Environmental and Genomic Insights into Marine Virus Populations and Communities

by

Jan Felix Finke

BSc, University of Applied Sciences Bonn-Rhein-Sieg, Germany, 2005
MSc, The University of Amsterdam, The Netherlands, 2009

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)

June 2017

Abstract

Marine viruses are the most abundant and genetically diverse biological entity in the oceans. Viruses infecting phytoplankton have a role in maintaining phytoplankton diversity, but also affect the cycling of carbon and nutrients through the microbial loop, which has substantial implications for the marine food chain and the planet's climate system. It has also become evident that viral replication is affected by environmental conditions. In turn, viruses appear to possess a repertoire of metabolic genes to compensate for environmental adversities.

However, it is not well understood how environmental variables affect viral replication in the environment or what the role of their genetic repertoire is in the selection to replicate. This thesis investigates the abundance and genetic diversity of viruses, the composition of viral communities and how the dynamics of viral replication is affected by in situ environmental conditions in four projects which are presented in Chapters 2, 3, 4 and 5.

Chapter 2 describes the influence of environmental variables on the variation in viral and host abundance, and how this dynamic changes among different environments.

Chapter 3 shows that phycodnaviruses infecting prasinophytes have a highly variable genetic repertoire with several metabolic genes of diverse origins. This genetic variability is reflected in their distribution in the environment, indicating selection on viruses.

Chapter 4 establishes an approach to study cyanomyovirus communities and their associated genetic repertoires in the environment. It shows that the distribution of cyanomyovirus ecotypes on temporal and spatial scales is a function of environmental variables.

Chapter 5 unveils a considerable mismatch between free cyanomyovirus communities, representing the seed bank, and replicating cyanomyoviruses in the cellular fraction. The emergence of replicating viruses out of the viral seed bank is highly variable and affected by environmental factors.

In conclusion, total viral abundance as well as the community composition of specific virus types show a relationship to environmental variables. The genetic repertoire of viruses appears to be an adaptation to selection pressure and specific viruses can

occupy environmental niches that are not only defined by the presence of susceptible hosts but also by a virus's ability to compensate for adversities.

Lay summary

Marine viruses are highly abundant and diverse, with millions of different viruses per millilitre of sea water. Viruses infect cells, replicate in them and eventually the cells lyse to release viral progenies. The infection of cells and their lysis has a role in maintaining cell diversity, but also affects the cycling of nutrients through the ecosystem, with substantial implications for the marine food chain and the planet's climate system.

However, it is not well understood how environmental conditions affect viral replication or what the role of the content of their genomes is. This thesis investigates the abundance and diversity of viruses in relation to environmental conditions.

Viral abundance as well as the community composition and the genomes of marine viruses show a relationship to various environmental conditions. The content of the genomes of viruses appears to be an adaptation to environmental niches, presumably to compensate for adversities.

Preface

In chapter 2 the database was supplemented with flow cytometry data generated by Christian Winter. The author, Jan F. Finke, conceived the project and performed the analysis under supervision by Curtis A. Suttle and with advice from Brian P.V. Hunt. Stilianos Louca advised on building an analytical script. Chapter 2 has been previously published in Finke, J F; Winget, D M; Chan, A M; Suttle, C A: Variation in the Genetic Repertoire of Viruses Infecting *Micromonas pusilla* Reflects Horizontal Gene Transfer and Links to Their Environmental Distribution (2017), *Viruses*, 9

Chapter 3 includes data generated by Danielle M. Winget and Amy M. Chan. Amy M. Chan isolated the *Micromonas pusilla* viruses MpV-PL1 and PpV-SP1, Danielle M. Winget built the sequencing libraries used in the project. Jan F. Finke conceived the project and performed the analysis under supervision by Curtis A. Suttle.

A modified version of chapter 3 has been published in Finke, J F; Winget, D M; Chan, A M; Suttle, C A: Variation in the Genetic Repertoire of Viruses Infecting *Micromonas pusilla* Reflects Horizontal Gene Transfer and Links to Their Environmental Distribution (2017), *Viruses*, 9

All other chapters represent independent work. The projects were conceived by Jan F. Finke and Curtis A. Suttle. Field sampling, experiments, lab work, data processing and analysis were performed by Jan F. Finke under the supervision of Curtis A. Suttle.

Table of contents

List of tables

List of figures

List of abbreviations

aa	Amino Acids
AIC	Akaike Information Criterion
AMG	Auxiliary Metabolic Gene
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
CA	Correspondence Analysis
CCA	Canonical Correspondence Analysis
cDNA	Complementary DNA
CDS	Coding Sequence
chl	Chlorophyll
CTD	Conductivity Temperature Depth (Sensor)
C3O	Canadian Three Oceans
DCM	Deep Chlorophyll Max
df	Degrees of freedom
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribonucleic Acid
DNApol	DNA polymerase
E	E-value
EDTA	Ethylenediaminetetraacetic Acid
EPA	Environmental Placement Algorithm
FCM	Flow Cytometer
FISH	Fluorescence In Situ Hybridisation
FSC	Forward Scatter

GC	GC content, guanine cytosine content (%)
GFL	Green Fluorescence
GOS	Global Ocean Survey
Gt	Giga Tons
HliP	High Light Inducible Protein
kb	Kilo Bases
kDa	Kilo Dalton
LDA	Linear Discriminate Analysis
m	Slope
Mb	Mega Bases
MCP	Major Capsid Protein
MLD	Mixed Layer Depth
n	Sample size
NCBI	National Centre for Biotechnology Information
NCLDV	Nucleocytoplasmic Large DNA Virus
Nif	Nitrogen Fixation (Regulon)
nr	Non redundant
nt	Nucleotide
OFL	Orange Fluorescence
ORF	Open Reading Frame
OTU	Operational Taxonomic Units
p	p-value
PAR	Photosynthetically Active Radiation

PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
PCR	Polymerase Chain Reaction
PebS	Phycoerythrin Synthase
PetE	Plastocyanine
PetF	Ferrodoxin
Pho	Phosphate Assimilation (Regulon)
PS	Photo System
Psb	Photoystem I protein
Psa	Photystem II protein
PtoX	Plastoquional Terminal Reductase
RFL	Red Fluorescence
RI	Rivers Inlet
RNA	Ribonucleic Acid
RT-PCR	Real Time Polymerase Chain Reaction, also Q-PCR
R^2	Determination coefficient
SAA	Saanich
SI	Saanich Inlet
SOG	Strait of Georgia
SSC	Side Scatter
TalC	Trans Aldolase
TE	Tris-EDTA
T_m	Melting temperature

Tris Tris hydroxymethyl Aminomethane

Glossary

amplicon

fragment of a sequence produced by amplification by polymerase chain reaction

annotate

in bioinformatics assign it a function if known, done by searching a database for homologues based on identity

anoxic

an environment fully depleted of oxygen; also see: hypoxic

auto-fluorescence

fluorescence of a particle or cell when excited by light without the use of fluorescent dye

autotroph

an organism that uses inorganic compounds for growth and light or inorganic compounds as an energy source

α-alpha diversity

describing the number of distinct organisms and their variability in relative abundances, product of richness and evenness

BLAST Basic Local Alignment Search Tool

bioinformatic tool to search databases for sequence homologues based on sequence similarity through pairwise alignment

bootstrapping

statistical method to measure confidence in a result by subsampling dataset by criteria, in phylogenetics to determine the confidence/ repeatability in the topology of a phylogenetic tree given in %

brackish

a mix of fresh and marine water with a PSU between 1-30 PSU

burst size

number of virions released from a cell at the end of the lytic cycle

β-beta diversity

describes the variability in alpha diversity over a set of samples

capsid, viral

a protein shell enclosing the genetic information of viruses

centroid

in bioinformatics the representative natural sequence from within a cluster; in contrast to an averaged, not natural consensus sequence

CA Correspondence Analysis

exploratory multivariate statistic for unimodally related data, often expanded by fitting of additional data to explore co-variations

CCA Canonical Correspondence Analysis

interpretive multivariate statistic to link unimodally related explanatory and response data, to identify patterns of cause and effect

CDS Coding Sequence

part of a gene that codes for a protein flanked by start and stop sequences

cluster

in bioinformatics a set of sequences having a similarity greater than a preset identity threshold

cyanobacteria, also cyanophyta

a phylum of prokaryotic, autotrophic organisms in the domain bacteria, they use light as their energy source in photosynthesis, possess characteristic red or blue-green (cyano) pigments to harvest a broader light spectrum

γ-gamma diversity

see: “gamma (γ) diversity”

DCM Deep Chlorophyll Maximum

the depth of highest chlorophyll concentration, usually where light intensity and nutrient accessibility are both sufficient for growth

dereplicate

in bioinformatics removing supernumerary, identical copies of a sequence leaving only one unique copy in a dataset

diversity

see: α-alpha, β-beta, γ-gamma diversity

E-value

statistical test applied for sequences of data, in bioinformatics the expectation value describes how likely it is that two sequences match by coincidence, does not correct for sequence length

ecotype

a distinct organism with a characteristic function or functionality and genome under specific environmental conditions, setting it apart from others

envelope, viral

a lipid bilayer protein structure surrounding a viral capsid in some viruses

EPA **Evolutionary Placement Algorithm**

algorithm to place short sequences in the context of a reference phylogeny based on longer sequences

estuary

semi enclosed body of water with a direct connection to the ocean at the mouth of a river in the tidal zone, transition zone between river and ocean

evenness

measurement of the similarity in relative abundances of distinct organisms in a sample

fitness

in ecology describing how adapted an organism is to a given environmental condition, often used in reference to other organisms

γ-gamma diversity

describes the overall (alpha) diversity over a set of samples

GC content

ratio of guanosine and cytodine nucleotides in relation to adenosine and thymidine given in %

genetic repertoire

set of genes of a biological entity, referring to genes with metabolic functions

genotype

a distinct organism with a characteristic genetic sequence

heterotroph

an organism that uses organic carbon as its carbon source and draws energy from other organisms

horizontal gene transfer

exchange of genetic material between organisms via transformation, transduction, conjugation or gene-transfer agents; also called lateral gene transfer

homologues

genes with sequence similarity with common ancestry and function

hypoxic

in seawater, an oxygen concentration below 1.5 ml l^{-1}

identity

in bioinformatics the similarity between two sequences expressed as the percentage of identical residues

Illumina

brand of so-called next-generation sequencing technologies, based on the principle of sequencing synthesis

Inlet

semi enclosed body of water with a direct connection to the ocean, which may or may not be brackish

integrating

in ecology and oceanography the equal mixing of samples over a defined range, for example depths or area

killing the winner

theory that viruses preferentially kill the dominant organisms of a population at their peak abundance, thus promoting eveness and diversity

lytic cycle

the consecutive steps of viral replication from infection of the host cell to release of virions

m (slope)

in a linear regression m describes the slope of the regression line of x and y variables

marker gene

gene used todetermine the phylogeny of organisms or viruses based on variations in the sequence

Maximum-Likelihood

In bioinformatics, a statistical method to find the most likely phylogeny among alternatives, considering evolutionary substitution models

Neighbor-Joining

in bioinformatics a method to build the simplest phylogenies from distance matrices

next-generation sequencing

new, high-throughput sequencing technologies capable of producing millions of reads at once

microphytoplankton

photosynthetic plankton cells between 2 and 5µm in diameter

mixed layer

in oceanography, the surface water layer that is well mixed by surface wind; constrained by the pycnocline, a zone of rapid density change as a function of salinity and temperature

myovirus

a member of the viral family *Myoviridae* that infects prokaryotes, has a dsDNA genome and characteristic long contractile tail. The family is in the order of *Caudovirales* along with the *Podoviridae* and *Siphoviridae*

NCLDVs Nucleocytoplasmic Large DNA Viruses

group of viruses with large, dsDNA genomes infecting eukaryotes and replicating in the nucleus and/ or cytoplasm, including six families of the proposed order *Megavirales*

ORF Open Reading Frame

string of nucleotides of predefined length between a start and a stop sequence (codon), possibly a gene

OTU Operational Taxonomic Unit

a sequence representing a defined organism or virus based on a preset sequence difference, precursor for a genotype

p-value

statistical test for the probability of the null hypothesis that a correlation between values is coincidence, the probability value describes the significance of a result

PAR Photosynthetically Active Radiation

Portion of the light spectrum that can be utilized for photosynthesis by known organisms, between 400 and 700nm wavelength

parse

sorting sequences to a set of reference sequences based on identity and a predefined identity level

PCA Principal Component Analysis

exploratory multivariate statistic for linearly related data

PCoA **Principal Coordinates Analysis**

exploratory multivariate statistic based on sample similarity

phage

specific term for a virus infecting prokaryotes; derived from bacteriophage

phenotype

an organism with characteristic feature(s)

photic zone

the zone of a water column that is well penetrated by light, depends on turbidity and the associated light attenuation and maximally extends to 150 to 200 meters, the bottom of the photic zone is marked by the compensation depth below which respiration exceeds primary production; usually approximated as the depth at which the light intensity dropped to 1 % of the surface irradiance

phycodnavirus

a member of the viral family *Phycodnaviridae* with large dsDNA genomes, one of the families that comprise the order of NCLDVs; all known phycodnaviruses infect phytoplankton

prasinovirus

genus of viruses in the family *Phycodnaviridae*, members of which infect prasinophytes

prasinophyte

member of the class *Prasinophyceae*, a class of paraphyletic, eukaryotic phytoplankton in the division chlorophyta

primary production

the anabolic process to generate energy-rich carbon molecules from inorganic carbon driven by physical or chemical energy

PSU Practical Salinity Units

salinity measurement based on conductivity of water, under standardized conditions approximately equal to ppt (parts per thousand) g/kg

R^2 coefficient of determination

in a linear regression R^2 describes the variation in the dependent variable that can be explained by the independent variable, the adjusted R^2 considers the number of independent variables upon which the model is built on

rarefy

in ecology, normalizing samples to the same total number of sequences by random subsampling

read

in bioinformatics, a raw array of nucleotide code from a sequencing reaction of an individual DNA molecule

replication cycle

the overall process of viral replication including the lytic cycle but also viral adsorption to and infection of the host cell

richness

the absolute number of distinct organisms in a sample, regardless of their relative abundance

Sanger sequencing

established reference sequencing method producing individual reads of high confidence, developed by Frederick Sanger

seed-bank

theory that the collective of viral genotypes present is stable among samples and that only/ mainly their relative abundance varies

stoichiometry

the ratio of nutrients to each other in cells or in water

stratified

describes a water column that has a well defined density gradient (pycnocline) with defined layer of lower density water on top, usually caused by high temperature or fresh water inflow

tail

a protein structure in viruses, phages that from the capsid

454 sequencing

an early next-generation sequencing technology by Roche, producing few but long reads; discontinued

Acknowledgements

The past years have been quite an endeavour and I would not have been able to complete this PhD project without the support from many people. Here I would like to express my gratitude to a few people who helped me over the years.

First of all, I would like to thank my supervisor Curtis Suttle for providing me with opportunities, his support, advice and guidance.

Furthermore, I would like to thank my supervisory committee, Philippe D. Tortell and Thomas J. Beatty for their thoughts, expertise, guidance and advice in pursuing this degree.

This thesis would not have been possible without the help of past and present members, and visiting scientists of the Suttle lab. Their insight, criticism, suggestions and expertise were invaluable. I would especially like to thank Amy Chan, Renat Adelshin, Anwar Al-Qattan, Christina Charlesworth, Caroline Chenard, Natacha Chenevoy, Anna Cho, Cheryl Chow, Jessie Clasen, Ricardo Cruz, Christoph Deeg, Matthias Fisher, Jessica Labonte, Gideon Mordecai, Jerome Payet, Jeff Strohm, Emma Shelford, Alvin Tian, Danielle Winget, Christian Winter, Jennifer Wirth, Marli Vlok, Kevin Zhong.

I would also like to thank Alyse Hawley, Brian Hunt, Stilianos Louca, Maite Maldonado, Chris Payne, Lora Pakhomova, Evgeny Pakhomov, Ania Posacka, Monica Torres-Beltran and many more for their help and advice.

The lab and this project have been supported by the discovery grant and equipment grant of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Foundation for Innovation and BCKDF. J.F. Finke received support from the Canadian Institute for Advanced Research (CIFAR) to attend conferences.

Many friends have supported me over the years. I want to express my gratitude to Adam, Allison, Anna, Anthony, Bonsai, Chris, Courtney, Emilie, Hildur, Ian, Joanna, Jonathan, Mac, Maria, Mathieu, Marie-Pier, Ned, Nora, Osmar, Paul, Sam, Ulrike and many more. Thank you for your support and advice, lending me an ear and lifting my spirits, distractions and adventures.

Last but not least, I want to thank my parents and my family for their continued support and encouragement not only throughout this degree, but my entire life. I would not be who I am without them. What a ride!

Chapter 1: Introduction

1.1 Marine viruses

Marine viruses are the smallest, most abundant and most diverse biological entities in the world's ocean with an estimated total number of 10^{30} particles (Suttle, 2007). They play crucial roles in marine ecosystems, promoting cell diversity and channeling nutrients.

Viral abundance across aquatic ecosystems ranges from 10^5 ml^{-1} to as high as 10^9 ml^{-1} (Proctor and Fuhrman, 1990; Paul *et al.*, 1993; Brum *et al.*, 2005; Ortmann and Suttle, 2005). Several studies have established that viral abundance is about an order of magnitude higher than bacterial abundance, but this virus to bacteria ratio (VBR) varies substantially among host-virus systems and environments (Fuhrman and Suttle, 1993; Wommack and Colwell, 2000; Knowles *et al.*, 2016; Wigington *et al.*, 2016). A meta-analysis of 25 studies by Wigington et al. (2016) confirmed that VBRs range widely, with 95% of the values falling between about 3.9 and 74.4 at depths less than 100 m, and between about 1.4 and 157.1 in deeper water. Median values for depths less-than and greater-than 100 m are about 10.5 and 16.0, respectively. They also showed the limitation of fixed 10:1 ratio models for estimating VBR and established the use of non-linear power functions to estimate viral abundance from bacterial abundance data. A different study by Knowles et al. (2016) documented linear correlations between viral abundance and bacterial abundance for various habitats and that there is a relative decrease in viral abundance with increasing bacterial abundance. Yet another study found that the relationships between bacterial and viral abundances differ significantly among samples from lakes, shallow and deep waters from the Pacific Ocean, and the Arctic Ocean (Clasen *et al.*, 2008). These papers demonstrate differences in the relationship between bacterial and viral abundances among locations and environments, implying that the environmental conditions influence these relationships. In turn, models show that a high viral abundance increases the contact rate with microbial hosts, so changes in the viral abundance and VBR could increase the chance of infection (Murray and Jackson, 1992; Mann, 2003).

As is the case of all viruses, those in the marine environment vary substantially in size, structure and genome content. The capsids of most marine viral particles range in size from 20 nm to 200 nm, although viruses have been isolated with capsid diameters up

to 440 nm (Arslan *et al.*, 2011) have been found in other environments. Most marine viruses appear to have icosahedral capsids, and a significant portion have tails (Brum, Schenck and Sullivan, 2013). Tailed viruses are a characteristic of many phages, which are viruses infecting bacteria. Genomes of viruses can be single-stranded or double-stranded RNA or DNA and range in size from <2 kb (Labonte and Suttle, 2013) to over 1.2 Mb in length (Raoult *et al.*, 2004; Arslan *et al.*, 2011). While DNA viruses are more extensively studied and commonly infect bacteria and phytoplankton, recent data suggest that RNA viruses which often infect protists are also an important part of the virosphere (Steward *et al.*, 2013). Every water sample harbors a diverse mix of viral morphotypes and genotypes. Even when focusing on a single virus group, a seawater sample can carry thousands of viral genotypes (Breitbart *et al.*, 2002; Filée *et al.*, 2005; Comeau and Krisch, 2008; Butina *et al.*, 2010; Goldsmith, Parsons and Beyene, 2015). Despite this high genetic diversity, some viral genotypes are widely distributed, and it is largely the relative abundance of viral genotypes that distinguishes communities (Angly *et al.*, 2006).

The elemental composition of viral particles is inherently different compared to that of their hosts (Jover *et al.*, 2014). Considering the wide range in size of viruses and the variation in nucleic-acid type, there also is considerable variation in the elemental composition among viral particles. Even within double-stranded DNA viruses there is considerable variation in their C:N:P stoichiometry, but it is estimated to be around 17:6:1 (Jover *et al.*, 2014). This is in contrast to the C:N:P stoichiometry of phytoplankton and heterotrophic marine plankton of about 106:16:1 and 69:16:1 (Redfield, Ketchum and Richards, 1963; Falkowski, 2000; Klausmeier *et al.*, 2004) and highlights the relatively high nitrogen and phosphorus requirement for viral replication. Consequently, it is estimated that up to 87% of cellular phosphorus can be assimilated into viral particles during replication (Jover *et al.*, 2014).

Classically, viruses are solely seen as pathogens of organisms, but their role is more complex and far reaching. By infecting and lysing cells, viruses are a major evolutionary force for plankton, altering microbial diversity directly by horizontal gene transfer and indirectly by selection (Sullivan, Waterbury and Chisholm, 2003; Brussaard *et al.*, 2008). Specifically infecting dominant, blooming cells, a concept termed “killing the

winner", viruses help maintain diversity and apply selection pressure (Thingstad and Lignell, 1997; Thingstad, 2000).

The other crucial role of viruses is to modulate biogeochemical processes in the oceans through the "viral shunt" (Wilhelm and Suttle, 1999). Viral lysis channels carbon and nutrients directly to pools of particulate and dissolved organic matter rather than through the food web, thus catalyzing nutrient regeneration (Figure 1.1). Viral lysis of phytoplankton, the most significant primary producers in the oceans, directly affects nutrient cycles and trophic transfer efficiency. Viral lysis is estimated to kill about 20-40 % of marine planktonic biomass daily (Fuhrman and Suttle, 1993; Suttle, 1994, 2005; Wilhelm and Suttle, 1999). Considering that about half of the annual global net primary production, 48.5 Gt of carbon, occurs in the oceans (Field *et al.*, 1998), it is apparent that viral lysis is not only crucial in marine systems but also for global carbon cycling. The impact of viral lysis on the flow of nutrients and trace elements is similar to that on carbon. Although viruses retain proportionally high amounts of nitrogen and phosphorus, lysate is still a rich source of macro and micro nutrients (Gobler *et al.*, 1997) and has been shown to supply nitrogen (Weinbauer *et al.*, 2011; Shelford *et al.*, 2012) and iron (Poorvin *et al.*, 2004, 2011) to primary producers. Clearly, marine viruses play a crucial role in marine ecosystems and biogeochemical cycling.

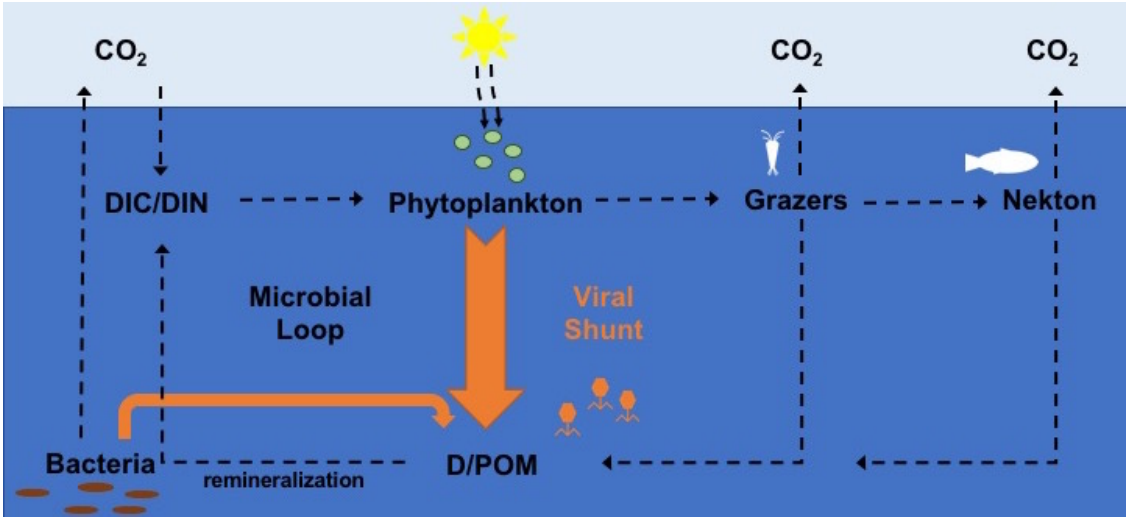

Figure 1.1: The viral shunt.
Simplified scheme showing the role of viruses and the viral shunt (orange) in the microbial loop, highlighting its impact on nutrient cycling through bacteria, phytoplankton. Dissolved inorganic carbon and nutrients (DIC/DIN) are directly funneled to the pool of dissolved and particulate organic matter (D/POM) and bypass higher trophic levels.

1.2 Phytoplankton and their viruses

Marine phytoplankton account for about 50% of the global carbon fixation, about half of which is by prokaryotic cyanobacteria and the other half by eukaryotic phytoplankton (Field *et al.*, 1998). Both prokaryotic and eukaryotic phytoplankton are infected by viruses with large dsDNA genomes with variable gene content.

Globally distributed and abundant marine cyanobacteria, primarily members of the genera *Synechococcus* and *Prochlorococcus* (Liu *et al.*, 1998), perform about 25% of the world's primary production (Li, William, 1994; Liu, Nolla and Campbell, 1997; Partensky, Hess and Vaulot, 1999; Weigele *et al.*, 2007). These cyanobacteria are constantly in the process of infection and lysis by viruses (cyanophages), which is driving their diversity and affecting biogeochemical cycles. Cyanophages infect dominant cyanobacteria in a community, causing their demise and thus maintain community diversity (Mühling *et al.*, 2005; DeLong *et al.*, 2006). Viral lysis also releases organic nutrients directly into the microbial loop through what is known as the viral shunt (Wilhelm and Suttle, 1999; Weitz and Wilhelm, 2012).

Cyanophages have circular, dsDNA genomes and are assigned to the *Caudovirales* (Weinbauer and Rassoulzadegan, 2003), an order of tailed phages with the families *Myoviridae*, *Podoviridae* and *Siphoviridae,* which in turn are split up into various sub-families (Marston and Sallee, 2003; Clokie and Mann, 2006; Lavigne *et al.*, 2008, 2009). This thesis focuses on cyanophages in the family *Myoviridae*, T4-like viruses referred to as cyanomyoviruses. These lytic viruses have characteristic long contractile tails and genomes ranging from 150 kb to >200 kb in size (Lu, Chen and Hodson, 2001; Sullivan, Waterbury and Chisholm, 2003; Lindell *et al.*, 2004; Clokie, Millard and Mann, 2010). They have a comparably large host range (Lu, Chen and Hodson, 2001; McDaniel, delaRosa and Paul, 2006; Hanson, Marston and Martiny, 2016) that can span genera and cyanophages usually carry a number of host-derived auxiliary metabolic genes (AMGs) (Breitbart *et al.*, 2007; Crummett *et al.*, 2016).

Among the eukaryote phytoplankton prasinophytes are widely distributed and arguably constitute the second most abundant group of phytoplankton after cyanobacteria. These picoplanktonic (<2 µm diameter) algae have a disproportional importance in primary production (Worden, Nolan and Palenik, 2004; Marin and

Melkonian, 2010), are prominent in coastal and oceanic communities, and comprise the genera *Micromonas*, *Ostreococcus* and *Bathyococcus*.

Prasinophytes are infected by large icosahedral double-stranded DNA viruses in the genus *Prasinovirus* (Van Etten *et al.*, 2002). Prasinoviruses are significant agents of mortality for prasinophytes that can influence nutrient fluxes through the viral shunt (Wilhelm and Suttle, 1999), maintain host diversity by lysis, and promote horizontal gene transfer (Moreau *et al.*, 2010; Filée, 2015). Prasinoviruses are related to viruses in the genus *Chlorovirus* that infect members of the genus *Chlorella* (Dunigan, Fitzgerald and Van Etten, 2006); both genera fall within the family *Phycodnaviridae* (ICTV 8th report), which belongs within the Nucleocytoplasmic Large DNA Viruses (NCLDVs) (Koonin and Yutin, 2010). Phycodnaviruses infect a wide range of eukaryotic microalgae, and have genomes ranging in length from about 100 to 400 kb (Koonin and Yutin, 2010). Genomes of the known prasionivruses infecting *Micromonas* spp., *Ostreococcus* spp. and *Bathyococcus* spp. are similar in length, ranging from about 180 to 200 kb (Derelle *et al.*, 2008, 2015; Moreau *et al.*, 2010), and to a certain degree show similarity in genome structure and content (Weynberg *et al.*, 2009; Moreau *et al.*, 2010). The function of most prasinovirus genes is still unknown, but subgroups of prasinoviruses share a common set of core genes that are essential for viral replication and viral structure (Santini *et al.*, 2013). Core genes include DNA polymerase, DNA topoisomerase and seven to eight capsid genes (Derelle *et al.*, 2008, 2015; Weynberg *et al.*, 2009). Similar to cyanophages, prasinoviruses can carry metabolic genes that are presumably host-derived. Prasinoviruses usually demonstrate strict host specificity, at least at the genus level of their hosts, but some viral isolates can broadly infect strains of *Micromonas pusilla* (Derelle *et al.*, 2015; Martínez *et al.*, 2015).

1.3 The genetic repertoire of phytoplankton viruses

Viruses can be affected by environmental stressors that influence their physical integrity, infectivity and replication, but they also encode proteins that may help overcome these challenges. These genes are part of the pan-genome that is distributed across the viral community and often appear to be host-derived genes, often referred to as auxiliary metabolic genes (AMGs) in cyanophages (Breitbart *et al.*, 2007). Hence, specific viruses

of the assemblage potentially carry genes that are advantageous for replication under certain environmental conditions, and allow them to be less reliant on host-encoded proteins. So far AMGs have primarily been identified in cyanophages, but similar genes are also found in prasinoviruses.

In cyanophages, AMGs are wide spread, can constitute a large part of the genome, influence the host's metabolism during replication, are expressed in host cells (Lindell *et al.*, 2005, 2007; Sullivan *et al.*, 2006; Thompson *et al.*, 2011) and potentially improve phage fitness (Dammeyer *et al.*, 2008; Hellweger, 2009). Moreover, their distribution in metagenomic sequences and isolates has been linked to environmental variables such as phosphate concentration or light intensity (Williamson *et al.*, 2008; Crummett *et al.*, 2016). Cyanophage AMGs include those encoding proteins associated with photosystem I (*psbA/D*) and photosystem II (*psa A-F*) (Mann *et al.*, 2003; Sullivan *et al.*, 2006; Sharon *et al.*, 2007, 2009), as well as phycoerythrobilin synthesis (*pebS*), a fusion gene of photosytem I (*psaJF*), and a range of other genes involved in photosynthetic processes (*petE*, *petF* and *ptoX*) (Dammeyer *et al.*, 2008; Millard *et al.*, 2009; Sharon *et al.*, 2009). Other potentially beneficial genes are associated with light protection (*hli*), the pentose phosphate pathway (*talC*), nitrogen fixation (*nifU*) and phosphate assimilation (*phoH, pstS*) (Lindell *et al.*, 2004, 2005, 2007; Millard *et al.*, 2004; Williamson *et al.*, 2008; Tetu *et al.*, 2009; Zeng and Chisholm, 2012).

Genes similar to AMGs are less studied in prasinoviruses, but have been found. Several prasinoviruses encode genes for functional K^+ channels (Siotto *et al.*, 2014), and , *Ostreococcus* viruses possess gene homologues associated with sugar metabolism, glycosyltransferases (*gtfS*), nucleotide modification, ribonucleotide reductase (*rrnR*), amino acid synthesis, acetolacetate synthase (*alS*), phosphate starvation (phoH) and many more (Weynberg *et al.*, 2009). Moreau et al. (2010) have also shown that some of these genes are widely shared among viruses infecting hosts from the genera *Bathyococcus*, *Ostreococcus* and *Micromonas*.

It has been suggested that viral PsbA proteins could account for around 10% of the global primary production, highlighting the potential impact of AMGs (Rohwer and Thurber, 2009). If AMGs are expressed during replication they have the potential to increase the fitness of a subset of the viral community under certain conditions.

Consequently, this should be reflected in the community composition and genetic content of viruses in relation to environmental conditions. However, the adaptation of the genetic content and community composition of viruses to e.g. environmental factors in situ remains to be demonstrated.

1.4 Environmental variables affecting virus-host interactions

Marine phytoplankton and their corresponding viruses are affected by a variety of environmental factors, including temperature, salinity, light and nutrient availability.

Temperature, salinity and light can be important factors influencing virus-host interactions. For example, high temperatures can inactivate viruses (Baudoux and Brussaard, 2005; Hardies *et al.*, 2013), or cause a switch from a lysogenic to a lytic life cycle (Williamson and Paul, 2006). Similarly, high salinity can hinder virus adsorption across a range of environments (Kukkaro and Bamford, 2009). Light can also be important in a number of ways, besides being essential for phytoplankton growth. For example, adsorption of viruses to *Synechococcus* cells is light dependent (Jia *et al.*, 2010). The duration of the lytic cycle and viral production in phytoplankton have been shown to be affected under low light levels (Juneau *et al.*, 2003; Brown, Campbell and Lawrence, 2007; Baudoux and Brussaard, 2008). Additionally, UV radiation causes dose dependent viral decay of viruses infecting bacteria, cyanobacteria and eukaryotic phytoplankton (Cottrell and Suttle, 1995a; Noble and Fuhrman, 1997; Garza and Suttle, 1998). Furthermore, decay rates depend on the viral GC% content (Kellogg and Paul, 2002), but damage can also be repaired by photoreactivation (Wilhelm, Weinbauer, Suttle and Jeffrey, 1998; Wilhelm, Weinbauer, Suttle, Ralph, *et al.*, 1998).

Nutrients influence the infection process, as would be expected given the high relative amount of phosphorus and nitrogen required by viruses compared to cells (Jover *et al.*, 2014). This is evident from the calculation that up to 87% of cellular phosphorus can be assimilated into viral particles during replication (Jover *et al.*, 2014). Phosphate has been shown to influence viral production by eukaryote phytoplankton, with less production under phosphate depletion compared to phosphate sufficiency (Bratbak *et al.*, 1998; Jacquet *et al.*, 2002; Maat *et al.*, 2014). The same has been shown for nitrate, but the effect was less pronounced than for phosphate (Bratbak *et al.*, 1998; Jacquet *et al.*,

2002). As well, nutrient availability can influence whether temperate viruses enter a lysogenic or lytic cycle (Wilson, Carr and Mann, 1996).

Altogether viral replication is affected by a myriad of environmental variables that change across timescales of hours to days, including weather patterns and water-column mixing by storms, which are similar to those influencing microbial community dynamics (Moore *et al.*, 2013). Moreover, rapid changes in climate are predicted to alter stratification, nutrient accessibility and light exposure for phytoplankton and their viruses (Sarmiento *et al.*, 1998, 2004; Danovaro *et al.*, 2011). Coastal habitats in particular, are expected to have an intensification of stratification and changes in vertical nutrient fluxes (Keeling, Körtzinger and Gruber, 2010; Capotondi *et al.*, 2012; Hordoir and Meier, 2012). Additionally, the coastal oceans are predicted to see shifts in the nitrogen-phosphate stoichiometry towards nitrogen limitation due to different anthropogenic inputs (Seitzinger *et al.*, 2010; Moore *et al.*, 2013). In light of these environmental changes, the response of virus communities and virus-host systems to them needs to be studied to better understand the impact in microbial ecosystems.

1.5 Assessing viral diversity

A first step to understand the response of viral communities to environmental conditions is to explore the differences in viral abundance and community composition across a range of environments. Studying viral community composition in the environment requires a meaningful genetic marker. Viruses lack a universal marker gene equivalent to the small ribosomal subunit RNA gene found in prokaryotes (16S rDNA) and eukaryotes (18S rDNA. So different marker genes are used to assess viral diversity and community composition, each having advantages and disadvantages. A key feature of a marker gene is its ubiquity in the genomes of the virus group that is being investigated; this restricts the possibilities to a few core genes. Typical marker genes used to assess viral diversity are those encoding DNA polymerase and capsid proteins.

DNA polymerase is essential for any DNA virus and hence a good candidate as a marker gene. In cyanomyoviruses, DNA polymerase (gp43) has been used as a marker gene, and although it is not cyanophage specific, a primer set developed based on cyanomyovirus isolates produces amplicons that are biased towards cyanophages

(Marston and Amrich, 2009). Extensive studies that used gp43 sequences to investigate cyanophage isolates collected from a range of locations and times revealed great diversity with hundreds of genotypes, that varied seasonally and spatially (Marston *et al.*, 2013). Marston and Martiny (2016) could furthermore separate cyanophage isolates of characteristic gene content and distribution into ecotypes. For prasinoviruses, DNA polymerase sequences have also been used to describe their diversity and distribution in the environment (Chen and Suttle, 1996; Chen, Suttle and Short, 1996; Short and Suttle, 2002, 2003). Additionally, DNA polymerase has been used to assess prasinovirus diversity in freshwater environments (Short and Short, 2008; Clasen and Suttle, 2009; Gimenes *et al.*, 2012).

Other frequently used markers are genes encoding capsid proteins. The capsid vertex portal protein, gp20, has been extensively used to describe cyanomyovirus communities seasonally and spatially (Wilson, Carr and Mann, 1996; Frederickson, Short and Suttle, 2003; Wang and Chen, 2004; Mühling *et al.*, 2005; Sandaa and Larsen, 2006). In some cases, changes in community composition were related to different physical conditions and host communities (Frederickson, Short and Suttle, 2003; Wang and Chen, 2004). However, while one study (McDaniel, delaRosa and Paul, 2006) could only amplify *gp20* sequences from about 60% of isolates, another study (Short and Suttle, 2005) found that identical *gp20* sequences were dispersed across a wide range of contrasting environments. This casts a doubt over the reliability and specificity of *gp20* as a marker gene. Similarly, sequences for the major capsid protein, gp23, have been widely used to describe diversity in myovirus communities (Tetart *et al.*, 2001; Filée *et al.*, 2005; Comeau and Krisch, 2008; Butina *et al.*, 2010), including high-frequency and in-depth sampling (Chow and Fuhrman, 2012; Needham *et al.*, 2013). The drawback of using both gp20 and gp23 to examine cyanomyovirus communities is the lack of specificity of the primers. For prasinoviruses, major capsid protein (MCP) sequences have been used to assess communities in some studies (Larsen *et al.*, 2008; Rowe *et al.*, 2011). Clerissi et al. (2014) have used MCP, and also DNA polymerase, to compare prasinoviruses to chloroviruses based on full length and partial gene sequence (Clerissi, Grimsley, Ogata, *et al.*, 2014). Similarly, Zhong and Jacquet (2014) have compared MCP to DNA polymerase in describing prasinoviruses. These two studies disagree on the congruency of DNA

polymerase and MCP in assessing communities, leaving uncertainty as to how suitable they are.

Given the limitations of using DNA polymerase and MCP sequences to assess viral diversity, some metabolic genes that are distributed across different types of viruses have been tried. For example, sequences associated with proteins involved in photosynthesis (PsbA) and phosphorus metabolism (PhoH) have been used, but it is questionable how well these sequences reflect diversity across broad groups of viruses (Mann *et al.*, 2003; Chénard and Suttle, 2008; Goldsmith *et al.*, 2011; Goldsmith, Parsons and Beyene, 2015). Perhaps the best approach is to build multi-gene phylogenies based on selected core genes; however, this is only possible with isolates (Derelle *et al.*, 2015). Similarly, phylogenetic inferences based on the presence and absence of genes in whole genomes provides a strong approach for comparing viral genomes (Snel, Bork and Huynen, 1999; Yutin, Wolf and Koonin, 2014; Legendre *et al.*, 2015), but it is again only applicable for studying isolates and not environmental virus communities. There still is no satisfying approach to study natural viral community compositions in relation to viral gene content and environmental variables.

1.6 Research problem

As summarized above, marine viruses and their hosts face a multitude of variables that influence the likelihood, progress and outcome of viral infections. However, it is not clear how the environment is related to the abundance and community composition of viruses. This thesis aims to elucidate the relationship between environmental variables and the abundance, gene content and community composition of important marine viruses.

1.7 Scientific objectives

The research problem stated above was addressed by examining differences in viral abundance, gene content and community composition in a range of samples with different underlying environmental conditions. The project is divided into the following four scientific objectives:

1) To develop a statistical model of viral abundances as a function of environmental parameters.

2) To investigate and compare the genetic repertoire of prasinoviruses and their distribution in the environment.
3) To characterize the composition of cyanomyovirus communities in relation to environmental variables
4) To assess the actively replicating cyanomyoviruses in comparison to the associated free virus community under differing environmental conditions.

1.8 Approach and methodological considerations

The four objectives were approached in four sub-projects which are presented in Chapters 2, 3, 4 and 5 of this thesis.

Water samples were collected at a range of locations and times to examine viral abundance and diversity. The samples were processed according to the different scientific objectives. Sampling locations included Saanich Inlet, the Juan de Fuca Strait, the Strait of Georgia, Queen Charlotte Sound, Rivers Inlet, in coastal British Columbia, and the North Pacific, the Arctic Ocean and the North Atlantic. Saanich Inlet on Vancouver Island B.C. is characterized by having seasonally occurring deep water that is low in oxygen. Rivers Inlet is comparable in size to Saanich Inlet, but does not have seasonally occurring low oxygen in the deep water. Juan de Fuca Strait, Strait of Georgia and Queen Charlotte Sound represent coastal waters between the mainland and Vancouver Island, and feature high flow rates, and are well mixed, while other locations represent sheltered and stratified inlets. Samples from the North Pacific, Arctic Ocean and North Atlantic are representative of arctic and sub-arctic conditions.

Each water sample represents a specific time and location and should not be viewed as representative of a specific location over time. Depending on the study, samples were taken at specific depths or integrated over depths; integrated samples were collected over the surface mixed layer. This assumes that a well mixed water mass with stable microbiologically relevant environmental variables does promote stable communities (Fuhrman, Cram and Needham, 2015). Spatial samples were selected to represent a range of environmental conditions and were separated by distance or geographical features. In open waters, samples were taken several kilometers apart to overcome the “patch size” in homogenous water masses (Hewson *et al.*, 2006). Temporal

samples by month or season were used to capture seasonal changes in weather, stratification and nutrient concentrations (Moore *et al.*, 2013; Fuhrman, Cram and Needham, 2015). Sample volumes varied from 20 to 200 liters; these volumes were chosen to minimize the randomizing effects of sample handling, avoid coincidently sampling uncharacteristic micro-niches and to collect sufficient genetic material. Furthermore, care was taken to process samples swiftly to minimize the effects of changing conditions during storage.

Depending on the study, samples were analyzed for viral abundance, viral community composition and associated environmental data such as temperature, salinity, oxygen, nitrate+nitrite, phosphate, silicate, oxygen, chlorophyll fluorescence and cell and viral abundances. Abundances of dsDNA viruses were measured by flow cytometry, which is an established high throughput method (Brussaard, 2004). Community composition was assessed by amplicon sequencing of marker genes that were selected carefully based on the viruses targeted in this study and the intended research question (Adriaenssens and Cowan, 2014), as detailed in the corresponding chapters. Amplicon sequencing using high-throughput technology has several advantages over metagenomic sequencing or studies based on viral isolation. By using appropriate primers, amplicon sequencing delivers a standardized, comparable sequence product with deep coverage, yielding high resolution data. However, one caveat of amplicon sequencing is the potential polymerase chain reaction (PCR) bias of the primers, which could favor or hinder certain products. In order to mitigate this effect, amplicons were drawn from pools of three separate, large volume PCRs with excess reagent concentrations and high starting template concentrations (Acinas *et al.*, 2005). To differentiate between genotypes of viruses, sequence identity levels were set for clustering of sequence reads into operational taxonomic units (OTUs). These identity levels were set depending on the marker gene and specific project. A microbiological concept is that a distinct “ecological species” is defined by its function (Fuhrman, Cram and Needham, 2015). Similarly, Shapiro and Polz (2014) propose to assess microorganisms based on genetic similarity, driven by gene flow and selection of phenotypes with a distinct ecological function. These concepts also apply to viruses. In the studies described here, genotypes were shown to

reflect the genetic similarity of viruses, and should provide an indication of their genetic potential.

Chapter 2 is focused on viral abundances determined by flow cytometry. The dataset is composed of samples from three different projects, covering a range of environments, seasons and depths. Thus the data reflect differences across a wide range of environmental conditions among samples, which help to reveal trends in the data. Physical and nutrient data supplement the data on virus and cell abundances. The use of flow cytometry enabled large amounts of data to be collected with high accuracy, characterized by low variability among replicates. This large sample number led to high statistical strength of the relationships between viral and cell abundances and environmental variables.

Chapter 3 examines the genetic repertoire of prasinoviruses, its variation and distribution in the environment. The genomes of two fully sequenced prasinoviruses that infect *Micromonas pusilla* were assembled and annotated, improving the annotations of genomes for these viruses. This made it possible to compare the variation in gene content among these and other prasinoviruses. Most ORFs could not be assigned a functional annotation as the annotation was strongly constrained by the sequence data in the database. However, it was still possible to compare the overall gene content, even if some ORFs were annotated as “hypothetical proteins”. Another crucial step in comparing the gene content among the viruses was setting an appropriate identity threshold for clustering, and the subsequent analysis. In this study, the threshold was set to 50% amino-acid identity, which is supported by the observation that there is over 50% amino-acid identity among *Ostreococcus* virus core genes (Derelle *et al.*, 2015). Nevertheless, this identity threshold was a compromise and may not fully reflect all shared genes. Also, the identity threshold used for environmental amplicons to compare natural communities was a compromise between missing diversity when the percent identity is set too low and inflating diversity when it is set too high. The selected marker gene and identity thresholds were chosen to reflect distance and similarity in the gene content of prasinoviruses and exploring its variation in the environment.

Chapter 4 examines temporal and spatial variation in the community composition of cyanomyoviruses. Temporal samples were taken in Saanich Inlet while spatial samples

were taken during several cruises in the Strait of Georgia and adjacent waters. These samples were used to build a database representing different temperatures, salinities, mixing regimes and trophic states. Similarity of reference viruses in genetic content was assessed by clustering at 50% amino-acid identity and related to the similarity of the corresponding DNA polymerase marker gene (*gp43*) sequences. The insight into this relation was then used to explore environmental cyanomyovirus communities. Cyanomyovirus community compositions were assessed by amplicon sequencing of *gp43*, which provided accurate data with high sample throughput and sequencing depth. Again, clustering was a crucial step in the amplicon read analysis. The amplicon identity threshold was chosen to reflect the relative distance and similarity in gene content of the viruses and used to compare the community compositions.

Chapter 5 compares the community composition of free cyanomyoviruses and cyanomyoviruses in the cellular fraction. The idea was that the viral communities derived from the cellular fraction represent replicating viruses. Due to the diel changes in growth and rapid responses to environmental changes (Jacquet *et al.*, 2001) one can anticipate that the differences between the free and cellular viral communities can be substantial. Cellular samples were collected by membrane filtration, which could also capture viruses attached to cell surfaces and detrital particles. However, this is expected to be small relative to the number of viral progeny in the cellular fraction. Moreover, cyanomyoviruses will typically only attach to cells they can infect (unpublished observation C. A. Suttle). Hence the communities from the cellular fraction were assumed to predominantly represent actively replicating viruses as concluded before (DeLong *et al.*, 2006). To characterize the actively replicating viral communities in the cellular fraction these samples were processed and analyzed identically to free virus samples.

1.9 Significance and rationale

Over the past 25+ years, findings in two fields of research that are associated with oceanography have made great advances. One is that marine viruses are very abundant and important components of marine ecosystems (Suttle, 2007). The other is the ever stronger trend in climate change and its impact on the oceans, as described in the 5th IPCC report (Cubasch *et al.*, 2013). These two seemingly unconnected fields of research

turn out to strongly overlap in the effect of environmental variables on viral replication and phytoplankton (Danovaro *et al.*, 2011). However, little is known how environmental change affects marine virus ecology.

This thesis investigated the correlation between environmental variables and viral abundance, community composition, and the genetic adaptation involved. Climate change affects many things including the temperature and salinity of the oceans. These changes affect water column mixing, nutrient availability to plankton, and the light exposure of phytoplankton. With phytoplankton being the primary producers in marine food webs and an essential part of the global carbon cycle their response to environmental change and viral infection is of key interest. The four projects reveal trends and correlations between viral abundance, their gene content, community composition and environmental variables. While these projects presumably just scratch the surface, they provide a detailed look at the variability of the genetic content of viruses, and how viral abundance and community composition vary across environments.

Chapter 2: Environmental variables affect the virus host relationship across marine environments

2.1 Summary

Marine viruses are highly abundant and generally outnumber their hosts by an order of magnitude. However, the relationship between viral abundance and host abundance does fluctuate and is affected by environmental variables. This project aims to describe how environmental variables influence the variation in viral and host abundance and how this dynamic changes among different environments.

Using flow cytometry, the relationship of viral abundance to bacterial abundance, nutrient concentrations and oceanographic variables was tested using multivariate models. Samples were taken over spatial and temporal ranges, in estuaries and inlets along the coast of British Columbia (Canada), the Pacific, the Arctic and the North Atlantic of various depths and seasons. Samples were classified into arctic, inlet, and hypoxic environments and generalized linear models of varying complexity were tested and compared for best fit with the Akaike Information Criterion (AIC).

Viruses and bacteria, as the numerically dominant host of viruses, showed significant relationships of varying strength for the arctic and inlet environments, but were insignificant in the hypoxic samples. Multivariate models of environmental variables showed high, significant explanatory power over the variation in viral abundance, matching or surpassing that of bacterial abundance. Combining environmental variables with bacterial abundance into multivariate models further improved the explanatory power over the variation in viral abundance in the models for all environments. Salinity, temperature and nutrients were significant variables across all three environments and for all models.

While no single environmental variable has strong explanatory power over the variation in viral abundance, a combination of them is in effect. The type and strength of significant variables incorporated in multivariate models differs among environments.

These findings help to understand the role of environmental variables in the virus to host relationship and how this relationship changes with environmental conditions.

2.2 Introduction

Viruses play crucial roles in aquatic ecosystems, controlling host diversity and the flux of nutrients and carbon through the viral shunt (Wilhelm and Suttle, 1999). They are highly abundant in aquatic ecosystems, and in different environments range in concentration from 10^6 ml^{-1} to as high as 10^8 ml^{-1} (Proctor and Fuhrman, 1990; Paul *et al.*, 1993; Ortmann and Suttle, 2005) with generally lower abundances in the deep sea and higher abundances at productive coastal sites. As contact rates between viruses and their potential hosts are proportional to viral abundance, increasing densities of viruses may lead to a greater impact on microbial host populations (Murray and Jackson, 1992; Mann, 2003).

Over the years it has been established that viral abundance is about an order of magnitude higher than bacterial abundance, but the virus to bacteria ratio (VBR) varies greatly among host virus systems and environments (Fuhrman and Suttle, 1993; Wommack and Colwell, 2000; Knowles *et al.*, 2016; Wigington *et al.*, 2016). In a metaanalysis of 25 studies Wigington et al. (2016) found that the median VBR ranged from 10.5 to 16 with depth. They also showed the limitation of models with a fixed VBR ratio of 10:1 and used non-linear power functions to relate viral and bacterial abundances. Knowles et al. (2016) did show a linear correlation between viral and bacterial abundances across a range of habitats, and that there was a relative decrease in viral abundance with increasing bacterial abundance. In Wigington et al. (2016) and Knowles et al. (2016) it is apparent that the relationships between viral abundance and bacterial abundance vary substantially among projects. As well, an earlier study found a significant difference in correlations between bacterial and viral abundances in samples from lakes, the Pacific, deep Pacific and Arctic oceans (Clasen *et al.*, 2008). Observations that the VBR changes under different conditions and among locations implies that it could be affected by environmental variables, with burst size, viral decay rates and photosynthetic host density potentially affecting the VBR (Clasen *et al.*, 2008; Wigington *et al.*, 2016). Hence, while viral and bacterial abundances for specific studies or locations are often highly correlated, deriving relationships that extend across biomes likely requires models that include environmental variables that affect the virus-host relationship.

Temperature and salinity are environmental variables that can directly affect virus-host interactions. For example, in the microalgae *Phaeocystis globosa* and *Heterosigma akashiwo*, lysis of infected cells occurred over a narrow temperature range and the different viruses were inactivated above temperatures ranging from 20 to 35°C (Nagasaki K, 1998; Baudoux and Brussaard, 2005). Similarly, inactivation at 40°C was shown for a phage of the prokaryote *Pseudoalteromonas marina* (Hardies *et al.*, 2013). Inactivation temperatures for marine viruses are usually above 20 °C, which is higher than what many virus-host systems in temperate regions are likely to encounter in nature. It can however play a role in microenvironments even in temperate waters. Furthermore, a rise in temperatures can favor the switch from a lysogenic to a lytic cycle in a marine phage-host system (Williamson and Paul, 2006), which would affect the total community viral production. Salinity has also been shown to interfere with the initial step of viral infection with high salt concentrations lowering infectivity and adsorption for a range of bacteria-virus host systems above 3-4 M NaCl in hypersaline environments, while marine phages have ionic requirements for stability (Kukkaro and Bamford, 2009; Mojica and Brussaard, 2014). However, another study showed an increase in viral abundance and drastic change in the viral community composition at hypersaline conditions above 240 ‰ (Bettarel *et al.*, 2011).

