissertation and knowledge of the virus life cycle still remains to be addressed. Experiments investigating the effect of temperature on S-EIV1 burst size and viral replication should be considered. I also hope that this dissertation will lead to future work on the characterization of the polar *Synechococcus* sp. strain PCCC-A2c (S-EIV1 host). Genomic information for *Synechococcus* sp. strain PCCC-A2c is essential to study host and phage gene expression during infection and for a broader understanding of co-evolutionary processes.

Finally, I hope this dissertation will inspire the use of DNA-SIP on viral assemblages with other complex carbon and nitrogen substrates. Although metagenomic analysis of viral assemblages has provided insights into the composition and diversity of whole viral assemblages, most of these sequences are unknown and cannot be categorized to specific types of viruses to hosts they infect. In environments where different hosts utilize different substrate,

coupling DNA-SIP with viral metagenomic analysis would help catalog viruses based on their host functions and facilitate the identification of viruses infecting new or alternate hosts.

6.5 Conclusion

Considering the very limited information available on viruses infecting polar cyanobacteria before I started this work, this dissertation greatly expanded our knowledge on the topic. Using culture dependent and independent approaches, I revealed new information concerning the cyanophages present in polar regions that can further be used to examine viral diversity. I also provided a new virus-host system, as well as developed a promising new technique to explore and identify active viruses in aquatic systems. Because Arctic freshwater systems exist in a delicate balance, it is vital now to determine a baseline of viral diversity in polar regions to better understand the changes that are occurring at an unprecedented rate due to climate change.

Bibliography

1. **Neufeld JD, Mohn WW**. 2005. Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils , revealed by serial analysis of ribosomal sequence tags. Appl. Environ. Microbiol. **71**:5710–5718.
2. **Van Hove P, Vincent WF, Galand PE, Wilmotte A**. 2008. Abundance and diversity of picocyanobacteria in High Arctic lakes and fjords. Arch. Hydrobiol. Suppl. Algol. Stud. **126**:209–228.
3. **Bottos EM, Vincent WF, Greer CW, Whyte LG**. 2008. Prokaryotic diversity of arctic ice shelf microbial mats. Environ. Microbiol. **10**:950–966.
4. **Pouliot J, Galand PE, Lovejoy C, Vincent WF**. 2009. Vertical structure of archaeal communities and the distribution of ammonia monooxygenase A gene variants in two meromictic High Arctic lakes. Environ. Microbiol. **11**:687–699.
5. **Mueller DR, Pollard WH**. 2004. Gradient analysis of cryoconite ecosystems from two polar glaciers. Polar Biol. **27**:66–74.
6. **Christner BC, Kvitko BH, Reeve JN**. 2003. Molecular identification of bacteria and Eukarya inhabiting an Antarctic cryoconite hole. Extremophiles **7**:177–183.
7. **Jungblut AD, Lovejoy C, Vincent WF**. 2010. Global distribution of cyanobacterial ecotypes in the cold biosphere. ISME J. **4**:191–202.
8. **Jeffries MO**. 1992. Arctic Ice shelves and ice islands: origin, growth and disintegration, physical characteristics, variability, and dynamics. Rev. Geophys. 245–267.
9. **Mueller DR**. 2003. Break-up of the largest Arctic ice shelf and associated loss of an epishelf lake. Geophys. Res. Lett. **30**:2031.
10. **Taton A, Grubisic S, Ertz D, Hodgson DA, Piccardi R, Biondi N, Tredici MR, Mainini M, Losi D, Marinelli F, Wilmotte A**. 2006. Polyphasic study of Antarctic cyanobacterial strains. J. Phycol. **42**:1257–1270.
11. **Strunecký O, Elster J, Komárek J**. 2010. Phylogenetic relationships between geographically separate Phormidium cyanobacteria: is there a link between north and south polar regions? Polar Biol. **33**:1419–1428.
12. **Kleinteich J, Wood SA, Küpper FC, Camacho A, Quesada A, Frickey T, Dietrich DR**. 2012. Temperature-related changes in polar cyanobacterial mat diversity and toxin production. Nature **2**:356–360.
13. **Vincent WF, Hobbie JE, Laybourn-Parry J**. 2008. Introduction to the limnology of high-latitude lake and river ecosystems, p. 1–23. *In* Vincent, WF, Laybourn-Parry, J (eds.), Polar Lakes and Rivers Limnology of Arctic and Antarctic Aquatic Ecosystems. Oxford Biology.
14. **Powell L, Bowman J, Skerratt J, Franzmann P, Burton H**. 2005. Ecology of a novel *Synechococcus* clade occurring in dense populations in saline Antarctic lakes. Mar. Ecol. Prog. Ser. **291**:65–80.
15. **Vincent WF, Gibson JAE, Pienitz R, Villeneuve V**. 2000. Ice Shelf Microbial Ecosystems in the High Arctic and Implications for Life on Snowball Earth. Naturwissenschaften **87**:137–141.

16. **Rankin I, Franzmann PD, McMeekin TA, Burton HR**. 1997. Seasonal distribution of picocyanobacteria in Ace Lake, a marine derived Antarctic Lake, p. 178–184. *In* Antarctic Communities, Species, Structure and Survival,. University of Cambridge Press.
17. **Robertson BR, Tezuka N, Watanabe MM**. 2001. Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. Int. J. Syst. Evol. Microbiol. **51**:861–871.
18. **Stockner JG**. 1988. Phototrophic picoplankton: An overview from marine and freshwater ecosystems. Limnol. Oceanogr. **33**:765–775.
19. **Li WKW, Subba Rao D V, Harrison WG, Smith JC, Cullen JJ, Irwin B, Platt T**. 1983. Autotrophic picoplankton in the Tropical Ocean. Science. **219**:292–295.
20. **Campbell L, Vaulot D**. 1993. Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (station ALOHA). Deep. Res. I **40**:2043–2060.
21. **Toledo G, Palenik B, Brahamsha B**. 1999. Swimming marine *Synechococcus* strains with widely different photosynthetic pigment ratios form a monophyletic group. Appl. Environ. Microbiol. **65**:5247–5251.
22. **Stomp M, Huisman J, Vörös L, Pick FR, Laamanen M, Haverkamp T, Stal LJ**. 2007. Colourful coexistence of red and green picocyanobacteria in lakes and seas. Ecol. Lett. **10**:290–298.
23. **Katano T, Nakano S, Ueno H, Mitamura O, Anbutsu K, Kihira M, Satoh Y, Drucker V, Sugiyama M**. 2005. Abundance, growth and grazing loss rates of picophytoplankton in Barguzin Bay, Lake Baikal. Aquat. Ecol. **39**:431–438.
24. **Rae R, Vincent WF**. 1998. Phytoplankton production in subarctic lake and river ecosystems: development of a photosynthesis-temperature-irradiance model. J. Plankton Res. **20**:1293–1312.
25. **Mueller DR, Vincent WF, Bonilla S, Laurion I**. 2005. Extremotrophs , extremophiles and broadband pigmentation strategies in a high arctic ice shelf ecosystem. FEMS Microbiol. Ecol. **53**:73–87.
26. **Rautio M, Vincent WF**. 2006. Benthic and pelagic food resources for zooplankton in shallow high-latitude lakes and ponds. Freshw. Biol. **51**:1038–1052.
27. **Vézina S, Vincent WF**. 1997. Arctic cyanobacteria and limnological properties of their environment: Bylot Island, Northwest Territories, Canada (73N, 80W). Polar Biol. **17**:523–534.
28. **Vincent WF, Downes MT, Castenholz RW, Howard-Williams C**. 1993. Community structure and pigment organisation of cyanobacteria-dominated microbial mats in Antarctica. Eur. J. Phycol. **28**:213–221.
29. **Elster J, Svoboda J, Komárek J, Marvan P**. 1997. Algal and cyanoprocaryote communities in a glacial stream, Sverdrup Pass, 79N, Central Ellesmere Island, Canada. Arch. Hydrobiol. Suppl. Algol. Stud. **85**:57–93.
30. **Fernández-Valiente E, Camacho A, Rochera C, Rico E, Vincent WF, Quesada A**. 2007. Community structure and physiological characterization of microbial mats in Byers Peninsula, Livingston Island (South Shetland Islands, Antarctica). FEMS Microbiol. Ecol. **59**:377–385.

