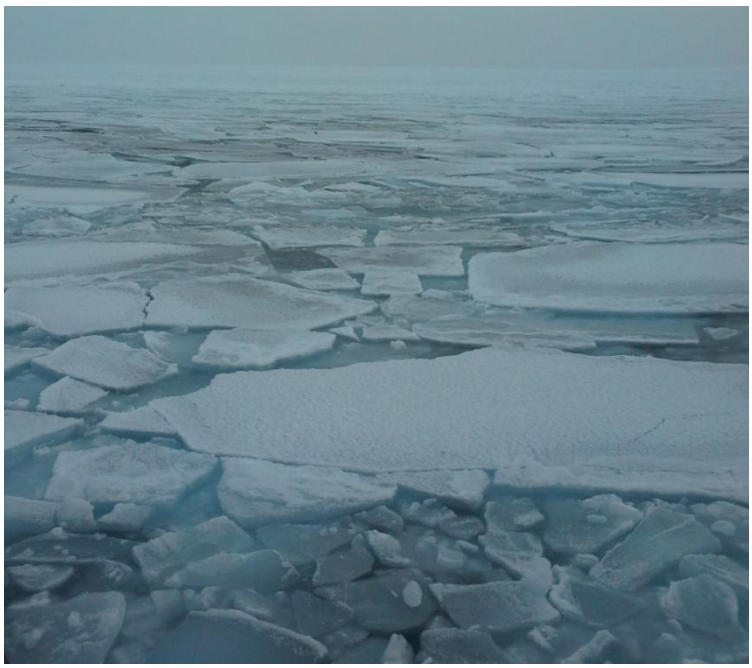
UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Diversity of dsDNA marine viral groups during winter in the Arctic Ocean north of Svalbard



Emily Olesin

Master's Thesis in Microbiology

2015

Table of Contents

Acknowledgements

I would like to acknowledge that financial support for this work was provided by the Norwegian Research Council through the MicroPolar Project (RCN 225956/E10).

This project has involved a steep learning curve, and literally took the aid of a small army of colleagues to complete. I would like to thank my main supervisor, Professor Ruth-Anne Sandaa, for her tireless work in guiding me in the production of this thesis and for her leadership role throughout my time in the graduate program at UiB: it has been a pleasure to be her master's student. I would also like to thank my secondary supervisor, Aud Larsen, for her invaluable input during the revision process and for being a wonderful cruise leader on the March 2014 research cruise I participated on with the MicroPolar team. Julia Storesund has my undying gratitude for her time spent over the past two years helping me to understand molecular viral ecology processes both in the lab and *in silico*. I owe most of my computational education to my bioinformatics advisor Bryan Wilson, who taught me everything I now know about using the command line and processing sequence data. Håkon Dahle also has my thanks for his assistance in this regard, especially in the Ion Torrent processing, and for our many productive discussions. I am grateful to Richard Telford for his help in decision-making for the statistics and for getting me started with R. Thanks to Louise Lindblom and Kenneth Meland at the Ion Torrent PGM facility in Bergen, whose guidance about the technology helped me better understand the post-processing. My thanks go to my many other excellent colleagues who answered my questions and gave advice, both within the Marine Microbiology Research Group and at the Centre for Geobiology: you are too numerous to list on one page, though I am grateful to all of you. Lastly, I thank my family for their patience and support during my writing process, especially my partner Alden.

I would like to dedicate this work to the scientist who got me started along this path, Professor Robert T. Wilce. Bob, I followed your advice and it took me to wonderful places and people I could not have possibly expected. Thank you.

Abbreviations and important terms

alpha rarefaction
resampling of data at stepped sampling depths to glean information about within-sample diversity (species richness and evenness).

alpha diversity
term used in this thesis to express within-sample diversity

ANOSIM
a non-parametric method to compare two or more groups of samples to test whether there are significant differences between those groups. ANOSIM uses permutations of the data to determine significance.

BBDuk
a fast and flexible opens source tool for adapter, quality, and contaminant trimming of sequence data, developed by Brian Bushnell (Joint Genome Institute)

beta diversity
term used in this thesis to express between-sample diversity (compositional dissimilarity of samples)

Chao1
an alpha diversity estimate of total species numbers within a sample

chimera
an artifact that occurs in sequence data as a result of two parent sequences from different sources fuse to one another during the amplification step. This artifact can artificially increase species diversity measures, especially in amplicon libraries when closely-related sequences are being amplified.

civil polar night
the condition, restricted to latitudes within the polar circle (above 72° 34'), when night lasts for more than 24 hours. In the territory of Svalbard, Norway civil polar night lasts from around 11 November to 30 January.

classical food chain
a term used by microbial ecologists to describe the complex higher trophic level food web that includes species whose interactions in the oceanic food web have been described previous to the advent of modern understanding of microbial input into the system.

CTD
acronym for conductivity/temperature/depth meter

DNA
acronym for deoxyribonucleic acid

dNTP
acronym for deoxynucleoside triphosphate.

DOC
acronym for "dissolved organic carbon".

DOM
acronym for "dissolved organic matter".

ds
double-stranded, in reference to DNA

g23
"gene product 23", referring to a major capsid protein gene of the double-stranded DNA viral family *Myoviridae*, which determines the form of the capsid head

grazer
term used microbial ecology to describe protistan or zooplanktonic species that prey upon prokaryotes.

Illumina
High throughput sequencing platform that utilizes the bridge amplification method and detects nucleotide addition using four fluorescent light signals

Ion Torrent
High throughput sequencing platform that detects nucleotide addition by changes in pH

phoH
an auxiliary gene with unknown function found in genomes across a diversity of double-stranded DNA viral families. Homologous gene in *Escherichia coli* is known to assist in phosphate uptake.

MCP
the major capsid protein gene of the algal viral family *Phycodnaviridae* and in the *Mimiviridae.*

HTS
acronym for high throughput sequencing, referring to massively parallel sequencing techniques.

Kill the Winner
A model for the population dynamics of phage–bacteria interactions where increase in a host population followed by an increase in its phage predator results in a more rapid rate at which the winner population is destroyed.

kmer
describes all possible substrings of a designated length (k) that are within a string

lateral gene transfer
describes transfer of genetic material within or between species through methods other than sexual or asexual reproduction.

lytic
a mode of viral lifestyle in which infection of a host results directly in cell lysis

lysogeny
a mode of viral lifestyle in which the genome of a virus integrates into the host genome or exists intracellularly as a plasmid.

metagenome
genomics of an entire environmental sample which includes the genetic material from many different organisms present.

microbial loop
a concept used to describe the significant contribution of marine microorganisms in the transport of carbon and other nutrients through the marine food web.

Mimiviridae
a viral family within the nucleocytoplasmic DNA viruses, which have amoebae hosts

mixotrophy
the ability some organisms have to use a combination autotrophic and heterotrophic modes of energy consumption

Myoviridae
a family of tailed dsDNA bacteriophages in the order *Caudovirales* that have prokaryotic hosts

NCLDV
nucleocytoplasmic large DNA viruses which include the *Phycodnaviridae* and *Mimiviridae* groups.

oligotrophic
a nutrient deplete environment.

omics
overarching term used to describe large-scale data-rich biology methods , such as metagenomics or proteomics

OTU
"operational taxonomic unit". Sequences are binned into the same OTU if sharing an indicated percent similarity at or above a threshold value (97% or 90% in this thesis). Commonly used in microbiology in place of a species definition.

paired-end sequencing
method of sequencing that analyzes both ends of a DNA fragment. Merging of paired-end reads created on the Illumina platform allows more coverage and accuracy by using forward and reverse reads for the same time and effort it takes to make a non-paired library preparation.

PCoA
acronym for principle coordinate analysis. PCoA is a multi-dimensional scaling method that with a given input distance matrix, will output a coordinate matrix that minimizes stress ultimately approximating the input distance matrix by reduction into only a few dimensions. This coordinate matrix can then be used in visualizations.

PCR
acronym for polymerase chain reaction

phage
shortened term used in place of bacteriophage or cyanophage, a term used to describe a virus that preys upon prokaryotes

phiX
control adapter-ligated library used by the Illumina platform that consists of fragments originating from the small and well-characterized genome of the phiX viral strain.

PHRED
a base-calling program that reads a DNA signal file (such as a chromatogram) and assigns a quality score based on its analysis of the peaks in the signal file.

Phycodnaviridae
A class of icosahedral algal dsDNA viruses, a group within the nucleocytoplasmic large DNA virsues

plasmid
a piece of intracellular DNA that is separate from the genomic DNA of the cell that can replicate independently.

POC
acronym for particulate organic carbon

prophage
temperate viral DNA integrated into the host genome

proteome
the entire set of proteins expressed by an organism, a cell, a group of organisms or a group of cells at any one time.

QIIME
"quantitative insights into microbial ecology" (pronounced "chime"). QIIME is an open-source bioinformatics pipeline designed for processing of raw prokaryotic sequence data from sequencing platform output files to a final statistical analysis and graphics.

rarefaction
method used to randomly resample a community to a common sequencing depth. This is used as a normalization technique within the QIIME pipeline.

RNA
acronym for ribonucleic acid

Roche/454
a HTS platform that performs sequencing by synthesis through which a light signal produced by a luciferin-luciferase reaction is detected upon nucleotide addition

R/V
acronym for research vessel

ss
single-stranded, in reference to DNA or RNA.

Unifrac
a method for comparing microbial communities that uses phylogenetic distance as a metric, which can be used to determine if communities are significantly different, and also in clustering and ordination techniques

UPARSE
a clustering algorithm designed by Dr. Robert Edgar to bin reads from microbial amplicon libraries into operational taxonomic units.

UPGMA
acronym for "unweighted pair group method with arithmetic mean"; a bottom-up hierarchical clustering method that defines dissimilarity between clusters as their average similarity. Used in this thesis to classify samples based on OTU composition.

temperate virus
a virus with a lysogenic replication cycle.

viral shunt
a process within the microbial loop in which the destruction of cells makes dissolved organic matter available for uptake only by microorganisms, thereby diverting energy that would have otherwise been passed up the classical food chain through predation on whole cells.

virion
the complete free living viral particle consisting of a protein capsid with the enclosed viral genome.

virulent virus
virus only capable of a lytic cycle.

virus
small acellular pathogenic agents (usually from 20 to 200 nm in size) that influence their hosts intracellularly via infection as obligate parasites.

WSC
acronym for the West Spitsbergen Current, which carries warm Atlantic-derived water north through Fram Strait.

Summary

Extreme changes in light and cold water temperatures throughout the annual cycle in the Arctic Ocean create a unique habitat that selects for particular microorganisms - including marine viruses. This study investigated diversity of ecologically significant viral groups at two marine sampling stations during the dark period in the Arctic Ocean north of the Svalbard archipelago through pyrosequencing of signature genes. Sequence data for three viral signature genes (*g23*, *phoH*, and *MCP*) were examined within the context of physical and biological environmental parameters to characterize the viral communities within several Arctic Ocean water masses of differing origin. Genotypic fingerprinting information from previous T4-like virus diversity investigations was used to explore phylogenetic relationships between Arctic Ocean *g23* genotypes examined in this thesis to a global diversity of T4-like viruses isolated from various environments. Our findings show that marine viral communities exhibit dominant and rare types that vary proportionally in abundance between water masses, and that the available prokaryotic host communities vary similarly. The biogeographic examination showed that many of the dominant Arctic Ocean T4-like genotypes from this study are possibly endemic to the arctic, while others show similarity to globally distributed types, supporting the paradigm that local viral diversity may be high while also being low globally.

Additionally, this study compared sequenced datasets of *g23* amplicons from the same water samples generated using three widely-implemented sequencing platforms (Roche/454, Illumina, and Ion Torrent) in order to assess comparability of data from newer platforms for viral diversity investigations to pyrosequencing data. The platform comparison revealed that clustering of signature gene sequences into OTUs based on 90% similarity resulted in preservation of broad patterns in between-sample diversity, and also that sequence read data generated using Illumina appear most similar to Roche/454. The author therefore recommends the Illumina platform for continued use of primers for amplification of viral signature genes developed for pyrosequencing.

1 Introduction

Viruses were historically thought of as relatively unimportant players in natural systems. In the late 1980s, however, Bergh et al. (1989) found that viruses of microorganisms are the most abundant biological entities in aquatic ecosystems. This discovery laid the foundation for today's rapidly evolving field of environmental virology. Viruses of microbial life are now known as vastly abundant, diverse, and universal agents that exert significant influence on host community structure and genetic composition. Through a wide array of interactions with hosts, viral infection of marine microbial communities ultimately affects biogeochemical cycling in the world oceans.

As global climate change continues to progress, questions arise about the consequences for ecosystems. Unique environments such as the Arctic Ocean are especially sensitive to climate change (IPCC 2001). Due to their high responsiveness to environmental conditions, shifts in the microbial populations of the ocean may be the most dramatically seen ecosystem variations as a result of climate change (Danovaro et al. 2011). As the base of the classical food chain, changes in the Arctic marine microbial community may therefore serve as a reliable signal to changes to come at higher trophic levels in the ecosystem. Few investigations of viruses within the microbial community in the Arctic Ocean exist to date; fewer studies still have examined marine viruses at high latitudes during the dark winter period. This study aims to assess the diversity of dsDNA viruses present at two sites in the Arctic Ocean north of Svalbard during the civil polar night.

Culture-free methodologies are necessary tools for environmental virologists: the majority of environmental marine microbes are not culturable and this issue is more accentuated in the case of marine viruses. Although developments in HTS technologies have produced the incredible capability to sequence all nucleic acids within an environment, gene fingerprinting tools remain useful to examine certain questions in marine microbial ecology. In this thesis, targeted gene (or tag amplicon) sequencing was used to characterize dsDNA virus groups known to infect hosts that play significant roles in the marine microbial community; namely the T4-like bacteriophage family *Myoviridae* (using a capsid protein gene *g23*), the unclassified algal virus family *Phycodnaviridae* (using the major capsid protein gene *MCP*), and a putative auxiliary metabolic gene of unknown function shared among a diversity of dsDNA bacteriophages called *phoH*.

As HTS has progressed as a technology, newer sequencing methods have surpassed one of the earliest HTS technologies, pyrosequencing, resulting in discontinuation of the pyrosequencing platform Roche/454 kit production in 2014. This transition requires testing of results obtained from current platforms against pyrosequencing results. Although platform comparisons exist for amplified prokaryotic (Claesson et al. 2010) and eukaryotic genes (Smith and Peay 2014), no comparison of environmental viral signature gene data exists. In this thesis identical samples of the *Myoviridae* capsid protein gene *g23* were used on three HTS platforms (Roche/454, Illumina MiSeq, and Ion

Torrent PGM) to investigate comparability between viral targeted gene amplification data outputs produced on the different sequencing platforms.

1.1 Viral characterization

Viruses are small acellular pathogenic agents (usually from 20 to 200 nm in size) that influence their hosts intracellularly via infection as obligate parasites (Suttle 2007). It is thought that every form of cellular life on Earth has at least one virus to which it is susceptible (Koonin and Dolja 2013). Even viruses infecting larger so-called "giant viruses" have been found in nature (Fischer and Suttle 2011; La Scola et al. 2008; Yau et al. 2011; Sun et al. 2010). The host range of a particular virus is often specific: for instance, a virus may only be able to infect few strains within a microbial species (Suttle and Chan 1994; Cottrell and Suttle 1995). In other cases, cells of related species may be susceptible to a virus with a broader host range (Wichels et al. 1998; Sullivan et al. 2003). One cellular species can be susceptible to infection by multiple viral types from phylogenetically distant viral families, which implies that viruses are the most diverse biological agents on Earth (Fuhrman 1999).

While genomes of cellular organisms are composed of DNA, viral genomes can be composed of DNA or RNA. Some viruses include both nucleic acid types during different life stages. Genomes of DNA and RNA viruses can be double-stranded (ds), single-stranded (ss), or a mixture of strand forms. The diversity of viral genomes includes linear, circular, or segmented arrangements. A virus existing outside of a cell (known as a virion) consists of a protein shell called the capsid which contains the viral genome. The structures of virions (Figure 1) are diverse in size, shape, and composition (Madigan 2012). Viral genomes mainly contain structural genes encoding capsids, tail proteins, insertion sites in the host genome or enzymes to lyse host cells (Weinbauer and Rassoulzadegan 2004). Most viral genes found through metagenomic studies do not originate from host cells but rather are unique to specific viral families and have no homology to any genes known within cellular life (Villarreal 2001).

Viruses rely on host cell machinery for their reproduction, which they accomplish via infection. Virus infection strategies includes the lytic and lysogenic phases (Figure 2). While some viruses are only capable of the lytic life cycle which leads to lysis of the host cell once infection begins (virulent viruses), others have the ability to enter an alternative phase called lysogeny. Viruses that can enter lysogeny are known as temperate viruses. Lysogeny is characterized by either integration of the virus genetic material into the host genome or existence of the viral genome in the cell cytoplasm as a plasmid (Lwoff 1953). Temperate viral DNA integrated into the host genome is termed a prophage, and can thus be passed on through generations of host cells before its eventual induction back to the lytic phase (Madigan 2012).

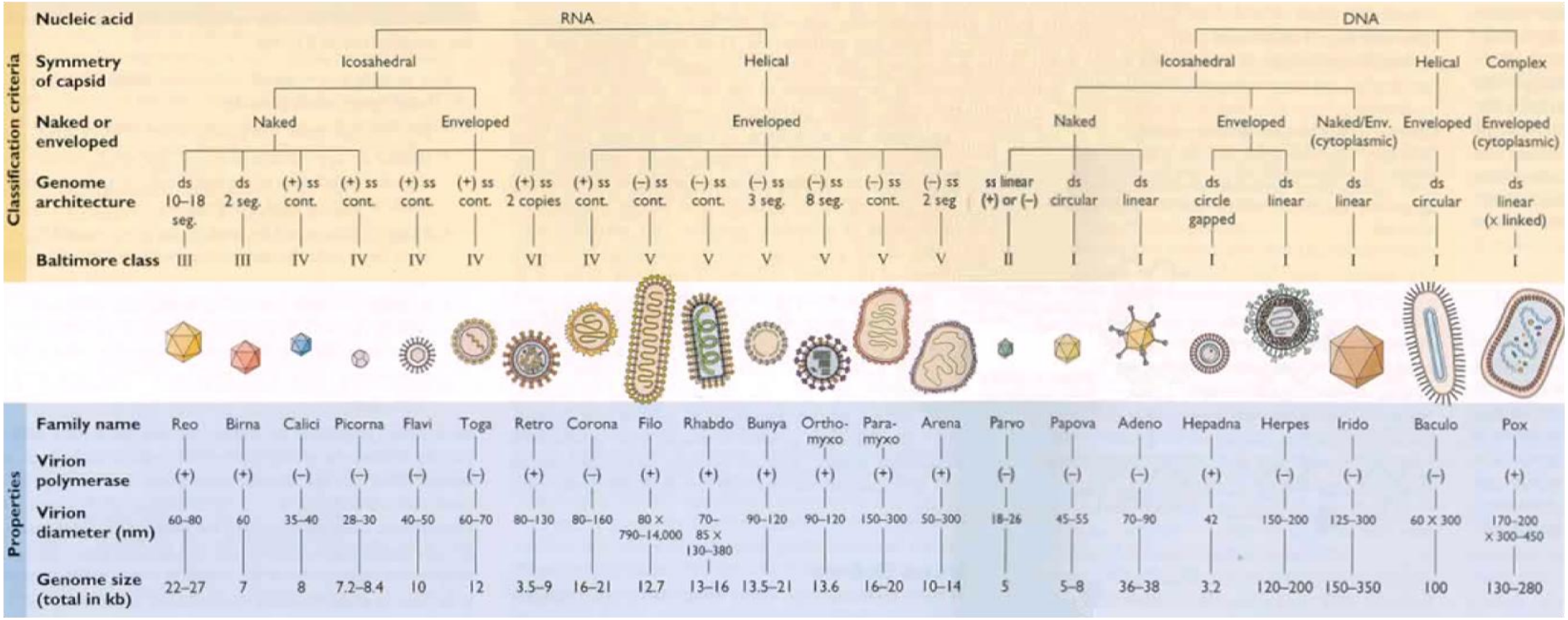


Figure 1. Morphology of different DNA and RNA viruses and their genomic diversity (sourced from http://www.nlv.ch/Virologytutorials/Classification.htm).

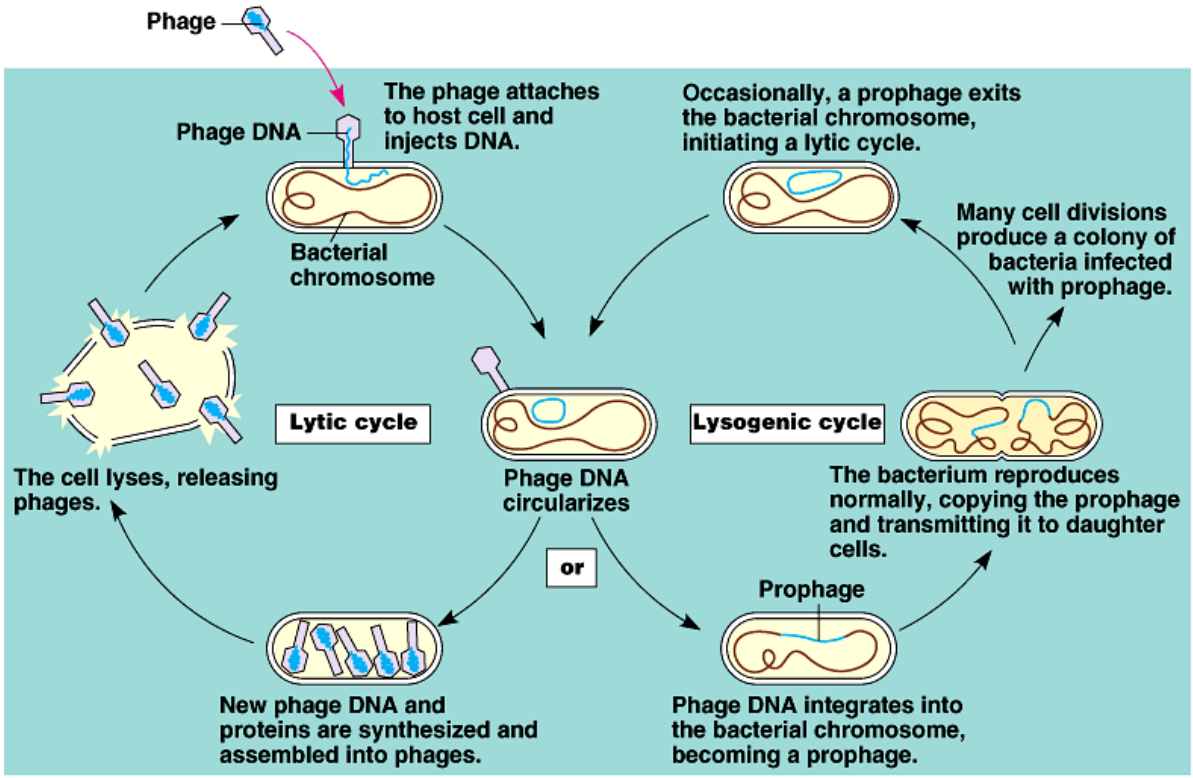


Figure 2. Graphic of lytic and lysogenic phases, both cycles are possible in temperate viruses while virulent viruses have only the lytic phase (Madigan 2012).

Whether or not virus particles qualify as living organisms has been debated for nearly a century. Distinguished pathologist and naturalist Professor Arthur Edwin Boycott's 1928 viewpoint about the nature of viruses may indeed be accurate:

"In this case 'live or dead' is a stupid question because it does not exhaust the possibilities. Our general notion of the structure of the universe leads us to expect that we shall meet with things that are not so live as a sunflower and not so dead as a brick, and a consideration of what we know about 'filterable viruses' and similar 'agents' brings us to the conclusion that they represent part of this intermediate group (Boycott 1928)".

The scientific community still has not come to a clear resolution of the placement of viruses in the context of evolution and the origin of life. Discoveries of new viral types continue to blur the lines between cellular life and viruses. Some workers have proposed that giant viruses may occupy a fourth branch of the tree of life (Boyer et al. 2010). Discovery of viral particles larger than many prokaryotes and larger even than some eukaryotes with translation machinery in their proteomes perpetuate this argument (Claverie and Abergel 2013). Other researchers have found phylogenetic evidence in contradiction to the fourth branch of life theory (Yutin et al. 2014). A recent deep and broad investigation of proteomes across the known spectrum of viral types and cellular life points to a common origin for modern cells and their viruses, and implies both viruses and cells have evolved commonly from

multiple ancient “protovirocell” types (Nasir and Caetano-Anolles 2015). The authors suggest that modern viruses were reduced to non-cellular entities over the course of their evolution. The proteomic data from the study strongly suggest that viruses are phylogenetically placed in the universal tree of life as entities constituting a fourth group (Nasir and Caetano-Anolles 2015).

1.2 Ecological impact of marine viruses

The estimated global abundance of viruses is approximately 10^{31} particles (Suttle 2005). In surface seawater viral particle abundances can be upwards of 10^{8} viruses mL^{-1} (Bergh et al. 1989). They are ubiquitous in the ocean, from the sea surface to deep marine sediments. Around 94% of total nucleic acid containing particles in the ocean are virus particles (Suttle 2007). The inconceivably high abundance and diversity of viruses in the ocean allows viral activity to significantly impact the marine ecosystem (Wilhelm and Suttle 1999). Marine viral activity influences microbial community structure (Longnecker et al. 2010; Thingstad et al. 2015; Thingstad et al. 2010), can terminate blooms of planktonic species (Bratbak et al. 1993; Larsen et al. 2001), and allow transfer of genetic material within and between both viruses and hosts (Sobecky and Hazen 2009). Through their predation on microbes, marine viruses ultimately affect evolution of organisms in the oceans and global biogeochemical cycling.

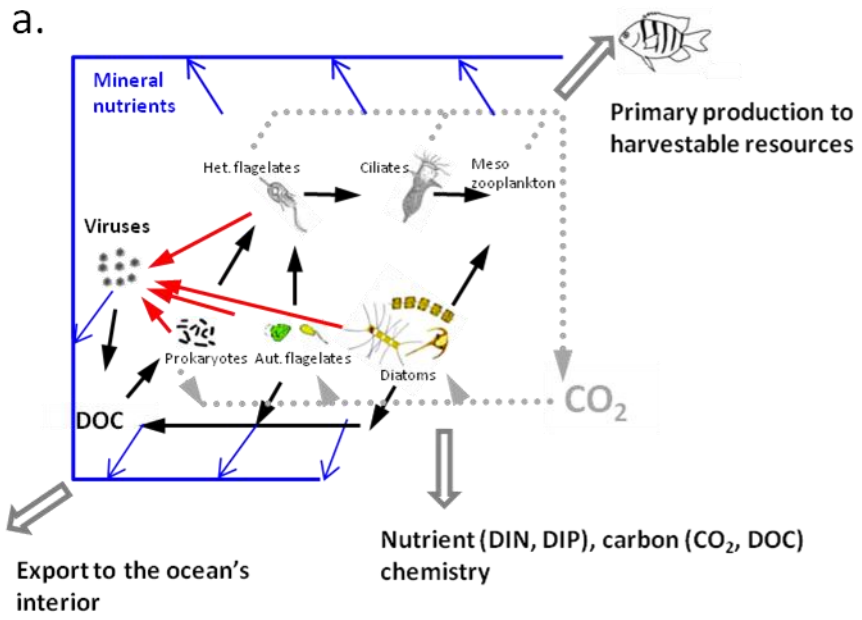
Viruses rely on the presence of a host for reproduction in their environment and the frequency of interaction between virus and host is a limiting factor to that reproduction. In fact, infection via diffusion would likely be improbable in the ocean without a high density of viruses and available microbial host cells, and without the water currents allowing movement of particles (Dennehy 2013). It is therefore not surprising that the majority of the virus particles in the natural environment infect the most consistently available hosts; prokaryotes (Fuhrman 1999; Wilhelm and Suttle 1999; Wommack and Colwell 2000). This is reflected in the fact that marine viral particles are often found in ratios of 5-10 particles per bacterial cell. Viruses that prokaryotes are susceptible to are known as bacteriophages or cyanophages (often abbreviated to “phages”). The microbial species best-adapted to the current environmental conditions have a strong selective pressure due to marine viral predation. In this way, marine viruses adjust not only microbial cell abundance and production but also affect the representation of microbial species within the local environment by suppressing dominant ecotypes, as described in the Killing the winner hypothesis (Thingstad 2000). This effect has been experimentally tested, confirming that bacterial community assemblages differ in the presence and absence of viral predation (Fuhrman and Schwalbach 2003; Bouvier and del Giorgio 2007).

Viruses contribute to the recycling of nutrients within the context of the microbial food web, also known as the “microbial loop” (Azam et al. 1983). In basic terms, the microbial loop concept describes cycling of dissolved organic matter (DOM) (mainly released from phytoplankton) within a microbial food chain consisting of a complex web of energy transfer between viruses, prokaryotes, diatoms, dinoflagellates and other micro-sized phytoplankton, and microzooplankton (Figure 3a). About half of the organic carbon fixed by phytoplankton in the world ocean passes through the microbial loop. The importance of the microbial loop on the overall food web varies between differing

local environmental conditions. The most marked influence of microbial processes is thought to occur in consistently oligotrophic waters (Munn 2011).

Viral lysis creates a majority of the DOM cycling through the microbial loop, though some DOM is also contributed through messy eating by grazers (Munn 2011). Viral infection of the plentiful bacteria, archaea, algae and protists is responsible for massive cell lysis in the ocean. This lysing influence is referred to in microbial ecology as the "viral shunt"(Wilhelm and Suttle 1999) (Figure 3b). The level of bacterial mortality due to the viral shunt is at least as large as that due to predation by grazers (Fuhrman 1999; Wommack and Colwell 2000). Phage alone are responsible for 10-50% of daily bacterial mortality through infection (Fuhrman 1999), and have been shown to consistently destroy bacteria on this scale across different environments (Suttle 1994).

Virus-host interactions drive an antagonistic co-evolution of predator and prey. Selection of resistance to viral infection in the host population requires evolution of a virus to maintain its virulence. This biodiversity-promoting "arms race" between viruses and their microbial hosts has been underway for billions of years (Buckling and Rainey 2002). Viruses also facilitate genetic exchange between and within host populations by horizontal and vertical gene transfer. Lateral gene transfer via viruses influences competitiveness between microbial host populations in the environment. For example, if a virus forms a mutualistic relationship with the host by conferring new metabolic traits to the host, this new trait may increase host fitness and the virus' chance of survival (Weinbauer and Rassoulzadegan 2004). Inter-species genetic exchange via viruses has been proposed to metaphorically "shake the tree of life" (Pennisi 1998), likely making a universal classification of organisms based on phylogenies impossible (Weinbauer and Rassoulzadegan 2004).



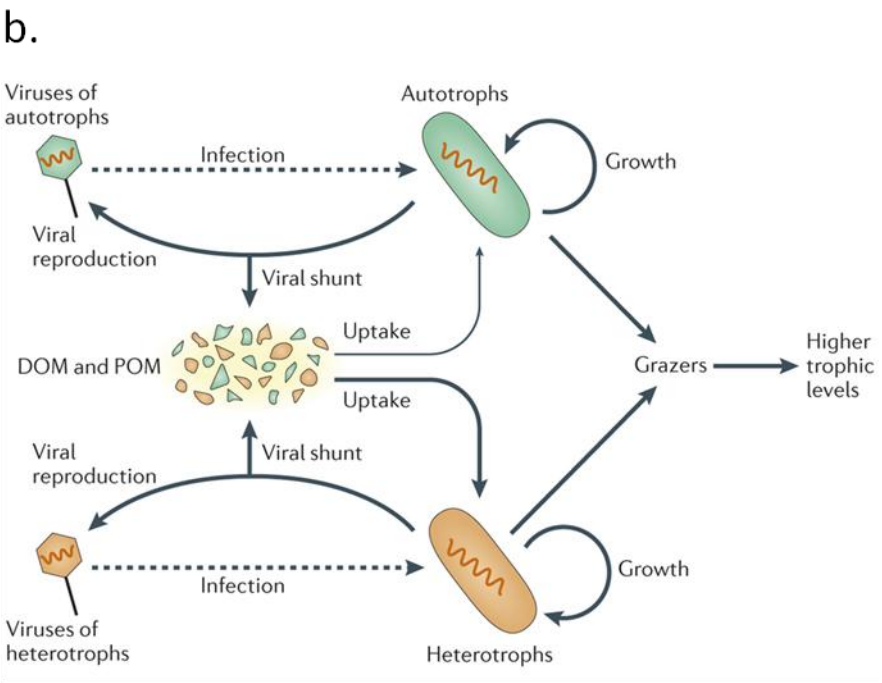


Figure 3a. Energy is transported to different trophic levels via the marine microbial food web, the players in which are comprised of prokaryotes, single-celled eukaryotes, and viruses. Red arrows indicate the viral lysis-mediated transformations of energy, black arrows refer to transport of dissolved and particulate organic carbon (DOC/POC), grey dotted arrows indicate the cycling pathways of CO_2, blue arrows refer to transport of mineral nutrients, and outlined grey arrows indicate net contributions to large scale ecosystem function (figure courtesy of Ruth-Anne Sandaa).

Figure 3b. The consequences of lysis through viral infection of autotrophic and heterotrophic prokaryotes are also known as the "viral shunt" (Jover et al. 2014).

1.3 Microbial communities in the Arctic

The famous 1934 hypothesis of Lourens Baas-Becking (Baas-Becking 1934) stating "Everything is everywhere, but the environment selects" has been a topic of heated debate regarding distribution of microbial species in the natural environment, including marine viruses. Studies of prokaryotic biogeography conclude that differences in species assemblage are influenced both by regional factors such as historical events in the environment, and also by local physical and biological aspects of the environment (reviewed in Lindström and Langenheder 2012). In the Arctic Ocean, studies of prokaryotes find abundant phylotypes remain abundant and rare phylotypes remain rare with changing season (Kirchman et al. 2010) and depth (Galand et al. 2009), contrary to Baas-Becking's hypothesis. If the rare biosphere serves as a "seed bank" in wait for conditions that favor these rare types, these observed patterns showing rare types remain rare and abundant types remain abundant would be unexpected. Instead, differences in the makeup of the microbial community assemblage appear to coincide with barriers within the marine environment which limit their dispersal (e.g. density gradients) (Galand et al. 2009; Gómez-Pereira et al. 2010; Varela et al. 2008).

1.3.1 The Arctic viral community

A paucity of information still exists regarding marine viral genomics to confirm or refute Baas-Becking's hypothesis for viruses in the ocean, and even less is known about Arctic Ocean viral community species identities and ecological functions. Studies of the microbial community in the Arctic Ocean are few in number due to the logistical challenges associated with sampling in the polar regions. Numerous studies including several in temperate and subarctic regions find that viral community assemblages follow biogeographic patterns (e.g. Pagarete et al. 2013; Needham et al. 2013; Sandaa and Larsen 2006; Goldsmith et al. 2015; Payet and Suttle 2014; Winter et al. 2013) while other studies find no such relationships, possibly due to broad passive viral dispersal (Snyder et al. 2007; Breitbart, Miyake, and Rohwer 2004; Short and Suttle 2005).

Bacteriophage virions are generally produced in greater numbers under environmental conditions favoring fast bacterial growth and productivity (Chibani-Chennoufi et al. 2004), which is generally not the case in the open Arctic Ocean. In a study of sea ice microbial communities, however, 10 to 100-fold higher viral abundances were observed within the ice than within the water column beneath during the spring ice algal bloom (Maranger et al. 1994). A majority of the viruses observed in the study were likely bacteriophages, based on their small capsid diameters.

The lysogenic phase is typically more common during times of low host abundance and productivity and in oligotrophic waters (McDaniel et al 2006). A metagenomic study of the marine viral community in the Arctic found a high abundance of temperate DNA bacteriophages (Angly et al. 2006). The indication of this finding is that viruses integrate into the genomes of their bacterial hosts as prophages to a greater degree in the Arctic than in warmer lower latitude waters. This conclusion has been further supported by studies examining the prevalence of prophage genes in low productivity environments such as Antarctic lakes (Laybourn-Parry et al. 2007) and also in mesopelagic

and deep waters (Weinbauer et al. 2003). When a virus is intracellular and integrated into the genome or exists as a plasmid, this lysogenic state provides advantages of UV protection and avoidance of destruction via enzymatic activity (Madigan 2012).

1.3.2 Arctic *Bacteria* and *Archaea* communities

The Arctic Ocean is dominated by bacteria although it is important to note that archaea are more abundant in cold, high latitude waters than in more temperate or tropical oceans (Wells et al. 2006). In the western Arctic Ocean, archaea are more abundant in layers near the seafloor containing suspended material from bottom sediments (Wells and Deming 2003). *Crenarchaeota* group Marine Group I (recently reclassified as *Thaumarcheota* (Brochier-Armanet et al. 2008)) has been observed as the most abundant archaeal group in the Canadian Arctic Ocean. Phylotypes within this group have also been shown to dominate the archaeal assemblages within sea ice and ice-influenced surface waters in the western Arctic Ocean (Collins et al. 2010).

The most abundant members of the prokaryotic community are thought to be well-adapted to the local environment and to contribute a majority of the biomass production (Cottrell and Kirchman 2003; Zhang et al. 2006). At high latitudes these species are adapted to low temperatures (Connelly et al. 2006). Arctic Ocean bacterial production was previously thought to be low in respect to that of other oceans (Rich et al. 1997). Evidence now indicates that production can be quite high depending on local environmental conditions (Wheeler et al. 1996) and that much of this activity is heterotrophic (Rich et al. 1997). This affects the carbon cycling and the food web structure of the Arctic Ocean (Kirchman et al. 2009). In a study near the Canadian Arctic (Arctic Ocean) it was shown that 53% of the bacterial species belonged to *Gammaproteobacteria*, and nearly all other clones were either from the *Bacteroidetes* or the *Alphaproteobacteria* (mainly SAR11) (Collins et al. 2010). SAR11 clade *Alphaproteobacteria* have also been observed to dominate bacterial assemblages of winter sea ice in the western Arctic Ocean (Collins et al. 2010).