Light can influence virus-host interactions in positive or negative ways. Photosynthetically active radiation (PAR) is required for phytoplankton growth and thus crucial for replication of phytoplankton viruses. Even adsorption of viral particles to their host can be light dependent (Jia *et al.*, 2010), as can be the duration of the viral replication cycle and the burst size (Brown, Campbell and Lawrence, 2007; Baudoux and Brussaard, 2008). Yet, some viruses infecting phytoplankton, including those infecting *H. akashiwo*, appear to be less sensitive to changes in the light regime (Juneau *et al.*, 2003; Lawrence and Suttle, 2004). Nonetheless, the final stage of virus replication is very energy demanding and can be especially vulnerable to light limitation in photosynthetic hosts (Mojica and Brussaard, 2014). Light can also have highly negative effects on viral replication. For example, UV radiation is a major factor causing viral decay, and decay rates for viruses of bacteria, cyanobacteria and eukaryotic phytoplankton increase in proportion to irradiance (Murray and Jackson, 1992; Cottrell and Suttle, 1995a; Noble and

Fuhrman, 1997; Garza and Suttle, 1998). In the ocean, light effects are restricted to the upper photic zone, with PAR influencing interactions of viruses of photosynthetic hosts, and UV radiation causing decay of all viruses.

There can also be profound effects of nutrients on virus-host interactions. Since viral particles mainly consist of a genome and a capsid, they have a different chemical composition than cellular organisms. A recent study (Jover *et al.*, 2014) calculated that the C:N:P stoichiometry of viruses is about 17:6:1, which is drastically different than that of their cellular hosts, which are typically 69:16:1 for heterotrophs and 106:16:1 for phototrophs (Redfield, Ketchum and Richards, 1963; Suttle, 2007; Jover *et al.*, 2014). Moreover, up to 87% of cellular phosphorus can be assimilated into viral particles during replication, highlighting the relatively high demand of viruses for nitrogen and phosphorus and the importance of these nutrients for viral replication (Jover *et al.*, 2014). For example, phosphorus depletion resulted in reduced viral production for a variety of prymnesiophytes and their viruses (Bratbak *et al.*, 1998; Maat *et al.*, 2014), and production of viruses infecting *Emiliania huxleyi* were affected by phosphate and nitrate availability (Jacquet *et al.*, 2002). In turn, phosphate addition can increase viral production (Motegi *et al.*, 2015). The limited available data indicate that nitrogen limitation either has no impact or reduces viral production (Bratbak, Egge and Heldal, 1993; Bratbak *et al.*, 1998). In summary, environmental factors affect viral replication, and thus would be expected to affect the relationship between virus and bacterial abundances. As well, there is mounting data that hosts and viruses adapt to environmental conditions (Chow *et al.*, 2013).

Despite the highlighted importance of environmental factors to virus-host interactions, their relationship to the relative abundances of viruses and bacteria in the environment has not been explored. This study addresses these influences by exploring which environmental variables influence the relative abundances of viruses and bacteria across a wide range of samples from diverse environments. In turn, this will lead to better predictions of how environmental changes will affect the relative abundances of viruses and bacteria.

2.3 Materials and methods

Sampling

Data from 515 samples were compiled from several years of data collected in Saanich Inlet (SI; 48°35' N, 123°30' W) (Torres-Beltran et al., submitted) and Rivers Inlet (RI; 51°26' N, 127°38' W) (Tommasi *et al.*, 2013), B.C., Canada, as well as along a cruise track from the Labrador Sea to the coast of British Columbia through the Canadian Arctic as part of the Canadian 3 Oceans project (C3O) (Carmack *et al.*, 2010) (Figure 2.1).

Water samples from depth profiles were collected with Go-Flo bottles and subsampled for various parameters as detailed below. Samples were taken from surface waters to a maximum depth of 1000 m.

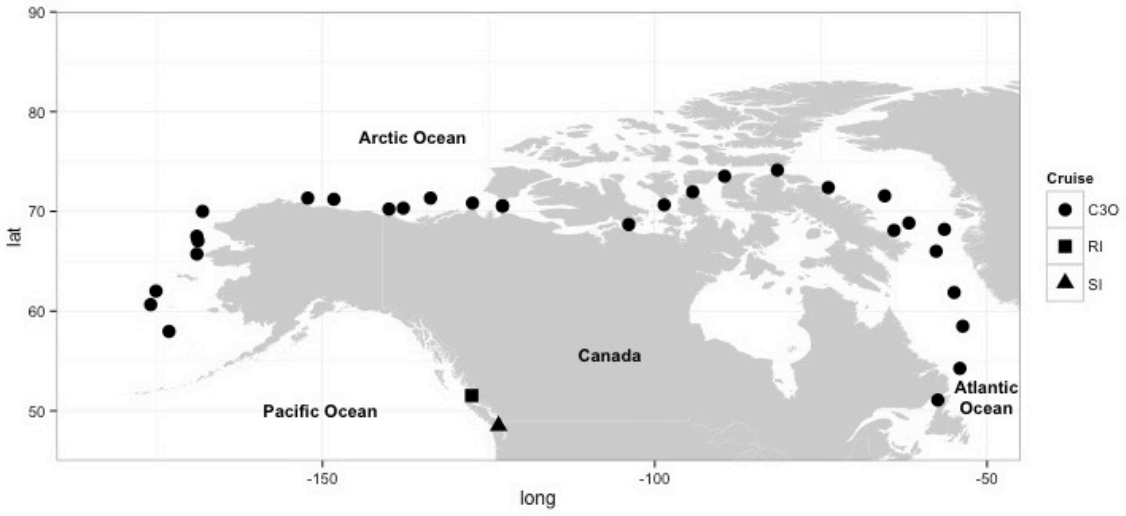

Figure 2.1: Sampling locations by project.
Each location represents multiple depths and/ or time points; C3O, Canadian three Oceans; RI, Rivers Inlet; SI, Saanich Inlet.

Viral abundance and bacterial abundance by flow cytometry

Abundances of dsDNA viruses and prokaryotes, referred to as the predominant bacteria hereafter, were determined in duplicate water samples using a Beckton Dickinson FACSCalibur flow cytometer with a 15 mW 488 nm air-cooled argon ion laser, as described in (Brussaard, 2004). Briefly, samples were fixed for 15 min at 4 °C in the dark with electron microscopy-grade glutaraldehyde (25 %), final concentration 0.5 %, followed by snap-freezing in liquid nitrogen and storage at -80 °C. Right before analysis, the

samples are thawed and diluted in 0.2 µm filtered, autoclaved TE 10:1 buffer (10 mM-Tris HCl; 1 mM EDTA pH 8.0) and stained with SYBR Green I (Invitrogen, Carlsbad CA) at a final concentration of $0.5x10^{-4}$ of the commercial stock, for 10 min at 80 °C in a water bath. Samples were diluted in TE buffer (pH 8.0) if necessary to reach 100 to 1000 events s^{-1}. Viruses were discriminated by plotting green fluorescence against side scatter, and the results analyzed with CYTOWIN version 4.31 (Vaulot, 1989).

Nutrient analysis

Nutrient samples were filtered through 0.22 µm pore-size PVDF syringe filters and stored at -20 °C until analysed. Total nitrate (reduced to nitrite and nitrite; referred to as the predominant nitrate hereafter), phosphate and silicate were analyzed with a Bran & Luebbe AutoAnalyzer 3 (SI and RI; SPX-Flow, Norderstedt, Germany) or a Technicon AutoAnalyzer 2 (C3O; SEAL-Analytical, Norderstedt, Germany) using air-segmented continuous-flow analysis. Colorimetry was used to measure the concentrations of reduced nitrate (Armstrong, Stearns and Strickland, 1967) and silicate at 550 nm, and reduced orthophosphate (Murphey and Riley, 1962) at 880 nm.

Physical properties

For physical data, in situ profiles of temperature, salinity and depth were measured with a SBE 25 (SI and RI) or SBE 911 (C3O) CTD (Seabird Electronics, Inc., Bellevue, WA). Chlorophyll concentration was estimated by a fast-repetition-rate fluorometer (FRRF), for SI and RI a WetStar fluorometer (Seabird Electronics, Inc., Bellevue, WA) for C3O a Seapoint Chlorophyll Fluorometer (Seapoint Sensors, Exeter, NH), were mounted to the CTD. Oxygen was measured with a SBE 43 oxygen sensor, and photosynthetically active radiation (PAR) was measured with a QSP-200PD sensor (Biospherical Instruments, San Diego, CA).

Statistical analyses

Of the 515 samples, 47 samples from Saanich Inlet were missing bacterial counts and 211 samples from Rivers Inlet did not have PAR data; these were left out of the analysis when applicable. Other irregularly missing data points, with <10 % missing per variable,

were filled by multiple imputation. The data were divided into three subsets of samples: Arctic, including some sub-Arctic samples, inlet, and hypoxic. Data from Saanich Inlet and Rivers Inlet comprised the inlet subset; data from C3O made up the Arctic subset, and samples with an oxygen concentration below 1.5 ml l^{-1} were pooled into the hypoxic subset. Statistical analysis was done in R, the statistical language (R, 2015). A linear discriminate analysis (LDA) of the samples based on scaled environmental variables was performed with the MASS package (version 7.3-40) to define the environments. Input variables for the LDA were temperature, salinity, chlorophyll, nitrate, phosphate, silicate and oxygen. Samples for one sampling day and one site were removed from the inlet subset due to extremely high viral counts, exceeding 1.5 times the interquartile range, and were thus considered to be outliers.

Single variable correlations were measured by linear models with viral and bacterial abundances being log transformed, while nitrate and phosphate data were not transformed, models were also built in the R environment. The explanatory power of the models were expressed as the coefficient of determination (R^2) and significances in p-values, m denoting the slope of the regression. Multivariate correlations were determined with generalized linear models, with Gaussian distribution and logarithmic link functions being run for viral abundance against environmental variables and / or log transformed bacterial abundance using the MASS package (Venables and Ripley, 2002). Models were run at a range of complexities, ranging from one input variable to all variables. For each complexity the optimal combination of variables was selected based on the Akaike Information Criterion (AIC) with the Stats package (R, 2015). Optimal models were then selected by comparing the AICs and considering improvements in explanatory power at different complexities; a relative drop in the AIC of two was considered relevant. Model fit was tested with a combined McFadden pseudo R^2 and significance was tested on z-values per coefficient, with a significance threshold of 0.05. Tests were performed with the BaylorEdPsych (version 0.5) package (Beaujean, 2012).

2.4 Results

The data used in this study were categorized into inlet samples from Saanich and Rivers Inlets, hypoxic samples, mainly from deep inlet water, and arctic samples from Canadian Arctic and sub-arctic; each sample had characteristic environmental conditions.

Viral abundances across the data that went into models ranged from $4.83x10^5$ to $1.40x10^8$ viruses ml^{-1}, and bacterial abundances ranged from $7.31x10^4$ to $7.40x10^7$ bacteria ml^{-1} (Table 2.1).

Table 2.1: Ranges and mean values of data included in the statistical analysis.

	min.	max.	mean	unit
Temperature	-1.710	15	7	°C
Salinity	3.060	35	31	PSU
Chlorophyll	0.030	44	2	mg m^-3
Oxygen	0.005	10	4	ml l^-1
PAR	0.000	669	25	µmol quanta m^-2 s^-1
Nitrite/Nitrate	0.010	54	15	µM
Phosphate	0.006	7	2	µM
Silicate	0.070	141	43	µM
Bacteria	7.31x10^4	7.40x10^7	1.66x10^6	# ml^-1
Viruses	4.83x10^5	1.40x10^8	8.35x10^6	# ml^-1

A set of outlier samples from June 2009 in Rivers Inlet had extraordinarily high viral abundances with $1.40x10^8$ viruses ml^{-1} at 10 m, which remained above $4x10^7$ viruses ml^{-1} until 320 m depth. Bacterial abundances were proportionally high and varied between $7.4x10^7$ and $2.04x10^7$ bacteria ml^{-1} over the same depths, but the environmental variables did not show a correlated pattern.

Environmental variables also varied widely across the data sets. Temperature ranged from -2 to 15 $^{\circ}C$ and salinity from 3 to 35 PSU, while chlorophyll and oxygen ranged from 0.03 mg m^{-3} and 0.005 ml l^{-1} to 44 mg m^{-3} and 10 ml l^{-1}, respectively. PAR data available for Saanich Inlet and C3O samples went down to zero in hypoxic water layers and reached a maximum of 669 µmol quanta m^{-2} s^{-1}. Nutrient values ranged from 0.01 to 54 µM for nitrate, 0.006 to 7 µM for phosphate and 0.07 to 141 µM for silicate.

Classification of samples into environments

Linear discriminate analysis (LDA) of all samples based on scaled environmental data, temperature, salinity, oxygen, nitrate, phosphate, silicate and chlorophyll, showed that the data can be separated into three distinct groups (Figure 2.2), reflecting arctic, inlet and hypoxic environments. The first dimension LD1 describes 92.6 % of the variation and the second dimension LD2 7.4 %. The arctic and inlet samples partially overlap in the LDA plot while the hypoxic samples are clearly separated.

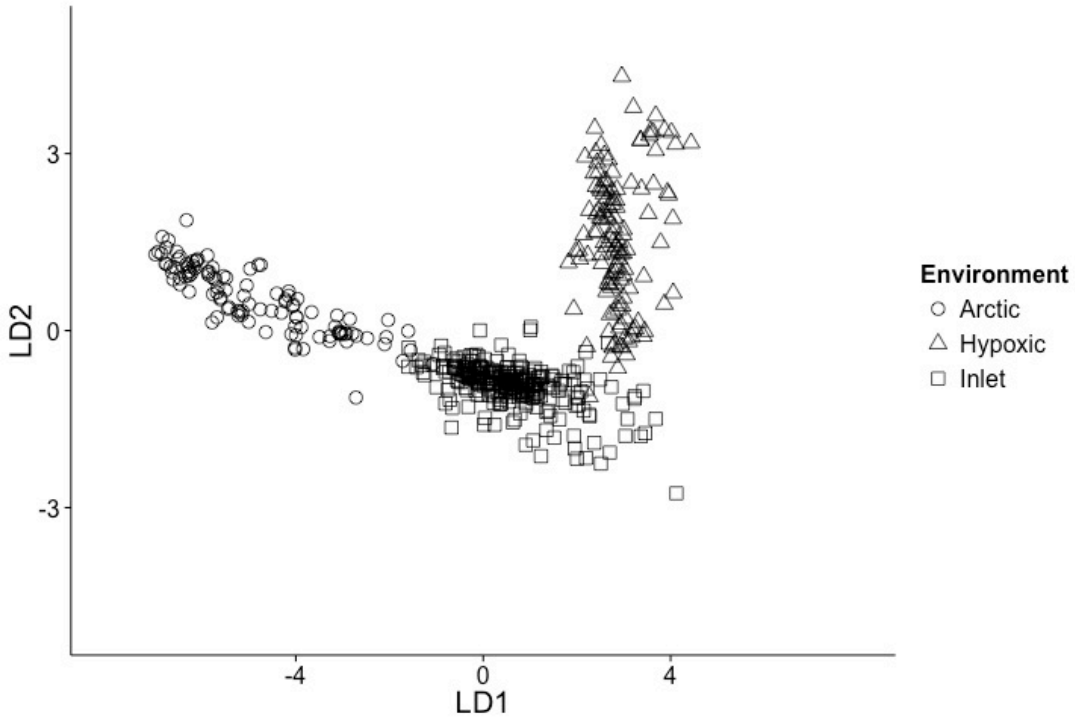

Figure 2.2: LDA of samples used in models.
The analysis is based on temperature, salinity, nitrate, phosphate, silicate, chlorophyll and oxygen. Symbols depict the classified environment, Arctic samples (open circles), Inlet samples (open squares), Hypoxic samples (open triangles).

Besides their variability in temperature and salinity, the three environments varied drastically in the availability of nitrate and phosphate (Figure 2.3). Nitrate to phosphate ratios in the inlet and coastal environments co-varied with a ratio of about 12:1, higher than the average elemental N:P stoichiometry of 5:1 for viral particles, but lower than the ratio of 16:1 associated with phytoplankton in balanced growth or heterotrophic bacteria

(Redfield, Ketchum and Richards, 1963; Jover *et al.*, 2014). Nutrient concentrations also co-varied with depth, with surface samples generally being low in nutrients. Furthermore, coastal samples generally showed lower nitrate concentrations than inlet samples. The majority of samples have a relatively low phosphate concentration when compared to nitrate concentrations. This trend was reversed in the hypoxic samples with nitrate and phosphate concentrations being negatively correlated.

Explanatory power of single variable linear models

Linear models (LM) showing the distribution of direct relationships of log transformed viral abundances vs. log transformed bacterial abundances for the Arctic, inlet and hypoxic data sets are shown in figure 2.4. For the inlet and Arctic data sets there were significant positive relationships between viral and bacterial abundances, explaining 48 % of the variation in viral abundance in the inlet and 66 % in the Arctic. In the hypoxic samples there was no significant relationship between viral and bacterial abundances (R^2=-0.01, p=0.79) (Table 2.2).

Nitrate and phosphate concentrations showed significant relationships to viral abundances in Arctic and inlet environments (Figures 2.5 and 2.6, Table 2.2). However, these relationships varied in strength and on average only explained ~ 20-40 % of the variation in viral abundances. For nitrate, the R^2 values were 0.37 for arctic samples and 0.33 for inlet samples, while for phosphate the values were 0.12 and 0.28, respectively. Relationships between viral abundances and nitrate or phosphate for the hypoxic samples were low in explanatory power (R^2=0.014; 0.009). Generally, viral abundance showed an inverse correlation to depth.

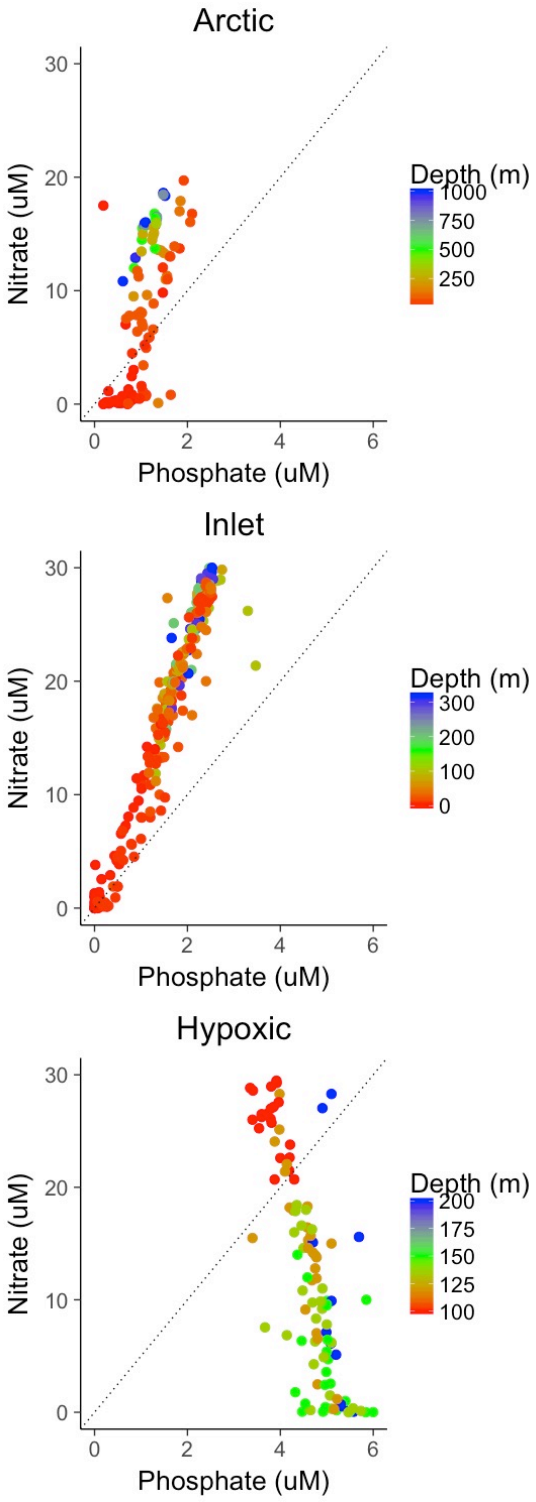

Figure 2.3: Nitrate to phosphate ratio for the samples from the three different environments. Colors indicate the sampling depth. The dashed lines show the elemental 5:1 stoichiometric N:P ratio of viral particles.

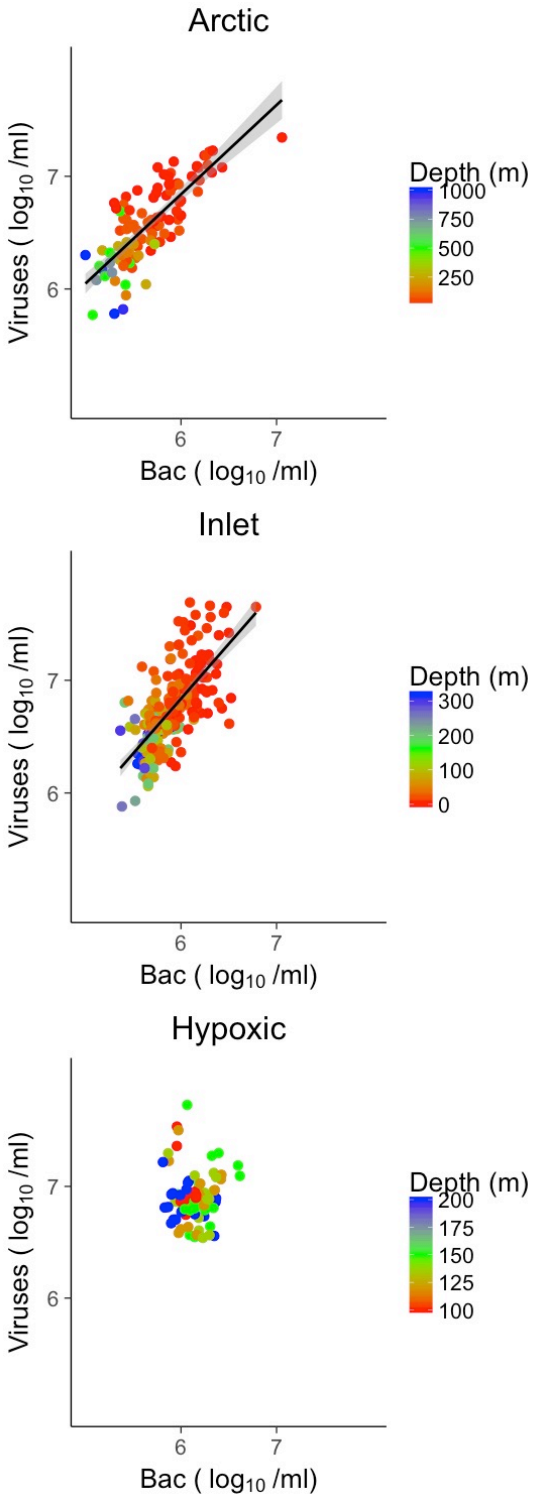

Figure 2.4: LMs of viral abundance to bacterial abundance.
Viral and bacterial abundances are log transformed, linear regression shown, grey shading indicates the 95 % interval, R^2 and slope (m) shown, significances (p) are $2.5e^{-27}$; $1.9e^{-37}$; 0.79.

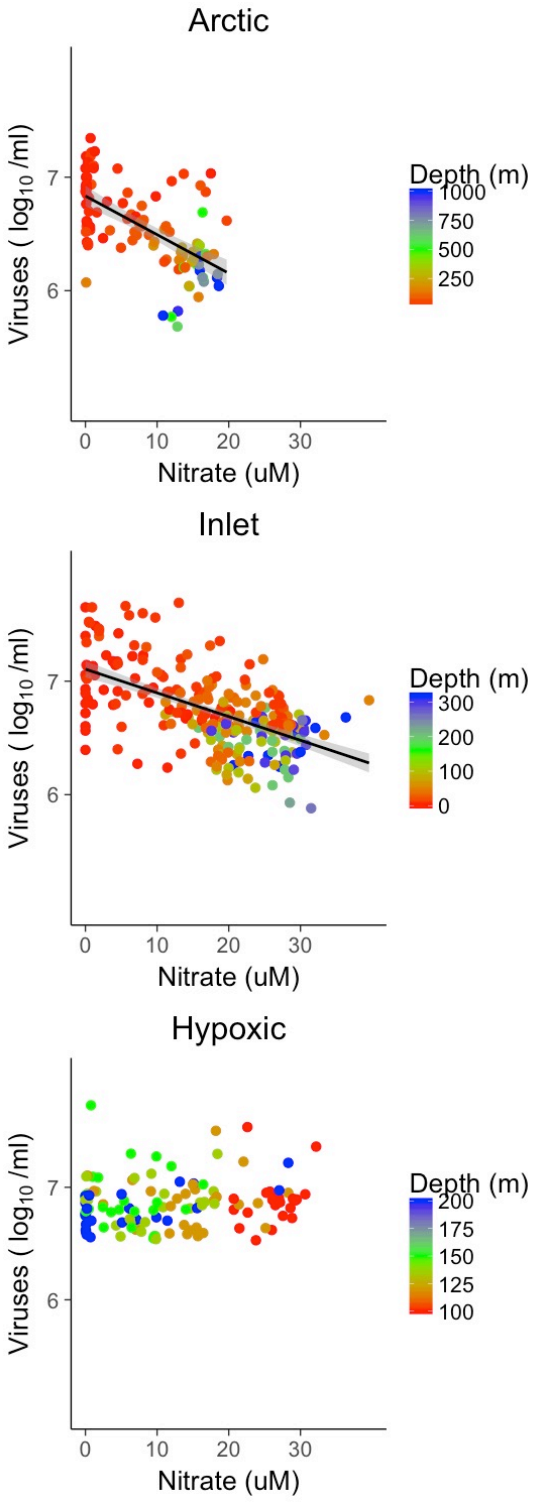

Figure 2.5: LMs of viral abundance to nitrate (μM) concentrations.
Viral abundance is log transformed, linear regression shown, grey shading indicates the 95 % interval, R^2 and slope (m) shown, significances (p) are $1.4e^{-12}$; $1.2e^{-24}$; 0.096.

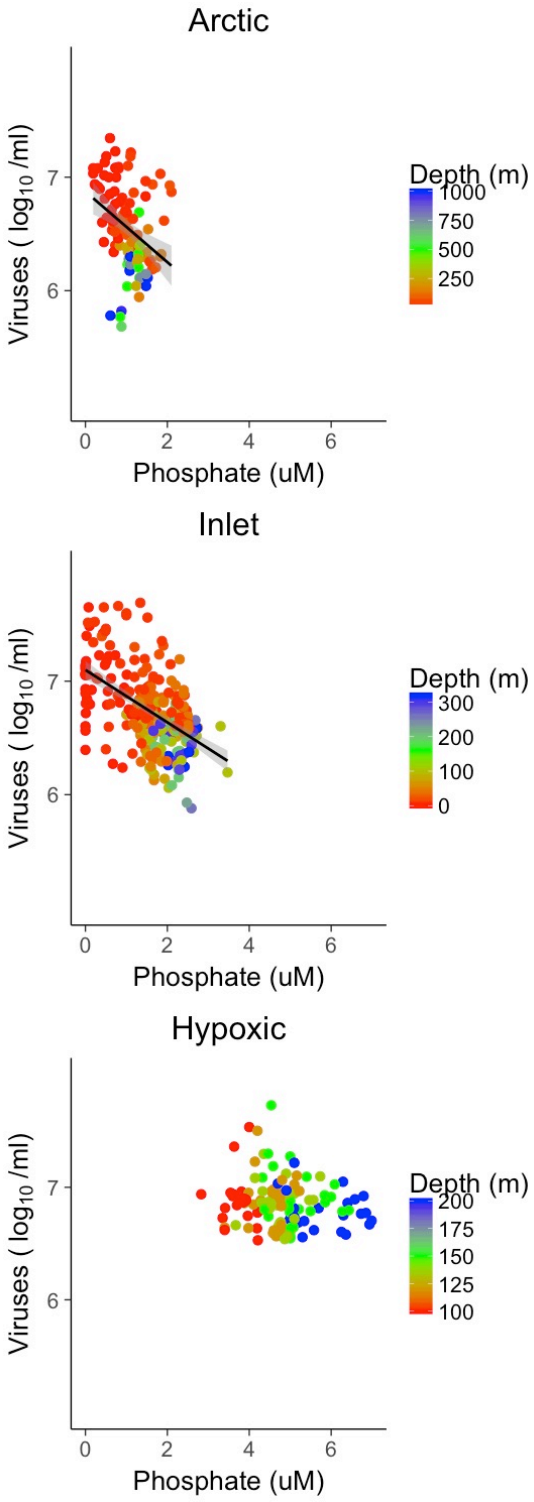

Figure 2.6: LMs of viral abundance to phosphate (μM) concentrations.
Viral abundance is log transformed, linear regression shown, grey shading indicates the 95 % interval, R^2 and slope (m) shown, significances (p) are $1e^{-04}$; $3.6e^{-20}$; 0.15.

Table 2.2: Model statistics for bivariate linear models.

		Arctic	Inlet
Bac. (log_{10})	R^2	0.66	0.48
	Slope	0.8	0.97
	p-value	2.50E-27	1.90E-37
NO_3	R^2	0.37	0.33
	Slope	-0.03	-0.02
	p-value	1.40E-12	1.20E-24
PO_4	R^2	0.12	0.28
	Slope	-0.31	-0.23
	p-value	1.00E-04	3.60E-20

Multivariate models show increased explanatory power

Multivariate models of viral abundance were based on generalized linear models (GLM) and logarithmic link functions. For each environment the best model was selected based on the AIC. Combining only environmental variables, excluding bacterial abundance, produced meaningful models in all three environments, roughly matching the explanatory power of bacterial abundance (Figure 2.7). The coefficient of determination for the three multivariate models was assessed by McFaddin pseudo R^2. Pseudo R^2 of the GLMs and viral abundance in arctic, inlet and hypoxic environments were 0.56, 0.47 and 0.31, respectively. Significant variables for all three environments was temperature, nitrate was significant in the Arctic and inlet environment (Table 2.3). Silicate was a significant variable for the Arctic and hypoxic environments, while phosphate was only significant in the hypoxic environment. Notably, for the inlet and hypoxic samples, the models using combined environmental variables created better correlations than models based on bacteria only.

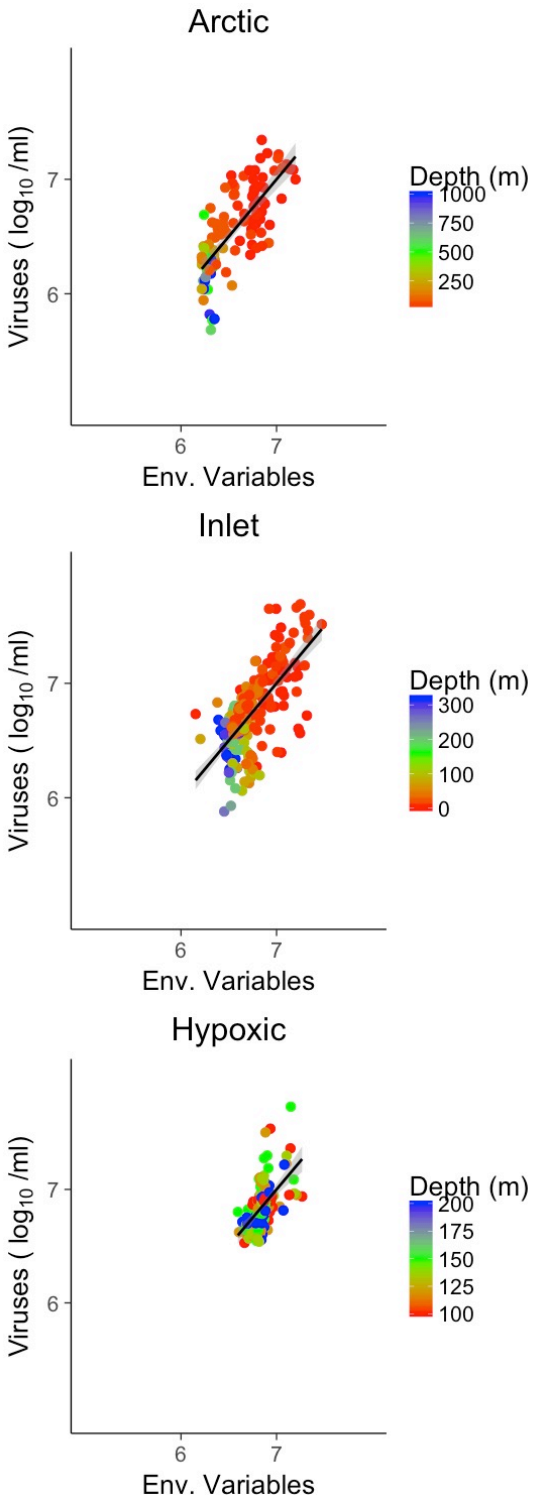

Figure 2.7: GLMs of viral abundance to environmental variables.
Gaussian model with a logarithmic link function. Linear regression shown, grey shading indicates the 95 % interval, significant pseudo R^2 and slope (m) are shown.

The combined models of bacterial abundance and environmental variables substantially improved the relationship to viral abundances across all environments (Figure 2.8). For the Arctic and inlet samples pseudo R^2s are high with 0.72, 0.59. The hypoxic model did not include bacterial abundance as a significant variable. Again, best models were identified by the AIC for each environment.

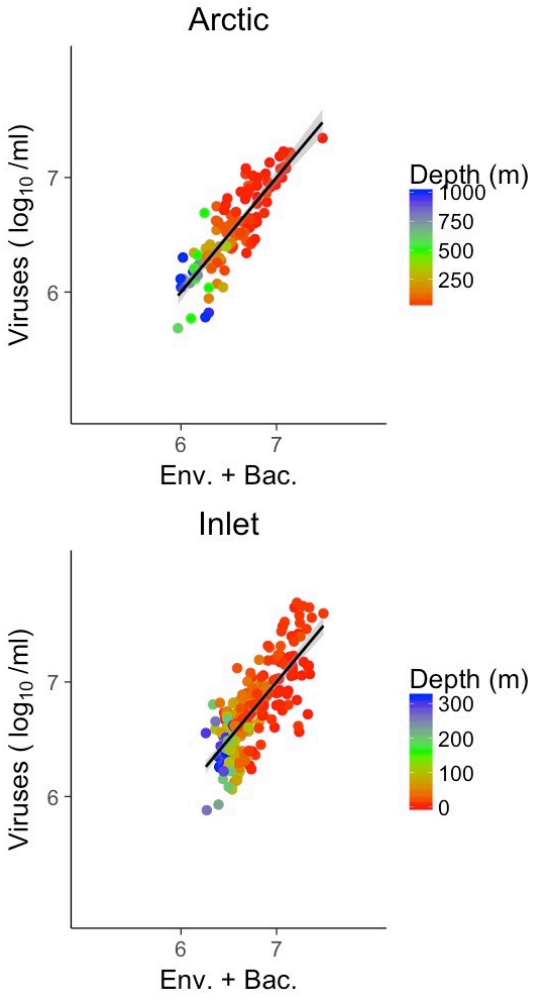

Figure 2.8: GLMs of viral abundance to bacterial abundance and environmental variables.
Bacterial abundance is log transformed. Gaussian model with a logarithmic link function. Linear regression shown, grey shading indicates the 95 % interval, significant pseudo R^2 and slope (m) are shown.

Besides bacterial abundance, nitrate was the only significant variable in the models for both environments (Table 2.4). The improvement over models solely based on bacterial abundances was stronger for the inlet samples than for arctic samples. GLMs

for samples where PAR data were available showed that PAR was not a significant parameter and did not improve the explanatory power of the models.

Table 2.3: Parameters of GLMs based on environmental variables.
AIC, pseudo R^2, sample size (n) and degrees of freedom (df) shown with the effect sizes for significant variables, fonts indicate the significance level.

Env.	Arctic	Inlet	Hypoxic
McFadden(R^2)	0.56	0.47	0.31
Slope	1.00	1.00	1.00
n / df	109/105	261/258	126/122
Intercept	**5.545**	**3.75**	**13.721**
Temperature	**0.141**	**1.068**	-2.763
Salinity		**0.199**	
Chlorophyll	**0.191**		**0.384**
Oxygen	**0.112**		
NO_3	**-0.018**	**-0.013**	
PO_4			**-0.078**
SiO_4	0.009		**0.004**
PAR			
Signif. Codes	**<0.01**	<0.05	*<0.1*

Table 2.4: Parameters of GLMs based on environmental variables and bacterial abundance.
AIC, pseudo R^2, sample size (n) and degrees of freedom (df) shown with the effect sizes for significant variables, fonts indicate the significance level.

Env.+Bac.	Arctic	Inlet	
McFadden(R^2)	0.73	0.59	
Slope	1.00	1.00	
n / df	109/105	252/249	
Intercept	**5.008**	1.020	
Temperature		**0.774**	
Salinity	*-0.523*		
Chlorophyll	0.098		
Oxygen			
NO_3	-0.010	*-0.003*	
PO_4			
SiO_4			
PAR			
Bac. (log_{10})	**0.607**	**0.665**	
Signif. Codes	**<0.01**	<0.05	*<0.1*

2.5 Discussion

It is well established that viral and bacterial abundances covary at a ratio of approximately 10:1 in marine waters (Wommack and Colwell, 2000). A recent meta-analysis of different studies, concluded that this relationship varies among studies and is usually better described by a power law than by a fixed ratio model (Wigington *et al.*, 2016). Furthermore, it is evident that the drivers of virus-host relationship differ among environments (Clasen *et al.*, 2008). Patterns in viral and bacterial abundances are relevant since an increase in either would increase contact rates of virus-host systems and thus infection rates, with effects on nutrient cycles. Lab and field experiments have identified several environmental variables that can influence viral abundance by affecting viral infection, replication and degradation (reviewed by Mojica and Brussaard, 2014). Here we examine the influence of environmental variables on the relationship between viral and bacterial abundances, and how the effects differ among environments.

Classification of samples into environments

The database was compiled from samples from several locations, seasons and depths. Environmental conditions, bacterial abundances and viral abundances were in the typical ranges for these habitats. The set of outlier samples from Rivers Inlet that was excluded from the models displayed excessively high viral and bacterial abundances which could not be related to any of the environmental variables and not explained in a model. Presumably this was due to a period of frequent lysis events and does show the difficulty in accounting for such extremes.

Based on environmental variables, samples were classified into Arctic, inlet and hypoxic environments. The LDA supported the approach to classify samples rather based on prevailing conditions than on geographic location or project. Arctic samples and inlet samples describe a gradual change in environmental conditions, while the hypoxic samples with dissolved oxygen concentrations below 1.5 ml l^{-1} represent a more drastically different environment, which is in line with other studies (Zaikova *et al.*, 2010; Moffitt *et al.*, 2015). All of the hypoxic samples were found at relatively great depths below 100 meters.

Given the stoichiometry of viral particles, nitrogen and phosphorus are key resources for viral replication and their concentrations could be expected to affect viral production. Nitrate to phosphate ratios were characteristic for each of the subsets of data, and were about 12:1 for the Arctic and inlet data, although concentrations reached higher values for the inlet samples. This ratio was higher than the estimated elemental ratio of 5:1 for viral particles (Jover *et al.*, 2014), but lower than the nitrate to phosphate ratio of ~15:1 found in marine samples (Tyrrell, 1999). The ratio of nitrate to phosphate was reversed to 1:12 in the hypoxic samples, as nitrate is used as an alternative electron acceptor by bacteria under anoxic conditions (Zaikova *et al.*, 2010). Arctic surface and hypoxic deep samples displayed potential nitrate limitation for viral replication with concentrations approaching zero. Nitrogen and phosphorus show a similar relationship between their dissolved inorganic ratio normalized to carbon in seawater and their elemental ratio normalized to carbon in cells (Moore *et al.*, 2013). Shifts in the nitrate to phosphate ratio in water are expected to reflect the nitrogen to phosphorus stoichiometry in cells. Thus, in low phosphate samples the high accumulation of cellular phosphorus in viruses (Jover *et al.*, 2014) could lead to a limitation in phosphorus supply during viral replication.

Explanatory power of single variable linear models

The strength of relationships between viral abundance and single variables differed among the subsets of data. The explanatory power of bacterial abundance was higher for the Arctic data (R^2=0.66) than for the inlets data (R^2=0.48), although both were comparable to relationships reported for other surface and sub-surface studies (Knowles *et al.*, 2016; Wigington *et al.*, 2016). Relationships of viral abundances to nitrate or phosphate were weaker than with bacterial abundance in the Arctic and inlet samples. But the significant explanatory power of nitrate (37 % and 33 %) in the Arctic and inlet environments comes close to that of bacterial abundance, highlighting the importance of nitrogen in viral replication. In the Arctic and inlet models viral abundance and depth showed covariation. However, within the scope of this study we treated depth as a co-variate for the environmental variables, e.g., light, rather than an independent variable. For the hypoxic data relationships for all three single variables failed, apparently this environment features different processes than the other two. Overall, bacterial abundance

showed high explanatory power of the variation in viral abundance in most cases, but nutrient concentrations alone were not adequate to explain viral abundance.

Multivariate models show increased explanatory power

When environmental variables are combined into multivariate models describing viral abundance, and the best models selected based on AIC, they represent a compromise between explanatory power, model fit and model complexity. Based on the pseudo R^2 values, the models for the Arctic and the inlet data explain over 50% of the variation in viral abundance. For the inlet data this matches the explanatory power of bacterial abundance alone, while for the hypoxic data the explanatory power was increased to 31 %, a substantial improvement compared to the insignificant correlations with bacterial abundance, nitrate or phosphate alone. Significant components of the models across all data sets were temperature and at least one of the nutrients. Notably, phosphate is a significant component of the model for the hypoxic model, where nitrate to phosphate ratios were inverted. The importance of nutrients, e.g. phosphate, to viral replication and infection is also highlighted by a reduced viral mortality of phytoplankton under phosphate limitation (Maat *et al.*, 2014). That phosphate was more significant than nitrate in the hypoxic model was presumably a result of the depletion of nitrate in samples that are truly anoxic. Chlorophyll being a significant variable in the hypoxic environment, however, must be a statistical artefact and should be disregarded in future studies. Based on the data presented chlorophyll, as a proxy for phytoplankton, could probably be neglected in future models, also since the majority of dsDNA viruses are produced by and infect heterotrophic bacteria. Multivariate models to explain viral abundance from environmental variables match or even exceed the explanatory power of bacterial abundance.

Overall, the combination of environmental variables and bacterial abundance data greatly improved the explanatory power of the models in all three datasets, with 73 % and 59 % of the variation in viral abundance explained by the multivariate models for the arctic and inlet data. For the hypoxic data, the explanatory power did not increase when including bacterial abundance from the multivariate model using environmental variables only, showing the disconnection between viral abundance and bacterial abundance in this environment. Across the multivariate models the strongest component besides bacterial

abundance was nitrate. While temperature and salinity were significant variables in the models for the Arctic and inlet environments, the hypoxic model only incorporated the nutrients in addition to bacterial abundance. Chlorophyll only was a significant variable in the Arctic environment where phytoplankton viruses feasibly are contributing of a substantial number of the viral community. Supporting this finding, studies found significant correlations between chlorophyll and viral abundance for seasonal samples in inlets of the Beaufort Sea (Payet and Suttle, 2008) and fresh water environments (Maranger and Bird, 1995). The influence of environmental variables on the relationship between viral and bacterial abundances, and the differences among environments was consistent with earlier observations from marine and freshwater environments (Clasen *et al.*, 2008). Yet this project extends these observations by a better understanding of the environmental variables in effect.

In conclusion, environmental variables appear to have a significant influence on the relationship between viral and bacterial abundances in marine samples. We provide a first attempt at generalized models that capture this relationship, and a first step towards a better ecological understanding of what controls virus abundance in the ocean. For the purpose of explanatory models, samples can be classified by environment rather than arbitrarily by project, cruise or station. While bacterial abundance is a good, established predictor for viral abundance, it fails in certain environments and can be substantially improved by incorporating environmental variables in more complex models. Individual environmental variables do not have great explanatory power for predicting viral abundances, but when combined in multivariate models can produce explanatory power comparable to that of bacterial abundance alone, or surpass it. Ultimately, the combination of bacterial abundance and environmental variables provides a better explanation of viral abundances across environments than bacterial abundances alone. Based on this study the best environmental variables to explain viral abundance are salinity and temperature, the key physical variables of sea water, and nutrient concentrations, specifically nitrate and phosphate. The three types of environments studied in this project are predicted to be strongly affected by climate change, with increased stratification in inlets, the North Atlantic, Arctic and Northeast Pacific, and associated changes in vertical nutrient fluxes and expanding oxygen minimum zones (OMZ) (Keeling, Körtzinger and Gruber, 2010;

Capotondi *et al.*, 2012; Hordoir and Meier, 2012). Understanding the interplay between viruses, hosts and environmental variables improves the capabilities in predicting how these environments will respond to environmental changes.

Chapter 3: Variation in the genetic repertoire of viruses infecting *Micromonas pusilla* reflects horizontal gene transfer and links to their environmental distribution

3.1 Summary

Prasinophytes, a group of eukaryotic phytoplankton in the division *Chlorophyta*, have a global distribution and are a major component of coastal and oceanic communities. Members of this group are infected by large double-stranded DNA viruses (prasinoviruses) of the *Phycodnaviridae* family, which can be significant agents of mortality. However, information on the genetic diversity of these prasinoviruses and their environmental distribution is limited.

This study examines the genetic repertoire, phylogeny and environmental distribution of phycodnaviruses infecting *Micromonas pusilla* and other prasinophytes and chlorophytes.

The genomes of viruses infecting *M. pusilla* were compared to viruses that infect other prasinophytes and chlorophytes, and a *M. pusilla* host genome. Presumed cell-derived genes were investigated for their closest non-viral homologue to identify their origin. A relationship between the genetic repertoire of viruses and their DNA polymerase phylogeny was established. Using this relationship the prevalence of phycodnavirus ecotypes was assessed in environmental samples.

The data showed that *M. pusilla* viruses share a limited set of core-genes, but vary strongly in their pan-genome, displaying a great diversity in genetic repertoire. This pan-genome contains numerous metabolic genes such as for amino acid synthesis and nucleotide sugar metabolism. Surprisingly few of these presumably host-derived genes are shared with *M. pusilla*, but rather have their closest non-viral homologue in bacteria and other eukaryotes, indicating horizontal gene transfer. This diversity of genetic repertoire was reflected in prasinoviruses communities across environmental samples.

This research highlights the variation in genetic repertoire encoded by prasinoviruses and their evolutionary history driven by horizontal gene transfer. It also reveals a high phylogenetic diversity and a connection to the distribution pattern of

prasinovirus ecotypes, deepening our understanding of the processes of selection on viruses.

3.2 Introduction

Prasinophytes are a divergent group of marine eukaryotic phytoplankton within the Division Chlorophyta (Leliaert *et al.*, 2012). They have a global distribution and are a major component of coastal and oceanic communities, and include the prominent genera *Micromonas*, *Ostreococcus* and *Bathyococcus*. They arguably constitute the second most abundant group of phytoplankton after cyanobacteria with a high importance in primary production (Worden, Nolan and Palenik, 2004; Marin and Melkonian, 2010).

Prasinophytes are infected by large icosahedral viruses with double-stranded DNA genomes (Van Etten *et al.*, 2002). These viruses can be significant agents of mortality and influence nutrient fluxes, host diversity or act in horizontal gene transfer (Brussaard, 2004; Dunigan, Fitzgerald and Van Etten, 2006). The genera *Prasinovirus* and *Chlorovirus* are within the family *Phycodnaviridae* (King *et al.*, 2012) which share properties with other Nucleocytoplasmic Large DNA Viruses (NCLDV) (Koonin and Yutin, 2010). NCLDVs include viruses infecting amoeba and mammals, but phycodnaviruses solely infect algae. While NCLDV genomes range from 100 kb to 2.5 Mb in size (Koonin and Yutin, 2010; Philippe *et al.*, 2015), characterized phycodnaviruses range from 150 to 560 kb (Van Etten, Lane and Dunigan, 2010).

In recent years, several prasinovirus genomes have been sequenced. The genome of *Micromonas pusilla* virus MpV1 is 184,095 bp; genomes of *Ostreococcus tauri* virus OtV5 and O. *lucimarinus* virus OlV2 are 186,234 and 196,300 bp, respectively, and those of the *Bathyococcus prasinos* viruses BpV2 and BpV1 are 187,069 are 198,519 bp, respectively (Derelle *et al.*, 2008, 2015; Moreau *et al.*, 2010). Generally, prasinoviruses are similar in genome structure and content, and show a high degree of orthology and synteny (Weynberg *et al.*, 2009; Moreau *et al.*, 2010).

Despite the similarity in genome architecture among prasinoviruses, they are typically host specific within a species, as shown for viruses infecting Os*treococcus* spp. (Derelle *et al.*, 2015). As well, viruses infecting *M. pusilla* do not infect *Ostreococcus* spp.; yet, they infect *M. pusilla* strains from different origins (Martínez *et al.*, 2015). However,

some viruses infect and potentially incorporate genes across genera (Iyer *et al.*, 2006); thus prasinoviruses have the potential to acquire genes from different host genera. Consequently, the genome of prasinoviruses is comprised of a small set of core genes and a larger flexible genome.

Although the functions of most prasinovirus genes are still unknown (Santini *et al.*, 2013), there is a set of core genes that are essential for viral replication and structure including DNA polymerase, DNA topoisomerase and seven to eight genes encoding capsid proteins (Derelle *et al.*, 2008, 2015; Weynberg *et al.*, 2009). In contrast, the flexible pan-genome comprises many genes of unknown function, but also metabolic genes similar in their role to the auxiliary metabolic genes (AMGs) found in cyanophages (Breitbart *et al.*, 2007). The genomes vary in the tRNAs (Moreau *et al.*, 2010) and K^+ channels (Siotto *et al.*, 2014) that are encoded. As well, *Ostreococcus* viruses possess genes of presumed cellular origin, some with homologues in their hosts. These include genes associated with sugar metabolism (glycosyltransferases), nucleotide modification (ribonucleotide reductase), amino-acid synthesis (acetolacetate synthase), phosphate starvation (phoH) and many more (Weynberg *et al.*, 2009). Moreau et al. (2010) have also shown that there are homologues of cell-derived genes in viruses that infect members of the genera *Bathyococcus*, *Ostreococcus* and *Micromonas*. Assuming that these genes are expressed during viral replication, the viruses carrying them will be more “fit” under conditions where these genes carry a selective advantage. Yet, the distribution of pan-genes among prasinoviruses remains to be studied.

Core genes have been used as targets to investigate the distribution and diversity of specific groups of viruses across environments. Moreover, because these genes are conserved and not laterally transferred, they can be used to build phylogenetic relationships within groups of viruses. Viruses which are most closely related would also be expected to be similar in terms of overall gene content. Hence, viruses that are most similar to each other with respect to the phylogeny of their core genes, would be expected to share a similar genetic repertoire.

Genes encoding DNA polymerase B (DNApol) and the Major Capsid Protein (MCP) have been used extensively to study phycodnavirus diversity. In particular, DNApol sequences have been used to infer diversity and phylogenetic relationships among phycodnaviruses

in marine (Chen and Suttle, 1996; Chen, Suttle and Short, 1996; Short and Suttle, 2002, 2003) and freshwater (Short and Short, 2008; Clasen and Suttle, 2009; Gimenes *et al.*, 2012) environments. Similarly, MCP has been used as a marker of phycodnavirus diversity (Larsen *et al.*, 2008; Rowe *et al.*, 2011). Clerissi, Grimsley, Ogata et al. (2014) used full and partial DNApol and MCP gene sequences to phylogenetically compare prasinoviruses and chloroviruses, and showed that full-gene phylogenies for DNApol and MCP were congruent. However, looking at diversity and phylogeny with amplicon sequences is compromised because of the specificity of the primers. For example, the primers typically used for DNApol (Short and Suttle, 2002) amplify MpV sequences (Short and Short, 2008; Clasen and Suttle, 2009), whereas, the primers used for MCP miss them (Larsen *et al.*, 2008). These differences were highlighted in a freshwater study (Zhong and Jacquet, 2014) in which primers for DNApol and MCP favored amplification of prasinovirus and prymnesiovirus sequences, respectively.

Another approach to examine the genetic relatedness among viruses is to build multi-gene phylogenies. This can be done based on selected core genes (Derelle *et al.*, 2015), or by comparing the presence and absence of genes across entire genomes. Gene presence or absence trees provide a rigorous way to examine evolutionary relationships among large DNA viruses (Yutin, Wolf and Koonin, 2014; Legendre *et al.*, 2015), but the approach is not amenable to comparing viruses based on environmental sequence data. Nonetheless, gene presence-absence trees can be used to construct robust phylogenetic relationships among sequenced virus isolates, which in-turn, can serve as a backbone for making predictions about virus gene content from environmental amplicon-based sequencing data. Therefore a relationship between a phylogeny based on core-gene sequences, such as for DNApol, and overall gene content would need to be established. In this way, environmental amplicon data for DNApol can be used to infer the gene content of prasinoviruses in nature. This approach is explored in this chapter.

The impact that the gene content of viruses has on their environmental distribution is unexplored, but marine viruses show biogeographic patterns (Breitbart and Rohwer, 2005; Chow and Suttle, 2015; Marston and Martiny, 2016), including viruses infecting O*streococcus tauri* that form distinct communities in contrasting environments (Bellec *et al.*, 2010). Furthermore, Clerissi, Grimsley, Subirana et al. (2014) showed that the diversity

and composition of prasinovirus communities is influenced by environmental factors, particularly the availability of phosphate. A recent study on cyanophage isolates linked the genome similarity and their environmental distribution, thus formulating a diversification of viruses into ecotypes (Marston and Martiny, 2016).