31. **Wharton RA, Parker BC, Simmons GM**. 1983. Distribution, species composition and morphology of algal mats in Antarctic dry valley lakes. Phycologia **22**:355–365.
32. **Hawes I, Schwarz A**. 1999. Photosynthesis in an extreme shade environment: Benthic microbial mats from Lake Hoare, a permanently ice-covered antarctic lake **459**:448–459.
33. **Vopel K, Hawes I**. 2006. Photosynthetic performance of benthic microbial mats in Lake Hoare, Antarctica. Limnol. Oceanogr. **51**:1801–1812.
34. **Jungblut A, Hawes I, Mountfort D, Hitzfeld B, Dietrich DR, Burns BP, Neilan BA**. 2005. Diversity within cyanobacterial mat communities in variable salinity meltwater ponds of McMurdo Ice Shelf , Antarctica. Environ. Microbiol. **7**:519–529.
35. **Paerl HW, Pinckney JL**. 1996. A mini-review of microbial consortia: their roles in aquatic production and biogeochemical cycling. Microb. Ecol. **31**:225–247.
36. **Kalff J, Welch HE**. 1974. Phytoplankton production in Char Lake, a natural polar lake, and in Meretta Lake, a polluted polar lake, Cornwallis Island, Northwest Territories. J. Fish. Res. board Canada **31**:621–636.
37. **Villeneuve V, Vincent WF, Komarek J**. 2001. Community structure and microhabitat characteristics of cyanobacterial mats in an extreme high Arctic environment: Ward Hunt Lake, p. 1999–224. *In* Elster, J, Seckbach, J, Vincent, W, Lhotsky, O (eds.), Algae and Extreme Environments. Nova Hedwigia Beihefte.
38. **Vincent WF**. 2000. Cyanobacterial dominance in the polar regions, p. 321–340. *In* Whitton, BA, Potts, M (eds.), Ecology of Cyanobacteria Their Diversity in Time and Space. Kluwer Academic Publishers.
39. **Komárek J, Elster J, Komárek O**. 2008. Diversity of the cyanobacterial microflora of the northern part of James Ross Island, NW Weddell Sea, Antarctica. Polar Biol. **31**:853–865.
40. **Bergh O, Borsheim KY**. 1989. High abundance of viruses found in aquatic environments. Nature **340**:467–468.
41. **Proctor LM, Fuhrman JA**. 1990. Viral mortality of marine bacteria and cyanobacteria. Nature **343**:60–62.
42. **Anesio AM, Mindl B, Laybourn-Parry J, Hodson AJ, Sattler B**. 2007. Viral dynamics in cryoconite holes on a high Arctic glacier (Svalbard). J. Geophys. Res. **112**:G04S31.
43. **Sawstrom C, Lisle J, Anesio AM, Priscu JC, Laybourn-Parry J**. 2008. Bacteriophage in polar inland waters. Extremophiles **12**:167–175.
44. **Fuhrman JA**. 1999. Marine viruses and their biogeochemical and ecological effects.
45. **Suttle CA**. 2005. Viruses in the sea. Nature **437**:356–361.
46. **Chibani-Chennoufi S, Bruttin A, Dillmann M-L, Brussow H**. 2004. Phage-Host Interaction : an Ecological Perspective. J. Bacteriol. **186**:3677–3686.
47. **Suttle CA**. 2007. Marine viruses-major players in the global ecosystem. Nat. Rev. Microbiol **5**:801–812.
48. **Jiang SC, Paul JH**. 1998. Significance of lysogeny in the marine environment: studies with isolates and a model of lysogenic phage production. Microb. Ecol **35**:235–243.
49. **Suttle CA, Chan AM**. 1994. Dynamics and distribution of cyanophages and their effect on Marine *Synechococcus* spp. Appl. Environ. Microbiol. **60**:3167–3174.
50. **Wilson WH, Joint IR, Carr NG, Mann NH**. 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp . strain WH7803. Appl. Environ. Microbiol. **59**:3736–3743.

51. **Waterbury JB, Valois FW**. 1993. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. Appl. Environ. Microbiol. **59**:3393–3399.
52. **Suttle CA**. 2000. Cyanophages and their role in the ecology of cyanobacteria. Chapter 20, p. 563–589. *In* Whitton, BA, Potts, M (eds.), The Ecology of Cyanobacteria: Their diversity in time and space. Kluwer Academic Publishers, Boston.
53. **Suttle CA, Chan AM**. 1993. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. Mar. Ecol. Prog. Ser. **92**:99–109.
54. **Sullivan MB, Waterbury JB, Chisholm SW**. 2003. Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. Nature **424**:1047–1051.
55. **Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW**. 2005. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol. **3**:e144.
56. **Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF, King J, Hatfull GF, Lawrence JG, Hendrix RW**. 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. Environ. Microbiol. **9**:1675–1695.
57. **Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch HM**. 2005. The Genome of S-PM2 , a " photosynthetic " T4-type bacteriophage that infects marine *Synechococcus* strains. J. Bacteriol. **187**:3188–3200.
58. **Huang S, Wang K, Jiao N, Chen F**. 2012. Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage-host genetic exchanges. Environ. Microbiol. **14**:540–558.
59. **Sabehi G, Shaulov L, Silver DH, Yanai I, Harel A, Lindell D**. 2012. A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. Proc. Natl. Acad. Sci. U. S. A. **109**:2037–2042.
60. **Safferman RS, Morris ME**. 1963. Algal virus: isolation. Science **140**:679–680.
61. **Safferman RS, Morris ME**. 1964. Growth characteristics of the blue-green algal virus LPP-1. J. Bacteriol. **88**:771–775.
62. **Adolph W, Haselkorn R**. 1971. Isolation and characterization of a virus infecting the blue-green alga Nostoc muscorum. Virology **208**:200–208.
63. **Liu X, Shi M, Kong S, Gao Y, An C**. 2007. Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium Phormidium foveolarum: complete genome sequence and DNA translocation. Virology **366**:28–39.
64. **Yoshida T, Takashima Y, Tomaru Y, Shirai Y, Takao Y, Hiroishi S, Nagasaki K**. 2006. Isolation and characterization of a cyanophage infecting the toxic cyanobacterium *Microcystis aeruginosa*. Appl. Environ. Microbiol. **72**:1239–1247.
65. **Dreher TW, Brown N, Bozarth CS, Schwartz AD, Riscoe E, Thrash C, Bennett SE, Tzeng S-C, Maier CS**. 2011. A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. Environ. Microbiol. **13**:1858–1874.
66. **Liu X, Kong S, Shi M, Fu L, Gao Y, An C**. 2008. Genomic analysis of freshwater cyanophage Pf-WMP3 Infecting cyanobacterium *Phormidium foveolarum*: the conserved elements for a phage. Microb. Ecol. **56**:671–680.