1.3.3 The Arctic phytoplankton community

In temperate oceans, the key phototrophic organisms are usually *Synecococcus sp.* and *Prochlorococcus sp.* In the high Arctic, however, small pico- and nano- eukaryotes dominate as the baseline phototrophs. In particular, the single-celled algal species, *Micromonas pusilla* (Butcher) Manton and Parke 1960 (Prasinophyceae (Chlorophyta)), is a key phototroph in the Arctic Ocean (Lovejoy et al. 2006). A cold and low-light adapted *M. pusilla* ecotype has been found in the Canadian Arctic (Connie Lovejoy et al. 2007). The over-wintering strategies of the pico- and nanoflagellar autotrophs are as yet unknown, though live cells of *M. pusilla* have been detected in surface waters down to 1,000 m in the middle of the civil polar night. It has been suggested that *M. pusilla* could be capable of alternate life strategies such as phagotrophy to maintain cell functions in the absence of light (Vader et al. 2015).

Although prokaryotic photosynthetic organisms dominate the world oceans, nanoplanktonic non-calcifying haptophytes are the most abundant and diverse group of picophototrophs in modern oceans, representing the "background" light harvesters of the world ocean (H. Liu et al. 2009). Their success in the marine environment may be due to their ability to prey upon bacteria as well as photosynthesize, known as a mixotrophic lifestyle. Noncalcifying haptopytes are closely related to the coccolithophores, the most well characterized group of haptophytes. Haptophyte populations appear to show geographic specificity of genotypes; a study found that certain lineages appear limited to the colder mixed waters of the subarctic (H. Liu et al. 2009). A small, bloom-forming haptophyte endemic to the Arctic, *Phaeocystis pouchetii* (Hariot) Lagerheim 1893 (Prymnesiophyceae), is an important player in biogeochemical cycles, especially sulphur cycling. A class of icosahedral algal dsDNA viruses known as the *Phycodnaviridae* includes previously isolated viruses able to infect *M. pusilla* (Cottrell and Suttle 1991) and *P. pouchetii* (Jacobsen et al. 1996) .This family of morphologically similar viruses is covered in greater detail below in section 1.7.2.

Ice algae are another component of the primary production in the Arctic. A study of sea ice in a subarctic fjord off Greenland including information on the winter season found the dominant algal groups within the ice to be cryptophytes, prasinophytes, and unidentified small flagellates in January and February, whereas later in the spring the microalgal community is dominated by pennate diatoms (Mikkelsen et al. 2008). Algal blooms form along the ice edge as a result of freshwater input from ice melt, and can sometimes extend for hundreds of kilometers behind the retreating ice (Perrette et al. 2011). It has long been questioned whether sea ice algal assemblages might inoculate these ice-edge phytoplankton blooms (Syvertsen 1991). Additionally, findings that diatom spores and dinoflagellate cysts are more abundant in sea ice than surface waters indicate that sea ice entrapment may serve as an overwintering strategy for some algal species (Różańska et al. 2008).

1.4 The Arctic environment

Research aimed at clarifying the relationships between physical oceanography and the marine microbes indicate that community structure of microorganisms in the oceans is highly determined by the mass of water in which the community resides (Galand et al. 2010). As microbes are highly sensitive to local environmental conditions (e.g. salinity, temperature, nutrient availability), it is important to consider the physical oceanography of the Arctic marine system to understand its microbial community.

The Arctic Ocean is essentially landlocked by the surrounding continents and as a consequence there are physical limitations on the entry of southern water masses to the Arctic basin. The Gulf Stream carries warmer Atlantic water north (red arrow in Figure 4), continuing as the West Spitsbergen Current (WSC) through the only deep gateway to the Arctic, the narrow 500 km wide Fram Strait, along the western coast of the Svalbard archipelago to either ultimately reach the Arctic Ocean or recirculate through Fram Strait. On its journey through the North Atlantic this

highly saline warmer water cools and sinks while also introducing energy into the Arctic Ocean. Fresher Arctic Ocean water and sea ice are exported south along the Greenland shelf through Fram Strait as the East Greenland Current. These two currents represent the main exchange of Arctic Ocean water with the rest of the world's oceans (Arctic Council 2013).

According to Rudels et al. (1991), the water masses formed or transformed in the Arctic Ocean are: 1) The 50 m deep Polar Mixed Layer, with freezing temperature and salinity of about 32.7 psu close to the Fram Strait. 2) The halocline found between 50 m and 250 m depth, with a salinity range from 33 to 34.4 psu and temperatures mostly close to freezing but increase at the lower boundary to 0°C. 3) The 400-600 m thick Atlantic Water layer with temperatures above 0°C and increasingly saline (34.4 to 34.9 PSU) with depth. 4) Deep Waters below 800-1000 m with salinities of 34.93-34.95 PSU and potential temperatures ranging from 0°C to -0.95°C at the bottom. The differing characteristics of these water masses may act as boundaries and serve as selective environments for certain groups of microbes (Galand et al. 2010).

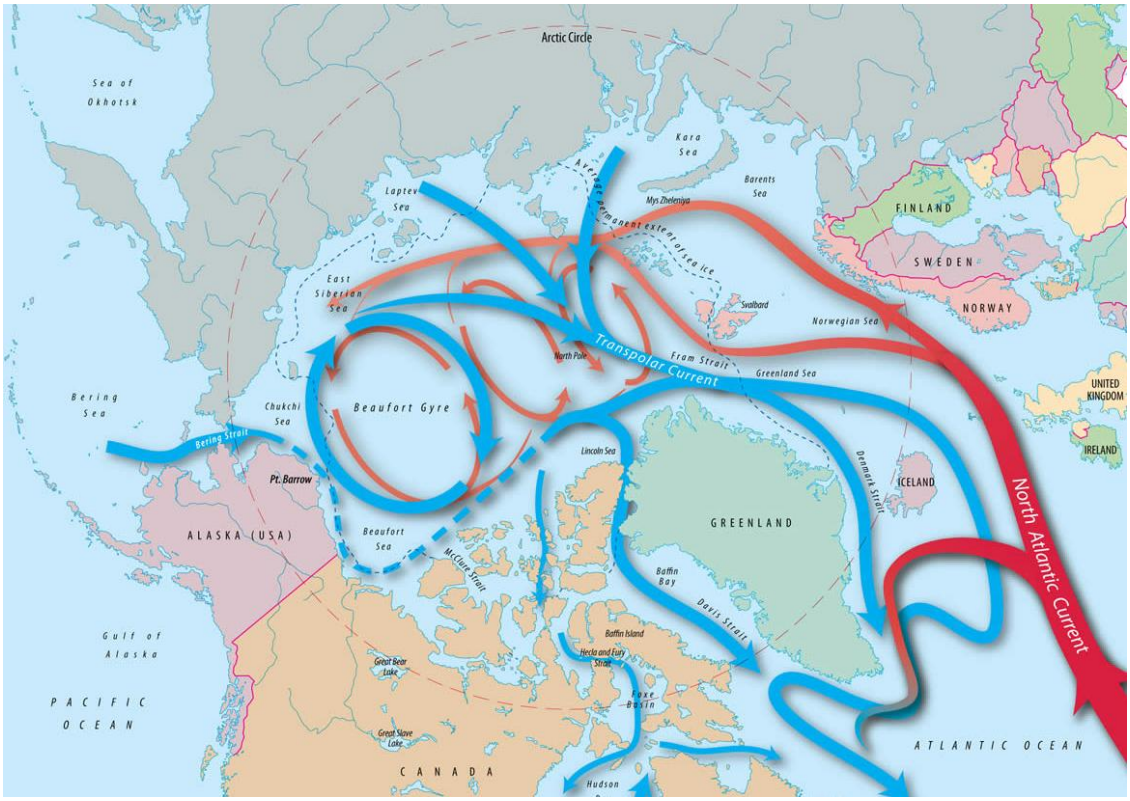


Figure 4. Illustration showing Arctic Ocean currents and those of surrounding seas (sourced from Cook (2015)).

While the Arctic Circle is completely devoid of sunlight during winter, for the rest of the annual cycle mixing influences light penetration through the water column. Another unique characteristic of polar oceans is the presence of sea ice. Wind-mixing of surface waters is constrained in parts of the Arctic Ocean where sea ice is perennial, creating conditions which would not otherwise exist in the open Arctic Ocean. Per volume production is higher within sea ice than in the pelagic zone beneath (K. R. Arrigo 1997) making it an ecologically important habitat in the Arctic Ocean.

The formation and subsequent melting of sea ice contribute to significant and continuous water-column stratification due to the salinity gradient. A cold, low density water layer forms from sea ice melt on top of warmer Atlantic water layer at the frontal zone north of Svalbard where ice meets the open ocean, creating a stark density gradient (Rudels et al. 1991) which could also create a niche environment selecting for certain microbes.

1.4.1 Climate change

Global climate change is expected to cause shifts in the unique Arctic Ocean ecosystem, including the structure and function of the marine microbial community. Changes in the microbial community are expected to be the most dramatic of all biological assemblage shifts in the ocean (Danovaro et al. 2011). Many components of global climate change could impact Arctic marine microbial ecology including sea ice melt, carbon cycle changes, and temperature fluctuations. Moreover, in terms of changes in microbial food web dynamics, changes in host communities will affect the structure and function of viral communities.

Increased annual sea ice melt could result in greater transport of Atlantic water masses, adjusting the currently restricted entry of lower-latitude waters into the Arctic. Species abundance and diversity in the microbial community may change from the present state if greater influence from warmer southern waters becomes the new norm and the availability of nutrients is altered. Stronger winds in areas where sea ice is no longer present year-round could result in greater mixing in the upper the water column, eradicating environmental niches for some microbes (Danovaro et al. 2011). Changes in the length of the ice melt season could also alter the extent and nature of ice-edge algal blooms (Arrigo 2013), and increase the incidence of timing mismatches that already occur between life cycles of microbial species and the higher trophic groups that prey upon them (Conover and Huntley 1991).

The global oceans contain ~95% of the mobile carbon reservoirs on the planet, most of which is stored inorganically in the form of HCO_3. Carbon dioxide is more soluble in colder water, thus the cold bottom water at the Poles is enriched with CO_2. As temperatures at the Poles increase, solubility of CO_2 decreases, resulting in CO_2 release into the atmosphere as water is upwelled from the deep ocean. The amount of dissolved inorganic and organic carbon in the ocean is high relative to the CO_2 in the atmosphere, therefore small changes in the oceans carbon cycling can result in an enormous disruption of annual CO_2 exchange with the atmosphere(Raven et al. 2005). It is currently unknown how the microbial community will influence and respond to the changing carbon cycling, as microbes of the world ocean heterogeneously provide a carbon source in some areas and act as a carbon sink in others, and this varies over time (Iversen and Seuthe 2011). Ocean acidification (resulting from anthropogenic input of CO_2 into the atmosphere and its subsequent absorption into the ocean) presents a problem for calcifying organisms, especially for important primary producers such as the calcifying phytoplankton group known as the coccolithophorids, as calcification rates are expected to lower under ocean acidified conditions (Beaufort et al. 2011).

Although there are conflicting reports about the levels of bacterial production in the Arctic Ocean, observations of lower levels of bacterial biomass production relative to primary production have been made in the Arctic Ocean and other polar environments compared with those of lower-latitude marine environments (Kirchman et al. 2009) (the reasons for such observations have been subject to debate (Brum et al. 2015)). Many biological processes may shift if lower-latitude waters bring more internal heat to the Arctic Ocean in the future, though heterotrophic processes are thought to be more sensitive to temperature than autotrophic processes (Wohlers et al. 2009). Some researchers have put forward the possibility of functional food web shifts at different trophic levels in response to rising temperature in the Arctic (e.g. Rose and Caron 2007; Pomeroy and Deibel 1986), though this inference is not agreed upon within the scientific community. For instance, a study based on mesocosm experiments in Kongsfjorden, Svalbard found the Arctic microbial system was predicted as adaptable to temperature increase when a mathematical model previously shown to reflect observations in the marine system was applied (Larsen et al. 2015).

1.5 Viral diversity through the lens of targeted gene sequencing

Metagenomic studies have indicated that thousands of dsDNA virus genotypes can be found within 10 -100 liters of seawater and that even the most abundant types comprise very little of the entire assemblage (Angly et al. 2006; Breitbart et al. 2004). While metagenomics are an excellent tool for analyzing overall biodiversity of a microbial population (Weinbauer 2004), species diversity may be examined using other tools at hand to the microbial ecologist. Although signature genes are an insufficient basis for in-depth viral identification, they provide a means to determine the number of phylotypes in an environment (Weinbauer and Rassoulzadegan 2004). Despite the challenges associated with characterizing such exceptionally diverse viral phylogenies, conserved marker genes have been identified to describe species diversity within groups of viruses considered to be dominant in marine systems. These include three genes capturing different viral groups, namely, a major capsid protein gene from the *Myoviridae* family (gene product 23 a.k.a. *g23*) (Tétart et al. 2001), a widely distributed auxiliary gene encoding a product with unknown function found within a diversity of phage families (phoH) (Goldsmith et al. 2011), and a gene encoding the major capsid protein from the *Phycodnaviridae* and *Mimiviridae* (*MCP*) (Larsen et al. 2008).

1.5.1 *Myoviridae*

Around 95% of dsDNA bacteriophage isolates from the marine environment have an icosahedral capsid and a filamentous tail attached at one of the icosahedron vertices (Wommack and Colwell 2000; Ackermann 2007). Tailed phages belong to the viral taxonomic order *Caudovirales*, which is further divided into three families based on their tail morphologies: the *Siphoviridae* (long non-contractile tailed), the *Podoviridae* (short tailed) and the *Myoviridae* (long contractile-tailed). The type species of the *Myoviridae*, simply called T4 (short for type 4) was first isolated from *Escherichia coli* around 1945. Although the origin of the discovery remains ambiguous, the isolate is likely sourced from feces or sewage (Abedon 2000). Metagenomic studies indicate that T4-like phages comprise a significant portion of the marine virus community (Breitbart et al. 2002; Angly et al. 2006). Myophage are known to have

broader host ranges than other tailed viral types. An example of this has been shown for a myophage that broadens its host range under low –light conditions (Chibani-Chennoufi et al. 2004). The *g23* major capsid protein gene is one of the more widely used genetic markers in environmental studies of virus communities (e.g. Filée et al. 2005; Bellas and Anesio 2013; Liu et al. 2012; Zheng et al. 2013; Fujihara et al. 2010; Wang et al. 2009; Wanget al. 2009; Needham et al. 2013; Chow and Fuhrman 2012; Pagarete et al. 2013; Butina et al. 2013, Chow et al. 2014). By using primers designed to amplify a conserved region of *g23,* diversity of T4-like bacteriophages within an environmental sample may be distinguished (Filée et al. 2005).

1.5.2 *Phycodnaviridae* and *Mimiviridae*

Viruses of eukaryotic algae have ecological significance as predators of one of the major groups of global primary producers. Two groups of dsDNA algae viruses known as the *Phycodnaviridae and Mimiviridae* includes members which have extraordinarily large genomes and a wide range of particle sizes. Phycodnaviruses are known to infect prasinophytes, chlorophytes, raphidophytes, phaeophytes, and haptophytes (Wilson et al.2009). The *Mimiviridae* family formally contains viruses isolated from heterotrophic protists (Mimivirus and *Cafeteria roenbergensis virus*) (Fischer et al. 2010; La Scola et al. 2003). Some viruses that infect prasinophytes and haptophytes, as well as some uncharacterized viruses with unknown hosts are also phylogenetically assigned to this family (Larsen et al. 2008; Sandaa et al. 2001; Johannessen et al. 2015). The *Phycodnaviridae* include members known to infect harmful algal bloom species such as the fish-killing raphidophyte alga *Heterosigma akashiwo*. Based on gathered evidence in microbial ecology, the activity of viruses within the *Phycodnaviridae* infecting bloom-forming species can be a determining factor in the initiation and termination of blooms, such as in the case of the coccolithophorid *Emiliania huxleyi.* Although these algae viruses may contribute to the boom and bust of phytoplankton blooms, perhaps their most important contribution is their role in maintenance of microbial community diversity and prevention of bloom formation (Wommack and Colwell 2000; Brussaard 2004). There are relatively few characterized members of the *Phycodnaviridae* to date, making them a challenging group to investigate for phylogenetic relationships. Additionally, the nucleocytoplasmic large DNA viruses (NCLDV) which include the *Phycodnaviridae* and *Mimiviridae* have been found to share only nine genes in common (Wilson et al. 2005; Van Etten et al. 2014; Iyer et al. 2006). Among these shared genes, the major capsid protein gene contains interspaced conserved regions used to fingerprint this family of viruses for investigations of their phylogenetic relationships and community diversity (Larsen et al. 2008).

1.5.3 Auxiliary metabolic genes

Auxiliary metabolic genes (AMGs) were once thought to be restricted to cellular life but have since been identified in many viral genomes through molecular methods. Groups of marine phages have been found to contain AMGs involved in nutrient limitation, carbon metabolism, nucleotide metabolism, and photosynthesis (Chenard and Suttle 2008; Sullivan et al. 2006; Lindell et al. 2005; Millard et al. 2004; Sullivan et al. 2005; Rohwer et al. 2000; Sullivan et al. 2009; Weigele et al. 2007). One AMG used in studies examining viral diversity is *phoH*, a gene of unknown function in viruses found in multiple families of dsDNA tailed phage. The *phoH* gene has been found in a diversity of

virus groups infecting a phylogenetically wide host range including groups of autotrophic and heterotrophic bacteria, and some autotrophic eukaryotes (Figure 5). The across-family diversity capture of *phoH* makes it a valuable signature gene for studies of marine viral community diversity, and studies have shown that the gene is widely spread in the viral fraction in marine environments (Goldsmith et al. 2011; Goldsmith et al. 2015). The primer set developed for *phoH* in viruses captures cyano and bacteriophages genes, and does not amplify known homologous bacterial *PhoH* genes (Goldsmith et al. 2011).

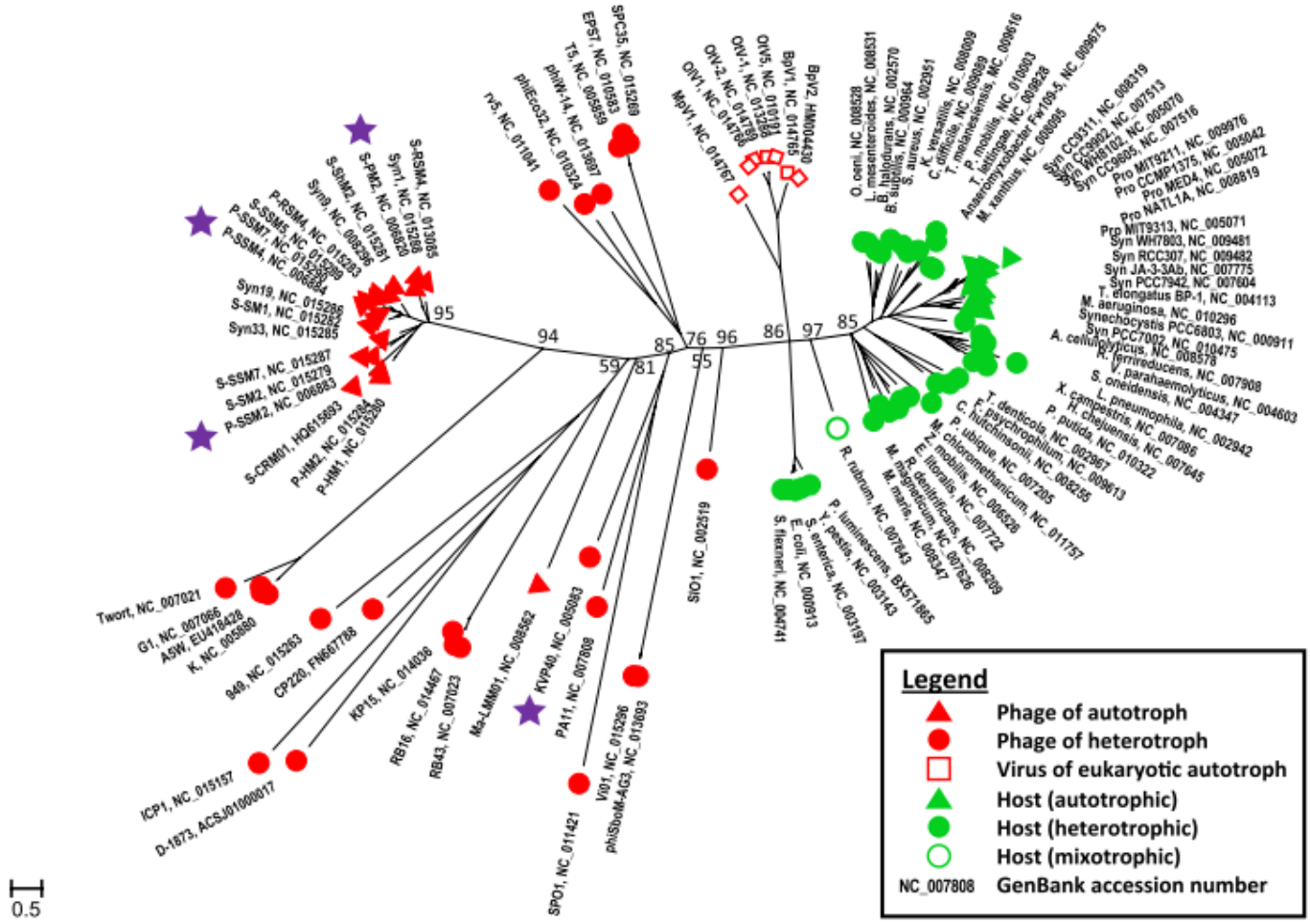


Figure 5. Phylogenetic tree showing the relationships of available *phoH* sequences within the NCBI database (as of 2011) found in a diversity of viruses and those of prokaryotes and some eukaryotes (figure from Goldsmith et al. 2011).

Some believe *phoH* interacts with the host Pho regulon in the uptake and metabolism of phosphate during phosphate-starved conditions, as the homologous gene in *E. coli* does (Hsieh et al. 2010, Wanner 1996), though homologs of the *E. coli PhoH* within different bacterial species have other possible functions (Kazakov et al. 2003), thus this putative function of the *phoH* gene in viruses has been contested.

1.6 HTS for viral diversity investigations

In the mid 2000's, HTS technologies forever changed standard operating procedures in genetic microbial ecology. The largest difference between traditional Sanger sequencing and HTS is throughput: a single Sanger sequencing run

generates in the region of 100s of sequences of 600- 900 bp length, while HTS technologies such as Roche/454 and Illumina can produce from 10^6 to 10^9 sequences of 100- 700 bp lengths per run (Table 1) (Logares et al. 2012). Each HTS technology platform available to date has advantages, drawbacks, and preferable applications (Table 2). Three platforms currently widely in use by microbial ecologists are the Roche/454 FLX Titanium, Illumina MiSeq, and Ion Torrent PGM platforms (hereon referred to as Roche/454, Illumina, and Ion Torrent). Comparing these platforms, Illumina has the highest throughput per run and the lowest error rates. The Roche/454 platform has the advantage of sequencing longer reads of up to 600 bases and is able to generate more contiguous assemblies. If Ion Torrent is run in 100 bp mode, it has higher throughput than the Illumina platform and boasts a very short run time. Some drawbacks of these platforms may guide a user's choice in which platform to choose for their research. The Roche/454 has the lowest throughput of the three, thus if sequence coverage is of high priority another platform may be more appropriate. The Illumina and Ion Torrent platforms do not handle longer reads as well as Roche/454, though this is changing as the chemistries of these two platforms are improving (Loman et al. 2012).

The sequencing chemistries behind Roche/454 (454 Life Sciences 1996), Illumina (Illumina 2010)and Ion Torrent (ThermoFischer Scientific 2012)sequencing are all different forms of massively parallel sequencing by synthesis. Roche/454 technology amplifies target DNA by emulsion PCR (emPCR) before flowing in dNTPs of one type at a time (T, A, C, or G) in a predefined order into the reaction wells. The reaction mixture is such that when one or more nucleotides concordant with complementary bases are incorporated, a luciferase-catalyzed reaction emits light (giving it the name pyrosequencing). The amount of light emitted (and the signal detected from the light) is proportional to the number of added nucleotides. Ion Torrent sequencing uses similar procedures to Roche/454, but instead of a light signal, the number of protons released upon addition acts as the detection signal for the incorporation of nucleotides (Logares et al. 2012). Inherent sources of error in both technologies include homopolymer errors. Homopolymer errors refer to the incidences when several of the same base must be incorporated into the synthesizing DNA strand during a single flow of dNTPs within Roche/454 or Ion Torrent platforms, which sometimes results in intermittent under- or over-calls the strength of the signal (Balzer et al. 2011). Homopolymer-associated errors are produced at 1.5 and 0.38 errors per 100 bases on the Roche/454 and Ion Torrent platforms, respectively (Loman et al. 2012).

Illumina sequencing uses the "bridge amplification" method instead of emPCR, creating amplified clusters of the target DNA on a glass flow cell. Instead of a single-wavelength light signal as in Roche/454, Illumina sequencing uses four differently colored fluorophore labeled dNTPs. The sequencer images the fluorescently labeled terminator of an incorporated nucleotide. Following each nucleotide addition, the terminator cleavage allows for incorporation of the next nucleotide, ensuring that each nucleotide addition is a unique event. Illumina sequencing can also obtain both ends of a template molecule through "paired-end" sequencing (Logares et al. 2012).

Much of the sequencing platform comparison work to date has focused on performance of each technology in terms

of error rates of assembled genomes or metagenomes (Loman et al. 2012; Jünemann et al. 2013; Li et al. 2014; Solonenko et al. 2013; Frey et al. 2014; Bolotin et al. 2012), with few studies comparing amplified gene datasets (Fuellgrabe et al. 2015; Salipante et al. 2014) or relating performance to the resulting captured microbial diversity (Claesson et al. 2010; Luo et al. 2012). A metagenomic comparison of a complex freshwater microbial sample found that data produced on Illumina and Roche/454 platforms captured the same fraction of total diversity in the system, and with comparable abundances of each contig (Luo et al. 2012). A similar assessment of viral signature gene data has not been done, though comparison of Roche/454, Illumina, and Ion Torrent platforms on a metagenomic sample of ocean viruses (that required amplification steps to have enough DNA for the work) found that the sequencing platforms produced comparable datasets (Solonenko et al. 2013).

Table 1. Prices and capabilities of high-throughput sequencing on platforms Roche/454, Illumina MiSeq, and Ion Torrent as of 2012 (table sourced from Loman et al. 2012).

Platform	Cost per run	Min throughput (read length)	Run time	Cost/MB	Mb/h
454 GS FLX	$1,100	600 Mb (750–800 bases)	8 h	$31	4.4
Ion Torrent PGM (314 chip)	$225	10 Mb (100 bases)	3 h	$22.5	3.3
(316 chip)	$425	100 Mb (100 bases)	3 h	$4.25	33.3
(318 chip)	$625	1,000 Mb (100 bases)	3 h	$0.63	333.3
MiSeq	$750	1,500 Mb (2 × 150 bases)	27 h	$0.5	55.5

Table 2. Descriptions of the leading 2nd and 3rd generation HTS technologies in order of commercial availability (table modified from Glenn 2011).

Platform	Current company	Former company	Sequencing method	Amplification method	Claim to fame	Primary applications
454	Roche	454	Synthesis (pyrosequencing)	emPCR	First Next-Gen Sequencer, Long reads	**1***, **2**, 3*, 4, **7**, 8*
Illumina	Illumina	Solexa	Synthesis	BridgePCR	First short-read sequencer; current leader in advantages†	1*, **2**, 3*, **4**, **5**, **6**, **7**, **8**
SOLiD	Life Technologies	Applied Biosystems	Ligation	emPCR	Second short-read sequencer; low error rates	3*, **5**, **6**, **8**
Ion Torrent	Life Technologies	Ion Torrent	Synthesis (H^+ detection)	emPCR	First Post-light sequencer; first system <$100 000	**1**, **2**, **3**, 4, **8**
PacBio	Pacific Biosciences	N/A	Synthesis	None	First real-time single-molecule sequencing	1, 2, 3, **7**, **8**

Bold indicates applications that are most often used, economical or growing.
1 = *de novo* BACs, plastids, microbial genomes.
2 = transcriptome characterization.
3 = targeted re-sequencing.
4 = *de novo* plant and animal genomes.
5 = re-sequencing and transcript counting.
6 = mutation detection.
7 = metagenomics.
8 = other (ChIP-Seq, µRNA-Seq, Methyl-Seq, etc.;
*Pooling multiple samples with sequence tags (i.e. MIDs or indexes) is required for efficient use of this application
†Illumina currently leads in number and percentage of error-free reads, Illumina HiSeqs with v3 chemistry lead in reads per run, GB/run, and cost/GB.

1.7 Project aims

The main objective of this study was to investigate the hitherto unstudied diversity of ecologically significant viral groups during the dark period in the Arctic Ocean north of the Svalbard archipelago. The present study is a contribution to the RCN project entitled "MicroPolar (225956/E10)" headed by the University of Bergen. The MicroPolar project aims to characterize the microbial populations in the Arctic Ocean at all trophic levels, including the viral community, over the course of an annual cycle.

In this study, we aimed to answer the following questions:

- Is the diversity of the Arctic Ocean viral community distinct from that of other geographic locations sampled to date?
- Is viral community composition distinguishable between water masses or other physical/chemical environmental factors, and does it reflect host community diversity?
- Does use of different sequencing platforms produce comparable diversity capture for the same environmental viral assemblages?

To answer these questions, three viral marker genes were sequenced (*g23*, *phoH* and *MCP*) from eastern Arctic Ocean samples to capture a broad diversity within fingerprinted viral groups. Bioinformatic analyses were used to group sequences into OTUs to investigate the biodiversity of samples through measures of OTU richness, evenness and phylogenetic distance. These results were examined in the context of environmental data (physical parameters, flow cytometry counts, nutrients, and bacterial diversity). Additionally, amplified signature gene *g23* sourced from identical aliquots of viral concentrates sequenced on three HTS platforms (Roche/454, Illumina, and Ion Torrent) compared the effects of sequencing method on viral diversity capture.

2 Materials and methods

2.1 Sampling locations, collection, and preparation

2.1.1 Sampling

Samples were collected on a joint cruise initiated by the related CarbonBridge project aboard the R/V *Helmer Hansen* between January 6th and 14th, 2014. The *Helmer Hansen* transited north from Longyearbyen, Svalbard to the Arctic Ocean north of the archipelago where water samples were taken at sites spanning the Atlantic water inflow to the Arctic Ocean. Samples assessed in this thesis were sourced from two sample sites, known as B16 and B8 (Figure 6). Sites were located along the northernmost transect of the cruise (designated Transect B) at 81° 46.04' N 19° 06.59' E and 81° 25.52' N 17° 49.60' E, respectively.

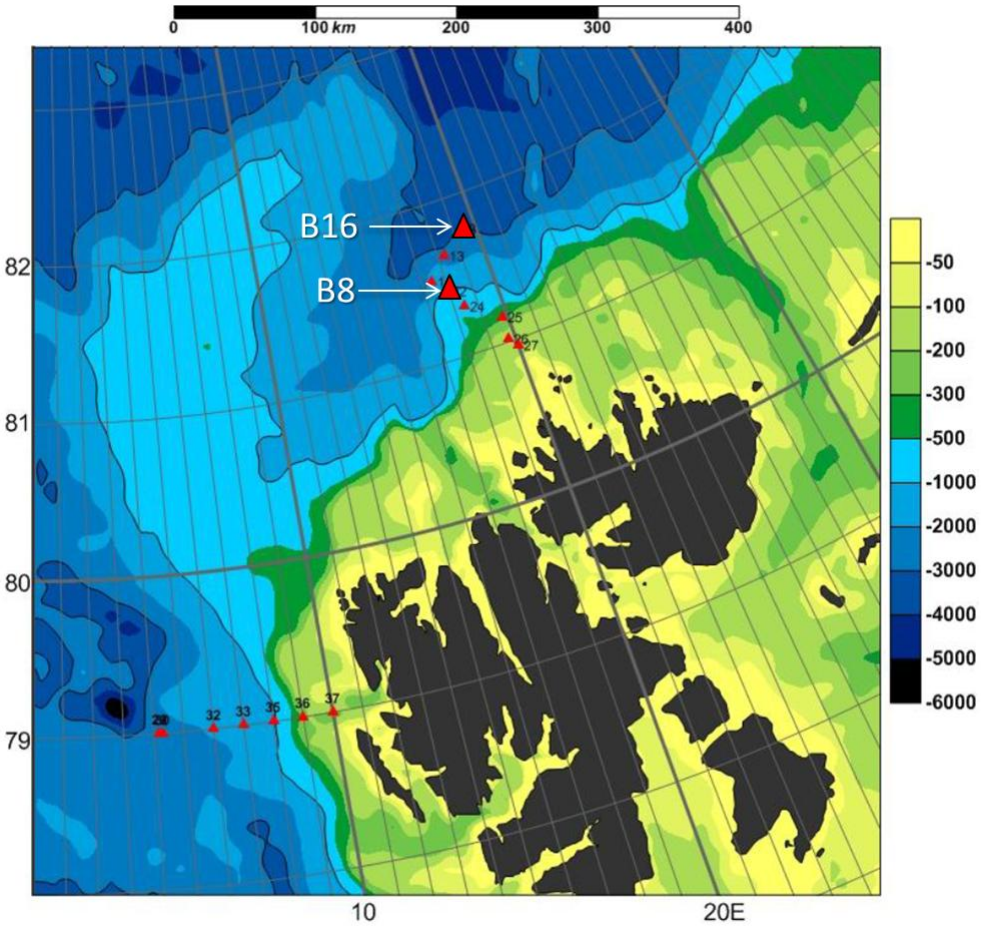


Figure 6. Map of cruise transects with seafloor topography. Stations B16 and B8 are noted in white along the northernmost transect. All stations are labeled with red triangles(sourced from CarbonBridge cruise report, January 2014).

Depth of collection and other physical and chemical parameters (salinity, temperature, density, fluorescence, and oxygen concentrations) were measured using a CTD mounted on a Niskin bottle rosette. The Niskin bottle rosette held 10 bottles of 5 L each to capture 50 L of water per cast. All bottles were fired at a single depth for each cast of the rosette. Water samples were taken at four depths at each site, in the order of 1000 m, 500 m, 20 m, and surface (Table 3). Colleagues at the University of Tromsø (CarbonBridge project) analyzed water samples from each site for

nutrient content. Oceanographer Arild Sundfjord of the Norwegian Polar Institute in Tromsø, Norway made determinations of water masses based on the physical parameters measured by the CTD (personal communication).

Table 3. Description of the locations (sample stations and depths) of the eight samples used in this thesis.

Station Name	Station Lat.	Station Long.	Depth at Station	Depth of collection
B16	81° 46.04' N	19° 06.59' E	3,165 m	surface 20 m 500 m 1000 m
B8	81° 25.52' N	17° 49.60' E	943 m	surface 20 m 500 m 1000 m

2.1.2 Viral sample filtration

The 50 L seawater samples from each depth were prefiltered through a 0.45 µm filter for phytoplankton capture. Due to possible low yield with increasing depth, the 50 L 0.45 µm filtered volume from 1000 m was combined with a 50 L volume prefiltered for bacteria at 1000 m using a 0.2 µm Sterivex filter (SVGPL10RC, Merk Millipore, Germany).

To capture the viral fraction, prefiltered water was concentrated by tangential flow filtration (Quickstand benchtop system, GE Healthcare Life Sciences) through a Polysulfone membrane with 100,000 NMWC pore size (UFP-100-C-4X2MA, GE Healthcare Life Sciences) connected to a peristaltic pump (Figure 7). The water samples were concentrated to a target volume of 50 mL for each sample. The filter was sanitized and rinsed according to the manufacturer's instructions before processing a new sample. Concentrated samples were divided into forty 1 mL aliquots, transferred into 2 mL cryotubes, and immediately frozen in liquid nitrogen.

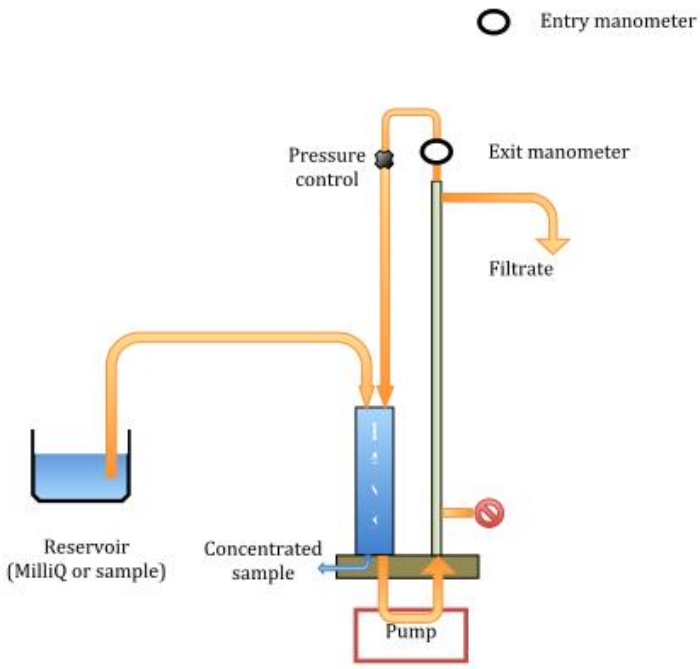


Figure 7. Tangential flow filtration system used to concentrate viral particles from seawater. The sample flows through the filter, resulting in retention of particles above the nanoscale filter matrix size (figure courtesy of Sven Le Moine Bauer).

2.2 Environmental parameter visualization

Data collected by sensors were visualized using the software program Ocean Data View (ODV). Density measurements extrapolated out from each CTD cast created a full view of water density along Transect B.

2.3 Flow cytometry

One 1 mL aliquot of each sample was preserved in glutaraldehyde (final concentration 0.5%) overnight in an unheated shipping container on deck then frozen in liquid nitrogen until required for viral particle enumeration by flow cytometry. Fixed samples were thawed to RT then serially diluted using 0.2 µm filtered 1xTE (10 mM Tris HCl with 1 mM EDTA, pH 8.0) to make 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4} dilutions. The sample dilutions (final volumes 500 µL each) were then stained using 5 µL of 100 x working solution of SYBR Green I (10,000x in DMSO, Sigma Aldrich) for 10 min in the dark at 80°C.