This project contrasts the genomes of prasinoviruses infecting *Micromonas pusilla* to those of other phycodnaviruses from a range of hosts and environments with the goal of describing their genetic composition in the context of their environmental distribution.

3.3 Materials and methods

Genomic analysis of Micromonas viruses

The *Micromonas pusilla* viruses MpV-PL1 and MpV-SP1 were isolated from the mixed layer in the Gulf of Mexico and coastal water of California, and propagated on *M. pusilla* strain UTEX991 (Cottrell and Suttle, 1995b). The viruses were purified from 15 mL of lysate by filtration through 0.45 and 0.22 μm pore-size Durpore membrane, ultracentrifugation and subsequent optiprep gradient centrifugation. The DNA was extracted and purified using Qiamp MinElut Virus DNA pin kit (Qiagen, Hilden, Germany) prior to sequencing to 10-fold depth and assembly by the Broad Institute, using the 454 GS FLX platform and Newbler 2.7 (Roche, Basel, CH). Read assembly resulted in two contigs per virus that were mapped in Mauve 2.3.1 (Darling *et al.*, 2004) to MpV1 as a reference genome. Sequencing gaps were closed by PCR amplification with customized primers (PL1 fwd-GAGGGTGGGCACGTTGGAG, rev-GTCTCTAGGACCCCCACCCT; SP1 fwd-GCTAATGACGAGTTCGGTCG, rev-ACTAAGTAACCGAAACTGTCCCC) to bridge the gaps, cloning of the product and subsequent Sanger sequencing (NAPS, University of British Columbia, Vancouver, BC). Final genomes were assembled in Geneious 6.0.5 (Biomatters Ltd., Auckland, NZ) based on sequence overlap.

To annotate the assembled genomes, Open Reading Frames (ORFs) were called using Artemis 14.0.0 (Carver *et al.*, 2005) using a minimum ORF length of 65 amino acids (195 nt) with start and stop codons. ORFs were translated into amino-acid sequences using the standard genetic code in three reading frames using Artemis. Putative coding sequences (CDS) were searched for homologues in the nr-database (NCBI) with a protein BLAST (BLAST-P). Annotations were manually selected based on a minimum E-value of

E^{-10} and minimum 50 % alignment length. tRNAs were determined with tRNAscan-SE v1.21 (Lowe and Eddy, 1997).

CDS for MpV-PL1, MpV-SP1, MpV1, MpV-12T and *M.pusilla* UTEX 991 were clustered in USEARCH 6.1.544 (Edgar, 2010) based on a 50 % pair-wise identity at the amino-acid level. Viral clusters were labeled based on the annotation of MpV-PL1 where applicable. Genome contents were compared based on a cluster presence-absence scheme and Venn diagrams produced in R (R, 2015). Core genes in the *M. pusilla* viruses were defined when a cluster contained CDS from all four genomes, or a CDS could be associated with a cluster based on functional annotation and BLAST-P analysis.

Deriving similarity in gene content from DNApol sequences

Prasinovirus and chlorovirus genomes were compared phylogenetically. The genomes were clustered using USEARCH 6.1.544, as above, and compared for gene presence or absence. Phylogenetic distances (Dij) among genomes were calculated as [Dij=-ln(Sij/sqrt(Ni*Nj))], where Sij, Ni and Nj are the number of shared genes, number of genes in one genome and number of genes in the other genome, respectively (Yutin, Wolf and Koonin, 2014). A Neighbor-Joining tree based on distance was constructed using the APE package (Paradis, Claude and Strimmer, 2004) in R, and visualized in FigTree 1.4.2 (Rambaut, 2014). Bootstrap values for branch support were calculated from 1000 iterations of random gene cluster sub-sampling. Reference DNApol sequences were extracted from the following genomes (accession numbers): MpV1 (NC_014767); MpV-12T (NC_020864); BpV1 (NC_014765); BpV2 (HM004430); OtV1 (NC_013288); OtV2 (NC_014789); OlV1 (NC_014766); OtV5 (NC_010191); OtV6 (JN225873); PBCV1 (NC_000852); PBCV158 (NC_009899). Amplicon equivalents of 140 aa length were extracted from the reference sequences and aligned with clustalo 1.2.3 (Sievers *et al.*, 2011). Phylogenetic distances among the reference viruses were calculated based on maximum likelihood with the WAG substitution model. The optimal substitution model was selected with prottest-3.4 (Darriba, Taboada and Posada, 2011) and distance calculated using RaxML 8.0 (Stamatakis, 2014). Phylogenetic distances of the gene presence-absence matrix and DNApol were compared with a Mantel Test in R.

Assessing environmental data and prasinovirus sequences

Amplicons of DNApol gene fragments were used to infer prasinovirus diversity in environmental samples. Samples of 20 to 72 liters of water were taken from the surface at three sites in the Strait of Georgia, Jericho Pier (JP) and Point Atkinson (PA), the Juan de Fuca Strait (JF) and in the surface layer and at 200 meters depth, several times per year in Saanich Inlet (SI) (Sampling details in Supplementary table A1). JP, PA and JF samples were sequentially filtered through 47 mm diameter GC50 (Advantec MFS Inc., Dublin, CA) and HVLP (Millipore Merck, Darmstadt, Germany) membrane filters (0.7 μm nominal pore-size for each filter). Similarly, Saanich Inlet samples were filtered through 2.7 μm nominal pore size GF/D filters (Whatman, Maidstone, UK) and 0.22-um pore-size Sterivex filters (Millipore, Billerica, MA). The remaining particulate matter in each filtrate was then concentrated by tangential flow filtration (TFF) with a 30 kDa molecular-weight cutoff cartridge filter (Prep-Scale; Millipore, Billerica, MA) to make a viral concentrate (VC) that was stored at 4 ºC in the dark. For DNA extraction, 12 ml VC subsamples were concentrated by ultracentrifugation for 4 h at 124000 g at 15 ºC, and the pellets eluted with 500ul Tris-HCl 1% SDS at 4 °C overnight. Samples from Saanich Inlet were pooled into surface layer and deep composites. The viral capsids were lysed with Proteinase K (Invitrogen, Carlsbad, CA) (100 μg ml^{-1}) and DNA extracted using phenol:chloroform. Partial DNA polymerase sequences were amplified with AVS1 and AVS2 primers (Chen and Suttle, 1995), and 500 ng of the PCR products used for library preparation and sequencing with a 454 GS FLX with Titanium Chemistry (Roche, Basel, CH) at the Broad Institute (Cambridge, MA). Reverse AVS sequences were denoised using QUIIME v1.4 (Caporaso *et al.*, 2010) and chimeras were removed using UCHIME v4.2.40. Denoised sequences were translated to amino acids with FragGeneScan v1.16 (Rho, Tang and Ye, 2010) and dereplicated using USERACH (v6.1.544). Reads from all environmental samples and reference sequences were pooled and clustered at 97% identity in USEARCH (v8.1). Clusters with only one member were discarded and centroids of the other clusters were aligned with clustalo 1.2.3. Gaps in the alignment were trimmed and a maximum-likelihood tree was built in RaxML 8.0. Environmental reads and reference sequences were parsed using USEARCH (v8.1) at 97 % identity, the phylogenetic tree was edited using iTOL v3.2.4 (Letunic and Bork, 2016). Frequency distribution of parsed

environmental reads were rarefied to the lowest number of cumulative reads per sample using the VEGAN package (Oksanen *et al.*, 2016) in R.

In situ measurements of temperature and salinity were made with electrodes mounted on a Seabird (Seabird, Bellevue, WA) CTD (Saanich Inlet) or YSI probe (YSI, Yellow Springs OH) probe (JP, PA, JF). As well, remote sensing data were extracted from Aqua MODIS data (NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group) to estimate chlorophyll a (chl, mg m^{-3}), daytime sea-surface temperature (SST, 4u, °C), photosynthetically active radiation (PAR, μmol photons m^{-2} s^{-1}) and particulate organic carbon (POC, mg m^{-3} 443/555) as a rolling 32-d composite pre-dating the sampling period, at a 4-km resolution. Data was processed and mapped in R.

3.4 Results

Origin and distribution of genes in Micromonas viruses

The genomes of *Micromonas* viruses MpV-PL1 and MpV-SP1 were completed, analyzed and annotated. Sequencing gaps were closed using custom-designed primers, and a BLAST-P analysis of MpV ORFs against the nr-database improved their annotations, although most still lack a putative function. This study focused on MpV-PL1 and MpV-SP1, and their comparison to MpV1 (NC_014767) (Moreau *et al.*, 2010) and MpV-12T (NC_020864). The viruses were isolated on three strains of *M. pusilla* (Table 3.1) and differ in genome size, the number and average length of their ORFs, GC content and tRNAs. The genome sizes range from 173,350 bp for MpV SP1 to 205,622 bp for MpV-12T, which does not correspond to the number or size of ORFs; MpV1 possesses the fewest (244) but, on average, longest ORFs (715 bp), while MpV-PL1 has the most (275) but not the shortest, on average (684 bp). MpV-12T also has the lowest GC content (39.8 %), while MpV-PL1 has the highest (43.3 %). Although six tRNAs are common in *Micromonas* viruses, MpV-PL1 lacks Leu-tRNA, while MpV-12T carries two copies of Asn-tRNA.

Table 3.1: General genome characteristics of *M. pusilla* viruses.
Comparison of MpV PL1 and SP1 to MpV1 and MpV-12T. Host refers to original host of isolation; ORF length is average; GC % for whole genomes; tRNAs present in genomes, copy number.

	MpV PL1	MpV SP1	MpV1	MpV12T
Genome size (bp)	196960	173350	184095	205622
Host	*Mp*UTEX991	*Mp*UTEX991	*Mp*RCC1109	*Mp*-LAC38
# ORF	275	248	244	253
ORF length	684	659	715	749
% GC	43.3	40.6	41.0	39.8
Asn-tRNA	1	1	1	2
Gle-tRNA	1	1	1	1
Ile-tRNA	1	1	1	1
Leu-tRNA	0	1	1	1
Thr-tRNA	1	1	1	1
Tyr-tRNA	1	1	1	1

The cluster analysis of four *Micromonas* viruses and a host (*M. pusilla* UTEX991) genome based on 50% amino-acid identity (Figure 3.1) revealed 80 ORFs shared by all viruses, 140 are shared in at least two genomes and 357 are unique. While MpV1, MpV-PL1 and MpV-SP1 share about half of their ORFs (130), MpV-12T has only 100 that occur in at least one other virus, and 153, not shared. In contrast, MpV-PL1 and MpV-SP1 have the highest overlap, with 194 ORFs in common. Only six ORFs are shared among the host, *M. pusilla* UTEX991, and the viruses at this similarity level. Three host ORFs occur in all the viruses (ribonucleotide reductase, dUTPase and a cell-division protein), while three others occur in a subset (DNA primase, a heat-shock protein and a hypothetical protein).

Combining the cluster analysis of putative viral genes with an additional BLAST-P analysis against the nr-database revealed a core-genome of 119 genes and 327 genes in a pan-genome (Table 3.2). Core genes include those essential for viral replication and virion structure, such as DNA polymerase type B (DNApol), DNA ligase, transcription initiation factor and seven capsid proteins. Most putative genes are in the flexible pan-genome, including genes which are functionally of cellular origin, such as those involved in carbon metabolism and DNA repair, yet most have no functional annotation. Other putative genes of presumable cellular origin associated with amino-acid synthesis, including acetaldehyde dehydrogenase, acetolacetate synthase and aminotransferase, occur in MpV1 (Moreau *et al.*, 2010), and also MpV-PL1, but are not found in the other

Micromonas viruses. Heat-shock-protein 70 is found in MpV-12T and MpV-PL1, and is also shared with *M. pusilla* UTEX991. The DNA methylase and DNA methyltransferases are site specific and differ among the viruses. Moreover, MpV-PL1 and MpV-SP1 have a putative host-derived gene for 6-phosphofructokinase, MpV1, MpV PL1 and MpV SP1 share dTDP-D-glucose 4,6-dehydratase. In contrast only MpV-12T carries UDP-glucose 6-dehydrogenase and only MpV-SP1 has two transketolase-related genes. Several other genes are shared among MpV-PL1, MpV-SP1 and MpV1, but not with MpV-12T, which also has the most genes without functional annotation.

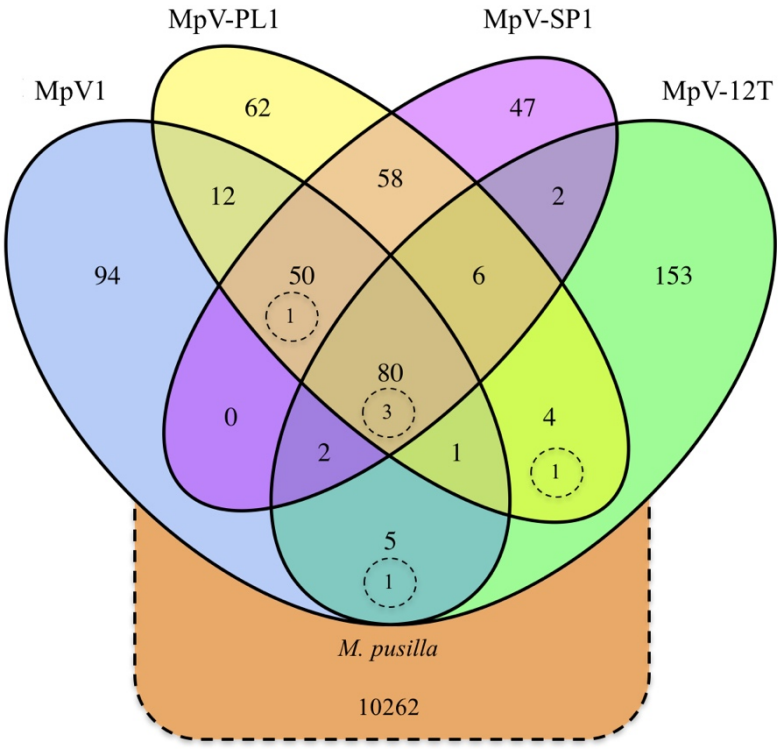

Figure 3.1:Shared genes of four *M. pusilla* viruses *and M. pusilla* UTEX991.
Venn diagram based on clusters by 0.5 amino-acid identity. Dashed circles represent host genes shared with viruses.

Table 3.2: Genes of the core and pan-genomes of four *M. pusilla* viruses.
Classification based on a combined cluster and BLAST-P analysis. The annotation is based on MpV-PL1.

Core-genes		Pan-genes	
Class	**Putative function**	**Class**	**Putative function**
DNA replication	DNA polymerase	AA synthesis	Acetolacetate synthase
	DNA topoisomerase		Acetolactate synthase
	DNA ligase		Aminotransferase
	DNA primase	DNA repair	Heat shock prot 70
Nucleotide metabolism	RNAse		DNA methylase
	Ribonuclease		DNA methyltransferase
	Ribonucleotide reductase		
Transcription	mRNA capping enzyme	Sugar manipulation	dTDP-D-glucose 4,6-dehydratase
	Trans. initiation		UDP-glucose 6-dehydrogenase
	Trans. elongation		6-phosphofructokinase
Structural genes	Capsid protein		transketolase N-term.
	MCP		transketolase B sub.
Metabolism	PhoH	Total Shared	108
Total Core	119	Total Unique	327

A BLAST-P analysis against the nr-database of putative coding sequences (CDS) with a functional annotation revealed that for most core-genes the closest hit is to other virus sequences, while for the flexible genome the close homologues often are cellular (Figure 3.2). However, few of the sequences of presumed cellular origin are found in *M. pusilla* UTEX991, but are rather more similar to sequences in other eukaryotes, bacteria, cyanobacteria or archaea.

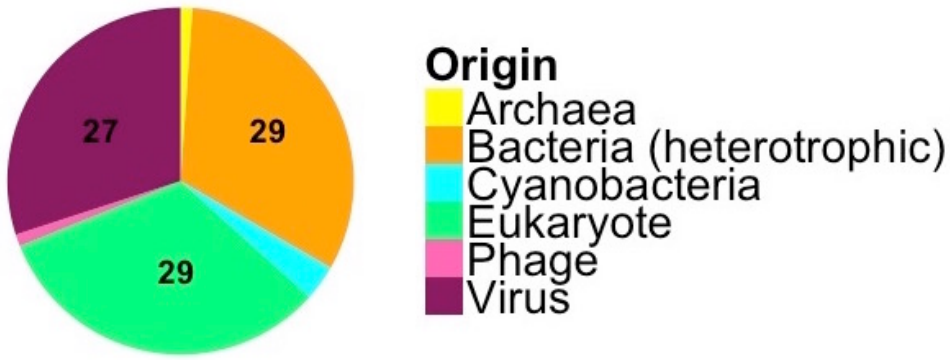

Figure 3.2: Presumed origin of viral genes.
Results for 90 genes with a functional annotation in the four *M. pusilla* viruses examined in this study. Presumed origin is based on BLAST-P hits against the nr-database, numbers of gene shown.

Deriving similarity in gene content from DNApol

A Neighbor-Joining (NJ) phylogenetic analysis of prasinoviruses and chloroviruses based on the presence and absence of putative genes shows the similarity of the viruses to each other (Figure 3.3). The more closely the viruses are related in their gene content, the closer they are on the tree, indicating that the *Chlorella*, *Bathyococcus* and most *Ostreococcus* viruses form well-defined groups; whereas, the *Micromonas* viruses form three distinct branches with MpV-PL1 and MpV-SP1 branching together, and MpV1 and MpV-12T being on separate branches.

Comparing the phylogenetic relationship among prasinoviruses and chloroviruses from analyses of gene presence and absence, and DNApol sequences draws a congruent picture with matching tree topology (Figure 3.3). Also the phylogenetic distances between pairs of viruses based on gene presence-absence data and full-length DNApol sequences (Table 3.3) were highly correlated (Mantel Test), whether chloroviruses were included in the analysis (r=0.99), or not (r=0.96). Amplicons from environmental samples will have to be clustered at an appropriate identity level that is specific to the full length DNApol. Correlating fragments of the reference DNApol sequences at different identity levels to their full length sequences showed decreasing variation with increasing stringency (Figure 3.5). The variation approached zero when clustering DNApol fragments at 97% identity, which was thus applied to the environmental sequences in this study.

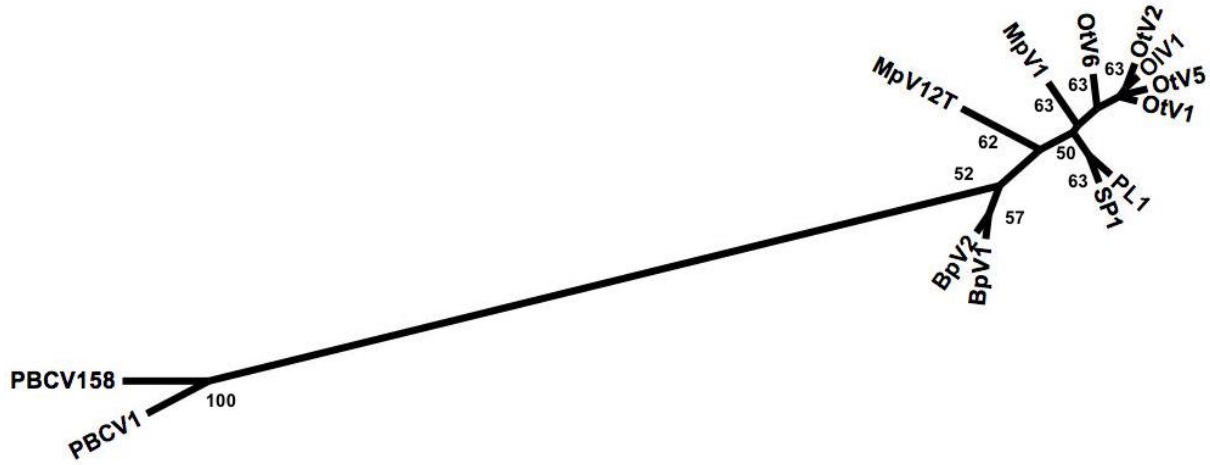

Figure 3.3:NJ phylogeny of prasinoviruses and chloroviruses based on gene content.
Representing isolated viruses infecting the genera *Ostreococcus* (OtV1, OtV2, OtV5, OtV6, OlV1), *Bathyococcus* (BpV1, BpV2), *Micromonas* (MpV1, MpV-12T, MpV-PL1, MpV-SP1) and *Chlorella* (PBCV1, PBV158). The neighbor-joining tree is based on the presence and absence of shared putative genes. Bootstrap values are based on 1000 iterations of sub-sampling.

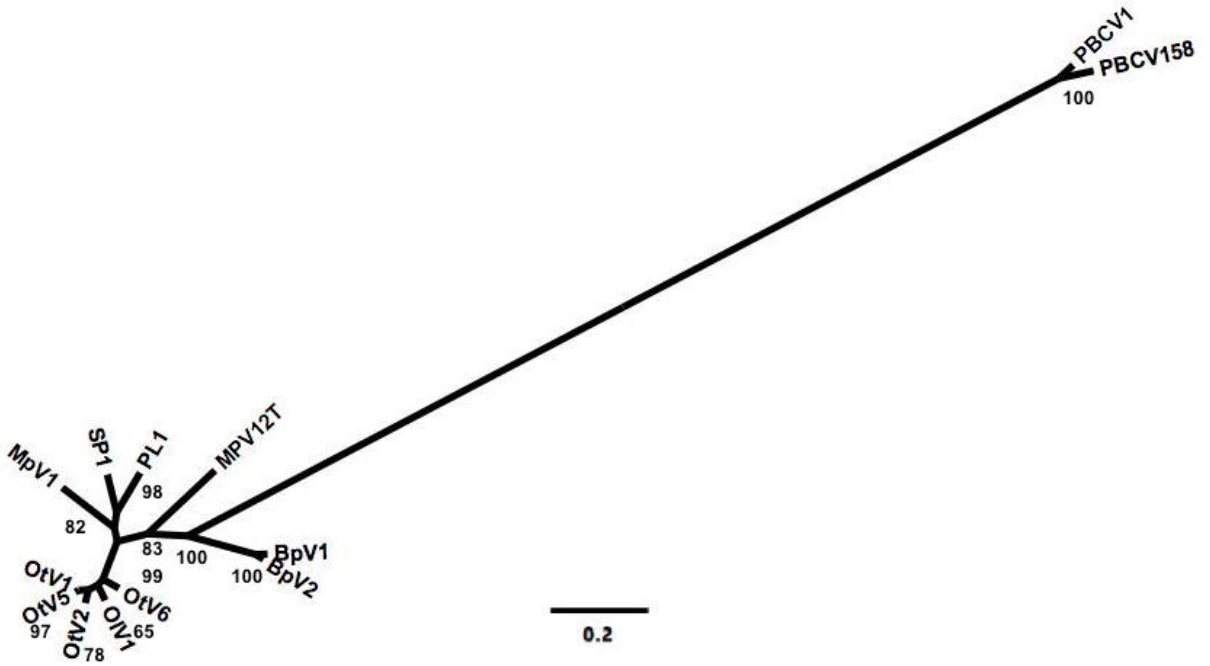

Figure 3.4: ML phylogeny of prasinoviruses and chloroviruses based on DNApol sequences.
Representing isolated viruses infecting the genera *Ostreococcus* (OtV1, OtV2, OtV5, OtV6, OlV1), *Bathyococcus* (BpV1, BpV2), *Micromonas* (MpV1, MpV-12T, MpV-PL1, MpV-SP1) and *Chlorella* (PBCV1, PBV158). The Maximum likelihood tree is based .on full-length DNApol sequences, bootstrap values based on 1000 iterations, scale bar represents substitution rate.

Table 3.3: Pairwise phylogenetic distances of prasinoviruses and chloroviruses.
Comparison of the phylogenetic distance between pairs of reference viruses based on full length DNApol (bottom left) and gene presence absence (top right). Italic numbers are number of CDS per genome. Mantel Test between the two distance matrices for all viruses and excluding chlorella viruses.

		Presence-Absence (aa id 50%)												
	CDS Clusters	MpV1	MPV12	PL1	SP1	BpV1	BpV2	OtV1	OtV2	OtV5	OtV6	OlV1	PBCV1	PBCV158
DNApol	MpV1	*244*	0.98	0.64	0.66	1.11	1.08	0.63	0.72	0.69	0.57	0.68	4.99	5.40
	MPV12	0.36	*252*	1.02	0.99	1.18	1.16	1.06	1.05	1.11	1.00	1.06	5.41	5.01
	PL1	0.26	0.36	*271*	0.30	1.12	1.10	0.71	0.75	0.75	0.67	0.73	5.04	5.45
	SP1	0.23	0.36	0.18	*244*	1.09	1.10	0.70	0.75	0.77	0.69	0.71	4.99	5.40
	BpV1	0.38	0.42	0.38	0.39	*202*	0.24	1.07	1.16	1.17	1.15	1.13	4.89	4.90
	BpV2	0.39	0.42	0.38	0.39	0.05	*209*	1.05	1.17	1.13	1.11	1.13	4.91	4.92
	OtV1	0.27	0.33	0.24	0.25	0.37	0.38	*230*	0.29	0.25	0.42	0.24	5.36	5.37
	OtV2	0.26	0.33	0.23	0.24	0.41	0.41	0.05	*235*	0.33	0.48	0.22	6.07	6.08
	OtV5	0.27	0.33	0.25	0.25	0.37	0.38	0.01	0.06	*260*	0.46	0.28	5.42	5.43
	OtV6	0.26	0.35	0.25	0.24	0.38	0.38	0.09	0.10	0.09	*249*	0.45	5.40	5.41
	OlV1	0.28	0.35	0.25	0.24	0.39	0.40	0.08	0.08	0.08	0.09	*246*	6.09	6.10
	PBCV1	2.09	2.15	2.19	2.05	2.03	2.05	2.21	2.21	2.22	2.13	2.16	*789*	0.81
	PBCV158	2.15	2.18	2.23	2.10	2.07	2.08	2.27	2.27	2.30	2.22	2.25	0.12	*806*
	Mantel Test						0.96, p=0.01							0.99, p=0.01

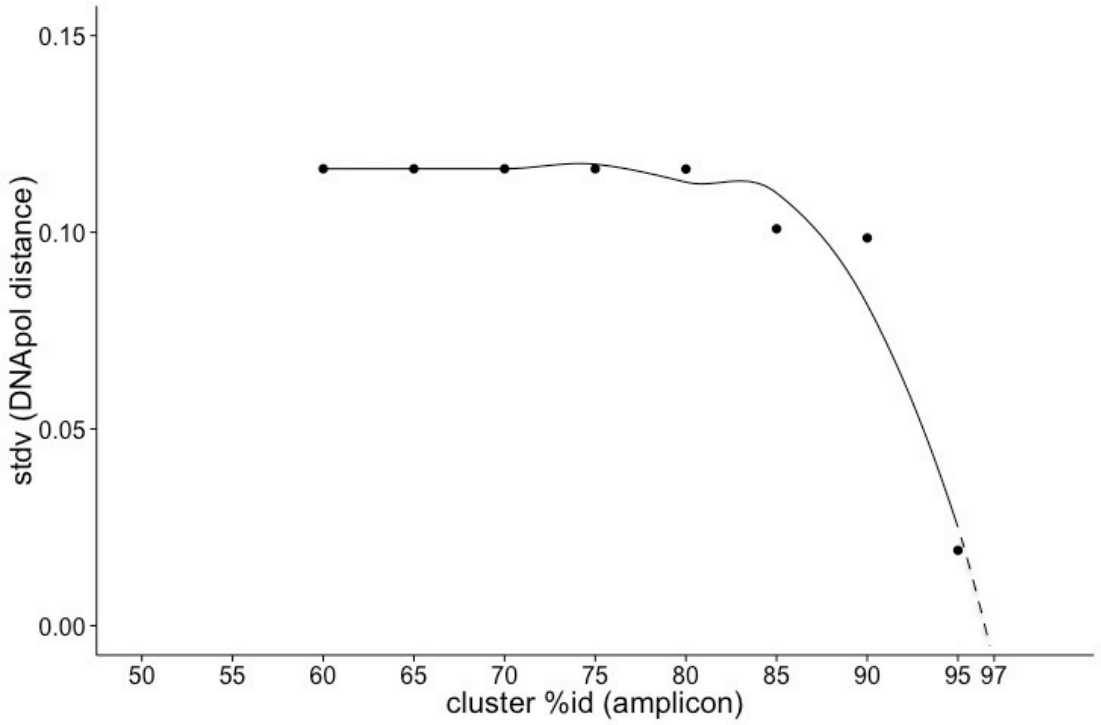

Figure 3.5: Variation in pairwise phylogenetic distance of DNApol.
Relationship between the variation in pairwise phylogenetic distance of full length reference virus DNApol sequences at different levels of % aa identity clustering of DNApol amplicons.

Prevalence of prasinoviruses is consistent with adaptation to environmental conditions

To study the distribution of prasinovirus ecotypes in the environment five environmental samples, JP, PA, JF and Saanich Inlet surface layer and deep (Figure 3.6) were compared. Samples from Saanich Inlet were collected at seven time points over the course of a year from 10 m (SI surface) 200 m (SI deep) and were analyzed as annual composites from the surface layer and the deep layer at this site (Supplementary table A1). Estimates from Aqua MODIS satellite data, averaged over 32 days around the sampling dates showed ambient SST of 18 °C, PAR of 45-50 μmol photons m^2 s^{-1} and Chlorophyll a (Chl) of 25-30 mg m^{-3} at JP and PA, while at JF the estimates were 10 °C SST, 55 μmol photons m^2 s^{-1} PAR and 10 mg chl m^{-3}. In situ salinity was 23 and PSU for PA and JF and 12 PSU for JP. Temperature and salinity for the SI surface and deep samples, measured in situ, averaged 7.5 and 9.4 °C and 30 and 31 PSU. Combined over all samples, environmental DNApol fragments from phycodnaviruses of about 129 aa length, pooled at 97 % similarity produced 197 OTUs including the references.

Phylogenetic analysis of these sequences revealed that they clustered into several groups, with most nodes being supported by bootstrap values above 75 % (Figure 3.7). The distribution of reference sequences on the tree matches the topology of trees based on gene presence or absence (Figure 3.3), and full-length DNApol sequences (Figure 3.4). Some of the environmental sequences groups were associated with sequences from prasinovirus isolates, while others were distant from known prasinovirus sequences. Moreover, the most abundant environmental sequence from each sample clustered relatively near a sequence from a prasinovirus, with the exception of the most abundant sequence from the SI deep sample, which lies on a distant branch that only contains environmental sequences. *Chlorella* viruses are on a distant branch, *Bathyococcus* viruses are clearly separated and the *Ostreococcus* viruses are clustered together. The *Micromonas* viruses MpV-PL1, MpV-SP1 and MpV1 branch closely together with SI surface and JF sequences, while MpV-12T is distant from the other *Micromonas* viruses. Numerous branches of environmental sequences are not represented by sequences from isolates. For each of the five environmental samples the dominant OTUs were placed on distinctive branches of the phylogenetic tree. Dominant OTUs in JP and SI-deep are represented on separate branches, while for PA and JF they overlap (Figure 3.7).

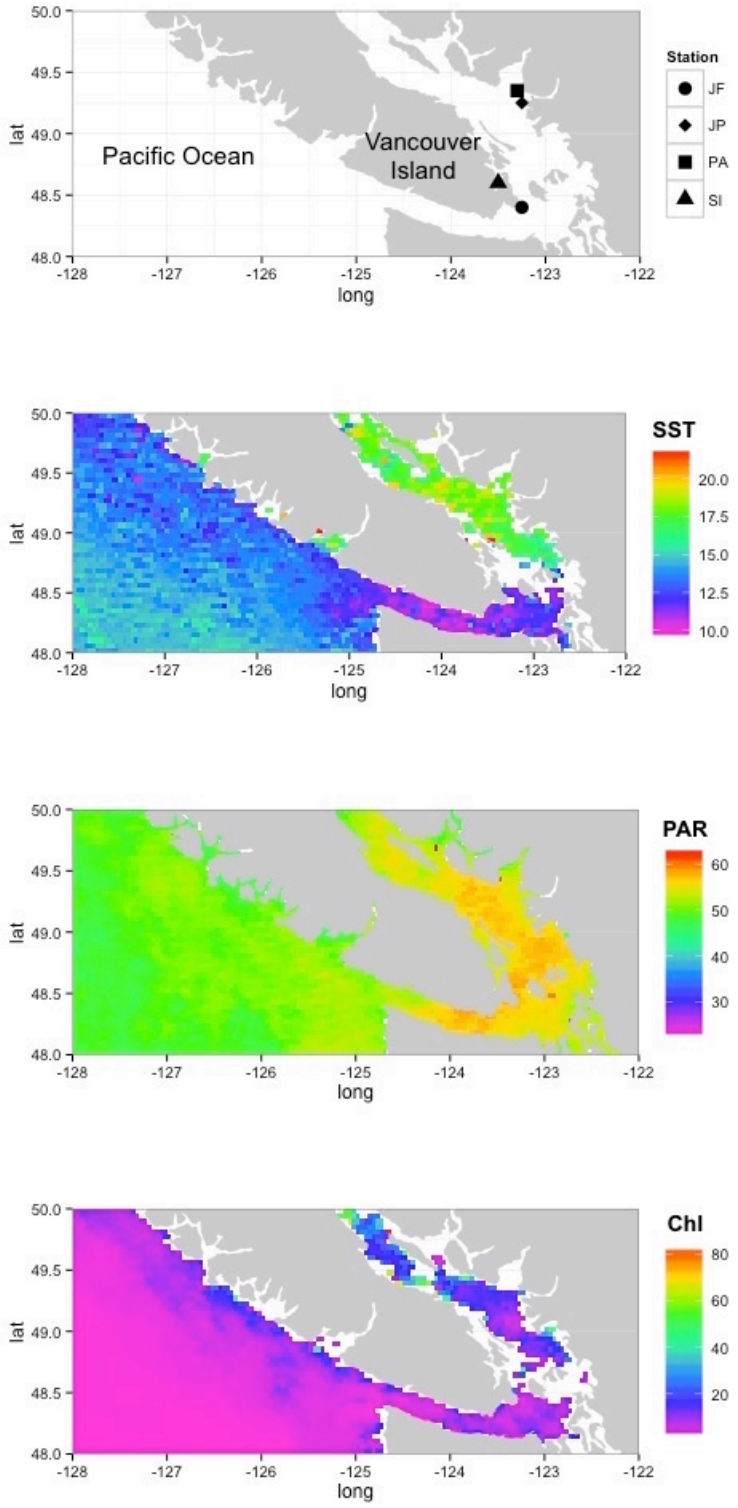

Figure 3.6: Sampling locations and in situ conditions for environmental DNApol samples.
Sampling locations for the five samples off the coast of British Columbia, Canada. Sea surface temperature (SST, °C), photosynthetically active radiation (PAR, μmol photons m^{-2} s^{-1}) and chlorophyll A (Chl, mg m^{-3}) concentration, based on 32 day composite data from the Aqua MODIS satellite.

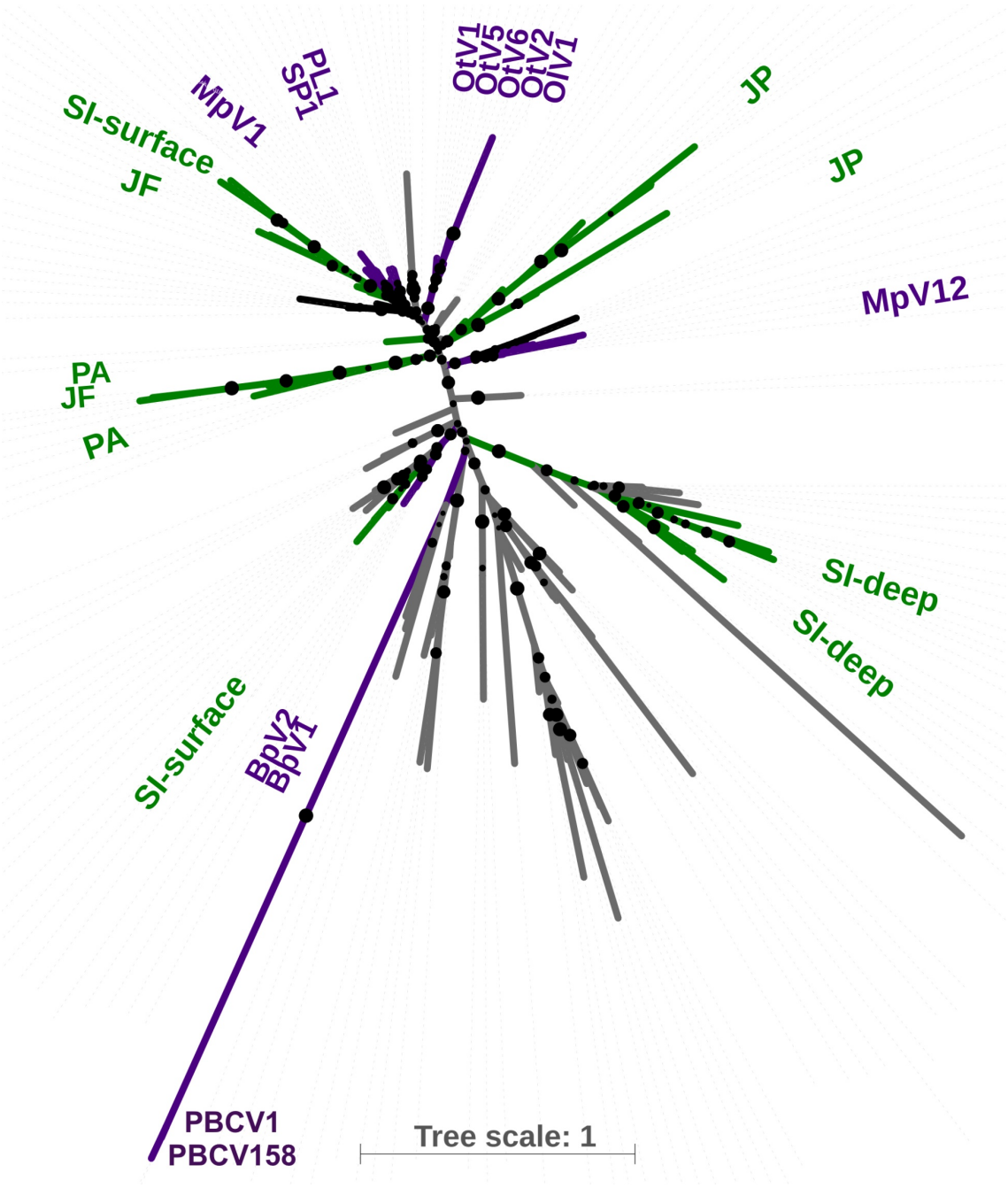

Figure 3.7: ML phylogeny of 197 phycodnavirus OTUs.
OTUs were recovered from five environmental samples of DNApol amplicons clustered at 97 % aa identity. Reference sequences are highlighted in purple, dominant OTUs and branches for the environmental samples are indicated in green JP (Jericho Pier), PA (Point Atkinson), JF (Juan de Fuca Strait), SI (Saanich Inlet). Grey branches represent rare OTUs. Bootstrap values indicate branch support, values from 50-100 % are shown as size dependent circles, the scale bar represents the substitution rate.

3.5 Discussion

This study highlights the similarities and differences among the genomes of *M. pusilla* viruses and other phycodnaviruses as well as their distribution in the environment. In particular, the results show that there is substantial overlap in the gene content among viruses infecting the genera *Micromonas*, *Ostreococcus* and *Bathyococcus*; however, there is also a large "flexible" component to their genomes. Moreover, there is considerable divergence among the *Micromonas* viruses, with the variation within these viruses being as large as it is among the sequenced prasinoviruses. Finally, an analysis of environmental DNApol sequences reveals expansive diversity of viruses closely related to prasinovirus isolates and niche-specific distribution of ecotypes in environmental samples. These findings are discussed in detail below.

Origin and distribution of genes in Micromonas viruses

The *Micromonas* viruses MpV1, MpV-PL1 and MpV-SP1 show a high degree of genome similarity to each other, as well as to *Ostreococcus* viruses in terms of the number of ORFs, ORF length, GC content and tRNAs, and in comparison to *Bathyococcus* viruses (Moreau *et al.*, 2010; Derelle *et al.*, 2015) and MpV-12T. Specifically, MpV-12T has a lower GC content and larger ORF length, and is more similar to *Bathyococcus* viruses, and was isolated on a different host strain than MpV-PL1 and MpV-SP1. Although MpV-12T has a wide host range (Martínez *et al.*, 2015), it does not infect the host of MpV-PL1 and MpV-SP1. Moreover, genomes were compared for homologues by clustering at an amino acid identity of 50%. This cut-off was selected based on identities among obvious homologues by annotation and based on the sensitivity of the UCLUST algorithm, which applies the same identity definition as BLAST (Edgar, 2010). MpV-PL1, MpV-SP1 and MpV1 share most of their genes, while more than half of the MpV-12T genome is not shared with the other *Micromonas* viruses (Figure 3.1). Having only 80 genes shared among the *Micromonas* viruses at this identity level is low relative to the seven sequenced *Ostreococcus lucimarinus* viruses (Derelle *et al.*, 2015), which share most of their genes and have pairwise nucleotide identities above 60 % for their core genes.

The identification of 80 ORFs with functional annotation that were shared among all the *Micromonas* viruses was the basis for defining a core genome among this group of

viruses, with the rest of the genes being assigned as the “flexible” pan-genome. These high similarity core genes (>50 % aa identity) were supplemented with results from an additional BLAST-P analysis. This resulted in a core genome of a combined 119 putative genes, including genes for viral replication and virion structure, as well as PhoH, which is induced under phosphate stress (Table 3.2). PhoH is widely distributed in marine phage and has been used as an alternative marker gene (Goldsmith *et al.*, 2011; Goldsmith, Parsons and Beyene, 2015) for phages and also eukaryote viruses in diversity studies, yet its exact function is not well defined. The core genes associated with viral replication are also found in *Ostreococcus* viruses, although the set of conserved genes in *Micromonas* viruses appears lower than described for *Ostreococcus* viruses (Derelle *et al.*, 2008, 2015; Weynberg *et al.*, 2009); however, it is much larger than that found in the NCLDV super group (Koonin and Yutin, 2010).

In contrast to the core-genome, there is also a shared but flexible pan-genome that varies among *Micromonas* viruses. Most of these ORFs have no functional annotation, and those that do have been seen in other prasinovirus genomes. The gene complex for amino-acid synthesis found in MpV1 and *Bathyococcus* viruses (Moreau *et al.*, 2010) is also present in MpV-PL1. Both MpV-PL1 and MpV-12T carry a copy of heat-shock protein 70 despite their otherwise limited genome overlap. Two transketolase genes, part of the Calvin cycle and pentose phosphate pathway, are only found in MpV-SP1, but have phycodnavirus homologues in metagenomes from Yellowstone Lake (Zhang *et al.*, 2015). A homologue of 6-phosphofructokinase, a key enzyme of glycolysis, is present in MpV-PL1 and MpV-SP1, similar to *Ostreococcus* viruses (Weynberg *et al.*, 2009). Given that these three genes were expressed during a transcriptional study of *M. pusilla* UTEX991 infected by MpV-SP1 (unpublished data), similarly to the expression of transaldolase, glucose-6-phosphate dehydrogenase and 6-phospohgluconate dehydrogenase in cyanophages (Thompson *et al.*, 2011), these genes may influence the host's metabolism during infection to boost viral replication. Furthermore the presence of dTDP-D-glucose 4,6-dehydratase and glycosyl transferase in MpV1, PL1 and SP1 indicates activity in nucleotide sugar manipulation and potential glycosylation of proteins, similar to findings of glycosyl transferase in the *Ostreococcus* virus OtV1 (Weynberg *et al.*, 2009). As well, UDP-glucose 6-dehydrogenase found in MpV-12T and MpV SP1 could feed products of

glycolysis into glycosylation of proteins of e.g. the capsid, similar to suggestions by Wang et al. (1993) and Weynberg et al. (2009). Altogether, these presumably cell-derived metabolic genes have the potential to be beneficial to viral production by boosting critical cell function for viral replication.

The search for host homologues of viral genes resulted in only six ORFs being shared between *Micromonas* viruses and the host strain for MpV-PL1 and MpV-SP1 at a similarity level of 50%. In contrast, *Ostreococcus* viruses share 11 genes with their host (Weynberg *et al.*, 2009), but often at lower amino-acid identities to host homologs. A further, more detailed BLAST-P analysis of ORFs in *Micromonas* viruses that have a functional annotation reveals that most have highest similarity to those typically found in other viruses, but especially that non viral hits at high similarity are to sequences from bacteria and eukaryotes that are not potential host taxa (Figure 3.2). This is similar to findings for the *Ostreococcus* virus OtV5 (Derelle *et al.*, 2008) and the Mollivirus, a NCLDV that infects *Acanthamoeba* (Legendre *et al.*, 2015). Another comparison of prasinoviruses of different hosts also revealed a pattern of shared metabolic genes with an origin outside the host range, suggesting horizontal gene transfer (Moreau *et al.*, 2010). Furthermore, horizontal gene transfer is believed to be the main mode to acquire novel genes for viruses of *Ostreococcus* and *Micromonas* (Filée, 2015), and to be beneficial for the virus (Monier *et al.*, 2009). The data presented here provide putative evidence that horizontal gene transfer from a range of sources is widespread among viruses of *Micromonas*, possibly under selection pressure to adapt to environmental conditions.

Deriving similarity in gene content from DNApol

Measuring the prevalence of viruses with specific genetic repertoires, i.e. ecotypes, in the environment poses a challenge. This problem was approached by first constructing a phylogenetic tree based on the presence and absence of genes, in order to infer how closely the viruses were related to each other (Figure 3.3) and then correlating it to the phylogeny based on full-length DNApol sequences (Figure 3.4) and PCR amplicons.

The phylogeny based on gene presence and absence data presents an overall view of the genetic similarity among the prasinoviruses and its relationship to *Chlorella* viruses. While *Ostreococcus* and *Bathyococcus* viruses form well-defined groups, the

Micromonas viruses are more scattered among the tree with MpV-12T on an isolated branch, suggesting substantial gene loss and transfer among these viruses. The relatively low bootstrap values in the gene presence and absence tree is similar to other phylogenies based on this technique (Yutin, Wolf and Koonin, 2014; Legendre *et al.*, 2015). This reflects that *M. pusilla* viruses generally share many genes, but MpV-PL1 shares more genes with OtV5 infecting *Ostreococcus* (125) than it does with MpV-12T (93) (Supplementary figure A1). Furthermore, the phylogenetic tree based on the presence and absence of genes is similar in topology to the phylogenetic relationship inferred from whole-gene DNApol sequences, as well as others based on DNApol sequences or the presence and absence of genes (Koonin and Yutin, 2010; Clerissi, Grimsley, Ogata, *et al.*, 2014; Zhong and Jacquet, 2014; Derelle *et al.*, 2015; Legendre *et al.*, 2015).

Comparing pairwise phylogenetic distances based on gene presence and absence and DNApol showed strong congruency in the Mantel Test (Table 3.3). This implies that DNApol sequences can be used to infer phylogenetic relationships among environmental sequences to assess the diversity of prasinoviruses in environmental samples as has been done (Short and Suttle, 2002; Clerissi, Grimsley, Ogata, *et al.*, 2014) and that it is a strong proxy to infer the similarity in gene content among prasinoviruses.

However, because PCR only amplifies a gene fragment, sequences need to be clustered at 97 % amino-acid identity to be specific to full length DNApol sequences. This is less stringent than Short and Short (2008), clustering at 97 % at the nucleotide level and Bellec et al. (2010) who considered differences by single nucleotides as defining a distinct *Ostreococcus* virus haplotype. In contrast, it is more stringent than clustering at 75 % identity which was used in another study on prasinovirus distribution (Clerissi, Grimsley, Subirana, *et al.*, 2014).

Prevalence of prasinoviruses is consistent with adaptation to environmental conditions

With a framework to infer the phylogenetic relationship and similarity in genetic repertoire among prasinoviruses based on DNApol amplicons, the approach was used to determine how well represented the sequenced prasinoviruses were across environmental samples. Since the four reference *Micromonas* viruses examined in this study were isolated from

widely separated geographic areas. MpV-SP1 and MpV-PL1 were isolated from water collected from Scripps Pier, San Diego CA, and Port Aransas, TX, respectively (Cottrell and Suttle, 1991), MpV1 was isolated from an eutrophic coastal lagoon in the northwestern Mediterranean (Moreau *et al.*, 2010), and MpV-12T was isolated off of the Dutch coast (Martínez *et al.*, 2015). Although *Micromonas* viruses occur in the coastal waters of British Columbia (Mayer and Taylor, 1979; Cottrell and Suttle, 1991), none of the sequenced isolates were from the region; hence, it was unknown if these genotypes would be well represented in these waters.

Five environmental samples from British Columbia coastal waters that reflect a range of conditions were analyzed for prasinovirus ecotypes and in situ conditions. Saanich Inlet is productive and stratified in spring and summer, and is isolated from deeper waters beyond the inlet because of a shallow sill; this leads to hypoxic deeper waters (Zaikova *et al.*, 2010). JP is strongly stratified, with a fresh water influence from English Bay that is adjacent to the city of Vancouver, while PA is more exposed and mixed with a higher salinity. JF is off the coast of Victoria in very exposed and mixed waters of the Juan de Fuca Strait (Masson and Pena, 2009). This is described by the prevailing salinity, temperature and chl concentrations at the sampling locations (Figure 3.6). While JP and PA are similar in their high SST of 18 °C and chl concentrations, JF is a much deeper mixed water body with a SST of only 10 °C and lower chl concentration. However, PA is more similar to JF in terms of salinity with both being 23 PSU, characteristic for well mixed and exposed coastal environments. The combined DNApol sequences from all samples produced 197 distinct OTUs which were used to build a diverse and well supported Maximum-Likelihood DNApol tree displaying the prasinovirus and chlorovirus diversity.

The multitude of well defined branches on the DNApol tree suggest a large diversity in prasinovirus ecotypes illustrates their distribution across environments (Figure 3.7). The distribution of reference viruses on the tree generally reflects the tree topology of the reference trees based on full-length DNApol sequences and the presence and absence of genes, confirming the approach. The environmental OTUs substantially increases the known richness of prasinoviruses and especially *Micromonas* viruses in the environment. Furthermore, the specific distribution of the representative OTUs for each of the five

environments suggests a specialization of the corresponding viral ecotypes to prevailing conditions. The Saanich Inlet samples, being long term integrated samples, should rather be seen in comparison to each other than to the other three samples. Despite the AquaMODIS data showing JP and PA being similar in temperature and JP, PA, JF having similar chl concentrations and PAR levels, PA and JF are more similar environments based on their *in situ* salinities and presumed mixing. This is also reflected in the dominant prasinovirus genotypes for the samples. Saanich Inlet deep sequences and the stratified, near shore JP sequences are on separate isolated branches. The dominant sequences in Saanich Inlet surface and especially the two mixed, more saline PA and JF samples share branches. This specialization of viruses to environments is congruent with findings that prasinovirus communities in the Northwest Mediterranian Sea are affected by environmental variables and especially nutrient availability (Clerissi, Grimsley, Subirana, *et al.*, 2014). Also, considering the relatively wide host range of these viruses within a genus (Cottrell and Suttle, 1991; Martínez *et al.*, 2015) the pattern likely represents a response by the prasinovirus community to the specific environmental conditions and not solely the host community. Altogether this could mean that prasinovirus ecotypes with similar genetic repertoires, approximated by DNApol similarity, dominate in similar environments.

In conclusion, this research highlights the genetic repertoire encoded by prasinoviruses infecting *M. pusilla* and other prymnesiophytes. We identified a core set of genes that are shared among *Micromonas* viruses despite their marked differences, and identified a large set of genes that make up a flexible part of the genome, implying that there is a large “pangenome” that is shared among prasinoviruses. Furthermore we set the *Micromonas* virus genomes in contrast to genomes of other prasinoviruses, phycodnaviruses and a host genome elucidating overlap in genetic repertoire. The presumed origin of shared genes and their distribution across viral clades shows a complex evolutionary history and horizontal gene transfer.

Chapter 4: Environmental variables shape cyanomyovirus communities

4.1 Summary

The globally distributed and numerically dominant cyanobacterial genera *Synechococcus* and *Prochlorococcus* account for a large proportion of the world's primary production. They are infected and lysed by viruses (cyanophages), a process that influences cyanobacterial diversity, as well as carbon and nutrient cycling. Known cyanophages belong to one of the families *Myoviridae*, *Podoviridae* and *Siphoviridae.* Their dsDNA genomes include a number of host-derived auxiliary metabolic genes (AMGs) which are expressed and potentially facilitate viral replication. However, it is not well understood how the variations in genetic repertoire interact with environmental variables to influence viral selection and shape cyanomyovirus communities.

This project aims to correlate the genetic repertoire of cyanomyoviruses with the marker gene *gp43* to investigate cyanomyovirus ecotype distribution as a function of environmental conditions across locations and seasons.