67. **Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ**. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region : implications for mechanisms of cyanophage evolution. Environ. Microbiol. **11**:2370–2387.
68. **Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, Defrancesco AS, Kern SE, Thompson LR, Young S, Yandava C, Fu R, Krastins B, Chase M, Sarracino D, Osburne MS, Henn MR, Chisholm SW**. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ. Microbiol. **12**:3035–3056.
69. **Chen F, Lu J**. 2002. Genomic Sequence and Evolution of Marine Cyanophage P60 : a New Insight on Lytic and Lysogenic Phages. Appl. Environ. Microbiol. **68**:2589–2594.
70. **Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, Sarracino D, Chisholm SW**. 2009. The genome and structural proteome of an ocean siphovirus : a new window into the cyanobacterial " mobilome " Environ. Microbiol. **11**:2935–2951.
71. **Labrie SJ, Frois-Moniz K, Osburne MS, Kelly L, Roggensack S, Sullivan MB, Gearin G, Zeng Q, Fitzgerald M, Henn MR, Chisholm SW**. 2013. Genomes of marine cyanopodoviruses reveal multiple orgins of diversity.. Environ. Microbiol **15**:1356–1376.
72. **Mann NH, Cook A, Bailey S, Clokie M, Amanullah A, Azam N, Balliet A, Hollander C, Hoffman B, Jr AF, Liebermann D, Zazzeroni F, Papa S, Smaele E De, Franzoso G**. 2003. Bacterial photosynthesis genes in a virus. Nature **424**:741–742.
73. **Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW**. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. U. S. A. **101**:11013–8.
74. **Yoshida T, Nagasaki K, Takashima Y, Shirai Y, Tomaru Y, Takao Y**. 2008. Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. J Bacteriol **190**:1762–1772.
75. **Gao E-B, Gui J-F, Zhang Q-Y**. 2012. A novel cyanophage with cyanobacterial non-bleaching protein A gene in the genome. J. Virol. **86**:236–245.
76. **Short CM, Suttle CA**. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Env. Microbiol **71**:480–486.
77. **Filée J, Tétart F, Suttle CA, Krisch HM**. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc. Natl. Acad. Sci. U. S. A. **102**:12471–12476.
78. **Wang G, Asakawa S, Kimura M**. 2011. Spatial and temporal changes of cyanophage communities in paddy field soils as revealed by the capsid assembly protein gene g20. FEMS Microbiol. Ecol. **76**:352–359.
79. **Marston MF, Amrich CG**. 2009. Recombination and microdiversity in coastal marine cyanophages. Environ. Microbiol. **11**:2893–2903.
80. **Baker AC, Goddard VJ, Davy J, Schroeder DC, Adams DG, Wilson WH**. 2006. Identification of a diagnostic marker to detect freshwater cyanophages of filamentous cyanobacteria. Appl. Environ. Microbiol **72**:5713–5719.
81. **Dorigo U, Jacquet S**. 2004. Cyanophage diversity , inferred from g20 gene analyses , in the largest natural lake. Appl. Environ. Microbiol. **70**:1017–1022.

82. **Sullivan MB, Coleman ML, Quinlivan V, Rosenkrantz JE, Defrancesco AS, Tan G**. 2008. Portal protein diversity and phage ecology. Environ. Microbiol **10**:2810–2823.
83. **Labonté JM, Reid KE, Suttle CA**. 2009. Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. Appl. Environ. Microbiol. **75**:3634–3640.
84. **Breitbart M, Miyake JH, Rohwer F**. 2004. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol. Lett. **236**:249–256.
85. **Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielaski J**. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. **4**:1344–1357.
86. **Chénard C, Suttle CA**. 2008. Phylogenetic Diversity of Sequences of Cyanophage Photosynthetic Gene psbA in Marine and Freshwaters. Appl Env. Microbiol **74**:5317–5324.
87. **Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, Béjà O**. 2005. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environ. Microbiol. **7**:1505–1513.
88. **Sandaa R, Clokie M, Mann NH**. 2008. Photosynthetic genes in viral populations with a large genomic size range from Norwegian coastal waters. FEMS Microbiol. Ecol. **63**:2–11.
89. **Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F**. 2006. The marine viromes of four oceanic regions. PLoS Biol. **4**:e368.
90. **Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, Mcdaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F**. 2008. Functional metagenomic profiling of nine biomes. Nature **452**:629–632.
91. **Mcdaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F, Paul JH**. 2008. Metagenomic analysis of lysogeny in Tampa Bay : Implications for prophage gene expression. PLoS One **3**:e3263.
92. **Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M**. 2009. Metagenomic analysis of viruses in reclaimed water. Environ. Microbiol. **11**:2806–2820.
93. **Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D**. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One **7**:e33641.
94. **Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards R a, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F**. 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature **452**:340–343.
95. **Edwards RA, Rohwer F**. 2005. Viral metagenomics. Nat. Rev. Microbiol. 6–12.
96. **Roux S, Krupovic M, Debroas D, Forterre P, Enault F**. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. **3**:130160.

97. **Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, Martin J, Verdú M, Deschamps P, Moreira D, López-García P, Mira A, Rodriguez-Valera F**. 2010. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J. **4**:1154–1166.
98. **Mizuno CM, Rodriguez-Valera F, Garcia-Heredia I, Martin-Cuadrado A-B, Ghai R**. 2013. Reconstruction of novel cyanobacterial siphovirus genomes from Mediterranean metagenomic fosmids. Appl. Environ. Microbiol. **79**:688–695.
99. **DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM**. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311**:496–503.
100. **Bergeron M, Vincent WF**. 1997. Microbial food web responses to phosphorus and solar UV radiation in a subarctic lake. Aquat. Microb. Ecol. **12**:239–249.
101. **Lawrence JG, Hatfull GF, Roger W, Hendrix RW**. 2002. Imbroglios of viral taxonomy : Genetic exchange and failings of phenetic approaches. J. Bacteriol. **184**:4891–4905.
102. **Fox JA, Booth SJ, Martin EL**. 1976. Cyanophage SM-2: A new blue-green algal virus. Virology **73**:557–560.
103. **Rippka R, Desruelles J, Waterbury JB, Herdman M, Stanier R**. 1979. Generic assigments, strain histories and properties of pure cultures of cyanobacteria. J. Gen. Microbiol. **111**:1–61.
104. **Suttle CA, Chan AM, Cottrell MT**. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Appl. Environ. Microbiol. **57**:721–726.
105. **Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J**. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. **13**:R122.
106. **Lukashin A V, Borodovsky M**. 1998. GeneMark . hmm : new solutions for gene finding. DNA Seq. **26**:1107–1115.
107. **Delcher AL, Bratke KA, Powers EC, Salzberg SL**. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23**:673–679.
108. **Lowe TM, Eddy SR**. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.
109. **Laslett D, Canback B**. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. **32**:11–16.
110. **Lavigne R, Sun WD, Volckaert G**. 2004. PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. Bioinformatics **20**:629–635.
111. **Grant JR, Stothard P**. 2008. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. **36**:W181–184.
112. **Pei J, Grishin N V**. 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics **23**:802–808.
113. **Pei J, Kim B-H, Tang M, Grishin N V**. 2007. PROMALS web server for accurate multiple protein sequence alignments. Nucleic Acids Res. **35**:W649–52.
114. **Drummond A, Ashton B, Buxton S, Cheung M, Cooper A**. 2011. Geneious v5.4.
115. **Stamatakis A, Hoover P, Rougemont J**. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. **57**:758–771.