Flow cytometry (FACSCalibur, Becton Dickinson) was carried out using a medium flow rate for 60 s. The flow cytometer setup is described in Marie et al. (1999). Viral particle groupings were discerned using plots of particles clustering based on side scatter versus the green DNA stain fluorescence using the CellQuest Pro software (BD Biosciences).

2.4 DNA extraction

A rapid virus DNA extraction method (Pagarete et al. 2013a) was implemented to isolate genomic DNA template material from viral concentrates for amplification of *g23* and *phoH*. For viral particle lysis, samples were thawed on a heat block at 90°C for 2 min, split into two 500 µL volumes, then further subjected to 2 cycles of alternately freezing at -20°C then thawing at 90°C for 2 min. Lysed samples were treated with 20 µL sterile filtered 0.5M EDTA of pH 8.0 and 5 µL of freshly made Proteinase K (10 mg mL^{-1}) before incubation for 10 min at 55°C. Samples were incubated on ice and 25 µL of 10% SDS added then gently inverted before further incubation at 55°C for 1 h. The lysate was cleaned using the Zymo DNA Cleanup and Concentration Kit (Zymo Genetics) (see Appendix A.2), then eluted in 20 µL of sterile filtered TE buffer heated to 70°C (10 mM Tris HCl, 0.1mM EDTA, pH 8.0) and stored at 4°C until use in PCR.

In preparation for amplification of *MCP*, template DNA material was taken directly from viral concentrates that were subjected to 2 cycles of alternately thawing at 90°C for 2 min then freezing at -20°C as per Larsen et al. 2008.

2.5 Amplifications for sequencing

The general workflow used to create all amplicon libraries for Roche/454, Illumina and Ion Torrent samples is outlined in Figure 8. The overall process of extraction, initial amplification without adapters, cleaning, secondary amplification with adapters, cleaning, and pooling was similar for all samples. The specific steps for each preparation varied in the extraction method, primers used, number of cycles, cleaning kits used, and adapter sequences as detailed below.

2.5.1 Amplification of *g23* and *phoH* for Roche/454 sequencing

The PCR reactions and preparation protocols for *g23* and *phoH* were modified from those specified in Filée et al. 2005 and Goldsmith et al. 2011, respectively. Six replicate 50 µL reaction mixtures for PCR were prepared per sample. The reaction mixtures are summarized in Table 4. Each 50 µL reaction contained 1U Ex Taq polymerase (Takara Bio), 1X Ex Taq buffer (Takara Bio), a 0.25 mM concentration of each primer (see Table 7), 0.2 mM deoxynucleoside triphosphates, 0.06% bovine serum albumin, 3% DMSO, and 1-10ng template DNA. Sterile milliQ H_2O was added to volume. The reaction conditions for *g23* and *phoH* amplification are summarized in Table 5.

Table 4. The reaction mixture components for amplification of the genes *g23* and *phoH*, totaling a reaction volume of 50 µL

PCR reaction mixture for g23/phoH amplification	
Taq polymerase	1U
Ex Taq buffer	1X
f-primer	0.25 M
r-primer	0.25 M
dNTPs	0.2 mM
BSA	0.06% of vol
DMSO	3% of vol
template DNA	1-10 ng
sterile milliQ H_2O	to volume
total	50 µL

Replicate PCR reactions from each respective sample were combined and cleaned using the Zymo DNA Cleanup and Concentration Kit (Appendix A.2). Amplification and cleaning steps were checked on 1% (w/v) agar electrophoresis gels (Appendix A.4). Successful amplicons were stored at 4°C until use.

Table 5. A summary of all PCR reactions performed for amplification of *g23*, *phoH* and *MCP*.

Gene Amplified	*Reaction Step*	*Time*	*Temperature °C*	*# Cycles PCR 1*	*# Cycles PCR 2*
g23	initial denaturation	5 min	95	1	1
	denaturation	45 s	95	20	10
	annealing	45 s	50		
	extension	1 min	72		
	final extension	7 min	72	1	1
	end	∞	4		
phoH	initial denaturation	5 min	95	1	1
	denaturation	45 s	95	20	15
	annealing	45 s	53		
	extension	1 min	72		
	final extension	7 min	72	1	1
	end	∞	4		
MCP	initial denaturation	15 min	95	1	1
	denaturation	30 s	95	20	N/A
	annealing	30 s	60 →50		
	extension	30 s	72		
	denaturation	30 s	95	35	25
	annealing	30 s	45		
	extension	30 s	72		
	final extension	7 min	72	1	1
	end	∞	4		

Two replicate reactions were made for each sample in a secondary amplification step. The reverse primer (see Table 7) contained an added adapter sequence for one-directional Roche/454 sequencing (Lib-L-adapter A sequence: CCTATCCCCTGTGTGCCTTGGCAGTCTCAG). The forward primer contained barcode sequences (also known as multiplex identifiers, or MIDs) which were unique to each of the eight samples (see Table 8). Each 50 µL reaction contained 1U Ex Taq polymerase (Takara Bio), 1X Ex Taq buffer (Takara Bio), a 0.1 mM concentration each of the barcoded forward primer and the adapter-containing reverse primer, 0.2 mM deoxynucleoside triphosphates, 0.06 % bovine serum albumin, 3% dimethyl sulfoxide, and a 10 µL volume of product of the first amplification as template DNA. Sterile milliQ H_2O was added to volume. The reaction conditions for each gene are summarized in Table 5. Success of amplifications was checked on 1% (w/v) agar electrophoresis gels (Appendix A.4). Samples were frozen at -20°C until ready to prepare for sequencing.

2.5.2 Amplification of *MCP*

The PCR preparation and thermocycling for *MCP* were identical to procedures published in Larsen et al. 2008. PCR template DNA material was taken directly from viral concentrates that were subjected to 2 cycles of alternately thawing at 90°C for 2 minutes then freezing to -20°C.

The reaction mixture for amplifications of *MCP* (Table 6) contained 1-10ng of template DNA, 10 µL of HotStar master mix (Qiagen, Germany), a 0.5M concentration of each primer (Table 7), and was adjusted final volume (20 µL) with sterile milliQ H_2O. The PCR reaction conditions (Table 5) consisted of a 15 min hot start before a 20 cycle touchdown PCR with a decrease of 0.5°C in the annealing step per cycle followed by a PCR with a consistent temperature in the annealing step for an additional 35 cycles.

Products were cleaned using an Agencourt AMPure XP magnetic bead kit (Beckman Coulter, USA) and a 6-Tube magnetic separation rack (New England Biolabs, USA)(Appendix A.3). Amplification and cleaning steps were checked on 1% (w/v) agar electrophoresis gels (Appendix A.4). Samples were frozen at -20°C until further use.

Table 6. The reaction mixture components for amplification of the *MCP* gene of the *Phycodnaviridae* are combined to make a total reaction volume of 20 µL.

PCR reaction mixture for MCP amplification

HotStar master mix	10 µL
f-primer	0.5M
r-primer	0.5M
template DNA	1-10 ng
sterile milliQ H_2O	to volume
total	20 µL

Primers used in the second amplification step contain added adapter sequences identical to those described above in section 2.5.1. Six replicate reactions were made for each sample using the product of the first amplification as template DNA. Identical reaction mixtures to those for the initial amplification were made, as summarized in Table 6. The reaction conditions are summarized in Table 5. Products were cleaned using an Agencourt AMPure XP magnetic bead kit and a 6-Tube magnetic separation rack (Appendix A.3). Amplification and cleaning steps were checked on 1% (w/v) agar electrophoresis gels (Appendix A.4). Samples were frozen at -20°C until ready to prepare for sequencing.

Table 7. Sequences in the 5 to 3 direction of degenerate primers used to amplify each marker gene.

Gene	*Primers*
g23	f: 5'- GATATTTGNGGNGTTCAGCCNATGA -3'
	r: 5'- CGCGGTTGAATTTCCAGCATGATTTC -3'
phoH	f: 5'- TGCRGGWACAGGTAARACAT -3'
	r: 5'- TCRCCRCAGAAAAYMATTTT -3'
MCP	f: 5'- GGYGGYCARCGYATTGA -3'
	r: 5'- TGIARYTGYTCRAYIAGGTA -3'

Table 8. Barcodes added during the second amplification step (PCR 2) to distinguish each sample within a pooled product sent for sequencing.

Sample	*Barcodes*
B16, surface	ATCAGACACG
B16, 20 m	ATATCGCGAG
B16, 500 m	CGTGTCTCTA
B16, 1000 m	CTCGCGTGTC
B8, surface	ACGAGTGCGT
B8, 20 m	ACGCTCGACA
B8, 500 m	AGACGCACTC
B8, 1000 m	AGCACTGTAG

2.5.3 DNA measurements

Reactions from each barcoded amplicon sample were thawed before combining replicate reactions and cleaning as described in section 2.5.1. DNA concentrations were measured for each cleaned sample using a Qubit 2.0 fluorometer (Invitrogen) and a Qubit HS dsDNA assay kit (Invitrogen) according to the manufacturer's instructions.

All eight samples were subsequently combined into a *g23* amplicon pool and a *phoH* amplicon pool so that nanogram amounts from each sample were contributed equally. The sample pool was measured again to confirm the DNA concentration on the Qubit fluorometer. Pooled *g23* samples were sent to the Norwegian Sequencing

Centre (NSC) in Oslo, Norway (http://www.sequencing.uio.no/) for Roche/454 pyrosequencing. The pooled *phoH* samples remained frozen at -20°C until ready to send for sequencing in combination with *MCP*.

All *phoH* samples were combined into a mixture containing equal nanogram amounts of all samples, then cleaned using the Zymo DNA Cleanup and Concentration kit (Appendix A.2) as described above. An identical process was performed for *MCP* samples. Sample pools were measured again for confirmation of DNA concentration on the Qubit fluorometer. The pools of *MCP* and *phoH* were then combined into one volume containing equal nanogram amounts of each pool. The *MCP*/*phoH* pooled samples were sent to Microsynth AG, Zurich, Switzerland (http://www.microsynth.ch/) for pyrosequencing on the Roche/454 GS FLX Titanium platform.

2.6 Illumina sequencing of *g23*

A preparation identical to the *g23* amplification for 454 sequencing (see sections 2.5.1 and 2.5.3) was sent for Illumina sequencing to the NSC in Oslo, Norway. The *g23* primer set used in this study was constructed for use on the Roche/454 platform, therefore additional adapters were added to the fragments at the NSC in order for the sample to be run using an Illumina MiSeq sequencer. Additionally, sample material for Illumina sequencing was spiked with a concentration of control DNA from the virus PhiX (Illumina, FC-110-3001) by a technician at the NRC before sequencing.

2.7 Ion Torrent sequencing of *g23*

A preparation similar to the *g23* amplification for Roche/454 sequencing (see sections 2.5.1 and 2.5.3) was sent for Ion Torrent sequencing at the University of Bergen, Norway. A larger number of replicates (6 from each sample) was required in the second amplification step (PCR 2 in Table 5) to get maximum yield of DNA as one microgram of total DNA was required by the Ion Torrent sequencer before a technician could prepare the library for sequencing. One other change was made for Ion Torrent sequencing preparation of *g23* from that described in section 2.4.1; the Lib-L adapter sequence located on the reverse primer used in Roche/454 and Illumina sequencing was replaced by the appropriate adapter sequence used in Ion Torrent sequencing (Adapter B sequence: CCTCTCTATGGGCAGTCGGTGAT) in order to obtain comparable results to Roche/454. The Ion Torrent sequencing facility at the University of Bergen, Norway performed the final sample dilution and the sequencing reaction.

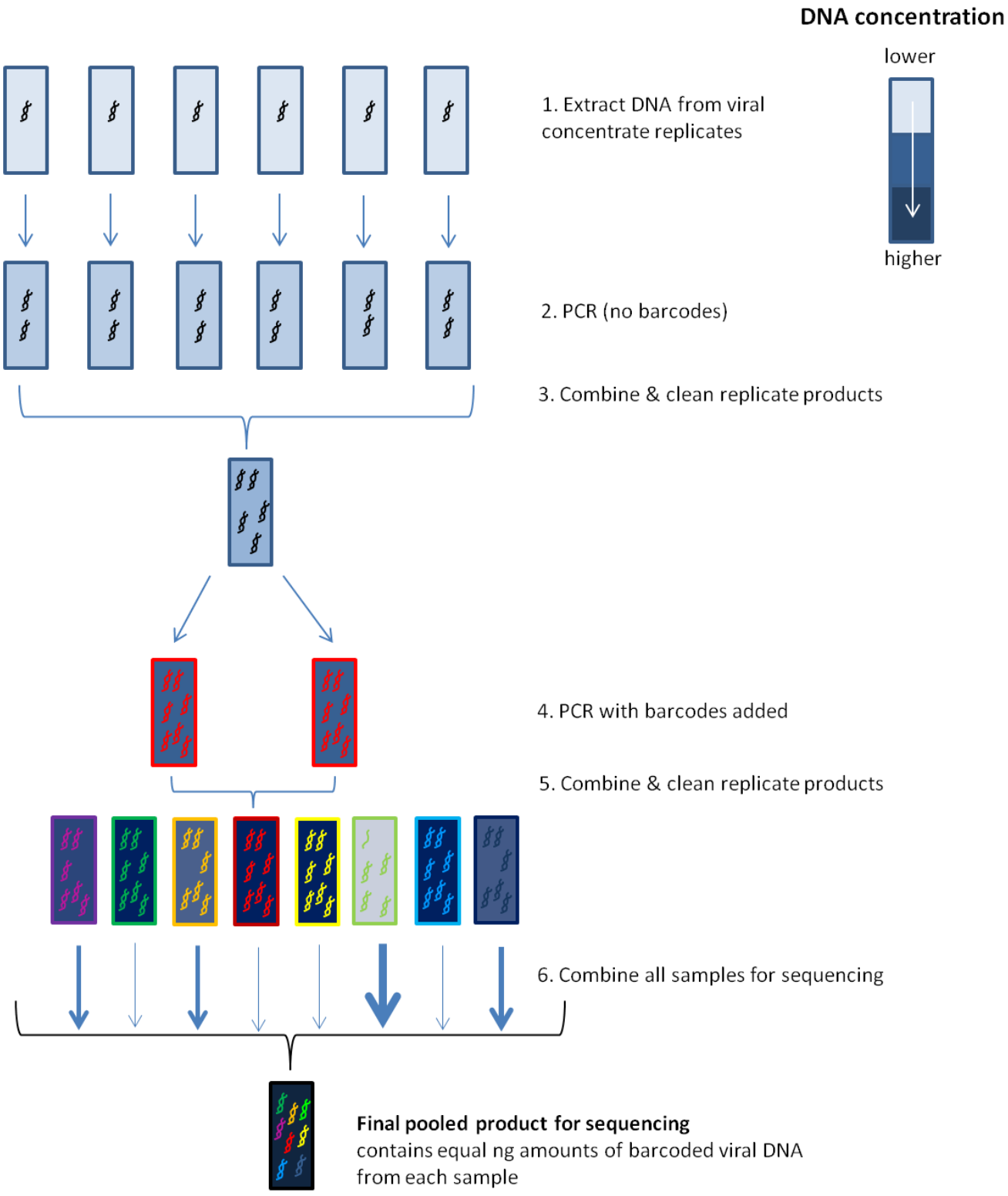


Figure 8. Amplification workflow for all samples in the study from extraction of viral concentrates isolated from seawater (step 1) to a final pooling of all eight samples with equal ng amounts of DNA (step 6).

2.8 Post-sequencing processing

A majority of the post-sequencing processing was accomplished on a 2-core personal computer with 8 GB memory using the specialty Ubuntu build BioLinux (Field et al. 2006). An annotated text file containing the pipeline providing all commands used to process and analyze all sequence data is provided in Appendix A.5. Below is an outline of the processing.

2.8.1 Illumina specific post-processing

The first two steps of post-sequencing processing were used only for the *g23* Illumina sequencing dataset because of the specific protocol necessary to post-process the data after the return of raw reads. Control phiX sequences were removed from Illumina data before downstream analysis, which was accomplished using the multi-functional script bbduk.sh within BBMap (Bushnell 2014). The script used the reference database of phiX genes from bowtie2-2.2.4 and a kmer length cutoff of 31 to scan the Illumina reads for contaminants with homology to phiX of at least 31 bp in length, allowing for one mismatch. This kmer length choice and allowed mismatches were identical to those given in an example by the script developer (Bushnell 2014) (See script commands in Appendix A.5). Subsequently, Illumina adapters were removed from the sequences using bbduk.sh. Specified parameters in this script looked for shorter kmers at the read ends down to a length of 12, and allowed one mismatch. Following this step, bbduk.sh was used to merge paired Illumina reads.

2.8.2 Ion Torrent specific post-processing

The output file delivered from the Ion Torrent sequencing facility was eight BAM format files (one for each sample) which had been demultiplexed by Ion Torrent software at the sequencing facility. BamTools (Barnett et al. 2011) script bam-to-fastq was used to convert files between BAM and FASTQ formats. The degenerate primer sequence was removed using the BBMap script bbduk.sh. This script was also used to quality trim sequences from both the left and right to exclude base calls with PHRED scores under 27. This step proved to create a systematic error and only eliminated primer sequence for some of the sequences because of the lack of the barcodes in the Ion Torrent sequences. Centre for Geobiology colleague Dr. Håkon Dahle added placeholder barcodes and successfully eliminated the primer sequence from all reads, while simultaneously trimming sequences based on expected error rather than PHRED score, according to the UPARSE processing pipeline (Edgar 2011). FASTA files resulting from the UPARSE primer and quality trimming steps were concatenated into a single FASTA file for further downstream analysis (2.8.2.2 onwards) the remainder of which was identical to all other dataset processing pipelines.

2.8.3 Post-sequencing processing of all datasets

The following protocol was generally used to process all datasets from Illumina, Roche/454, and Ion Torrent sequencing used in this study. Several sections are specific only to the three platform comparison of *g23* datasets, however these have been indicated as such with asterisks.

2.8.3.1 Sequence data quality checking and trimming

An initial check of the quality of pyrosequencing reads was visualized using the program FastQC (Andrews 2010). The raw SFF formatted files were converted to QUAL, FASTQ and FASTA files using the QIIME script process_sff.py. The BBMap script bbduk.sh was used to quality trim sequences from both the left and right to exclude base calls with PHRED scores under 27. Quality trimmed sequences were subsequently demultiplexed according to a manually-generated mapping file containing barcodes and the forward primer sequence using the QIIME script split_libraries.py. Demultiplexing removed bases corresponding to adapter A at the beginning of the read along with the barcode sequence and Linker/Primer sequence, leaving only the desired sequence of biological origin (Figure 9) for further analyses. The length of sequencing reads is not expected to extend into the reverse primer, and therefore no step was included for its removal.

Figure 9. Layout of a 454 Roche sequence amplicon. This scheme describes the layout used for target gene amplifications.

2.8.3.2 OTU picking and elimination of sequencing artifacts

Chimera checking and OTU picking were performed via the QIIME script pick_OTUs.py using 97% similarity for OTU picking and utilizing de novo chimera detection provided by USEARCH (R. C. Edgar 2010). The script generated an OTU table BIOM file in HDF5 format. The OTU table was then split by sample using QIIME script filter_samples_from_otu_table.py. Each sample-specific OTU table was filtered for singletons using QIIME script filter_otus_from_otu_table.py. The sample-specific OTU tables were merged back into a single OTU table using QIIME script merge_otu_tables.py.

Before OTU picking steps for the comparison of Roche/454, Illumina, and Ion Torrent *g23* datasets, all three quality filtered datasets were sub sampled to even sequencing depth for improved comparability and globally trimmed to 200 bp per sequence to minimize gaps in the downstream sequence alignment. OTU picking was performed using 90% sequence similarity rather than 97% to account for the global trimming step, which results in a requirement of 20 bp differences rather than 6 differences to distinguish OTUs from one another. The platform comparison did not filter singletons.

2.8.3.3 Diversity analyses and heatmaps

OTU tables were converted from .BIOM format to tab-delimited .TXT file format to be used in the R environment version 3.2.0 (R Development Core Team 2011) for diversity analyses and were transposed to meet requirements for

scripts within the R package Vegan version 2.2-1 (Oksanen et al. 2013). Within-sample diversity measures, known as alpha diversity, were calculated using QIIME script alpha_diversity.py, with Chao1 (Chao 1984) species richness estimates, PD (phylogenetic distance) (Walker and Faith 1994), and observed OTUs as the user-chosen outputs. Alpha rarefaction curves were generated in R using the Vegan script rarecurve to create plots of estimates of species richness at different sampling depths within each sample. Pielou's evenness (Pielou 1977) values were calculated by:

$$E=D/\log(S)$$

Where D is the Shannon diversity index value of the sample, and S is the number of observed OTUs for the sample.

The diversity between sample datasets, known as beta diversity, was analyzed using QIIME script jackknifed_beta_diversity.py. This script resampled the data at even sequencing depth per sample at the minimum sequencing depth of all samples in each dataset, then used Unifrac distances (Lozupone and Knight 2005) to create weighted UPGMA cluster dendrograms. Bootstrapping of the UPGMA dendrograms was accomplished using the QIIME script make_bootstrapped_tree.py, which created a PDF file of the dendrogram with colored branches indicating level of bootstrap support. The script also created principal coordinate analysis (PCoA) plots based on Unifrac distance matrices. Statistical comparison of categories using ANOSIM was accomplished using the script compare_categories.py.

Heatmaps were generated describing normalized relative abundances of the most abundant OTUs representing ≥1% of the total sequence datasets using QIIME script make_otu_heatmap.py.

2.8.3.4 Phylogenetic analyses

A list of representative sequences from each OTU was generated using the QIIME script ref_seqs.py. Only OTUs that represented ≥ 1% of the total sequence datasets were included in the final phylogenetic identity analyses. Each OTU representative sequence was queried against NCBI BLAST (Altschul et al. 1997). The top hits were examined for percent sequence identity and query coverage. Sequence identities ≥ 70% were reported, and listed as no BLAST hit if no identity ≥ 70% was found or if in addition the query coverage was ≤ 50%. Only the *g23* Roche/454 dataset contained enough information to create a phylogeny, therefore the following steps were only performed for that dataset. Dereplicated BLAST hits and the representative sequences from all OTUs were aligned using the alignment tool Muscle (Edgar 2004) through the QIIME script align_seqs.py. Phylogenies were made using this alignment file as input to the QIIME script make_phylogeny.py (which uses FastTree (Price et al. 2009) by default for tree construction). Trees were visualized and annotated using the tree editing software FigTree (Rambaut 2008).

3 Results

3.1 Environmental parameters

Visualization of transect B physical data revealed a cold, fresher water layer at the surface at station B16 (Figure 10, far left vertical profile; Figure 11, top panel) not seen at station B8 (Figure 10, right vertical profile; Figure 11, bottom panel). The density range of the water column (low at surface, high at depth) was between 27.6 and 28.1 σ_T at station B16 and between 27.9 and 28.1 σ_T at station B8.

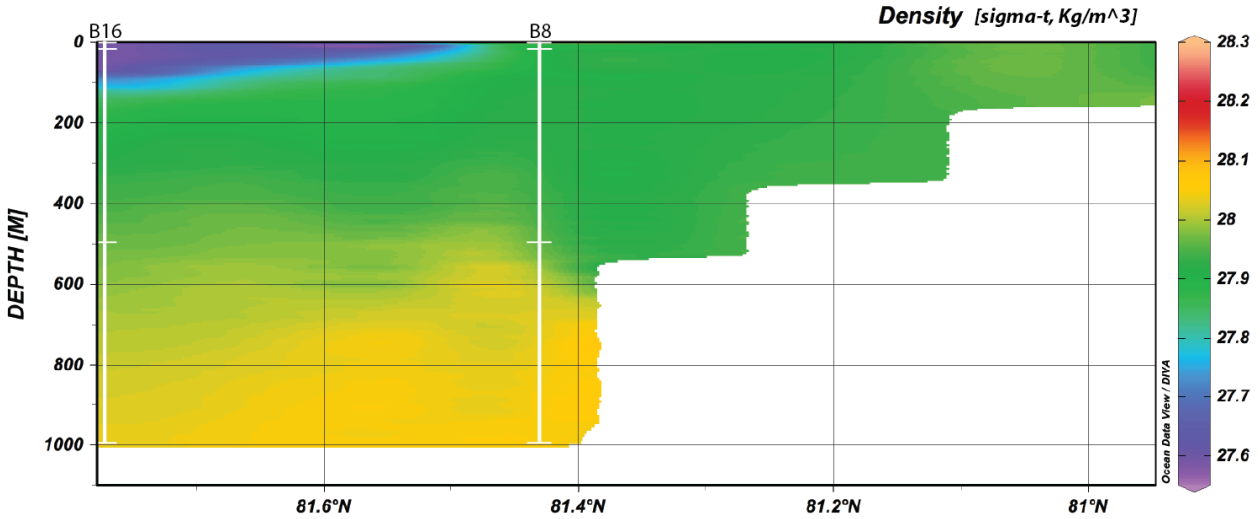
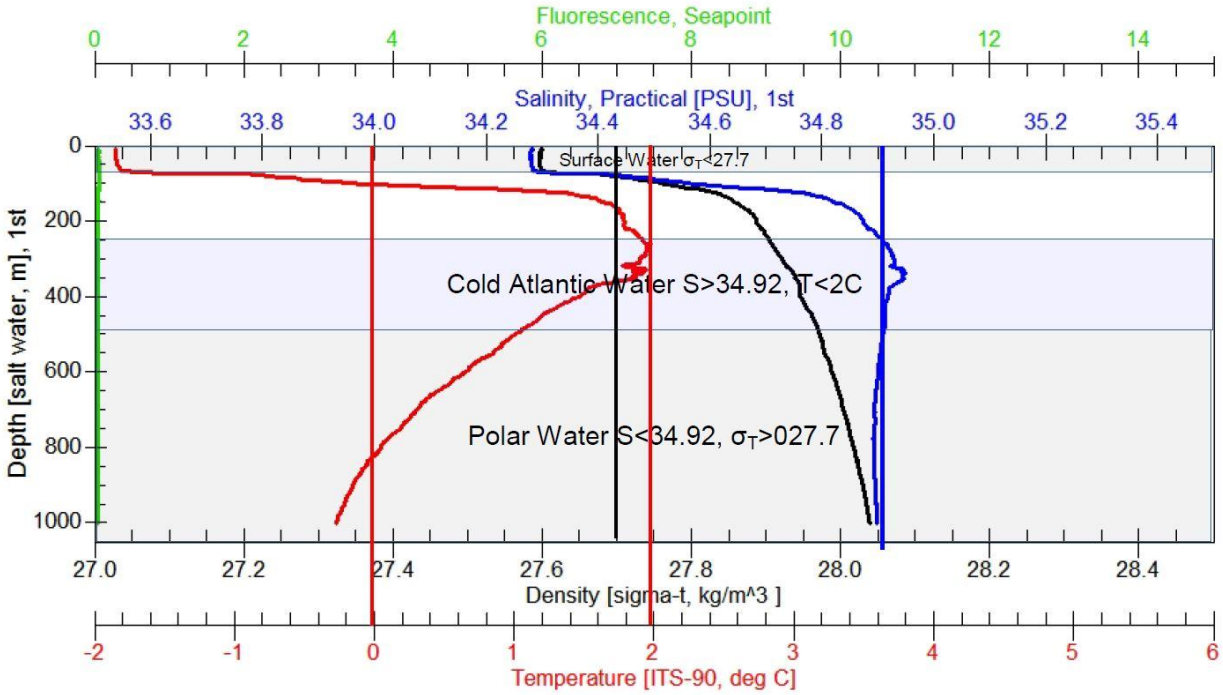


Figure 10. Density of the water column extrapolated from data collected at CTD sample sites along transect B from northernmost (left) to southernmost (right). White vertical lines represent B16 and B8, and short horizontal white lines represent sampling depths.

Each genomic sample was determined as sourced from a mass of water as described in Table 9 (and as graphed in Figure 11). Genomic samples at station B16 are sourced from Surface Water (B16.surface and B16.20m) and Polar Water (B16.500m and B16.1000m), while samples at station B8 are sourced from Atlantic Water above (B8.surface and B8.20m) and below the pycnocline (B8.500m) or from Arctic Intermediate Water (B8.1000m).

B16



B8

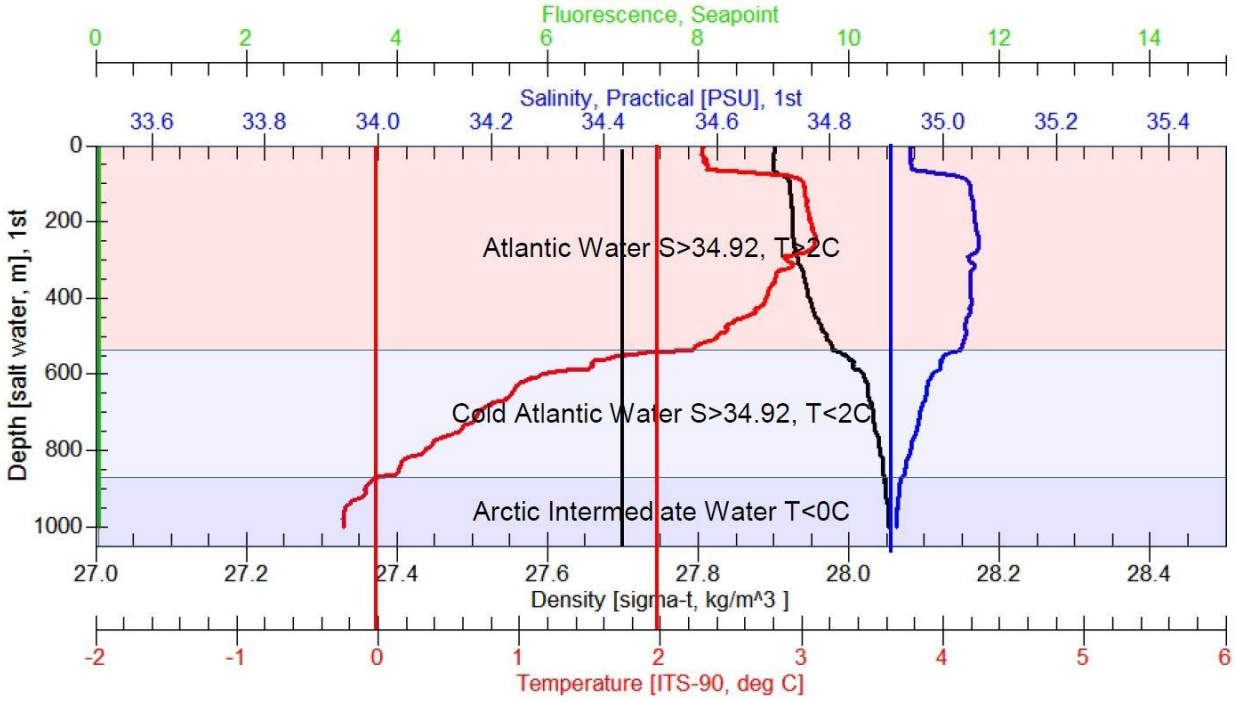


Figure 11. Physical characterization of depth profiles at B16 (top chart) and B8 (bottom chart) collected using the CTD and auxiliary sensors, annotated with water mass assignments.

Table 9. Sample names and the associated water mass within which the source water was taken as determined by physical parameters.

Genomic Sample ID	Assigned Water Mass
B16.surface B16.20m	Surface Water
B16.500m B16.1000m	Polar Water
B8.surface B8.20m	Atlantic Water (above pycnocline)
B8.500m	Atlantic Water (below pycnocline)
B8.1000m	Arctic Intermediate Water

Nutrient data (Figure 12) indicate that station B8 surface water was higher in nitrate/nitrite (10 µM) and phosphate (0.69 µM) than station B16 surface water (5 µM and 0.43 µM, respectively). Nitrate/nitrite and phosphate increased with increasing depth at both sites, though 20 m depth at station B8 had slightly lower values (1-2 µM difference) for both nutrients.

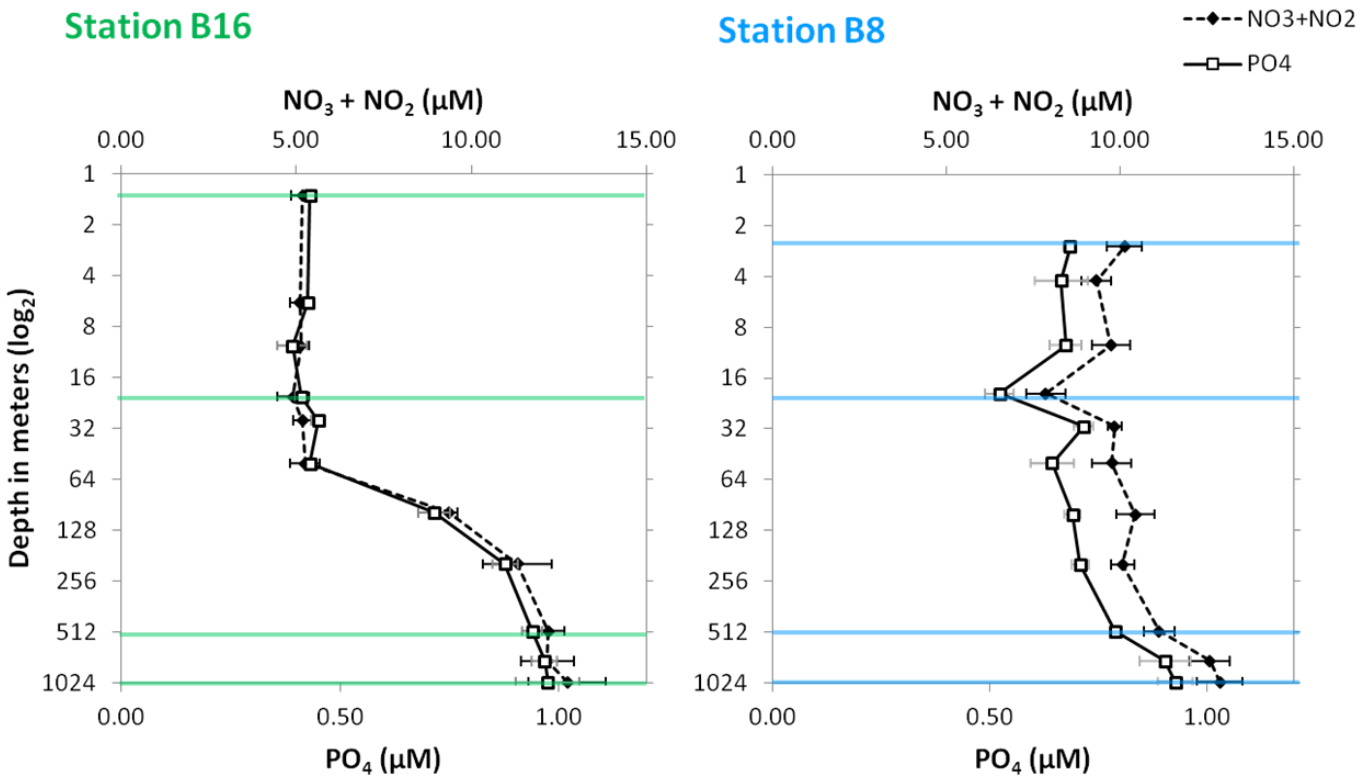


Figure 12. Concentrations of nitrate/nitrite (dotted black lines) and phosphate (solid black lines) profiles at stations B16 and B8 with the y-axis as depth in meters presented in log_2 scale. Standard deviations are represented by error bars. Colored horizontal lines refer to genomic sampling depths. Colleagues at the University of Tromsø analyzed nutrient data.

Flow cytometry counts of microbial particles (Figure 13) reveal virus, bacteria, and eukaryotes were generally present at both stations at relative abundances on the order of 1 x 10^6 particles mL^{-1}, 1 x 10^5 cells mL^{-1}, and 1 x 10^2 cells mL^{-1}, respectively. All particle counts generally decreased with increasing depth. The virus to bacteria ratio

(VBR) increased with depth at both sites, ranging from 14 to 37 at station B16 and from 12 to 18 at station B8 (not shown). Viral and bacterial counts correlated tightly at both stations throughout the water column. At station B16 maximum bacteria (1.61 x 10^5 mL^{-1}) and maximum virus (2.22 x 10^6 mL^{-1}) enumerations were measured at 30 m depth while at station B8 maximum bacteria counts (1.36 x 10^5 mL^{-1}) and maximum viral counts (1.7 x 10^6 mL^{-1}) were measured at 200 m depth. The range between the minimum and maximum viral counts was 1.22 x 10^6 mL^{-1} at station B16 and 7.5 x 10^5 mL^{-1} at station B8. The range between the minimum and maximum bacterial cell counts was 6.72 x 10^4 mL^{-1} at station B16 and 7.73 x 10^4 mL^{-1} at station B8. The range between minimum and maximum eukaryotic cell enumerations was 1.48 x 10^2 mL^{-1} at station B16 and 8.6 x 10^1 mL^{-1} at station B8.