The marker gene phylogeny could be correlated to similarity in genetic repertoire. The data revealed spatial and seasonal patterns in the viral community which are related to environmental variables. The mixing regime of the water column as defined by salinity and temperature, and the associated nutrient availability, proved to be significant predictors of cyanomyovirus richness, diversity and community composition. Since environmental variables shape viral communities from the resident viral seed bank, different seed banks will respond differently to environmental changes. It is evident that environmental variables do shape cyanomyovirus communities and that viral ecotypes with corresponding genetic repertoires underlie selection pressure. However, all the mechanisms involved in viral selection remain to be fully understood.

The data further the understanding of the temporospatial variation of cyanomyovirus community compositions, their genetic repertoire, and their relationship with environmental variables. This understanding helps to better predict the response and ecological impact of cyanomyoviruses against a backdrop of environmental change.

4.2 Introduction

Cyanobacteria are globally distributed and abundant pico-phytoplankton that are estimated to account for around 25% of the world's primary production (Li, William, 1994; Liu, Nolla and Campbell, 1997; Field *et al.*, 1998; Partensky, Hess and Vaulot, 1999; Weigele *et al.*, 2007). In the oceans, they are numerically dominated by members of the genera *Synechococcus* and *Prochlorococcus* (Liu *et al.*, 1998). Cyanobacteria are infected and lysed by viruses (cyanophages), a process that influences their diversity, as well as carbon and nutrient cycling (Wilhelm and Suttle, 1999; Mühling *et al.*, 2005; DeLong *et al.*, 2006; Weitz and Wilhelm, 2012).

Cyanophages infecting members of the *Synechococcus* spp. and *Prochlorococcus* spp. have circular, dsDNA genomes (Suttle, 2000) and appear to be primarily lytic (Marston and Sallee, 2003; Clokie and Mann, 2006). All cyanophages isolated to date belong to the order *Caudovirales* (Weinbauer and Rassoulzadegan, 2003) and, depending on morphology, are assigned to one of the three families *Myoviridae*, *Podoviridae* or *Siphoviridae*. Cyanophages within the family *Myoviridae*, are referred to as cyanomyoviruses, many of which are genetically related to the coliphage, T4, and are referred to as T4-like phages. It is this group of viruses on which this project focuses. T4-like phages are characterized by having long contractile tails and a comparably large host range that can span across genera (Suttle and Chan, 1993; Lu, Chen and Hodson, 2001; Sullivan, Waterbury and Chisholm, 2003; Lindell *et al.*, 2004). Their genome sizes range from 150 kb to >200 kb and include a number of host-derived auxiliary metabolic genes (AMGs) (Breitbart *et al.*, 2007; Clokie, Millard and Mann, 2010). These AMGs include genes coding for proteins involved in photosynthesis (PsbA, PebS), nutrient uptake (PhoH, PstS, NifU) and carbon metabolism (TalC) (Mann *et al.*, 2003; Sullivan *et al.*, 2005; Dammeyer *et al.*, 2008; Williamson *et al.*, 2008). Studies have shown that these AMGs are expressed and benefit viral replication (Bragg and Chisholm, 2008; Thompson *et al.*, 2011). An *in silico* model with phages of *Prochlorococcus* showed a fitness advantage for phages carrying AMGs for photosynthesis (Hellweger, 2009). The differences in genetic repertoire and the associated adaptations to niches diversifies cyanophages into ecotypes (Marston and Martiny, 2016). Most genes in cyanomyoviruses are not functionally

annotated, yet have the potential to also be beneficial during replication under specific conditions.

Virus replication is a resource-intensive process with a high demand for nutrients compared to cellular organisms (Jover *et al.*, 2014), and several environmental factors have been shown to influence viral replication or degradation. For example, phosphate availability affects viral production (Suttle and Chen, 1992; Wilson, Carr and Mann, 1996; Sullivan *et al.*, 2005; Jia *et al.*, 2010) and a putative gene that is associated with phosphate stress was predominantly found in cyanophage isolates from nutrient limited regions (Sullivan *et al.*, 2010), indicating selection pressure. As well, light affects adsorption and infection by cyanophages, but UV radiation affects viral degradation and DNA damage (Suttle and Chen, 1992; Noble and Fuhrman, 1997; Weinbauer *et al.*, 1999). Hence, variation within the genetic repertoire of cyanophages may result in some viral genotypes being more fit than others under different environmental conditions, thus shaping the genetic composition of cyanophage communities. However, the relationship among the genetic repertoire and community composition of viruses, and environmental conditions is not well understood and challenging to study.

Over the years several genes have been used as markers to describe the biogeography and seasonal patterns of cyanomyovirus communities. The earliest and most extensively used marker gene for cyanomyoviruses is the capsid assembly protein gp20. Based on sequence analysis of isolates, PCR primers were designed that targeted cyanomyovirus *gp20* sequences. Using these primers and a DGGE fingerprint analysis showed that *gp20* sequences changed dramatically along a south-north transect in the surface water of the Atlantic ocean, but showed less variation with depth (Wilson *et al.*, 1999). Similar approaches revealed that community changes can also occur across small distances and depths, as well as across seasons (Frederickson, Short and Suttle, 2003; Wang and Chen, 2004). Both patterns were associated with changes in the physical environments and host communities. In another seasonal study, Mühling et al. (2005) examined changes in *gp20* sequences and cyanobacterial diversity in the Red Sea and concluded that cyanomyoviruses controlled the composition of the host communities. A similar seasonal pattern was shown in Norwegian coastal waters by Sandaa and Larson (2006) that were also associated with surprisingly large variations in cyanomyovirus

genome size. However, despite pronounced seasonal and spatial variation in cyanomyovirus communities, some genotypes are widely distributed across sharply different environments (Short and Suttle, 2005). There are also concerns about using *gp20* sequences to estimate cyanomyovirus diversity. McDaniel et al. (2006) revealed that *gp20* sequences could only be amplified from about 60% of cyanomyovirus isolates. As well, Short and Suttle (2005) amplified *gp20* sequences from samples collected at 2.5 km deep water, a biome not expect to be dominated by myoviruses infecting cyanobacteria. Consequently, amplified *gp20* sequences may not be suitable for examining diversity changes in T4-like cyanophages.

An alternative marker gene for T4-like phages is *gp23* that encodes the major capsid protein. The gene is highly conserved, and has been used to look at a wide diversity of myoviruses infecting a broad range of hosts (Tetart *et al.*, 2001). The diversity of *gp23* sequences across marine and fresh water environments is extensive (Filée *et al.*, 2005; Comeau and Krisch, 2008; Butina *et al.*, 2010), and a study with high sampling frequency on myoviruses and bacteria showed resilience of taxa and a covariation between viruses and bacteria if a two-day time lag was incorporated (Needham *et al.*, 2013). Moreover, *gp23* data show that while there is persistence of some OTUs there are strong seasonable patters in the composition of viral communities (Chow and Fuhrman, 2012). However, because *gp23* is highly conserved across a broad range of phages, it is less useful for specifically targeting cyanophages (Chow and Fuhrman, 2012).

The AMGs *psbA* and *phoH* have also been used as marker genes for a broader range of viruses. The photosynthesis protein PsbA is common in phages infecting *Synechococcus*, and can be used to distinguish freshwater and marine cyanophages (Mann *et al.*, 2003; Chénard and Suttle, 2008). An advantage of using *psbA* as a marker is that it occurs in myoviruses and podoviruses, but being a host-derived gene the phylogeny potentially reflects more on the origin of the gene than the viral phylogeny. The phosphate stress-induced protein PhoH is present in the genomes of several groups of viruses and can be found in marine samples across a range of locations and depths. Moreover, viral and host sequences can be clearly distinguished from each other (Goldsmith *et al.*, 2011). Community composition assessed with *phoH* shows that some viruses persist, but also that there is variation across depths and seasons (Goldsmith,

Parsons and Beyene, 2015). Patterns that were again confirmed for several types of viruses based on *gp23* and *phoH* (Goldsmith *et al.*, 2015). Both *psbA* and *phoH* are useful marker genes, but are neither essential nor exclusive to cyanomyoviruses and thus difficult to use in studying this group of viruses specifically.

The DNA polymerase gene, *gp43*, is relatively new as a marker gene for myoviruses, and while the primers were not designed to be cyanophage specific they amplify cyanomyoviruses well (Marston and Amrich, 2009). An extensive study of cyanophage isolates from a range of locations and times, showed that there was great diversity, seasonality and geographic variability in gp43 sequences (Marston *et al.*, 2013). The data also highlighted stark differences in community composition between contrasting environments, and that seasonal composition varied more gradually. Marston and Martiny (2016) have furthermore used gp43 to describe temporal patterns for cyanomyovirus isolate ecotypes.

One problem marker genes generally share is how well they reflect overall genome composition. Phylogenies based on comparing the presence and absence of genes have been used to compare closely related viral families and are the best way to compare the overall gene content of viruses (Yutin, Wolf and Koonin, 2014). However, this approach requires full genome sequencing and annotation of isolates and thus cannot be used to study complex natural communities.

The research described here establishes a correlation between the genetic repertoire of cyanomyoviruses and the marker gene *gp43*. This correlation is then used to interpret the distribution of cyanomyovirus genotypes and their genetic repertoire as a function of environmental conditions across locations and seasons. This data reveals a spatial and a seasonal pattern in the viral communities related to differences in environmental variables. Understanding the relationship between environmental variables and viral community composition can be used to help predict the response and ecological impact of cyanomyoviruses against a backdrop of environmental change.

4.3 Materials and methods

Sampling

Samples were taken for 12 months from the mixed surface layer and at 10 meters depth in Saanich Inlet, and throughout the mixed layer at 18 sites in the Strait of Georgia over the course of three years. Twenty to 200 liters of water were sampled with Niskin bottles (General Oceanics, Miami, FL) and processed immediately or stored at 4 °C in the dark until processing within 24 h. Samples were pre-filtered through 2.7 µm nominal pore-size GF/D glass-fiber filters (Whatman GE Health Care, Little Chalfont, UK) and 0.22 µm pore-size Sterivex filters (Merck Millipore, Billerica, MA) for Saanich Inlet samples and 47-mm diameter 0.7 µm pore-size GC50 glass-fiber and 0.45 µm pore-size HVLP filters (Merck Millipore, Billerica, MA) for the Strait of Georgia samples. The remaining virus-size particulate matter in the samples was then concentrated to 500 ml volume by tangential flow ultrafiltration (TFF, Prep-Scale) with a 30 kDa cutoff (Merck Millipore, Billerica, MA). Viral concentrates (VCs) were stored at 4 °C until further use.

Environmental data collection and processing

Depth profiles of temperature and salinity were measured with a rosette mounted or cable deployed CTD SBE 25 (Seabird Electronics Inc., Bellevue, WA). Chlorophyll concentration was estimated using a fast repetition rate fluorometer (FRRF), and oxygen concentration was measured by a SBE 43 oxygen sensor (Seabird Electronics). Photosynthetically active radiation (PAR) was measured with a QSP-200PD sensor (Biospherical Instruments, San Diego, CA).

On board, nutrient samples were filtered through 0.22 µm pore-size PVDF syringe filters and the filtrate stored at -20 °C for later analysis with a Bran & Luebbe AutoAnalyzer 3 (SPX-Flow, Norderstedt, Germany) using air-segmented continuous-flow analysis. Combined nitrate (reduced to nitrite) and nitrite, silicate and phosphate were measured by absorbance following established protocols (Murphey and Riley, 1962; Armstrong, Stearns and Strickland, 1967).

Viral and bacterial abundances were measured using a Beckton Dickinson FACSCalibur flow cytometer with a 15 mW 488 nm air-cooled argon ion laser. Samples were fixed for 15 min at 4 °C in the dark with 25 % electron-microscopy grade

glutaraldehyde (final concentration 0.5 %), followed by snap-freezing in liquid nitrogen and storage at -80 °C. Prior to measurement, samples were thawed and diluted in 0.2 µm filtered, autoclaved TE 10:1 buffer (10 mM-Tris HCl; 1 mM EDTA pH 8.0) and stained with SYBR Green I (Invitrogen, Carlsbad, CA) at a final dilution of $0.5x10^{-4}$ of the commercial stock, incubated for 10 min at 80 °C or 15 min at room temperature for viruses and bacteria respectively (Brussaard, 2004). Samples were diluted in TE buffer (pH 8.0) to ascertain 100 to 1000 events s^{-1}. Viruses and bacteria were discriminated by plotting green fluorescence against SSC signals and data was analyzed with CYTOWIN version 4.31 (Vaulot, 1989) and WEASEL version 3.3 (Battye, 2015).

DNA extraction, PCR and sequencing library preparation

For DNA extraction, 25 ml of VC were syringe filtered through 0.22 µm pore-size GV PVDF Millex filters (Merck Millipore, Billerica, MA) and centrifuged for 6 h at 120,000 g and 8 °C. The supernatant was discarded and viral pellets were resuspended in 500 µl TE buffer at 4 °C over night. Free DNA was treated with 5 µl DNase I (Invitrogen, Carlsbad, CA) at 37 °C for 15 min and inactivated with 10 µl EDTA (0.25M) at 65 °C for 15 min. Viral capsids were lysed with 60 µl Proteinase K (Invitrogen, Carlsbad, CA) at 56°C for 15 min, viral DNA was extracted with Pure Link Viral RNA/DNA columns (Invitrogen, Carlsbad, CA) following the manufacturer's instructions and eluted in UltraPure water (Invitrogen, Carlsbad, CA).

To ensure an equal amount of template in the PCR DNA was first quantified with a Qubit 2.0 using the dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA). Viral DNA polymerase gene (*gp43*) fragments of about 475 bp length were amplified by PCR using primers from Marston et al. (2013). For each sample 1-2 ng of template DNA were used in a two-step, large scale PCR with a total of 35 cycles. PCR conditions were an initial denaturing step at 94 °C for 3 min, denaturing at 94 °C for 45 s, annealing at 50 °C for 45 s, extension at 72 °C for 45 s and a final extension at 72 °C for 10 min. Triplicates PCR products were pooled and run on a 0.8% Ultrapure LMP Agarose gel (Invitrogen, Carlsbad, CA). Bands in the appropriate size range around 475 bp were excised and the DNA was extracted with the Zymoclean Gel DNA Recovery Kit (Zymo, Irvine, CA) and

eluted in UltraPure water (Invitrogen, Carlsbad, CA). DNA products were quantified by Qubit, aliquoted and stored at -20 °C.

For library preparation, 500 ng of DNA product was used with the NxSeq Low DNA AmpFREE kit (Lucigen, Middleton, WI) following the manufacturer's protocol with NextFlex-96 sequencing adapters (Bioo, Austin, TX). Libraries were purified and size selected (~600 bp) with Agencourt AMPureXP beads (Beckman Coulter, Pasadena, CA) and eluted in low TE. Library construction was confirmed with a Bioanalyzer 2100 using High Sensitivity DNA Chips (Agilent, Santa Clara, CA).

Sequencing

For pooling, libraries were quantified by Q-PCR with SSoFast Eva Green Supermix (BioRad, Hercules, CA) and KAPA DNA Standard (KAPA Biosystems, Boston, MA) on a C10000 Touch PCR block with a CFX 96 head (BioRad, Hercules, CA), and pooled for equal template concentration. Libraries were sequenced in two rounds at the UCLA (Los Angeles, CA) and McGill (Montreal, QC) sequencing facilities using 2x300 HiSeq paired-end technology (Illumina, San Diego, CA).

Bioinformatic processing

Sequences were trimmed using TRIMMOMATIC 0.33 (Bolger, Lohse and Usadel, 2014), applying a quality thread score of 30 and a minimum length of 36 nt. Paired reads were merged with USEARCH 8.1 (Edgar, 2010), translated with FragGeneScan 1.20 (Rho, Tang and Ye, 2010), cleaned up and size selected for a minimum length of 140 amino acids. Reads from all samples were pooled and dereplicated, and then clustered and chimera tested with USEARCH 8.1 at 97 % amino-acid identity. Singletons were removed and OTUs were selected for cyanophage similarity using BLAST-P with a cut-off E-value of 10^{-3}. Per sample, reads were parsed to representative OTUs with UPARSE 8.1 (Edgar, 2013) at an amino-acid identity of 97 %.

A whole genome reference phylogeny was built from fully sequenced and annotated genomes of 19 cyanomyoviruses that were retrieved from NCBI. Genome accession numbers are S-SM1, NC_015282; S-SM2, NC_015279; S-ShM2, NC_015281; S-SSM7, NC_015287; S-SSM5, NC_015289; S-PM2, NC_006820; S-RSM4,NC_013085;

Syn1, NC_015288; Syn9, NC_008296; Syn19, NC_015286; Syn33, NC_015285; P-HM1, NC_015280; P-HM2, NC_015284; P-SSM2, NC_006883; P-SSM7, NC_015290; P-SSM4, NC_006884; P-RSM1, NC_021071; P-RSM4, NC_015283; P-TIM40, NC_028663. CDS were clustered with USEARCH 8.1 at a 50% amino-acid identity. Phylogenetic distances between genomes calculated based on the presence or absence of genes with the formula Dij=-ln(Sij/sqrt(Ni*Nj)), with Sij, Ni and Nj as the number of shared genes, number of genes in one genome and number of genes in the other genome, respectively (Yutin, Wolf and Koonin, 2014) and a Neighbor-Joining (NJ) tree was built in the ape package (Paradis, Claude and Strimmer, 2004) in R (R, 2015). Amino-acid sequences for gp43 were extracted from the genomes above and aligned in Clustal (Sievers *et al.*, 2011), and a Maximum-Likelihood (ML) tree was built in RaxML 8.0.0. (Stamatakis, 2014) using the WAG substitution model, with the optimal substitution model selected in prottest-3.4 (Darriba, Taboada and Posada, 2011). A Mantel Test to compare distance matrices based on gene presence/absence vs. gp43 was performed in the ade4 package (Dray and Dufour, 2007) using R. Identities of partial DNA polymerase sequences from reference viruses were determined in RaxML 8.0.0., their specificity to distances of full length sequences at different identity levels compared in R. Environmental OTUs were placed on the full length gp43 reference tree by the Evolutionary Placement Algorithm (EPA) in RaXML 8.0.0., and the phylogenetic tree was edited using iTOL v3.2.4 (Letunic and Bork, 2016), clades were defined by eye.

Statistical analyses

Statistical analyses were performed in R. Scaling and principal component analysis (PCA) analysis of the environmental data was done using the FactoMineR package (Le, Josse and Husson, 2008). Parsed reads per sample were rarefied to the lowest read number of the samples and the VEGAN package (Oksanen *et al.*, 2016) was used to determine diversity indices, and conduct the principal coordinate analysis (PCoA) and canonical correspondence analysis (CCA). Environmental classification and subsequent indicator species analysis were done in the IndecSpecies package (Caceres and Legendre, 2009).

4.4 Results

Deriving similarity in gene content from gp43 sequences

To assess how the phylogenetic analyses based on gp43 sequences reflect the overall gene content of T4-like cyanomyoviruses the gene content of 19 reference cyanomyoviruses was compared with CDSs being clustered at 50 %. This produced a Neighbor-Joining (NJ) phylogenetic tree with well supported branches based on gene content (Figure 4.1). In comparison, the tree displays a similar architecture to a Maximum-Likelihood (ML) tree based on full-length DNA polymerase gene sequences (Figures 4.2).

A Mantel Test of the pair wise distances among reference viruses between the gene content and the DNA polymerase phylogeny proved significant congruency at 0.87 (Table 4.1). Furthermore, when the pairwise phylogenetic distances of the full-length reference gp43 sequences were compared to distances determined from the fragments amplified by the PCR primers, the variation in full length gp43 phylogenetic distance approached zero when gp43 amplicons were clustered above a 95% amino-acid (aa) identity level (Figure 4.3). In a one-on-one comparison of reference virus amplicons the highest pairwise identity for these sequences was 96.4%. Accordingly, the environmental amplicons were clustered at 97% aa identity.

Spatial and temporal variation in environmental reads and sampling conditions

A total of 42 environmental samples were taken during 2010, 2011 and 2012. Of these, 18 were in the Strait of Georgia and adjacent waters (SOG), 24 were taken in Saanich Inlet (SAA) over a 12-month period from the mixed surface layer (2-5 m) and at 10 m depth (Figure 4.4). The 18 samples from SOG were from 13 locations; seven were from open waters between Vancouver Island and the mainland, five were from inlets and one from Queen Charlotte Sound, off the northern tip of Vancouver Island. For the SOG samples six discrete samples were collected from the surface through the mixed layer to the subsurface chlorophyll maximum, and combined to provide an integrated sample, representative of the mixed layer; the bottom sampling depth varied from eight to 18 m. The 24 samples from SAA represent a seasonal cycle at two depths over one year.

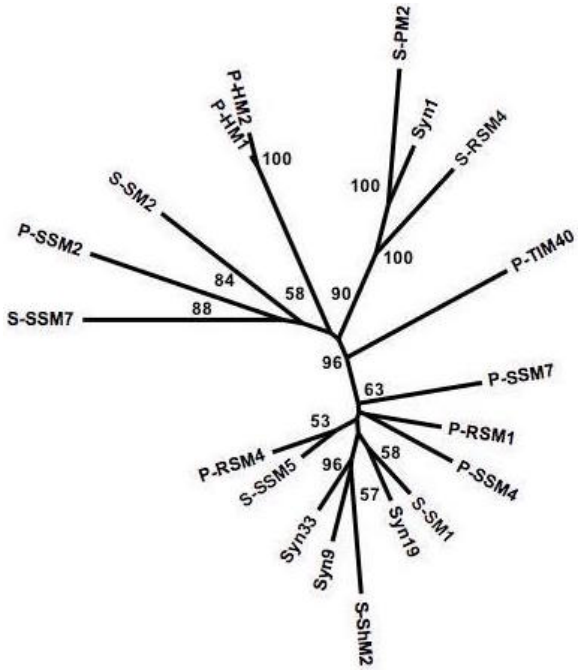

Figure 4.1: NJ phylogeny of reference cynaomyoviruses based on gene content.
Phylogenetic distance is based on the gene content, ORFs clustered at 50% aa identity. Bootstrap values over 50 % are shown.

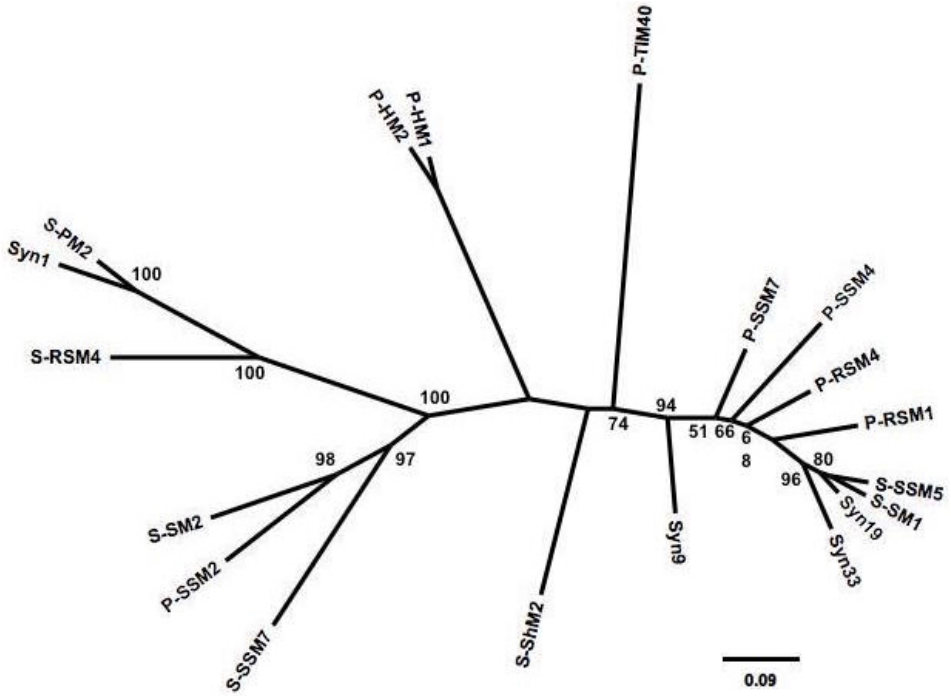

Figure 4.2: ML phylogeny of reference cynaomyoviruses based on gp43 sequences.
Phylogenetic distance is based in the full length gp43 amino acid sequences. Bootstrap values above 50 % are shown, scale is the substitution rate.

Table 4.1: Pairwise phylogenetic distances of cyanomyoviruses.
Phylogenetic distance is based on gene content and full length gp43. Congruence was tested with a Mantel Test. Numbers of CDS per virus used in the cluster analysis are shown along the diagonal in italics.

		Presence-Absence (aa id 50%)																		
	CDS Clusters	S-SM1	S-SM2	S-ShM2	S-SSM7	S-SSM5	S-PM2	S-RSM4	Syn1	Syn9	Syn19	Syn33	P-HM1	P-HM2	P-SSM2	P-SSM7	P-SSM4	P-RSM1	P-RSM4	P-TIM40
gp43	S-SM1	*234*	1.44	1.05	2.05	0.44	1.97	1.38	1.63	0.76	0.52	0.75	1.75	1.86	2.00	1.01	0.96	0.81	0.71	1.48
	S-SM2	0.67	*267*	1.62	1.70	1.63	2.01	1.64	1.59	1.47	1.49	1.65	1.82	1.95	1.69	1.94	1.96	1.71	1.91	1.88
	S-ShM2	0.44	0.63	*229*	2.07	1.09	1.96	1.58	1.35	0.85	1.03	0.77	1.87	1.90	2.13	1.34	1.30	1.19	1.29	1.53
	S-SSM7	0.82	0.49	0.73	*317*	1.89	2.25	1.98	2.08	2.00	1.95	2.00	1.67	1.84	1.71	1.78	1.77	1.65	1.64	1.94
	S-SSM5	0.10	0.67	0.44	0.81	*223*	1.98	1.59	1.60	0.83	0.54	0.77	1.68	1.81	1.90	0.92	0.89	0.70	0.47	1.40
	S-PM2	0.86	0.70	0.77	0.74	0.85	*239*	1.25	0.85	1.92	1.99	1.95	2.01	2.08	2.31	2.04	2.03	2.02	1.94	1.97
	S-RSM4	0.77	0.55	0.81	0.70	0.78	0.40	*237*	1.00	1.38	1.55	1.55	1.81	1.98	2.08	1.91	1.85	1.75	1.83	1.82
	Syn1	0.91	0.75	0.86	0.76	0.88	0.16	0.44	*234*	1.41	1.58	1.32	1.80	1.94	2.20	1.97	1.90	1.74	1.77	1.74
	Syn9	0.24	0.56	0.35	0.74	0.25	0.77	0.66	0.83	*224*	0.74	0.68	1.78	1.89	2.09	1.10	1.08	0.99	1.01	1.52
	Syn19	0.09	0.69	0.44	0.84	0.08	0.86	0.77	0.89	0.26	*215*	0.68	1.74	1.79	1.95	1.04	0.83	0.71	0.91	1.42
	Syn33	0.15	0.68	0.41	0.82	0.16	0.84	0.80	0.85	0.26	0.14	*224*	1.81	1.81	1.92	0.92	0.93	0.76	0.91	1.58
	P-HM1	0.63	0.68	0.66	0.66	0.64	0.82	0.83	0.82	0.57	0.67	0.64	*240*	0.20	1.80	1.62	1.31	1.47	1.38	1.66
	P-HM2	0.63	0.68	0.68	0.70	0.64	0.83	0.85	0.85	0.58	0.66	0.65	0.10	*241*	1.91	1.72	1.33	1.55	1.47	1.75
	P-SSM2	0.67	0.33	0.64	0.47	0.64	0.76	0.66	0.76	0.60	0.68	0.66	0.69	0.66	*334*	1.31	1.86	1.80	1.90	1.92
	P-SSM7	0.22	0.64	0.46	0.73	0.22	0.77	0.71	0.82	0.25	0.23	0.23	0.54	0.52	0.59	*237*	1.02	0.93	0.96	1.45
	P-SSM4	0.29	0.67	0.49	0.72	0.28	0.79	0.76	0.79	0.34	0.29	0.29	0.57	0.56	0.61	0.24	*219*	0.75	0.71	1.39
	P-RSM1	0.20	0.66	0.50	0.78	0.20	0.81	0.76	0.84	0.30	0.19	0.22	0.61	0.62	0.65	0.24	0.28	*212*	0.73	1.31
	P-RSM4	0.19	0.65	0.49	0.74	0.18	0.77	0.76	0.81	0.28	0.21	0.24	0.54	0.53	0.63	0.19	0.25	0.20	*236*	1.41
	P-TIM40	0.57	0.76	0.63	0.84	0.56	0.91	0.86	0.95	0.59	0.58	0.60	0.75	0.74	0.73	0.54	0.62	0.58	0.55	*233*
	Mantel Test	0.87, p=0.01																		

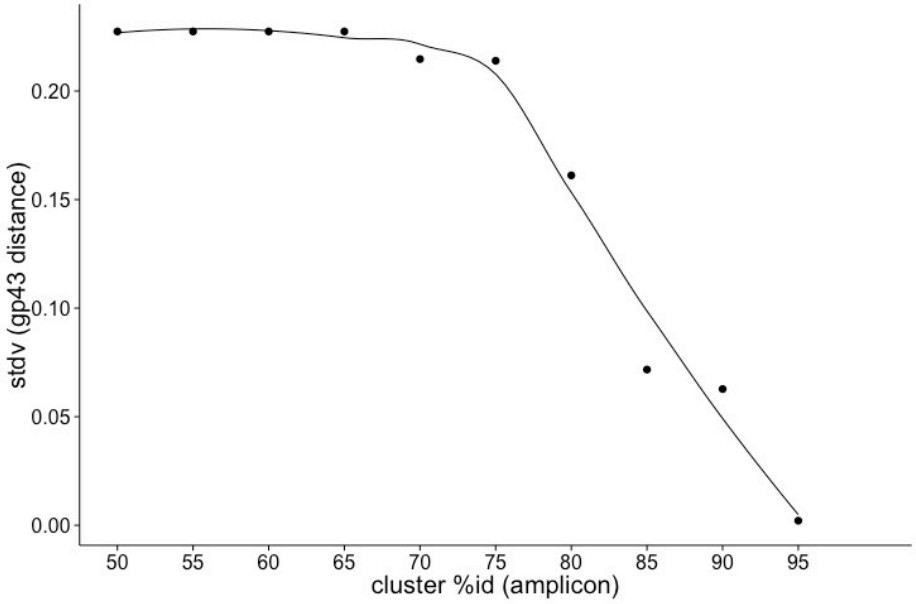

Figure 4.3: Variation in pairwise phylogenetic distance of gp43.
The variance is for full length gp43 sequences at different % clustering thresholds of the gp43 amplicons. Based on 19 reference cyanomyoviruses.

The samples represent a range of environmental conditions (Supplementary table A2). Saanich Inlet is seasonally stratified, and partially separated from adjacent waters outside the inlet by an 80 m deep sill at its entrance. The mixed layer depth in SAA ranged from two to below 10 m with deeper mixing usually occurring in the fall and spring. The SOG samples ranged in salinity from 23 to 31 PSU; whereas in SAA the salinity remained around 28 PSU. Temperatures varied seasonally in SAA between 7 and 14 °C. Temperature (T) versus salinity (S) plots for SOG and SAA (Supplementary figures A2 and A3), show that density differences among samples in SOG are mainly driven by salinity while in SAA it is a combination of salinity and temperature. For SAA, nutrient concentrations and viral and bacterial abundance data were only available for the 10 m samples. Picophytoplankton abundances measured by flow cytometry generally showed high abundances of *Synechococcus* spp. with some eukaryotes.

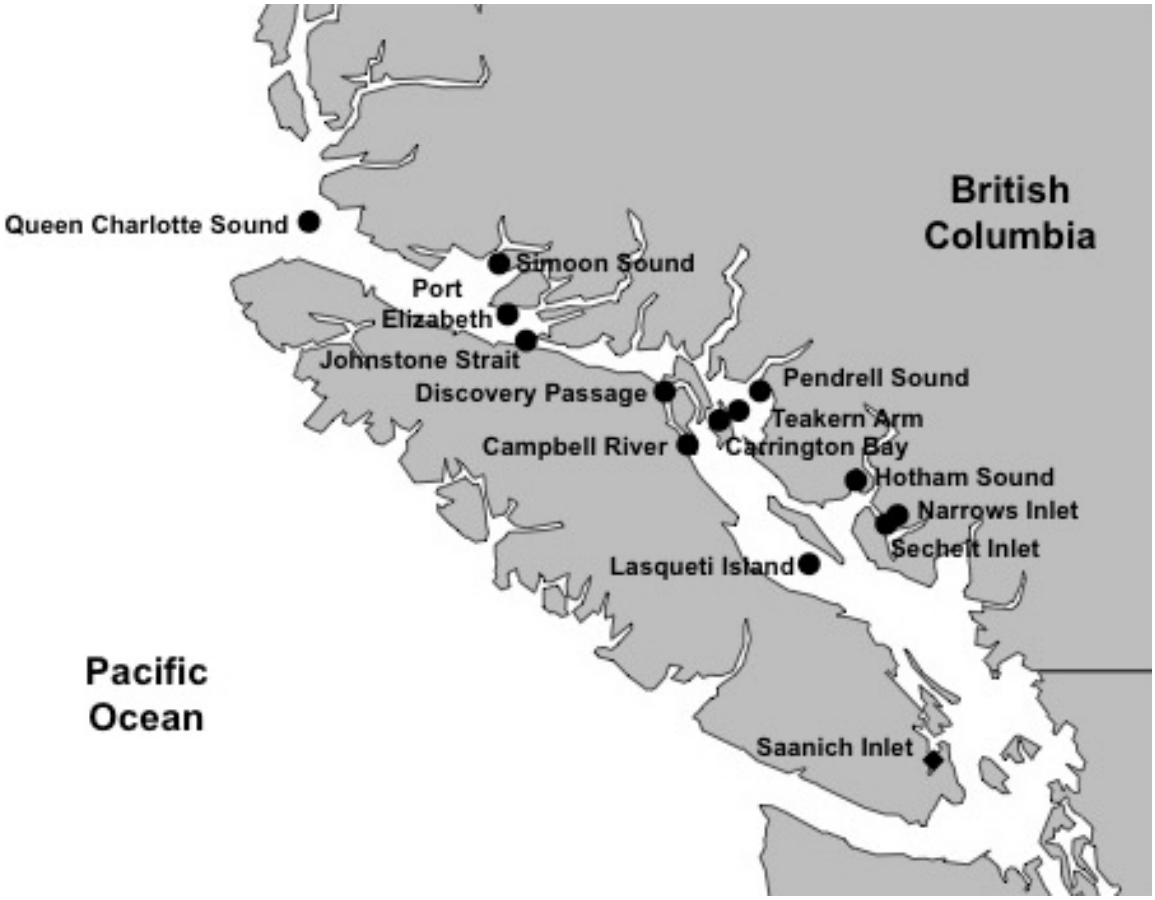

Figure 4.4: Sampling locations for the Strait of Georgia and Saanich Inlet.
Samples were taken from 2010-2012 at several time points and depths Saanich Inlet samples (SAA; solid diamonds) or over an integrated depth for Strait of Georgia samples (SOG; solid circles).

The combined sequencing data resulted in 9.84 million unique reads after quality control and dereplication; subsequent clustering at 97% aa identity produced 12,200 gp43 OTUs. A minimum cluster size of 500 reads resulted in 667 OTUs, representing 97% of all initial, unique reads. BLAST-P analysis revealed that 606 of the reads (90.9%) were associated with cyanophages. An EPA tree constructed by placing these OTUs on a maximum-likelihood reference tree of full length gp43 sequences revealed that most OTUs belonged to clades that were not represented by the reference sequences (Figure 4.5). Overall, the tree produced 15 coherent clades, indicated as I-XV.

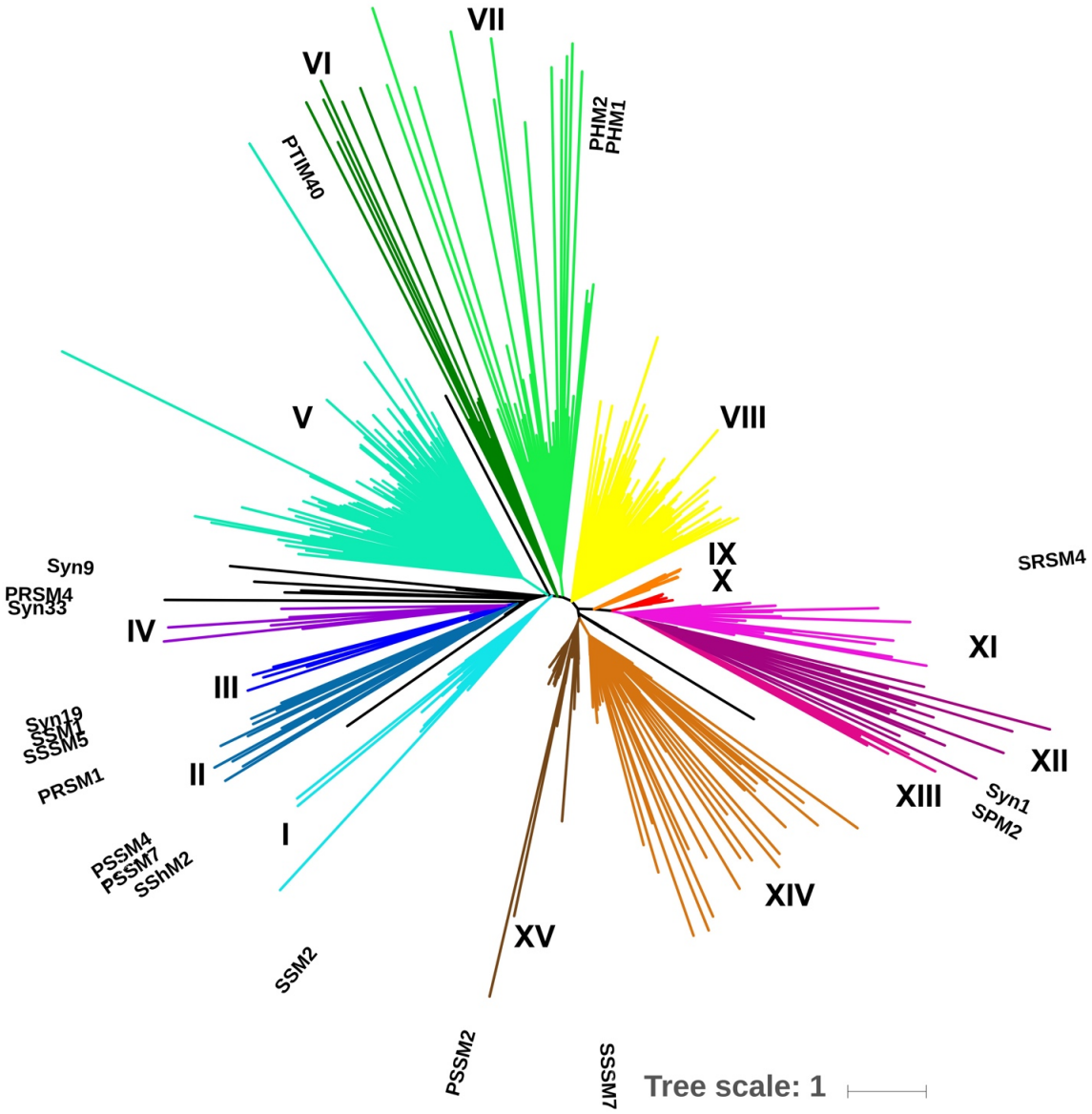

Figure 4.5: EPA phylogeny of 625 gp43 OTUs.
A ML tree of full-length gp43 sequences from 19 viruses served as a reference tree and 606 OTUs at 97 % aa identity from environmental amplicons were mapped onto the reference tree. Reference sequences are labeled; each tip represents an OTU. Coherent clades are color coded and labeled in roman numerals; secluded OTUs are in black, the scale bar represents the substitution rate.

Spatial variation among samples from the Strait of Georgia (SOG)

The 18 SOG samples covered a range of environmental conditions as shown in the PCA based on environmental variables (figure 4.6). Samples clustered into open straits, Johnstone Strait, Queen Charlotte Sound, Port Elizabeth, Discovery Passage and Campbell River, and sheltered inlet samples. The analysis also showed that samples from Narrows Inlet, Sechelt Inlet and Pendrell Sound, that were collected over multiple years did not necessarily cluster close together, indicating that the environmental conditions differed among years. The first and second dimensions of the PCA accounted for 70.88% and 11.68% of the variation, respectively. Variations were strongly driven by nutrients, salinity and temperature (figure 4.6b), with temperature and salinity offsetting each other.

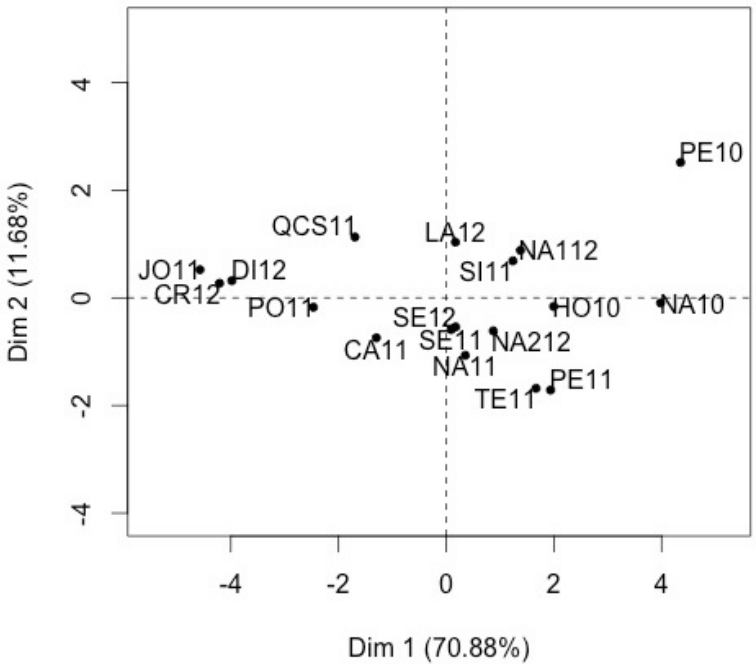

Figure 4.6: PCA of SOG samples based on environmental variables.
Labels show sampling year and location: CA, Carrington Bay; CR, Campell River; DI, Discovery Passage; HO, Hotham Sound; JO, Johnstone Strait; LA, Lasqueti Island; NA (1,2), Narrows Inlet; PE, Pendrell Sound; PO, Port Elizabeth; SE, Sechelt Inlet; SI, Simoon Sound; Teak, Teakern Arm; QC, Queen Charlotte Sound.

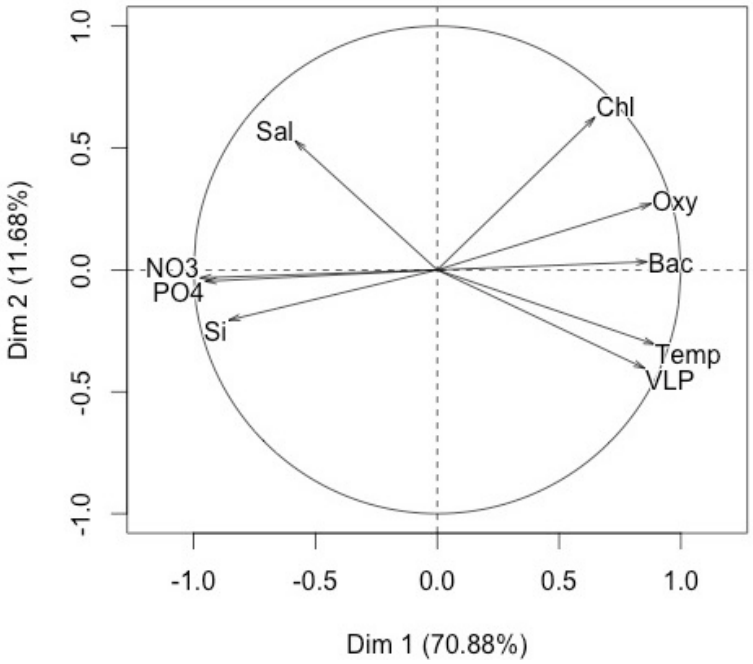

Figure 4.6.b: Environmental factors of the PCA for SOG.
Vector direction and length indicate the covariation and relative strength of environmental variables. Labels: Sal, salinity; Temp, temperature; Oxy, Oxygen; NO3, reduced nitrate and nitrite; PO4, phosphate; Si, silicate; VLP, viral abundance; Bac, bacterial abundance.

Rarefied community composition of the 625 OTUs for the 18 SOG samples showed that viral communities have dominant and persistent OTUs within phylogenetic groups and across the phylogenetic tree (Figure 4.7), samples are arranged in columns and OTUs are sorted in rows by clades derived from the phylogeny in figure 4.5. However, there were striking differences among the communities from Teakern Arm, Queen Charlotte Sound, Simoon Sound and Carrington Bay. Among the dominant OTUs for the SOG samples were those in the phylogenetic clades IV, V, VIII, XIV, XV.

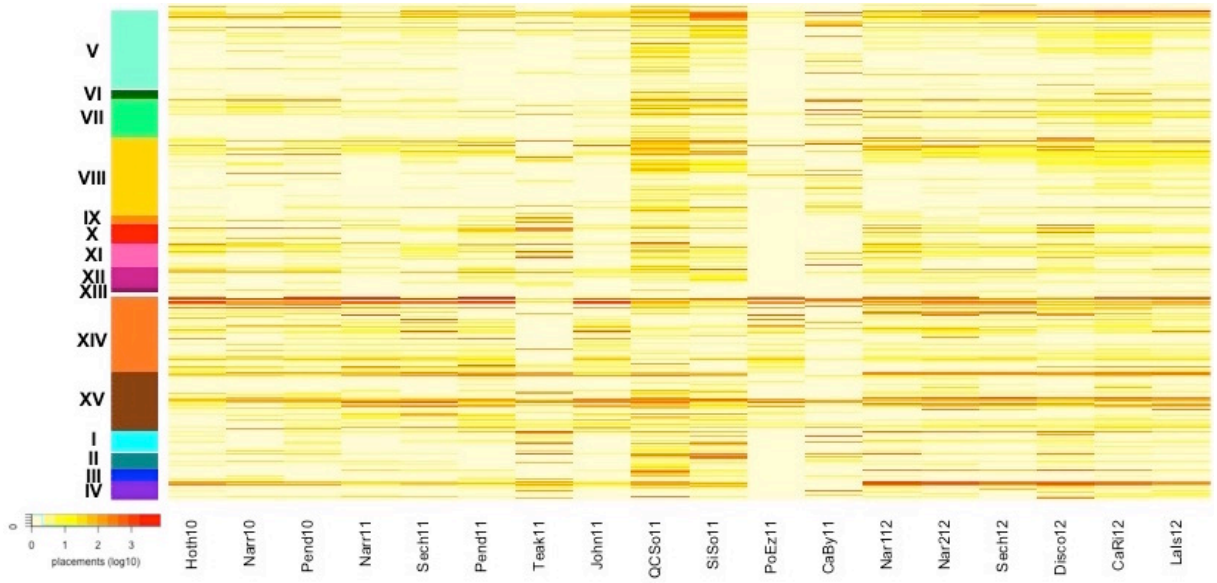

Figure 4.7: Community composition of SOG samples.
Communities were rarefied, heat indicates relative abundance. Samples are arranged in columns by year, OTUs are arranged in rows with clades indicated by color and number in correspondence to figure 5. Labels: CaBy, Carrington Bay; CaRi, Campell River; Disco, Discovery Passage; Hoth, Hotham Sound; John, Johnston Strait; Lals, Lasqueti Island; Narr, Narrows Inlet; Pend, Pendrell Sound; PoEz, Port Elizabeth; Sech, Sechelt Inlet; SiSo, Simoon Sound; QCSo, Queen Charlotte Sound; numbers indicate the sampling year.

A PCoA of the SOG virus communities grouped samples into three main clusters (Figure 4.8). The 2012 samples cluster together regardless of where they were collected. Another cluster comprises samples from 2010 and 2011, from inlet and strait locations. The four samples that fall outside of the clusters are from Teakern Arm and Carrington Bay in the southeast and Queen Charlotte Sound and Simoon Sound in the north. What also stands out is that the community composition in Narrows Inlet varies over the sampling years 2010, 2011 and 2012.

Diversity also varies across samples, with Port Elizabeth having the lowest diversity estimate (2.27) and Queen Charlotte Sound the highest (4.98), beta diversity (Shannon) for SOG was 1.96. Species richness was lowest in Narrows Inlet in 2010 (142) and highest in Discovery Passage in 2012 (494) (Table 4.2). Across all samples, salinity had significant explanatory power over alpha diversity and species richness with R^2 values of 0.2 and 0.3 (Figure 4.09).

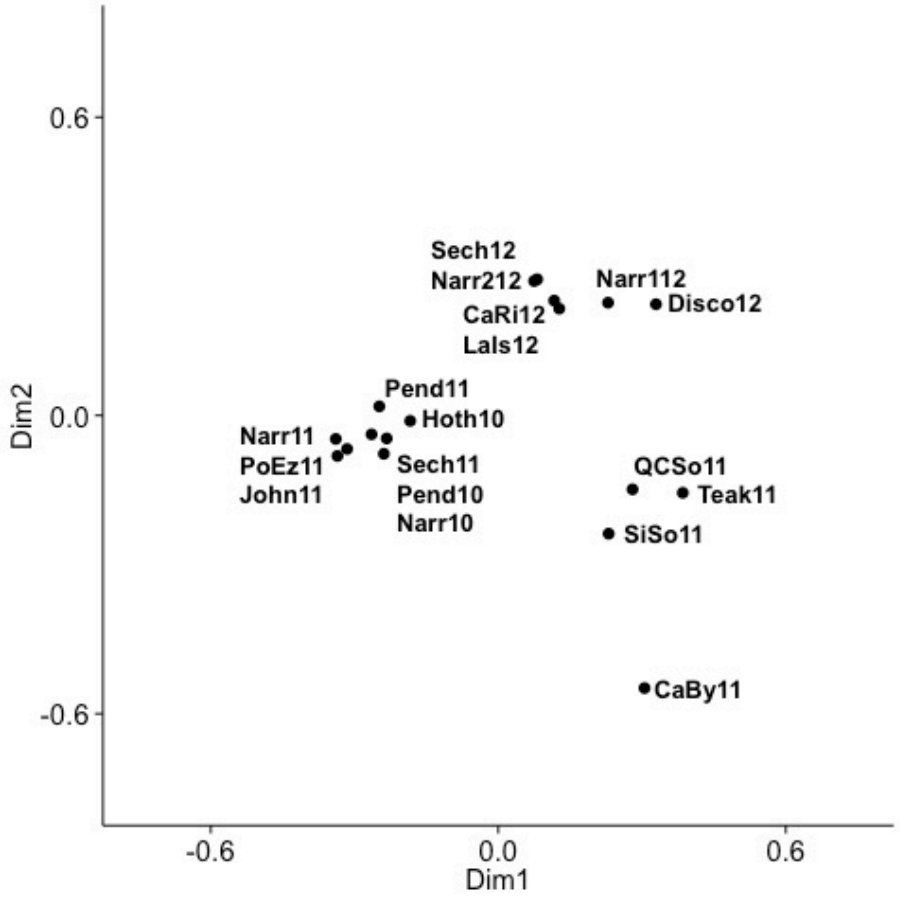

Figure 4.8: PCoA of SOG cyanomyovirus community composition.
Labels show year and location: CaBy, Carrington Bay; CaRi, Campell River; Disco, Discovery Passage; Hoth, Hotham Sound; John, Johnston Strait; LaIs, Lasqueti Island; Narr, Narrows Inlet; Pend, Pendrell Sound; PoEz, Port Elizabeth; Sech, Sechelt Inlet; SiSo, Somoon Sound; QCSo, Queen Charlotte Sound.

Table 4.2: Diversity indices for SOG and SAA communities.
Diversity is Shannon alpha diversity, richness is defined as species richness. SOG samples arranged by site, SAA by month for surface and 10 m samples.