116. **Schmidt HF, Sakowski EG, Williamson SJ, Polson SW, Wommack KE**. 2014. Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. ISME J. **8**:103–114.
117. **Pajunen MI, Elizondo MR, Skurnik M, Kieleczawa J, Molineux IJ**. 2002. Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3. J. Mol. Biol. **319**:1115–1132.
118. **Choi M, Miller A, Rothman-denes LB**. 1995. Identification , cloning , and characterization of the bacteriophage N4 gene encoding the single-stranded DNA-binding protein **270**:22541–22547.
119. **Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ**. 2013. Abundant SAR11 viruses in the ocean. Nature **494**:357–360.
120. **Rohwer F, Segall A, Steward G, Seguritan V, Breitbart M, Wolven F, Azam F**. 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. Limnol. Oceanogr. **45**:408–418.
121. **Wang K, Chen F**. 2008. Prevalence of highly host-specific cyanophages in the estuarine environment. Environ. Microbiol. **10**:300–312.
122. **Ackermann HW.** 2006. Classification of bacteriophages, p. 8–16. *In* Calendar, R (ed.), The bacteriophages, 2nd ed. Oxford University Press, New York, NY.
123. **Raytcheva D a, Haase-Pettingell C, Piret JM, King J a**. 2011. Intracellular assembly of cyanophage Syn5 proceeds through a scaffold-containing procapsid. J. Virol. **85**:2406–15.
124. **Millard A, Clokie MRJ, Shub DA, Mann NH**. 2004. Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. Proc. Natl. Acad. Sci. U. S. A. **101**:11007–11012.
125. **Kotzabasis K, Strasser B, Navakoudis E, Senger H, Dornemann**. 1999. The regulatory role of polyamines in structure and functioning of the photosynthetic apparatus during photoadaptation. J. Photochem. Photobiol. B Biol. **50**:45–52.
126. **Mulo P, Laakso S, Mäenpää P, Aro EM**. 1998. Stepwise photoinhibition of photosystem II. Studies with Synechocystis species PCC 6803 mutants with a modified D-E loop of the reaction center polypeptide D1. Plant Physiol. **117**:483–490.
127. **Clokie MR, Millard AD, Mann NH**. 2010. T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. Virol. J. **7**:291.
128. **Ignacio-Espinoza JC, Sullivan MB**. 2012. Phylogenomics of T4 cyanophages: lateral gene transfer in the "core" and origins of host genes. Environ. Microbiol. **14**:2113–2126.
129. **Pen MMO, Bullerjahn GS**. 1995. The DpsA protein of *Synechococcus* sp . strain PCC7942 is a DNA-binding hemoprotein. J. Biol. Chem. **270**:22478–22482.
130. **Storz G, Tartaglia LA, Farr SB, Ames BN**. 1990. Bacterial defenses against oxidative stress. Trends Genet. **6**:363–8.
131. **Farr SB, Kogoma T**. 1991. Oxidative stress responses in *Escherichia coli* and *Salmonella* typhimurium . Microbiol. Mol. Biol. Rev. **55**:561–585.
132. **Martinez A, Kolter R**. 1997. Protection of DNA during oxidative stress by the nonspecific DNA-binding protein Dps. J. Bacteriol. **179**:5188–5194.
133. **Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW**. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc. Natl. Acad. Sci. U. S. A. **108**:E757–E764.

134. **Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R, Herdman M, Sivonen K, Coursin T, Laurent T, Goodwin L, Nolan M, Davenport KW, Han CS, Rubin EM, Eisen JA, Woyke T, Gugger M, Kerfeld CA**. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc. Natl. Acad. Sci. U. S. A. **110**:1053–1058.
135. **Rao V, Feiss M**. 2008. The bacteriophage DNA packaging motor. Annu. Rev. Genet. **42**:647–681.
136. **Kang I, Oh H-M, Kang D, Cho J-C**. 2013. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. Proc. Natl. Acad. Sci. U. S. A. **110**:12343–8.
137. **Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, Verberkmoes NC**. 2013. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc. Natl. Acad. Sci. U. S. A. **11031**:12798–12803.
138. **Agawin NSR, Agustf S**. 1997. Abundance, frequency of dividing cells and growth rates of *Synechococcus* sp . (Cyanobacteria) in the stratified Northwest Mediterranean Sea. J. Plankton Res. **19**:1599–1615.
139. **Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen J a, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC**. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5**:e77.
140. **López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A**. 2009. High diversity of the viral community from an Antarctic lake. Science **326**:858–861.
141. **Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D**. 2008. Assembly of viral metagenomes from Yellowstone hot springs. Appl. Environ. Microbiol. **74**:4164–4174.
142. **Breitbart M**. 2012. Marine Viruses: Truth or Dare. Ann. Rev. Mar. Sci. **4**:425–448.
143. **Comeau AM, Arbiol C, Krisch HM**. 2010. Gene network visualization and quantitative synteny analysis of more than 300 marine T4-like phage scaffolds from the GOS metagenome. Mol. Biol. Evol. **27**:1935–1944.
144. **Priscu JC, Adams EE, Paerl HW, Fritsen CH, Dore JE, John T, Wolf CF, Mikucki JA**. 2005. Perennial antarctic lake ice: a refuge for cyanobacteria in an extreme environment, p. 22–49. *In* Castello, JD, Rogers, SO (eds.), Life in Ancient Ice. Princeton Press.
145. **Moreno J, Rodriguez H, Vargas MA, Rivas J, Guerrero MG**. 1995. Nitrogen-fixing cyanobacteria as source of phycobiliprotein pigments. Composition and growth performance of ten filamentous heterocystous strains J. Appl. Phycol. 17–23.
146. **Kim J-D**. 2006. Screening of Cyanobacteria (blue-green algae) from rice paddy soil for anti-fungal activity against plant pathologenic fungi. Mycobiology **34**:138–142.
147. **Potts M**. 2000. Nostoc, p. 465–504. *In* Whitton, B, Potts, M (eds.), The Ecology of Cyanobacteria.

148. **Koz'yakow S**. 1977. Cyanophages of the series A(L) specific for the blue-green alga Anabaena variabilis. Eksp. naya al' gologiya. **25**:151–175.
149. **Adolph KW, Haselkorn R**. 1972. Photosynthesis and the development of blue-green algal virus N-1. Virology **47**:370–374.
150. **Sambrook J, Russell D**. 2001. Molecular Cloning: A Laboratory ManualMolecular Cloning: A Laboratory Manual (3rd ed.).3rd editio. Cold Spring Harbor Laboratory Press.
151. **Crooks GE, Hon G, Chandonia J, Brenner SE**. 2004. WebLogo : a sequence logo generator. Genome Res. 1188–1190.
152. **Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ru W**. 2003. Bacteriophage T4 Genome. Microbiol. Mol. Biol. Rev. **67**:86–156.
153. **Comeau AM, Tremblay D, Moineau S, Rattei T, Kushkina AI, Tovkach FI, Krisch HM, Ackermann H-W**. 2012. Phage morphology recapitulates phylogeny: the comparative genomics of a new group of myoviruses. PLoS One **7**:e40102.
154. **Yin Y, Fischer D**. 2008. Identification and investigation of ORFans in the viral world. BMC Genomics **9**:24.
155. **Enav H, Béjà O, Mandel-Gutfreund Y**. 2012. Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. ISME J. **6**:619–628.
156. **Hu N-T, Thiel T, Giddings TH, Wolk CP**. 1981. New *Anabaena* and *Nostoc* cyanophages from sewage settling ponds. Virology **246**:236–246.
157. **Mann NH, Clokie MRJ**. 2012. Cyanophages, p. 535–557. *In* Whitton, BA (ed.), Ecology of Cyanobacteria II. Springer Netherlands, Dordrecht.
158. **Hertveldt K, Lavigne R, Pleteneva E, Sernova N, Kurochkina L, Korchevskii R, Robben J, Mesyanzhinov V, Krylov VN, Volckaert G**. 2005. Genome comparison of *Pseudomonas aeruginosa* large phages. J. Mol. Biol. **354**:536–545.
159. **Bancroft I, Smith R**. 1988. An analysis of restriction endonuclease sites in cyanophages infecting the heterocystous cyanobacteria *Anabaena* and *Nostoc*. J. Gen. Microbiol. **69**:739–743.
160. **Kanesaki Y, Suzuki I, Allakhverdiev SI, Mikami K, Murata N**. 2002. Salt stress and hyperosmotic stress regulate the expression of different sets of genes in *Synechocystis* sp. PCC 6803. Biochem. Biophys. Res. Commun. **290**:339–348.
161. **Tetu SG, Brahamsha B, Johnson DA, Tai V, Phillippy K, Palenik B, Paulsen IT**. 2009. Microarray analysis of phosphate regulation in the marine cyanobacterium *Synechococcus* sp. WH8102. ISME J. **3**:835–849.
162. **Paul JH, McDaniel LD**. 2006. Temperate and lytic cyanophages from the Gulf of Mexico. J. Mar. Biol. Assoc. United Kingdom **86**:517–527.
163. **Ortmann AC, Lawrence JE, Suttle CA**. 2002. Lysogeny and lytic viral production during a bloom of the cyanobacterium *Synechococcus* spp. Microb. Ecol. **43**:225–231.
164. **Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW**. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet. **3**:e231.
165. **Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP, Paulsen IT, de Marsac NT, Wincker P, Dossat C, Ferriera S, Johnson J, Post AF, Hess WR,**