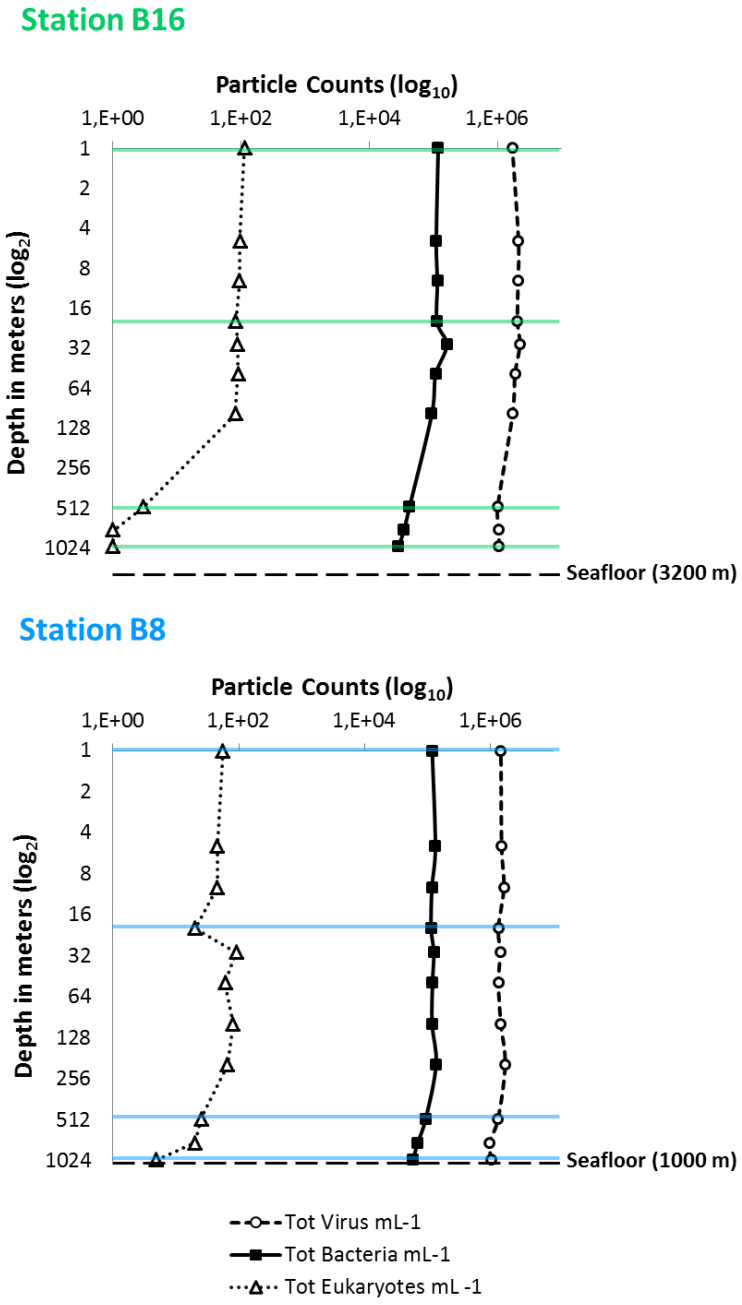


Figure 13. Flow cytometry enumerations at stations B16 and B8 of microbial cells and virus particles per milliliter of seawater. Colored horizontal lines refer to depths of tagged sequencing samples at either station B16 (turquoise) or B8 (blue). Particle counts courtesy of Maria Lund Paulsen, UiB.

3.2 Verification of amplification

Amplification of the viral signature genes following the second PCR and cleaning steps resulted in purified products of varying concentrations as summarized in Table 10. Agarose gels confirmed that barcoded product sizes ranged from 300 – 600 bp sequence length (Appendix B.1, Figures B-1 to B-5). Total nanogram amounts in DNA mixtures sent for pyrosequencing of each marker gene were 160 ng of *g23* (20 ng from each sample), 240 ng of *phoH* (30 ng from each sample), and 222 ng of *MCP* (~ 28 ng from each sample). The total ng amounts in *g23* samples sent for Illumina MiSeq and Ion Torrent sequencing were 240 ng (30 ng from each sample) and 4800 ng (600 ng from each sample), respectively. It should be noted, however, that technicians made dilutions of these libraries before sequencing was accomplished, therefore the sent DNA material does not reflect the amount of sample ultimately used in any of the sequencing reactions.

Table 10. DNA concentrations of each of 8 samples of *g23*, *phoH* and *MCP* amplicons as measured on the Qubit 2.0 Fluorometer. Products were later pooled in equal ng amounts before sending to the Norwegian Sequencing Centre for Roche/454 sequencing.

Sample ID	Concentration of *g23* PCR product (ng/µL)	Concentration of *phoH* PCR product (ng/µL)	Concentration of *MCP* PCR product (ng/µL)
B16.surface	15	10.4	9.83
B16.20m	3.39	10.9	9.04
B16.500m	12.3	3.24	5.85
B16.1000m	6.9	3.26	8.32
B8.surface	18.7	6.29	19.6
B8.20m	15.8	31.2	13.2
B8.500m	14.6	6.35	13.4
B8.1000m	19.5	3.25	11.1

3.3 Sequencing run diagnostics

Generated FastQC reports (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) of raw pyrosequencing data received from NSC (Appendix B.2.1) and Microsynth (Appendix B.2.2) show the *g23* Roche/454 run resulted in a total of 1,171,251 reads with an average length of 450 bp and an average per read PHRED score of 38 (Figure B-7). The combined *phoH* and *MCP* sequencing run resulted in a total of 232,305 reads with an average length of 411 bp (Figure B-11) and an average per read PHRED score of 38 (Figure B-10).

FASTQC reports of Illumina paired-end data received from NSC show *g23* sequencing runs 1 and 2 both resulted in a total of ~22 million reads each with length 301 bp and an average per read PHRED score of 37. The numbers of reads were significantly reduced to 5,872,874 sequences following contamination cleaning and paired read merging steps (Appendix B.2.3), after which sequences ranged from 35 – 590 bp in length (Figure B-14).

Ion Torrent data received from the UiB Ion Torrent PGM facility resulted in a total of 2,688,165 reads (Appendix

B.2.4) with a majority of sequences having 420 – 439 bp (Figure B-17) and a majority of per read PHRED scores at 28 (Figure B-16), with a second, less abundant peak at 18.

3.4 Characterization of OTU tables from Roche/454 data

De novo picking of OTUs at 97% similarity level and exclusion of singletons resulted in 1,696 OTUs of *g23*, 140 OTUs of *phoH*, and 233 OTUs of *MCP*. Of these OTUs, 396 (23%) were shared between all 8 samples in the *g23* dataset (Appendix B.6, Figure B-20), 33 (24%) OTUs were shared between all samples in the *phoH* dataset, and none of the OTUs were shared between all samples in the *MCP* dataset. Among the 10 OTUs (4%) shared by either 6 to 7 of the *MCP* samples, all OTUs except one did not include sequences from sample B8.1000m. Almost 52% of all OTUs in the *MCP* dataset are unique to any one sample. Of those, 59% contain from 3-10 sequences and 30% contain 20-850 sequences.

Rank abundance plots (Appendix B.5, Figures B-18 and B-19) show 31% of *g23* OTU assignments are extremely rare (arbitrarily defined as 10 or fewer reads). Including those extremely rare OTUs, 70% of the OTUs in the dataset belong to OTUs containing 100 reads or less. *g23* OTUs containing 100 – 1,000 reads comprised 20.5% of the total OTUs, and OTUs containing 1,000 – 65,000 reads comprised 9.5% of the dataset. OTUs with 100 reads or fewer comprised 74% of the *phoH* dataset and 77.2% of the *MCP* dataset. OTUs with 100 – 1,000 reads comprised 18.6% of the *phoH* dataset and 17.6% of the *MCP* dataset. OTUs with 1,000 – 10,000 reads comprised 7.4% of the *phoH* dataset and 5.2% of the *MCP* dataset.

3.5 Diversity analyses

Pyrosequenced samples of *g23*, *phoH* and *MCP* were sequenced to between 86 -96%, 85-97%, and 77-95.8% of the Chao1 species richness estimates, respectively (Tables 11, 12, and 13). The indication that sequencing had not reached full saturation for most samples is reflected in alpha rarefaction curves (Figure 14). Richness was especially elevated in station B16 samples sourced from the cold Surface Water layer: the highest OTU richness in the *g23* dataset was from B16.20m (1126 OTUs). This sample has a 3-fold higher sequencing depth than the other *g23* samples. B16.1000m represents the second highest OTU richness value in the *g23* dataset (1025 OTUs).

For the *MCP* dataset, evenness (Appendix B.3, Table B-2) is generally greater at 1,000 m depth than at 20 m or surface. The range of Pielou's evenness index values for the rarefied *g23* (0.729-0.82) and *phoH* (0.589 – 0.665) samples are narrower than for the *MCP* samples (0.339 – 0.771).

Table 11. Alpha diversity indices of pyrosequenced *g23* samples. Metrics included are the Chao1 species richness estimate, the total phylogenetic distance (PD), and observed numbers of OTUs.

	Sample Name	Chao1	PD Whole Tree	Observed OTUs
g23	B16.surface	1056	193	949
	B16.20m	1297	252	1233
	B16.500m	1007	179	921
	B16.1000m	1126	213	1091
	B8.surface	1044	178	900
	B8.20m	1098	186	945
	B8.500m	1184	203	1050
	B8.1000m	1085	186	972

Table 12. Alpha diversity indices of pyrosequenced *phoH* samples. Metrics included are the Chao1 species richness estimate, the total phylogenetic distance (PD), and observed numbers of OTUs.

	Sample Name	Chao1	PD Whole Tree	Observed OTUs
phoH	B16.surface	117	13	106
	B16.20m	100	12	97
	B16.500m	97	11	86
	B16.1000m	76	8	70
	B8.surface	119	10	102
	B8.20m	95	12	85
	B8.500m	98	10	88
	B8.1000m	71	8	66

Table 13. Alpha diversity indices of pyrosequenced *MCP* samples. Metrics included are the Chao1 species richness estimate, the total phylogenetic distance (PD), and observed numbers of OTUs.

	Sample Name	Chao1	PD Whole Tree	Observed OTUs
MCP	B16.surface	97	24	90
	B16.20m	107	30	99
	B16.500m	80	22	69
	B16.1000m	62	20	48
	B8.surface	60	14	49
	B8.20m	79	22	71
	B8.500m	89	21	69
	B8.1000m	48	19	46

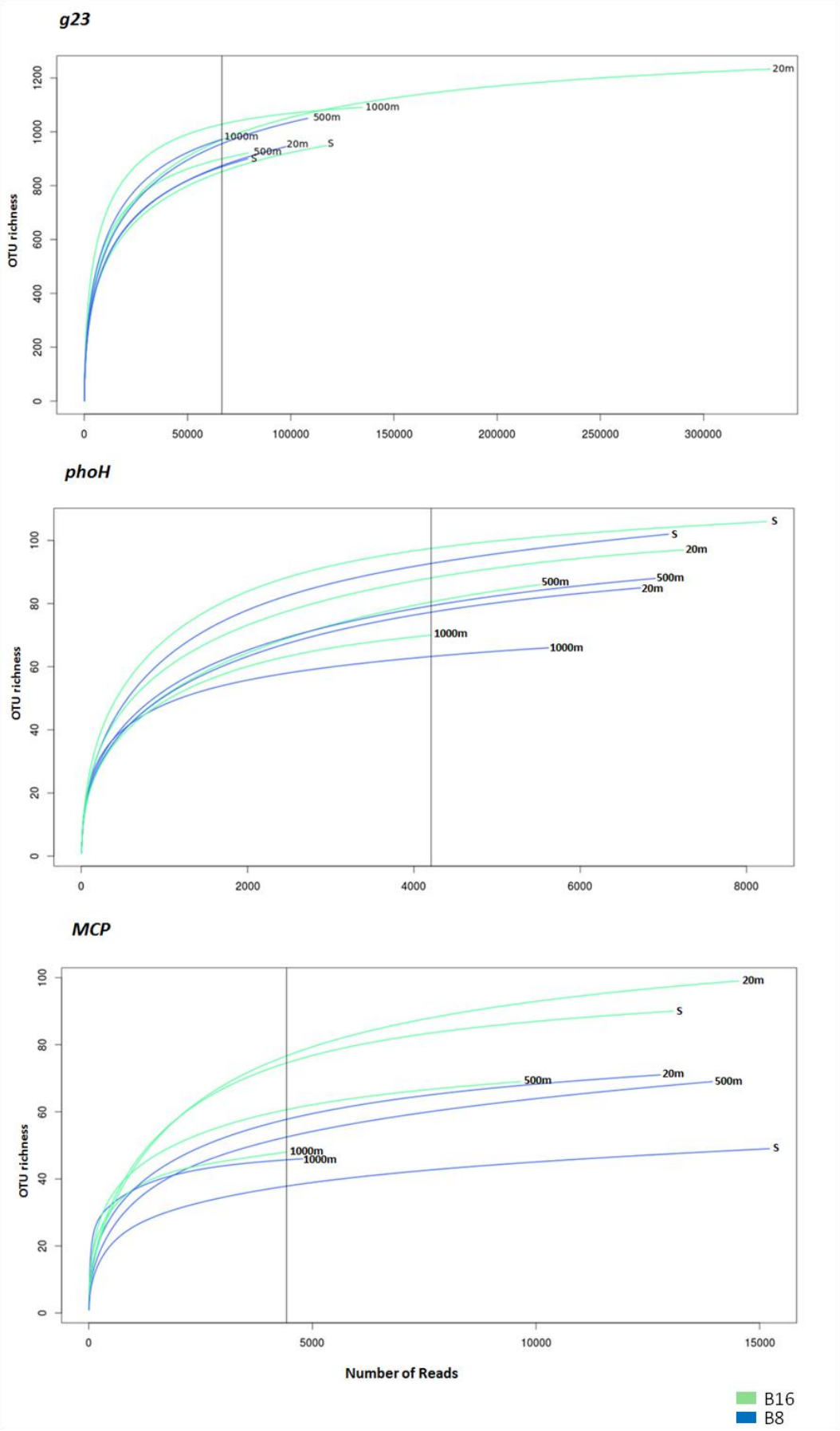


Figure 14. Alpha rarefaction curves of *g23, phoH and MCP* samples on the Roche/454 platform. Black vertical lines refer to maximum rarefaction depth sampled for each dataset. Turquoise lines represent B16 samples and blue lines represent B8 samples.

Jackknife support of the weighted UPGMA dendrograms based on the three pyrosequenced genes (Figures 15, 16, and 17) resulted in 75-100% confidence for all nodes in all three datasets. Within the *g23* and *phoH* datasets, weighted UPGMA dendrograms show a cluster of all station B8 samples, a cluster comprised of B16.surface and B16.20m samples, and a cluster comprised of B16.500m and B16.1000m samples. Within the *g23* station B8 cluster, samples at B8.surface and B8.20m are the most closely associated. In the *phoH* B8 cluster, samples B8.500m and B8.1000m are the most closely associated. Samples B16.500m and B16.1000m samples clustered equidistantly from all other samples within the *g23* dataset. In the *phoH* dataset, B16.500m occupies the furthest distance and B16.1000m is secondarily distant from all other samples. The dendrogram of *MCP* samples B8.surface, B8.20m, and B8.500m form a cluster which is most similar to the cluster comprised of samples B16.surface and B16.20m. B16.500m is more similar to each of the aforementioned clusters than to either of the samples sourced from 1,000 m depth, which cluster together as most dissimilar to all other *MCP* samples.

Categorical comparison of samples by water mass using ANOSIM on weighted Unifrac distance matrices (Appendix B.7) revealed that the grouping of samples by water mass is statistically significant for the *g23* dataset (Table B-8) with an R value of 0.913 (closer to +1 indicates stronger grouping) and a p-value of 0.003. ANOSIM on the *phoH* and *MCP* datasets (Tables B-9 and B-10) did not result in statistically significant grouping by water mass, with R values of 0.096 and 0.304, and p-values of 0.313 and 0.144, respectively.

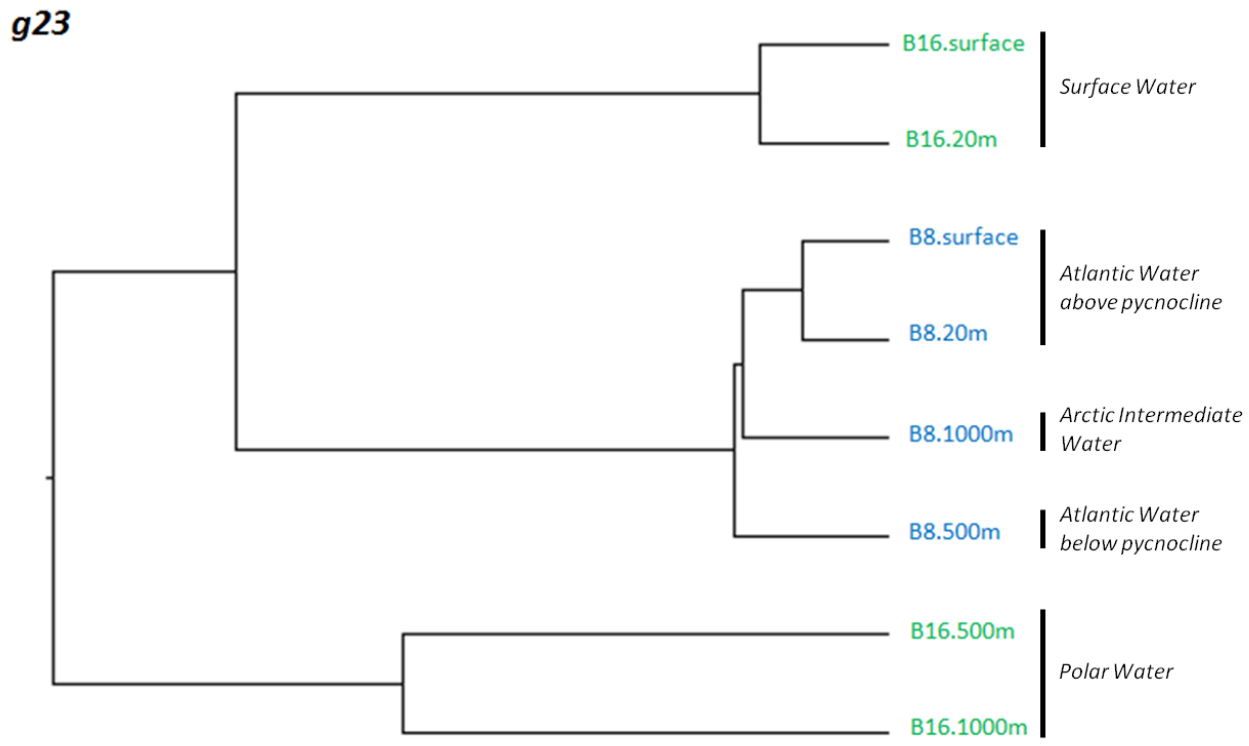


Figure 15. Weighted UPGMA dendrogram of *g23* Roche/454 samples with even sequence depth of all samples (64K sequences) and singleton OTUs removed. Blue labels are B16 samples and turquoise labels are B8 samples. Water masses of origin are noted in italics.

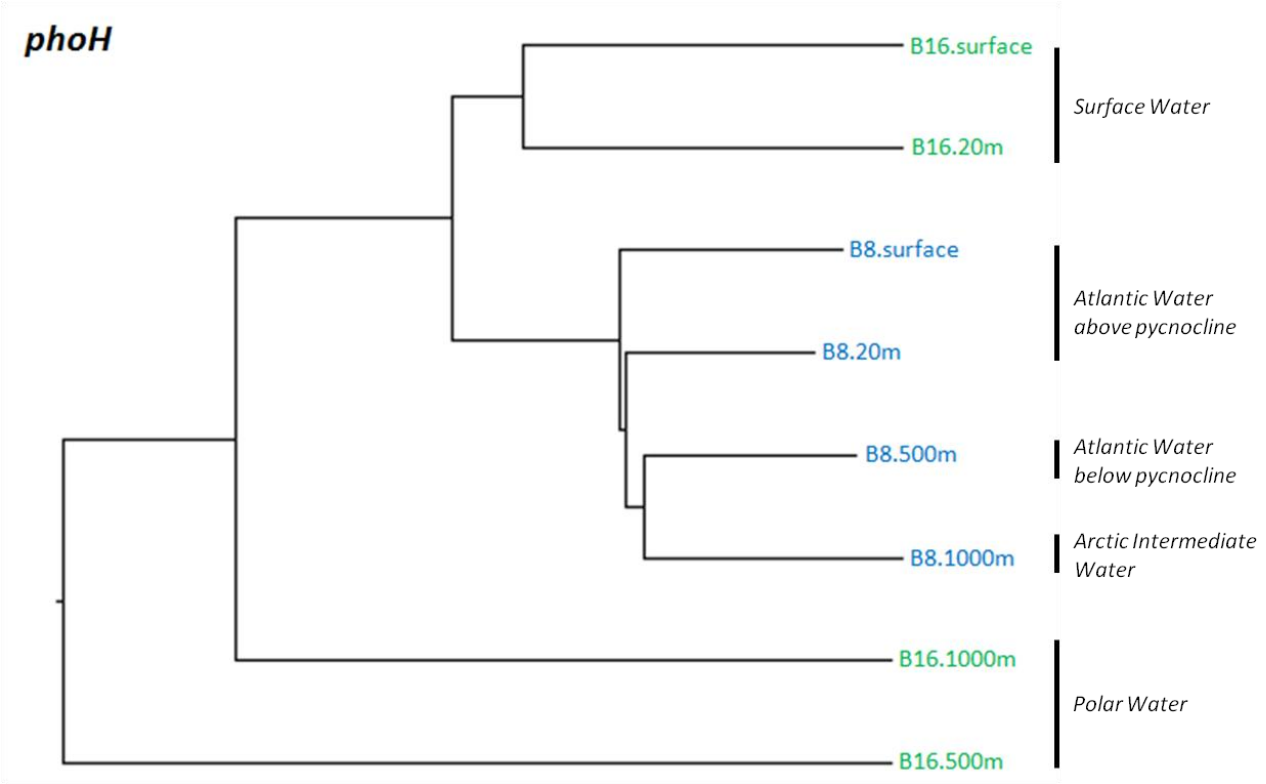


Figure 16. Weighted UPGMA dendrogram of *phoH* Roche/454 samples with even sequence depth of all samples (4K sequences) and singleton OTUs removed. Blue labels are B16 samples and turquoise labels are B8 samples.

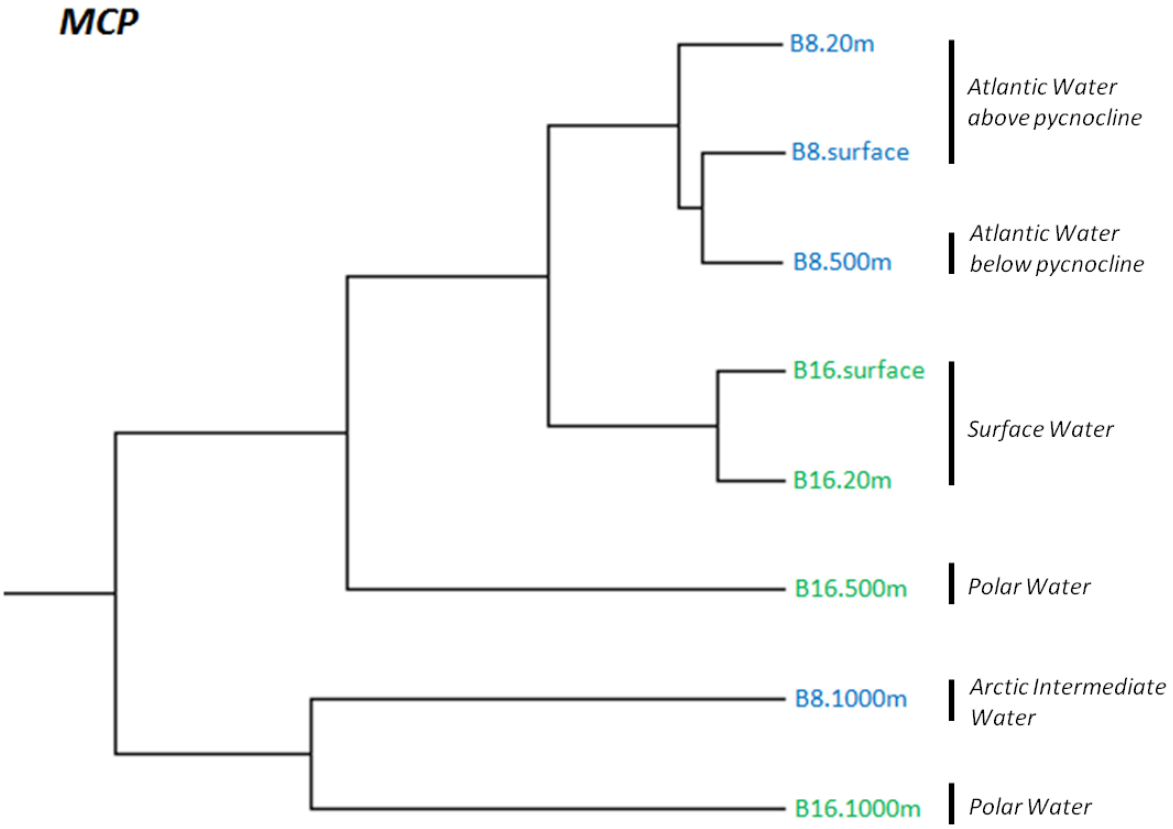


Figure 17. Weighted UPGMA dendrogram of *MCP* Roche/454 samples with even sequence depth of all samples (4K sequences) and singleton OTUs removed. Blue labels are B16 samples and turquoise labels are B8 samples.

3.6 OTU heatmaps and homologous sequences in NCBI BLAST

The heatmap of the twenty-six OTUs representing ≥ 1% of the *g23* dataset (Figure 18) shows that samples which cluster according to the weighted UPGMA have varying proportions of certain OTUs. The *g23* B16.surface/B16.20m cluster shares high proportions (a range of 19 – 68% abundance) of sequence reads belonging to seven OTUs that have closest sequence similarity to samples deposited in the NCBI BLAST database sourced from either the Arctic Ocean or marine waters in Norway. All station B8 samples contained high abundances (11 – 27%) of sequences belonging to twelve OTUs with similarity to BLAST entries from one of the following: the Arctic (1 OTU), Norway (3 OTUs), coastal California during winter (3 OTUs) or summer (1 OTU), Gulf of Mexico (1 OTU), a Chinese lake (1 OTU, low homology) or no BLAST hit (2 OTUs). Sample B16.500m contained 73% of sequence abundance from OTU 23, which had no BLAST hit. Sample B16.1000m contained a similarly high abundance (71%) of sequences from OTU 21, which also had no BLAST hit. Both B16.500m and B16.1000m contained high abundance of sequences belonging to four OTUs with BLAST hits from SPOT in winter (2 OTUs) and summer (1 OTU), and Norway (1 OTU).

The heatmap of the fourteen OTUs representing ≥ 1% of the *phoH* dataset (Figure 19) shows fewer clear relationships associating to the weighted UPGMA clusters than the *g23* dataset. The UPGMA cluster comprising all station B8 samples contains sequences for all the OTUs representing ≥ 1% of the dataset except OTUs 11 and 5. OTU 11 was only found in high proportions within ice-influenced Surface Water samples, with a 42% abundance of reads from B16.surface and 37% from B16.20m. OTU 5 reads are disproportionately abundant in sample B16.500m, which contains 76% of reads for this OTU. Within the B8 cluster, samples sourced from Atlantic Water (B8.surface, B8.20m, and B8.500m) contain the largest proportions of sequences belonging to OTU 7 (24%, 21%, and 16% abundance, respectively). OTU 7 reads are present in lower abundances in all other samples in the dataset, except in samples sources from Polar Water (B16.500m and B16.1000m), in which it is absent. The closest BLAST hit matching the entire length of the representative sequence for OTU 7 (with only 85% identity) is a *phoH* sample from Raunefjorden, Norway. The highest percentage of reads clustering to OTU 9 (23% abundance) are sourced from sample B16.1000m, and are also present in all other samples in the dataset (from 6 - 14% abundance of OTU 9).

The heatmap of the twelve OTUs representing ≥ 1% of the *MCP* dataset (Figure 20) shows the differing percent abundances of the OTUs reflecting the UPGMA clusters. UPGMA clustered samples B8.surface, B8.20m and B8.500m share high proportions (a range from 14 – 29% abundance) of reads from OTUs 0, 1, 3, and 10. The B16.surface/20m UPGMA cluster also had high proportions of the same four OTUs (13 – 24% abundance) but also has large proportions of OTU 2 (33% abundance in B16.20m, 40% in B16.surface) and OTU 9 (69% abundance in B16.20m, 39% in B16.surface). The B8.1000m/B16.1000m cluster shares a large proportion of OTU 7 (45% abundance in B8.1000m, 55% in B16.1000m), which is not present in any of the other samples. The B16.500m sample has a high proportion of sequences from OTU 4 (86% abundance), which is only present in one other sample (B8.1000m, with 11%

abundance). Sample B8.500m and B8.1000m are the only two samples containing OTU 5 sequences, and the proportion of these is very high for B8.1000m (71% abundance). Five of the OTUs included in the *MCP* heatmap had only poor homology to previously cultured algal viral species from Norway.OTU 9 had 72% sequence similarity to a *Pyramimonas orientalis* virus (76% query cover), OTUs 1 and 6 had 86% and 83% similarity to *Prymnesium kappa* virus, respectively (19% and 36% query cover), and OTUs 2 and 4 were 84 and 86% sequence similarity to *Haptolina ericina* virus, respectively (40% and 35% query cover). No OTU matched to *Micromonas pusilla* virus sequences in the BLAST database.

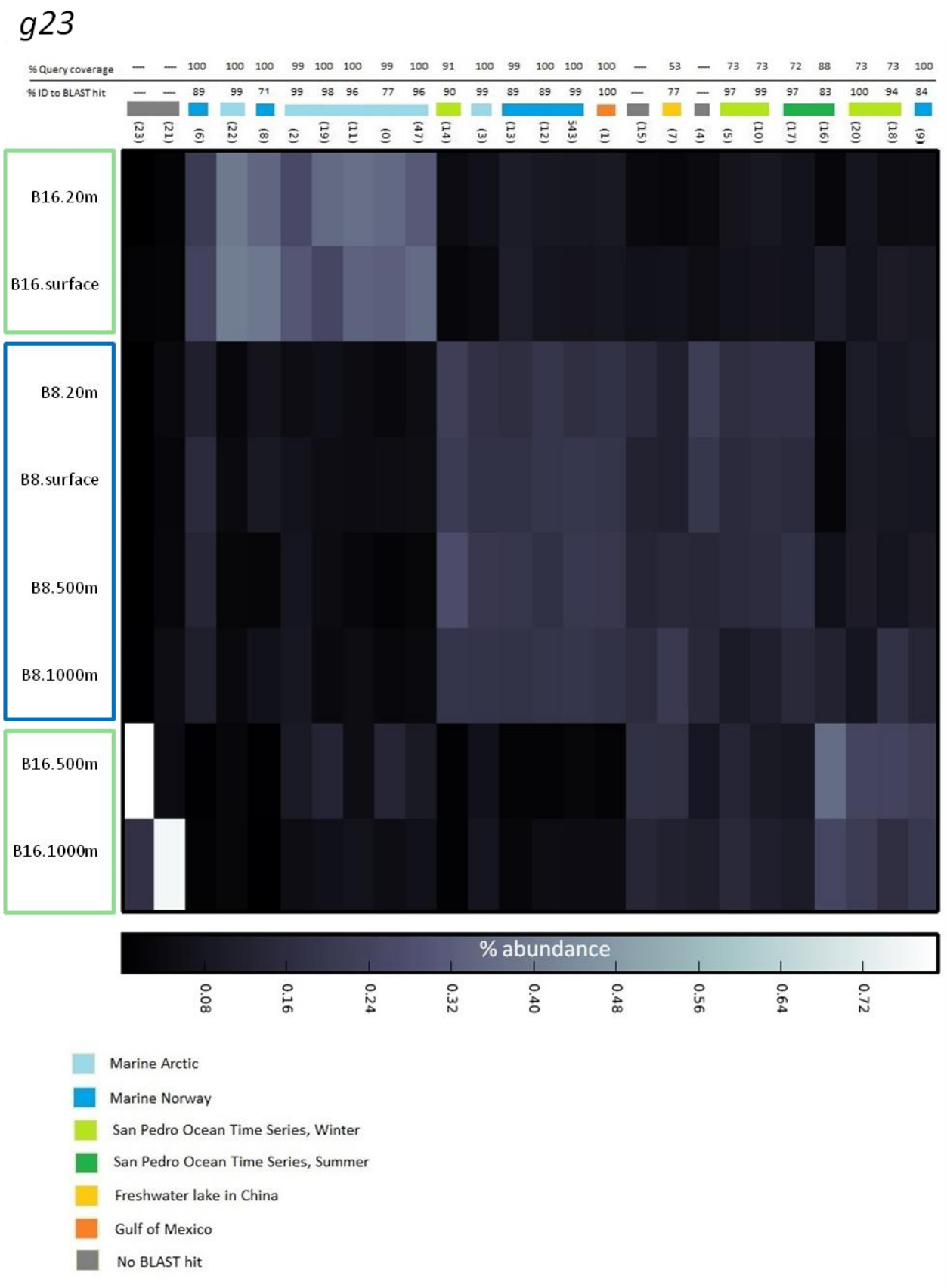


Figure 18. OTU heatmap of ≥1% OTUs in the *g23* Roche/454 dataset rarefied to even depth (28K sequences each). Open turquoise and blue rectangles indicate weighted UPGMA clusters of samples using the entire dataset. OTU names are numbers along the top of the heatmap in parentheses. Colors associated with OTUs indicate the origin of the top BLAST hit to the OTU representative sequence.

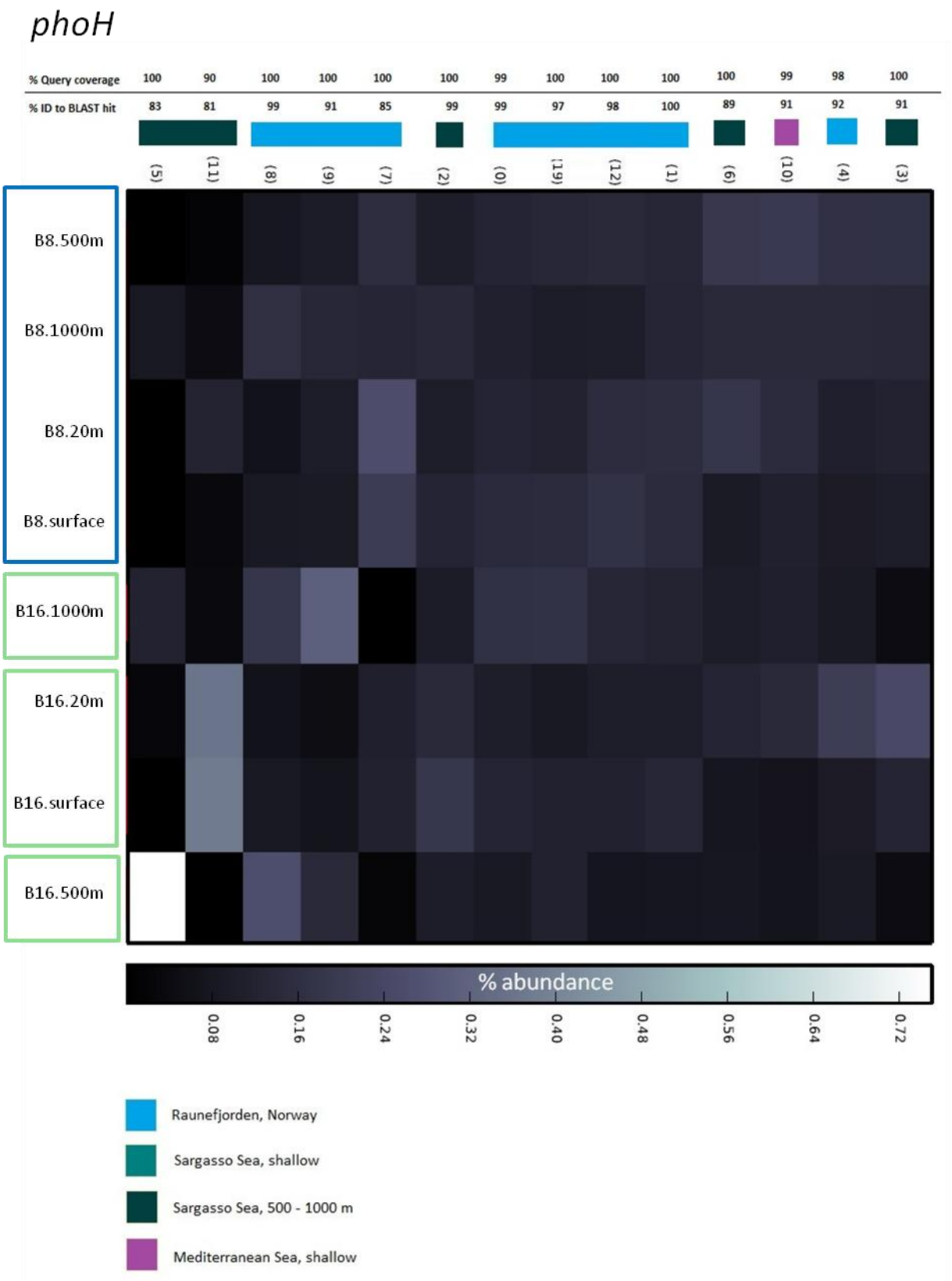


Figure 19. OTU heatmap of ≥1% OTUs in the *phoH* Roche/454 dataset rarefied to even depth (3. 7K sequences each). Open turquoise and blue rectangles indicate weighted UPGMA clusters of samples using the entire dataset. OTU names are numbers along the top of the heatmap in parentheses. Colors associated with OTUs indicate the origin of the top BLAST hit to the OTU representative sequence.