SOG

Site	Year	diversity	richness
Hotham Sound	2010	3.17	426
Narrows Inlet	2010	3.02	142
Pendrell Sound	2010	3.18	348
Narrows Inlet	2011	2.52	270
Sechelt Inlet	2011	2.86	363
Pendrell Sound	2011	3.08	355
Teakern Arm	2011	4.11	280
Jonhstone Strait	2011	2.90	329
QC Sound	2011	4.98	475
Simoon Sound	2011	4.49	403
Port Elizabeth	2011	2.27	193
Carrington Bay	2011	3.43	188
Narrows Inlet 1	2012	4.07	426
Narrows Inlet 2	2012	3.83	416
Sechelt Inlet	2012	3.54	428
Discover Passage	2012	4.12	494
Campbell River	2012	3.73	455
Lasqueti Island	2012	3.66	462

SAA

Month	Year	surface	10 m	surface	10 m
May	2011	4.48	3.8	349	368
June	2011	4.55	4.75	379	456
July	2011	3.66	4.37	328	462
September	2011	3.82	4.05	293	448
November	2011	3.33	3.92	303	358
December	2011	4.42	4.87	383	498
January	2012	3.67	4.59	256	440
February	2012	4.61	4.74	394	453
March	2012	4.56	4.23	424	353
May	2012	4.22	4.39	300	433
June	2012	3.47	3.87	318	464
August	2012	3.09	4.63	310	435

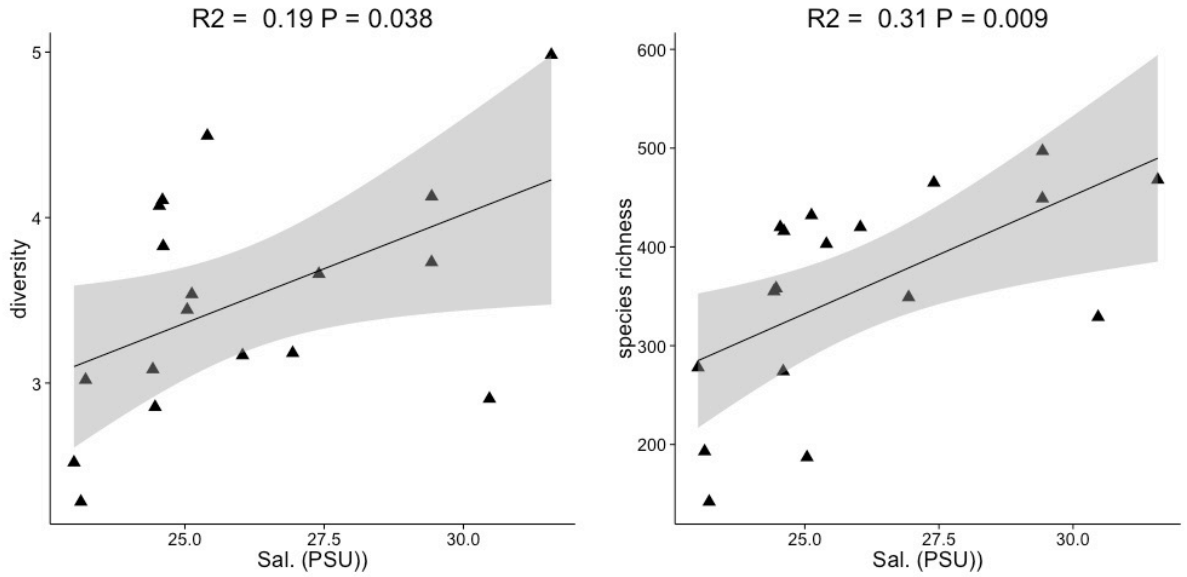

Figure 4.9: Diversity and richness of SOG samples in relation to salinity.
Diversity is Shannon alpha diversity, richness is species richness, linear regression shown, grey shading is the 95 % confidence interval, R^2 and significance (p) are shown.

Temporal variation among samples from Saanich Inlet (SAA)

The 12 monthly samples from Saanich Inlet for 10 m covered all seasons for a year. A PCA analysis of the environmental variables showed clustering of samples by season (Figure 4.10). Samples from November to March and from June to September formed two clusters. The spread within and between clusters is driven by nutrients and temperature (Figure 4.10b). Additionally, samples from May 2011 and 2012 cluster away from the main clusters, apparently based on chlorophyll and oxygen.

Community composition over 12 months for the surface and 10 m samples shows that certain OTUs are dominant throughout the year (Figures 4.11 and 4.11b), samples are sorted in columns by month and OTUs are sorted in rows by clades derived from the phylogeny in figure 4.5. These dominant OTUs are from clusters throughout the phylogenetic tree, namely clades number IV, V, VIII, X, XI, XIV, XV.

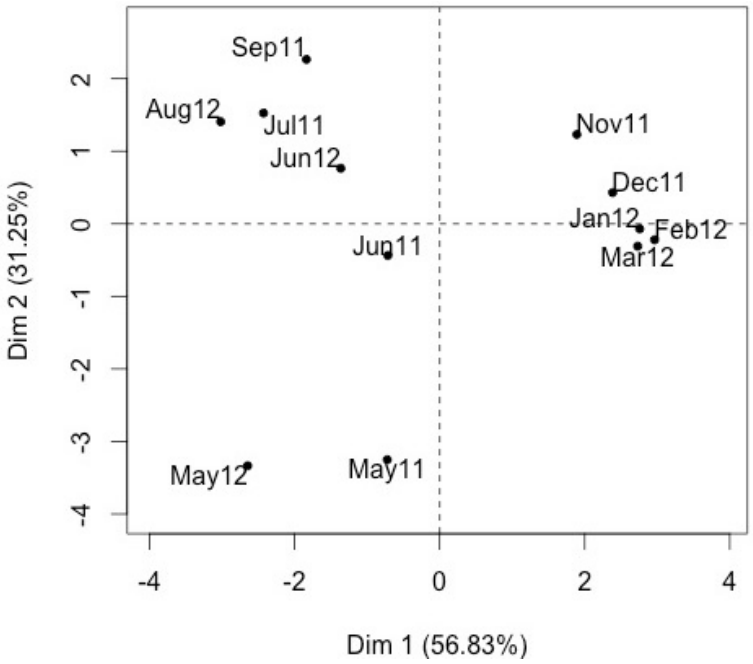

Figure 4.10: PCA of SAA 10 m samples based on environmental variables.
Labels indicate sampling month and year.

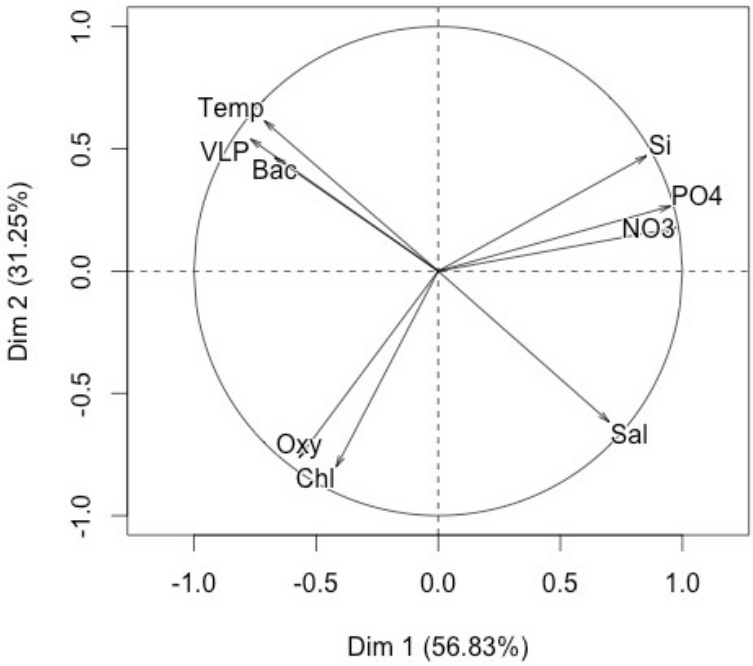

Figure 4.10.b: Environmental factors of the PCA for SAA 10m.
Vector direction and length indicate the covariation and relative strength of environmental variables. Labels: Sal, salinity; Temp, temperature; Oxy, Oxygen; NO3, reduced nitrate and nitrite; PO4, phosphate; Si, silicate; VLP, viral abundance; Bac, bacterial abundance.

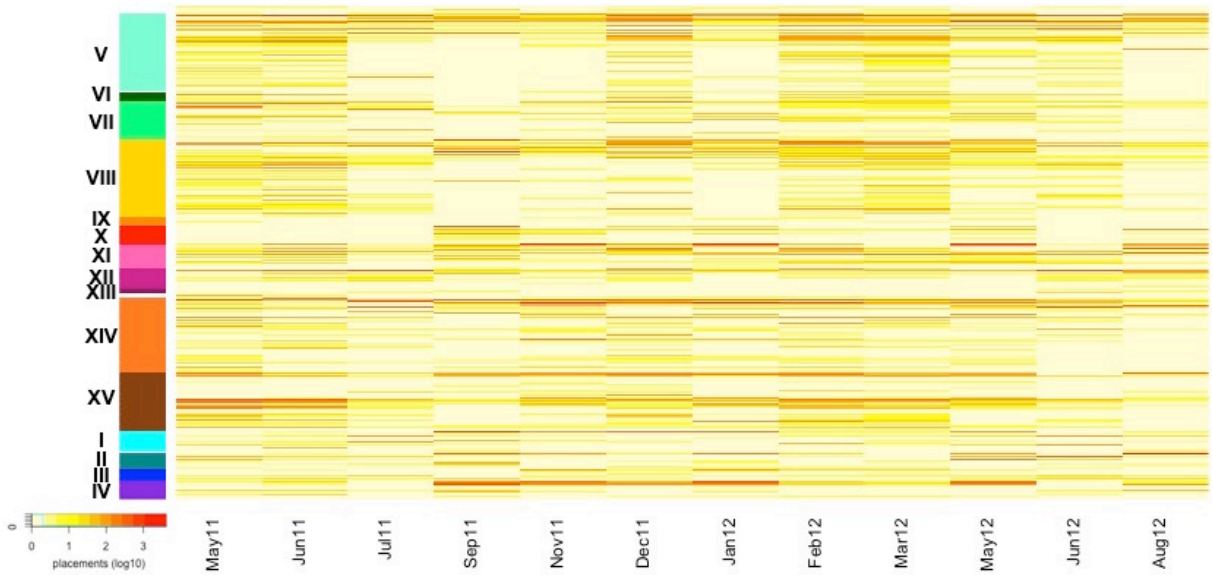

Figure 4.11: Community composition of SAA samples.
Samples from the surface layer, communities were rarefied, heat indicates relative abundance. Samples are arranged in columns by year, OTUs are arranged in rows with clades indicated by color and number in correspondence to figure 5, labels indicate sampling month and year.

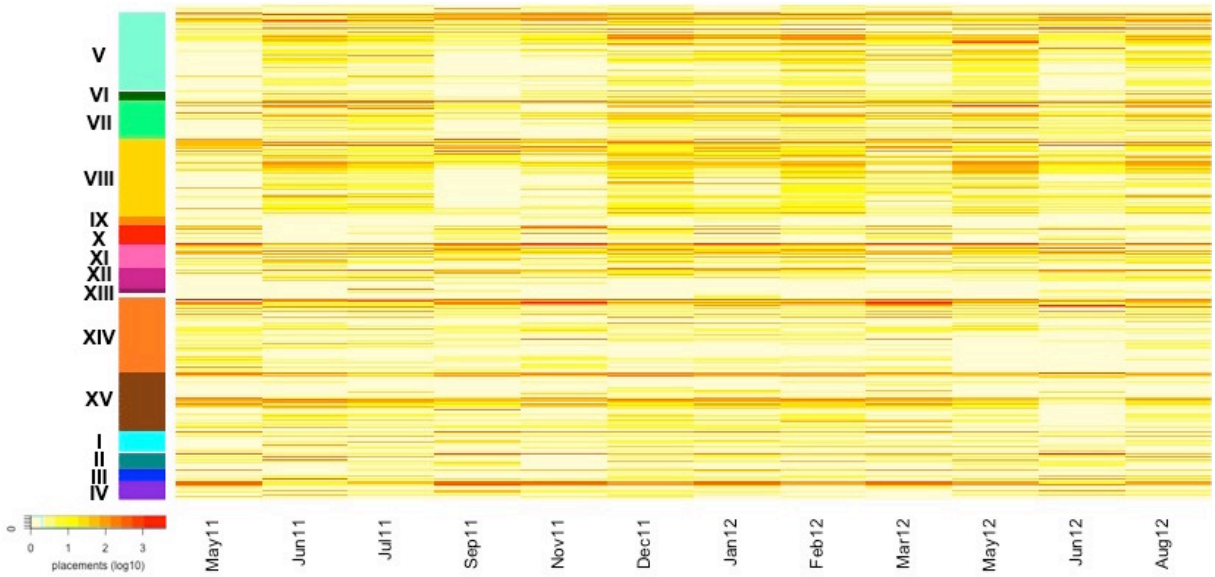

Figure 4.11.b: Community composition of SAA samples.
Samples from 10 m depth, communities were rarefied, heat indicates relative abundance. Samples are arranged in columns by year, OTUs are arranged in rows with clades indicated by color and number in correspondence to figure 5, labels indicate sampling month and year.

Comparison of community composition across months for the surface layer and 10 m samples by PCoA revealed a seasonal pattern (Figures 4.12 and 4.12b). Communities from November to February cluster together; whereas, communities from March through September are more spread out. Generally, the surface layer communities (Figure 4.12) show more variation than communities from 10 m (Figure 4.12b). Diversity indices for the SAA surface layer samples ranged from 3.09 in August, to 4.61 in February, while species richness ranged from 256 in January, to 424 in March. At 10 m, the diversity ranged from 3.80 in May, to 4.87 in December, and richness varied from 353 in March, to 498 in December (Table 4.2,). The beta diversity (Shannon) was 2.19 and 1.75 for the surface and 10 m samples.

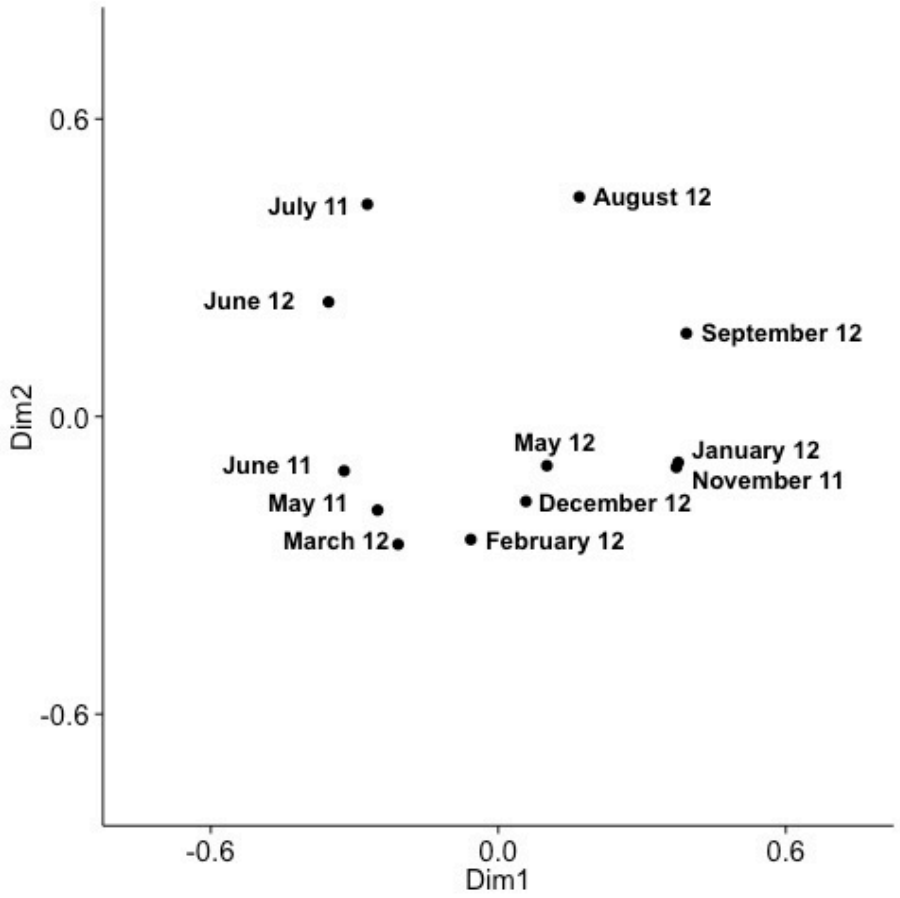

Figure 4.12: PCoA of SAA cyanomyovirus communities.
Samples were from the surface layer, the PCoA is based in relative OTU abundances. Communities were rarefied, labels indicate sampling month and year.

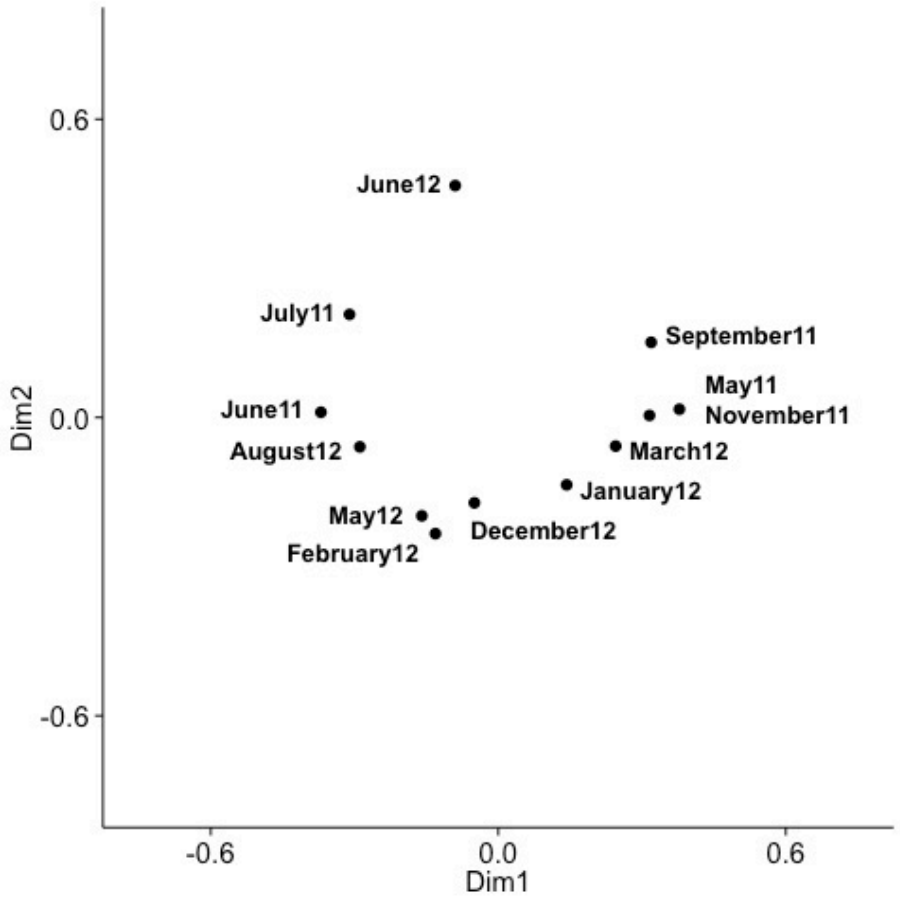

Figure 4.12.b: PCoA of SAAcyanomyovirus communities.
Samples from 10 m depth, the PCoA is based in relative OTU abundances. Communities were rarefied, labels indicate sampling month and year.

Combined analysis and the effect of environmental variables

In a wider approach data from SOG and the SAA 10 m samples, which have all the corresponding environmental data available, were combined for analysis. In comparison SOG samples show the largest range in diversity, followed by SAA surface layer and 10 m samples (Figure 4.13). The beta diversity is highest in the surface layer samples from SAA (2.19), followed by SOG (1.96) and SAA 10 m (1.75). The range in species richness is highest in the SOG samples while the average richness is highest in SAA 10 m samples (Figure 4.14) Regression of the combined alpha diversity and species richness of SOG and SAA 10 m samples against environmental variables showed a strong relationship to salinity with significant R^2 values of 0.38 and 0.39 (Figure 4.15).

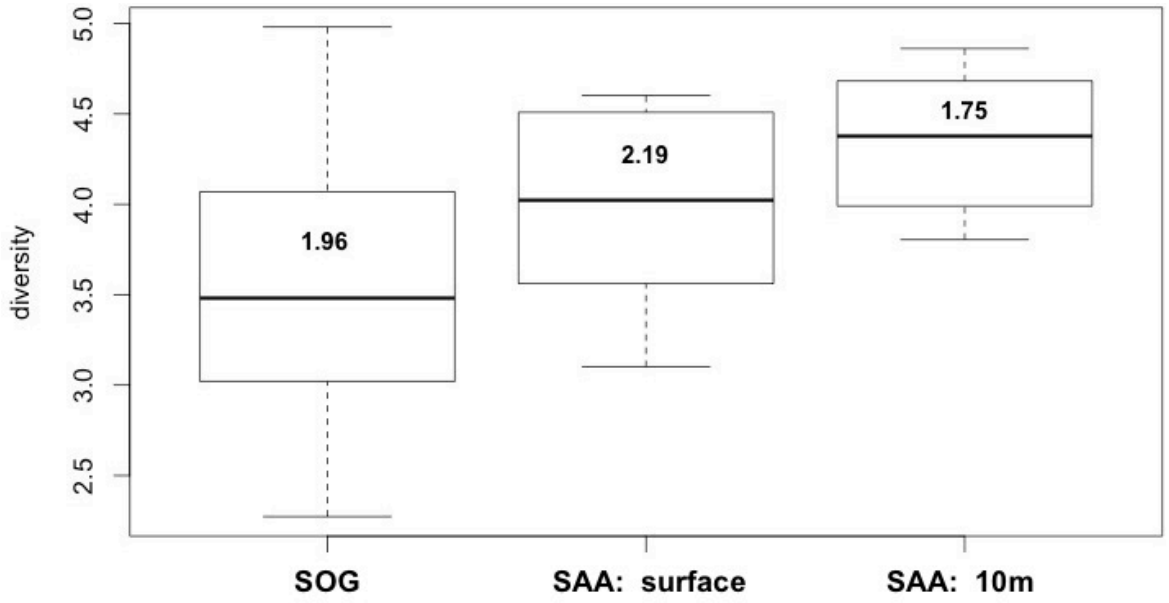

Figure 4.13: Range of diversity for SOG, SAA surface and 10 m communities.
Diversity is Shannon alpha diversity, overall beta diversity per subset is shown in the box. Whiskers indicate the range, box 50 % of data points with median.

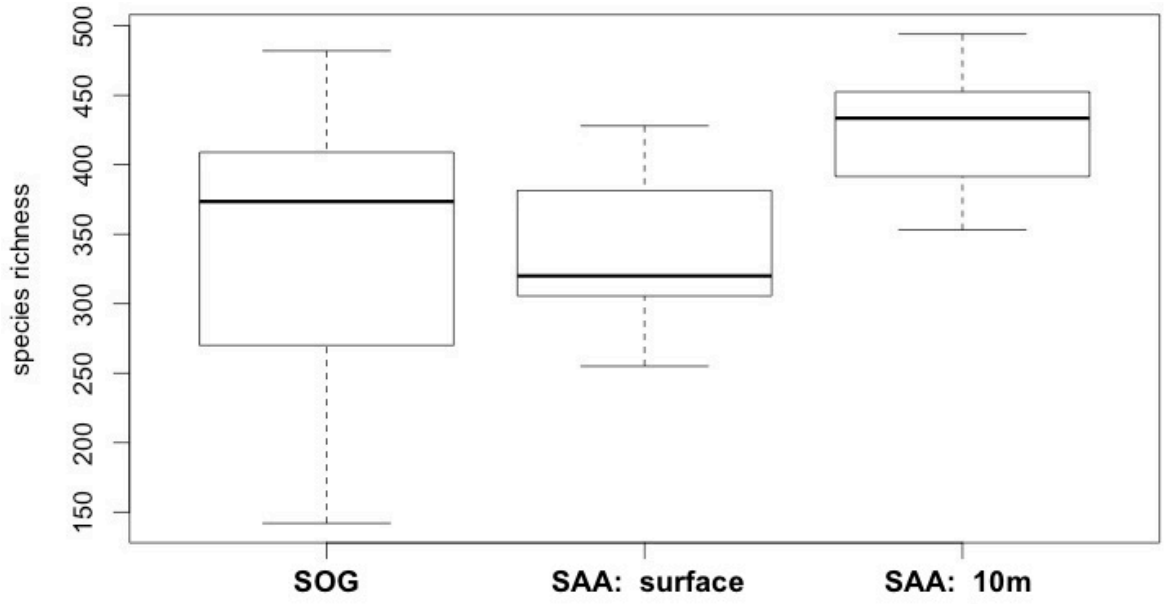

Figure 4.14: Range of richness for SOG, SAA surface and 10 m communities.
Whiskers indicate the range, box 50 % of data points with median.

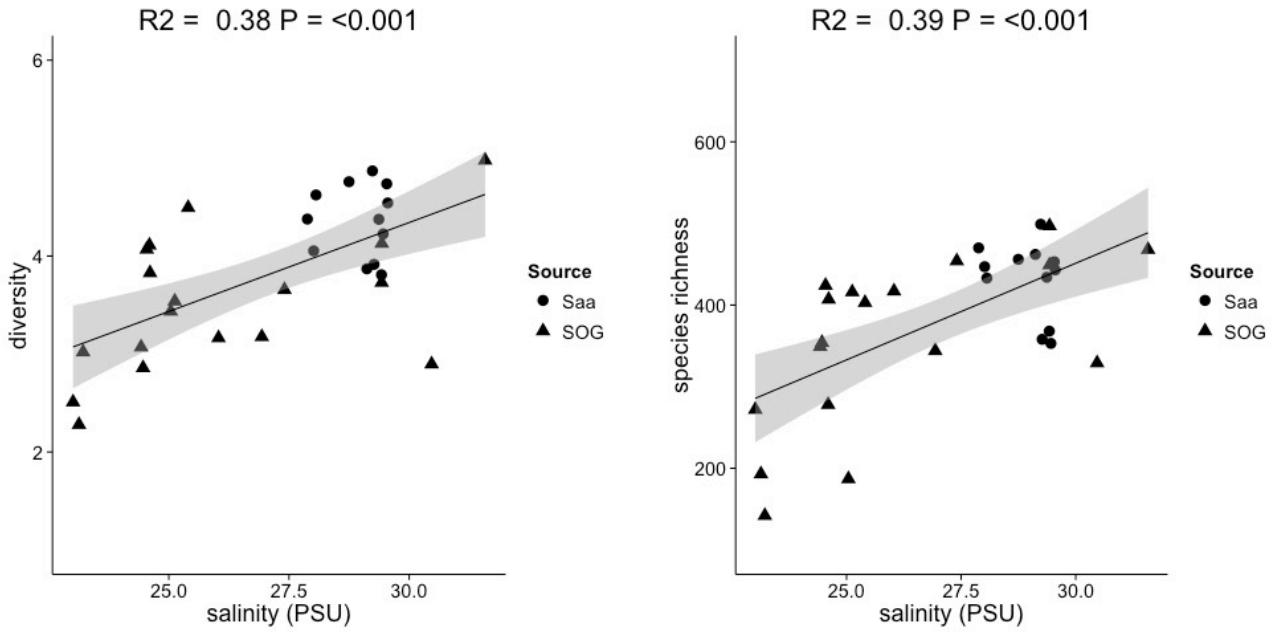

Figure 4.15: Diversity and richness of combined samples in relation to salinity.
Diversity is Shannon alpha diversity and richness is species richness. Sample source is indicated, SOG (solid circles) and SAA 10 m (solid triangle), linear regression shown, grey shading is the 95 % confidence interval, R^2 and significance (p) are shown.

Community composition data from SOG and 10 meters SAA samples (samples accompanied by all environmental data) were compared to each other based on pair-wise Bray-Curtis similarity. A subsequent PCoA analysis of the combined SOG and SAA communities showed a trend for clustering by source and that the SOG and SAA samples generally vary along two distinct trajectories (Figure 4.16). However, there is also overlap between communities from SOG and SAA, with Teakern Arm, Queen Charlotte Sound, Simoon Sound and Carrington Bay clustering with the SAA communities.

To identify how environmental variables are shaping the community composition, communities were compared for their similarity in community composition, covariation of OTUs and the effect of environmental variables. The combined data from SOG and SAA (10 m) was tested in a constrained correspondence analysis (Figure 4.17). Samples are scattered in two dimensions; dimension one (CCA1) described 9.99% of the community variation, 27.09% accumulated variation and dimension two (CCA2) described 5.76% of the community variation, 15.61% accumulated variation. Communities are loosely spread

out by SOG and SAA samples and by season, especially Port Elizabeth is distant to the other communities.

The majority of OTUs cluster mainly in the center, with some OTUs grouping with the Saanich Inlet summer samples. The combined environmental variables accounted for 36.9% of the constrained variation in a significant model (p=0.04). Salinity appears to be the strongest parameter, followed by temperature; both vary primarily along the first dimension, but in opposing directions. The nutrients co-vary primarily along the second dimension. A stepwise regression of the model revealed temperature, salinity and nitrate as significant model parameters (p=0.012, 0.018, 0.047) in influencing community composition.

To further investigate whether specific OTUs are crucial in defining communities an indicator species analysis was executed. Building on the results from the CCA, the combined samples were divided into three classes based on the strongest environmental variables, temperature, salinity and nitrate concentration. Environmental class 1, 2 and 3 were predominantly characterized by high, medium and low mean nitrate concentrations, class 1 was relatively cold and saline while classes 2 and 3 were on average warmer and less saline. Each of the classes were about equally comprised of 10, 11 and 9 communities. Samples in the first class are the winter samples from SAA and well mixed SOG samples. The second class is comprised of sheltered inlet samples from SOG and the third class includes the spring and summer samples from SAA plus Simoon Sound and Pendrell Island 2010. An indicator species analysis on the three classes assigned a total of 104 significant OTUs at a p=0.05 cut-off, and revealed a number of indicator species with a significant association to one of the classes (Table 4.3). Most indicator species were assigned to the 3rd class (80), only 6 and 18 were assigned to the 2nd and 1st classes, respectively.

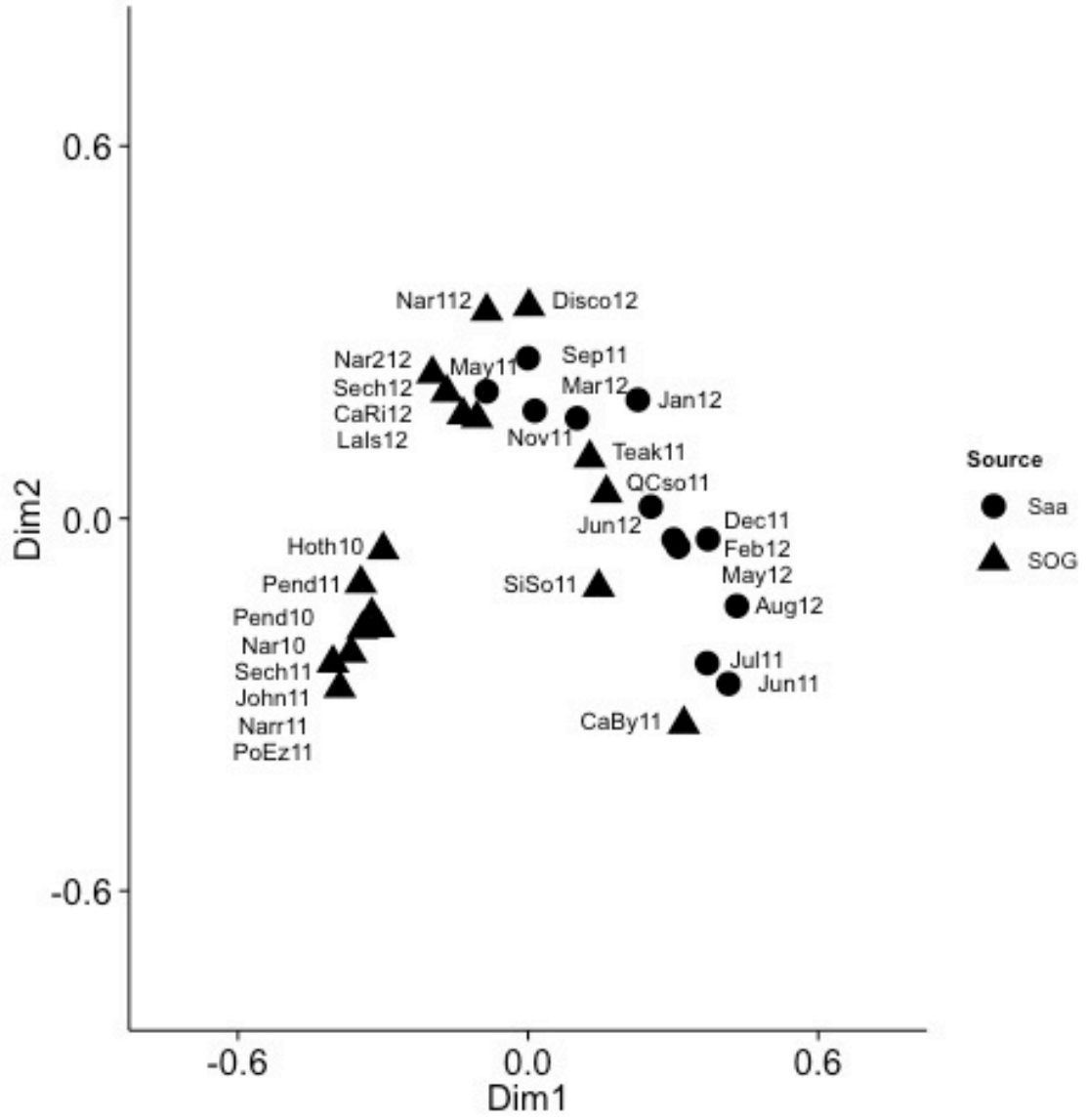

Figure 4.16: PCoA of combined SOG and SAA 10 m cyanomyovirus communities.
PCoA based on Bray-Curtis similarity of the community composition. SOG labels: CaBy, Carrington Bay; CaRi, Campell River; Disco, Discovery Passage; Hoth, Hotham Sound; John, Johnstone Strait; Lals, Lasqueti Island; Narr, Narrows Inlet; Pend, Pendrell Sound; PoEz, Port Elizabeth; Sech, Sechelt Inlet; SiSo, Simoon Sound; Teak, Teakern Arm; QCSo, Queen Charlotte Sound. SAA labels denote sampling month and year.

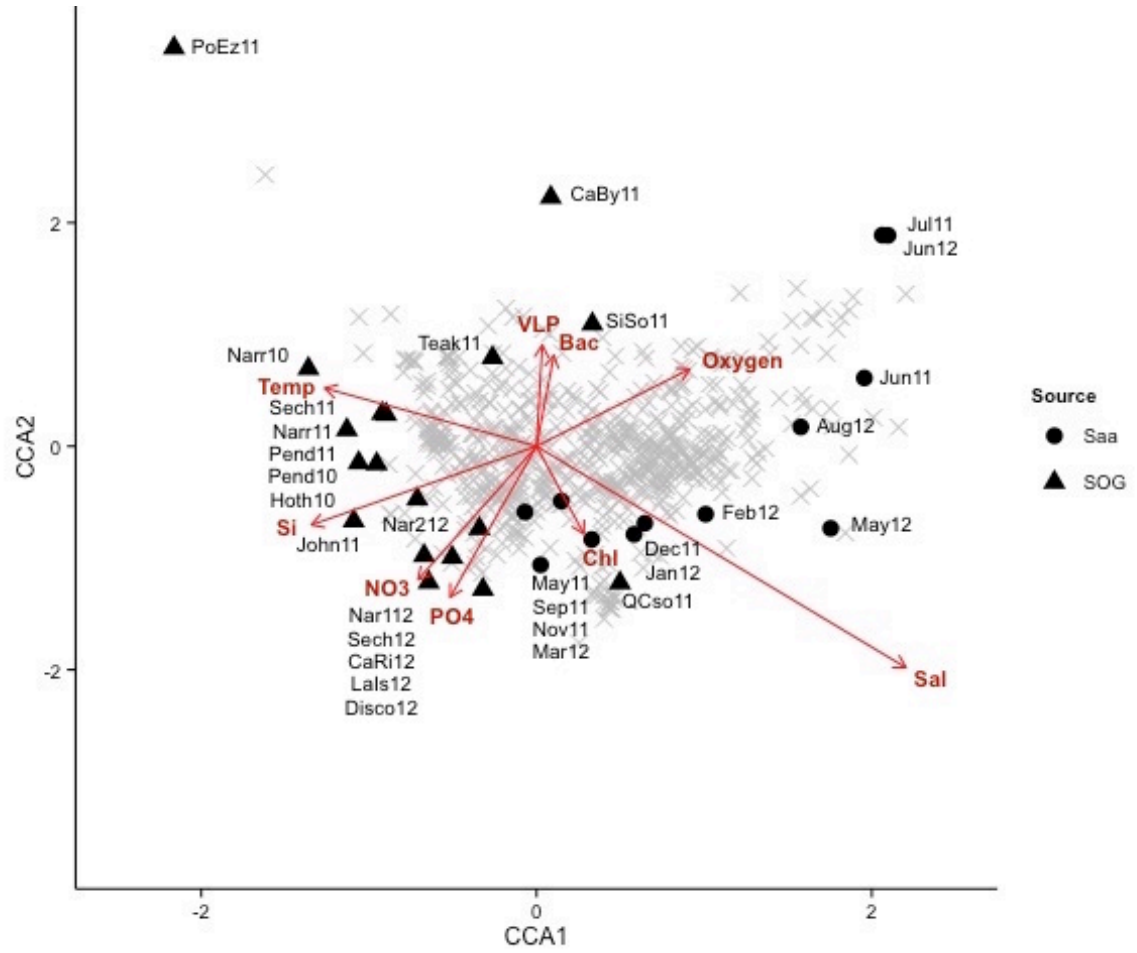

Figure 4.17: CCA of the combined SOG and SAA 10 m cyanomyovirus communities.
SAA samples are shown as solid circles and SOG samples as solid triangles, OTUs are in grey heathers. The direction and strength of the constraining environmental variables are shown in red. SOG labels: CaBy, Carrington Bay; CaRi, Campell River; Disco, Discovery Passage; Hoth, Hotham Sound; John, Johnston Strait; LaIs, Lasqueti Island; Narr, Narrows Inlet; Pend, Pendrell Sound; PoEz, Port Elizabeth; Sech, Sechelt Inlet; SiSo, Somoon Sound; Teak, Teakern Arm; QCSo, Queen Charlotte Sound. SAA labels denote sampling month and year.

Table 4.3: Cyanomyovirus indicator species analysis.
Environmental classes one, two and three with the mean temperature, salinity and nitrate concentrations, and the class sample size and number of assigned OTUs per class.

Class	1st	2nd	3rd
Temperature (°C)	9.1	12.9	11.2
Salinity (PSU)	29.1	24.8	28.1
Nitrate (µM)	25.1	11.8	4.7
Samples	10	11	9
OTUs	18	6	80

4.5 Discussion

Phages infecting cyanobacteria influence marine ecosystems by affecting nutrient cycles, primary production and host diversity, but cyanophages also are susceptible to environmental variables throughout their replication. The marine environment is inherently variable, and subject to the long-term effects of climate change. This environmental variability, in turn, will interact with the genetic repertoire of the cyanophage ecotypes to influence their distribution and relative abundance. This study examines differences in cyanophage community composition across a range of environmental conditions. By linking the genetic repertoire of cyanophages to genetic distance in a marker gene, temporal and spatial changes in cyanophage community composition in the context of environmental conditions is explored.

Deriving similarity in gene content from gp43 sequences

Determining whether the genetic repertoire of cyanomyoviruses can be inferred from gp43 sequences first required a phylogenetic distance analysis based on the presence and absence of genes. This is a powerful approach to compare genomes of closely related viruses, as was done for viruses infecting eukaryotic phytoplankton (Yutin, Wolf and Koonin, 2014; Legendre *et al.*, 2015). A neighbor-joining tree of 19 reference cyanomyovirus genomes clearly separated viruses into distinct branches, and produced reasonable bootstrap support values that are comparable to those obtained for comparison of viruses infecting eukaryotic phytoplankton (Yutin, Wolf and Koonin, 2014;

Legendre *et al.*, 2015) (Figure 4.1). Generally, the tree topology matched that of a phylogeny based on full-length gp43 sequences (Figure 4.2), the same viruses clustered together on branches or were isolated on both trees. This suggests that gp43 sequences reflect the gene content of cyanomyoviruses and is supported by the results from the Mantel Test, which showed a 0.87 congruence of pairwise phylogenetic distances based on genome content and full-length gp43 sequences (Table 4.1).

Given that the relative gene content of cyanomyoviruses can be inferred from gp43 sequences, the next step was to determine how gp43 amplicon sequences can provided an accurate representation of full-length gp43 sequences. Therefore the variance in pairwise distances of full-length gp43 sequences was set in correlation to different identity levels of amplicon sequences. That the variance approached zero at an identity level above 95 % and a maximum identity of 96.4 % among references indicated that amplicon sequences clustered at an aa identity of 97 % are specific to full length sequences. An identity of 97 % is also commonly used for clustering 16S rRNA sequences into taxonomic groups in bacteria (Buttigieg et al. 2015). It is relatively stringent compared to the 90 % identity that was used to cluster prasinovirus gp43 sequences (Clerissi, Grimsley, Ogata, et al. 2014), but less stringent than the 99 % identity used for comparing gp43 sequences of myovirus isolates (Marston & Amrich 2009; Marston et al. 2013). This level of inferred amino-acid identity for gp43 amplicon sequences strikes a balance between sufficient specificity to full length sequences of genotypes and a risk of inflating diversity of environmental amplicons. With this approach, environmental gp43 amplicon sequences can be used as a proxy for the similarity gene content of cyanomyoviruses, and thus create meaningful OTUs.

With a robust method of inferring relative similarity in gene content from gp43 amplicon sequence data, it was possible to investigate the diversity in natural cyanomyovirus communities and the presumed genetic repertoire. Clustering the combined environmental gp43 amplicon sequences from all samples at 97 % aa identity revealed 12,200 OTUs. While earlier studies have shown high phylogenetic diversity of myophages (Comeau and Krisch, 2008; Butina *et al.*, 2010) these results have uncovered unprecedented depths in richness, relative to studies on cyanophages or myoviruses using *phoH* or gp23 amplicon sequences (Filée *et al.*, 2005; Goldsmith, Parsons and

Beyene, 2015). However, only 667 (5.5 %) of the OTUs encompassed ~97 % of the initial sequences, showing the extreme unevenness of cyanomyovirus genotypes, and the overwhelming preponderance of rare taxa. A BLAST-P analysis positively identified a cyanophage as the top hit for 606 (90.9 %) of these 667 OTUs, with the rest presumably being other T4-like phages not infecting cyanobacteria. The data confirmed that gp43 is suitable as a marker gene for examining cyanomyovirus diversity when using the designated primers (Marston and Amrich, 2009). However, the sequence length of the amplicons is not sufficient to represent the large diversity of sequences in a well-supported maximum-likelihood tree. Therefore, an EPA phylogenic approach was adopted; in this method full-length gp43 sequences were used to construct a reference tree, on which the 606 amplicon sequences were placed in a phylogenetic context (Figure 4.5). Given that gp43 sequences reflect the similarity in gene content of cyanomyviruses, the great richness depicted in the tree also implies substantial variation in their genetic repertoire. The clades associated with the 19 reference viruses are well represented by numerous environmental reads, but there are also many clades with no known isolates indicating a great genomic richness of cyanomyoviruses that remains to be studied.

Environmental variables are defining the spatial and temporal samples

The 42 samples in this study spanned a range of conditions found in temperate, coastal environments in which *Synechococcus* spp. and their phages would be expected to occur. PCA analyses for the environmental variables in the SOG samples shows that locations were spread along a gradient mainly driven by nutrient concentrations and salinity; whereas, the SAA samples were separated into summer and winter clusters, that were mainly defined by nutrient concentration and temperature. The SAA samples from May 2011 and 2012 clustered separately and were linked to increased chlorophyll and oxygen concentrations associated with the spring bloom. The assessed environmental variables cover those that have been shown to influence virus host interactions (Mojica and Brussaard, 2014), excluding PAR because of its high daily variability. The flow cytometric signature does show cyanobacterial communities to be dominated by *Synechococcus* spp. throughout the samples, as was expected in these environments (Partensky, Blanchot and Vaulot, 1999; Scanlan and West, 2002). Considering the wide host range

documented for cyanomyoviruses (McDaniel, delaRosa and Paul, 2006; Hanson, Marston and Martiny, 2016), host variation is most likely not the sole driver of cynanomyovirus community composition. Overall, the link between nutrient availability and salinity and temperature indicate that the mixing regimes are what is ultimately defining these samples and have an effect on viruses directly or through the host.

Dominant phylogenetic groups prevail across samples

Examining the community composition across locations and seasons for SOG and SAA shows that some closely related OTUs were relatively abundant, widespread and persistent (Figures 4.7, 4.11, 4.11b). Other studies have reported that some phage taxa are widely distributed and commonly found over a range of samples (Wilson *et al.*, 1999; Short and Suttle, 2005; Goldsmith *et al.*, 2011). Recent studies on the dynamics of virus communities showed some genotypes to be persistent over time or consistently dominant (Chow and Fuhrman, 2012; Needham *et al.*, 2013; Goldsmith, Parsons and Beyene, 2015). It also is evident that communities show large differences across locations in SOG and more gradual shifts over seasons in SAA which could be based on contrasting environmental conditions and matches observations on cyanophage communities (Chow and Fuhrman, 2012; Marston *et al.*, 2013). Another observed pattern especially in the SOG data is that there are alternating communities, with only one or the other community pattern possible. A detailed look furthermore shows a fine variation of OTUs within the dominant clades, similar to the fine scale diversification observed within clusters of cyanomyovirus isolates over time and space (Marston and Martiny, 2016). Given that different phylogenetic groups of phages dominate in the communities, a difference in host community would be one explanation. However, a flexible host range has been documented for cyanomyoviruses (Lu, Chen and Hodson, 2001; McDaniel, delaRosa and Paul, 2006; Hanson, Marston and Martiny, 2016), thus alternatively it is likely that differing gene content among cyanomyoviruses affects their distribution and the community composition.

Diversity indices show relationships to environmental variables

One factor that may explain the observed pattern of distribution of the cyanophage communities and associated diversity indices is the mixing regime. Diversity is lowest in the sheltered, stratified, SOG Port Elizabeth sample (2.27) and during the stratified waters in the SAA August sample (3.09) and SAA May sample (3.80), for the surface and 10m samples respectively (Table 4.2). In contrast, the highest diversity is found in well-mixed SOG Queen Charlotte Sound sample (4.98) and for the well-mixed SAA samples in February (4.61) and December (4.87), for samples from the surface and 10 m respectively. This translates into a relationship between alpha diversity, richness and salinity of cyanomyovirus communities in the SOG samples and alpha diversity and salinity, temperature in the surface SAA samples (Figures 4.11, 4.15).

The comparison among subsets in the beta-diversity analysis shows that the surface samples in SAA, which are more exposed to seasonal changes and stratification than the 10 m samples, have the highest beta-diversity describing overall community variation (Figure 4.9). Especially in temperate environments, water-column mixing regimes that are related to season or weather, shape the prevailing conditions for plankton. The relationship of diversity and salinity even strengthens when SOG and SAA (10 m) samples are combined as it now describes the difference in environmental conditions on a geographical as well as a seasonal scale (Figure 4.16). While no relationship between diversity or richness and salinity has been shown for viruses to date, similar relationships between diversity or richness and salinity have been made for bacteria and phytoplankton. Estrada et al. (2004) showed an increase in phytoplankton and picoplankton diversity in a salinity range from 15-25 PSU, Herlemann et al. (2011) showed a slight increase in bacterial richness at salinities from 5-13 PSU. However, this relationship of bacterial diversity or richness to salinity was not confirmed by Campbell and Kirchman (2012) who found the highest diversity and richness of bacteria at low salinity. Another study only saw a very weak increase in bacterial richness at salinities above 15 PSU in lakes (Wang *et al.*, 2011). The relationship of diversity and temperature in the SAA surface samples is weak, but a large scale study on bacteria in geothermal environments revealed such a relationship with high confidence, albeit at a wider temperature range and stable salinity (Sharp *et al.*, 2014).

Connell (1978) proposed that diversity is highest when disturbance is intermediate. To a certain degree this has been shown for phytoplankton communities in a lake, with high diversity being a transient state, occurring at times of relatively high mixing depths (Weithoff, Walz and Gaedke, 2001). Applying this concept to the here presented data means that environments with higher salinity and thus a more mixed water column promote a higher diversity in cyanomyovirus communities by exposing communities to more varying conditions.

Differences in community composition correlate with environmental variables

A comparison of communities from the SOG samples was examined by PCoA. Samples with similar community composition resolved into three main clusters (Figure 4.8) which were neither well described by geographic proximity or the type of environment, nor did they match the pattern of environmental conditions described in the PCA. The similarity of the Queen Charlotte Sound and Simoon Sound communities can be explained by their proximity and because they are both relatively cold and well mixed. However, Teakern Arm and Carrington Bay, which are environmentally similar to each other but very different than Queen Charlotte Sound and Simoon Sound, also clustered in that group. All samples from 2012 clustering together is also surprising since they span a range of environments from sheltered Sechelt and Narrows Inlets to the more exposed Strait of Georgia. Consequently, seasonal environmental conditions are more important than location in dictating cyanophage community composition. Also the samples from Narrows Inlet over three years are all substantially different in their community composition, which further indicates a strong environmental component. The strong shifts in community composition observed at some stations are consistent with earlier observations in B.C. coastal waters and geographically distant locations (Frederickson, Short and Suttle, 2003; Marston *et al.*, 2013). Nonetheless, the observed pattern of OTU distribution and community composition did not match that of environmental variables in the SOG data.

Comparing changes in community composition of the SAA surface and 10 m samples by PCoA was consistent with community composition following a seasonal pattern (Figures 4.14, 4.14b). A similar pattern was seen in the PCA of environmental variables, which suggests that the cyanophage communities follow a seasonal pattern

governed by environmental variables, in particular temperature, mixing regime and thus nutrient availability. Changes in the community composition in the surface samples are more pronounced than those from 10m, presumably due to a stronger exposure of the communities to seasonal changes. Similarly, at the Bermuda Atlantic Time-series Study site viral communities showed higher similarity between surface and 100 m samples in the well mixed winter than under stratified summer conditions (Goldsmith *et al.*, 2015). Significant drivers were salinity and temperature and to a certain degree chlorophyll. The displayed seasonal pattern is similar to what Chow and Fuhrman (2012) showed for myovirus communities over a multi-year cycle, in which communities were similar when sampled 12 months apart and has also been reported for bacterioplankton (El-Swais et al. 2015). Defined, but gradual temporal shifts with seasonal re-occurrence of community composition (Marston et al. 2013) and community structures corresponding to seasons and the associated temperatures (Wang & Chen 2004) have also been observed. However, other studies have found very pronounced changes in community composition with season (Sandaa and Larsen, 2006).

To understand larger patterns in the community composition samples from SOG and SAA 10 m were analyzed together. In a PCoA of the combined samples, SOG and SAA samples vary primarily along two distinct trajectories and the differences in community composition among locations appear to be stronger than those among seasons (Figure 4.17). Yet the four SOG communities Teakern Arm, Queen Charlotte Sound, Simoon Sound and Carrington Bay that cluster with the SAA communities show that under certain conditions, communities from geographically distant samples can be surprisingly similar.

To further identify key environmental variables associated with shifts in community composition, a canonical correspondence analysis (CCA) was performed on the combined SOG and SAA 10m data (Figure 4.18). Overall, it shows that 36% of the variation in community composition across all samples is explained by a combination of all environmental variables. This is a significant result; another large scale study found a weak correlation between environmental variables and cyanomyovirus community composition (Huang *et al.*, 2015), but the study mainly focused on temperature. Another study on viral community composition assessed by metagenomics data showed that

sampling location, season and depth, and presumably the associated conditions, are strong predictors for viral community structure (Hurwitz *et al.*, 2014). In the present study, temperature, salinity and nitrate were significant variables, which is in line with the aforementioned finding that salinity and temperature relate to diversity indices. A study on prasinovirus distribution in the Mediterranean Sea found a similar pattern, with phosphate being the strongest variable to explain viral community composition (Clerissi, Grimsley, Subirana, *et al.*, 2014). Considering the relatively high requirements of viruses for nitrogen and phosphorus, and virions accumulating a substantial amount of the hosts nutrients (Jover *et al.*, 2014), their impact on viral replication is intuitive. In the SOG and SAA samples nitrate and phosphate co-varied strongly; hence, nitrate might simply have stronger statistical power, and thus generally describes nutrient availability. Moreover, because temperature and salinity control stratification, this also affects nutrient availability. As well, in the CCA most OTUs are clustered near the center, differences in community composition are more related to the relative abundance of each OTU in a community than by their presence or absence, as suggested before (Rodriguez-Brito *et al.*, 2010; Needham *et al.*, 2013). Counterintuitively, biological variables (viral abundance, bacterial abundance and chlorophyll concentration) were not significant factors affecting diversity indices or community composition. Some studies have shown correlations between cyanophages diversity and viral or cellular abundances while others have not. One study found in a north-south transect in the Atlantic Ocean that cyanomyovirus diversity did not correlate to any environmental variable, but found a correlation to host diversity (Jameson *et al.*, 2011). A similar correlation was found in the Red Sea where cyanophage abundance and diversity covaried with the abundance and diversity of the cyanobacteria (Mühling *et al.*, 2005). Huang et al. (2015) found a correlation between cyanomyovirus communities and host communities based on reads recruited to isolates, but not with reads recruited to genotypes. Notably, for cyanopodoviruses, which appear to have a narrower host range, that relationship was strong in both cases. Another transect in the Atlantic Ocean found an impact of temperature and increased mixing on phytoplankton virus dynamics (Mojica *et al.*, 2016).

Finally, an indicator species analysis was done to identify OTUs that are characteristic of specific environmental conditions (Table 4.3). Samples were divided into

three environmental classes based on salinity, temperature and nitrate, the strongest variables in the CCA. The first class represents well-mixed environments with high nutrient availability, the second and third classes represent gradually more stratified environments with lower nutrient availability. The distribution of indicator species to the three environmental classes shows that many phage OTUs were only associated with well-mixed environments with high nutrient availability. The few OTUs which were strongly correlated to environment two and especially three suggest that some phages however are specifically associated with envronments of lower nutrient concentrations, suggesting they may have traits that allow them to compensate for low-nutrient availability. The association of the genetic repertoire of a cyanomyovirus to its relative abundance in communities under environmental conditions formulates the presence of viral ecotypes. This is in line with the recent definition of viral ecotypes for isolates (Marston and Martiny, 2016).