Partensky F. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. **9**:R90.
166. **Horvath P, Barrangou R**. 2010. CRISPR/Cas, the immune system of Bacteria and Archaea. Science **327**:167–170.
167. **Bhaya D, Grossman AR, Steunou A, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF**. 2007. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. 703–713.
168. **Deveau H, Garneau JE, Moineau S**. 2010. CRISPR/Cas system and its role in phage-bacteria interactions. Annu. Rev. Microbiol. **64**:475–493.
169. **Bhaya D, Davison M, Barrangou R**. 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu. Rev. Genet. **45**:273–297.
170. **Marraffini LA, Sontheimer EJ**. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature **11**:181–190.
171. **Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin E V**. 2011. Evolution and classification of the CRISPR-Cas systems. Nature **9**:467–777.
172. **Labrie SJ, Samson JE, Moineau S**. 2010. Bacteriophage resistance mechanisms. Nature **8**:317–327.
173. **Van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJJ**. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. **34**:401–407.
174. **Lillestøl RK, Redder P, Garrett RA, Brügger KIM**. 2006. A putative viral defence mechanism in archaeal cells 59–72.
175. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero D A, Horvath P**. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315**:1709–1712.
176. **Lillestøl RK, Shah SA, Brügger K, Redder P, Phan H, Christiansen J, Garrett RA**. 2009. CRISPR families of the crenarchaeal genus Sulfolobus : bidirectional transcription and dynamic properties **72**:259–272.
177. **Kunin V, Sorek R, Hugenholtz P**. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. **8**:R61.
178. **Cai F, Axen S, Kerfeld CA**. 2013. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. RNA Biol. **10**:1–7.
179. **Seed KD, Lazinski DW, Calderwood SB, Camilli A**. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature **494**:489–491.
180. **Godde JS, Bickerton A**. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J. Mol. Evol. **62**:718–729.
181. **Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MTG, Wright A, Churcher C, Quail M a, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S,**

Dupuy B, Dougan G, Barrell B, Parkhill J. 2006. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nature **38**:779–786.
182. **Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD**. 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. **21**:1616–1625.
183. **Grissa I, Vergnaud G, Pourcel C**. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. **35**:W52–7.
184. **Garcia-pichel F, Nu U, Muyzer G**. 1997. PCR Primers To Amplify 16S rRNA Genes from Cyanobacteria. Appl. Environ. Microbiol. **63**:3327–3332.
185. **Makarova KS, Aravind L, Wolf YI, Koonin E V**. 2011. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct **6**:38.
186. **Agervald Å, Camsund D, Stensjö K, Lindblad P**. 2012. CRISPR in the extended hyp-operon of the cyanobacterium Nostoc sp. strain PCC 7120, characteristics and putative function(s). Int. J. Hydrogen Energy **37**:8828–8833.
187. **Horvath P, Coûté-Monvoisin A-C, Romero DA, Boyaval P, Fremaux C, Barrangou R**. 2009. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int. J. Food Microbiol. **131**:62–70.
188. **Pougach K, Semenova E, Bogdanova E, Datsenko K a, Djordjevic M, Wanner BL, Severinov K**. 2010. Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol. Microbiol. **77**:1367–1379.
189. **Mullineaux CW, Mariscal V, Nenninger A, Khanum H, Herrero A, Flores E, Adams DG**. 2008. Mechanism of intercellular molecular exchange in heterocyst-forming cyanobacteria. EMBO J. **27**:1299–1308.
190. **Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A**. 2009. Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. **7**:828–836.
191. **Varin T, Lovejoy C, Jungblut AD, Vincent WF, Corbeil J**. 2011. Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. Appl. Environ. Microbiol.**72**:549-559.
192. **Rappé MS, Giovannoni SJ**. 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57**:369–94.
193. **Neufeld JD, Wagner M, Murrell JC**. 2007. Who eats what, where and when? Isotope-labelling experiments are coming of age. ISME J. **1**:103–110.
194. **Dumont MG, Murrell JC**. 2005. Stable isotope probing: linking microbial identity to function. Nature **3**:499–504.
195. **Frias-Lopez J, Thompson A, Waldbauer J, Chisholm SW**. 2009. Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. Environ. Microbiol. **11**:512–525.
196. **Neufeld JD, Vohra J, Dumont MG, Lueders T, Manefield M, Friedrich MW, Murrell JC**. 2007. Protocol DNA stable-isotope probing. Nature **2**:860–866.
197. **Jungblut AD, Vincent WF, Lovejoy C**. 2012. Eukaryotes in Arctic and Antarctic cyanobacterial mats. FEMS Microbiol. Ecol.**82**: 416-428.
198. **Williamson KE, Radosevich M, Wommack KE**. 2005. Abundance and diversity of viruses in six Delaware soils. Appl. Environ. Microbiol. **71**: 3119-3125.

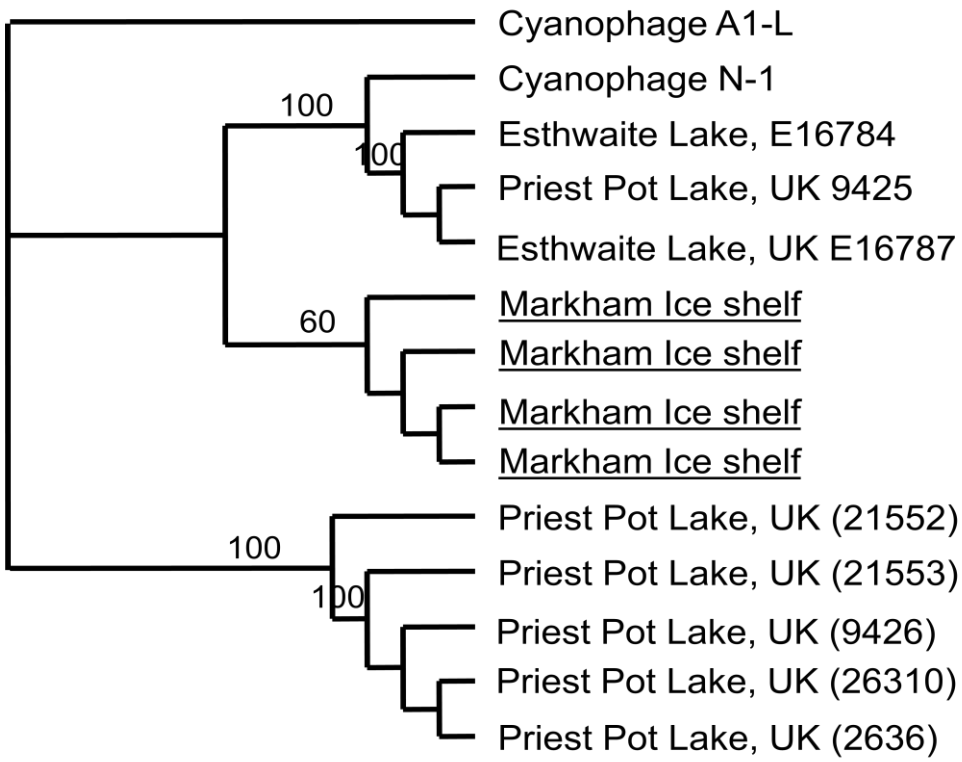
199. **Charif D, Lobry J**. 2007. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis, p. 207–232. *In* Bastolla, MU, Porto, M, Roman, H, Vendruscolo, M (eds.), Structural approaches to sequence evolution: Molecules, networks, populations. Springer Verlag, New York, NY.
200. **Wickham H**. 2009. ggplot2: elegant graphics for data analysis. Springer New York, 2009. Springer New York.
201. **Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S**. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. **8**:205–13; 227–53.
202. **Taylor P, Lee CG, Watanabe T, Fujita Y, Asakawa S, Kimura M**. 2013. Heterotrophic growth of cyanobacteria and phage- mediated microbial loop in soil : Examination by stable isotope probing (SIP) method. Soil Sci. Plant Nutr. **58**:37–41.
203. **Paul JH**. 2008. Prophages in marine bacteria : dangerous molecular time bombs or the key to survival in the seas ? ISME J. **2**:579–589.
204. **Breitbart M, Thompson LR, Suttle CA, Sullivan MB**. 2007. Exploring the vast diversity of marine viruses. Oceanography **20**:135–139.
205. **Lionard M, Péquin B, Lovejoy C, Vincent WF**. 2012. Benthic cyanobacterial mats in the high arctic: multi-layer structure and fluorescence responses to osmotic stress. Front. Microbiol. **3**:140.
206. **Wommack KE, Colwell RR**. 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. **64**:69–114.
207. **McDaniel L, Paul JH**. 2005. Effect of nutrient addition and environmental factors on prophage induction in natural populations of marine Synechococcus species. Appl. Environ. Microbiol. **71**:842–850.
208. **Weinbauer MG, Suttle CA**. 1999. Lysogeny and prophage induction in coastal and offshore bacterial communities. Aquat. Microb. Ecol. **18**:217–225.
209. **Williamson SJ, Houchin LA, Mcdaniel L, Paul JH**. 2002. Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay , Florida. Appl. Environ. Microbiol. **68**:4307–4314.
210. **Payet JP, Suttle CA**. 2013. To kill or not to kill : The balance between lytic and lysogenic viral infection is driven by trophic status. Limnol. Oceanogr. **58**:465–474.
211. **Zobell C, Anderson DQ**. 1936. Observations on the multiplication of bacteria in different volumes of store sea water and the influence of oxygen tension and solid surfaces. Biol. Bull. **71**:324–342.
212. **Marrasé C, Lim LE, Caron DA**. 1992. Seasonal and daily changes in bacterivory in a coastal plankton community. Mar. Ecol. Prog. Ser. **82**:281–289.
213. **Thurber R V, Haynes M, Breitbart M, Wegley L, Rohwer F**. 2009. Laboratory procedures to generate viral metagenomes. Nature **4**:470–483.
214. **Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE**. 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl. Environ. Microbiol. **77**:8071–8079.