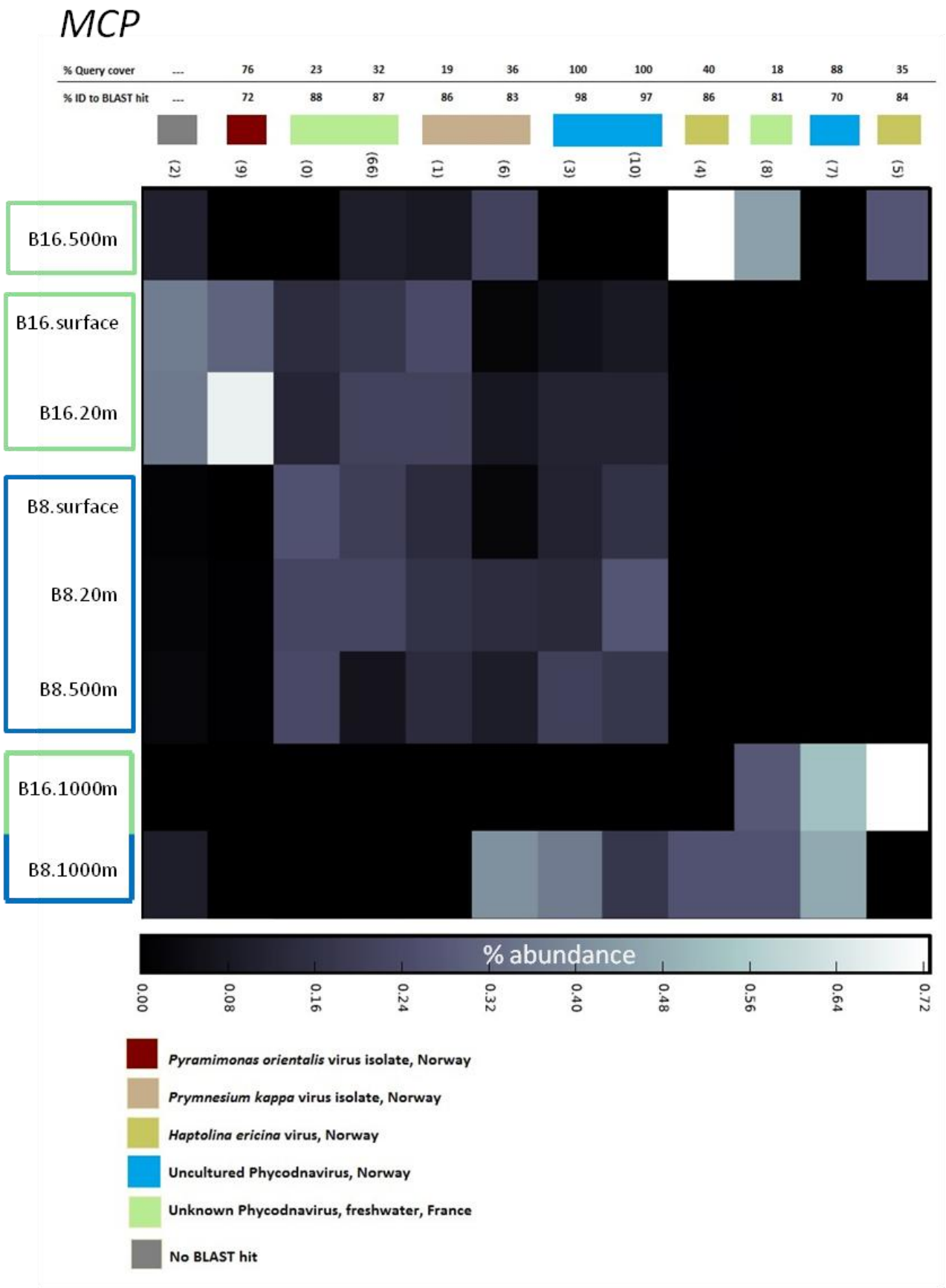


Figure 20. OTU heatmap of ≥1% OTUs in the *MCP* Roche/454 dataset rarefied to even depth (1.8K sequences each). Open turquoise and blue rectangles indicate weighted UPGMA clusters of samples using the entire dataset. OTU names are numbers along the top of the heatmap in parentheses. Colors associated with OTUs indicate the origin of the top BLAST hit to the OTU representative sequence.

3.7 *g23* OTU diversity

The phylogeny produced using the Muscle alignment of *g23* OTU representative sequences and the top BLAST hits of the 369 OTUs present in all eight samples reveals four large clusters (Figure 21). Clusters 2 and 4 contain few (3 and 4, respectively) of the OTUs representing ≥ 1% of the dataset. These clusters also contain BLAST hits to sequences originating from a variety of environments around the world. Cluster 3 contains 11 out of the 33 OTUs representing ≥ 1% of the *g23* dataset mainly contains BLAST sequences from Norway, the Arctic, or San Pedro Ocean Time Series station during winter (defined here as November to February). Subcluster 1a contains the most abundant OTU in the dataset, and only aligned BLAST sequences sourced from Norwegian, Arctic, or SPOT station samples (winter only) group to this subcluster. Subcluster 1a is adjacent to other branches within Cluster 1 that contain BLAST hit sequences that are less geographically limited.

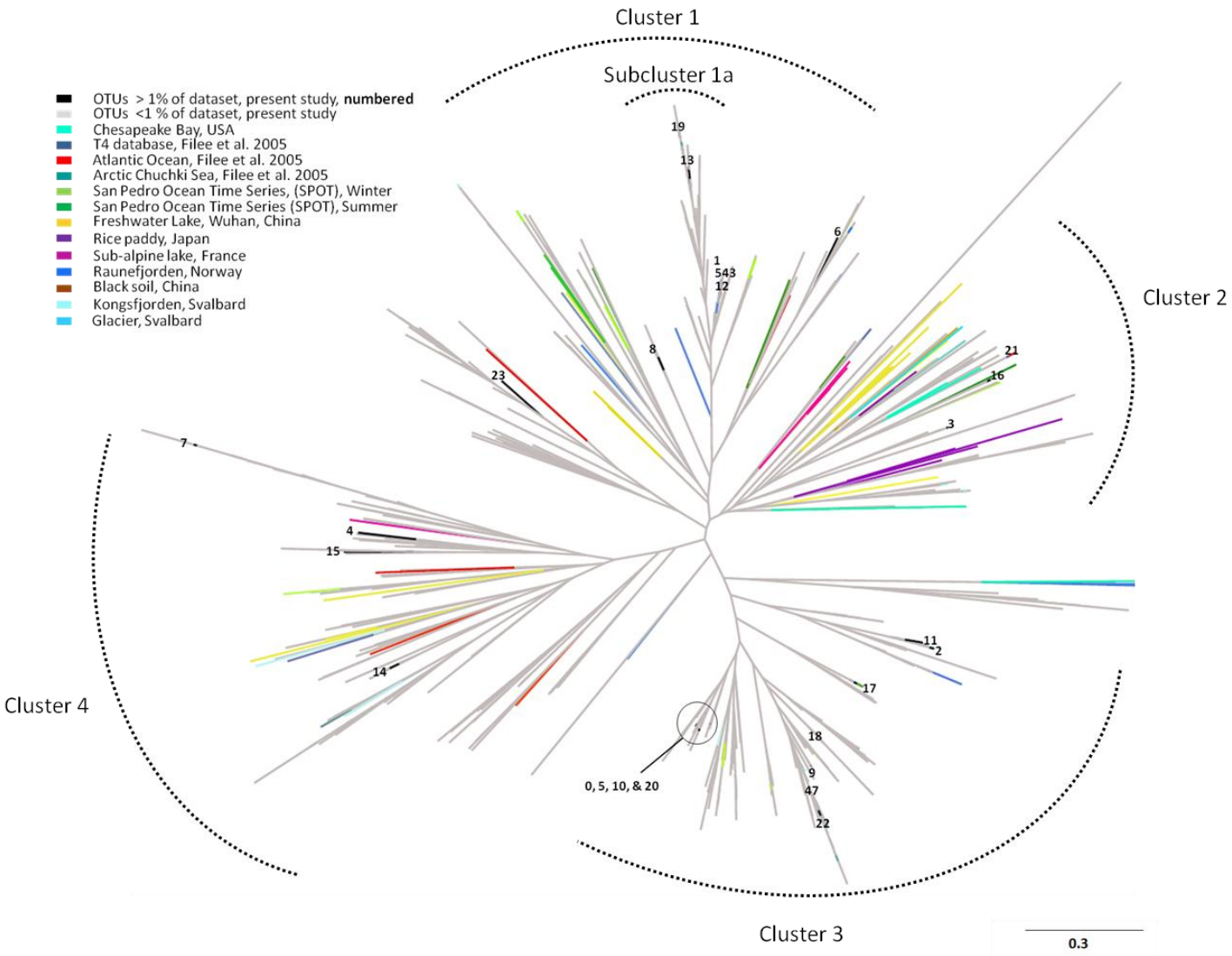


Figure 21. Phylogeny based on alignment of *g23* OTU representative sequences and non-redundant sequences sourced from top BLAST hits of the 369 OTUs present in all 8 samples. OTUs representing ≥ 1% of the dataset are labeled in black (at least 7,000 sequences per OTU). The 5 OTUs with the largest number of reads in the dataset are OTUs 1 (65,454 reads), 0 (51,147 reads), 2 (36,500 reads), 3 (23,808 reads), and 4 (19,433).

3.8 Comparison of Roche/454, Illumina, and Ion Torrent platforms

The three-platform comparison of *g23* data generated on Roche/454, Illumina, and Ion Torrent platforms resulted in 2,634 total OTUs when compared at 90% sequence similarity. The rank abundance curve of the OTUs (Figure 22) showed an agreement of abundances for a majority of the OTUs in the dataset. Exclusively Ion Torrent data exhibited divergences in abundances above the curve, especially for the OTUs that were least abundant across all three datasets.

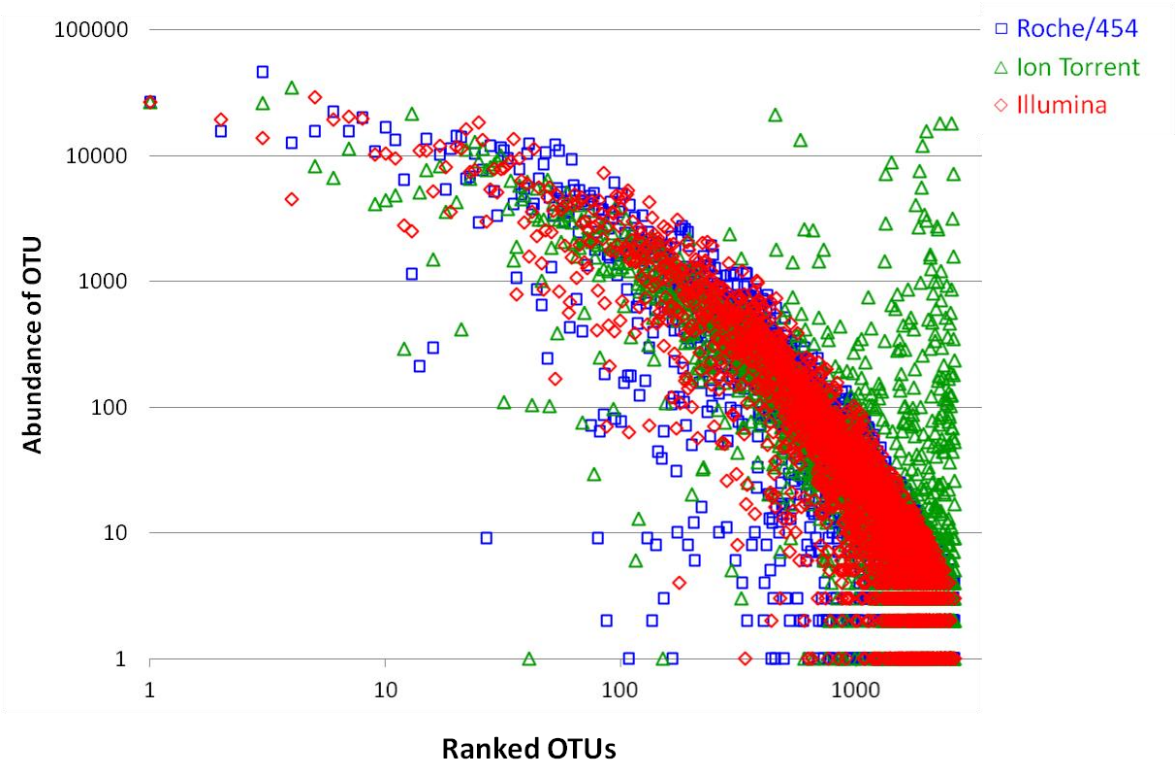


Figure 22. Rank abundance curves of Roche/454, Illumina, and Ion Torrent datasets of *g23* OTUs (based on 90% sequence similarity) from datasets randomly sub sampled to equal depth (913K sequences each).

Alpha rarefaction curves of *g23* samples showing within-sample diversity for each sample amongst the three datasets (Figure 23) indicate that Ion Torrent sequencing resulted in the highest numbers of OTUs (indicated by "Species" on y-axis) in all samples. A majority of samples show that Roche/454 and Illumina datasets have similar richness and evenness, indicated by their well-matched curvatures.

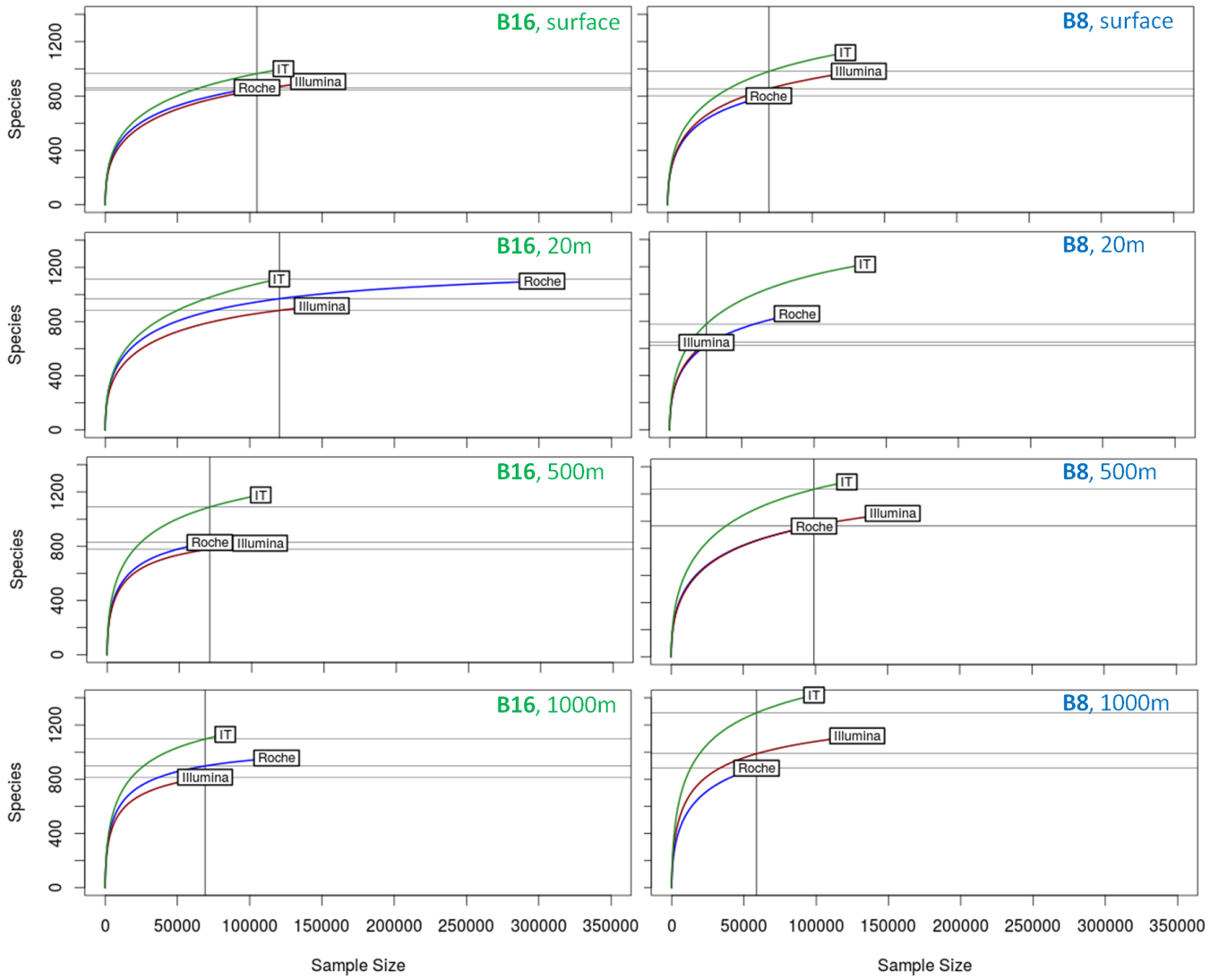


Figure 23. Alpha rarefaction curves of each *g23* sample from sequencing runs performed on the Roche/454 (blue), Illumina (red), and Ion Torrent (IT, green) platforms. Sample names are indicated in the upper right corner of each plot, and labels are colored by station. Datasets were normalized to the number of sequences of the smallest dataset (913K sequences). Black vertical lines represent the maximum sampling depth in the rarefaction curve, and black horizontal lines indicate the species number at maximum rarefaction depth.

Figure 24. Plot in three-dimensional space using PCoA of the weighted UniFrac distance matrix of *g23* samples analyzed on the Roche/454 (top panel), Illumina (middle panel), and Ion Torrent (lower panel) platforms. The same plot is shown in each panel, highlighting samples from each platform separately for clarity. Transparent points within each panel are the *g23* samples from the other two platforms. Each point has a different color, designating each water sample. Sequencing platforms are labeled along with sample names. Tightly clustered red, orange and dark blue points to the left side of the figure are station B8 surface, 20 m, and 500 m samples from all platforms.

The PCoA plot (Figure 24) based on weighted Unifrac distance matrix (a phylogeny-based method) of the three *g23* datasets reveals that PC1 describes 91.17% of the variation, whereas PC2 and PC3 describe 5.41% and 1.06% of the variation, respectively. Atlantic Water samples from station B8 at surface (red sphere), 20 m (blue sphere) and 500 m (orange sphere) cluster together tightly within all three datasets. Station B16 samples originating from the ice-influenced Surface Water are more closely related to one another than to any other sample, though the position of these samples in the PCoA plot appear to vary by platform rather than by depth. The same is true for all B16.1000m samples, which originate from Polar Water. The clustering of Polar Water samples B16.500m and B16.1000m in the UPGMA dendrogram of Roche/454 data only is not preserved in this PCoA of the platform comparison: instead, sample B16.500m appears to cluster as most similar to station B8 Atlantic Water samples for all three platform datasets. Arctic Intermediate Water sample B8.1000m clusters with station B8 samples from Atlantic Water for the Roche/454 dataset, but is separated from Atlantic Water samples in the Illumina and Ion Torrent datasets.

Comparison of categories using the weighted Unifrac distance matrix of all samples in the platform comparison tested using ANOSIM (Appendix B.7) revealed that grouping of samples by sequencing platform was not statistically significant, with an R value of -0.027 and a p-value of 0.55 (Table B-6). A second ANOSIM (Table B-7) on samples grouped by water mass showed categorical grouping with statistical significance, with an R value of 0.688 and a p-value of 0.001.

4 Discussion

4.1 Is diversity within the Arctic Ocean viral community distinct from that of other geographic locations sampled to date?

Our results of the viral diversity in these Arctic Ocean samples during the polar night were explored to determine the phylogenetic relationships of viral sequences in this study to other viral samples globally. The connectivity and physical mixing of oceans allows for transport of viruses and their hosts within and between marine ecosystems globally, suggesting that everything might actually be everywhere, as in Baas-Becking's hypothesis. At the same time, available nutrient heterogeneity, light, temperature and salinity gradients, and bottom depth differences are known to create limitations to marine microbial dispersal and act simultaneously as the forces with which the environment selects for certain types, contributing to beta diversity of microbial communities across spatial and temporal scales (Zinger et al. 2011).

The use of the *g23* gene in numerous and globally disparate previous studies of *Myoviridae* diversity (Filée et al. 2005; Pagarete et al. 2013; Chow and Fuhrman 2012; Short and Suttle 2005; Butina et al. 2013; Liu et al. 2012; Bellas and Anesio 2013; Zheng et al. 2013; Bench et al. 2007) provides pre-existing information to make some inferences about the geographic novelty of *g23* genotypes found during a single sampling effort in the Arctic Ocean. This is not the case for *phoH* (Goldsmith et al. 2011; Goldsmith et al. 2015) and *MCP* (Larsen et al. 2008; Zhong and Jacquet 2014): relatively few prior studies have used these markers to investigate viral diversity , therefore the following discussion regarding geographic ecology of the Arctic viral assemblage is limited to information based on the *g23* dataset only.

Our data appear in accordance with findings from numerous previous studies that find viral community assemblages are composed of a mixture of globally distributed and geographically constrained types (reviewed in Breitbart and Rohwer 2005)). Within the *g23* tree based on FastTree phylogeny assignments (Figure 21) Clusters 2 and 4 contain fewer of the OTUs with abundances >1% of the dataset, and the sampling origin of BLAST hit sequences are widely distributed across ecosystems and regions. The origins of BLAST hits in Clusters 2 and 4 are sourced from globally distributed terrestrial, freshwater, and marine sampling sites, of which very few are either Arctic or recorded as collected during winter. Our hypothesis is that the distribution of genotypes within these clusters is not endemic to a particular locale or environmental condition, but rather that genotypes in these clusters originate from viral types that are successful across a wide range of environments.

It is possible that these cosmopolitan viruses are able to infect globally distributed host species, though the presence of commonly available host cells is not the only possible explanation for the wide distribution of these viral

types. Previous studies have indicated that identical viruses are found in vastly different ecosystems, implying mobility between environments e.g. between soil, marine, and freshwater habitats (Breitbart et al. 2004; Short and Suttle 2005, Sano et al. 2004). The demonstrated ability of some viruses to broaden their host range upon exposure to new hosts (Chibani-Chennoufi et al. 2004)enhances the aptitude of a virus to successfully move between environments, and does not require presence of the same host species across environments. As mentioned earlier, Chibani-Chennoufi and colleagues (2004) found a myophage type able to broaden its host range during low-light conditions. Myophage under polar night conditions may also be able to achieve this feat in situ. Until virus – host interactions are explored further with these globally distributed types, however, it is impossible to know the reason for their ubiquity.

BLAST hit sequences falling within two phylogenetically distinct clusters in the *g23* tree containing the majority of the most numerically abundant OTUs (Subcluster 1a and Cluster 3) belong to samples previously collected in Arctic or marine waters off of Norway, or else they are from surface samples collected during winter at a coastal California sampling station known as SPOT (defined as November to February). The genotypes in Cluster 3 and Subcluster 1a appear to represent two groups of myophage mainly found in the Arctic environment, but are less phylogenetically distant from the two globally distributed clusters than from one another. Member genotypes within Cluster 3 and Subcluster1a may therefore represent two myophage types selected for in the high Arctic. The eight most abundant OTUs in the ice-influenced Surface Water samples at station B16 all fall within these Arctic-specific clusters, whereas only half of the most abundant OTUs in the deep Polar Water samples at station B16 and the Atlantic Water samples at B8 fall within these clusters. It is possible that these Arctic OTUs are specific to ice-associated communities (Borriss et al. 2003), and that vertical transport of viruses via sinking particles (Fuhrman 1999) has contributed to the dispersal of these viral types to deep water masses.

Although query sequence coverage and percent identity to BLAST hits were generally lower for queries matching coastal California samples than for queries matching Arctic or marine samples from Norway, it is interesting that sequences from the eastern Pacific Ocean taken in winter cluster as similar to both Arctic BLAST sequences and the majority of high abundance OTUs found in the present study taken during the polar night. Unfortunately logistical issues often prevent polar scientists from winter sampling due to sea ice cover, harsh sea conditions, and constant darkness, which has resulted in datasets mainly collected in summer in this region. Thus, clustering of samples from the temperate Pacific and the Arctic Ocean with no obvious relationship except season of sampling may reflect insufficient resolution due to sparsely sampled *g23* data across locations and seasons, but could also suggest the success of these myophage genotypes across marine environments during the lower production winter season.

Alternatively, the clustering of Arctic OTUs with Pacific Ocean samples could be due to the significant transport of biota between the Pacific, Atlantic, and Arctic Oceans. Although the gateways to the Arctic are limited by surrounding continents, persistent and strong transport of water and biota to the Arctic Ocean occurs via advective

processes from both the Pacific Ocean and, to a greater extent, the Atlantic Ocean (Wassmann et al. 2015). Microbial species will inevitably be transported along these currents, which could explain the presence of viruses (and likely host species) previously collected in Pacific Ocean water samples. If indeed this is the case, the presence of free-living Pacific Ocean originated viruses in such a high percentage within the viral community would require their successful reproduction upon entry into Arctic Ocean environment (Wassmann et al. 2015) rather than an inactive contribution to the total diversity. The rate of spread of these viral types is an aspect to consider: high mobility of herpes viruses alongshore have been reported, with dispersal rates of about 10,000 km per year (Lawrence 2008). Once passing into the Arctic Circle during the polar night, the obstacle of degradation due to solar radiation would not exist, though the longevity of virus particles under these conditions remains untested. It is possible that viruses with extended longevity that are transported along such currents do not have many barriers to their dispersal (McCallum et al. 2003).

Viral community assemblages in these Arctic Ocean samples are made up of both globally distributed and local types, as has been shown not only in other viral diversity investigations (Short and Suttle 2005; Breitbart et al. 2004; Huang et al. 2015), but also for prokaryotic (Zinger et al. 2011; Ghiglione et al. 2012; Pommier et al.2005; Massana et al.2000)and eukaryotic (Montresor et al. 2003; Darling et al. 2000) species investigations. Huang et al. 2015 also found closely related phage types with either habitat-specific relative abundances or globally consistent abundances across viral communities within open ocean, coastal marine, estuarine, and coral reef viral communities. Habitat-specific and globally consistent types formed separate phylogenetic clusters (Huang et al. 2015) similar to the patterns exhibited by myophages examined in the present study. These results imply heterogeneity of distribution of viral types within and between environments, indicating presence of rare and abundant viral taxa. The presence of global and local types supports that viral diversity is high locally but perhaps low globally, and that these patterns occur simultaneously within individual viral assemblages (Breitbart et al. 2005).

4.2 Is viral community composition distinguishable between water masses or other physical/chemical environmental factors, and does it reflect host community diversity?

4.2.1 Viral communities within different water masses

The sampling accomplished in this work captured viral concentrates originating from ice-influenced Surface Water, deep Polar Water, warm Atlantic Water carried north by the WSC, and deep Arctic Intermediate Water. Viral assemblages originating from different water masses investigated in this thesis are for the most part distinguishable from one another, with the exception of the Atlantic and Arctic Intermediate Waters. Water mass assignments were based on density gradients throughout the water column, though the density differences at station B8 are less dramatic between surface and deep samples compared to station B16 (Figures 10 and 11) and therefore a lack of physical barriers may exist between the water masses at station B8. Repeated sampling is required to characterize the viral assemblages within these water masses, however, this pilot investigation of the notoriously less stratified

winter Arctic Ocean water column is a promising step towards this end.

The viral assemblages in our study also support viral transport via water mass as shown in UPGMA clustering of both *g23* (Figure 15) and *phoH* samples (Figure16), which group station B16 samples based on water mass (either Surface Water or Polar Water), though only the grouping of *g23* samples by water mass had statistical support when using ANOSIM. The relationship between water masses is less apparent in station B8 samples originating from Atlantic Water and Arctic Intermediate Water. The dominant dispersal method for microbes in the ocean is by movement of water masses (Winter et al. 2013), and other studies have found planktonic eukaryote, prokaryote, and viral community assemblages to vary in relation to the water masses samples are sourced from ((Winter et al. 2013; Agogué et al. 2011; Gómez-Pereira et al. 2010; Varela et al. 2008; Galand et al. 2010; Fu et al. 2013; Monier et al. 2014). The formation of Arctic Intermediate Water is in fact complex, as it is sourced from cooling Atlantic Water mixed with Arctic bottom waters and, to a lesser degree, Polar Water (Blindheim 1990). The very nature of intermediate water masses adds complexity, and it is therefore not surprising that distinguishing between viral communities in this layer and the water masses contributing to its formation is less concrete than between other, more distinct water masses.

In addition to the physical bounds created and transport mediated by water masses, viral production represents another facet of viral diversity within the context of these water masses. The abundance of viruses in aquatic systems is affected by viral production and loss rates, which in turn can vary with burst size, frequency of host infection, host diversity, and rate of viral decay (Clasen et al. 2008). Viral infection rate estimates in the pelagic Arctic Ocean are extremely low compared to other oceans due to low virus concentrations and elevated viscosity of cold arctic waters (Steward et al. 2007) such as the Polar Water and Arctic Intermediate Water masses. As mentioned earlier, infection rate in the marine environment is density-dependent (Dennehy 2013), and host-virus contact rates have been shown to be ten times lower in the Arctic Ocean than in temperate waters (Steward et al. 2007). From the results of their study and several previous investigations of Arctic Ocean virus abundance, Steward et al. 2007 hypothesize that because infections persist despite less frequent contact with hosts, Arctic Ocean viruses may have reduced decay rates, which partially compensate for density-dependent limitations. Decay rates are known to vary widely between algal viral species or even between strains of a single algal viral species (Tomaru et al. 2005), and rates of decay or deactivation are sometimes temperature sensitive (Baudoux and Brussaard 2005). Thus it should be considered that reduced host contact rates and longevity of free-living viral particles due to the environmental conditions within these water masses may play a role in the relative abundance of viral OTUs in our study.

The clear distinction between viral communities in the fresher Surface Water and the more saline deep Polar Water samples at station B16 is probably linked to sea ice influence in the subfreezing surface layer and lack thereof at 500 - 1000 meters depth. Sea ice harbors many microbial cells and creates a niche environment that is a biomass and production hotspot even during the polar night (Bachy et al. 2011). It should be noted that OTU richness was also

highest in samples originating from sea-ice influenced Surface Water in both the *g23* (Table 11) and *MCP* (Table 13) datasets. The most abundant *g23* OTUs in the ice-influenced Surface Water samples at station B16 all have high homology to other myophage only previously found in the Arctic or in marine sampling sites off of Norway and are rare or absent from all other samples in this study (Figure 18). This indicates that genotypes within the Surface Water myophage assemblage may be endemic to the Arctic. Borriss et al. 2003 found tailed bacteriophages (including myophages) infecting psychrophilic bacteria isolated from sea ice appear even more severely cold-adapted than their bacterial hosts, with growth maxima below 14°C and successful plaque formation at 0°C (Borriss et al. 2003). Phages of psychrophilic bacteria have also been isolated from marine waters (Delisle and Levin 1969), though no direct in-depth study has yet been done to characterize the differences in phage genotypes between sea ice and surface water communities.

Polar Water samples contained extremely high abundances of reads from two *g23* OTUs (23 and 21), neither of which had any significant BLAST hit (no hits >70% sequence ID and over 50% query coverage) and group to Clusters 1 and 2 respectively in the *g23* phylogeny (Figure 21). Deep ocean sampling is less common in studies of marine microbial ecology: much of the work done to date is limited to the highly productive euphotic zone of the water column (Zinger et al. 2011). While viral data in the deep sea is still extremely sparse, it is also possible that this lack of genotype identity to sequences in the BLAST database may signify the novelty of these OTUs within deep arctic-derived Polar Water.

The relative similarity of all *g23* and *phoH* samples at station B8 to Surface Water samples of sea ice origin at station B16 may be a result of deep winter mixing and lack of a density gradient at station B8. Winter convection in the Arctic Ocean can homogenize the water column down to 200 m depth in some places, and possibly to the bottom along the shelf (Rudels et al. 1991). The relatively small density gradient at station B8 may reflect this winter convection, which could explain the absence of a cold Surface Water layer at station B8 and create homogeneity of the water column, allowing greater vertical particle movement and resulting in homogeneity of the viral population between the Atlantic Water and Arctic Intermediate Water below. Unfortunately, no sample in this study captured a cross-section of the Atlantic Water layer at station B16, which may have clarified the relationship between the Atlantic Water samples at station B8 and the Surface Water samples at station B16.

The difference between the range of viral families captured by *g23* and *phoH* using Roche/454 sequencing may explain UPGMA clustering variations between these datasets. The closest relationship in the *g23* dataset among the station B8 samples is between samples from Atlantic Water above the pycnocline (Figure 15), whereas the closest relationship between station B8 samples in the *phoH* dataset is that of Atlantic Water below the pycnocline and Arctic Intermediate Water samples (Figure 16). The *g23* primers used in this thesis target a wide diversity of the *Myoviridae* assemblage while *phoH* primers target viral types originating from an assortment of phylogenetically unrelated viral families. As a result, the two datasets may not necessarily share the same relationships between

closely related samples. It is likely that the *phoH* and *g23* datasets contain only some of the same viruses, as *phoH* can be found within *Myoviridae* genomes but not all (Goldsmith et al. 2011).

The *phoH* dataset is dominated by only fourteen OTUs: 85.5% of the total *phoH* reads are members of these OTUs, making this a highly uneven dataset. Of these OTUs, the three most numerically abundant *phoH* OTUs contribute more than half of the sequences in the dataset. In contrast, nearly twice as many highly abundant OTUs (26) account for 49% of the total *g23* reads, and no single OTU contributes more than 6% of sequences in the dataset. Such large representation of only a few *phoH* OTUs may account for the difficulty in relating the *phoH* UPGMA clusters directly to relative abundances in the OTU *phoH* heatmap of these fourteen OTUs (Figure 19). A 2015 study by Goldsmith and colleagues showed that deep sequencing of *phoH* at the Bermuda Atlantic Time Series (BATS) station also resulted in dominance of few (only five) OTUs comprising more than half of the sequences in their dataset which spanned over multiple seasons and years. These abundant OTUs were found across a majority of the samples, while all remaining OTUs had fewer member sequences and were rarely found in individual samples. The proportional representation of these dominant OTUs in each sample, however, exhibited considerable variation spatially and temporally. With this largely uneven dataset, the authors introduced a median rank parameter to glean sufficient resolution of the community dynamics of OTUs within the top half of each sample community (Goldsmith et al. 2015). By analyzing more than just the between-sample abundances of the OTUs with the most reads, an investigation of the top half of the *phoH* community may further inform the distinction between water masses evidenced by the UPGMA clustering of samples in our severely uneven dataset.

MCP samples originating from Polar Water (B16.1000m) and Arctic Intermediate Water (B8.1000m) form a cluster within the *MCP* UPGMA dendrogram (Figure 17). Similarities between these *Phycodnaviridae* and *Mimiviridae* assemblages may be more dependent on depth of sampling rather than resident water mass. It is plausible that viruses associated with sinking particles have contributed to this effect. Non-growing or dying cultures of flagellated cells and diatoms demonstrably sink at more rapid rates than growing cultures, and different species of algae sink at different rates depending on cell density and aggregate formation (Eppley et al. 1967). Sunken algal cells or aggregates may therefore influence the algal virus communities found in the deepest samples in this thesis, whereas algal viruses predating on active members of the algal community may dominate surface water samples. The Pielou's evenness values for the *MCP* dataset (Appendix B.3, Table B-2), which generally increase with increasing depth (possibly due to dominance of fewer actively growing types at surface than in the lower water column), may support this inference.

Throughout the water column at both sample stations, chlorophyll a levels were below the detection limit (Figure 11). The polar night is a challenging period for autotrophs: density of autotrophic cells may be reduced during this time due to cell death and grazing pressure (Bachy et al. 2011). Mixotrophy has been shown to be an ecologically

important mode of energy production among marine pico and nanoplankton globally (including in Arctic waters) and has been implicated from other investigations as an important survival mechanism for a diversity of protists during the polar night (Bachy et al. 2011; Baldisserotto et al. 2005; Bell and Laybourn-Parry 2003; Gradinger 2015). In the absence of light under polar night conditions, the dominant algal virus types in all of these samples are likely infecting active cells capable of mixotrophic means of energy production.

The only abundant OTU in the ice-influenced Surface Water samples in the *MCP* dataset with a species-identified BLAST hit was OTU 9, with 72% sequence similarity to a nanoflagellate *Pyramimonas orientalis* virus (Figure 20), previously isolated in Norway (Sandaa et al. 2001). While sequences clustering to this OTU were highly abundant in the ice-influenced samples, they were entirely absent in all other water masses. *Pyramimonas* species are frequently present in polar waters, and some have been reported to be capable of mixotrophy (Gast et al. 2014). Previous findings show that *Pyramimonas* can be found both in the Bering Sea water column and trapped in Arctic sea ice (Różańska et al. 2008). Although 72% identity represents a very limited homology to any of these viral sequences and any such interpretation is highly speculative, BLAST identities to putative viruses of mixotrophs in the cold Surface Water layer may indicate the presence of active mixotrophic cells within this water mass or otherwise may be sourced from nearby sea ice.

4.2.2 Trends in viral community and host community diversity

While the relationships between phages and their bacterial hosts remain unknown for the environmental samples assessed in this thesis, it is evident upon examination of the 16S rRNA gene assemblages (which describes a very broad diversity of prokaryotes) within the same water samples that diversity of both myophages and their potential host cells vary similarly. Of the two genes assessed in this thesis for fingerprinting of prokaryotic viruses (*g23* and *phoH*), the UPGMA dendrogram of *g23* appears to have greatest similarity to the UPGMA dendrogram based on *16S* data at these sampling sites (Figure 25, *16S* data courtesy of Oliver Müller, UiB) , suggesting a relationship exists between myophage and prokaryotic communities in these samples. This is an interesting result, as Payet and Suttle's study in the Canadian Arctic Ocean (2014) did not find correlations between T4-like bacteriophage assemblages and their prokaryotic host communities when tested using ANOSIM on Bray-Curtis distance matrices of their datasets. The UPGMA dendrograms in the present study are based on weighted UniFrac distance matrices from *16S* and *g23*, however, use of the Mantel Test (generally a more powerful method than ANOSIM (Anderson and Walsh 2013)) on the Bray-Curtis distance matrices for these datasets also did not yield a significant relationship (see Appendix B.7). Further investigation of associations between *16S* and *g23* data should be performed based on UniFrac distance (which is phylogenetically based) to ascertain whether statistical support exists for the similarities in the UPGMA topologies shown above.

It is also interesting that our work shows virus assemblages varying similarly to their potential prokaryotic host assemblage, as patterns of co-variation of marine viral and host populations are observed more often over time

rather than spatially (reviewed in Huang et al. 2015). For instance, multi-seasonal time series investigations of cyanomyophage assemblages in the Red Sea and in Raunefjorden, Norway found that the dominant cyanomyophages co-varied with the *Synechococcus* population (Mühling et al. 2005; R.-A. Sandaa and Larsen 2006; Pagarete et al. 2013b), while a different study of cyanomyophage assemblages along a transect did not find any significant relationship to host cyanobacterial abundance or diversity (Jameson et al. 2011). As another example, Needham et al. 2013 found *g23* and *16S* OTUs correlated only when they varied in their relative abundances over time and were also abundant in the overall dataset (Needham et al. 2013).