This study showed that *gp43* reflects the similarity in gene content between pairs of cyanomyoviruses. In this study gp43 sequences portrayed enormous genetic diversity in cyanomyoviruses. Yet, only a fraction of this diversity represented the large majority of phages in all samples, and only a few OTUs were dominant.

The data on gp43 diversity were combined with information on location, season and environmental conditions to infer their relative importance in defining cyanomyovirus community composition. Environmental conditions varied substantially among sampling times and locations. While location was related to community composition in some samples, the pattern was stronger among seasons. Variance in cyanomyovirus community richness, diversity and composition were tied to salinity, temperature and nitrate concentration. Salinity can directly affect phage infectivity (Kukkaro and Bamford, 2009), but can also affect adsorption kinetics and influence the distribution of potential hosts. Given the range of environmental conditions sampled across this study, salinity is likely to be a descriptor for a complex combination of factors, including mixing and the associated availability of nutrients and light. This set of environmental variables appears to shape cyanomyovirus communities. Moreover, based on an indicator species analysis, some viral OTUs are associated with specific environmental conditions and, considering their representation of the genetic repertoire, thus describe viral ecotypes.

The mixing regime of the water column as defined by salinity and temperature, and the associated nutrient availability, are significant predictors of cyanomyovirus community composition. However, there are likely many other factors including the resident viral seed bank and particle dispersal (Chow and Suttle, 2015).

In conclusion, interactions between cyanophages and their hosts can be influenced by a variety of environmental factors. In part, these interactions may be affected by the genetic repertoire encoded by viral ecotypes which differs substantially and has the potential to carry modes to cope with environmental adversaries. Consequently, environmental variables shape cyanomyovirus communities from the viral seed bank, but all the processes involved remain to be fully understood. This research establishes a link between a marker gene phylogeny and corresponding similarity in genetic repertoire. The resulting data furthers the understanding of how the community composition of cyanomyoviruses and their genetic repertoires vary over time and space and how communities respond to environmental variables. This understanding will enable better predictions about the response of viral communities under changing environmental conditions.

Chapter 5: Cyanomyovirus communities show variability in their replication

5.1 Summary

Infection of cyanobacteria by cyanophages can effectively terminate host blooms, a concept termed "killing the winner". In turn, the host-virus interaction is affected by environmental conditions at any stage of the replication cycle. Many of these environmental conditions can be transient, including mixing of the water column and light radiation. This could lead to equally transient viral communities emerging out of the viral "seed bank".

This project investigates how free viral communities (the putative seed bank) and viral communities in the cellular fraction (considered to be the actively replicating viruses) compare to each other under differing environmental conditions. The aim is to elucidate the response of virus host systems to environmental variables. The research was approached on an operational taxonomic unit level using DNA polymerase (*gp43*) as a marker gene, that also serves as a proxy for the genetic repertoire of viruses. The composition of free cyanomyovirus communities and cyanomyovirus communities in the cellular fraction were compared for several sites.

In comparison, the composition of free cyanomyovirus communities varied strongly across sampling sites while the communities in the cellular fraction showed a surprisingly high degree of similarity. Furthermore, corresponding free and cellular communities from the same site generally showed a high degree of dissimilarity. Environmental conditions covary with sites of high free-to-cellular dissimilarity, light and nutrients, and sites of low free-to-cellular dissimilarity, salinity and bacterial abundance.

The observations could be described by three scenarios, trans-regional synchronisation of infection, high degradation of free viruses, or stalled infections. While none of these scenarios fully explains the results, the observations of free-cellular community dissimilarity are robust and need to be explained.

5.2 Introduction

At any given time, it is estimated that around 10 % of plankton cells in the ocean are infected by viruses, and about 20 % of the plankton biomass is lysed on a daily basis (Proctor and Fuhrman, 1990; Weinbauer, Brettar and Ho, 2003; Suttle, 2007). This lysis rate and the high abundance of viruses show the immense impact viruses have on marine ecosystems, nutrient cycles and plankton diversity. Considering that about half of the world's primary production is performed by marine phytoplankton (Field *et al.*, 1998), the importance of marine viruses at a global scale cannot be overstated (Wilhelm and Suttle, 1999; Mühling *et al.*, 2005; DeLong *et al.*, 2006; Weitz and Wilhelm, 2012). In addition to the sheer abundance of marine viruses and their ecological importance, they can have complex interactions with their hosts and be influenced by environmental variables.

Viral infection of phytoplankton can severely impact phytoplankton community composition. Experiments have shown the simultaneous infection of phytoplankton by viruses can lead to the termination of blooms and environmental studies show evidence for control of host community composition by cyanophages (Bratbak, Egge and Heldal, 1993; Mühling *et al.*, 2005). Furthermore, it was recognized that many viral genotypes are widely, even globally, present and that abundant genotypes which make up a large proportion of viral communities at a given time occur in response to host blooms (Breitbart and Rohwer, 2005). An early theory termed this dynamic “killing the winner”, suggesting that viruses respond to blooms of hosts, diminishing them and thus maintaining host diversity over time (Thingstad, 2000). Complementary to this, the stable viral resident community has been described as the “seed-bank population” from which transient communities emerge (Short, Rusanova and Short, 2011).

A range of studies have demonstrated that environmental factors influence host-virus interactions, including initial infection, replication, burst size and duration of the replication cycle. Initial adsorption and infection of phytoplankton by viruses is affected by temperature, salinity and light radiation. High temperatures inactivate viruses infecting eukaryotic phytoplankton and bacterioplankton (Baudoux and Brussaard, 2005; Hardies *et al.*, 2013). Adsorption of viruses to host cells is decreased by increased salinity for some phage systems, and is light-dependent for viruses infecting cyanobacteria of the genus *Synechococcus* (Kukkaro and Bamford, 2009; Jia *et al.*, 2010). Furthermore, the duration

of the lytic cycle and viral production are affected by light regimes in phytoplankton virus systems (Brown, Campbell and Lawrence, 2007; Baudoux and Brussaard, 2008).

Viral production by eukaryotic phytoplankton is furthermore influenced by nutrient availability. For example, viral production was lowered under phosphate depletion and increased under phosphate addition (Bratbak *et al.*, 1998; Jacquet *et al.*, 2002; Maat *et al.*, 2014; Motegi *et al.*, 2015). A similar effect was seen for nitrate availability, but the effect was less pronounced (Bratbak *et al.*, 1998; Jacquet *et al.*, 2002). The effect of nutrient availability on viral production can be explained by the relatively high proportion of phosphorus and nitrogen in viral particles relative to carbon (Jover *et al.*, 2014).

Viral decay is affected by UV radiation, with higher decay rates of viruses infecting heterotrophic bacteria, cyanobacteria, and eukaryotic phytoplankton under the higher radiation in the photic zone (Cottrell and Suttle, 1995a; Noble and Fuhrman, 1997; Garza and Suttle, 1998). The rate of the decay is also correlated to the GC content of viral genomes (Kellogg and Paul, 2002) and can be compensated for by photoreactivation (Wilhelm, Weinbauer, Suttle and Jeffrey, 1998; Wilhelm, Weinbauer, Suttle, Ralph, *et al.*, 1998). In addition to these factors influencing lytic viruses, the switch from a lysogenic to a lytic state can be prompted by increased temperatures, nutrient availability, and UV radiation (Williamson and Paul, 2006; Payet and Suttle, 2013).

Many environmental factors that affect host-virus interactions change rapidly including ocean mixing, weather patterns, and light, which could lead to transient viral populations emerging out of seed-bank populations. To a degree, this pattern has been shown in myovirus communities from varying environments (Filée *et al.*, 2005; Comeau and Krisch, 2008). In contrast, other studies have shown persistence and gradual, long-term variations in viral communities. For example, using the *gp20* marker gene, the same myovirus genotypes were found in wide ranging environments including the Arctic and Pacific oceans, as well as in freshwaters (Short and Suttle, 2005). A high frequency sampling study of *gp23* sequences also showed persistence of some genotypes over long periods (Needham *et al.*, 2013), as well as gradual changes and recurring seasonal patterns (Goldsmith *et al.*, 2011; Chow and Fuhrman, 2012; Goldsmith, Parsons and Beyene, 2015). Similarly, the *gp43* marker gene revealed great diversity and recurring

seasonal patterns in myovirus communities infecting *Synechococcus* spp. (Marston *et al.*, 2013).

However, the aforementioned studies were all based on communities of free virus particles. Arguably, the composition of free virus communities represents the accumulated viruses that have been produced by past lytic events, perennial communities, and may not reflect the viruses that are momentarily replicating under the existing conditions. The viruses that are replicating at a specific moment, under specific conditions would be better represented by viruses in the cellular fraction, rather than the free virus fraction.

A good model to examine the dynamics between viruses in the free and cellular fractions are myoviruses infecting *Synechococcus* spp., an abundant and widely distributed marine phytoplankton genus (Liu *et al.*, 1998; Scanlan and West, 2002). Members of the genus *Synechococcus* are found from the poles to the tropics and from coastal to oligotrophic oceanic environments, and show strong seasonal dynamics. Moreover, cyanomyoviruses have a diverse genetic repertoire including proteins for photosynthesis (PsbA/D, PebS, HliP), carbon metabolism (TalC) and phosphate stress (PhoH, PstS) (Mann *et al.*, 2003; Lindell *et al.*, 2005; Sullivan *et al.*, 2005; Dammeyer *et al.*, 2008; Thompson *et al.*, 2011) that potentially confer fitness advantages under different environmental conditions. Furthermore, some of these genes are widely distributed among isolates while others are rare (Sullivan *et al.*, 2005; Clokie, Millard and Mann, 2010; Puxty *et al.*, 2014; Crummett *et al.*, 2016), and some studies have linked gene content to the environmental conditions when the viruses were isolated (Williamson *et al.*, 2008; Kelly *et al.*, 2013; Crummett *et al.*, 2016). Cyanomyoviruses also have wide host ranges (Lu, Chen and Hodson, 2001; McDaniel, delaRosa and Paul, 2006; Hanson, Marston and Martiny, 2016), reducing the impact of host community variations on the composition of virus communities. Cyanomyovirus phylogeny can be studied by DNA polymerase (*gp43)* which has been developed and used as a reliable marker gene (Marston and Amrich, 2009; Marston *et al.*, 2013). Additionally, an earlier project demonstrated that gp43 amino acid sequences are not only effective to describe cyanomyovirus communities, but also represent relative similarity in the gene content of viruses (Chapter 4 of this thesis). This is in line with the recent formulation of viral ecotypes (Marston and Martiny, 2016) and makes it possible to assess the replication dynamic of such ecotypes in the environment.

This project investigates the composition of free and cellular cyanomyovirus communities under differing environmental conditions, with the assumption that the cellular communities reflect replicating viruses. This is done at the Operational Taxonomic Unit (OTU) level using *gp43* as a marker gene. The goal is to assess the mismatch between free and cellular cyanomyovirus communities and to identify the environmental variables that influence the composition of the cellular and free cyanomyovirus communities. In turn, this elucidates how subsets of the viral community are drawn from the free community, the seed-bank, and provides insights into virus host dynamics and their response to environmental variables.

5.3 Materials and methods

Sampling

Integrated samples from the surface throughout the mixed layer to the deep chlorophyll maximum (DCM) were taken at nine sites from contrasting environments in the Strait of Georgia in 2011. A total of 200 liters of water were sampled with Niskin bottles (General Oceanics, Miami, FL) and processed immediately. For the viral fraction, samples were pre-filtered through 47 mm GC50 and 0.45 µm HVLP filters (Millipore), concentrated by tangential flow filtration (TFF, Prep-Scale) with a 30 kDa cutoff (Merck Millipore, Billerica, MA) to 500 ml and stored at 4 °C until further use. For the cellular fraction three to five liters of water from the cyanobacterial peak, as determined by flow cytometry (see below) within the mixed layer were filtered through 47 mm 0.45 µm HVLP filters (Millipore), flash frozen in liquid nitrogen and stored at -80 °C until processing.

Environmental data collection and processing

To determine sampling depths and sampling conditions, *in situ* profiles of temperature, salinity and depth were measured with a rosette-mounted CTD SBE 25 (Seabird Electronics, Inc., Bellevue, WA). *In situ* chlorophyll concentration was assessed with a fast-repetition-rate fluorometer (FRRF). Oxygen concentration was measured by a SBE 43 (Seabird Electronics Inc., Bellevue, WA) oxygen sensor and photosynthetically active radiation (PAR) with a QSP-200PD sensor (Biospherical Instruments, San Diego, CA).

Samples for nutrient analyses were filtered through 0.22 μm pore-size PVDF syringe filters and stored at -20 °C. Nitrate and nitrite, phosphate and silicate were analyzed by a Bran & Luebbe AutoAnalyzer3 (SPX-Flow, Norderstedt, Germany) using air-segmented continuous-flow analysis. Briefly, reduced nitrate and silicate were detected by a colorimeter at 550 nm (Armstrong, Stearns and Strickland, 1967) while reduced orthophosphate was read at 880 nm (Murphey and Riley, 1962).

Cyanobacterial, viral and heterotrophic bacterial abundance were measured using a Beckton Dickinson FACSCalibur flow cytometer with a 15 mW 488 nm air-cooled argon ion laser. If necessary, samples were diluted in TE buffer (pH 8.0) to reach 100 to 1000 events s^{-1}. While cyanobacterial measurements were performed onboard, samples for viruses and heterotrophic bacteria were fixed for 15 min. at 4 °C in the dark with electron-microscopy-grade glutaraldehyde (25 %), final concentration 0.5 %, followed by snap-freezing in liquid nitrogen and storage at -80 °C. Cyanobacteria were measured based on chlorophyll autofluorescence and discriminated from picoeukaryotes based on their phycoerythrin signal.

Viral and bacterial samples were thawed and diluted in 0.2 μm filtered, autoclaved TE 10:1 buffer pH 8.0 (10 mM-Tris HCl; 1 mM EDTA) and stained with SYBR Green I (Invitrogen, Carlsbad CA) at a final concentration of $0.5x10^{-4}$ of the commercial stock, incubated for 10 min at 80 °C for virual samples and for 15 min at room temperature for bacterial samples (Brussaard, 2004). Viruses and bacteria were discriminated by plotting green fluorescence against side scatter. Flow cytometry data were analyzed with CYTOWIN version 4.31 (Vaulot, 1989) and WEASEL version 3.3 (Battye, 2015).

DNA extraction

Viral DNA was extracted from 25 ml of viral concentrate. The sample was syringe filtered through 0.22-μm pore-size GV PVDF Millex filters (Merck Millipore, Billerica, MA) and centrifuged for 6 h at 120,000 g and 8 °C. After discarding the supernatant, viral pellets were resuspended in 500 μl 1xTE buffer at 4 °C overnight. To remove free DNA, the suspension was treated with 5 μl DNase I (Invitrogen, Carlsbad, CA) at 37 °C for 15 min, followed by inactivation with 10 μl EDTA (0.25 M) at 65 °C for 15 min. Viral capsids were lysed by adding 60 μl Proteinase K (Invitrogen, Carlsbad, CA) at 56 °C for 15 min, viral

DNA was extracted with Pure Link Viral RNA/DNA columns (Invitrogen, Carlsbad, CA) following the manufacturer's instructions. DNA was eluted in UltraPure water (Invitrogen, Carlsbad, CA).

DNA from the cellular fraction was extracted with the Power Water DNA Isolation Kit (MoBio, Carlsbad, CA) following the manufacturer's instructions. Briefly, frozen filters were cut in half under sterile conditions and half was processed by bead beating for 5 min in lysis buffer. The debris was separated by centrifugation and the cellular DNA in the supernatant was bound and washed in the proprietary column; DNA was eluted in 100 µl of elution buffer.

Marker gene (gp43 & rpoC1) amplification

Samples from the free viral fraction and the cellular fraction were treated identically in the following steps. Prior to PCR, DNA concentration in the eluent was quantified with a Qubit 2.0 using the dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA). To assess cyanomyovirus communities, a fragment *gp43* was amplified from viral and cellular fractions by PCR with the primers from Marston et al. (2013) producing amplicons of about 475 bp in length, in a two-step, large scale PCR with an accumulative 35 cycles based on 1-2 ng of template DNA. The PCR cycle consisted of an initial denaturing step at 94 °C for 3 min, denaturing at 94 °C for 45 s, annealing at 50 °C for 45 s, extension at 72 °C for 45 s and a final extension at 72 °C for 10 min. Triplicate PCR products were pooled and run on a 0.8 % Ultrapure LMP agarose gel (Invitrogen, Carlsbad, CA). DNA bands of the appropriate size were plugged and extracted with the Zymoclean Gel DNA Recovery Kit (Zymo, Irvine, CA). Purified DNA was eluted in UltraPure water (Invitrogen, Carlsbad, CA) and the DNA concentration quantified by Qubit; aliquots were stored at -20 °C. To estimate cyanobacterial diversity a section of the *rpoC1* gene was amplified by PCR using the N5 and C-terminal primers and PCR protocols previously published (Palenik and Haselkorn, 1992; Mühling *et al.*, 2006).

Sequencing library preparation

For library preparation, 500 ng of DNA was used with the NxSeq Low DNA AmpFREE kit (Lucigen, Middleton, WI) following the manufacturer's protocol with NextFlex-96

sequencing adapters (Bioo, Austin, TX). Purification and size selection of libraries was done with Agencourt AMPureXP beads (Beckman Coulter, Pasadena, CA), followed by elution in low TE. The success of library construction was confirmed with a Bioanalyzer 2100 using High-Sensitivity DNA Chips (Agilent, Santa Clara, CA).

Sequencing

Libraries were quantified by Q-PCR with SSoFast Eva Green Supermix (BioRad, Hercules, CA) and KAPA DNA Standard (KAPA Biosystems, Boston, MA) on a C10000 Touch PCR block with a CFX 96 head (BioRad, Hercules, CA). They were pooled for equal template concentration and sequenced in two rounds at UCLA (Los Angeles, CA) and McGill (Montreal, QC) using the 2x300 HiSeq paired-end technology (Illumina, San Diego, CA).

Bioinformatic processing

Sequences were trimmed to a quality thread score of 30 and a minimum length of 36 nt using TRIMMOMATIC 0.33 (Bolger, Lohse and Usadel, 2014). Paired, overlapping reads were merged with USEARCH 8.1 (Edgar, 2010) and translated with FragGeneScan 1.20 (Rho, Tang and Ye, 2010). Cleaned and size-selected reads with a minimum length of 140 amino acids were then de-replicated per sample with USEARCH 8.1 before being pooled. Pooled gp43 reads were again de-replicated, clustered for OTUs and tested for chimeras with USEARCH 8.1 at 97% amino-acid identity; singletons were removed. Representative OTUs were selected based on similarity to cyanophages using BLAST-P with a cut-off E-value of 10^{-3}. A phylogeny was built by placing OTUs on a full-length DNA polymerase reference tree made with the Evolutionary Placement Algorithm (EPA) in RaXML 8.0.0. (Berger and Stamatakis, 2011; Stamatakis, 2014). Environmental reads per sample were parsed to the OTUs with UPARSE 8.1 (Edgar, 2013) at an amino-acid identity of 97 %.

Cyanobacterial reads (*rpoC1*) were de-replicated, chimera-checked and clustered in USEARCH 8.1 to define OTUs. OTUs were selected for cyanobacteria similarity using best hits in a BLAST-P analysis with a cut-off E-value of 10^{-3}. Confirmed cyanobacterial OTUs were aligned with reference sequences using Clustal (Sievers *et al.*, 2011) and

trimmed in trimAl v1.2 (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009). The optimal substitution model for the aligned reads was selected in prottest-3.4 (Darriba, Taboada and Posada, 2011) and the maximum-likelihood tree was built in RaxML version 8.0.0. (Stamatakis, 2014) using the VT substitution model, including partial *rpoC1* sequences from reference *Synechococcus* and *Prochlorococcus* spp.

Statistical analyses

Environmental variables associated with the samples were scaled and examined by PCA using the FactoMineR package (Le, Josse and Husson, 2008). Community compositions were rarefied to the lowest total read number using VEGAN (Oksanen *et al.*, 2016). Diversity indices, Bray-Curtis similarities and the correspondence analysis (CA) were also computed in VEGAN. All statistical analyses were performed in R (R, 2015).

5.4 Results

Variability of sampling sites and samples

To compare free and cellular cyanomyovirus communities, nine samples were taken in the Strait of Georgia during a one-week cruise in 2011. Samples were taken from the mixed surface layer, extending to the DCM at locations with contrasting environmental conditions (Figure 5.1). The maximum sampling depth and the corresponding DCM ranged from 8 m to 18 m in the sheltered Simoon Sound and the well-mixed Johnstone Strait, respectively. These samples also correspond to the minimum and maximum nutrient concentrations in the samples.

Similarity and difference among the environmental conditions at the sample sites are described by PCA (Figure 5.2), the first and second dimensions account for 64.49 % (Dim 1) and 19.12 % (Dim 2) of the variation, respectively. Stations are spread out along the first dimension from inlets to more exposed stations. Johnstone Strait presents very different conditions than the other stations. Differences are mainly driven by nutrient concentrations, salinity and temperature and their co-variation with viral and bacterial abundances (Figure 5.2b).

Flow cytometry showed that the picophytoplankton communities were generally dominated by *Synechococcus* spp. cells (Figure 5.3). Abundances of eukaryote picophytoplankton and *Synechococcus* cells ranged from $9.18x10^2$ to $1.59x10^4$ and from $4.89x10^2$ to $2.31x10^5$ cells ml^{-1}, respectively (Figure 5.4). Viral and bacterial abundances ranged from $5.42x10^6$ to $6.86x10^7$ and from $5.35x10^5$ to $2.13x10^6$ ml^{-1}, respectively and are generally higher than picophytoplankton abundances. Cyanobacterial diversity across all samples based on rpoC1 sequences resolved 25 distinct OTUs, most of which branch off in proximity to reference sequences on a ML tree (Figure 5.5). However, the tree also shows two deep-branching clades made up solely of environmental sequences.

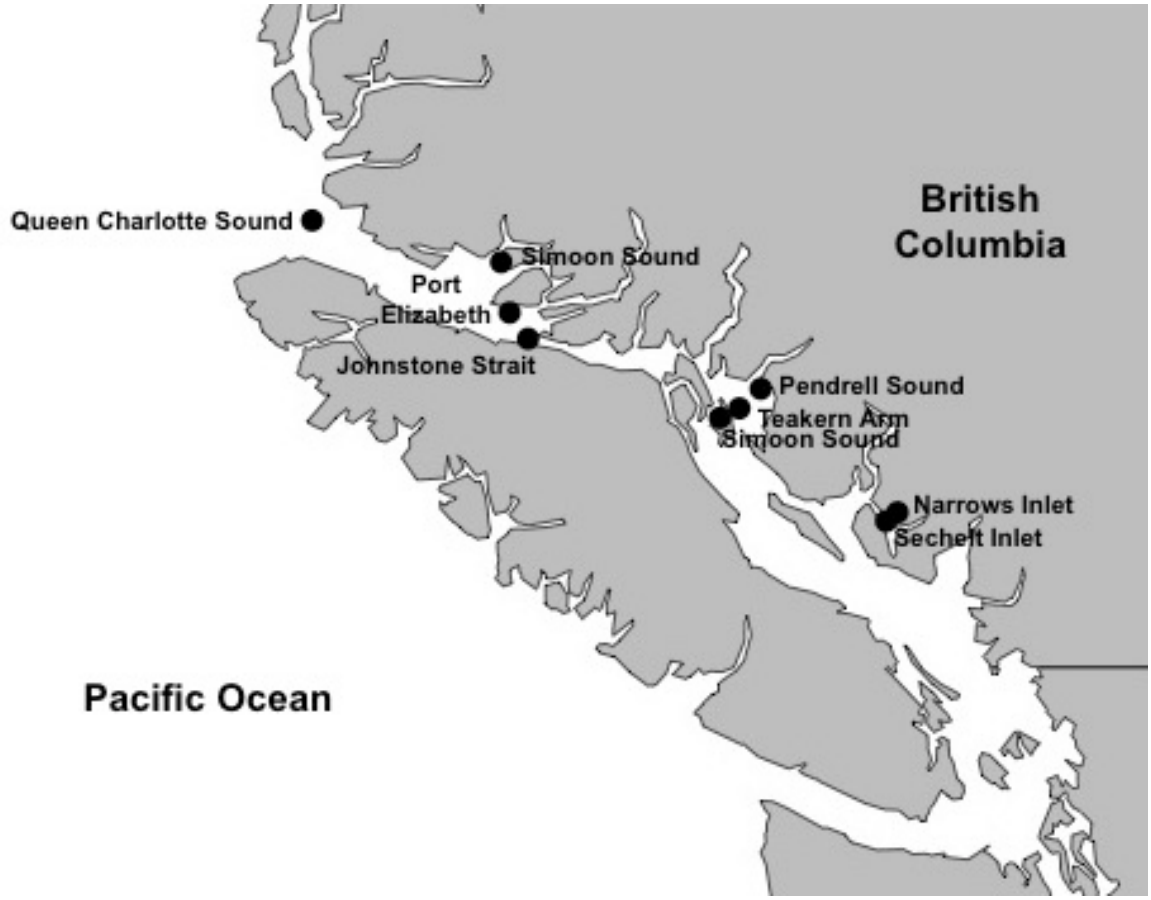

Figure 5.1: Sampling locations in the Strait of Georgia and adjacent waters.
Samples were taken in the Strait of Georgia and adjacent waters during a one-week research cruise in 2011.

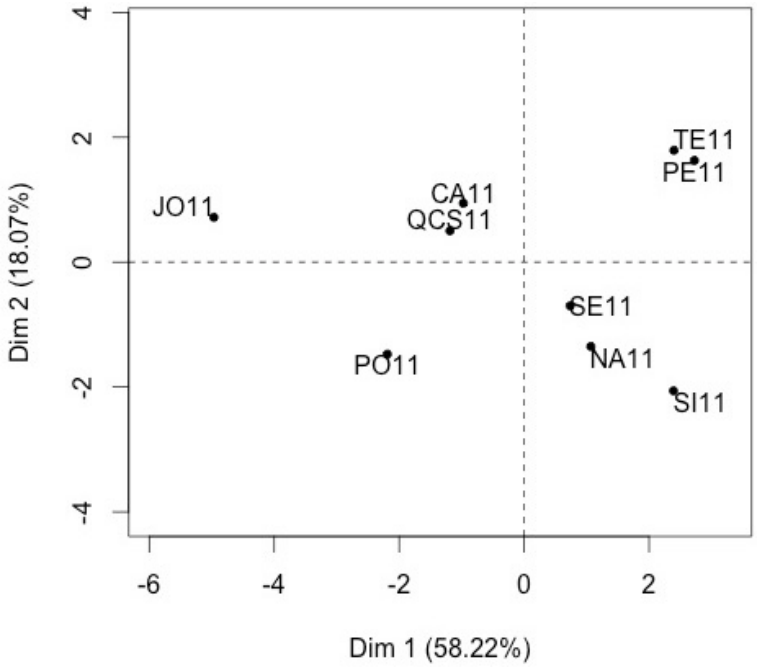

Figure 5.2: PCA of the samples based on environmental variables.
Distance indicates their similarity in environmental conditions. Labels: CA, Carrington Bay; JO, Johnstone Strait; NA, Narrows Inlet; PE, Pendrell Sound; PO, Port Elizabeth; SE, Sechelt Inlet; SI, Simoon Sound; TE, Teakern Arm; QC, Queen Charlotte Sound.

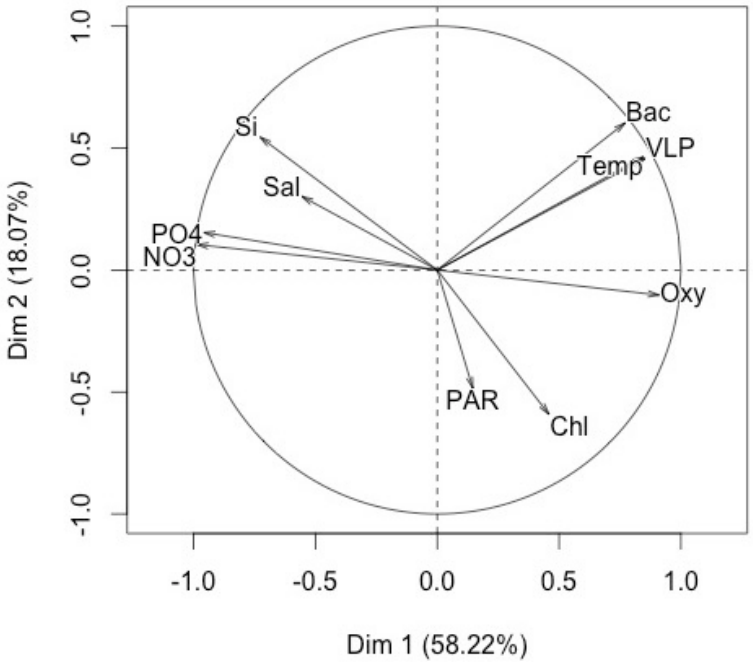

Figure 5.2.b: Environmental factors of the PCA.
Vector direction and length indicate the covariation and relative strength of environmental variables. Labels: Sal, salinity; Temp, temperature; Oxy, Oxygen; NO3, reduced nitrate and nitrite; PO4, phosphate; Si, silicate; VLP, viral abundance; Bac, bacterial abundance.

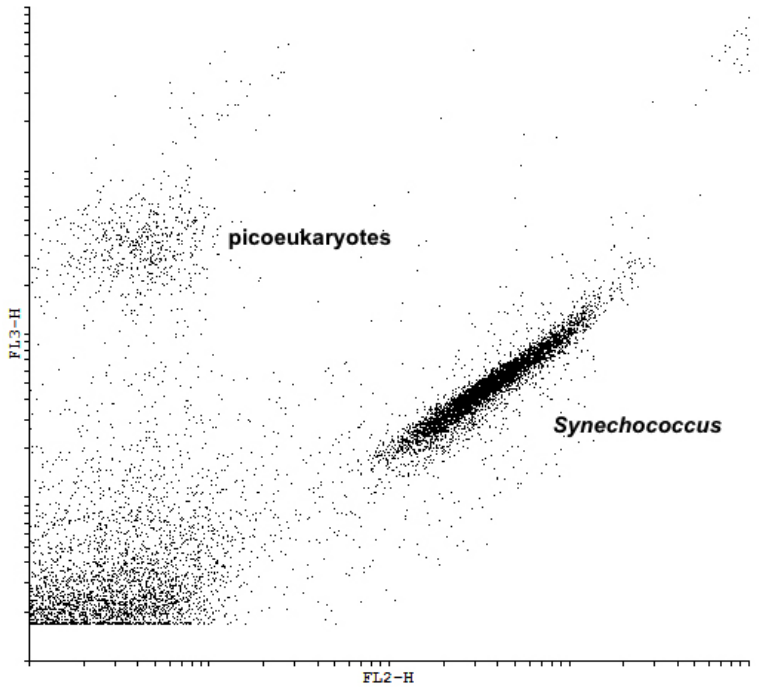

Figure 5.3: Flow cytometry scatter plot of phytoplankton.
Characteristic plot of sample measuring phytoplankton composition and abundance. Excited with a 488 nm argon laser, FL-2 orange fluorescence (phycoerythrin), FL-3 red fluorescence (chlorophyll).

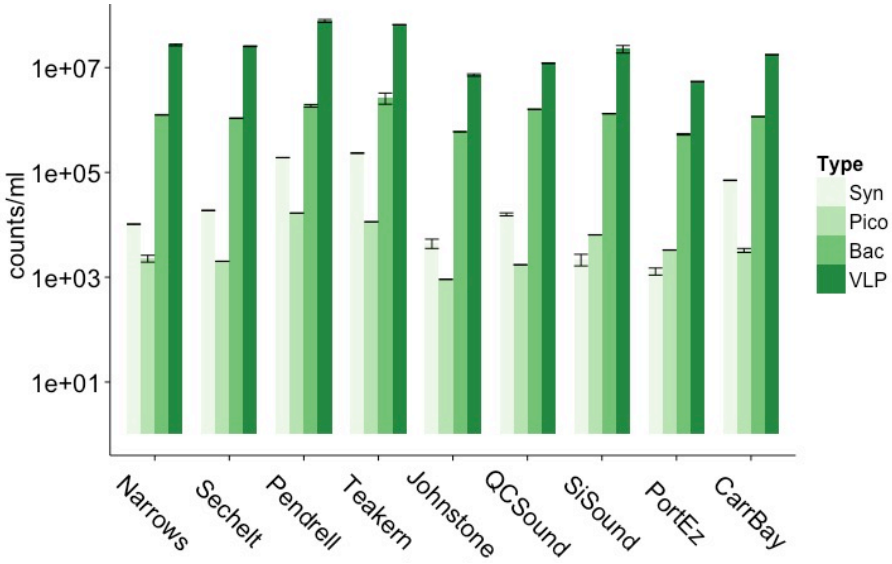

Figure 5.4: Plankton counts by flow cytometry.
Abundances of viruses, heterotrophic bacteria, cyanobacteria (*Synechococcus*) and picoeukaryote phytoplankton in the samples. Counts are averages of replicates, error bars show the standard deviation. Labels: CarrBay, Carrington Bay; Johnstone, Johnstone Strait; Narrows, Narrows Inlet; Pendrell, Pendrell Sound; PortEz, Port Elizabeth; Sechelt, Sechelt Inlet; SiSound, Simoon Sound; QCSound, Queen Charlotte Sound; Syn, *Synechococcus*; Pico, Picoeukaryotes,; Bac, Bacteria; VLP, Virus like particles.

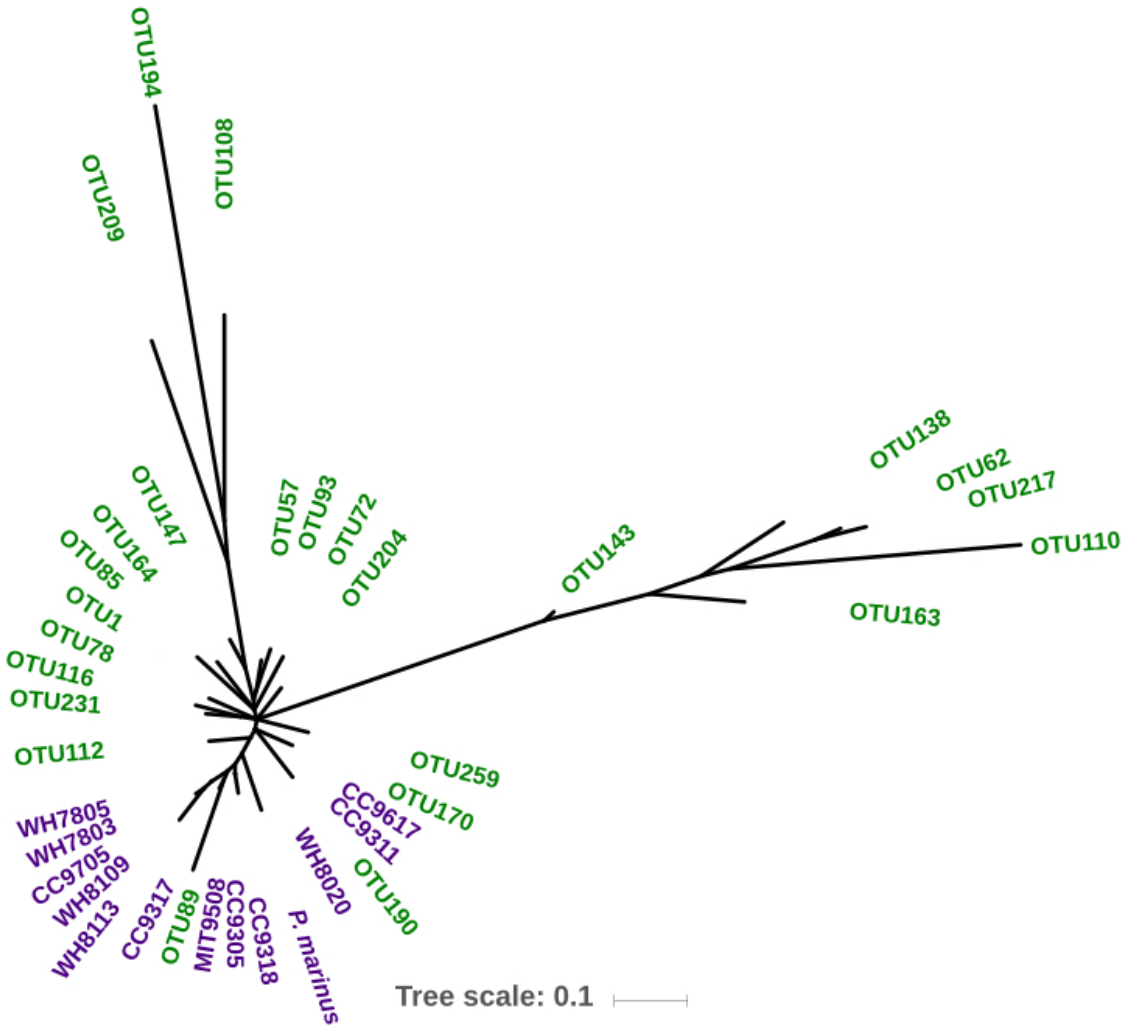

Figure 5.5: ML phylogeny of cyanobacteria OTUs based on rpoC1.
Phylogenetic tree of partial rpoC1 sequences from environmental cyanobacterial reads and reference sequences. Reference sequences are highlighted in purple, novel clades with no known reference are highlighted in green. Scale bar represents the substitution rate.

Free and cellular cyanomyovirus communities vary in composition and diversity

To compare free and cellular cyanomyovirus communities, environmental sequences were recruited to 625 OTUs, the 606 most abundant environmental OTUs and 19 reference sequences. These OTUs were derived from a larger pool of 12,200 OTUs, but represent 97% of the initial sequences (Chapter 4). The phylogenetic relationship of the OTUs was described by a EPA tree based on 97 % aa identity (Chapter 4, Figure 4.5).

The community compositions show that certain clades are dominant across samples from different geographical locations, but there is also shifts of dominant OTUs within the clades (Figure 5.6). Dominant clades across locations for free communities (A) and cellular communities (B) are I, IV, V, VIII, X and XIV. Furthermore, the dominant and abundant OTUs in the cellular samples are often differing from their free counterparts.

In order to assess how large a fraction of the free communities is actively replicating, diversity indices of free and cellular communities were compared. Diversity indices of the free and the cellular communities showed a wider range of alpha-diversities for the free communities, from 2.28 to 4.98, compared to 3.73 to 4.14 in the cellular fraction (Figure 5.7). Yet, the overall beta-diversity (Shannon) proved to be slightly higher in the cellular fraction (2.02) than in the free communities (1.82). Similarly, the range in species richness was larger in the free communities, 193 to 484, than in the cellular fraction, 273 to 357 (Figure 5.8). The discrepancy in alpha diversity between free and cellular communities was most pronounced in Narrows Inlet, Sechelt Inlet, Johnstone Strait and Port Elizabeth (Table 5.1).

A pair-wise Bray-Curtis similarity analysis of free and cellular cyanomyovirus communities was performed to identify samples of high similarity and dissimilarity. Bray-Curtis similarity values range from the most similar at 0.16 (cellular Sechelt Inlet vs. Narrows Inlet) to the most dissimilar at 0.95 (Port Elizabeth cellular vs. free). The results showed that cellular communities are more similar (green) to each other than are the free communities (Figure 5.9). This especially showed for Carrington Bay, where the free community is very different from most other free communities. What also stands out is that the distances between the corresponding free and cellular communities from the same samples (diagonal cells) are relatively different, especially Narrows Inlet, Sechelt Inlet, Pendrell Sound, Johnstone Strait and Port Elizabeth.

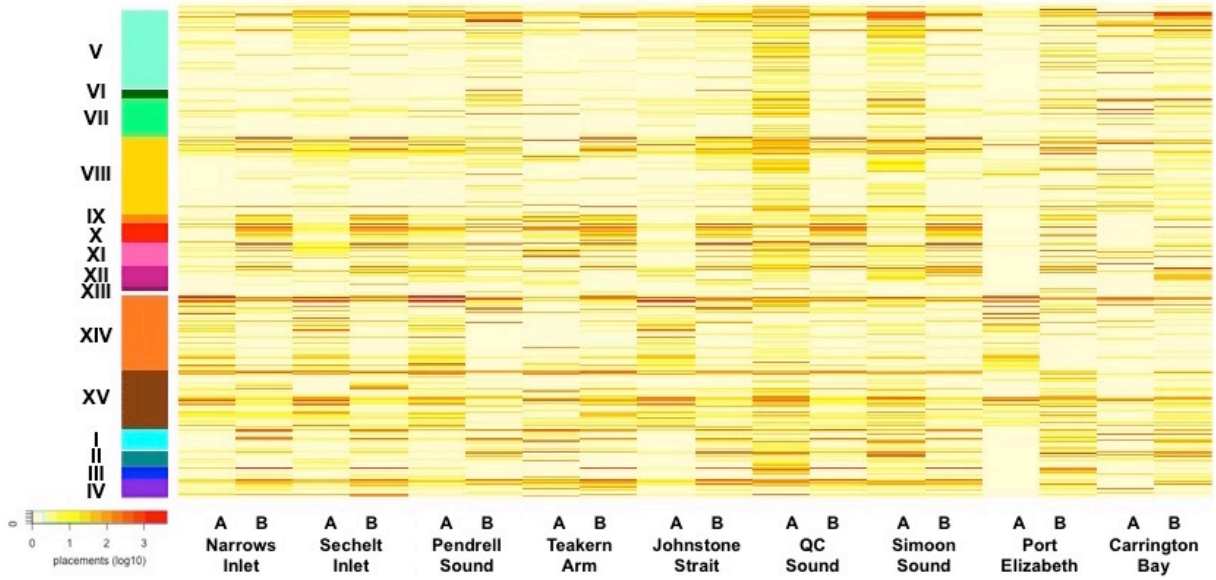

Figure 5.6: Community composition of free and cellular samples.
Samples from the free virus communities (A) and virus communities in the cellular fraction (B), communities are rarefied, heat indicates relative abundance. Samples are arranged in columns by year, OTUs are arranged in rows with clades indicated by color and number in correspondence to figure 5, labels indicate sampling month and year.

Table 5.1: Diversity indices for free and cellular communities.
Assessed per sampling site for cyanomyovirus communities in the free fraction and the cellular fraction, diversity is Shannon alpha diversity, richness is species richness.

	free		cellular	
	diversity	richness	diversity	richness
Narrows Inlet	2.52	287	3.93	275
Sechelt Inlet	2.86	369	4.08	323
Pendrell Sound	3.08	369	3.55	314
Teakern Arm	4.12	283	4.06	297
Jonhstone Strait	2.90	329	4.04	365
QC Sound	4.98	482	3.73	284
Simoon Sound	4.49	403	3.86	338
Port Elizabeth	2.28	193	4.15	279
Carrington Bay	3.43	188	3.89	359

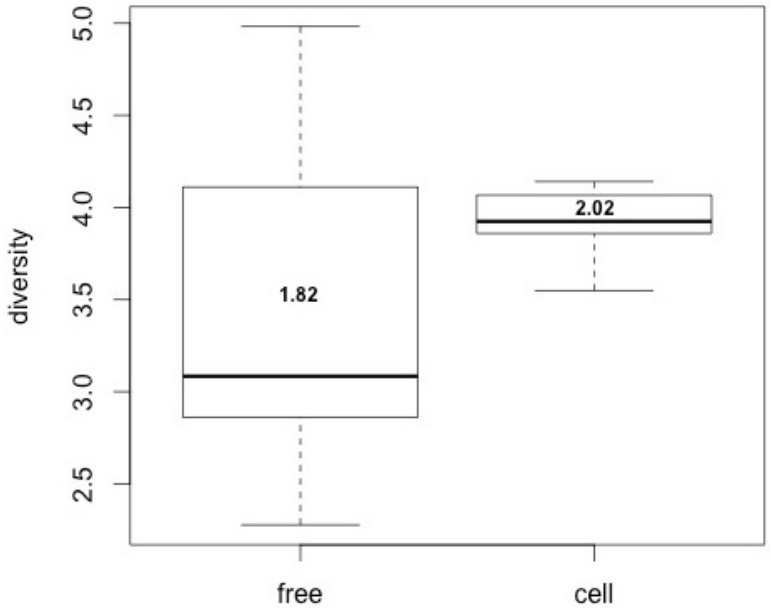

Figure 5.7: Range of diversity for free and cellular cyanomyovirus communities.
Diversity is the Shannon alpha diversity assessed over nine samples each; numbers in boxes are Shannon beta diversities for the free vs. cellular subsets. Whiskers indicate the range, box 50 % of data points with median.

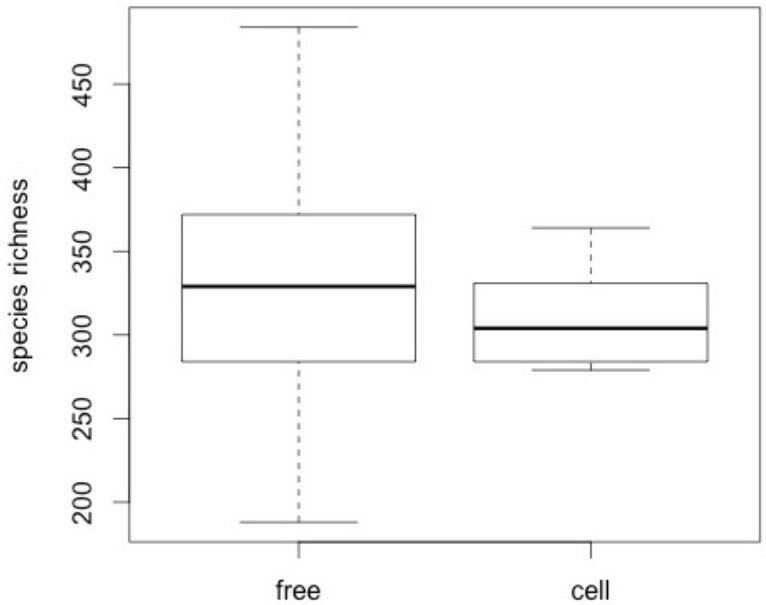

Figure 5.8: Range of richness for free and cellular cyanomyovirus communities.
Richness is defined as species richness and assessed over nine samples each. Whiskers indicate the range, box 50 % of data points with median.

To compare all communities with each other, a correspondence analysis was performed. The first dimension in the correspondence analysis (CA1) explains 16.4 % of variation and the second dimension (CA2) accounts for 15.2 %. The results showed that the free communities generally form two distinct clusters and vary strongly in the first and second dimension (Figure 5.10). In contrast, the cellular communities show high similarity and mainly vary in the second dimension. Also, the cellular communities appear more similar to each other than to their corresponding free communities. The free community for Carrington Bay is very distant from the other free communities, with its cellular equivalent being the most similar. Overall only the free and cellular communities for Teakern Arm clustered closely together, showing high similarity. The individual OTUs shown as solid grey circles are mainly concentrated around the cluster of cellular communities, but some OTUs show an abundance pattern characteristic for the cluster of free communities and for Carrington Bay.

Next, a canonical correspondence analysis with fitted environmental data was performed to identify variables that covary with sites that show high and low dissimilarities between their free and cellular cyanomyovirus communities. Therefore, the ratios between relative free abundance and relative cellular abundance at the OTU level were compared per site in a correspondence analysis (Figure 5.11). Sites are placed based on the mismatch between relative OTU abundances in the free and cellular community compositions. The first dimension (CA1) and the second dimension (CA2) explain 19.0 and 18.1 % of the variance, respectively. Narrrows Inlet, Sechelt Inlet, Pendrell Sound, Johnstone Strait and Port Elizabeth cluster together, showing a similar degree of mismatch between their free and cellular communities. Queen Charlotte Sound, Simoon Sound and Teakern Arm form a second cluster, while Carrington Bay is again separate to the other sites. The individual OTUs are mainly spread out between the two aforementioned clusters of sites, while some are clustered around Carrington Bay. Covariation of environmental variables to the sites are indicated by vectors based on fitted data. Vectors were scaled by the strength of their determination coefficient (R^2). Hence the direction and length of the vectors indicate the covariation of environmental variables with sites and the strength of the covariation. The vectors describe three axes, PAR extends towards the first cluster of sites, Narrrows Inlet, Sechelt Inlet, Pendrell Sound,

Johnston Strait and Port Elizabeth, while virus abundance, bacterial abundance, chlorophyll, oxygen, temperature and salinity extend towards the second cluster. The nutrients follow a separate axis. The environmental variables with the strongest covariation to sites proved to be bacterial abundance (0.28), salinity (0.26), PAR (0.26) and silicate (0.5). However, all these relationships have low significance values.

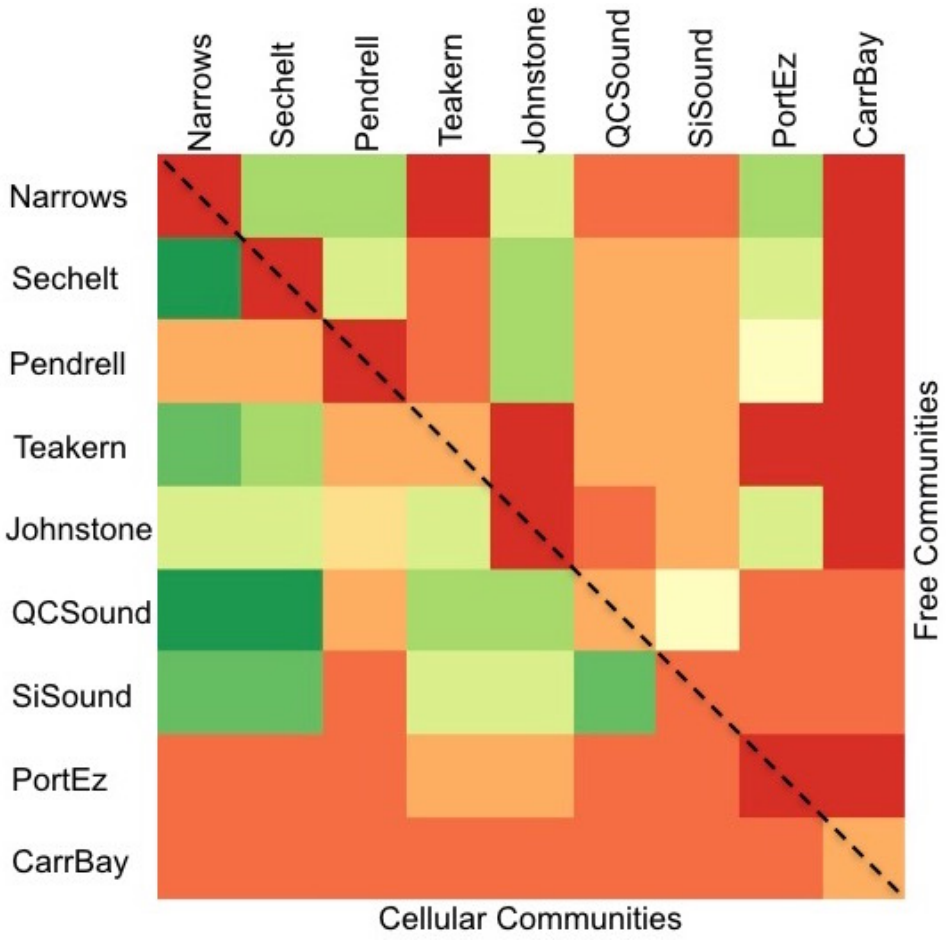

Figure 5.9: Pairwise Bray-Curtis similarities of free and cellular cyanomyovirus communities.
Free communities top-right and cellular communities bottom-left. Diagonal cells are pairwise free vs. cellular communities of the same site. Green indicates relatively high similarity, red relatively high dissimilarity. Labels: CarrBay, Carrington Bay; Johnstone, Johnstone Strait; Narrows, Narrows Inlet; Pendrell, Pendrell Sound; PortEz, Port Elizabeth; Sechelt, Sechelt Inlet; Si Sound, Simoon Sound; Teakern, Teakern Arm; QCSound, Queen Charlotte Sound.