215. **Culley AI**. 2013. Insight into the unknown marine virus majority. Proc. Natl. Acad. Sci. U. S. A. **110**:12166–12167.

Appendix A Filamentous cyanobacterial strains tested for virus isolation

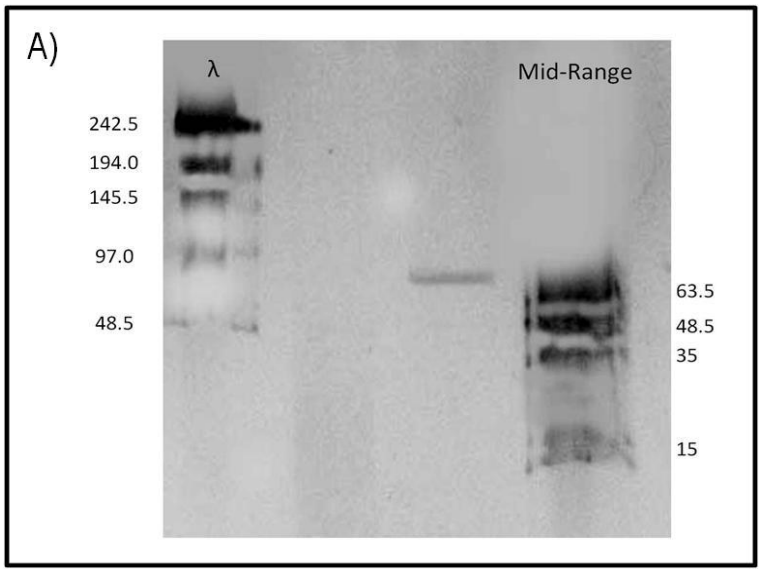
Strain ID	Location	Latitude	Longitude
PCCC-MIS15	Markham Ice Shelf	83°03'N	71°27'W
PCCC-WHL74	Ward Hunt Lake	83°05'N	74°10'W
PCCC-WHL66	Ward Hunt Lake	83°05'N	74°10'W
PCCC-A10	Lake A	83º05’N	75 º30’W
PCCC-WHL105	Ward Hunt Lake	83°05'N	74°10'W
PCCC-WHI	Ward Hunt Ice Shelf	83°01'N	71°30'W
PCCC-PA	Antoniates pond	82º58’N	75 º24’W

Appendix B Phylogenetic analysis of MCP sequences showing the presence of cyanophages related to cyanophage A-1 and N-1 in High Arctic cyanobacterial mats.



Maximum Likelihood analysis of the major capsid protein genes (MCPF5/R5) for the *Nostoc*-like cyanophages. Bootstrap for 100 replicates are marked at the nodes. Sequences from Esthwaite Lake and Priest Pot Lake were retrieved from NCBI. Sequences from Markham Ice shelf (underlined in the tree) were amplified and sequenced from nucleic acid extraction of a high artic cyanobacterial mats using the MCPF5/R5 primer as described in Baker et al (80)**.**

Appendix C Pulse-field gel electrophoresis of the S-EIV1 genome



A pulse field gel electrophoresis (PFGE) was used to determine the size of S-EIP1. Seventy ml of lysate previously filtered through a 0.45 μm pore-size filter (HVLP; Millipore,) was centrifuged for 6 hrs at 119,577 xg in a Beckman Coulter ultracentrifuge (45Ti rotor, 8°C). A subsample of resuspended viruses (100 μL) was mixed with to an equal volume of molten low melt agarose solution (1% agarose), and dispensed into a plug mold. After solidification, plug was incubated in a Proteinase K digestion buffer (250 mM EDTA, 1% sodium dodecyl sulfate, 1 mg of proteinase K ml^{-1}) overnight at room temperature. After incubation, the buffer was decanted, and the plug was washed by submerging it in 10 mM Tris–1 mM EDTA, pH 8.0 for 30 mins. The washing step was repeat 3 times. The agarose plug was then loaded into a 1% agarose PGFE gel (0.5x TBE buffer-45 mM Tris-borate, 1 mM EDTA pH 8.0). PFGE was performed under the following conditions: 0.5× TBE tank buffer (45 mM Tris-borate, 1 mM EDTA pH 8.0), 1- to 15-s pulse ramp, 120° included angle, 6.0 V cm^{-1} 14°C, and 22 h. After electrophoresis, the gel was stained in 0.1x SYBR Green solution (Invitrogen) for 60 mins, then visualized and photographed with an Alpha Imager 3400 system. . The number to the left indicates the size of the lambda marker (λ) and the number on the right indicates the size of the mid-range marker (mid-range).

Appendix D Environmental sequences from the Global Ocean Survey, their accession number in the CAMERA database.