Previous investigations find that myophage contribute to a majority of marine dsDNA viral populations infecting a broad range of hosts belonging to bacteria, archaea and cyanobacteria (Breitbart et al. 2002; Angly et al. 2006; Brum et al. 2015) although new metagenomic methods allowing for amplification of single-stranded DNA and RNA viruses have led to evidence that tailed dsDNA bacteriophages may not dominate viral communities (reviewed in Kim et al. 2013). The dynamics between virus and host cannot yet be known due to lack of information about which myophages infect which prokaryotes, however, it is hypothesized that bacteria-virus pairs may exhibit a time-lagged association (Needham et al. 2013): abundant viral particles may be associated with host cells that were previously dominant but have since been destroyed as a result of a Kill the Winner dynamic (Thingstad and Lignell 1997) and are therefore more rare at time of sampling.

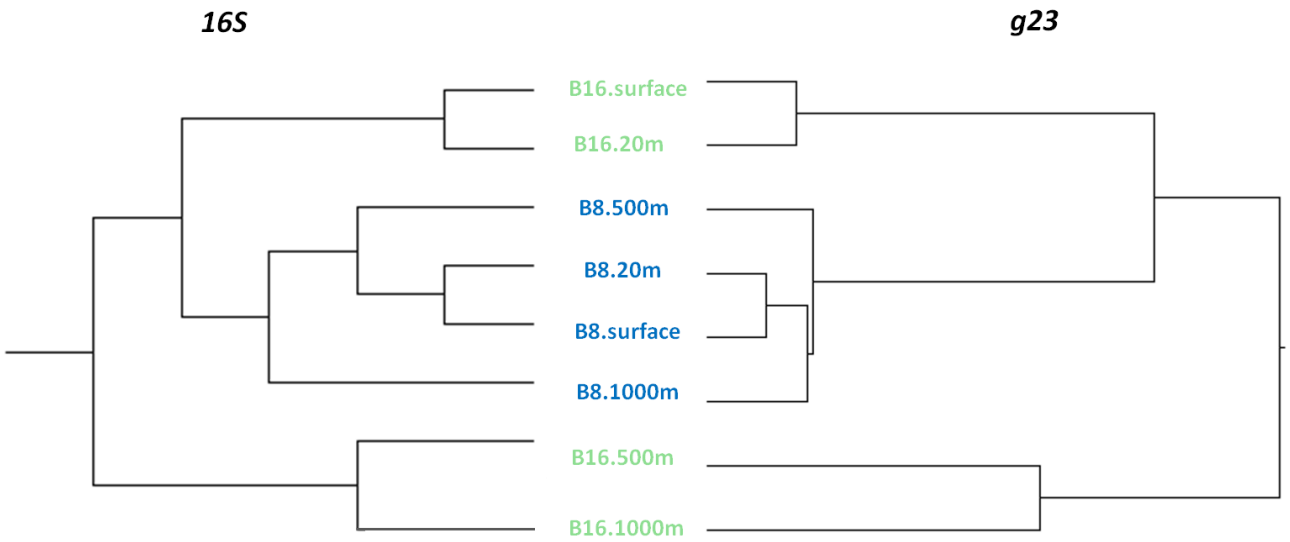


Figure 25. Weighted UPGMA dendrogram of prokaryotic 16S data (left dendrogram) with mirrored *g23* data (right dendrogram) from water samples at the same sites (both based on Unifrac distance matrices). Station B16 samples are labeled in turquoise and station B8 samples are labeled in blue. The *16S* dataset was collected by colleagues on the MicroPolar project and analyzed by Oliver Müller.

This thesis did not examine whether algal host communities were reflected by their potential viruses, though the algal virus BLAST hits in the present study may provide some insight to the ecologically important algal groups present in the environment during the polar night, with the knowledge that entries in the BLAST database for algal *MCP* are extremely underrepresented. Several of the putatively identifiable virus types in the *MCP* dataset included

OTUs with limited sequence similarity to *Haptolina ericina* and *Prymnesium kappa* viruses. The two isolates these BLAST hits originate from are haptophyte viruses sampled in the fjords of Norway. Haptophyte species of the *Haptolina* and *Prymnesium* groups often occur in low abundances in the marine environment but can form blooms (Edvardsen and Paasche 1998). Johannessen et al. 2015 investigated the host ranges for haptophyte viruses from two Norwegian fjords, and found several *H. ericina* virus isolates and a *P. kappa* virus isolate that were able to infect the same hosts, including several strains of *H. ericina* and one *P. kappa* strain, while other viral isolates had more specific host ranges (Johannessen et al. 2015). As the percent identities to any of the isolated viruses in the study by Johannessen and colleagues were very poor, the host range of these arctic algal viruses remains unknown. Thus the possibilities for haptophyte viral host range for the viruses isolated in the present study could be broad or narrow, and further investigation of these arctic types should be investigated to understand the host-viral relationships for this ecologically important microalgal group.

It was unexpected that none of the *MCP* OTUs exhibited homology to a BLAST entry for *MCP* from viruses known to infect *Micromonas pusilla* (full batch BLAST results not shown). This phototrophic prasinophyte has been previously observed as the most abundant and active photosynthetic cell type in western Arctic Ocean waters (Lovejoy et al. 2007), persisting and growing even throughout the polar night in both the western Arctic Ocean and in the waters north of the Svalbard archipelago (Lovejoy et al. 2007; Vader et al. 2014), possibly through the ability to carry out phagotrophy (McKie-Krisberg and Sanders 2014). The absence of *MCP* OTUs with homology to *M. pusilla* viral predators may simply reflect the paucity of information in the NCBI BLAST database, as algal viruses have not yet been deeply sequenced at this high latitude.

4.3 Does use of different sequencing platforms produce comparable diversity capture for the same environmental viral assemblages?

As found in several recent platform comparisons of different types of amplicon libraries (Luo et al. 2012; N. J. Loman et al. 2012), the Roche/454 and Illumina platforms are comparable in diversity metrics, though the Illumina dataset yielded higher alpha diversity estimates of *g23* in some cases (Figure 23). All three platforms yielded similar broad patterns in beta diversity, despite the Ion Torrent dataset reporting higher overall alpha diversity of samples. Although Roche/454 sequencing of the *g23* samples clustered the Atlantic Water and Arctic Intermediate Water samples both in the UPGMA dendrogram of Roche/454 data alone (Figure 15) and in the PCoA of the platform comparison (Figure 24), greater Unifrac distances were observed between Atlantic Water and Arctic Intermediate Water samples sequenced on the Illumina and Ion Torrent platforms. All other between-sample relationships were similar for all three datasets, thus this discrepancy is unexpected (especially because the *phoH* dataset also does not differentiate these water masses). If it were possible, Roche/454 sequencing of *g23* amplicons should be repeated to confirm this result.

Additionally, the clustering of Polar Water samples B16.500m and B16.1000m seen in the weighted UniFrac matrix described in the UPGMA dendrogram of the *g23* Roche/454 dataset (Figure 15) is not preserved in the PCoA based on the weighted Unifrac matrix of the platform comparison (Figure 24). Instead, sample B16.500m appears to cluster as most similar to station B8 Atlantic Water samples for all three platforms. The B16.500m water sample was collected at the interface between the Cold Atlantic Water and Polar Water masses (Figure 11, top panel). Further, the *phoH* and *MCP* UPGMA dendrograms (Figures 16 and 17, respectively) also indicate similarity of station B8 Atlantic Water samples to B16.500m. Therefore the hierarchical clustering based on the eight samples in the Roche/454 *g23* dataset alone is likely insufficient basis on which to characterize this viral community, and inclusion within a broader dataset (as with our 24-sample platform comparison) clarifies its grouping.

From the rank-abundance curve of the three platform data (Figure 22), the Ion Torrent reads are visibly more numerous for OTUs that are less abundant in the other two datasets, with Ion Torrent being the only dataset straying above the curve. Several necessary differences in sample preparation and/or in the post-processing likely contributed to these disparities. Two notable differences were made in the preparation of the Ion Torrent sample, as well as one important difference in the post-processing: 1) six replicates of each initial PCR reaction (without adapters) were combined as template material for second PCR (with adapters) rather than two replicates as in the preparations for Roche/454 and Illumina platforms. This was necessary because such a high yield (1µg) of DNA was required by the UiB Ion Torrent PGM facility. Normally high total sample DNA is achieved by pooling many samples together before running them, but the present study incorporated only eight samples. A recent study using fungal ITS region templates found that pooling of greater numbers of PCR reaction replicates did not change the diversity metrics of samples; only greater sequencing depth had a clear effect on beta diversity (Smith and Peay 2014). Therefore we expect that the number of pooled PCR replicates to have had minimal effect on the results. 2) The adapter sequence was exchanged on the reverse primer to integrate the adapter specific to the Ion Torrent platform. This was done to make the sequencing reaction was as comparable as possible to pyrosequencing, which also uses emPCR. The use of two adjacent adapters on the reverse primer rather than just one is impractical and not as comparable. One oversight in the Illumina dataset was use of Roche/454 adapters adjacent to Illumina adapters. For more comparable performance, Illumina adapters would have ideally replaced Roche/454 adapters. 3) Use of an effective post-processing pipeline (UPARSE) step to successfully trim the primer from the reads necessitated cleaning of sequence data based on expected error rather than on a PHRED score cutoff. It is possible that this method of data cleaning altered the proportions of kept sequences in the Ion Torrent dataset differently from that of the other two platforms. We find this scenario is unlikely as disparities seen in the Ion Torrent dataset are preserved from the author's earlier attempts of relating the data even without the primer sequence cleaning (data not shown).

4.4 Discussion of methods

4.4.1 DNA sample collection and extraction

A major limiting factor in data collection was time taken to perform the tangential flow filtration of 50 liters of seawater for a single viral concentrate. Each filtration, which can take upwards of six hours for each sample, must be done on site before the community assemblage has time to alter. Time and equipment dedicated to this processing were limited resources, which led to the collection of only eight total samples for viral community fingerprinting. Such limited sample size restricts the statistical evaluations possible with respect to the environmental parameters. Instead, the power of these datasets is in the sequencing depth achieved and community diversity.

The tangential flow filtration method itself is biased, variably capturing between 2 – 98% of viruses (John et al. 2011). In a study examining isolation methods of viral DNA, an iron flocculate method used to treat freshwater for virus removal was found to capture 94% of the virus particles present, while tangential flow filtration methods captured 23% of virus particles (John et al. 2011). Additionally, as half of the source water for the deepest samples in this work was prefiltered through 0.2 µM filters rather than 0.45 µM as for all the other samples, it is possible that the relative abundance of virus particles larger than 0.2 µM in the viral community could be non-representative of the actual relative abundance in the assemblage. Thus, the efficacy of the viral particle concentration method limited the sample number and may have contributed to sample bias in the present work.

4.4.2 PCR bias and quality trimming

The author is aware of the inherent biases of PCR amplification which have been reviewed thoroughly in many publications and their importance in microbial community investigations (e.g. Pinto and Raskin 2012; Polz and Cavanaugh 1998; Sipos et al. 2010). The consistently low yield of viral DNA in seawater samples presents limitations that cannot be avoided: e.g. in this work, the *MCP* gene required 80 total cycles of PCR for the amplification to produce any significant product. This cycle number is well above the recommended maximum of 35 cycles (New England BioLabs 2015), however, six different reactions were pooled for each sample to mitigate this bias and all samples were all treated equally throughout the amplicon library preparation, and are therefore still considered comparable. The excess cycles of PCR may explain the lack of any OTUs found universally in the *MCP* dataset, especially at the later cycles in the reaction, when templates are known to anneal to one another and prevent template-primer annealing (Suzuki and Giovannoni 1996).

In addition to PCR bias, the degenerate primers used in this study were developed based on known viral examples of the genes of interest, and therefore will amplify known subsets of genes previously seen in marine viruses but may not amplify the full assemblage of present types within the viral groups studied. Primer design is a balance between specificity/efficacy and product yield (Sipos et al. 2010), and although proportions of OTUs may not represent the true relative abundances of the microbial community in these samples due to amplification biases, without PCR

steps viral DNA fingerprinting would not be possible.

Quality trimming of both sequence ends by eliminating ambiguous bases, homopolymers, and PHRED score calls below a threshold value were done in tandem within the BBMap package created by Brian Bushnell. From the per base quality scores in the FastQC reports for each dataset (Appendix B.2, Figures B-6, B-9, and B-12) and also based on evidence presented in previous studies (reviewed in Del Fabbro 2013) quality trimming of ends is a superior method to averaging quality scores over the length of a read as the quality score tapers sharply only at the 3' end of sequences.

4.4.3 OTU picking and chimera checking

Use of a 97% sequence similarity threshold for defining an OTU is a well-established method for defining "species" from molecular sequence data from prokaryotes (Stackerbrandt and Goebel 1994). It might be argued that 3% dissimilarity might be too stringent for viral diversity, and some workers have used thresholds of 5% or 10% dissimilarity. The rate of mutation in dsDNA phage genomes is $10^{-7} – 10^{-8}$ changes per base pair per generation (Drake et al. 1998; Drake and Holland 1999), which is orders of magnitude higher than microbial host mutation rates, but not as high as for other viral types such those with RNA genomes. Until a threshold value for such viral datasets can be agreed upon within the scientific community, divergence from standard operating procedures and other bioinformatic tools designed for prokaryotic data is risky business for the average user without full examination of the consequences for such choices. In the case of this thesis, because no large databases of viral sequence information are centrally curated as are tools like GreenGenes (DeSantis et al. 2006) or SILVA (Pruesse et al. 2007), this work relied on de novo picking of OTUs and chimera checking. The algorithm for *de novo* OTU picking available within QIIME is Robert Edgar's algorithm known as USEARCH (Edgar 2010). Edgar recommends against expansion of the OTU radius of dissimilarity greater than 3% because doing so discards more true biological sequences that will not contribute to *de novo* chimera checking, which relies on the most common sequences as a reference (http://www.drive5.com/usearch/manual/uparse_otu_radius.html).

4.4.4 Rarefaction choices

Some contention in the scientific community exists about the most useful application of rarefaction for statistical comparison of samples to assess diversity (Hughes and Hellmann 2005). To be able to quickly and easily assess all the samples in a dataset, most molecular ecologists have chosen to rarefy all samples to the depth of sequencing corresponding to the sample with the least number of reads, or to another equally arbitrary level which also results in loss of data (reviewed in Cárcer et al. 2011). Some workers propose a mixed model approach rather than the minimum sequenced depth. They argue that the rarefaction of data to even sequencing depth is inadmissible because it overlooks biologically valid data, though they comment that datasets with large enough sampling size that can withstand the loss of data may not necessarily apply to this disadvantage (McMurdie and Holmes 2014). Instead, McMurdie and Holmes promote analyzing microbiome data using differential abundances: "like differentially

expressed genes, a species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design". Although future works should consider incorporating different methods to correct for sampling depth disparities in microbiome datasets, the industry standard remains rarefaction to even sequencing depth, though adherence to typical statistical procedures may be an advantage to microbial ecologists moving forward.

5 Conclusion

This single sampling effort informs us that Arctic Ocean viral communities are extremely diverse, and are distinguishable by water mass or depth, though more samples would allow for further characterization of these water masses to validate environmental factors determining these viral community differences. The abundant genotypes within these communities are composed of a mixture of types geographically constrained to the Arctic Ocean and some that are globally distributed, the assemblage of which may be determined by the properties and history of the resident water mass (Galand et al. 2010). Although the prokaryotic communities and viral communities cluster with similar dendrogram topologies within each respective dataset, little can be said about the interactions between these populations without repeated sampling over time and co-culture of host isolates and their viruses.

It would appear that Illumina Miseq paired-end data is more comparable to results from the Roche/454 than the Ion Torrent PGM for our *g23* amplicon datasets, although broad patterns of beta diversity are for the most part conserved between platforms. Therefore, we recommend that future studies using viral fingerprinting to employ the Illumina platform for comparable results, with the knowledge that platform biases most likely exist (Bolotin et al. 2012; Frey et al. 2014).

The tools used in this study have the potential to aid in unraveling the ecological role of Arctic marine viruses in a changing climate. Because of the integral role of viruses in marine systems, knowledge about viruses in the Arctic Ocean may be vital to current monitoring and future mitigation efforts in the light of the inevitable changes the Arctic Ocean will experience in years to come. While this addition to basic science is valuable in itself as so little is known about Arctic Ocean viruses, it is hoped that the characterization of microbial biodiversity in polar water masses may act as a signal for downstream effects resulting from microbial community shifts.

6 Future work

The long-term intention of this work is to serve as a method development test for the MicroPolar project yearlong viral dataset taken in 2014, to which this information will be added for a broader future multi-seasonal investigation. Additional information on diversity of Arctic Ocean virus communities throughout the yearlong sampling period of the MicroPolar project will add to our knowledge of the marine viral community seasonal dynamics. Several previous examinations of viral diversity have shown that certain OTUs are more abundant in winter than in summer at the same location, and vice versa (Pagarete et al. 2013; Chow and Fuhrman 2012; Sandaa and Larsen 2006; Brum et al. 2015; Zhong and Jacquet 2014). Repeated sampling of this area may allow us to make characterizations of the current viral populations within Arctic Ocean water masses, which in turn may serve as a monitoring tool for the microbial communities that live there.

Future analyses of viral data collected throughout the annual cycle during the MicroPolar Project may also add to our knowledge about global diversity of microbes, and will inform further on the significance of dominant viral groups during winter in the Arctic Ocean. A recent study examining global biodiversity of marine bacteria found that species richness was highest during the winter months in temperate regions, with richness rising with increasing latitude in the Northern Hemisphere (Ladau et al. 2013). Shorter photoperiods also associated with higher bacterial richness globally in the model (Ladau et al. 2013). If host specificity of viruses is high, the viral community richness could also be elevated during winter.

Without a widely-used and curated viral signature gene database, taxonomic assignments of environmental viral sequences will remain difficult if not impossible. Although taxonomy of viruses may be a self-defeating endeavor, detailed metadata associated with sequence entries may prove helpful in the quest to understand viral biogeography.

Viromics has potential to be a more descriptive and higher resolution methodology to use when examining viral communities (Sullivan 2014). Together with transcriptomics, whole-sample sequencing could reveal more complete answers to the main questions posed by microbial ecologists; “who is there?” and “what are they doing?”. The viral research group within the MicroPolar project intends to use environmental sample metagenomic sequencing data in parallel with the gene fingerprinting techniques featured in this study to examine the broader viral population within samples collected throughout the annual cycle. In addition to molecular methods, culturing of more novel viruses from Arctic Ocean samples can also improve our current understanding of host-viral interactions and complement the information gleaned through omics.

7 References

Abedon, S.T. 2000. “Anecdotal , Historical and Critical Commentaries on Genetics: The Murky Origin of Snow White and Her T-Even Dwarfs.” *Genetics* 155 (2): 481–86.

Ackermann, H.W. 2007. “5500 Phages Examined in the Electron Microscope.” *Archives of Virology*. doi:10.1007/s00705-006-0849-1.

Agogué, H., D. Lamy, P. R. Neal, M.L. Sogin, and G. J, Herndl. 2011. “Water Mass-Specificity of Bacterial Communities in the North Atlantic Revealed by Massively Parallel Sequencing.” *Molecular Ecology* 20 (2): 258–74. doi:10.1111/j.1365-294X.2010.04932.x.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Anderson, M. J., and D. C. I. Walsh. 2013. “PERMANOVA, ANOSIM, and the Mantel Test in the Face of Heterogeneous Dispersions: What Null Hypothesis Are You Testing?” *Ecological Monographs* 83 (4). Ecological Society of America: 557–74. doi:10.1890/12-2010.1.

Andrews, S. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” *Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M.Haynes, S. Kelley, H. Liu, J.M. Mahaffy, J.E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C.A. Suttle, F. Rohwer. 2006. “The Marine Viromes of Four Oceanic Regions.” *PLoS Biology* 4 (11): 2121–31. doi:10.1371/journal.pbio.0040368

Arctic Council. 2013. “Arctic Ocean.” *The Columbia Encyclopedia*. Encyclopedia.com.

Arrigo, K. R. 1997. “Primary Production in Antarctic Sea Ice.” *Science* 276 (5311). American Association for the Advancement of Science: 394–97. doi:10.1126/science.276.5311.394.

Arrigo, K. R. 2013. “The Changing Arctic Ocean.” *Elementa: Science of the Anthropocene* 1 (1). BioOne: 000010. doi:10.12952/journal.elementa.000010.

Azam, F., T. Fenchel, J.G. Field, J.S. Gray, L.A. Meyer-Reil, and F. Thingstad. 1983. “The Ecological Role of Water-Column Microbes in the Sea .” *Marine Ecology Progress Series*. doi:10.3354/meps010257.

Baas-Becking, L. G. M. 1934. Geobiologie of Inleiding Tot de Milieukunde. The Hague, Neth.: Van Stockum and Zoon.

Bachy, C., P. López-García, A. Vereshchaka, and D. Moreira. 2011. “Diversity and Vertical Distribution of Microbial Eukaryotes in the Snow, Sea Ice and Seawater Near the North Pole at the End of the Polar Night.” *Frontiers in Microbiology* 2 (January): 106. doi:10.3389/fmicb.2011.00106.

Baldisserotto, C., L. Ferroni, C. Andreoli, M. P. Fasulo, A. Bonora, and S. Pancaldi. 2005. “Dark-Acclimation of the Chloroplast in Koliella Antarctica Exposed to a Simulated Austral Night Condition.” *Arctic, Antarctic, and Alpine Research* 37 (2). Institute of Arctic and Alpine Research (INSTAAR), University of Colorado: 146–56. doi:10.1657/1523-0430(2005)037[0146:DOTCIK]2.0.CO;2.

Balzer, S., K. Malde, and I. Jonassen. 2011. “Systematic Exploration of Error Sources in Pyrosequencing Flowgram Data.” *Bioinformatics (Oxford, England)* 27 (13): i304–9. doi:10.1093/bioinformatics/btr251.

Barnett, D. W, E. K Garrison, A.R. Quinlan, M. P. Strömberg, and G. T. Marth. 2011. “BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” *Bioinformatics (Oxford, England)* 27 (12): 1691–92. doi:10.1093/bioinformatics/btr174.

Baudoux, A.-C., and C.P.D. Brussaard. 2005. “Characterization of Different Viruses Infecting the Marine Harmful Algal Bloom Species Phaeocystis Globosa.” *Virology* 341 (1): 80–90. doi:10.1016/j.virol.2005.07.002.

Beaufort, L., I. Probert, T. de Garidel-Thoron, E. M. Bendif, D. Ruiz-Pino, N. Metzl, C. Goyet, N. Buchet, P. Coupel, M. Greland, B. Rost, R.E.M. Rickaby, C. De Vargas. 2011. “Sensitivity of Coccolithophores to Carbonate Chemistry and Ocean Acidification.” *Nature* 476 (7358). Nature Publishing Group: 80–83. doi:10.1038/nature10295.

Bell, E.M., and J. Laybourn-Parry. 2003. “Mixotrophy in the Antarctic Phytoflagellate, Pyramimonas gelidicola (Chlorophyta: Prasinophyceae)1” *Journal of Phycology* 39 (4): 644–49. doi:10.1046/j.1529-8817.2003.02152.x.

Bellas, C. M., and A. M. Anesio. 2013. “High Diversity and Potential Origins of T4-Type Bacteriophages on the Surface of Arctic Glaciers.” *Extremophiles : Life under Extreme Conditions* 17 (5): 861–70. doi:10.1007/s00792-013-0569-x.

Bench, S. R., T. E. Hanson, K. E. Williamson, D. Ghosh, M. Radosovich, K. Wang, and K. E. Wommack. 2007. “Metagenomic Characterization of Chesapeake Bay Virioplankton.” *Applied and Environmental Microbiology* 73 (23): 7629–41. doi:10.1128/AEM.00938-07.

Bergh, O., K. Y. Børsheim, G. Bratbak, and M. Heldal. 1989. “High Abundance of Viruses Found in Aquatic Environments.” *Nature* 340: 467–68. doi:10.1038/340467a0.

Blindheim, J. 1990. “Arctic Intermediate Water in the Norwegian Sea.” *Deep Sea Research Part A. Oceanographic Research Papers* 37 (9): 1475–89. doi:10.1016/0198-0149(90)90138-L.

Bolotin, D. A., I. Z. Mamedov, O. V. Britanova, I.V. Zvyagin, D. Shagin, S.V. Ustyugova, M.A. Turchaninova, S. Lukyanov, Y.B. Lebedev, and D. M. Chudakov. 2012. “Next Generation Sequencing for TCR Repertoire Profiling: Platform-Specific Features and Correction Algorithms.” *European Journal of Immunology* 42 (11): 3073–83. doi:10.1002/eji.201242517.

Borriss, M., E. Helmke, R. Hanschke, and T. Schweder. 2003. “Isolation and Characterization of Marine Psychrophilic Phage-Host Systems from Arctic Sea Ice.” *Extremophiles* 7 (5): 377–84. doi:10.1007/s00792-003-0334-7.

Bouvier, T., and P. A. del Giorgio. 2007. “Key Role of Selective Viral-Induced Mortality in Determining Marine Bacterial Community Composition.” *Environmental Microbiology* 9 (2): 287–97. doi:10.1111/j.1462-2920.2006.01137.x.

Boycott, A.E. 1928. “The Transition from Live to Dead: The Nature of Filtrable Viruses.” *Journal of the Royal Society of Medicine* 22 (1). SAGE Publications: 55–69. doi:10.1177/003591572802200121.

Boyer, M., M.-A. Madoui, G. Gimenez, B. La Scola, and D. Raoult. 2010. “Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4 Domain of Life Including Giant Viruses.” *PloS One* 5 (12). Public Library of Science: e15530. doi:10.1371/journal.pone.0015530.

Bratbak, G., J. K. Egge, and M. Heldal. 1993. “Viral Mortality of the Marine Alga Emiliania Huxleyi (Haptophyceae) and Termination of Algal Blooms.” *Marine Ecology Progress Series*. doi:10.3354/meps093039.

Breitbart, M., J.H. Miyake, and F. Rohwer. 2004. “Global Distribution of Nearly Identical Phage-Encoded DNA Sequences.” *FEMS Microbiology Letters* 236: 249–56. doi:10.1016/j.femsle.2004.05.042.

Breitbart, M., and F. Rohwer. 2005. “Here a Virus, There a Virus, Everywhere the Same Virus?” *Trends in Microbiology* 13 (6): 278–84. doi:10.1016/j.tim.2005.04.003.

Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer. 2002. “Genomic Analysis of Uncultured Marine Viral Communities.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14250–55. doi:10.1073/pnas.202488399.

Brochier-Armanet, C., Bastien B., S. Gribaldo, and P. Forterre. 2008. “Mesophilic Crenarchaeota: Proposal for a Third Archaeal Phylum, the Thaumarchaeota.” *Nature Reviews. Microbiology* 6 (3): 245–52. doi:10.1038/nrmicro1852.

Brum, J. R., B.L. Hurwitz, O. Schofield, H. W. Ducklow, and M. B. Sullivan. 2015. “Seasonal Time Bombs: Dominant Temperate Viruses Affect Southern Ocean Microbial Dynamics.” *The ISME Journal*, August. doi:10.1038/ismej.2015.125.

Brum, J. R, J C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. de Vargas, J.M. Gasol, G. Gorsky, A.C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B.T. Poulos, S.M.S. Schwenck, S. Speich, C. Dimier, S.Kandels-Lewis, M. Picheral, S. Searson, Tara Oceans Coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, M.B. Sullivan “Patterns and Ecological Drivers of Ocean Viral Communities.” *Science (New York, N.Y.)* 348 (6237): 1261498. doi:10.1126/science.1261498.

Brussaard, C. P. D. 2004. “Viral Control of Phytoplankton Populations-a Review1.” *The Journal of Eukaryotic Microbiology* 51 (2): 125–38. doi:10.1111/j.1550-7408.2004.tb00537.x.

Buckling, A., and P. B. Rainey. 2002. “Antagonistic Coevolution between a Bacterium and a Bacteriophage.” *Proceedings. Biological Sciences / The Royal Society* 269 (1494): 931–36. doi:10.1098/rspb.2001.1945.

Bushnell, B. 2014. “BBMap.” http://sourceforge.net/projects/bbmap/.

Butina, T. V., O.I. Belykh, S.A. Potapov, and E.G. Sorokovikova. 2013. “Diversity of the Major Capsid Genes (*g23*) of T4-like Bacteriophages in the Eutrophic Lake Kotokel in East Siberia, Russia.” *Archives of Microbiology* 195 (7): 513–20. doi:10.1007/s00203-013-0884-8.

Cárcer, D. A. d., S.E. Denman, C. McSweeney, and M. Morrison. 2011. “Evaluation of Subsampling-Based Normalization Strategies for Tagged High-Throughput Sequencing Data Sets from Gut Microbiomes.” *Applied and Environmental Microbiology* 77 (24): 8795–98. doi:10.1128/AEM.05491-11.

Chao, A. 1984. “Nonparametric Estimation of the Number of Classes in a Population.” *Scandinavian Journal of Statistics* 11 (4): 265–70.

Chenard, C., and C. A. Suttle. 2008. “Phylogenetic Diversity of Cyanophage Photosynthetic Genes (psbA) in Marine and Fresh Waters.” *Applied and Environmental Microbiology* 74 (17): 5317–24. doi:10.1128/AEM.02480-07.

Chibani-Chennoufi, S., A. Bruttin, M.-L. Dillmann, and H. Brüssow. 2004. “Phage-Host Interaction: An Ecological Perspective.” *Journal of Bacteriology* 186 (12): 3677–86. doi:10.1128/JB.186.12.3677-3686.2004.

Chow, C.-E. T., and J. A. Fuhrman. 2012. “Seasonality and Monthly Dynamics of Marine Myovirus Communities.” *Environmental Microbiology* 14: 2171–83. doi:10.1111/j.1462-2920.2012.02744.x.

Chow, C.-E. T., D.Y. Kim, R. Sachdeva, D. A Caron, and J. A Fuhrman. 2014. “Top-down Controls on Bacterial Community Structure: Microbial Network Analysis of Bacteria, T4-like Viruses and Protists.” *The ISME Journal* 8 (4). International Society for Microbial Ecology: 816–29. doi:10.1038/ismej.2013.199.

Claesson, M. J., Q. Wang, O. O’Sullivan, R. Greene-Diniz, J. R . Cole, R. P. Ross, and P. W. O’Toole. 2010. “Comparison of Two next-Generation Sequencing Technologies for Resolving Highly Complex Microbiota Composition Using Tandem Variable 16S rRNA Gene Regions.” *Nucleic Acids Research* 38 (22). Oxford University Press: e200. doi:10.1093/nar/gkq873.

Clasen, J.L., S. M. Brigden, J. P. Payet, and C. A. Suttle. 2008. “Evidence That Viral Abundance across Oceans and Lakes Is Driven by Different Biological Factors.” *Freshwater Biology* 53 (6): 1090–1100. doi:10.1111/j.1365-2427.2008.01992.x.

Claverie, J.-M., and C. Abergel. 2013. “Open Questions about Giant Viruses.” *Advances in Virus Research* 85 (January): 25–56. doi:10.1016/B978-0-12-408116-1.00002-1.

Connelly, T. L., C. M. Tilburg, and P. L. Yager. 2006. “Evidence for Psychrophiles Outnumbering Psychrotolerant Marine Bacteria in the Springtime Coastal Arctic.” *Limnology and Oceanography*. doi:10.4319/lo.2006.51.2.1205.

Conover, R.J., and M. Huntley. 1991. “Copepods in Ice-Covered seas—Distribution, Adaptations to Seasonally Limited Food, Metabolism, Growth Patterns and Life Cycle Strategies in Polar Seas.” *Journal of Marine Systems* 2 (1-2): 1–41. doi:10.1016/0924-7963(91)90011-I.

Cook, J. (Woods Hole Oceanographic Institute). 2015. “Arctic Ocean Circulation.” Accessed August 5. http://www.whoi.edu/main/topic/arctic-ocean-circulation.

Cottrell, M. T., and D. L. Kirchman. 2003. “Contribution of Major Bacterial Groups to Bacterial Biomass Production (thymidine and Leucine Incorporation) in the Delaware Estuary.” *Limnology and Oceanography*. doi:10.4319/lo.2003.48.1.0168.

Cottrell, M. T., and C.A. Suttle. 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: 1–9. doi:http://dx.doi.org/10.3354/meps078001.

Cottrell, M. T., and C. A. Suttle. 1995. “Dynamics of Lytic Virus Infecting the Photosynthetic Marine Picoflagellate Micromonas Pusilla.” *Limnology and Oceanography* 40 (4): 730–39. doi:10.4319/lo.1995.40.4.0730.

Danovaro, R., C. Corinaldesi, A. Dell’Anno, J. A. Fuhrman, J. J. Middelburg, R. T. Noble, and C. A. Suttle. 2011. “Marine Viruses and Global Climate Change.” *FEMS Microbiology Reviews*. doi:10.1111/j.1574-6976.2010.00258.x.

Darling, K. F., C. M. Wade, I. A. Stewart, D. Kroon, R. Dingle, and A. J. Brown. 2000. “Molecular Evidence for Genetic Mixing of Arctic and Antarctic Subpolar Populations of Planktonic Foraminifers.” *Nature* 405 (6782). Macmillian Magazines Ltd.: 43–47. doi:10.1038/35011002.

Del Fabbro C, S. Scalabrin, M. Morgante, F.M. Giorgi (2013) An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE 8(12): e85024. doi:10.1371/journal.pone.0085024

Delisle, A. L., and R. E. Levin. 1969. “Bacteriophages of Psychrophilic Pseudomonads. II. Host Range of Phage Active againstPseudomonas Putrefaciens.” *Antonie van Leeuwenhoek* 35 (1): 318–24. doi:10.1007/BF02219152.

Dennehy, J. J. 2013. “What Ecologists Can Tell Virologists.” *Annual Review of Microbiology*, no. May: 117–35. doi:10.1146/annurev-micro-091313-103436.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. “Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB.” *Appl Environ Microbiol* 72: 5069–72.

Drake, John W., and J.J. Holland. 1999. “Mutation Rates among RNA Viruses.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (24): 13910–13. doi:10.1073/pnas.96.24.13910.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. “Rates of Spontaneous Mutation.” *Genetics* 148 (4): 1667–86. doi: 610966.

Edgar, R. C. 2011. “Usearch User Guide 5.0.” doi:10.1016/S0022-3913(12)00047-9.

Edgar, R. C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucl. Acids Res.* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Edgar, R. C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics* 26 (19): 2460–61. doi:10.1093/bioinformatics/btq461.

Edvardsen, B., and E Paasche. 1998. “Bloom Dynamics and Physiology of Prymnesium and Chrysochromulina.” *NATO ASI Series, G Ecological Sciences* 41: 193–208.

Eppley, R. W., R.W. Holmes, and J. D.H. Strickland. 1967. “Sinking Rates of Marine Phytoplankton Measured with a Fluorometer.” *Journal of Experimental Marine Biology and Ecology* 1 (2): 191–208. doi:10.1016/0022-0981(67)90014-7.

Collins, E. R., G. Rocap, and J. W. Deming. 2010. “Persistence of Bacterial and Archaeal Communities in Sea Ice through an Arctic Winter.” *Environmental Microbiology* 12 (7): 1828–41.

Field, D., B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston. 2006. “Open Software for Biologists: From Famine to Feast.” *Nature Biotechnology* 24 (7). Nature Publishing Group: 801–3. doi:10.1038/nbt0706-801.

Filée, J., F. Tétart, C. A. Suttle, and H. M. Krisch. 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): 12471–76. doi:10.1073/pnas.0503404102.

Fischer, M. G., M. J. Allen, W. H. Wilson, and C. A. Suttle. 2010. “Giant Virus with a Remarkable Complement of Genes Infects Marine Zooplankton.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (45): 19508–13. doi:10.1073/pnas.1007615107.

Fischer, M. G., and C. A. Suttle. 2011. “A Virophage at the Origin of Large DNA Transposons.” *Science (New York, N.Y.)* 332: 231–34. doi:10.1126/science.1199412.

Thingstad, T.F., E. Strand, and A. Larsen. 2010. “Stepwise Building of Plankton Functional Type (PFT) Models: A Feasible Route to Complex Models?” *Progress in Oceanography* 84 (1-2): 6–15. doi:10.1016/j.pocean.2009.09.001.

Frey, K. G., J. E.Herrera-Galeano, C. L. Redden, T. V. Luu, S. L. Servetas, A. J. Mateczun, V. P. Mokashi, and K. A. Bishop-Lilly. 2014. “Comparison of Three next-Generation Sequencing Platforms for Metagenomic Sequencing and Identification of Pathogens in Blood.” *BMC Genomics* 15 (1): 96. doi:10.1186/1471-2164-15-96.

Fu, Y., K. F. Keats, R. B. Rivkin, and A. S. Lang. 2013. “Water Mass and Depth Determine the Distribution and Diversity of Rhodobacterales in an Arctic Marine System.” *FEMS Microbiology Ecology* 84 (3): 564–76. doi:10.1111/1574-6941.12085.

Fuellgrabe, M.W., D. Herrmann, H. Knecht, S. Kuenzel, M. Kneba, C. Pott, and M. Brüggemann. 2015. “High-Throughput, Amplicon-Based Sequencing of the CREBBP Gene as a Tool to Develop a Universal Platform-Independent Assay.” Edited by Ramy K. Aziz. *PLOS ONE* 10 (6). Public Library of Science: e0129195. doi:10.1371/journal.pone.0129195.

Fuhrman, J. A. 1999. “Marine Viruses and Their Biogeochemical and Ecological Effects.” *Nature* 399: 541–48. doi:10.1038/21119.

Fuhrman, J. A., and M. Schwalbach. 2003. “Viral Influence on Aquatic Bacterial Communities.” In *Biological Bulletin*, 204:192–95.