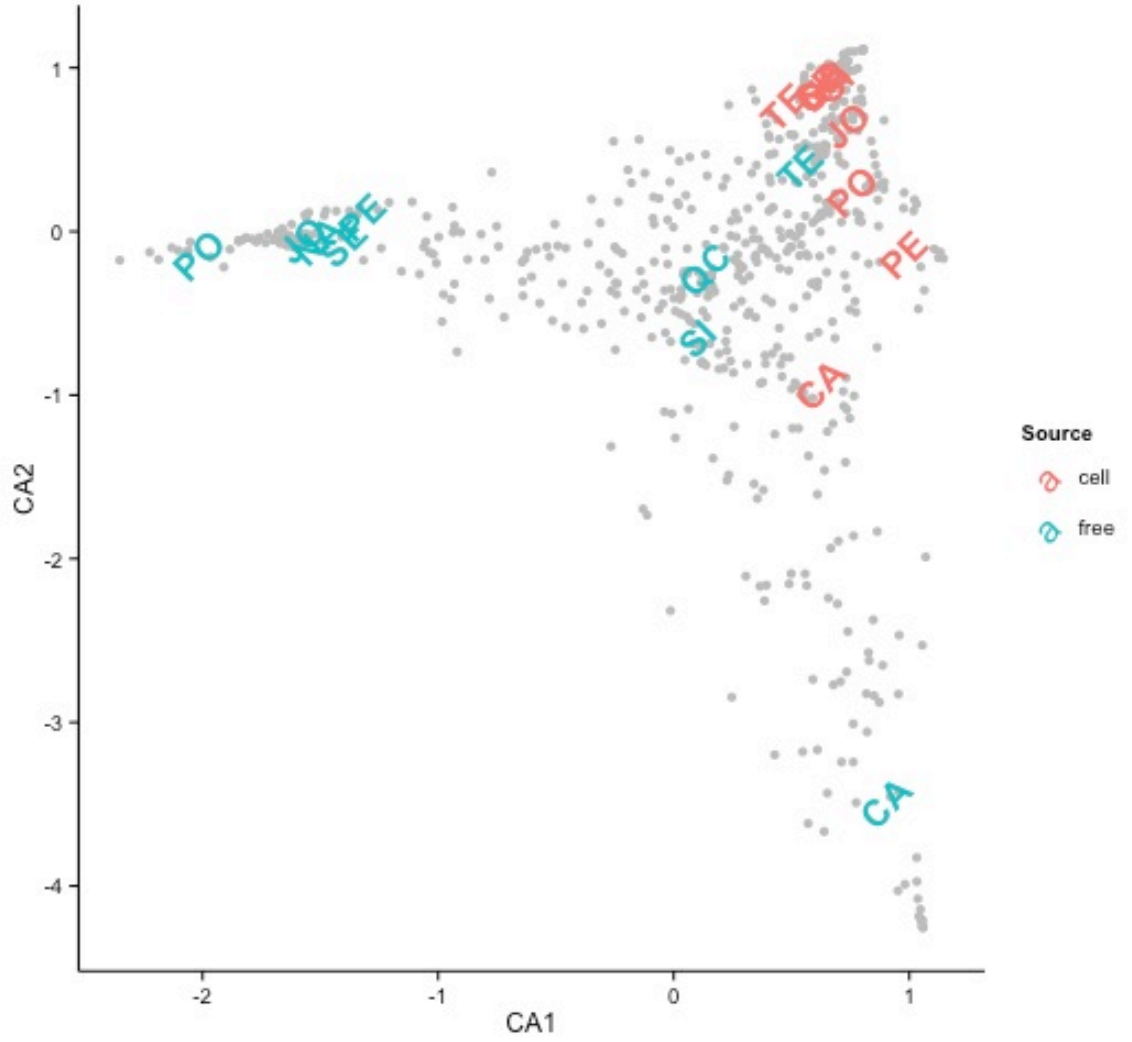

Figure 5.10: CA of free and cellular cyanomyovirus communities.
Distance indicates community similarity, characteristic distribution of OTUs to communities shown. Free communities in blue, cellular communities in red, OTUs as solid grey circles. Labels: CA, Carrington Bay; JO, Johnstone Strait; NA, Narrows Inlet; PE, Pendrell Sound; PO, Port Elizabeth; SE, Sechelt Inlet; SI, Simoon Sound; TE, Teakern Arm; QC, Queen Charlotte Sound.

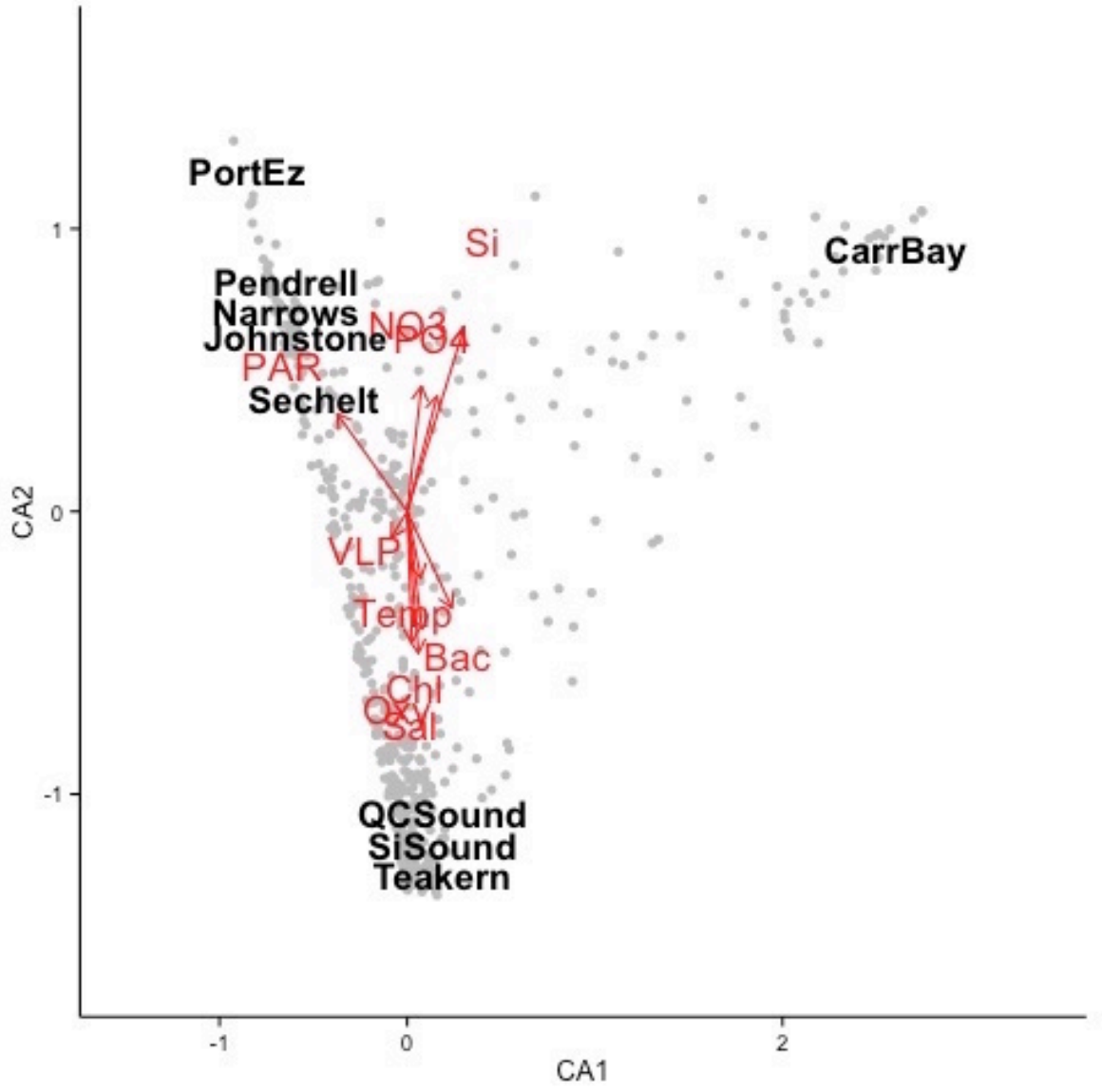

Figure 5.11: CCA of the OTU ratios of free to cellular cyanomyovirus communities.
OUT ratios were assessed per site with characteristic OTU distribution shown (solid grey circles). Environmental variables were fitted to the ratio data and scaled by the strength of covariation. Arrow direction and length describe the covariation of an environmental variable with a site and the strength of covariation. Labels: CarrBay, Carrington Bay; Johnstone, Johnstone Strait; Narrows, Narrows Inlet; Pendrell, Pendrell Sound; Port Eliz, Port Elizabeth; Sechelt, Sechelt Inlet; SiSound, Simoon Sound; Teakern, Teakern Arm; QCSound, Queen Charlotte Sound; Sal, salinity; Temp, temperature; Oxy, oxygen; Chl, chlorophyll; PAR, photosynthetically active radiation; NO3, nitrate; PO4, phosphate; Si, silicate; VLP, viral particle abundance; Bac, bacterial abundance.

5.5 Discussion

Cyanomyoviruses and their hosts exist under a range of environmental conditions that influence viral infectivity, replication and degradation rates. However, the interplay between environmental variables and which genotypes within the viral seed bank infect and replicate within host cells is unknown. This study demonstrates that, in general, viral communities in the cellular fraction are distinct from free communities in the water, and that the similarity among cellular communities across locations can be greater than between corresponding cellular and free communities at the same location. This implies that there is widespread similarity across locations in terms of which viruses from the seed bank actively replicate within the cellular communities. Hypothetically, this could be caused by wide reaching regional factors selecting on viral infection or variable levels of degradation of viral particles. These findings and their implications are discussed in detail below.

Variability of sampling sites and samples

The sites sampled in this study spanned a variety of environments including sheltered inlets, sounds, and open straits and coastal areas with a range of mixed-layer depths (Figure 5.1). The differences in environmental conditions among locations are reflected by the PCA analysis (Figure 5.2), which emphasizes that deeply mixed Johnstone Strait differs from other locations because of the associated high salinity and nutrient concentrations in the surface layer. The other well-mixed sites, Queen Charlotte Sound and Carrington Bay, cluster nearest to Johnstone Strait. In contrast, Narrows Inlet, Sechelt Inlet, Pendrell Sound, Teakern Arm and Simoon Sound, which are more stratified, are characterized by lower salinities and nutrient concentrations (Figure 5.2b). Port Elizabeth, with its low salinity, high nutrient concentrations and low surface temperature is likely affected by the high freshwater outflow from the glaciers at the head of Knight Inlet.

Despite the wide range of environmental conditions sampled, cells with a fluorescence signature consistent with *Synechococcus* spp. dominated the picophytoplankton in flow cytometry counts (Figures 5.3 and 5.4). The overall low number of distinct OTUs based on rpoC1 sequences indicates relatively low richness in the host-cell communities across samples (Figure 5.5). This matches the findings of a study on

Synechococcus communities over an annual cycle, which showed overall only 40 *Synechococcus* types and a low diversity and variability across samples (Mühling *et al.*, 2006). Most sequences clustered closely to reference sequences from isolates of *Synechococcus* spp., consistent with members of this genus being the predominant cyanobacteria in the sampling area. The one reference sequence belonging to an isolate of *Prochlorococcus* spp. (*P. marinus*) had no environmental match, but members of this genus would not be expected to occur in the temperature range of the sampling locations. The distant OTUs on the protruding branches also proved to be cyanobacteria, but given their phylogenetic distance they potentially are of fresh water origin.

Cyanomyovirus community composition and diversity indices

The community composition of cyanomyoviruses was assessed based on 606 environmental OTUs which, based on 97 % aa identity, captured ~97 % of the 9.84 million initial sequences (Chapter 4). All these OTUs showed high similarity to cyanomyoviruses by BLAST-P analysis, and their phylogenetic relationship serves as a proxy of similarity in genetic content. Furthermore, it is assumed that viruses in the cellular fraction represent replicating viruses, while free viruses represent a seed bank of viruses from past lytic events and dispersal that serves as a reservoir from which new infections emerge. This enables one to compare viral communities in the free and cellular fractions, allowing to deduce the interplay of viral genetic repertoires and environmental variables.

The composition of the free and cellular virus communities reveals dominant sets of phylogenetic clusters of OTUs within the free fraction and within the cellular fraction that persist across many samples (Figure 5.6). Moreover, there are alternative sets of OTUs that define other samples. In light of the correlation between phylogenetic distance between pairs of cyanomyoviruses assessed by DNApol and their similarity in gene content, which was established in chapter 4, viruses with different gene content seem to prevail in different environments. The consistency of these patterns suggests that there are specific selection pressures that dictate which virus OTUs dominate.

The ranges in diversity and species richness are larger in the free communities than in the cellular communities, indicating a higher degree of community variation among free communities. Yet the mean diversity and beta-diversity (Shannon) are only slightly

higher in the combined cellular communities (Figure 5.7). Most samples showed a discrepancy in diversity and richness between free and cellular communities. That the free communities show more variability in their diversity and richness than the cellular communities seems counterintuitive when considering that the replicating communities arise out of the dormant free communities and indicates a strong impact of environmental factors on viral communities.

Comparison of community compositions and processes driving them

In pairwise Bray-Curtis similarity analyses of the community compositions, cellular communities showed higher coherence than free viral communities (Figure 5.9). As well, the corresponding free and cellular communities at each site are dissimilar. This is also evident by comparing the pairwise similarities of the communities by CA (Figure 5.10), which shows the high similarity among the cellular fraction and the wide scatter among the free virus communities. Together these results suggest a strong and wide-reaching selection factor that influences which viruses replicate. However, among the free communities there are clusters of similar groups. For example, the free viral communities in Narrows Inlet, Sechelt Inlet, Pendrell Sound, Johnstone Strait and Port Elizabeth show a relatively high degree of similarity, even though the environmental conditions in Johnstone Strait and Port Elizabeth are very different from those at the other locations. The dissimilarities between most corresponding free and cellular communities, except for Teakern Arm, are striking.

To understand the underlying processes, the dissimilarity between the corresponding free and cellular communities was assessed by examining the ratio in the relative abundances of each OTU in the free and cellular fractions in a CA with fitted environmental variables (Figure 5.11). The sites with the highest dissimilarity are grouped together, and are also the ones with the highest discrepancy in alpha diversity. In contrast, Queen Charlotte Sound, Simoon Sound and Teakern Arm were in another cluster, and have more similar free and cellular communities. These also are the sites that displayed a higher alpha diversity in the free communities than in the cellular communities. Interestingly, PAR and nutrient concentrations covary with the sites that have dissimilar cellular and free communities, while salinity and bacterial abundance covary with sites

that show higher similarities. However, the covariation of environmental variables only is a tentative observation due to the low sample size.

Intuitively, one would expect that viruses that are abundant in the cellular fraction would be well represented in the free viral fraction, primarily because lysis would be expected to produce the same viruses that dominate the cellular fraction. The mismatch in diversity and composition between the free and cellular communities at several locations suggests regional effects that select for a subset of the free viral community to replicate. The observation that similar phylogenetic clusters of viruses occur in the cellular fraction across substantially different environments, while they mismatch the corresponding free communities is puzzling.

Regional effects could include the composition of the host community and environmental effects such as nutrients or light. The effects would need to cause synchrony across locations in terms of which viruses from the seed bank of free viruses infect and lyse the host community. This would represent a wide ranging selection pressure for viruses adapted to specific conditions, mirroring the concept of viral ecotypes (Marston and Martiny, 2016). Alternatively, viruses that dominate the cellular fraction are very unstable once released, degrading below the detection limit. Bratbak et al. (1996) measured viral turn over rates of just 10 to 20 min and associated it to synchronized lysis and fast degradation. As a third explanation, viral replication is stalled and particles are not released at all, resembling a form of pseudolysogeny as it has been reported for T4 phage of *Escherichia coli* (Los, Wegrzyn and Neubauer, 2003).

As discussed before, several environmental variables can affect viral production and thus shape communities. Light radiation can influence cyanophage infection dynamics in several ways, and hence could potentially play a role in the observations. For example, cyanophage adsorption to host cells is light dependent (Cseke and Farkas, 1979; Jia *et al.*, 2010) as is completion of the lytic cycle (Suttle and Chen, 1992; Kao *et al.*, 2005). Light may also interact with common AMGs in cyanomyoviruses (Crummett *et al.*, 2016), boosting photosynthesis and viral production (Mann *et al.*, 2003; Lindell *et al.*, 2005; Hellweger, 2009). Furthermore, light radiation causes rapid decay of free viral particles (Suttle and Chen, 1992; Garza and Suttle, 1998; Weinbauer *et al.*, 1999; Wilhelm *et al.*, 2003). Alternatively, while cyanomyoviruses do demonstrate wide host ranges (Lu,

Chen and Hodson, 2001; McDaniel, delaRosa and Paul, 2006; Hanson, Marston and Martiny, 2016) and overall the host community shows little variability, the replication efficiency might vary among different hosts in the environment and could be part of the explanation.

There were two notable findings in this study. The composition of the free and cellular viral communities was highly variable, varying across the sampling locations and environmental conditions. In most cases the community in the cellular fraction, which should represent the actively replicating viruses, was substantially different from the corresponding free community, the putative seed bank. The observations suggest three possible scenarios. First, the observation implies synchrony across locations in terms of which members of the free viral community are infecting and replicating in the cellular fraction, indicating a regional effect on which viruses are infecting and replicating in the host communities. It also suggests that similar host communities are in place in the different environments, a conclusion that is supported by the data. Second, either the infection must be so well synchronized across the broad range of environments sampled that the cellular viruses have yet to populate the free-virus community, or the cellular viruses are very unstable once released. Alternatively, the viruses may remain in the cellular fraction and are not released at all.

It should be emphasized that for each location the cellular fraction was sampled from the depth of maximal *Synechococcus* abundance, while the free fraction was integrated over several depths in the mixed layer. This was done to ensure that the free fraction represents the stable, long term cyanomyovirus community and the cellular fraction represents the actively replicating cyanomyoviruses under specific conditions. There is no reason that the difference in the sampling approach for the cellular and free fraction should lead to the observed higher similarity among cyanomyovirus communities from the cellular fraction than among cyanomyovirus communities from the free fraction across locations.

None of the described scenarios seem very plausible; yet, the observations of similar viruses in the cellular fraction across locations, and typically low similarity with the associated free virus fraction is robust, and needs to be explained.

Chapter 6: Conclusion

6.1 Summary

This thesis investigated the variability of the genetic content of viruses and how environmental variables covary and potentially impact viral community compositions and viral abundance. The results show that viral abundance and community composition are affected by temperature, salinity and nutrient availability. Prasinoviruses and cyanomyoviruses show a high degree of variability in their genetic content which also is reflected by their diversity in the environment. The research was conducted in four projects corresponding to four chapters. The findings and their implications are discussed in detail below.

Chapter 2 studied the effect of environmental variables on the relationship between viral and bacterial abundance, and how this differs among environments. Viral abundance was measured as total indiscriminate abundance of DNA viruses, which was partly due to methodological constraints. Viral and bacterial abundances showed a strong relationship, as has been shown in other studies, but the explanatory power of environmental variables matched that of bacterial abundance when multivariate models were used. Furthermore, environmental variables drastically improved the explanatory power of bacterial abundance data when they were included in multivariate models. The findings indicate that environmental variables such as water temperature, salinity and nutrient accessibility influence virus-host dynamics and thus affect total viral abundance. The set of variables significantly related to viral abundance do, however, differ among environments. In conclusion, environmental variables affect viral production and multivariate models that include environmental variables improve the explanatory power for viral abundance in the environment.

Chapter 3 examined the genetic repertoire of prasinoviruses and how it varied across environments. Viral gene annotation was restricted by limited databases with many genes having an unknown function. Consequently, genomes were compared through clustering by sequence identity which provided a relative similarity in gene content. The data established a core genome for prasinoviruses infecting *M. pusilla*, and also identified a large flexible pan genome that contains cellular derived metabolic genes, spanning

across prasinovirus genera. This variability in genetic repertoire among prasinoviruses was furthermore reflected by their prevalence in environmental samples. The findings indicate a complex evolutionary history of prasinoviruses with frequent genetic exchange and viral genomes being under selection pressure by environmental conditions.

Chapter 4 explored the relationship between environmental variables and the distribution of cyanomyoviruses with different gene content. Viral genotypes were defined from environmental sequences as OTUs. Not being derived from viral isolates, these OTUs represented arbitrary viral genotypes. However, a statistical correlation was established between the gene content of cyanomyoviruses, based on clustering, and the DNA polymerase gene (*gp43*). The data revealed enormous, and unprecedented diversity of cyanomyoviruses in the environment and uncovered that cyanomyovirus community composition is a function of environmental variables. Differences in mixing regimes, driven by salinity and temperature, and the associated nutrient availability were linked to the cyanomyovirus community composition and prevalence of specific viruses. Combined, these data imply that cyanomyoviruses are under selective pressure that is reflected in their gene content and their distribution in the environment. Thus, environmental variables are shaping viral community composition.

Chapter 5 compared free cyanomyovirus communities, representing the seed bank, and replicating cyanomyoviruses in the cellular fraction over a range in environmental conditions. Community compositions were assessed based on cyanomyovirus OTUs, as defined in chapter 4. There was a great diversity in cyanomyoviruses in the cellular fraction that in most cases differed substantially from the composition of the associated free virus community. Furthermore, environmental variables covaried with the degree of mismatch of free and cellular cyanomyovirus communities. Only certain members of the free viral community appeared to be successful at replicating under specific conditions. The observations uncover that the emergence of replicating viruses out of free virus communities is highly variable and under the influence of environmental factors.

The results and insights from these projects advance the field of viral ecology by 1) showing variability and signs of adaptation for the gene content of viral genomes, 2) demonstrating that the similarity of the genetic content of viruses is reflected by marker

gene sequences, so that it can be used to study their distribution, 3) linking environmental variables with viral abundance and community composition and 4) uncovering an unexpected variability between the viral seed bank and actively replicating viruses in the environment. This chapter highlights the contributions made to the field of viral ecology and draws directions for future research.

1) Viral genomes show high variability and signs of adaptation

The thesis shows that phylogenetic diversity is matched by the diversity in gene content. Chapters 3 and 4 identify the core genomes of prasinoviruses and cyanomyoviruses, and show that there are large, variable pan-genomes in both viral groups. These “pan-genes” include metabolic genes, and their presence differs among viruses, and could potentially affect viral replication. Chapter 3 also shows that the closest homologues of metabolic genes in prasinoviruses are associated with a variety of eukaryotes and bacteria.

This research emphasizes the flexibility of viral gene content, including metabolic genes and unidentified genes and indicate a complex path for gene acquisition.

2) The diversity in genetic content of viruses can be assessed through marker genes

Several studies have used a range of viral marker genes to assess the diversity and community composition of marine viruses. A crucial step in creating phylogenetically and ecologically meaningful OTUs is to establish an identity threshold for sequences. Confronted with the challenge of measuring the success of ecotypes, viruses with a specific genetic content fitting an ecological niche, in the environment, chapters 3 and 4 developed an approach that links the similarity in gene content to the similarity in marker gene phylogeny. When a marker gene is carefully selected, it can not only be a phylogenetic tool, but also give an approximation of the pairwise similarity in gene content between viruses. Using an identity threshold for amplicon sequences that is congruent with full-length gene sequences, makes it possible to infer full-length sequences from environmental amplicon sequencing data.

This work establishes a correlation between marker gene diversity and the genetic repertoire of viruses, providing an approximation of metabolic capabilities and thus viral

ecotypes in environmental samples. This understanding opens new interpretations of the composition and shifts of viral communities under consideration of their genetic content.

3) Environmental variables affect viral abundance and shape community composition

The correlation between viral and bacterial abundances is well established. As well, the variability in viral community composition has been documented by several research projects. This research, especially chapter 2 and 4, show that total viral abundance and the community composition of specific phytoplankton viruses vary with environmental variables. Combining several environmental variables in multivariate models showed significant relationships with viral abundance, diversity and community composition. No single variable was significant in its effect, however, oceanographic patterns of stratification and nutrient concentrations were important. Chapter 4 also showed that the composition of cyanomyovirus communities changed gradually throughout the water column and over seasons. This was in contrast to more drastic shifts among different locations. This supports the seed-bank theory of viruses being persistent in their presence and absence, but variable in their relative importance. With a stable seed-bank of viruses, environmental variables drive the success of specific viruses which shapes the community.

The data further the understanding of the pace and scale of shifts in marine virus communities and how environmental variables shape communities. Those community shifts are not driven by isolated variables, but rather by overall environmental conditions.

4) Actively replicating viruses differ from their free virus communities

Several studies have identified variables and processes that affect virus-host interactions at all stages of the replication cycle (Mojica and Brussaard, 2014). However, changes in the composition of free viral communities are limited in studying how viral infections are influenced by transient environmental changes. The results in chapter 5 exposed stark differences in most samples between the composition of communities from the viral and cellular fractions. Furthermore, viral communities in the cellular fraction were typically more similar to each other than to their free equivalents and the degree of community mismatch varied with environmental variables.

The results are among the first to describe actively replicating viruses under the immediate influence of environmental variables, and uncovers a level community dynamic which has been overlooked by previous research.

6.2 Implications

This project approached the connections among environmental variables, viral abundance, viral genome content and community composition from different perspectives. Combining the insights from this and other recent research manifests a picture of viral adaptation and competition. When applying the concept of fitness to viruses (Orr, 2009; Meyer *et al.*, 2012), viruses with a specific genetic repertoire would be more fit than others under specific conditions, resulting in higher infection rates, replication rates and higher abundances. With a better understanding of viral fitness and competition this can be included in advanced ecological, dynamic models on genome content and competition as has been done for bacteria (Louca and Doebeli, 2015). Figure 6.1 summarizes the variables involved in infection, replication, production and degradation of viruses infecting phytoplankton. Temperature and salinity affect stratification of the water column and determine mixed-layer depth, light exposure and accessibility to nutrients. However, auxiliary metabolic genes (AMGs) may provide a means for viruses to compensate for damage to their genome and resource limitation. This work has several implications for viral ecology that are summarized below.

-Community composition of viruses and their abundances are shaped by a complex set of environmental variables. Environmental variables are not isolated in their effect.

-The selection pressure for beneficial metabolic genes in viruses is strong enough to drive frequent exchange of these genes from origins beyond the host range of a virus.

-Viral ecotypes succeed within a community under specific conditions by outcompeting others. Beyond appropriate hosts their success depends on the match between their genetic repertoire and the environmental conditions.

-Free viral communities describe relatively stable communities shaped from the seed-bank. The actively replicating viruses differ from the composition of the free communities and are best described by communities derived from the cellular fraction. Consequently, current research on free virus communities overlooks part of the transient community dynamics.

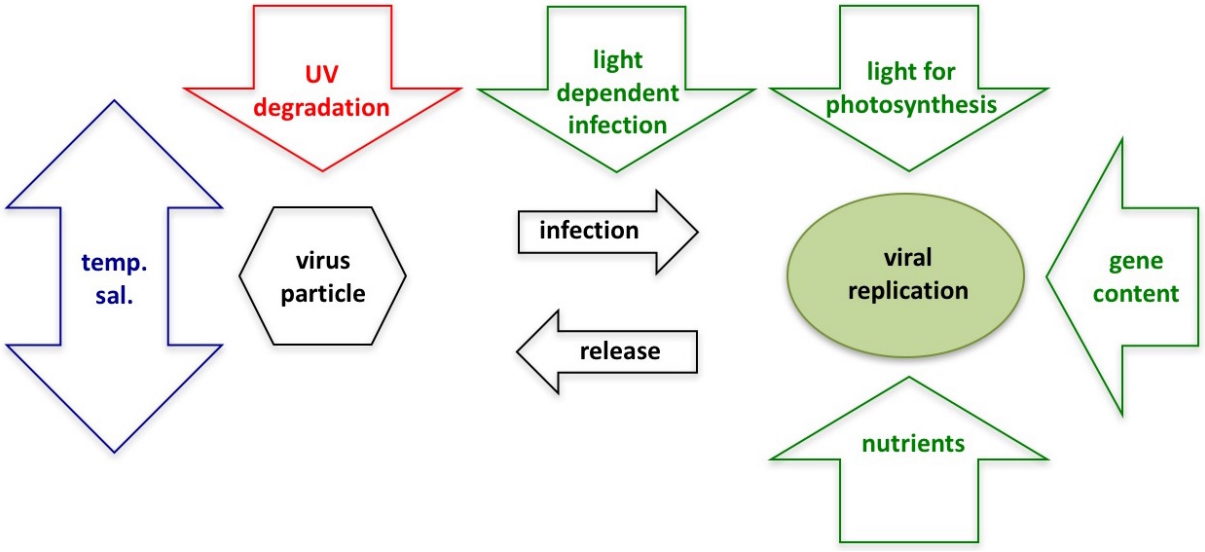

Figure 6.1: Summary of effects of environmental variables on viral replication.
Arrows indicate positive (green) and negative (red) effects of variables at various stages of viral replication. The blue arrow indicates the vertical shift of phytoplankton and their viruses in the water column due to stratification. Temperature and salinity affect stratification and thus the depth of the habitat of phytoplankton cells and their viruses. This alters the exposure to light radiation from the surface or access to remineralized nutrients from deep waters. Metabolic genes, if present, can partially offset environmental adversities.

6.3 Future directions

With the steady advances in sequencing technology and the associated drop in cost it is feasible that future studies on similar questions will result in more samples, more sequences and thus higher statistical significance. More sequencing data would make it possible to build models not on total viral abundance, but on specific abundances of genotypes or groups of viruses. Another measure that would produce a clearer picture is sampling from a wider range of contrasting environments. With the gained understanding from this study, selecting specific areas of nutrient limitation for phytoplankton such as the North Atlantic gyre, where nitrogen and phosphorus are co-limiting (Moore *et al.*, 2013) would be a strong approach to study the effect of these key nutrients on viral replication. As well, sampling programs that keep some variables relatively constant (e.g. light radiation by latitude), but others varying strongly (e.g. nitrogen concentration in the Pacific vs. Atlantic), could reveal isolated effects. As indicated in Chapter 4 and 5, and to a certain degree by Needham et al. (2013), shorter sampling intervals of days or fractions of days would uncover high-frequency community dynamics. Also, improved sequencing technology producing longer amplicons would increase the confidence in the phylogeny and the correlation between environmental sequences and the genetic repertoire.

This work, specifically Chapters 3 and 4, highlights the correlation between gene content and viral distribution. Yet the limited number of genes with a functional annotation resulted in a relative comparison of genetic content among viruses. Each viral genome that is fully annotated increases the size of the databases and in turn enables better annotation of currently unidentified genes. The annotation databases are best increased with information from virus isolates and experimentally determined gene functions; while metagenomic data provide information on the prevalence of genes. Better annotations will help the interpretation of community dynamics and potentially identify genes that are tied to specific environmental conditions. This process is accelerated through new efficient tools for annotation that also incorporate pathway prediction such as MetaPathways (Konwar *et al.*, 2015).

To assess the immediate effect of environmental variables on infection rates and viral “success”, research should focus on dynamics in the cellular fraction. A relatively new and promising method to study this comprehensively is single-cell sequencing. This

approach has been spearheaded by a recent project studying the impact of viruses inside cells in the marine environment (Labonté *et al.*, 2015). While this is not yet a high-throughput method, if combined with RNA transcriptome sequencing, it could give a very concise look at viruses infecting a cell, as well as the expression levels of viral genes.

Without a doubt, advances in molecular, statistical and computational methods open up new ways to study complex ecological interactions from the genome to the global level. Combining high-throughput sequencing, new tools in gene annotation, advanced ecological models and huge swaths of environmental data from remote sensing will improve the understanding of marine ecosystems and enable better predictions about their response to changing environments.

6.4 Conclusion

This project studied the effect of environmental variables on the total viral abundance and the community composition of phytoplankton viruses at various temporal and spatial scales. By combining analysis of environmental conditions with viral abundance and dynamics in community composition, insight was obtained into the effects of environmental variables on viral distribution. Samples were analyzed from temperate, arctic, marine and coastal environments by flow cytometry, high-throughput sequencing, environmental metadata and multivariate statistics. A novel approach to correlate similarity in gene content with marker gene phylogeny enabled a look at virus community dynamics in relation to viral gene content and environmental variables. The study of actively replicating viruses in host cells uncovered an apparent selection on viruses from free communities for replication. This thesis furthers the understanding of the variability of the genetic content of viruses and how environmental variables influence viral abundance and shape viral community composition.

Bibliography

Acinas, S. G., Sarma-rupavtarm, R., Klepac-ceraj, V. and Polz, M. F. (2005) 'PCR-Induced Sequence Artifacts and Bias : Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample', *Applied and Environmental Microbiology*, 71(12), pp. 8966–8969. doi: 10.1128/AEM.71.12.8966.

Adriaenssens, E. M. and Cowan, D. A. (2014) 'Using Signature Genes as Tools To Assess Environmental Viral Ecology and Diversity', *Applied and Environmental Microbiology*, 80(15), pp. 4470–4480. doi: 10.1128/AEM.00878-14.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A. and Rohwer, F. (2006) 'The Marine Viromes of Four Oceanic Regions', *PLoS Biology*, 4(11). doi: 10.1371/journal.pbio.0040368.

Armstrong, F. A. J., Stearns, C. R. and Strickland, J. D. H. (1967) 'The measurement of upwelling and subsequent biological processes by means of the Technicon AutoAnalyzer^{TM} and associated equipment.', *Deep-Sea Res*, 14(3).

Arslan, D., Legendre, M., Seltzer, V., Abergel, C. and Claverie, J.-M. (2011) 'Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(42), pp. 17486–91. doi: 10.1073/pnas.1110889108.

Battye, F. (2015) 'Weasel'.

Baudoux, A. C. and Brussaard, C. P. D. (2005) 'Characterization of different viruses infecting the marine harmful algal bloom species Phaeocystis globosa', *Virology*, 341, pp. 80–90. doi: 10.1016/j.virol.2005.07.002.

Baudoux, A. C. and Brussaard, C. P. D. (2008) 'Influence of irradiance on virus-algal host interactions The role of viruses in marine phytoplankton mortality', *Journal of Phycology*, 44. doi: 10.1111/j.1529-8817.2008.00543.x.

Beaujean, A. A. (2012) 'BaylorEdPsych'.

Bellec, L., Grimsley, N., Derelle, E., Moreau, H. and Desdevises, Y. (2010) 'Abundance , spatial distribution and genetic diversity of *Ostreococcus tauri* viruses in two different environments', *Environmental microbiology reports*, (July 2016). doi: 10.1111/j.1758-2229.2010.00138.x.

Berger, S. A. and Stamatakis, A. (2011) ‘Aligning short Reads to Reference Alignments and Trees — Aligning short reads to reference alignments and trees — Supplementary Data’, *Bioninformatics*, pp. 1–8.

Bettarel, Y., Bouvier, T., Bouvier, C., Carr, C., Desnues, A., Domaizon, I., Jacquet, S., Agnes, R. and Sime-Ngando, T. (2011) ‘Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient’, *Fems Micriobiology Ecology*, 76, pp. 360–372. doi: 10.1111/j.1574-6941.2011.01054.x.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: A flexible trimmer for Illumina sequence data’, *Bioinformatics*, 30, p. 2114.

Bragg, J. G. and Chisholm, S. W. (2008) ‘Modeling the Fitness Consequences of a Cyanophage- Encoded Photosynthesis Gene’, *October*, 3(10), pp. 1–9. doi: 10.1371/journal.pone.0003550.

Bratbak, G., Egge, J. K. and Heldal, M. (1993) ‘Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms’, *Marine Ecology Progress Series*, 93.

Bratbak, G., Heldal, M., Thingstad, T. F. and Tuomi, P. (1996) ‘Dynamics of virus abundance in coastal seawater’, *FEMS Micriobiology Ecology*, 19, pp. 263–269.

Bratbak, G., Jacobsen, A., Heldal, M., Nagasaki, K. and Thingstad, F. (1998) ‘Virus production in Phaeocystis pouchetii and its relation to host cell growth and nutrition’, *Aquatic Microbial Ecology*, 16, pp. 1–9.

Breitbart, M. and Rohwer, F. (2005) ‘Here a virus, there a virus, everywhere the same virus?’, *Trends in microbiology*, 13(6), pp. 278–84. doi: 10.1016/j.tim.2005.04.003.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. and Rohwer, F. (2002) ‘Genomic analysis of uncultured marine viral communities.’, *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), pp. 14250–5. doi: 10.1073/pnas.202488399.

Breitbart, M., Thompson, L. R., Suttle, C. A. and Sullivan, M. B. (2007) ‘Exploring the vast diversity of marine viruses’, *Oceanography*, 20(2), pp. 135–139.

Brown, C. M., Campbell, D. A. and Lawrence, J. E. (2007) ‘Resource dynamics during infection of Micromonas pusilla by virus MpV-Sp1’, *Environmental Microbiology*, 9(11), pp. 2720–2727.

Brum, J. R., Schenck, R. O. and Sullivan, M. B. (2013) 'Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses', *The ISME Journal*. Nature Publishing Group, 7(9), pp. 1738–1751. doi: 10.1038/ismej.2013.67.

Brum, J. R., Steward, G. F., Jiang, S. C. and Jellison, R. (2005) 'Spatial and temporal variability of prokaryotes , viruses , and viral infections of prokaryotes in an alkaline , hypersaline lake', *Aquatic Microbial Ecology*, 41, pp. 247–260.

Brussaard, C. P. D. (2004) 'Optimization of Procedures for Counting Viruses by Flow Cytometry', *Applied and Environmental Microbiology*, 70(3), pp. 1506–1513. doi: 10.1128/AEM.70.3.1506.

Brussaard, C. P. D. (2004) 'Viral Control of Phytoplankton Populations, a Review', *The Journal of Eukaryotic Microbiology*, 51(April), pp. 125–138.

Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. a, Middelboe, M., Nagasaki, K., Paul, J. H., Schroeder, D. C., Suttle, C. a, Vaqué, D. and Wommack, K. E. (2008) 'Global-scale processes with a nanoscale drive: the role of marine viruses.', *The ISME journal*, 2(6), pp. 575–8. doi: 10.1038/ismej.2008.31.

Butina, T. V, Belykh, O. I., Maksimenko, S. Y. and Belikov, S. I. (2010) 'Phylogenetic diversity of T4-like bacteriophages in Lake Baikal, East Siberia', *FEMS Microbiology Letters*, 309(2), pp. 122–129.

Buttigieg, P. L., Ramette, A. and Algar, C. K. (2015) 'Biogeographic patterns of bacterial microdiversity in Arctic deep-sea sediments (HAUSGARTEN , Fram Strait)', *Frontiers in Microbiology*, 5, pp. 1–12. doi: 10.3389/fmicb.2014.00660.

Caceres, M. De and Legendre, P. (2009) 'Associations between species and groups of sites : indices and statistical inference', *Ecology*, 90(12), pp. 3566–3574.

Campbell, B. J. and Kirchman, D. L. (2012) 'Bacterial diversity , community structure and potential growth rates along an estuarine salinity gradient', *Isme Journal*. Nature Publishing Group, 7(1), pp. 210–220. doi: 10.1038/ismej.2012.93.

Capella-Gutiérrez, S., Silla-Martínez, J. M. and Gabaldón, T. (2009) 'trimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics*, 25(15), pp. 1972–1973. doi: 10.1093/bioinformatics/btp348.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Jeremy, E., Ley, R. E., Lozupone, C. A., Mcdonald, D., Muegge, B. D., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J. and Walters, W. A. (2010) 'QIIME allows analysis of high-throughput community sequencing data', *Nature Methods*, 7(5), pp. 335–336. doi: 10.1038/nmeth.f.303.QIIME.

Capotondi, A., Alexander, M. A., Bond, N. A., Curchitser, E. N. and Scott, J. D. (2012) 'Enhanced upper ocean stratification with climate change in the CMIP3 models', *Journal of Geophysical Research: Oceans*, 117(4), pp. 1–23. doi: 10.1029/2011JC007409.

Carmack, E. C., Mclaughlin, F. A., Vagle, S., Melling, H. and Williams, W. J. (2010) 'Structures and Property Distributions in the Three Oceans Surrounding Canada in 2007 : A Basis for a Long-Term Ocean Climate Monitoring Strategy', *Atmosphere-Ocean*, 48(4). doi: 10.3137/OC324.2010.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M., Barrell, B. G. and Parkhill, J. (2005) 'ACT : the Artemis comparison tool', *Bioinformatics*, 21(16), pp. 3422–3423. doi: 10.1093/bioinformatics/bti553.

Chen, F. and Suttle, C. A. (1995) 'Amplification of DNA polymerase gene fragments from viruses infecting microalgae', *Applied and Environmental Microbiology*, 61(4), pp. 1274–1278.

Chen, F. and Suttle, C. A. (1996) 'Evolutionary Relationships among Large Double-Stranded DNA Viruses That Infect Microalgae and Other Organisms as Inferred from DNA Polymerase Genes', *Virology*, 219(219), pp. 170–178.

Chen, F., Suttle, C. A. and Short, S. M. (1996) 'Genetic Diversity in Marine Algal Virus Communities as Revealed by Sequence Analysis of DNA Polymerase Genes', *Applied and Environmental Microbiology*, 62(8), pp. 2869–2874.

Chénard, C. and Suttle, C. a (2008) 'Phylogenetic diversity of sequences of cyanophage photosynthetic gene psbA in marine and freshwaters.', *Applied and environmental microbiology*, 74(17), pp. 5317–24. doi: 10.1128/AEM.02480-07.

Chow, C. E. T. and Fuhrman, J. A. (2012) 'Seasonality and monthly dynamics of marine myovirus communities', *Environmental Microbiology*, 14(8), pp. 2171–2183. doi: 10.1111/j.1462-2920.2012.02744.x.

Chow, C. T., Kim, D. Y., Sachdeva, R. and Caron, D. A. (2013) 'Top-down controls on bacterial community structure : microbial network analysis of bacteria , T4-like viruses and protists', *ISME Journal*. Nature Publishing Group, 8(4), pp. 816–829. doi: 10.1038/ismej.2013.199.

Chow, C. T. and Suttle, C. A. (2015) 'Biogeography of Viruses in the Sea', *Annu Rev Virol*, 2. doi: 10.1146/annurev-virology-031413-085540.

Clasen, J. L., Brigden, S. M., Payet, J. P. and Suttle, C. A. (2008) 'Evidence that viral abundance across oceans and lakes is driven by different biological factors', *Freshwater Biology*, 53, pp. 1090–1100. doi: 10.1111/j.1365-2427.2008.01992.x.

Clasen, J. L. and Suttle, C. A. (2009) 'Identification of Freshwater Phycodnaviridae and Their Potential Phytoplankton Hosts , Using DNA pol Sequence Fragments and a Genetic-Distance Analysis', *Applied and Environmental Microbiology*, 75(4), pp. 991–997. doi: 10.1128/AEM.02024-08.

Clerissi, C., Grimsley, N., Ogata, H., Hingap, P., Poulain, J. and Desdevises, Y. (2014) 'Unveiling of the Diversity of Prasinoviruses (Phycodnaviridae) in Marine Samples by Using High-Throughput Sequencing Analyses of PCR-Amplified DNA Polymerase and Major Capsid Protein Genes', *Applied and environmental microbiology*, 80. doi: 10.1128/AEM.00123-14.

Clerissi, C., Grimsley, N., Subirana, L., Maria, E., Oriol, L., Ogata, H., Moreau, H. and Desdevises, Y. (2014) 'Prasinovirus distribution in the Northwest Mediterranean Sea s affected by the environment and particularly by phosphate availability', *Virology*.

Clokie, M. R. J. and Mann, N. H. (2006) 'Marine cyanophages and light.', *Environmental microbiology*, 8(12), pp. 2074–82. doi: 10.1111/j.1462-2920.2006.01171.x.

Clokie, M. R. J., Millard, A. D. and Mann, N. H. (2010) 'T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology.', *Virology journal*, 7, p. 291. doi: 10.1186/1743-422X-7-291.

Comeau, A. M. and Krisch, H. M. (2008) 'The capsid of the T4 phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere', *Molecular Biology and Evolution*, 25(7), pp. 1321–1332. doi: 10.1093/molbev/msn080.

Connell, J. H. (1978) 'Diversity in Tropical Rain Forests and Coral Reefs', *Science*, 199(4335), pp. 1302–1310.

Cottrell, M. T. and Suttle, C. A. (1991) 'Wide-spread occurrence and clonal variation in viruses which cause lysis of a cosmopolitan , eukaryotic marine phytoplankter , Micromonas pusilla', *Marine Ecology Progress Series*, 78.

Cottrell, M. T. and Suttle, C. A. (1995a) 'Dynamics of a Lytic Virus Infecting the Photosynthetic Marine Picoflagellate Micromonas pusilla', *Limnology Oceanography*, 40(4), pp. 730–739.

Cottrell, M. T. and Suttle, C. A. (1995b) 'Genetic diversity of algal viruses which lyse the photosynthetic picoflagellate Micromonas pusilla (Prasinophyceae)', *Applied and Environmental Microbiology*, 61(8), pp. 3088–3091.

Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F. and Martiny, J. B. H. (2016) 'The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses', *Virology*. Elsevier, 499, pp. 219–229. doi: 10.1016/j.virol.2016.09.016.

Cseke, C. S. and Farkas, G. L. (1979) 'Effect of light on the attachment of cyanophage AS-1 to Anacystis nidulans', *Journal of Bacteriology*, 137, pp. 667–669.

Cubasch, U., Wuebbles, D., Chen, D., Facchini, M. C., Frame, D., Mahowald, N. and Winther, J.-G. (2013) 'Climate Change 2013 The Physical Science Basis', *IPCC*, 5.

Dammeyer, T., Bagby, S. C., Sullivan, M. B., Chisholm, S. W. and Frankenberg-Dinkel, N. (2008) 'Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria.', *Current biology : CB*, 18(6), pp. 442–8. doi: 10.1016/j.cub.2008.02.067.

Danovaro, R., Corinaldesi, C., Dell'anno, A., Fuhrman, J. a, Middelburg, J. J., Noble, R. T. and Suttle, C. a (2011) 'Marine viruses and global climate change.', *FEMS microbiology reviews*, 35(6), pp. 993–1034. doi: 10.1111/j.1574-6976.2010.00258.x.

Darling, A. C. E., Mau, B., Blattner, F. R. and Perna, N. T. (2004) 'Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements', *Genome research*, 14, pp. 1394–1403. doi: 10.1101/gr.2289704.tion.

Darriba, D., Taboada, G. L. and Posada, D. (2011) 'ProtTest 3 : fast selection of best-fit models of protein evolution', *Bioinformatics*, pp. 1–4. doi: 10.1093/bioinformatics/btr088.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W. and Karl, D. M. (2006) 'Community genomics among stratified microbial assemblages in the ocean's interior.', *Science (New York, N.Y.)*, 311(5760), pp. 496–503. doi:

10.1126/science.1120250.

Derelle, E., Ferraz, C., Escande, M.-L., Eychenie, S., Cooke, R., Piganeau, G., Desdevises, Y., Bellec, L., Moreau, H. and Grimsley, N. (2008) 'Life-Cycle and Genome of OtV5 , a Large DNA Virus of the Pelagic Marine Unicellular Green Alga Ostreococcus tauri', *PloS one*, 3(5). doi: 10.1371/journal.pone.0002250.

Derelle, E., Monier, A., Cooke, R., Worden, A. Z., Nigel, H. and Moreau, H. (2015) 'Diversity of viruses infecting the green micro-alga Ostreococcus lucimarinus', *Journal of virology*, 10. doi: 10.1128/JVI.00246-15.

Dray, S. and Dufour, A. B. (2007) 'The ade4 package: implementing the duality diagram for ecologists', *Journal of Statistical Software*, 22.

Dunigan, D. D., Fitzgerald, L. A. and Van Etten, J. L. (2006) 'Phycodnaviruses: A peek at genetic diversity', *Virus Research*, 117(1), pp. 119–132. doi: 10.1016/j.virusres.2006.01.024.

Edgar, R. C. (2010) 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 26(19), pp. 2460–2461. doi: 10.1093/bioinformatics/btq461.

Edgar, R. C. (2013) 'UPARSE: highly accurate OTU sequences from microbial amplicon reads.', *Nature Methods*, 10(10), pp. 996–8. doi: 10.1038/nmeth.2604.

El-Swais, H., Dunn, K. A., Bielawski, J. P., Li, W. K. W. and Walsh, D. A. (2015) 'Seasonal assemblages and short-lived blooms in coastal north-west Atlantic Ocean bacterioplankton', *Environmental Microbiology*, 17(10), pp. 3642–3661. doi: 10.1111/1462-2920.12629.

Estrada, M., Henrisken, P., Gasol, J. M., Casamayor, E. O. and Pedros-Alio, C. (2004) 'Diversity of planktonic photoautotrophic microorganisms along a salinity gradient as depicted by microscopy , flow cytometry , pigment analysis and DNA-based methods', *FEMS Microbiology Ecology*, 49, pp. 281–293. doi: 10.1016/j.femsec.2004.04.002.

Van Etten, J. L., Graves, M. V, Boland, W. and Delaroque, N. (2002) 'Phycodnaviridae – large DNA algal viruses', *Archives of Virology*, 147, pp. 1479–1516. doi: 10.1007/s00705-002-0822-6.

Van Etten, J. L., Lane, L. C. and Dunigan, D. D. (2010) 'DNA Viruses: The Really Big Ones (Giruses)', *Annu Rev Microbiol*, 64, pp. 83–99. doi: 10.1146/annurev.micro.112408.134338.DNA.

Falkowski, P. G. (2000) ‘Minireview: Rationalizing Elemental Ratios in Unicellular Algae’, *J. Phycol.*, 36, pp. 3–6.

Field, C. B., Behrenfeld, M. J., Randerson, J. T. and Falkowski, P. (1998) ‘Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components’, *Science*, 281(5374), pp. 237–240. doi: 10.1126/science.281.5374.237.

Filée, J. (2015) ‘Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution’, *Frontiers in Microbiology*, 6(June), pp. 1–13. doi: 10.3389/fmicb.2015.00593.

Filée, J., Tétart, F., Suttle, C. a and Krisch, H. M. (2005) ‘Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere.’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(35), pp. 12471–6. doi: 10.1073/pnas.0503404102.

Frederickson, C. M., Short, S. M. and Suttle, C. a (2003) ‘The physical environment affects cyanophage communities in British Columbia inlets.’, *Microbial ecology*, 46(3), pp. 348–57. doi: 10.1007/s00248-003-1010-2.

Fuhrman, J. A., Cram, J. A. and Needham, D. M. (2015) ‘Marine microbial community interpretation’, *Nature Publishing Group*. Nature Publishing Group, 13(3), pp. 133–146. doi: 10.1038/nrmicro3417.

Fuhrman, J. A. and Suttle, C. A. (1993) ‘Viruses in marine planktonic systems’, *Oceanography*, 6(2), pp. 51–63.

Garza, D. R. and Suttle, C. A. (1998) ‘The Effect of Cyanophages on the Mortality of Synechoccocus spp. and Selection for UV Resistant Viral Communities’, *Microbial Ecology*, 36, pp. 281–292.

Gimenes, M. V., Zanotto, P. M. de A., Suttle, C. A., Cunha, H. B. da and Mehnert, D. U. (2012) ‘Phylodynamics and movement of Phycodnaviruses among aquatic environments’, *Isme Journal*, 6, pp. 237–247. doi: 10.1038/ismej.2011.93.

Gobler, C. J., Hutchins, D. A., Fisher, L. N. S., Cosper, E. M., Wilhelmy, S. A. S.- and Dom, I. (1997) ‘Release and bioavailability of C , N , e Se , and Fe following viral lysis of a marine chrysophyte’, *Limnology Oceanography*, 42(7), pp. 1492–1504.

Goldsmith, D. B., Brum, J. R., Hopkins, M., Carlson, C. A. and Breitbart, M. (2015) ‘Water column stratification structures viral community composition in the Sargasso Sea’, *Aquatic*

Microbial Ecology, 76, pp. 85–94. doi: 10.3354/ame01768.

Goldsmith, D. B., Crosti, G., Dwivedi, B., Mcdaniel, L. D., Varsani, A., Suttle, C. A., Weinbauer, M. G., Sandaa, R. and Breitbart, M. (2011) 'Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity', *Applied and Environmental Microbiology*, 77(21), pp. 7730–7739. doi: 10.1128/AEM.05531-11.

Goldsmith, D. B., Parsons, R. J. and Beyene, D. (2015) 'Deep sequencing of the viral phoH gene reveals temporal variation , depth-specific composition , and persistent dominance of the same viral phoH genes in the Sargasso Sea', *PeerJ*. doi: 10.7717/peerj.997.

Hanson, C. A., Marston, M. F. and Martiny, J. B. H. (2016) 'Biogeographic Variation in Host Range Phenotypes and Taxonomic Composition of Marine Cyanophage Isolates', *Frontiers in Microbiology*, 7(June), pp. 1–14. doi: 10.3389/fmicb.2016.00983.

Hardies, S. C., Hwang, Y. J., Hwang, C. Y., Jang, G. I. and Cho, B. C. (2013) 'Morphology, Physiological Characteristics, and Complete Sequence of Marine Bacteriophage RIO-1 Infecting Pseudoalteromonas marina', *Journal of virology*, 87(16), pp. 9189–9198. doi: 10.1128/JVI.01521-13.

Hellweger, F. L. (2009) 'Carrying photosynthesis genes increases ecological fitness of cyanophage in silico.', *Environmental microbiology*, 11(6), pp. 1386–94. doi: 10.1111/j.1462-2920.2009.01866.x.

Herlemann, D. P. R., Labrenz, M., Ju, K., Bertilsson, S., Waniek, J. J. and Andersson, A. F. (2011) 'Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea', *Isme Journal*, pp. 1571–1579. doi: 10.1038/ismej.2011.41.

Hewson, I., Steele, J. A., Capone, D. G. and Fuhrman, J. A. (2006) 'Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters', *Marine Ecology Progress Series*, 311, pp. 67–77.

Hordoir, R. and Meier, H. E. M. (2012) 'Effect of climate change on the thermal stratification of the baltic sea: A sensitivity experiment', *Climate Dynamics*, 38(9–10), pp. 1703–1713. doi: 10.1007/s00382-011-1036-y.

Huang, S., Zhang, S., Jiao, N. and Feng, C. (2015) 'Marine Cyanophages Demonstrate Biogeographic Patterns throughout the Global Ocean', *Applied and environmental microbiology*, 81(1), pp. 441–452. doi: 10.1128/AEM.02483-14.

Hurwitz, B. L., Westveld, A. H., Brum, J. R. and Sullivan, M. B. (2014) 'Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses', *PNAS*, 111(29). doi: 10.1073/pnas.1319778111.

Iyer, L. M., Balaji, S., Koonin, E. V and Aravind, L. (2006) 'Evolutionary genomics of Nucleo-Cytoplasmic Large DNA Viruses', *Virus Research*, 117(March 2016). doi: 10.1016/j.virusres.2006.01.009.