Name	Accession number
env101	EDB49145
env102	EDF47865
env103	EDF79875
env104	EDB58248
env105	EBJ16138
env106	ECT85583
env107	ECB63617
env108	ECP52028
env109	ECR67198
env110	EBN59854
env111	EDJ46665
env112	ECR82609
env113	EDW44887
env114	EDD72074
env115	EDI36158
env116	ECW76514
env117	EDV30784
env118	EBH20789
env119	EDA72585

Appendix E Predicted ORFs of Cyanophage S-EIV1

ORF	Length (bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
2	1803	+	hypothetical protein	*Synechococcus* phage S-RIP1	$7e^{-66}$	43% (183)
3	306	+	hypothetical protein	*Synechococcus* phage S-CBP3	$1e^{-17}$	44% (45)
5	1107	+	Lysozyme	*Synechococcus* phage S-CBP3	e^{-115}	55% (205)
6	235	-	hypothetical protein (gp22)	*Synechococcus* phage S-CBS3	$7e^{-08}$	73% (25)
8	888	-	PurM	uncultured phage MedDCM-OCT-S04-C348	$9e^{-78}$	60% (135)
9	936	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$2e^{-10}$	50%(35)
10	438	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$6e^{-19}$	37% (54)
13	1266	-	glycosyl transferase group 1	uncultured phage MedDCM-OCT-S04-C348	e^{-121}	53% (209)
14	156	-	S-adenosylmethionine decarboxylase proenzyme (DUF206)	*Prochlorococcus marinus* str. AS9601	$2e^{-05}$	44%(22)
15	2640	-	P4 phage primase	uncultured phage MedDCM-OCT-S04-C348	e^{-161}	70% (265)
17	849	-	hypothetical protein PCC7424_2405	uncultured phage MedDCM-OCT-S04-C348	e^{-110}	71% (190)
19	588	+	hypothetical protein GDI_2993	*Gluconacetobacter diazotrophicus* PAl	$1e^{-25}$	39% (74)
22	219	+	hypothetical protein COCOR_02333	*Corallococcus coralloides* DSM2259	$1e^{-15}$	42% (23)
27	1464	-	sulfate adenylyltransferase	*Psychromonas ingrahamii* 37	$7.2e^{-2}$	28.4% (23)
28	552	-	DNA-binding ferritin-like protein	*Opitutaceae bacterium* TAV1	$4e^{-04}$	30% (39)
29	243	+	hypothetical protein CPKG_00047	Cyanophage KBS-S-2A	$1e^{-16}$	48% (77)
30	1719	+	hypothetical protein HMPREF1025_01337	*Lachnospiraceae bacterium*	$3e^{-12}$	37% (65)
33	621	+	hypothetical protein CPKG_00052	Cyanophage KBS-S-2A	$5e^{-23}$	48% (49)
37	261	-	hypothetical protein	Uncultured phage MedDCM-OCT-S04-C348	$6e^{-19}$	58% (43)
38	219	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$1e^{-08}$	43%(30)
39	717	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$2e^{-26}$	35% (75)
44	963	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$5e^{-36}$	44% (84)
46	216	+	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$5.3e^{-5}$	48.9% (44)
47	1104	-	hypothetical protein alr7568	Nostoc sp. PCC 7120]	$3e^{-07}$	21% (52)
50	807	-	ssDNA-binding protein	*Thermo* uncultured phage MedDCM-OCT-S04-C348	$8.7e^{-67}$	45%(237)
52	1869	-	DNA polymerase family A	uncultured phage MedDCM-OCT-S04-C348	e^{-148}	70% (245)
53	759	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$3e^{-108}$	56% (71)
55	627	-	FAD dependent thymidylate synthase	uncultured phage MedDCM-OCT-S04-C348	$3e^{-57}$	53% (116)

56	378	-	hypothetical protein HMPREF1487_04334 (endodeoxyribonuclease)	*Pseudomonas* sp. HPB0071	$3e^{-05}$	47% (26)
63	378	-	hypothetical protein SXDG_00159	*Synechococcus* phage S-RIM8	$8e^{-05}$	83 %(20)
64	1755	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$1e^{-42}$	46% (98)
67	228	-	hypothetical protein	uncultured phage MedDCM-OCT-S04-C348	$7e^{-10}$	48 %(25)
68	282	-	hypothetical protein	uncultured phage MedDCM-OCT-S09-C37	$3e^{-09}$	48% (34)
69	372	-	hypothetical protein	uncultured bacterium	$7e^{-08}$	68% (28)
81	216	+	hypothetical protein P23p111	*Thermus* phage P23-45	$7e^{-06}$	42% (24)
84	960	+	hypothetical protein	*Nodosilinea nodulosa*	$2.e^{-6}$	25.8% (219)
92	483	+	hypothetical protein Ava_D0016	*Anabaena variabilis* ATCC 29413	2.0^{-3}	35% (23)
95	2538	+	putative terminase large subunit Ava_D0014	*Anabaena variabilis* ATCC 29413	$1e^{-67}$	43%(137)
97	183	+	hypothetical protein SWPG_00102	*Synechococcus* phage S-CBM2	$2e^{-05}$	47% (23)
99	2301	+	hypothetical protein Ava_D0012	*Anabaena variabilis* ATCC 29413	$4e^{-97}$	32% (249)
100	399	+	hypothetical protein	*Nodosilinea nodulosa*	$2.93\ e^{-05}$	45.3% (39)
103	468	+	hypothetical protein Ava_D0009	*Anabaena variabilis* ATCC 29413	$6e^{-16}$	36% (63)
105	429	-	hypothetical protein	*Nodosilinea nodulosa*	$6.1e^{-4}$	30% (134)
106	744	+	peptidase, M23 family	*Acinetobacter* sp. WC-743	$3e^{-07}$	33%(44)
107	405	+	hypothetical protein Ava_D0006	*Anabaena variabilis* ATCC 29413	$4e^{-19}$	42% (58)
109	657	+	hypothetical protein	*Burkholderia ambiforia*	$3e^{-18}$	39% (48)
114	219	-	hypothetical protein	*Synechococcus* sp. Cb0101	$1.5e^{-3}$	31% (59)
115	360	+	HNH nuclease	*Synechococcus* sp. CC9902	$4e^{-16}$	42% (40)
117	303	-	hypothetical protein	*Synechococcus* phage S-CBS2	$4e^{-27}$	75 %(60)
119	663	+	Lysozyme	*Acinetobacter* sp. RUH2624	$6e^{-18}$	37% (57)
121	237	-	protein of unknown function DUF3310	*Clostridium* sp.	$6e^{-06}$	47% (34)
124	585	-	deoxycytidine triphosphate deaminase	*Synechococcus* sp. WH 7803	$2e^{-53}$	53% (104)
125	591	-	hypothetical protein VOLCADRAFT_106473	*Volvox carteri*	$6e^{-09}$	25% (50)

Appendix F Predicted ORFs for Cyanophage A-1

ORFs	Length(bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
5	2016	R	DNA polymerase B	*Cyanothece* PCC7425	$7.00e^{-115}$	40.4% (237)
11	1428	R	Type III restriction enzyme,DEAD-box helicase	Lactobacillus phage phiadh	$2.00e^{-21}$	25.8%(91)
24	939	R	putative ant AntA/AntB antirepressor	*Leptolyngbya* sp. PCC 7375	$1.90e^{-24}$	45%(51)
28	618	R	DNA N-6-adenine-methyltransferase	*Synechocystis* sp. PCC 7509	$1.00e^{-06}$	27.3%(44)
38	243	R	ASCH domain protein	*Clostridium symbiosum* ATCC 14940	$9.87e^{-16}$	44%(35)
40	564	R	Hypothetical protein	*Bacillus megaterium* WSH-002	$1.03e^{-08}$	27.7%(44)
44	3066	R	putative DNA primase	*Streptococcus thermophilus* CNRZ1066	$7.73e^{-03}$	20.4%(80)
51	1209	F	Transposase	*Nostoc* sp. PCC7120	0	388(100%)
53	285	F	Hypothetical protein	*Anabeana variabilis* ATCC 29413	$5.00e^{-31}$	
54	165	R	putative antirepressor			
56	639	F	Thymidylate kinase	*Raphidopsis brooki* D9	$1.5e^{-31}$	38.6% (78)
59	417	F	Hypothetical protein	*Calothrix* sp. PCC 7103	$2.59e^{-48}$	58% (98)
62	1044	F	DNA-cytosine methyltransferase	*Thrichodesmium erythraeum* IMS101	$3.0e^{-62}$	37.8%(140)
71	507	R	Hypothetical protein	Roseophage DSS3p2	$5.00e^{-29}$	43.2%(64)
73	726	R	Hypothetical protein	*Acidovorax avenae* subsp. *citrulli* AAC00-	$1.00e^{-07}$	29.6%(32)
76	600	F	dCTP deaminase/dUTPase superfamily	*Cyanothece* PCC7425	$2.00e^{-76}$	67.3%(134)
81	888	F	DNA methylase N-4/N-6 domain protein	*Arthrospira maxima*	$1.00e^{-65}$	49.4%(133)
82	303	R	5'-3' exonuclease		$8.00e^{-04}$	28.4%(29)
83	432	F	endodeoxyribonuclease RusA	Cyanothece PCC7425		38.4 %(86)
86	780	R	Thymidylate synthase complementing protein	Chlorobium phaeobacteroides BSI	$4.38e^{-39}$	
88	642	R	Helix-turn-helix motif			
93	1359	F	Terminase large subunit	Nostoc sp. PCC 7524 AFY48994.1	$2.28e^{-54}$	30% (129)
98	1605	F	Hypothetical protein	Nostoc sp. PCC 7524 AFY48995.1	$7.16e^{-17}$	21%(113)
100	702	F	Outer membrane protein (OmpH-like)			
101	1098	F	putative Major capsid protein	Cyanophage AN-15	$7.00e^{-159}$	77.6(288)
104	327	F	Hypothetical protein	Lactobacillus phagage KC5a	$3.94e^{-03}$	29.1%(25)
107	1521	F	Tail sheath protein	*Nostoc* sp. PCC 7524 AFY49006.1	$1.14e^{-69}$	38% (137)
110	504	F	T4-like virus tail tube protein			
112	2070	F	Phage-related tail	*Burkholderia cenocepacia* MC0-3	$2.1e^{-08}$	38%(50)
114	1182	R	Zinc metallopeptidase M23-family/Phage late control gene D	*Sagittula stellata* E-37	$2.11e^{-08}$	50.7%(38)