Fujihara, S., J. Murase, C. C. Tun, T. Matsuyama, M. Ikenaga, S. Asakawa, and M. Kimura. 2010. “Low Diversity of T4-Type Bacteriophages in Applied Rice Straw, Plant Residues and Rice Roots in Japanese Rice Soils: Estimation from Major Capsid Gene (*g23*) Composition.” *Soil Science and Plant Nutrition* 56 (6): 800–812. doi:10.1111/j.1747-0765.2010.00513.x.

Galand, P. E., E.O. Casamayor, D. L. Kirchman, and C. Lovejoy. 2009. “Ecology of the Rare Microbial Biosphere of the Arctic Ocean.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (52): 22427–32. doi:10.1073/pnas.0908284106.

Galand, P. E., M. Potvin, E.O. Casamayor, and C. Lovejoy. 2010. “Hydrography Shapes Bacterial Biogeography of the Deep Arctic Ocean.” *The ISME Journal* 4 (4). Nature Publishing Group: 564–76. doi:10.1038/ismej.2009.134.

Gast, R. J., Z. M. McKie-Krisberg, S. A. Fay, J. M. Rose, and R. W. Sanders. 2014. “Antarctic Mixotrophic Protist Abundances by Microscopy and Molecular Methods.” *FEMS Microbiology Ecology* 89 (2). The Oxford University Press: 388–401. doi:10.1111/1574-6941.12334.

Ghiglione, J.-F., P. E. Galand, T. Pommier, C. Pedros-Alio, E. W. Maas, K. Bakker, S. Bertilson, D. L. Kirchman, C. Lovejoy, P.L. Yager, and A. E. Murray "Pole-to-Pole Biogeography of Surface and Deep Marine Bacterial Communities." *Proceedings of the National Academy of Sciences* 109 (43): 17633–38. doi:10.1073/pnas.1208160109.

Glenn, T. C. 2011. "Field Guide to next-Generation DNA Sequencers." *Molecular Ecology Resources* 11 (5): 759–69. doi:10.1111/j.1755-0998.2011.03024.x.

Goldsmith, D. B., G. Crosti, B. Dwivedi, L. D. McDaniel, A. Varsani, C.A. Suttle, M. G. Weinbauer, R. -A. Sandaa, and M. Breitbart. 2011. "Development of *phoH* as a Novel Signature Gene for Assessing Marine Phage Diversity." *Applied and Environmental Microbiology* 77 (21): 7730–39. doi:10.1128/AEM.05531-11.

Goldsmith, D. B., R. J. Parsons, D. Beyene, P. Salamon, and M. Breitbart. 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* 3: e997. doi:10.7717/peerj.997.

Gómez-Pereira, P. R., B. M. Fuchs, C. Alonso, M. J. Oliver, J. E. E. van Beusekom, and R.Amann. 2010. "Distinct Flavobacterial Communities in Contrasting Water Masses of the North Atlantic Ocean." *The ISME Journal* 4 (4): 472–87. doi:10.1038/ismej.2009.142.

Gradinger, R. 2015. "Competition within the Marine Microalgae over the Polar Dark Period in the Greenland Sea of High Arctic." *Acta Oceanologica Sinica*. http://www.sciencemeta.com/index.php/HYXBEN/article/view/372540.

Huang, S., S. Zhang, N. Jiao, and F.Chen. 2015. "Marine Cyanophages Demonstrate Biogeographic Patterns throughout the Global Ocean." *Applied and Environmental Microbiology* 81 (1): 441–52. doi:10.1128/AEM.02483-14.

Hughes, J. B, and J.J Hellmann. 2005. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity." *Methods in Enzymology* 397 (January): 292–308. doi:10.1016/S0076-6879(05)97017-1.

Illumina. 2010. "Technology Spotlight: Illumina Sequencing Technology." https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

IPCC. 2001. Climate Change 2001: The Scientific Basis. doi:10.1256/004316502320517344.

Iversen, K. R., Seuthe, L. 2011. "Seasonal microbial processes in a high-latitude fjord (Kongsfjorden, Svalbard): I. Heterotrophic bacteria, picoplankton and nanoflagellates." *Polar Biology* (34): 731 – 749. doi:10.1007/s00300-010-0929-2.

Iyer, L.M., S Balaji, E. V. Koonin, and L. Aravind. 2006. "Evolutionary Genomics of Nucleo-Cytoplasmic Large DNA Viruses." *Virus Research* 117 (1): 156–84. doi:10.1016/j.virusres.2006.01.009.

Jacobsen, A., G.Bratbak, and M.Heldal. 1996.
"Isolation and Characterization of a virus infecting Phaeocystis pouchetii (Prymnesiophyceae)"*Journal of Phycology* 32 (6): 923–27. doi:10.1111/j.0022-3646.1996.00923.x.

Jameson, E., N.H. Mann, I. Joint, C. Sambles, and M. Mühling. 2011. "The Diversity of Cyanomyovirus Populations along a North-South Atlantic Ocean Transect." *The ISME Journal* 5 (11). International Society for Microbial Ecology: 1713–21. doi:10.1038/ismej.2011.54.

Johannessen, T. V., G. Bratbak, A. Larsen, H.Ogata, E. S. Egge, B. Edvardsen, W. Eikrem, and R.-A. Sandaa. 2015. "Characterisation of Three Novel Giant Viruses Reveals Huge Diversity among Viruses Infecting Prymnesiales (Haptophyta)." *Virology* 476 (February): 180–88. doi:10.1016/j.virol.2014.12.014.

John, S. G., C. B. Mendez, L. Deng, B. Poulos, A. K. M. Kauffman, S. Kern, J. Brum, M. F. Polz, E. A. Boyle, and M. B. Sullivan. 2011. "A Simple and Efficient Method for Concentration of Ocean Viruses by Chemical Flocculation." *Environmental Microbiology Reports* 3 (2): 195–202. doi:10.1111/j.1758-2229.2010.00208.x.

Jover, L. F., C. Effler, A. Buchan, S. W. Wilhelm, and J. S. Weitz. 2014. “The Elemental Composition of Virus Particles: Implications for Marine Biogeochemical Cycles.” *Nature Reviews. Microbiology* 12 (7). Nature Publishing Group: 519–28. doi:10.1038/nrmicro3289.

Jünemann, S., F. J. Sedlazeck, K. Prior, A. Albersmeier, U. John, J. Kalinowski, A. Mellmann, A. Goesmann, A.von Haeseler, and J. Stoye. 2013. “Updating Benchtop Sequencing Performance Comparison.” *Nature Biotechnology* 31 (4): 294–96. doi:10.1038/nbt.2522.

Kazakov, A.E., O.Vassieva, M. S. Gelfand, A. Osterman, and R.Overbeek. 2003. “Bioinformatics Classification and Functional Analysis of *PhoH* Homologs.” *In Silico Biology* 3 (1-2): 3–15.

Kim, M.-S., T. W. Whon, and J.-W.Bae. 2013. “Comparative Viral Metagenomics of Environmental Samples from Korea.” *Genomics & Informatics* 11 (3): 121–28. doi:10.5808/GI.2013.11.3.121.

Kirchman, D. L., X. A.G. Morán, and H. Ducklow. 2009. “Microbial Growth in the Polar Oceans — Role of Temperature and Potential Impact of Climate Change.” *Nature Reviews Microbiology* 7 (6): 451–59. doi:10.1038/nrmicro2115.

Kirchman, D. L., M. T. Cottrell, and C. Lovejoy. 2010. “The Structure of Bacterial Communities in the Western Arctic Ocean as Revealed by Pyrosequencing of 16S rRNA Genes.” *Environmental Microbiology* 12 (5): 1132–43. doi:10.1111/j.1462-2920.2010.02154.x.

Koonin, E. V., and V. V. Dolja. 2013. “A Virocentric Perspective on the Evolution of Life.” *Current Opinion in Virology* 3 (5): 546–57. doi:10.1016/j.coviro.2013.06.008.

La Scola, B., S.Audic, C. Robert, L. Jungang, X. de Lamballerie, M. Drancourt, R. Birtles, J.-M. Claverie, and D. Raoult. 2003. “A Giant Virus in Amoebae.” *Science (New York, N.Y.)* 299 (5615): 2033. doi:10.1126/science.1081867.

La Scola, B., C. Desnues, I. Pagnier, C. Robert, L. Barrassi, G. Fournous, M. Merchat, M. Suzan-Monti, P. Forterre, E. V. Koonin, D. Raoult. 2008. “The Virophage as a Unique Parasite of the Giant Mimivirus.” *Nature* 455: 100–104. doi:10.1038/nature07218.

Ladau, J., T. J. Sharpton, M. M Finucane, G. Jospin, S. W. Kembel, J. O’Dwyer, A. F Koeppel, J. L Green, and K. S. Pollard. 2013. “Global Marine Bacterial Diversity Peaks at High Latitudes in Winter.” *The ISME Journal* 7 (9). International Society for Microbial Ecology: 1669–77. doi:10.1038/ismej.2013.37.

Larsen, A., J. K. Egge, J. C. Nejstgaard, I. Di Capua, R.Thyrhaug, G. Bratbak, and F. Thingstad. 2015. “Contrasting Response to Nutrient Manipulation in Arctic Mesocosms Are Reproduced by a Minimum Microbial Food Web Model.” *Limnology and Oceanography* 60 (2): 360–74. doi:10.1002/lno.10025.

Larsen, A., R.-A. Sandaa, C.P.D. Brussaard, J. Egge, M. Heldal, A. Paulino, R. Thyrhaug, E.J. Vanhannen, and G. Bratbak. 2001. “Population Dynamics and Diversity of Phytoplankton, Bacteria and Viruses in a Seawater Enclosure.” *Marine Ecology Progress Series* 221: 47–57.

Larsen, J. B., A. Larsen, G. Bratbak, and R. A. Sandaa. 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): 3048–57. doi:10.1128/AEM.02548-07.

Lawrence, J. E. 2008. “Furtive Foes: Algal Viruses as Potential Invaders.” *ICES Journal of Marine Science* 65 (5): 716–22. doi:10.1093/icesjms/fsn024.

Laybourn-Parry, J., W. A. Marshall, and N. J. Madan. 2007. “Viral Dynamics and Patterns of Lysogeny in Saline Antarctic Lakes.” *Polar Biology* 30 (3): 351–58. doi:10.1007/s00300-006-0191-9.

Li, S., S. W. Tighe, C. M. Nicolet, D. Grove, S. Levy, W. Farmerie, A. Viale, C. Wright, P. A. Scheitzer, Y. Gao, D. Kim, J. Boland, B. Hicks, R. Kim, S. Chhangawala, N. Jafari, N. Raghavachari, J. Gandara, N. Garcia-Reyero, C. Hendrickson, D. Roberson, J.A. Rosenfeld, T. Smith, J. G. Underwood, M. Wang, P. Zumbo, D.A. Baldwin, G.S. Grills and C.E. Mason. 2014. “Multi-Platform Assessment of Transcriptome Profiling Using RNA-Seq in the ABRF next-Generation Sequencing Study.” *Nature Biotechnology* 32 (9): 915–25. doi:10.1038/nbt.2972.

Lindell, D., J. D Jaffe, Z. I. Johnson, G. M Church, and S. W Chisholm. 2005. “Photosynthesis Genes in Marine Viruses Yield Proteins during Host Infection.” *Nature* 438 (7064): 86–89. doi:10.1038/nature04111.

Lindström, E. S., and S. Langenheder. 2012. “Local and Regional Factors Influencing Bacterial Community Assembly.” *Environmental Microbiology Reports* 4 (1): 1–9. doi:10.1111/j.1758-2229.2011.00257.x.

Liu, H., I. Probert, J. Uitz, H. Claustre, S. Aris-Brosou, M. Frada, F. Not, and C. de Vargas. 2009. “Extreme Diversity in Noncalcifying Haptophytes Explains a Major Pigment Paradox in Open Oceans.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (31): 12803–8. doi:10.1073/pnas.0905841106.

Liu, J., G. Wang, Q. Wang, J. Jin, and X. Liu. 2012. “Phylogenetic Diversity and Assemblage of Major Capsid Genes (*g23*) of T4-Type Bacteriophages in Paddy Field Soils during Rice Growth Season in Northeast China.” *Soil Science and Plant Nutrition* 58 (4): 435–44. doi:10.1080/00380768.2012.703610.

Logares, R., T. H. A. Haverkamp, S.Kumar, A. Lanzén, A. J. Nederbragt, C. Quince, and H. Kauserud. 2012. “Environmental Microbiology through the Lens of High-Throughput DNA Sequencing: Synopsis of Current Platforms and Bioinformatics Approaches.” *Journal of Microbiological Methods*. doi:10.1016/j.mimet.2012.07.017.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. 2012. “Performance Comparison of Benchtop High-Throughput Sequencing Platforms.” *Nature Biotechnology* 30 (5). doi:10.1038/nbt.2198.

Loman, N. J., C. Constantinidou, J. Z. M. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson, and M. J. Pallen. 2012. “High-Throughput Bacterial Genome Sequencing: An Embarrassment of Choice, a World of Opportunity.” *Nature Reviews Microbiology*. doi:10.1038/nrmicro2850.

Longnecker, K., M. J. Wilson, E. B. Sherr, and B. F. Sherr. 2010. “Effect of Top-down Control on Cell-Specific Activity and Diversity of Active Marine Bacterioplankton.” *Aquatic Microbial Ecology* 58 (2): 153–65. doi:10.3354/ame01366.

Lovejoy, C., R. Massana, and C. Pedro. 2006. “Diversity and Distribution of Marine Microbial Eukaryotes in the Arctic Ocean and Adjacent Seas Diversity and Distribution of Marine Microbial Eukaryotes in the Arctic Ocean and Adjacent Seas †‡.” *Applied and Environmental Microbiology* 72 (5): 3085–95. doi:10.1128/AEM.72.5.3085.

Lovejoy, C., W. F. Vincent, S. Bonilla, S. Roy, M. J. Martineau, R. Terrado, M. Potvin, R. Massana, and C. Pedrós-Alió. 2007. “Distribution, Phylogeny, and Growth of Cold-Adapted Picoprasinophytes in Arctic Seas.” *Journal of Phycology* 43 (1): 78–89. doi:10.1111/j.1529-8817.2006.00310.x.

Lozupone, C., and R. Knight. 2005. “UniFrac: A New Phylogenetic Method for Comparing Microbial Communities.” *Applied and Environmental Microbiology* 71 (12): 8228–35. doi:10.1128/AEM.71.12.8228-8235.2005.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis. 2012. “Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample.” *PloS One* 7 (2): e30087. doi:10.1371/journal.pone.0030087.

Lwoff, A. 1953. “Lysogeny.” *Microbiol. Mol. Biol. Rev.* 17 (4): 269–337.

Madigan, MT. 2012. Brock Biology of Microorganisms, 13th Edn. International Microbiology. doi:10.1016/B978-1-4832-3136-5.50010-3.

Maranger, R., D. R. Bird, and S. K. Juniper. 1994. “Viral and Bacterial Dynamics in Arctic Sea Ice during the Spring Algal Bloom near Resolute, N.W.T., Canada.” *Marine Ecology Progress Series* 111 (1-2): 121–27. doi:10.3354/meps111121.

Massana, R., E. F. DeLong, and C. Pedros-Alio. 2000. “A Few Cosmopolitan Phylotypes Dominate Planktonic Archaeal Assemblages in Widely Different Oceanic Provinces.” *Applied and Environmental Microbiology* 66 (5): 1777–87. doi:10.1128/AEM.66.5.1777-1787.2000.

McCallum, H., D.Harvell, and A. Dobson. 2003. “Rates of Spread of Marine Pathogens.” *Ecology Letters* 6 (12): 1062–67. doi:10.1046/j.1461-0248.2003.00545.x.

McKie-Krisberg, Z. M., and R. W. Sanders. 2014. “Phagotrophy by the Picoeukaryotic Green Alga Micromonas: Implications for Arctic Oceans.” *The ISME Journal* 8 (10). International Society for Microbial Ecology: 1953–61. doi:10.1038/ismej.2014.16.

McMurdie, P. J., and S. Holmes. 2014. “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.” *PLoS Computational Biology* 10 (4): e1003531. doi:10.1371/journal.pcbi.1003531.

Mikkelsen, D. M., S. Rysgaard, and R. N. Glud. 2008. “Microalgal Composition and Primary Production in Arctic Sea Ice: A Seasonal Study from Kobbefjord (Kangerluarsunnguaq), West Greenland.” *Marine Ecology Progress Series* 368: 65–74. doi:10.3354/meps07627.

Millard, A., M. R. J. Clokie, D. A. Shub, and N. H. Mann. 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): 11007–12. doi:10.1073/pnas.0401478101.

Monier, A., J. Comte, M. Babin, A. Forest, A. Matsuoka, and C. Lovejoy. 2014. “Oceanographic Structure Drives the Assembly Processes of Microbial Eukaryotic Communities ´,” 1–13. doi:10.1038/ismej.2014.197.

Montresor, M., C. Lovejoy, L. Orsini, G. Procaccini, and S. Roy. 2003. “Bipolar Distribution of the Cyst-Forming Dinoflagellate Polarella Glacialis.” *Polar Biology* 26 (3). Springer-Verlag: 186–94. doi:10.1007/s00300-002-0473-9.

Mühling, M., N. J. Fuller, A. Millard, P. J. Somerfield, D. Marie, W. H. Wilson, D. J. Scanlan, A. F. Post, I. Joint, and N. H. Mann. 2005. “Genetic Diversity of Marine Synechococcus and Co-Occurring Cyanophage Communities: Evidence for Viral Control of Phytoplankton.” *Environmental Microbiology* 7 (4): 499–508. doi:10.1111/j.1462-2920.2005.00713.x.

Munn, C. 2011. *Marine Microbiology: Ecology and Applications*. 2nd ed. Garland Science, Taylor & Francis Group, LLC.

Nasir, A., and G. Caetano-Anolles. 2015. “A Phylogenomic Data-Driven Exploration of Viral Origins and Evolution.” *Science Advances* 1 (8). American Association for the Advancement of Science: e1500527–e1500527. doi:10.1126/sciadv.1500527.

Needham, D. M., C.-E. T. Chow, J..A Cram, R. Sachdeva, A. Parada, and J. A. Fuhrman. 2013. “Short-Term Observations of Marine Bacterial and Viral Communities: Patterns, Connections and Resilience.” *The ISME Journal* 7 (7). International Society for Microbial Ecology: 1274–85. doi:10.1038/ismej.2013.19.

New England BioLabs Inc. 2015. “PCR Protocol for Taq DNA Polymerase with Standard Taq Buffer (M0273).” https://www.neb.com/protocols/1/01/01/taq-dna-polymerase-with-standard-taq-buffer-m0273.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O’Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2013. “Package ‘vegan.’” *R Package Ver. 2.0–8*.

Pagarete, A., C. E T Chow, T. Johannessen, J. A. Fuhrman, T. F. Thingstad, and R. -A. Sandaa. 2013. “Strong Seasonality and Interannual Recurrence in Marine Myovirus Communities.” *Applied and Environmental Microbiology* 79 (20): 6253–59. doi:10.1128/AEM.01075-13.

Payet, J. P., and C. A. Suttle. 2014. “Viral Infection of Bacteria and Phytoplankton in the Arctic Ocean as Viewed through the Lens of Fingerprint Analysis.” *Aquatic Microbial Ecology* 72 (1): 47–U100. doi:10.3354/ame01684.

Pennisi, E. 1998. “Genome Data Shake Tree of Life.” *Science (New York, N.Y.)*. doi:10.1126/science.280.5364.672.

Perrette, M., A. Yool, G. D. Quartly, and E. E. Popova. 2011. “Near-Ubiquity of Ice-Edge Blooms in the Arctic.” *Biogeosciences* 8 (2): 515–24. doi:10.5194/bg-8-515-2011.

Pinto, A. J., and L. Raskin. 2012. “PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets.” *PloS One* 7 (8). Public Library of Science: e43093. doi:10.1371/journal.pone.0043093.

Polz, M. F., and C. M. Cavanaugh. 1998. “Bias in Template-to-Product Ratios in Multitemplate PCR.” *Applied and Environmental Microbiology* 64 (10): 3724–30. http://aem.asm.org/content/64/10/3724.full.

Pomeroy, L .R., and D Deibel. 1986. “Temperature Regulation of Bacterial Activity during the Spring Bloom in Newfoundland Coastal Waters.” *Science (New York, N.Y.)* 233 (4761): 359–61. doi:10.1126/science.233.4761.359.

Pommier, T., J. Pinhassi, and A. Hagstroem. 2005. “Biogeographic Analysis of Ribosomal RNA Clusters from Marine Bacterioplankton.” *Aquatic Microbial Ecology* 41 (1): 79–89.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. 2007. “SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” *Nucleic Acids Research* 35 (21): 7188–96. doi:10.1093/nar/gkm864.

R Development Core Team, R. 2011. “R: A Language and Environment for Statistical Computing.” Edited by R Development Core Team. *R Foundation for Statistical Computing*. R Foundation for Statistical Computing. R Foundation for Statistical Computing. doi:10.1007/978-3-540-74686-7.

Rambaut, A. 2008. “FigTree v1.1.1: Tree Figure Drawing Tool.” http://tree.bio.ed.ac.uk/software/figtree/.

Raven, J, K Caldeira, and H Elderfield. 2005. *Ocean Acidification due to Increasing Atmospheric Carbon Dioxide*. The Royal Society.

Rich, J., M. Gosselin, E. Sherr, B. Sherr, and D. L. Kirchman. 1997. “High Bacterial Production, Uptake and Concentrations of Dissolved Organic Matter in the Central Arctic Ocean.” *Deep Sea Research Part II: Topical Studies in Oceanography* 44 (8): 1645–63. doi:10.1016/S0967-0645(97)00058-1.

Rodriguez-Brito, B., L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, J. Buchanan, C. Desnues, E. Dinsdale, R. Edwards, B. Felts, M. Haynes, H. Liu, D. Lipson, J. Mahaffy, A. B. Martin-Cuadrado, A. Mira, J. Nulton, L. Pasic, S. Rayhawk, J. Rodriguez-Mueller, F. Rodriguez-Valera, P. Salamon, S. Srinagesh, T. F. Thingstad, T. Tran, R. V. Thurber, D. Willner, M. Youle, F. Rohwer. 2010. “Viral and Microbial Community Dynamics in Four Aquatic Environments.” *The ISME Journal* 4 (6): 739–51. doi:10.1038/ismej.2010.1.

Rohwer, F., A. Segall, G. Steward, V. Seguritan, M. Breitbart, F. Wolven, and F. Azam. 2000. “The Complete Genomic Sequence of the Marine Phage Roseophage SIO1 Shares Homology with Nonmarine Phages.” *Limnology and Oceanography* 45 (2): 408–18. doi:10.4319/lo.2000.45.2.0408.

Rose, J. M., and D. A. Caron. 2007. “Does Low Temperature Constrain the Growth Rates of Heterotrophic Protists? Evidence and Implications for Algal Blooms in Cold Waters.” *Limnology and Oceanography* 52 (2): 886–95. doi:10.4319/lo.2007.52.2.0886.

Różańska, M., M. Poulin, and M. Gosselin. 2008. “Protist Entrapment in Newly Formed Sea Ice in the Coastal Arctic Ocean.” *Journal of Marine Systems* 74 (3-4): 887–901. doi:10.1016/j.jmarsys.2007.11.009.

Rudels, B., A.-M. Larsson, and P.-I. Sehlstedt. 1991. “Stratification and Water Mass Formation in the Arctic Ocean: Some Implications for the Nutrient Distribution.” *Polar Research*, 19–32. doi:http://dx.doi.org/10.3402/polar.v10i1.6724.

Salipante, S. J., T. Kawashima, C. Rosenthal, D. R. Hoogestraat, L. A. Cummings, D. J. Sengupta, T. T. Harkins, B.T. Cookson, and N. G. Hoffman. 2014. “Performance Comparison of Illumina and Ion Torrent next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling.” *Appl. Environ. Microbiol.* 80 (24): 7583–91. doi:10.1128/AEM.02206-14.

Sandaa, R. -A., M. Heldal, T. Castberg, R. Thyrhaug, and G. Bratbak. 2001. “Isolation and Characterization of Two Viruses with Large Genome Size Infecting Chrysochromulina Ericina (Prymnesiophyceae) and Pyramimonas Orientalis (Prasinophyceae).” *Virology* 290 (2): 272–80. doi:10.1006/viro.2001.1161.

Sandaa, R.-A., and A. Larsen. 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): 4610–18. doi:10.1128/AEM.00168-06.

Sano, E., S. Carlson, L. Wegley, and F. Rohwer. 2004. “Movement of Viruses between Biomes.” *Applied and Environmental Microbiology* 70 (10): 5842–46. doi:10.1128/AEM.70.10.5842-5846.2004.

Short, C. M., and C. A. Suttle. 2005. “Nearly Identical Bacteriophage Structural Gene Sequences Are Widely Distributed in Both Marine and Freshwater Environments.” *Applied and Environmental Microbiology* 71 (1): 480–86. doi:10.1128/AEM.71.1.480-486.2005.

Sipos, R., A. Székely, S. Révész, and K. Márialigeti. 2010. “Addressing PCR Biases in Environmental Microbiology Studies.” In *Bioremediation SE - 3*, edited by Stephen P Cummings, 599:37–58. Methods in Molecular Biology. Humana Press. doi:10.1007/978-1-60761-439-5_3.

Smith, D. P., and K. G. Peay. 2014. “Sequence Depth, Not PCR Replication, Improves Ecological Inference from next Generation DNA Sequencing.” *PloS One* 9 (2): e90234. doi:10.1371/journal.pone.0090234.

Snyder, J. C., B. Wiedenheft, M. Lavin, F. F. Roberto, J. Spuhler, A. C. Ortmann, T. Douglas, and M. Young. 2007. “Virus Movement Maintains Local Virus Population Diversity.” *Proceedings of the National Academy of Sciences* 104 (48): 19102–7. doi:10.1073/pnas.0709445104.

Sobecky, P. A., and T. H. Hazen. 2009. “Horizontal Gene Transfer and Mobile Genetic Elements in Marine Systems.” *Methods in Molecular Biology (Clifton, N.J.)*. doi:10.1007/978-1-60327-853-9_25.

Solonenko, S.A., J. C. Ignacio-Espinoza, A. Alberti, C. Cruaud, S. Hallam, K. Konstantinidis, G.Tyson, P. Wincker, and M. B. Sullivan. 2013. “Sequencing Platform and Library Preparation Choices Impact Viral Metagenomes.” *BMC Genomics* 14: 320. doi:10.1186/1471-2164-14-320.

Stackerbrandt, E., and B. M. Goebel. 1994. “Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology.” *International Journal of Systematic Bacteriology* 44 (4): 846–49. doi:10.1099/00207713-44-4-846.

Steward, G. F., L. B. Fandino, J. T. Hollibaugh, T. E. Whitledge, and F. Azam. 2007. “Microbial Biomass and Viral Infections of Heterotrophic Prokaryotes in the Sub-Surface Layer of the Central Arctic Ocean.” *Deep Sea Research Part I: Oceanographic Research Papers* 54 (10): 1744–57. doi:10.1016/j.dsr.2007.04.019.

Sullivan, M. B. 2014. “Viromes, Not Gene Markers for Studying dsDNA Viral Communities.” *Journal of Virology*, no. December. doi:10.1128/JVI.03289-14.

Sullivan, M. B, J. B. Waterbury, and S. W. Chisholm. 2003. “Cyanophages Infecting the Oceanic Cyanobacterium Prochlorococcus.” *Nature* 424 (6952): 1047–51. doi:10.1038/nature01929.

Sullivan, M. B., M. L. Coleman, P.Weigele, F. Rohwer, and S.W. Chisholm. 2005. “Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations.” *PLoS Biology* 3 (5): 0790–0806. doi:10.1371/journal.pbio.0030144.

Sullivan, M. B., B. Krastins, J. L. Hughes, L. Kelly, M. Chase, D.Sarracino, and S. W. Chisholm. 2009. “The Genome and Structural Proteome of an Ocean Siphovirus: A New Window into the Cyanobacterial ‘Mobilome.’” *Environmental Microbiology* 11 (11): 2935–51. doi:10.1111/j.1462-2920.2009.02081.x.

Sullivan, M. B., D. Lindell, J. A. Lee, L. R. Thompson, J. P. Bielawski, and S. W. Chisholm. 2006. “Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts.” *PLoS Biology* 4 (8): 1344–57. doi:10.1371/journal.pbio.0040234.

Sun, S., B. La Scola, V. D. Bowman, C. M. Ryan, J. P. Whitelegge, D. Raoult, and M. G. Rossmann. 2010. “Structural Studies of the Sputnik Virophage.” *Journal of Virology* 84: 894–97. doi:10.1128/JVI.01957-09.

Suttle, C. A., and A. M. Chan. 1994. “Dynamics and Distribution of Cyanophages and Their Effect on Marine Synechococcus Spp.” *Applied and Environmental Microbiology* 60 (9): 3167–74.

Suttle, C. A. 1994. “The Significance of Viruses to Mortality in Aquatic Microbial Communities.” *Microbial Ecology* 28 (2): 237–43. doi:10.1007/BF00166813.

Suttle, C. A. 2005. “Viruses in the Sea.” *Nature* 437 (7057): 356–61. doi:10.1038/nature04160.

Suttle, C. A. 2007. “Marine Viruses--Major Players in the Global Ecosystem.” *Nature Reviews. Microbiology* 5 (10). Nature Publishing Group: 801–12. doi:10.1038/nrmicro1750.

Suzuki, M.T., and S.J. Giovannoni. 1996. “Bias Caused by Template Annealing in the Amplification of Mixtures of 16S rRNA Genes by PCR.” *Applied and Environmental Microbiology* 62 (2): 625–30.

Syvertsen, E. E. 1991. “Ice Algae in the Barents Sea: Types of Assemblages, Origin, Fate and Role in the Ice-Edge Phytoplankton Bloom.” *Polar Research*.

Tétart, F., C. Desplats, M. Kutateladze, C. Monod, H. W. Ackermann, and H. M. Krisch. 2001. “Phylogeny of the Major Head and Tail Genes of the Wide-Ranging T4-Type Bacteriophages.” *Journal of Bacteriology* 183 (1). American Society for Microbiology: 358–66. doi:10.1128/JB.183.1.358-366.2001.

ThermoFischer Scientific. 2012. “Ion PGM System for Next Generation Sequencing.” https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing.html.

Thingstad, T. F., and R. Lignell. 1997. “Theoretical Models for the Control of Bacterial Growth Rate, Abundance, Diversity and Carbon Demand.” *Aquatic Microbial Ecology* 13: 19–27. doi:10.3354/ame013019.

Thingstad, T. F., B. Pree, J. Giske, and S. Våge. 2015. “What Difference Does It Make If Viruses Are Strain-, rather than Species-Specific?” *Frontiers in Microbiology* 6 (April): 320. doi:10.3389/fmicb.2015.00320.

Tomaru, Y., H. Tanabe, S. Yamanaka, and K. Nagasaki. 2005. “Effects of Temperature and Light on Stability of Microalgal Viruses, HaV, HcV and HcRNAV.” *Plankton Biology and Ecology (Japan)*.

Vader, A., M. Marquardt, A. R. Meshram, and T. M. Gabrielsen. 2014. “Key Arctic Phototrophs Are Widespread in the Polar Night.” *Polar Biology* 38: 13–21. doi:10.1007/s00300-014-1570-2.

Van Etten, J. L., M. V. Graves, D. G. Müller, W. Boland, and N. Delaroque. 2014. “Phycodnaviridae– Large DNA Algal Viruses.” *Archives of Virology* 147 (8): 1479–1516. doi:10.1007/s00705-002-0822-6.

Varela, M.M., H. M. Van Aken, E. Sintes, and G.J. Herndl. 2008. “Latitudinal Trends of Crenarchaeota and Bacteria in the Meso- and Bathypelagic Water Masses of the Eastern North Atlantic.” *Environmental Microbiology* 10: 110–24. doi:10.1111/j.1462-2920.2007.01437.x.

Walker, P. A, and D. P Faith. 1994. “Diversity-PD: Procedures for Conservation Evaluation Based on Phylogenetic Diversity.” *Biodiversity Letters* 2 (5): 132–39. doi:10.2307/2999777.

Wang, G., J. Jin, S. Asakawa, and M. Kimura. 2009. “Survey of Major Capsid Genes (*g23*) of T4-Type Bacteriophages in Rice Fields in Northeast China.” *Soil Biology and Biochemistry* 41 (2): 423–27. doi:10.1016/j.soilbio.2008.11.012.

Wang, G., J. Murase, K.Taki, Y. Ohashi, N. Yoshikawa, S. Asakawa, and M. Kimura. 2009. “Changes in Major Capsid Genes (*g23*) of T4-Type Bacteriophages with Soil Depth in Two Japanese Rice Fields.” *Biology and Fertility of Soils* 45 (5): 521–29. doi:10.1007/s00374-009-0362-2.

Wassmann, P., K.N. Kosobokova, D. Slagstad, K.F. Drinkwater, R.R. Hopcroft, S.E. Moore, I. Ellingsen, R. J. Nelson, E. Popova, J. Berge, and E. Carmack. 2015. “The Contiguous Domains of Arctic Ocean Advection: Trails of Life and Death.” *Progress in Oceanography*, July. doi:10.1016/j.pocean.2015.06.011.

Weigele, P. R., W. H. Pope, M.L. Pedulla, J. M. Houtz, A. L. Smith, J. F. Conway, J. King, G. F. Hatfull, J. G. Lawrence, and R.W. Hendrix. 2007. “Genomic and Structural Analysis of Syn9, a Cyanophage Infecting Marine Prochlorococcus and Synechococcus.” *Environmental Microbiology* 9 (7): 1675–95. doi:10.1111/j.1462-2920.2007.01285.x.

Weinbauer, M. G. 2004. “Ecology of Prokaryotic Viruses.” *FEMS Microbiology Reviews*. doi:10.1016/j.femsre.2003.08.001.

Weinbauer, M. G., I. Brettar, and M. G. Höfle. 2003. “Lysogeny and Virus-Induced Mortality of Bacterioplankton in Surface, Deep, and Anoxic Marine Waters.” *Limnology and Oceanography*. doi:10.4319/lo.2003.48.4.1457.

Weinbauer, M. G., and F. Rassoulzadegan. 2004. “Are Viruses Driving Microbial Diversification and Diversity?” *Environmental Microbiology*. doi:10.1046/j.1462-2920.2003.00539.x.

Wells, L. E., M.Cordray, S. Bowerman, L. A. Miller, W. F. Vincent, and J. W. Deming. 2006. “Archaea in Particle-Rich Waters of the Beaufort Shelf and Franklin Bay, Canadian Arctic: Clues to an Allochthonous Origin?” *Limnology and Oceanography*. doi:10.4319/lo.2006.51.1.0047.

Wells, L. E., and J. W. Deming. 2003. “Abundance of Bacteria, the Cytophaga-Flavobacterium Cluster and Archaea in Cold Oligotrophic Waters and Nepheloid Layers of the Northwest Passage, Canadian Archipelago.” *Aquatic Microbial Ecology* 31 (1): 19–31. doi:10.3354/ame031019.

Wheeler, P. A., M. Gosselin, E. Sherr, D. Thibaultc, D. L. Kirchman, R. Benner, and T. E. Whitledge. 1996. “Active Cycling of Organic Carbon in the Central Arctic Ocean.” *Nature* 380 (6576): 697–99. doi:10.1038/380697a0.

Wichels, A., S. S. Biel, H. R. Gelderblom, T. Brinkhoff, G. Muyzer, and C. Schütt. 1998. “Bacteriophage Diversity in the North Sea.” *Applied and Environmental Microbiology* 64 (11): 4128–33.

Wilhelm, S. W., and C. A. Suttle. 1999. “Viruses and Nutrient Cycles in the Sea.” *BioScience*. doi:10.2307/1313569.

Wilson, W. H., J. L. Van Etten, and M. J. Allen. 2009. “The Phycodnaviridae: The Story of How Tiny Giants Rule the World.” *Current Topics in Microbiology and Immunology* 328: 1–42. doi:10.1007/978-3-540-68618-7-1.

Wilson, W. H, D. C. Schroeder, M.J. Allen, M.T. G. Holden, J. Parkhill, B. G. Barrell, C. Churcher, N. Hamlin, K. Mungall, H. Norbertczak, M.A. Quail, C. Price, E. Rabbinowitsch, D. Walker, M. Craigon, D. Roy, and P. Ghazal. 2005. “Complete Genome Sequence and Lytic Phase Transcription Profile of a Coccolithovirus.” *Science (New York, N.Y.)* 309 (5737): 1090–92. doi:10.1126/science.1113109.

Winter, C., B. Matthews, and C. A. Suttle. 2013. “Effects of Environmental Variation and Spatial Distance on Bacteria, Archaea and Viruses in Sub-Polar and Arctic Waters.” *The ISME Journal* 7 (8). International Society for Microbial Ecology: 1507–18. doi:10.1038/ismej.2013.56.

Wohlers, J., A. Engel, E. Zöllner, P. Breithaupt, K. Jürgens, H.-G. Hoppe, U. Sommer, and U. Riebesell. 2009. “Changes in Biogenic Carbon Flow in Response to Sea Surface Warming.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (17): 7067–72. doi:10.1073/pnas.0812743106.

Wommack, K. E., and R. R. Colwell. 2000. “Virioplankton: Viruses in Aquatic Ecosystems.” *Microbiology and Molecular Biology Reviews : MMBR* 64 (1): 69–114. doi:10.1128/MMBR.64.1.69-114.2000.

Yau, S., F. M. Lauro, M. Z. DeMaere, M. V. Brown, T. Thomas, M. J. Raftery, C. Andrews-Pfannkoch, M. Lewis, J. M. Hoffman, J. A. Gibson, and R. Cavicchioli. 2011. “Virophage Control of Antarctic Algal Host-Virus Dynamics.” *Proceedings of the National Academy of Sciences of the United States of America* 108: 6163–68. doi:10.1073/pnas.1018221108.

Yutin, N., Y. I. Wolf, and E. V. Koonin. 2014. “Origin of Giant Viruses from Smaller DNA Viruses Not from a Fourth Domain of Cellular Life.” *Virology* 466-467 (October): 38–52. doi:10.1016/j.virol.2014.06.032.