Jacquet, S., Iglesias-rodriguez, D., Wilson, W., Jacquet, S., Heldal, M., Iglesias-rodriguez, D. and Larsen, A. (2002) 'Flow cytometric analysis of an Emiliania huxleyi bloom terminated by viral infection', *Aquatic Microbial Ecology*, 27(March 2002), pp. 111–124. doi: 10.3354/ame027111.

Jacquet, S., Partensky, F., Lennon, J. and Vaulot, D. (2001) 'Diel patterns of growth and division in marine picoplankton in culture', *Journal of Phycology*, 37(June). doi: 10.1046/j.1529-8817.2001.037003357.x.

Jameson, E., Mann, N. H., Joint, I., Sambles, C. and Mühling, M. (2011) 'The diversity of cyanomyovirus populations along a North-South Atlantic Ocean transect.', *The ISME journal*, 5(11), pp. 1713–21. doi: 10.1038/ismej.2011.54.

Jia, Y., Shan, J., Millard, A., Clokie, M. R. J. and Mann, N. H. (2010) 'Light-dependent adsorption of photosynthetic cyanophages to Synechococcus sp. WH7803.', *FEMS microbiology letters*, 310(2), pp. 120–6. doi: 10.1111/j.1574-6968.2010.02054.x.

Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W. and Weitz, J. S. (2014) 'The elemental composition of virus particles: implications for marine biogeochemical cycles', *Nature Reviews Microbiology*, 12. doi: 10.1038/nrmicro3289.

Juneau, P., Suttle, C. A., Harrison, P. J., Juneau, P., Lawrence, J. E., Suttle, C. A. and Harrison, P. J. (2003) 'Effects of viral infection on photosynthetic processes in the bloom-forming alga Heterosigma akashiwo . Aquat Microb Ecol 31 : Effects of viral infection on photosynthetic processes in the bloom-forming alga Heterosigma akashiwo', *Aquatic Microbial Ecology*, (May 2016), pp. 9–17. doi: 10.3354/ame031009.

Kao, C. C., Green, S., Stein, B. and Golden, S. S. (2005) 'Diel Infection of a Cyanobacterium by a Contractile Bacteriophage', *Society*, 71(8), pp. 4276–4279. doi: 10.1128/AEM.71.8.4276.

Keeling, R. E., Körtzinger, A. and Gruber, N. (2010) 'Ocean deoxygenation in a warming

world.', *Annual review of marine science*, 2, pp. 199–229. doi: 10.1146/annurev.marine.010908.163855.

Kellogg, C. A. and Paul, J. H. (2002) 'Degree of ultraviolet radiation damage and repair capabilities are related to G+C content in marine vibriophages', *Aquatic Microbial Ecology*, 27, pp. 13–20.

Kelly, L., Ding, H., Huang, K. H., Osburne, M. S. and Chisholm, S. W. (2013) 'Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent', *The ISME Journal*. Nature Publishing Group, 7(9), pp. 1827–1841. doi: 10.1038/ismej.2013.58.

King, A. M. Q., Adams, M. J., Carstens, E. B. and Lefkowitz, E. J. (2012) *Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses*.

Klausmeier, C. A., Litchman, E., Daufresen, T. and Levin, S. A. (2004) 'Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton', *Nature*, 429(May), pp. 171–174. doi: 1.1029/2001GL014649.

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Little, M., Luque, A., Mcnair, K., Oliveira, L. S. De, Quistad, S. D., Robinett, N. L., Sanchez, S. E., Sandin, S., Silva, G. G. Z., Smith, J., Sullivan, C., Thompson, C., Vermeij, M. J. A., Sala, E., Youle, M., Young, C., Zgliczynski, B., Brainard, R., Edwards, R. A., Nulton, J., Thompson, F. and Rohwer, F. (2016) 'Lytic to temperate switching of viral communities', *Nature*, 531. doi: 10.1038/nature17193.

Konwar, K. M., Hanson, N. W., Bhatia, M. P., Kim, D., Wu, S., Hahn, A. S., Morgan-lang, C., Cheung, H. K. and Hallam, S. J. (2015) 'Genome analysis MetaPathways v2 . 5 : quantitative functional , taxonomic and usability improvements', *Bioinformatics*, 31(June), pp. 3345–3347. doi: 10.1093/bioinformatics/btv361.

Koonin, E. V and Yutin, N. (2010) 'Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses', *Intervirology*, 53, pp. 284–292. doi: 10.1159/000312913.

Kukkaro, P. and Bamford, D. H. (2009) 'Virus – host interactions in environments with a wide range of ionic strengths', *Environmental microbiology reports*, 1, pp. 71–77. doi: 10.1111/j.1758-2229.2008.00007.x.

Labonte, J. M. and Suttle, C. A. (2013) 'Previously unknown and highly divergent ssDNA

viruses populate the oceans', *Isme Journal*, 7, pp. 2169–2177. doi: 10.1038/ismej.2013.110.

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., Sullivan, M. B., Woyke, T. and Wommack, K. E. (2015) 'Single-cell genomics-based analysis of virus – host interactions in marine surface bacterioplankton', *ISME Journal*, pp. 2386–2399. doi: 10.1038/ismej.2015.48.

Larsen, J. B., Larsen, A., Bratbak, G. and Sandaa, R. (2008) 'Phylogenetic Analysis of Members of the Phycodnaviridae Virus Family , Using Amplified Fragments of the Major Capsid Protein Gene', *Applied and Environmental Microbiology*, 74(10), pp. 3048–3057. doi: 10.1128/AEM.02548-07.

Lavigne, R., Darius, P., Summer, E. J., Seto, D., Mahadevan, P., Nilsson, A. S., Ackermann, H. W. and Kropinski, A. M. (2009) 'Classification of Myoviridae bacteriophages using protein sequence similarity.', *BMC microbiology*, 9, p. 224. doi: 10.1186/1471-2180-9-224.

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W. and Kropinski, A. M. (2008) 'Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools.', *Research in microbiology*, 159(5), pp. 406–14. doi: 10.1016/j.resmic.2008.03.005.

Lawrence, J. E. and Suttle, C. A. (2004) 'Effect of viral infection of sinking rates of Heterosigma akashiwo and its implications for bloom termination', *Aquatic Microbial Ecology*, 37(May), pp. 1–7. doi: 10.3354/ame037001.

Le, S., Josse, J. and Husson, F. (2008) 'FactoMineR: An R Package for Multivariate Analysis', *Journal of Statistical Software*, 25.

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M. and Claverie, J.-M. (2015) 'In-depth study of Mollivirus sibericum, a new 30,000-y- old giant virus infecting Acanthamoeba', *PNAS*. doi: 10.1073/pnas.1510795112.

Leliaert, F., Smith, D. R., Herron, M. D., Verbruggen, H., Delwiche, C. F. and Clerck, O. De (2012) 'Phylogeny and Molecular Evolution of the Green Algae', *Critical Reviews in Plant Sciences*, 31, pp. 1–46. doi: 10.1080/07352689.2011.615705.

Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3 : an online tool for the display and annotation of phylogenetic and other trees', *Nucleic Acids Research*, 2006,

pp. 1–4. doi: 10.1093/nar/gkw290.

Li, William, K. W. (1994) ‘Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultrphytoplankton: Measurements vrom flow cytometric sorting’, *Limnology Oceanography*, 0.

Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T., Kettler, G., Sullivan, M. B., Steen, R., Hess, W. R., Church, G. M. and Chisholm, S. W. (2007) ‘Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution.’, *Nature*, 449(7158), pp. 83–6. doi: 10.1038/nature06130.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. and Chisholm, S. W. (2005) ‘Photosynthesis genes in marine viruses yield proteins during host infection.’, *Nature*, 438(7064), pp. 86–9. doi: 10.1038/nature04111.

Lindell, D., Sullivant, M. B., Johnson, Z., Tolonen, A. C., Rohwer, F. and Chisholm, S. W. (2004) ‘Transferr of photosynthesis genes to and from Prochiorococcus viruses’, *PNAS*, 0(I).

Liu, H., Campbell, L., Landry, M. R., Nolla, H. A., Brown, S. L. and Constantinou, J. (1998) ‘Prochlorococcus and Synechococcus growth rates and contributions to production in the Arabian Sea during the 1995 Southwest and Northeast Monsoons’, *Deep-Sea Research Part II*, 45.

Liu, H., Nolla, H. A. and Campbell, L. (1997) ‘Prochlorococcus growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean’, *Aquat Microb Ecol*, 12, pp. 39–47. doi: 10.3354/ame012039.

Los, M., Wegrzyn, G. and Neubauer, P. (2003) ‘A role for bacteriophage T4 rI gene function in the control of phage development during pseudolysogeny and in slowly growing host cells’, *Research in Microbiology*, 154, pp. 547–552. doi: 10.1016/S0923-2508(03)00151-7.

Louca, S. and Doebeli, M. (2015) ‘Calibration and analysis of genome-based models for microbial ecology’, *elife*, pp. 1–17. doi: 10.7554/eLife.08208.

Lowe, T. M. and Eddy, S. R. (1997) ‘tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence’, *Nucleic Acids Research*, 25(5), pp. 955–964.

Lu, J., Chen, F. and Hodson, R. E. (2001) ‘Distribution , Isolation , Host Specificity , and Diversity of Cyanophages Infecting Marine Synechococcus spp . in River Estuaries’,

Applied and environmental microbiology, 67(7), pp. 3285–3290. doi: 10.1128/AEM.67.7.3285.

Maat, D. S., Crawfurd, K. J., Timmermans, K. R. and Brussaard, C. P. D. (2014) 'Elevated CO 2 and Phosphate Limitation Favor Micromonas pusilla through Stimulated Growth and Reduced Viral Impact', *Applied and environmental microbiology*, 80(10), pp. 3119–3127. doi: 10.1128/AEM.03639-13.

Mann, N. H. (2003) 'Phages of the marine cyanobacterial picophytoplankton', *FEMS microbiology reviews*, 27. doi: 10.1016/S0168-6445(03)00016-0.

Mann, N. H., Cook, A., Millard, A., Bailey, S. and Clokie, M. (2003) 'Bacterial photosynthesis genes in a virus', *Nature*, 424.

Maranger, R. and Bird, D. (1995) 'Viral abundance in aquatic systems:a comparison between marine and fresh waters', *Marine Ecology Progress Series*, 121, pp. 217–226. doi: 10.3354/meps121217.

Marin, B. and Melkonian, M. (2010) 'Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons', *Protist*, 161(2), pp. 304–336. doi: 10.1016/j.protis.2009.10.002.

Marston, M. F. and Amrich, C. G. (2009) 'Recombination and microdiversity in coastal marine cyanophages.', *Environmental microbiology*, 11(11), pp. 2893–903. doi: 10.1111/j.1462-2920.2009.02037.x.

Marston, M. F. and Martiny, J. B. H. (2016) 'Genomic diversification of marine cyanophages into stable ecotypes', *Environmental Microbiology*, 18(11), pp. 4240–4253. doi: 10.1111/1462-2920.13556.

Marston, M. F. and Sallee, J. L. (2003) 'Genetic Diversity and Temporal Variation in the Cyanophage Community Infecting Marine Synechococcus Species in Rhode Island ' s Coastal Waters', *Applied and environmental microbiology*, 69(8), pp. 4639–4647. doi: 10.1128/AEM.69.8.4639.

Marston, M. F., Taylor, S., Sme, N., Parsons, R. J., Noyes, T. J. E. and Martiny, J. B. H. (2013) 'Marine cyanophages exhibit local and regional biogeography', *Environmental microbiology*, 15, pp. 1452–1463. doi: 10.1111/1462-2920.12062.

Martínez, J. M., Boere, A., Gilg, I. and Lent, J. W. M. Van (2015) 'New lipid envelope-

containing dsDNA virus isolates infecting Micromonas pusilla reveal a separate phylogenetic group', *Aquatic Microbial Ecology*, 74(JANUARY). doi: 10.3354/ame01723.
Masson, D. and Pena, A. (2009) 'Chlorophyll distribution in a temperate estuary : The Strait of Georgia and Juan de Fuca Strait', *Estuarine, Coastal and Shelf Science*, 82, pp. 19–28. doi: 10.1016/j.ecss.2008.12.022.
Mayer, J. A. and Taylor, F. J. R. (1979) 'A virus which lyses the marine nanofalgellate Micromonas pusilla', *Nature*, 281, pp. 299–301.
McDaniel, L. D., delaRosa, M. and Paul, J. H. (2006) 'Temperate and lytic cyanophages from the Gulf of Mexico', *Journal of the Marine Biological Association of the UK*, 86(3), p. 517. doi: 10.1017/S0025315406013427.
Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T. and Lenski, R. E. (2012) 'Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda', *Science*, 335(January).
Millard, A., Clokie, M. R. J., Shub, D. a and Mann, N. H. (2004) 'Genetic organization of the psbAD region in phages infecting marine Synechococcus strains.', *Proceedings of the National Academy of Sciences of the United States of America*, 101(30), pp. 11007–12. doi: 10.1073/pnas.0401478101.
Millard, A. D., Zwirglmaier, K., Downey, M. J., Mann, N. H. and Scanlan, D. J. (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.', *Environmental microbiology*, 11(9), pp. 2370–87. doi: 10.1111/j.1462-2920.2009.01966.x.
Moffitt, S. E., Moffitt, R. A., Sauthoff, W., Davis, C. V., Hewett, K. and Hill, T. M. (2015) 'Paleoceanographic insights on recent oxygen minimum zone expansion: Lessons for modern oceanography', *PLoS ONE*, 10(1), pp. 1–39. doi: 10.1371/journal.pone.0115246.
Mojica, K. D. A. and Brussaard, C. P. D. (2014) 'Factors affecting virus dynamics and microbial host-virus interactions in marine environments', *FEMS Microbiol Ecol*, 89, pp. 495–515. doi: 10.1111/1574-6941.12343.
Mojica, K. D. A., Huisman, J., Wilhelm, S. W. and Brussaard, C. P. D. (2016) 'Latitudinal variation in virus-induced mortality of phytoplankton across the North Atlantic Ocean', *Isme Journal*. Nature Publishing Group, 10(2), pp. 500–513. doi: 10.1038/ismej.2015.130.

Monier, A., Pagarete, A., de Vargas, C., Allen, M. J., Read, B., Claverie, J.-M. and Ogata, H. (2009) 'Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus.', *Genome research*, 19(8), pp. 1441–9. doi: 10.1101/gr.091686.109.

Moore, C. M., Mills, M. M., Arrgio, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E. D., Geider, R. J., Guieu, C., Jaccard, S. L., Jickells, T. D., La Roche, J., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito, M. A., Thingstad, T. F., Tsuda, A. and Ulloa, O. (2013) 'Processes and patterns of oceanic nutrient limitation', *Nature Publishing Group*. Nature Publishing Group, 6(9), pp. 701–710. doi: 10.1038/ngeo1765.

Moreau, H., Piganeau, G., Desdevises, Y., Cooke, R., Derelle, E. and Grimsley, N. (2010) 'Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer.', *Journal of virology*, 84(24), pp. 12555–63. doi: 10.1128/JVI.01123-10.

Motegi, C., Kaiser, K., Benner, R. and Weinbauer, M. G. (2015) 'SHORT COMMUNICATION Effect of P-limitation on prokaryotic and viral production in surface waters of the Northwestern Mediterranean Sea', *Journal of Plankton Research*, 37(1), pp. 16–20. doi: 10.1093/plankt/fbu089.

Mühling, M., Fuller, N. J., Somerfield, P. J., Marie, D., Wilson, W. H., Scanlan, D. J., Post, A. F., Joint, I. and Mann, N. H. (2005) 'Genetic diversity of marine Synechococcus and co-occurring cyanophage communities : evidence for viral control of phytoplankton', *Environmental Microbiology*, 7, pp. 499–508. doi: 10.1111/j.1462-2920.2004.00713.x.

Mühling, M., Fuller, N. J., Somerfield, P. J., Post, A. F., Wilson, W. H., Scanlan, D. J., Joint, I. and Mann, N. H. (2006) 'High resolution genetic diversity studies of marine Synechococcus isolates using rpoC1 -based restriction fragment length polymorphism', *Aquatic Microbial Ecology*, 45, pp. 263–275.

Murphey, J. and Riley, J. P. (1962) 'A modified single solution method for the determination of phosphate in natural waters', *Anal. Chem. Acta*, 27.

Murray, A. G. and Jackson, G. A. (1992) 'Viral dynamics : a model of the effects of size , shape , motion and abundance of single-celled planktonic organisms and other particles', *Marine Ecology Progress Series*, 89, pp. 103–116.

Nagasaki K, Y. M. (1998) ‘Effect of temperature on the algicidal activity and the stability of HaV (Heterosigma akashiwo virus)’, *Aquatic Microbial Ecology*, 15, pp. 211–216. doi: 10.3354/ame015211.

Needham, D. M., Chow, C.-E. T., Cram, J. A., Sachdeva, R., Parada, A. and Fuhrman, J. A. (2013) ‘Short-term observations of marine bacterial and viral communities: patterns, connections and resilience.’, *The ISME journal*. Nature Publishing Group, 7(7), pp. 1274–85. doi: 10.1038/ismej.2013.19.

Noble, R. T. and Fuhrman, J. A. (1997) ‘Virus Decay and Its Causes in Coastal Waters’, *Applied and environmental microbiology*, 63(1), pp. 77–83.

Oksanen, J., Blanchet, G. F., Kindt, R., Legendre, P., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. and Wagner, H. (2016) *vegan: Community Ecology Package. R package version 2.3-3.* Available at: http://cran.r-project.org/package=vegan.

Orr, H. A. (2009) ‘Fitness and its role in evolutionary genetics’, *Nature reviews. Genetics*, (June). doi: 10.1038/nrg2603.

Ortmann, A. C. and Suttle, C. A. (2005) ‘High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality’, *Deep-Sea Research*, 52, pp. 1515–1527. doi: 10.1016/j.dsr.2005.04.002.

Palenik, B. and Haselkorn, R. (1992) ‘Multiple evolutionary origins of prochlorophytes, the chlorophyll b-containing prokaryotes’, *Nature*, 355.

Paradis, E., Claude, J. and Strimmer, K. (2004) ‘APE: analyses of phylogenetics and evolution in R language’, *Bioinformatics*, 20.

Partensky, F., Blanchot, J. and Vaulot, D. (1999) ‘Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters : a review’, *Bulletin de l’Institute Oceanographic Monaco*.

Partensky, F., Hess, W. R. and Vaulot, D. (1999) ‘Prochlorococcus , a Marine Photosynthetic Prokaryote of Global Significance’, *Microbiology and Molecular Biology Reviews*, 63(1).

Paul, J. H., Rose, J. B., Jiang, S. C., Kellogg, C. A. and Dickson, L. (1993) ‘Distribution of Viral Abundance in the Reef Environment of Key Largo , Florida’, *Applied and environmental microbiology*, 59(3), pp. 718–724.

Payet, J. P. and Suttle, C. A. (2008) 'Physical and biological correlates of virus dynamics in the southern Beaufort Sea and Amundsen Gulf', *journal of marine systems*, 74. doi: 10.1016/j.jmarsys.2007.11.002.

Payet, J. P. and Suttle, C. A. (2013) 'To kill or not to kill : The balance between lytic and lysogenic viral infection is driven by trophic status', *Limnology Oceanography*, 58(2), pp. 465–474. doi: 10.4319/lo.2013.58.2.0465.

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J. and Abergel, C. (2015) 'Pandoraviruses : Amoeba Viruses with Genomes Up to 2 . 5 Mb Reaching That of Parasitic Eukaryotes', *Science*, 281(March). doi: 10.1126/science.1239181.

Poorvin, L., Rinta-kanto, J. M., Hutchins, D. A. and Wilhelm, S. W. (2004) 'Viral release of iron and its bioavailability to marine plankton', *Limnology Oceanography*, 49(5), pp. 1734–1741.

Poorvin, L., Sander, S. G., Velasquez, I., Ibisanmi, E., Lecleir, G. R. and Wilhelm, S. W. (2011) 'A comparison of Fe bioavailability and binding of a catecholate siderophore with virus-mediated lysates from the marine bacterium Vibrio alginolyticus PWH3a', *journal of experimental marine biology ecology*, 399, pp. 43–47. doi: 10.1016/j.jembe.2011.01.016.

Proctor, L. M. and Fuhrman, J. A. (1990) 'Viral mortality of marine bacteria and cyanobacteria', *Nature*, 343.

Puxty, R. J., Millard, A. D., Evans, D. J. and Scanlan, D. J. (2014) 'Shedding new light on viral photosynthesis', *Photsynth Research*. doi: 10.1007/s11120-014-0057-x.

R, C. T. (2015) *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* Available at: http://www.r-project.org.

Rambaut, A. (2014) 'Tree Figure Drawing Tool Version 1.4.2'. Institute of Evolutionary Biology, University of Edinburgh.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogate, H., La Scola, B., Suzan, M. and Claverie, J.-M. (2004) 'The 1 . 2-Megabase Genome Sequence of Mimivirus', *Science*, 306(November), pp. 1344–1351.

Redfield, A. C., Ketchum, B. H. and Richards, F. A. (1963) 'The composition of seawater: Comparative and descriptive oceanography.', in N, H. M. (ed.) *The sea: ideas and observations on progress in the study of the seas*. Interscience, pp. 26–77.

Rho, M., Tang, H. and Ye, Y. (2010) 'FragGeneScan: Predicting genes in short and error-prone reads', *Nucleic Acids Research*, 38(20), pp. 1–12. doi: 10.1093/nar/gkq747.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., Felts, B., Haynes, M., Liu, H., Lipson, D., Mahaffy, J., Martin-Cuadrado, A. B., Mira, A., Nulton, J., Pašić, L., Rayhawk, S., Rodriguez-Mueller, J., Rodriguez-Valera, F., Salamon, P., Srinagesh, S., Thingstad, T. F., Tran, T., Thurber, R. V., Willner, D., Youle, M. and Rohwer, F. (2010) 'Viral and microbial community dynamics in four aquatic environments.', *The ISME journal*, 4(6), pp. 739–751. doi: 10.1038/ismej.2010.1.

Rohwer, F. and Thurber, R. V. (2009) 'Viruses manipulate the marine environment.', *Nature*, 459(7244), pp. 207–12. doi: 10.1038/nature08060.

Rowe, J. M., Gobena, D., Wilson, W. H. and Wilhelm, S. W. (2011) 'Application of the major capsid protein as a marker of the phylogenetic diversity of Emiliania huxleyi viruses', *Fems Micriobiology Ecology*, 76. doi: 10.1111/j.1574-6941.2011.01055.x.

Sandaa, R.-A. and Larsen, A. (2006) 'Seasonal variations in virus-host populations in Norwegian coastal waters: focusing on the cyanophage community infecting marine Synechococcus spp.', *Applied and environmental microbiology*, 72(7), pp. 4610–8. doi: 10.1128/AEM.00168-06.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C. and Barbe, V. (2013) 'Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes', *PNAS*. doi: 10.1073/pnas.1303251110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1303251110.

Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. and Manabe, S. (1998) 'Simulated response of the ocean carbon cycle to anthropocgenic climate warming', *Nature*, 393(May), pp. 1–2.

Sarmiento, J. L., Slater, R., Barber, R., Bopp, L., Doney, S. C., Hirst, A. C., Kleypas, J., Matear, R., Mikolajewicz, U., Monfray, P., Soldatov, V., Spall, S. A. and Stouffer, R. (2004) 'Response of ocean ecosystems to climate warming', *Global Biogeochemical Cycles*, 18(August 2003). doi: 10.1029/2003GB002134.

Scanlan, D. J. and West, N. J. (2002) 'Molecular ecology of the marine cyanobacterial genera Prochlorococcus and Synechococcus.', *FEMS microbiology ecology*, 40(1), pp.

1–12. doi: 10.1111/j.1574-6941.2002.tb00930.x.

Seitzinger, S. P., Mayorga, E., Bouwman, A. F., Kroeze, C., Beusen, A. H. W., Billen, G., Drecht, G. Van, Dumont, E., Fekete, B. M., Garnier, J. and Harrison, J. A. (2010) 'Global river nutrient export : A scenario analysis of past and future trends', *Global Biogeochemical Cycles*, 24. doi: 10.1029/2009GB003587.

Shapiro, B. J. and Polz, M. F. (2014) 'Ordering microbial diversity into ecologically and genetically cohesive units', *NIH Public Access*, 22(5), pp. 235–247. doi: 10.1016/j.tim.2014.02.006.Ordering.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R. Y., Partensky, F., Koonin, E. V, Wolf, Y. I., Nelson, N. and Béjà, O. (2009) 'Photosystem I gene cassettes are present in marine virus genomes.', *Nature*, 461(7261), pp. 258–62. doi: 10.1038/nature08284.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D. B., Yooseph, S., Zeidner, G., Golden, S. S., Mackey, S. R., Adir, N., Weingart, U., Horn, D., Venter, J. C., Mandel-Gutfreund, Y. and Béjà, O. (2007) 'Viral photosynthetic reaction center genes and transcripts in the marine environment.', *The ISME journal*, 1(6), pp. 492–501. doi: 10.1038/ismej.2007.67.

Sharp, C. E., Brady, A. L., Sharp, G. H., Grasby, S. E., Stott, M. B. and Dunfield, P. F. (2014) 'Humboldt ' s spa : microbial diversity is controlled by temperature in geothermal environments', *The ISME Journal*. Nature Publishing Group, 8(6), pp. 1166–1174. doi: 10.1038/ismej.2013.237.

Shelford, E., Middelboe, M., Møller, E. F. and Suttle, C. A. (2012) 'Virus-driven nitrogen cycling enhances phytoplankton growth', *Aquatic Microbial Ecology*, 66, pp. 41–46. doi: 10.3354/ame01553.

Short, C. M., Rusanova, O. and Short, S. M. (2011) 'Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada.', *The ISME journal*. Nature Publishing Group, 5(5), pp. 810–21. doi: 10.1038/ismej.2010.183.

Short, C. M. and Suttle, C. A. (2005) 'Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments', *Applied and environmental microbiology*. Am Soc Microbiol, 71(1), pp. 480–486. doi: 10.1128/AEM.71.1.480.

Short, S. M. and Short, C. M. (2008) 'Diversity of algal viruses in various North American freshwater environments', *Aquatic Microbial Ecology*, 51, pp. 13–21. doi: 10.3354/ame01183.

Short, S. M. and Suttle, C. A. (2002) 'Sequence Analysis of Marine Virus Communities Reveals that Groups of Related Algal Viruses Are Widely Distributed in Nature', *Applied and Environmental Microbiology*, 68(3), pp. 1290–1296. doi: 10.1128/AEM.68.3.1290.

Short, S. M. and Suttle, C. A. (2003) 'Temporal dynamics of natural communities of marine algal viruses and eukaryotes', *Aquatic Microbial Ecology*, 32, pp. 107–119. doi: 10.3354/ame032107.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. and Higgins, D. G. (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.', *Molecular systems biology*, 7(1), p. 539. doi: 10.1038/msb.2011.75.

Siotto, F., Martin, C., Rauh, O., Etten, J. L. Van, Schroeder, I., Moroni, A. and Thiel, G. (2014) 'Viruses infecting marine picoplancton encode functional potassium ion channels', *Virology*. Elsevier, 466–467, pp. 103–111. doi: 10.1016/j.virol.2014.05.002.

Snel, B., Bork, P. and Huynen, M. A. (1999) 'Genome phylogeny based on gene content', *Nature*, 21(january), pp. 108–110.

Stamatakis, A. (2014) 'Stamatakis - 2014 - RAxML version 8 a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, pp. 2010–2011.

Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M. and Poisson, G. (2013) 'Are we missing half of the viruses in the ocean ?', *The ISME journal*, 7, pp. 672–679. doi: 10.1038/ismej.2012.121.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. and Chisholm, S. W. (2005) 'Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations.', *PLoS biology*, 3(5), p. e144. doi: 10.1371/journal.pbio.0030144.

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A. M., Kelly, L., Weigele, P. R., Defrancesco, A. S., Kern, S. E., Thompson, L. R., Young, S., Yandava, C., Fu, R., Krastins, B., Chase, M., Sarracino, D., Osburne, M. S., Henn, M. R. and Chisholm, S. W. (2010) 'Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments.', *Environmental microbiology*, 12, pp.

3035–3056. doi: 10.1111/j.1462-2920.2010.02280.x.

Sullivan, M. B., Lindell, D., Lee, J. a, Thompson, L. R., Bielawski, J. P. and Chisholm, S. W. (2006) 'Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts.', *PLoS biology*, 4(8), p. e234. doi: 10.1371/journal.pbio.0040234.

Sullivan, M. B., Waterbury, J. B. and Chisholm, S. W. (2003) 'Cyanophages infecting the oceanic cyanobacterium Prochlorococcus.', *Nature*, 424(6952), pp. 1047–51. doi: 10.1038/nature01929.

Suttle, C. a (2005) 'Viruses in the sea.', *Nature*, 437(7057), pp. 356–61. doi: 10.1038/nature04160.

Suttle, C. a (2007) 'Marine viruses--major players in the global ecosystem.', *Nature reviews. Microbiology*, 5(10), pp. 801–12. doi: 10.1038/nrmicro1750.

Suttle, C. A. (1994) 'The significance of viruses to mortality in aquatic microbial communities', *Microbial Ecology*, 28(2), pp. 237–243.

Suttle, C. A. (2000) 'Cyanopahges and Their Role in the Ecology of Cyanobacteria', in Whitton, B. A. and Potts, M. (eds) *The Ecology of Cyanobacteria*. Kluwer Academic Publishers, pp. 563–589.

Suttle, C. A. and Chan, A. M. (1993) 'Marine cyanophages infecting oceanic and coastal strains of Synechococcus: abundance, morphology, cross-infectivity and growth characteristics', *Marine Ecology Progress Series*, 92(1/2), pp. 99–109. doi: 10.3354/meps092099.

Suttle, C. A. and Chen, F. (1992) 'Mechanisms and rates of decay of marine viruses in seawater.', *Applied and environmental microbiology*, 58(11), pp. 3721–9.

Tetart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H.-W. and Krish, H. M. (2001) 'Phylogeny of the Major Head and Tail Genes of the Wide-Ranging T4-Type Bacteriophages', *Journal of Bacteriology*, 183(1), pp. 358–366. doi: 10.1128/JB.183.1.358.

Tetu, S. G., Brahamsha, B., Johnson, D. a, Tai, V., Phillippy, K., Palenik, B. and Paulsen, I. T. (2009) 'Microarray analysis of phosphate regulation in the marine cyanobacterium Synechococcus sp. WH8102.', *The ISME journal*, 3(7), pp. 835–49. doi: 10.1038/ismej.2009.31.

Thingstad, T. F. (2000) ‘Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems’, *Limnology and Oceanography*, 45(6), pp. 1320–1328.

Thingstad, T. F. and Lignell, R. (1997) ‘Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand’, *Aquatic Microbial Ecology*, 13, pp. 19–27.

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J. and Chisholm, S. W. (2011) ‘Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism.’, *Proceedings of the National Academy of Sciences of the United States of America*, pp. 1–8. doi: 10.1073/pnas.1102164108.

Tommasi, D., Hunt, B. P. V, Pakhomov, E. A. and Mackas, D. L. (2013) ‘Mesozooplankton community seasonal succession and its drivers : Insights from a British Columbia , Canada , fjord’, *journal of marine systems*, 116, pp. 10–32.

Tyrrell, T. (1999) ‘The relative influences of nitrogen and phosphorus on oceanic primary production’, *Nature*, 400, pp. 525–531.

Vaulot, D. (1989) ‘CYTOPC : Processing Software for Flow Cytometric Data’, *Signal and Noise*, 2, p. 292.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S-Plus*. Fourth. Springer New York.

Wang, I., Li, Y., Que, Q., Bhattacharya, M., Lane, L. C., Chaney, W. G. and Etten, J. L. V. A. N. (1993) ‘Evidence for virus-encoded glycosylation specificity’, *Proceedings of the National Academy of Sciences*, 90(May), pp. 3840–3844.

Wang, J., Yang, D., Zhang, Y., Shen, J., Gast, C. Van Der, Hahn, M. W. and Wu, Q. (2011) ‘Do Patterns of Bacterial Diversity along Salinity Gradients Differ from Those Observed for Macroorganisms ?’, *PloS one*, 6(11). doi: 10.1371/journal.pone.0027597.

Wang, K. and Chen, F. (2004) ‘Genetic diversity and population dynamics of cyanophage communities in the Chesapeake Bay’, *Aquatic Microbial Ecology*, 34(2), pp. 105–116.

Weigele, P. R., Pope, W. H., Pedulla, M. L., Houtz, J. M., Smith, A. L., Conway, J. F., King, J., Hatfull, G. F., Lawrence, J. G. and Hendrix, R. W. (2007) ‘Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus.’, *Environmental microbiology*, 9(7), pp. 1675–95. doi: 10.1111/j.1462-2920.2007.01285.x.

Weinbauer, M. G., Bonilla-Findji, O., Chan, A. M., Dolan, J. R., Short, S. M., Simek, K., Wilhelm, S. W. and Suttle, C. A. (2011) 'Synechococcus growth in the ocean may depend on the lysis of heterotrophic bacteria', *Journal of Plankton Research*, 33(10), p. fbr041-. doi: 10.1093/plankt/fbr041.

Weinbauer, M. G., Brettar, I. and Ho, M. G. (2003) 'Lysogeny and virus-induced mortality of bacterioplankton in surface , deep , and anoxic marine waters', *Limnology Oceanography*, 48(4), pp. 1457–1465.

Weinbauer, M. G. and Rassoulzadegan, F. (2003) 'Are viruses driving microbial diversification and diversity?', *Environmental Microbiology*, 6(1), pp. 1–11. doi: 10.1046/j.1462-2920.2003.00539.x.

Weinbauer, M. G., Wilhelm, S. W., Suttle, C. A., Pledger, R. J. and Mitchell, D. L. (1999) 'Sunlight-induced DNA damage and resistance in natural viral communities', *Aquatic Microbial Ecology*, 17(2), pp. 111–120. doi: 10.3354/ame017111.

Weithoff, G., Walz, N. and Gaedke, U. (2001) 'The intermediate disturbance hypothesis - species diversity or functional diversity ?', *Journal of plankton*, 23(10).

Weitz, J. S. and Wilhelm, S. W. (2012) 'Ocean viruses and their effects on microbial communities and biogeochemical cycles', *Faculty of 1000*, 8(September), pp. 2–9. doi: 10.3410/B4-17.

Weynberg, K. D., Allen, M. J., Ashelford, K., Scanlan, D. J. and Wilson, W. H. (2009) 'From small hosts come big viruses : the complete genome of a second Ostreococcus tauri virus , OtV-1', *Environmental microbiology*, 11, pp. 2821–2839. doi: 10.1111/j.1462-2920.2009.01991.x.

Wigington, C. H., Sonderegger, D., Brussaard, C. P. D., Buchan, A., Finke, J. F., Fuhrman, J. A., Lennon, J. T., Middelboe, M., Suttle, C. A., Stock, C., Wilson, W. H., Wommack, K. E., Wilhelm, S. W. and Weitz, J. S. (2016) 'marine virus and microbial cell abundances', *Nature Microbiology*, (January), pp. 4–11. doi: 10.1038/nmicrobiol.2015.24.

Wilhelm, S. W., Jeffrey, W. H., Dean, A. L., Meador, J., Pakulski, J. D. and Mitchell, D. L. (2003) 'UV radiation induced DNA damage in marine viruses along a latitudinal gradient in the southeastern Pacific Ocean', *Aquatic Microbial Ecology*, 31, pp. 1–8.

Wilhelm, S. W. and Suttle, C. A. (1999) 'Viruses and Nutrient Cycles in the Sea', *Bioscience*, 49(October).

Wilhelm, S. W., Weinbauer, M. G., Suttle, C. A. and Jeffrey, W. H. (1998) 'The role of sunlight in the removal and repair of viruses in the sea', *Limnology and Oceanography*, 43(4), pp. 586–592. doi: 10.4319/lo.1998.43.4.0586.

Wilhelm, S. W., Weinbauer, M. G., Suttle, C. A., Ralph, R. J. and Mitchell, D. L. (1998) 'Measurements of DNA damage and photoreactivation imply that most viruses in marine surface waters are infective', *Aquatic Microbial Ecology*, 14, pp. 215–222.

Williamson, S. J. and Paul, J. H. (2006) 'Environmental Factors that influence the Transition from Lysogenic to Lytic Existence in the HSIC/Listonella pelagia Marine Phage–Host System', *Microbial ecology*, 52, pp. 217–225. doi: 10.1007/s00248-006-9113-1.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M. and Venter, J. C. (2008) 'The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples.', *PloS one*, 3(1), p. e1456. doi: 10.1371/journal.pone.0001456.

Wilson, W. H., Carr, N. G. and Mann, N. H. (1996) 'The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium Synechococcus SP. WH7803', *Journal of Phycology*, 32, pp. 506–516.

Wilson, W. H., Fuller, N. J., Joint, I. R. and Mann, N. H. (1999) 'Analysis of Cyanophage Diversity in the Marine Environment Using Denaturing Gradient Gel Electrophoresis', *Microbial biosystems: new frontiers*.

Wommack, K. E. and Colwell, R. R. (2000) 'Virioplankton: Viruses in aquatic ecosystems', *Microbiology and Molecular Biology Reviews*, 64(1), pp. 69–114.

Worden, A. Z., Nolan, J. K. and Palenik, B. (2004) 'Assessing the dynamics and ecology of marine picophytoplankton : The importance of the eukaryotic component', *Limnology and Oceanography*, 49(1), pp. 168–179.

Yutin, N., Wolf, Y. I. and Koonin, E. V (2014) 'Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life', *Virology*. Elsevier, 466–467, pp. 38–52. doi: 10.1016/j.virol.2014.06.032.

Zaikova, E., Walsh, D. a, Stilwell, C. P., Mohn, W. W., Tortell, P. D. and Hallam, S. J. (2010) 'Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British

Columbia.', *Environmental microbiology*, 12(1), pp. 172–91. doi: 10.1111/j.1462-2920.2009.02058.x.

Zeng, Q. and Chisholm, S. W. (2012) 'Report Marine Viruses Exploit Their Host's Two-Component Regulatory System in Response to Resource Limitation', *Current Biology*. Elsevier Ltd, 22(2), pp. 124–128. doi: 10.1016/j.cub.2011.11.055.

Zhang, W., Zhou, J., Liu, T., Yu, Y., Pan, Y. and Yan, S. (2015) 'Four novel algal virus genomes discovered from Yellowstone Lake metagenomes', *Nature Publishing Group*. Nature Publishing Group, pp. 1–13. doi: 10.1038/srep15131.

Zhong, X. and Jacquet, S. (2014) 'Contrasting diversity of phycodnavirus signature genes in two large and deep western European lakes', *Environmental microbiology*, 16, pp. 759–773. doi: 10.1111/1462-2920.12201.

Appendix A supplementary tables

Table A. 1: Sampling details for DNApol environmental samples.
Samples from Jericho Pier, Point Atkinson, Juan de Fuca Strait and Saanich Inlet. Sampling date, location and depth given with *in situ* temperature (Temp.), salinity (Sal.). Field sampling volume (Vol.), type of pre-filter used in the lab, volume of viral concentrate (VC) and volume used for DNA extraction (Ext.) used are stated.

Site	Year	Month	Lat, Long	Depth (m)	Temp. (C)	Sal. (PSU)	Volume (l)	Pre-filter	VC (ml)	Extr. (ml)
Jericho Pier (JP)	2006	June	49.28N, 123.20W		18	12	41		900	650
Point Atkinson (PA)	2006	June	49.32N, 123.25W	1	14	23	45	GC50- HVLP	795	650
Juan de Fuca Strait (JF)	2006	July	48.45N, 123.32W		8.3	33	72		620	420
Saanich Inlet (SI)	2007	April	48.58N, 123.5W		na	na	unknown	GF/D- Sterivex	245	1.7
	2008	February			7.5	31	18		287	7.8
	2008	March			7.5	31	18		168	5.9
	2008	April		10	7.6	31	16		230	2.9
	2008	June			9.8	33	14		185	1.1
	2008	August			12	34	18		190	3.1
	2008	December			9	32	15		210	11
	2007	April			na	na	17		220	10
	2008	February			9.4	34	17		300	8.9
	2008	March			9.4	34	16		220	23
	2008	April		200	9.4	34	19		250	8.3
	2008	June			9.4	34	18		235	15
	2008	August			9.4	34	19		160	78
	2008	December			9.2	34	16		186	41

Table A. 2: Sampling details for gp43 environmental samples.
Sampling locations and years are given, environmental parameters were measured *in situ* by CTD or analyzed in the lab.

Project	Site	Year	Month	Depth (m)	Lat (°N)	Long (°W)	Temp (°C)	Sal (PSU)	Chl (mg m^-3)
SOG	Hotham	2010	july	16	49.868	-124.042	13.97	26.03	4.38
SOG	Narrows	2010	july	14	49.731	-123.739	14.29	23.22	9.09
SOG	Pendrell	2010	july	13	50.296	-124.727	13.40	26.93	15.40
SOG	Narrows	2011	september	10	49.732	-123.738	12.41	23.01	1.35
SOG	Sechelt	2011	september	11	49.675	-123.85	12.24	24.46	3.99
SOG	Pendrell	2011	september	10	50.273	-124.713	14.36	24.42	1.05
SOG	Teakern	2011	september	10	50.184	-124.863	14.04	24.60	0.85
SOG	Johnson1	2011	september	18	50.501	-126.35	9.28	30.47	0.35
SOG	QC Sound	2011	september	10	51.04	-127.87	11.06	31.58	0.91
SOG	Simon Sound	2011	september	8	50.85	-126.54	11.20	25.40	5.81
SOG	Port Ellizabeth	2011	september	15	50.62	-126.48	9.13	23.13	0.78
SOG	Carrington Bay	2011	september	10	50.14	-125	11.36	25.04	1.63
SOG	Narrows2	2012	september	10	49.75	-123.73	12.31	24.54	12.90
SOG	Narrows3	2012	september	15	49.71	-123.75	12.78	24.61	4.32
SOG	Sechelt	2012	september	12	49.67	-123.83	12.90	25.12	4.92
SOG	Disco Passage	2012	september	11	50.27	-125.38	9.93	29.43	1.01
SOG	Campbell	2012	september	11	50.03	-125.22	9.57	29.43	1.13
SOG	Lasqueti Island	2012	september	10	49.49	-124.37	11.71	27.41	10.62
SAA	Saanich	2011	may	5	48.5	-123.5	9.74	29.20	15.54
SAA	Saanich	2011	june	5	48.5	-123.5	12.03	27.95	0.89
SAA	Saanich	2011	july	5	48.5	-123.5	12.57	27.69	5.48
SAA	Saanich	2011	september	5	48.5	-123.5	14.68	27.88	4.09
SAA	Saanich	2011	november	5	48.5	-123.5	9.71	28.83	2.96
SAA	Saanich	2011	december	5	48.5	-123.5	9.06	28.74	3.39
SAA	Saanich	2012	january	5	48.5	-123.5	7.80	28.50	1.58
SAA	Saanich	2012	february	5	48.5	-123.5	7.54	29.29	2.56
SAA	Saanich	2012	march	5	48.5	-123.5	7.28	29.46	2.03
SAA	Saanich	2012	may	5	48.5	-123.5	9.55	29.02	0.85
SAA	Saanich	2012	june	5	48.5	-123.5	11.23	29.03	4.83
SAA	Saanich	2012	august	5	48.5	-123.5	13.71	27.64	5.73
SAA	Saanich	2011	may	10	48.5	-123.5	8.99	29.42	13.74
SAA	Saanich	2011	june	10	48.5	-123.5	10.38	28.75	0.80
SAA	Saanich	2011	july	10	48.5	-123.5	11.85	27.88	1.57
SAA	Saanich	2011	september	10	48.5	-123.5	13.14	28.01	1.42
SAA	Saanich	2011	november	10	48.5	-123.5	10.55	29.27	2.18
SAA	Saanich	2011	december	10	48.5	-123.5	8.73	29.24	1.74
SAA	Saanich	2012	january	10	48.5	-123.5	8.12	29.55	1.11
SAA	Saanich	2012	february	10	48.5	-123.5	7.69	29.53	1.91
SAA	Saanich	2012	march	10	48.5	-123.5	7.28	29.46	1.98
SAA	Saanich	2012	may	10	48.5	-123.5	8.62	29.37	25.18
SAA	Saanich	2012	june	10	48.5	-123.5	10.02	29.12	2.62
SAA	Saanich	2012	august	10	48.5	-123.5	12.99	28.06	4.72

Table A.2 Continued

Oxy (ml l^-1)	NO3 (µM)	PO4 (µM)	Si (µM)	PAR (µmol quanta m^-2 s^-1)	Bac (# ml^-1)	VLP (# ml^-1)	VC#	Vol. (L)
5.19	8.95	0.98	30.98	383.11	2.15E+06	3.02E+07	1121	20
6.02	7.37	0.85	28.64	2429.33	4.60E+06	5.62E+07	1122	20
6.91	1.47	0.37	24.95	149.48	2.85E+06	2.60E+07	1126	20
4.55	11.83	1.26	32.63	514.74	8.88E+05	2.15E+07	1294	20
4.42	14.38	1.44	35.90	204.86	1.09E+06	2.51E+07	1295	20
4.85	8.42	0.95	37.49	3.02	1.72E+06	6.86E+07	1296	20
4.68	9.23	1.03	39.66	88.10	2.13E+06	6.06E+07	1298	20
3.41	27.02	2.23	51.69	3.44	5.75E+05	6.82E+06	1301	20
4.56	18.11	1.66	35.57	10.59	1.05E+06	1.06E+07	1306	20
4.88	6.34	0.77	24.23	6.49	9.87E+05	1.86E+07	1309	20
4.33	19.76	1.59	42.38	179.33	5.35E+05	5.41E+06	1311	20
4.37	18.36	1.71	48.34	3.05	1.20E+06	1.67E+07	1312	20
4.29	10.14	1.33	25.98	66.20	1.46E+06	2.34E+07	1401	20
4.51	11.08	1.24	33.26	963.99	1.27E+06	3.18E+07	1402	20
4.59	15.64	1.59	39.15	2264.07	1.30E+06	2.86E+07	1403	20
3.71	24.68	2.23	52.62	782.92	6.17E+05	7.63E+06	1409	20
3.49	26.32	2.37	56.30	1491.46	9.53E+05	7.44E+06	1410	20
4.49	14.01	1.41	40.34	144.77	1.58E+06	2.19E+07	1412	20
7.87				291.25	4.97E+05	2.51E+06	1242	200
5.89				2327.80	4.66E+05	1.54E+07	1250	200
5.47				694.20	2.01E+06	4.60E+07	1264	200
5.10				182.26	1.55E+06	3.61E+07	1289	200
3.57				612.91	7.87E+05	1.20E+07	1321	200
3.76				612.91	6.02E+05	6.82E+06	1328	200
4.50				234.59	6.04E+05	4.86E+06	1335	200
4.55				30.84	5.48E+05	4.52E+06	1347	200
4.64				67.73	7.05E+05	4.56E+06	1354	200
8.13				1050.70	1.15E+06	2.06E+07	1362	200
5.33				280.01	3.50E+06	3.01E+07	1369	200
5.65				969.75	2.30E+06	4.19E+07	1383	200
6.92	4.50	0.50	11.40	47.71	4.97E+05	2.51E+06	1242	200
5.87	8.00	1.20	17.30	669.14	4.66E+05	1.54E+07	1250	200
5.57	5.60	0.80	29.30	137.39	2.01E+06	4.60E+07	1264	200
3.82	6.10	1.00	30.70	55.16	1.55E+06	3.61E+07	1289	200
3.33	25.70	2.50	46.80	162.66	7.87E+05	1.20E+07	1321	200
3.78	27.80	2.30	46.50	162.66	6.02E+05	6.82E+06	1328	200
4.16	27.40	2.40	47.50	75.85	6.04E+05	4.86E+06	1335	200
4.33	27.10	2.40	53.90	11.83	5.48E+05	4.52E+06	1347	200
4.63	27.30	2.30	52.30	28.28	7.05E+05	4.56E+06	1354	200
7.11	0.20	0.30	1.40	366.85	1.15E+06	2.06E+07	1362	200
4.67	8.10	1.00	23.60	67.26	3.50E+06	3.01E+07	1369	200
5.07	1.90	0.50	20.70	167.18	2.30E+06	4.19E+07	1383	200

Appendix B supplementary figures

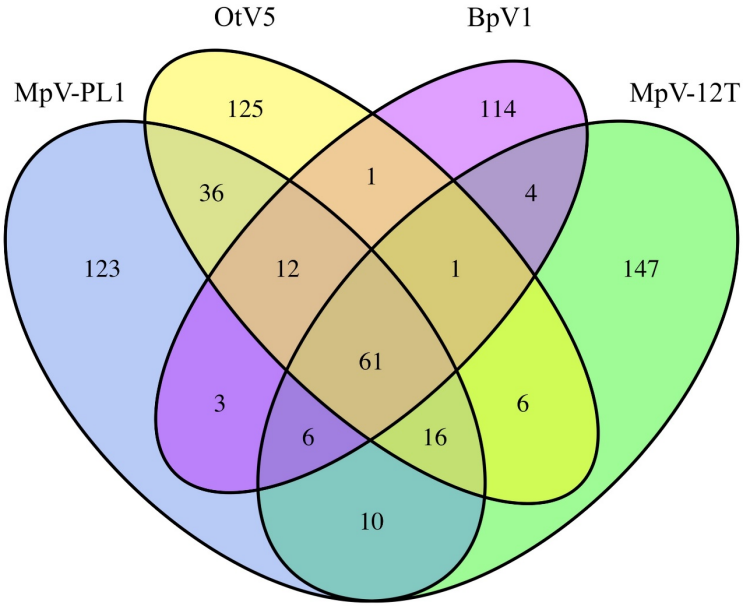

Figure A. 1: Shared genes of four prasinoviruses.
Venn diagram of the CDS of the four viruses infecting the three genera of *Ostreo*coccus *Bathyococcus*- and *Micromonas* in comparison, based on clusters at 50 % aa identity.

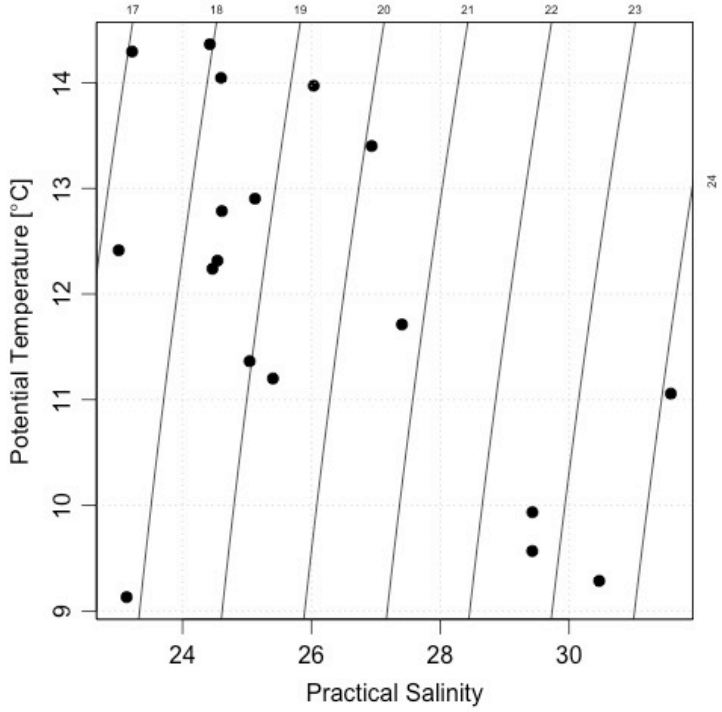

Figure A. 2: Temperature-Salinity (TS) plot of the SOG samples.

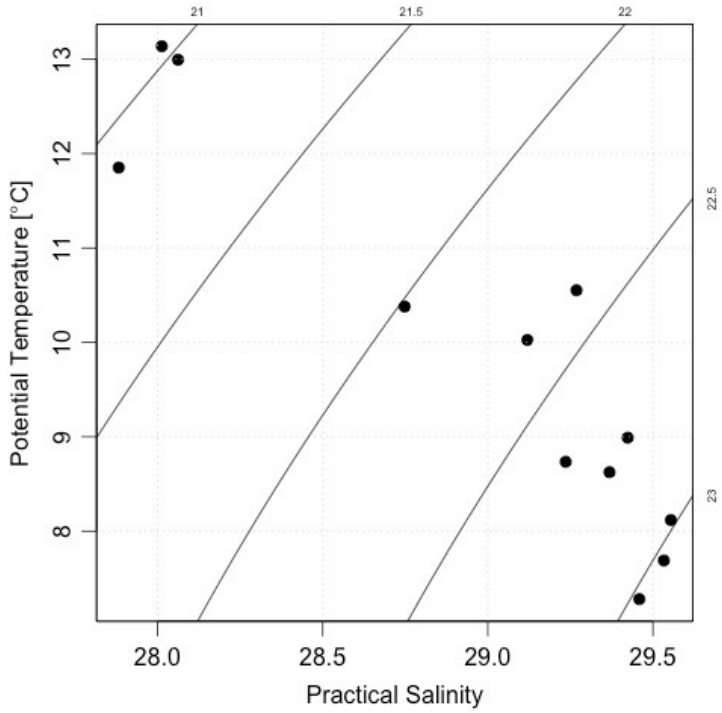

Figure A. 3: Temperature-Salinity (TS) plot of the SAA 10 meters samples.