			protein			
115	1416	R	Lysozyme-like domain,Rare lipoprotein A	*Anabeana variabilis* ATCC 29413	$2.3e^{-19}$	57.7%(60)
121	738	R	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49010.1)	$9.39e^{-11}$	29%(67)
124	891	R	Exonuclease RNase T and DNA polymerase	*Cyanobacterium aponinum* PCC10605	$1.63e^{-08}$	27% (48)
127	438	R	Hypothetical protein	*Geobacillus* sp Y412MC10	$5.6e^{-15}$	43.6%(44)
128	405	F	LuxR-family regulator protein, helix-turn-helix motif	*Clostridium acetobutylicum* ATCC 824	$8.10e^{-05}$	42.6(54)
129	813	F	Phage-related baseplate assembly protein	*Nostoc* sp PCC7524 (AFY49015.1)	$1.46e^{-32}$	34%(91)
136	738	F	BaseplateJ phage tail	*Nostoc* sp PCC7524 (AFY49018.1)	$3.76e^{-33}$	41% (94)
139	576	F	Phage tail protein (tail_P2_I)	*Nostoc* sp PCC7524 (AFY49020.1)	$5.63e^{-40}$	59%(77)
140	1389	F	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49021.1)	$4.83e^{-57}$	37%(157)
142	1140	F	Tail collar protein	*Nostoc* sp PCC7524 (AFY49022.1)	$2.1e^{-52}$	50% (149)

Appendix G Predicted ORFs for Cyanophage N-1

ORFs	Length (bp)	Strand	Significant hit	Organism	e-value	%identity (shared aa)
2	1896	R	DNA polymerase B	*Cyanothece* sp. PCC 7424	7.0e-^{118}	41.5%(243)
7	1410	R	Putative DEAH-family helicase	Lactobacillus phage phiadh	2.61e^{-31}	28.1%(123)
23	600	R	Hypothetical protein	*Nostoc punctiforme* PCC 73102	1.10e^{-06}	21.4%(40)
30	633	R	C-5 cytosine-specific DNA methylase	*Nostoc punctiforme* PCC 73102	1.81e^{-05}	36.2%(34)
37	240	R	ASCH domain protein	*Clostridium* phage phiMMP02	4.06e^{-12}	42.3%(33)
40	588	R	Hypothetical protein	*Bacillus* phage SPBc2	7.44^{e-05}	22.6%(35)
43	2739	R	putative DNA primase	*Listeria welshimeri serovar* 6b str. SLCC5334	6.16e^{-04}	20.2% (68)
51	627	F	Thymidylate kinase	*Lyngbya* PCC8106	2.45e^{-23}	36%(74)
61	1047	F	DNA-cytosine methyltransferase	*Anabaena variabilis* ATCC 29413]	1.02e^{-56}	33.2(127)
70	579	R	Hypothetical protein	*Thermoanaerobacter italicus* Ab9	8.67e^{-03}	32.7%(33)
72	507	R	Hypothetical protein	*Calothrix* sp. PCC 7103	1.73e-^{47}	56%(94)
74	726	R	Hypothetical protein	*Calothrix* sp. PCC 7103	4.94e^{-06}	30.4%(78)
76	585	F	dCTP deaminase	*Synechococcus* sp. PCC 7335	1.45e^{-69}	65.5%(131)
81	411	F	endodeoxyribonuclease RusA	*Desulfotomaculum reducens* MI-1]	3.99e^{-01}	37.7%(29)
84	741	R	Thymidylate synthase complementing protein/FAD-dependent thymidylate synthase	*Chlorobaculum parvum* NCIB 8327	8.35e^{-33}	39.1%(86)
92	1359	F	phage terminase, large subunit	*Nostoc* sp PCC7524	2.32e^{-55}	31%(134)
95	1602	F	Hypothetical protein	*Nostoc* sp PCC7524 (AFY48995)	3.55e^{-18}	22%(106)
98	696	F	Outer membrane protein (OmpH-like)			
100	1089	F	putative major capsid protein	Cyanophage AN-15	1.94e^{-134}	72%(254)
103	330	F	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49001)	9.62e^{-06}	30%(30)
106	1521	F	Tail sheath protein	*Nostoc* sp PCC7524 (AFY49006)	3.74e^{-74}	39%(138)
108	498	F	T4-like virus tail tube			
112	2079	F	phage-related tail trans could also be phage tail tape measure protein	Vibrio phage VHML	1.22e^{-7}	34%(42)
116	2694	R	Lysozyme-like domain, rare lipoprotein A (RlpA)-like double psi beta barrel	*Nostoc* sp PCC7524 (AFY49014)	1.48e^{-37}	28 % (137)

121	849	R	Hypothetical protein	*Nostoc* sp PCC7524 (AFY49010)	$9.82e^{-10}$	27.9%(51)
125	918	F	Exonuclease RNase T and DNA polymerase III	*Thauera* sp MZIT	$4.14e^{-04}$	28.4%(38)
131	387	F	LuxR family regulatory protein, helix-turn-helif motif	*Streptomyces albus*	$3.15e^{-06}$	40.3% (25)
134	813	F	gp5 baseplate hub subunit and tail lysozyme	Acinetobacter phage Ac42	$4.24e^{-07}$	28.6%(30)
144	330	R	Lysosyme	*Nostoc* sp PCC7524 (AFY49017)	$2.51e^{-03}$	24%(32)
145	1167	F	Baseplate J phage tail protein	*Nostoc* sp PCC7524 (AFY49018)	$4.18\ e^{-68}$	43%(144)
146	576	F	Phage tail protein	*Nostoc* sp. PCC 7524 (AFY49020)	$2.32e^{+00}$	26.3% (45)
147	1395	F	Phage tail fiber protein	*Nostoc* sp. PCC 7524 (AFY49021)	$2.97e^{-64}$	37%(172)
148	1155	F	Tail collar protein	*Nostoc* sp. PCC 7524 (AFY49022)	$1.62e^{-47}$	33%(137)

Appendix H Reciprocal dot-plots of the Nostoc myoviruses A-1(x-axis) and N-1 (y-axis) based on whole genome nucleotide sequences.