Zhang, Y., N. Jiao, M. T. Cottrell, and D. L. Kirchman. 2006. “Contribution of Major Bacterial Groups to Bacterial Biomass Production along a Salinity Gradient in the South China Sea.” *Aquatic Microbial Ecology* 43 (3): 233–41. doi:10.3354/ame043233.

Zheng, C., G. Wang, J. Liu, C. Song, H. Gao, and X. Liu. 2013. “Characterization of the Major Capsid Genes (*g23*) of T4-Type Bacteriophages in the Wetlands of Northeast China.” *Microbial Ecology* 65 (3): 616–25. doi:10.1007/s00248-012-0158-z.

Zhong, X., and S. Jacquet. 2014. “Differing Assemblage Composition and Dynamics in T4-like Myophages of Two Neighbouring Sub-Alpine Lakes.” *Freshwater Biology* 59 (8): 1577–95. doi:10.1111/fwb.12365.

Zinger, L., L. A. Amaral-Zettler, J. A. Fuhrman, M. C. Horner-Devine, S. M. Huse, D. B. M.Welch, J. B. H. Martiny, M. Sogin, A. Boetius, and A. Ramette. 2011. “Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems.” *PloS One* 6 (9). Public Library of Science: e24570. doi:10.1371/journal.pone.0024570.

Appendix A: Protocols

A.1 Rapid Protocol for DNA Isolation

LYSIS

1. Incubate 500 µL viral concentrate at 90 °C for 2 x 2 minutes, on ice in between.
2. Add 20 µL 0.5M EDTA (pH 8.0)
3. Add 5 µL Proteinase K (freshly made 10 mg/mL), incubate 10 minutes at 55 °C.
4. Add 25 µL 10% SDS, incubate further for 1 h at 55 °C.

PURIFICATION

1. Clean the lysate using ZYMO DNA clean up and concentration kit (Appendix 1.2)
2. Eluate in sterile destilled water (20 µL).

A.2 ZYMO DNA Cleanup and ConcentratorTM -5 (D4003, Zymo Research) Protocol

1. In a 1.5 mL microcentrifuge tube, add 2-7 volumes of DNA Binding Buffer to each volume of DNA sample. Mix briefly by vortexing.
2. Transfer mixture to a provided Zymo-Spin^{TM} Column in a Collection Tube.
3. Centrifuge for 30 seconds. Discard the flow-through.
4. Add 200 µL DNA Wash Buffer to the column. Centrifuge for 30 seconds. Repeat the wash step.
5. Add ≥ 6 µL DNA Elution Buffer or water directly to the column matrix and incubate at room temperature for one minute.
6. Transfer the column to a 1.5 mL microcentrifuge tube and centrifuge for 30 seconds to elute the DNA.

A.3 Agencourt AMPure XP magnetic bead kit (Beckman Coulter, USA)

1. Shake the Agencourt AMPure XP bottle to resuspend any magnetic particles that may have settled. Then add 1.8 x the sample reaction volume of Agencourt AMPure XP.
2. Pipette mix reagent and sample 10 times. Let the mixed sample incubate for 5 minutes at RT for maximum recovery.
3. Place the reaction tubes onto the Agencourt Super Magnet Rack for 2 minutes to separate beads from the solution. Wait for the solution to clear before proceeding to the next step.
4. Aspirate the cleared solution from the reaction tubes and discard. Leave 5 µL of supernatant behind, otherwise beads are drawn out with the supernatant.
5. Dispense 200 µL of 70% ethanol to each reaction tube and incubate for 30 seconds at RT. Aspirate out the ethanol and discard. Repeat for a total of two washes.
6. Remove the reaction tubes from the magnetic rack, and then add 40 µL of elution buffer to each reaction tube and pipette mix 10 times. Incubate for 2 minutes.
7. Place the reaction tubes onto the Agencourt Super Magnet Rack for 1 minute to separate beads from the solution.
8. Transfer the eluate to a new reaction tube.

A.4 DNA Electrophoresis Preparation and Protocol

1. Add 60 mg of SeaKem® LE Agarose (50004, Lonza) to a glass container filled with 60 mL of TAE (consisting of 40mM Tris, 20mM acetic acid, and 1mM EDTA).
2. Loosely cap the container and microwave the mixture for 30 seconds then swirl for 20 seconds, repeating this step until the agarose is completely dissolved in the TAE.
3. Add 2 µL of 10 000X Gel Red™ stain (41003, Biotium, USA) to gel and mix until transparent.
4. Allow gel to cool at least 20 minutes (must be cool enough to touch), then pour gel into gel rack on flat surface and insert well combs.
5. Allow 25 minutes for the gel to solidify and fill electrophoresis chamber with TAE.
6. Insert the solidified gel into the electrophoresis chamber so that the gel surface is submerged in TAE.
7. Mix sample on a sterile parafilm surface with 1 µL DNA stain and load into gel. Load the Mass DNA Ladder (10496-016, Life Technologies) and positive and negative controls.
8. Run the electrophoresis reaction at 300 Volts for 10 minutes.
9. Load the completed gel into BIO RAD Molecular Imager® (ChemiDoc XRS™) and obtain a fluorescent image using Image Lab™ Software.

A.5 Annotated bioinformatics pipeline

Formatting and Evaluation of the sequencing run

#The output from the sequencing run preferably arrives in SFF or in FASTQ format (for the QIIME pipeline).
#This file type includes not only the base pair information, but also the quality scores (PHRED scores).
Convert SFF to QUAL, FASTA files (in QIIME)
> process_sff.py -i infile.sff
#Convert FASTA and QUAL files to FASTQ (in QIIME)
>convert_fastaqual_fastq.py -f infile.fasta -q infile.qual
#
Some sequencing facilities will instead provide a BAM file, instead convert to a FASTQ file using bamtools.
This is was the format the Ion Torrent dataset arrived in. All other runs were in SFF format.
#BAM is a sequence alignment file type, but conversion to FASTQ is also possible
>bedtools bamToFastq -i infile.BAM -fq seqfile.fastq
#
#Once in FASTQ format, the dataset is run through the program FASTQC (author used interface-based version)
#This is a good program for assessing the success of the sequencing run.

Illumina-specific contamination cleaning, adapter removal, and merging of paired reads

#BBduk is used to remove the control phiX sequences from the dataset
#each of the two runs are simultaneously cleaned for phiX. The script looks for kmers of length 31 to match #phiX sequence. The script creates files containing reads matching to phiX and the remainder of the clean data
>sh /filepath/bbduk.sh -Xmx1g in1=R1.fastq in2=R2.fastq out1=unmatched1.fastq out2=unmatched2.fastq outm1=matched1.fastq outm2=matched2.fastq ref=/filepath/bowtie2-2.2.4/indexes/phiX.fasta k=31 hdist=1 stats=stats.txt
#
#BBduk is used to remove the adapters**.** This run had not only Illumina adapters, but also Roche/454 adapters
#inside of Illumina adapters on the amplicon ends. This required this removal step from left end of sequences.

>sh /filepath/bbduk.sh -Xmx1g in1=unmatched1.fastq in2=unmatched2.fastq out1=lclean1.fastq out2=lclean2.fastq literal=CCATCTCATCCCTGCGTGTCTCCGACTCAG ktrim=l k=30 mink=12 hdist=1
#
#BBduk is then used to merge paired cleaned reads
>sh /filepath/bbduk.sh in1= lclean1.fastq in2= lclean2.fastq out=merged.fastq
#
#From this point Illumina sequences were processed exactly the same as all other datasets.

Quality trimming with BBduk
#quality trimming was accomplished using a package BBMaps developed by Brian Bushnell
#both ends of sequence were trimmed to exclude sequence which had PHRED scores of 27 or less
#Xmx1g is passed to Java to set the memory usage, in this case it specifies 1 gig of RAM
#This method retains more generally good quality information that can sometimes be lost when
#trimming sequences based on average whole-sequence PHRED scoring.
>/filepath/bbduk.sh -Xmx1g in=seqfile.fastq out=seqfile_trim.fastq qtrim=rl trimq=27
#
the remainder of the QIIME pipeline does not utilize quality scores, so we convert to FASTA format
>python /filepath/fastq_to_fasta.py -n seqfile_trim.fastq -o seqfile_trim.fasta

Demultiplexing in QIIME
#our barcoded samples must be separated out from one another. We use QIIME for this.
#The script requires a mapping file about sequence dataset, including barcode and primer sequences.
#Helpful example Map file @ http://qiime.org/_static/Examples/File_Formats/Example_Mapping_File.txt
#requires user to tell the script how many base pairs our barcodes have (10 in this case)
>split_libraries.py -m MapFile.txt - f seqfile_trim.fasta -b 10 -o Split_library_output/

Pick OTUs based on sequence similarity with USEARCH in QIIME
#Cluster based on the USEARCH algorithm as it has a de novo chimera checking function.
#The reason for algorithm choice is that open source viral databases are sparsely populated for targeted
#amplicon data compared to 16S or other prokaryotic genes.
#USEARCH author Robert Edgar does not recommend expanding the dissimiliarity radius beyond 3%.
#Choose to suppress chimera detection based on a reference file. The script instead does de novo
#chimera checking. Chimeras occur when two parent sequences fuse to one another during the
#PCR amplification process. The most numerous sequences in the dataset become the “reference” for this check.
>pick_otus.py -i Split_library_output/seqs.fna -s 0.97 -m usearch --
suppress_reference_chimera_detection -o picked_otus_97/

Elimination of singletons
#Although QIIME does not produce OTUs with fewer than 2 member reads, any one sample may only
#contribute one read to that OTU. In this case, it is considered by some to be a singleton.
To take care of this, the OTU table split by sample
>filter_samples_from_otu_table.py -i otu_table.biom -o otu_table_samp1.biom -e samp1_otus.txt
Each sample-specific OTU table was then filtered for singletons.
> filter_otus_from_otu_table.py -i otu_table_samp1.biom -o otu_table_samp1_nosing.biom -n 2
The sample-specific OTU tables were merged back into a single OTU table
> merge_otu_tables.py -i otu_table_samp1.biom, otu_table_samp2.biom -o merged_otu_table.biom

Pick representative sequences in QIIME

#Take a representative sequence of each created OTU to cut down on the sequences to align for
#phylogenetic interpretation. Default chooses the first encountered sequence from each OTU.
>pick_rep_set.py -i picked_otus_97/seqs_otus.txt -f Split_library_output/seqs.fna -o rep_set_97.fna

Align representative sequences in QIIME
#Uses algorithm MUSCLE to align sequences
#Alignment of representative sequences and sequences pulled from the NCBI BLAST database that matched as
#top hits to all OTUs shared between all samples in a dataset.
>align_seqs.py -m muscle -i rep_set_with_BLAST

Create a phylogeny of representative sequences in QIIME
#this script uses the default tree building method, FastTree, to build a phylogeny from the alignment
#this script is also used to create a tree with NCBI BLAST sequences included
>make_phylogeny.py -i /filepath/muscle_aligned_rep_set.fna -o rep_set.tre

Make OTU table in QIIME
#this script creates an HDF5 BIOM formatted otu table
>make_otu_table.py -i picked_otus_97/seqs_otus.txt -o otu_table_97.biom

Alpha diversity measures

1. **Estimated species richness indices: made in QIIME**
   # for this command, use the biom formatted OTU table(s) and the phylogenetic tree to run multiple metrics
   # select chosen richness indices. Here we have chosen Chao1 and PD whole tree.
   >alpha_diversity.py –i otu_tables/tables.biom –m chao1, PD_whole_tree -t rep_set.tre -o adiv_output/
2. **Alpha rarefaction curves: made in R using package "vegan"**
   # load the package
   >lib(vegan)
   #import the OTU table as a csv formatted text file and transpose for use in vegan
   >SampleName <- read.csv("~/Filepath/Filename.csv", header=T, row.names=1)
   >SampleName <- t(SampleName)
   #
   #define parameters of line color, the rarefaction maximum, and line weight
   >col <- c("blue","darkred","forestgreen")
   >lwd <- c(2,2)
   >raremax <-min(rowSums(SampleName)
   #
   #run the rarefaction, tell the script to rarefy in steps of 100 reads, from 100 reads to the (raremax).
   #Define labels and axes limits as desired.
   >rarecurve(SampleName, step=100, xlab="Sample Size", ylab= "Species", col=col, lwd=lwd, xlim=c(0,350000), ylim=c(0,1400))

Beta-diversity measures

1. **Beta-diversity measures with jackknife support.**
   #This script rarefies all samples to specified sequence depth
   # calculates Unifrac distance matrices, and also creates a diversity of plots including
   #UPGMA dendrograms and PCoA based on Unifrac distance matrices

#jackknifing (repeated resampling of data) is performed to test robustness of UPGMA clustering
#and compares UPGMA trees to the full or consensus trees to generate jackknife support for nodes
>jackknifed_beta_diversity.py -i otu_table.biom -o bdiv_raredepth/ -e 25000 -m MapFile.txt -t rep_set.tre

2. **ANOSIM**
   #run ANOSIM on groups of samples categorized by water mass, as indicated in the mapping file
   >compare_categories.py --method anosim -i /filepath/weighted_unifrac_distmatrix.txt -m MapFile.txt -c Treatment -o ANOSIM_out/

Comparison of *16S* and *g23* datasets using Mantel Test in R package "vegan"

#the Mantel statistic tests the correlation between two dissimilarity matrices. Permutations of the observed rows
#and columns are used to assess significance.
#
#load the package
>library(vegan)
#read the distance matrices into R
>bact <- read.table("~/filepath/otu_table_16S.txt", header=T, row.names=1)
>*g23* <-read.table("~/filepath/otu_table_*g23*.txt", header=T, row.names=1)
#transpose the data for using in vegan
>bact <- t(bact)
>*g23* <- t(*g23*)
#create Bray-Curtis distance matrices
>bact.dist <- vegdist(bact)
>*g23*.dist <- vegdist(*g23*)
#run the Mantel test
>mantel(bact.dist, *g23*.dist, method="spear")

Appendix B: Results

B.1 Electrophoresis gels

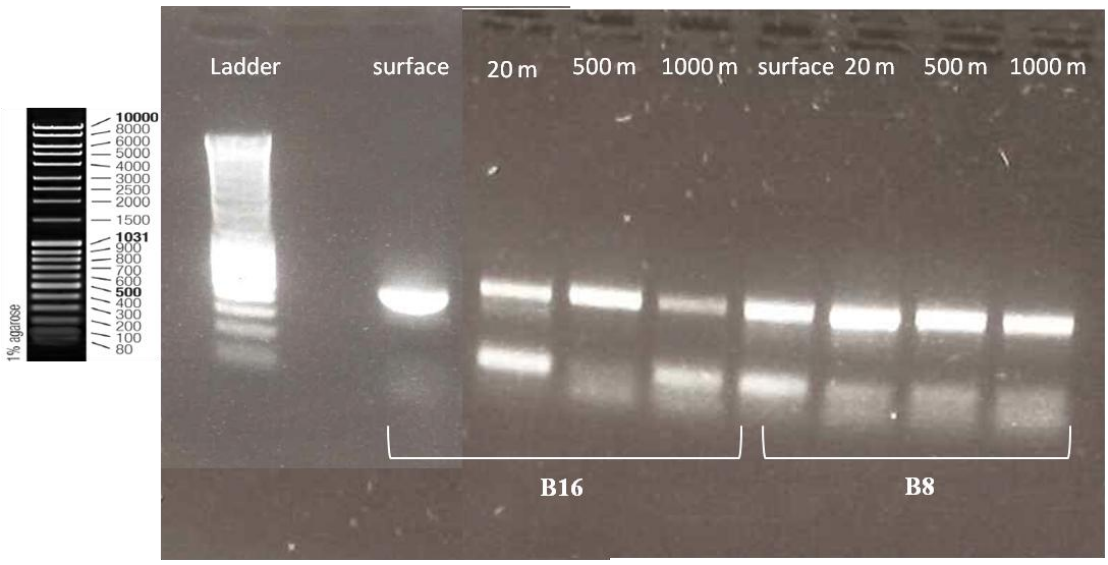


Figure B-1. Images of *g23* amplicons with barcodes added sent for sequencing (image is before purification step to remove primer), run on a 1% (w/v) agarose gel with DNA Mass Ladder for size reference.

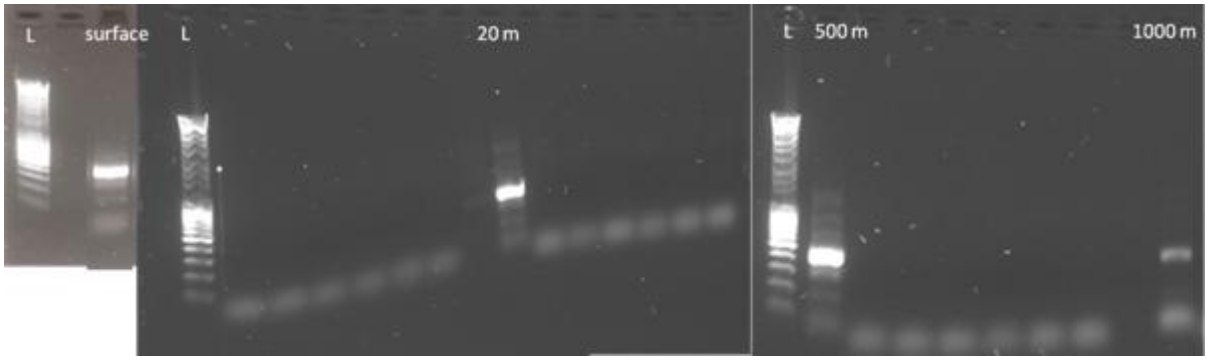


Figure B-2. Images of *phoH* amplicons from station B16 with barcodes added sent for sequencing, run on 1% (w/v) agarose gels with DNA Mass Ladder for size reference. Unlabeled lanes with weak bands are the initial amplification products without barcodes, labeled lanes are the final products sent for sequencing (image is before purification step to remove primer).

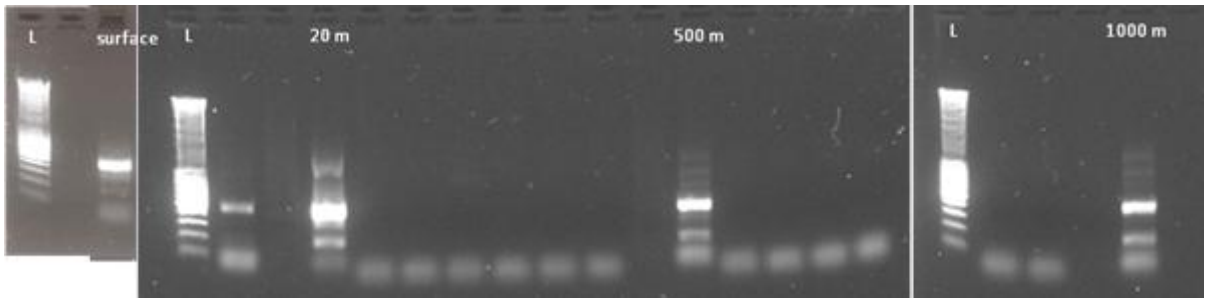


Figure B-3. Images of *phoH* amplicons from station B8 with barcodes added sent for sequencing, run on 1% (w/v) agarose gels with DNA Mass Ladder for size reference. Unlabeled lanes with weak or absent bands are the initial amplification products without barcodes, labeled lanes are the final products sent for sequencing (image is before purification step to remove primer).

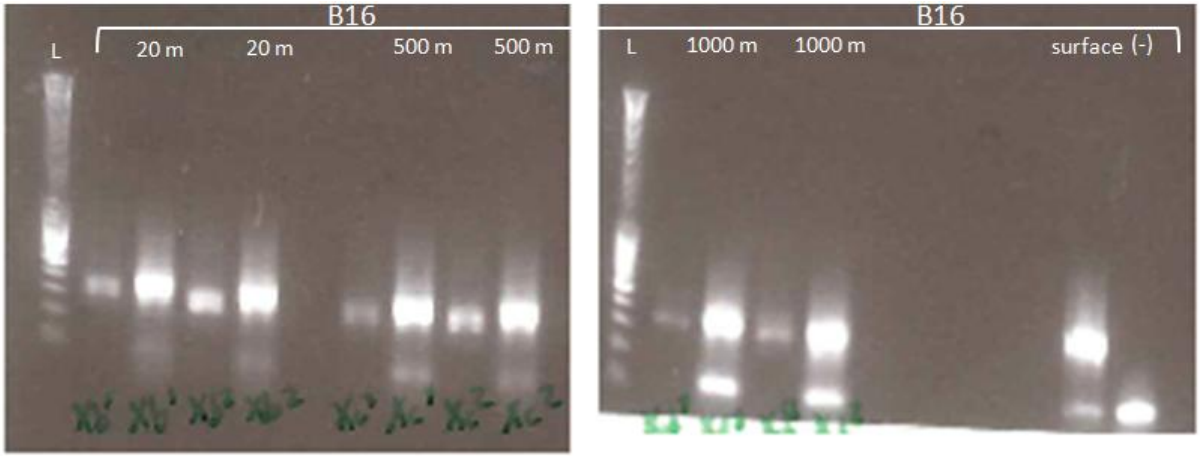


Figure B-4. Images of *MCP* amplicons from station B16 (duplicates from each depth) with barcodes added, run on 1% (w/v) agarose gels with DNA Mass Ladder for size reference. Unlabelled lanes with weak bands are the initial amplification products without barcodes, labelled lanes are the final products sent for sequencing (image is before purification step to remove primer).

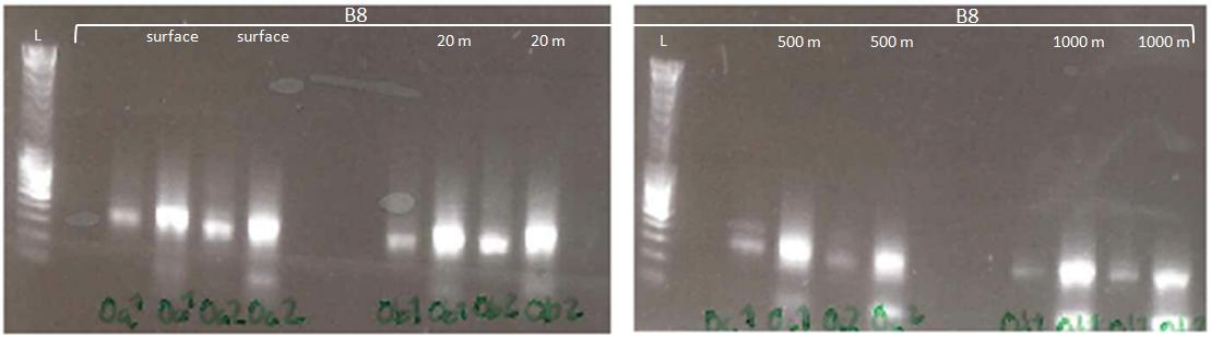


Figure B-5. Images of *MCP* amplicons from station B8 (duplicates from each depth) with barcodes added, run on 1% (w/v) agarose gels with DNA Mass Ladder for size reference. Unlabelled lanes with weak bands are the initial amplification products without barcodes, labelled lanes are the final products sent for sequencing (image is before purification step to remove primer).

B.2 Quality control reports

B.2.1 FASTQC report on raw sequencing run of *g23* on Roche/454

Filename	**Emily_*G23*.fastq**
File type	**Conventional base calls**
Encoding	**Sanger/Illumina 1.9**
Total Sequences	**1,171,251**
Filtered Sequences	**0**
Sequence length	**41 - 986**
%GC	**47**

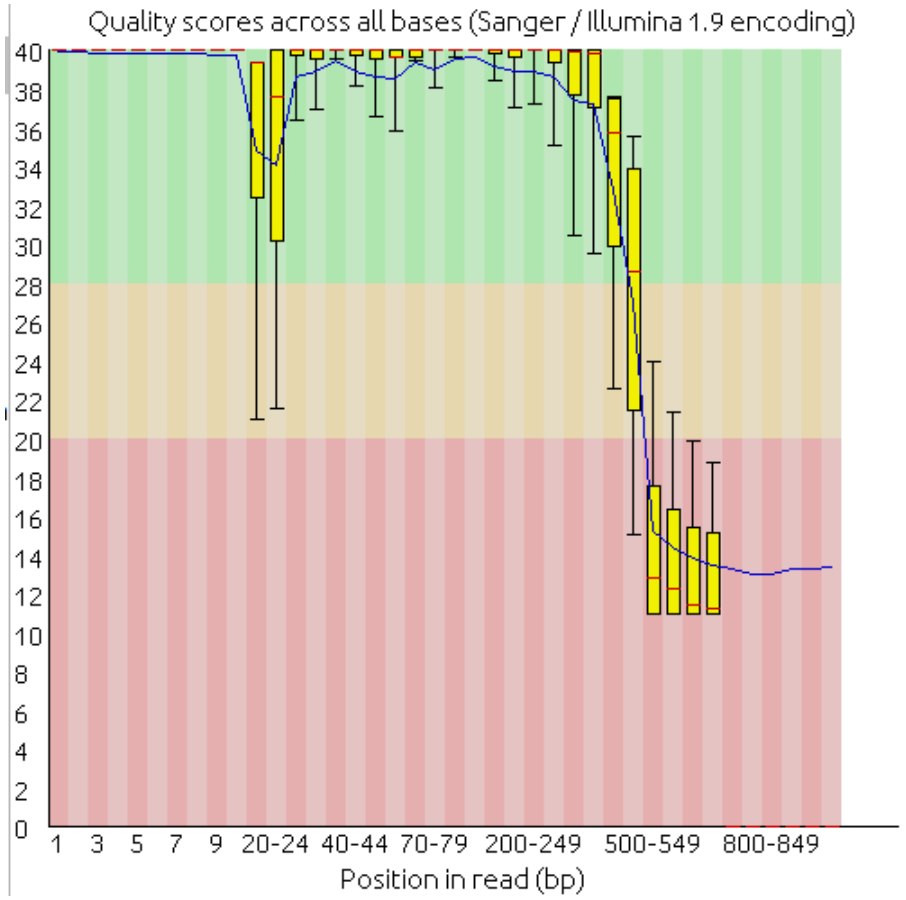


Figure B-6. PHRED scores per base position of reads in the Roche/454 *g23* sequencing run.

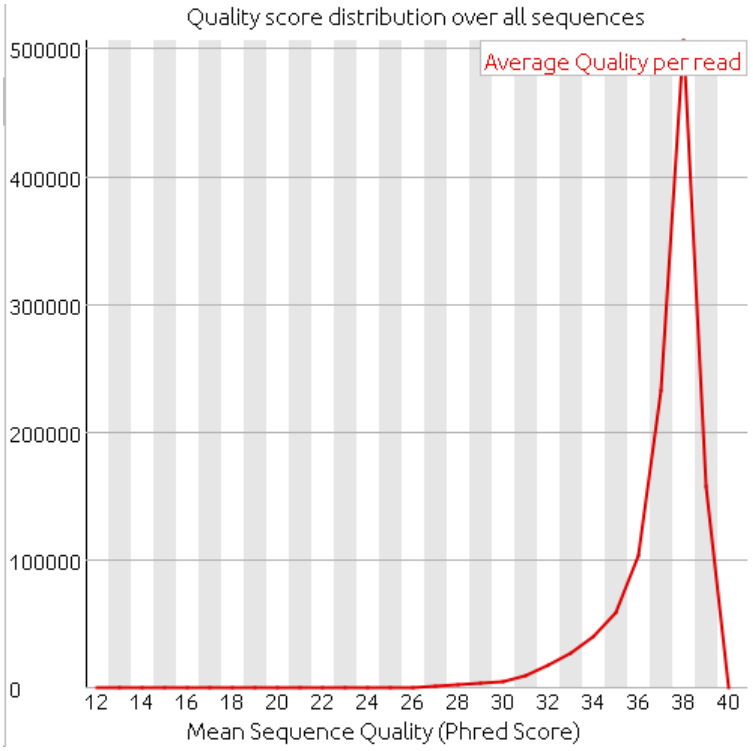


Figure B-7. Abundance of average per read PHRED scores for the Roche/454 *g23* dataset.

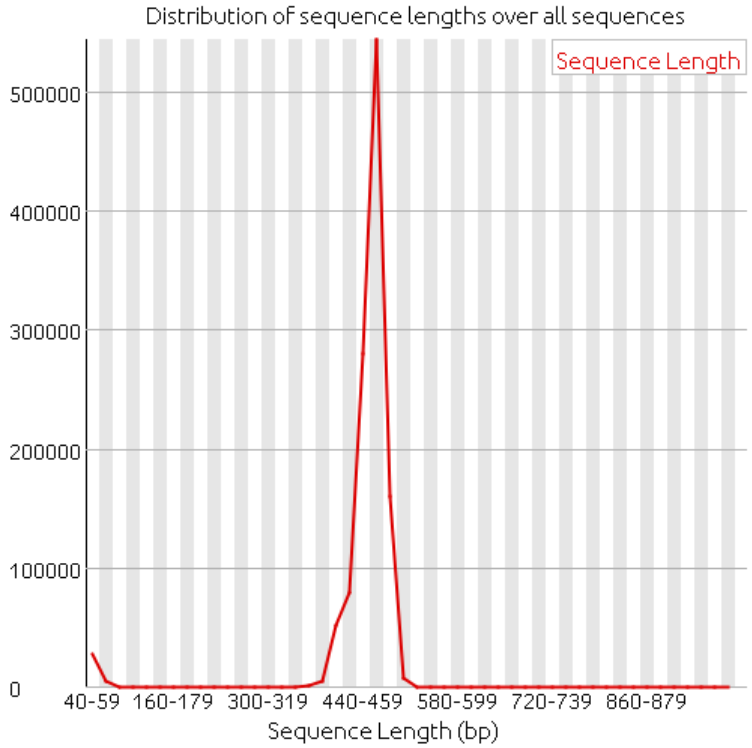


Figure B-8. Distribution of lengths from the Roche/454 *g23* dataset.

B.2.2 FASTQC Report on combined sequencing run including *phoH* and *MCP* on Roche/454

Filename	pooledReads.fastq
File type	Conventional base calls
Encoding	Sanger/Illumina 1.9
Total Sequences	232,305
Filtered Sequences	0
Sequence length	43 - 1625
%GC	45

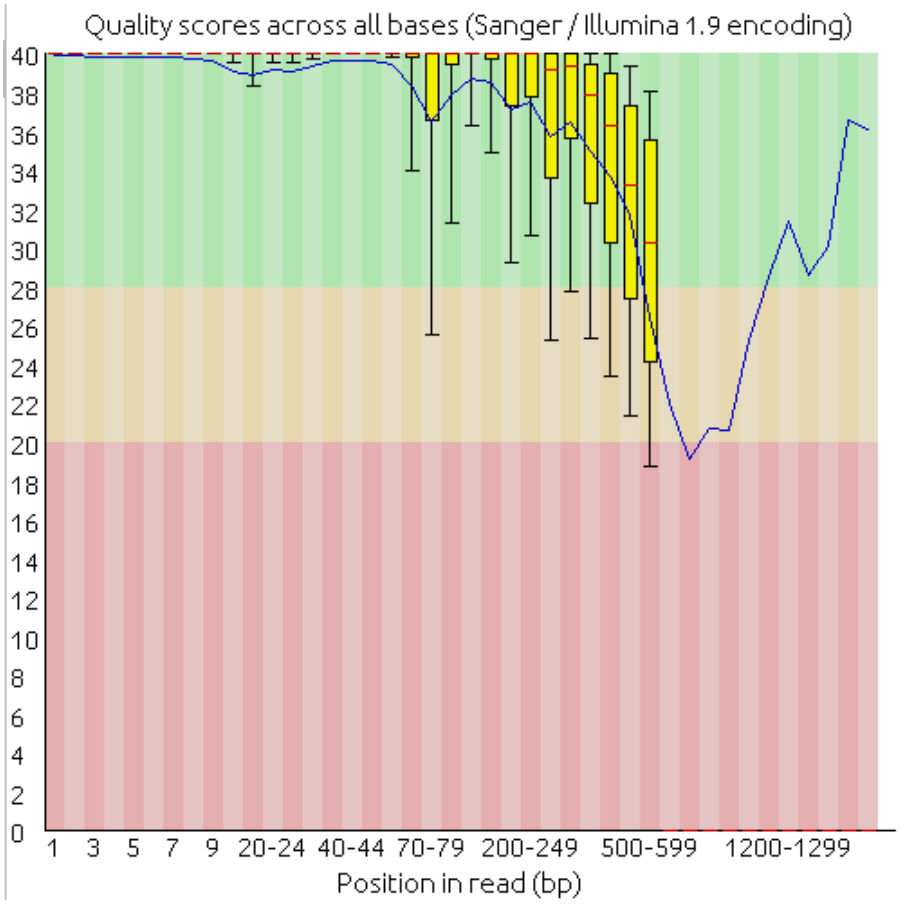


Figure B-9. PHRED scores per base position from the pooled *phoH* and *MCP* Roche/454 sequencing run

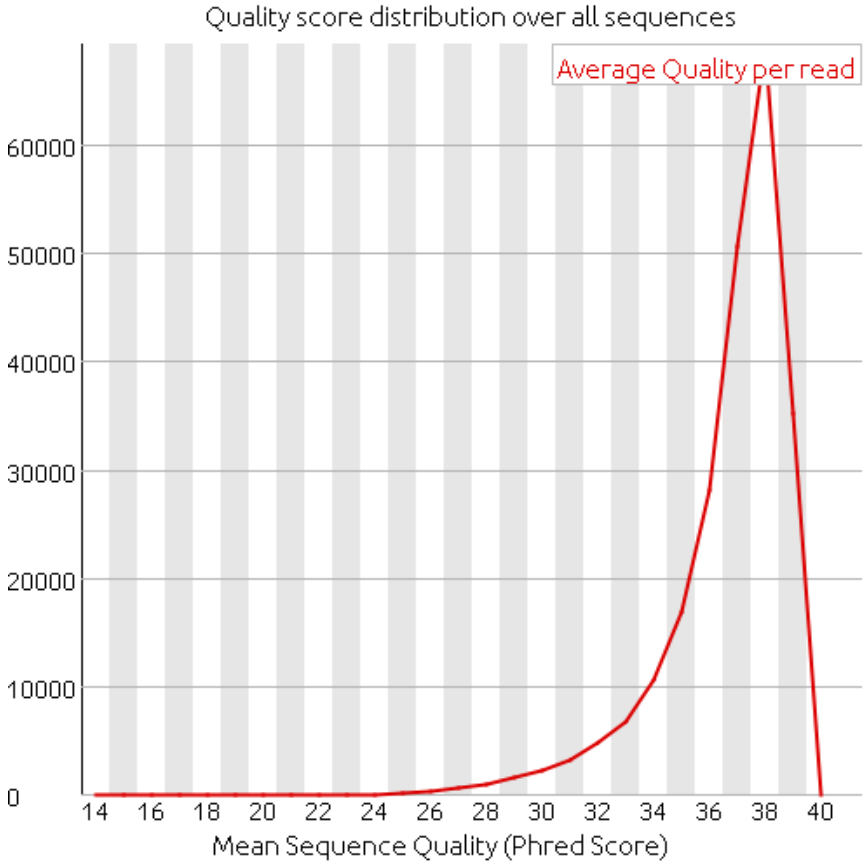


Figure B-10. Abundance of average per read PHRED scores for the pooled *phoH* and *MCP* Roche/454 sequencing run

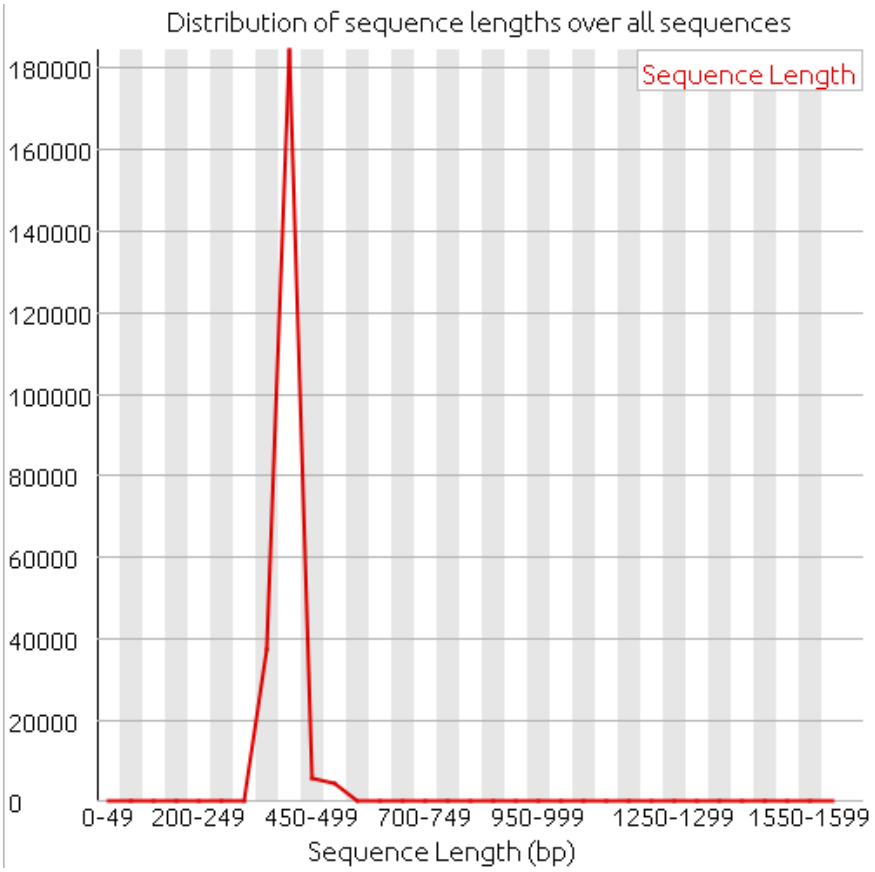


Figure B-11. Distribution of lengths in the pooled *phoH* and *MCP* Roche/454 sequencing run

B.2.3 FASTQC Report on merged paired reads of *g23* data on Illumina MiSeq

Filename	**Merged.fastq**
File type	**Conventional base calls**
Encoding	**Sanger/Illumina 1.9**
Total Sequences	**22,372,496**
Filtered Sequences	**0**
Sequence length	**35 - 590**
%GC	**48**

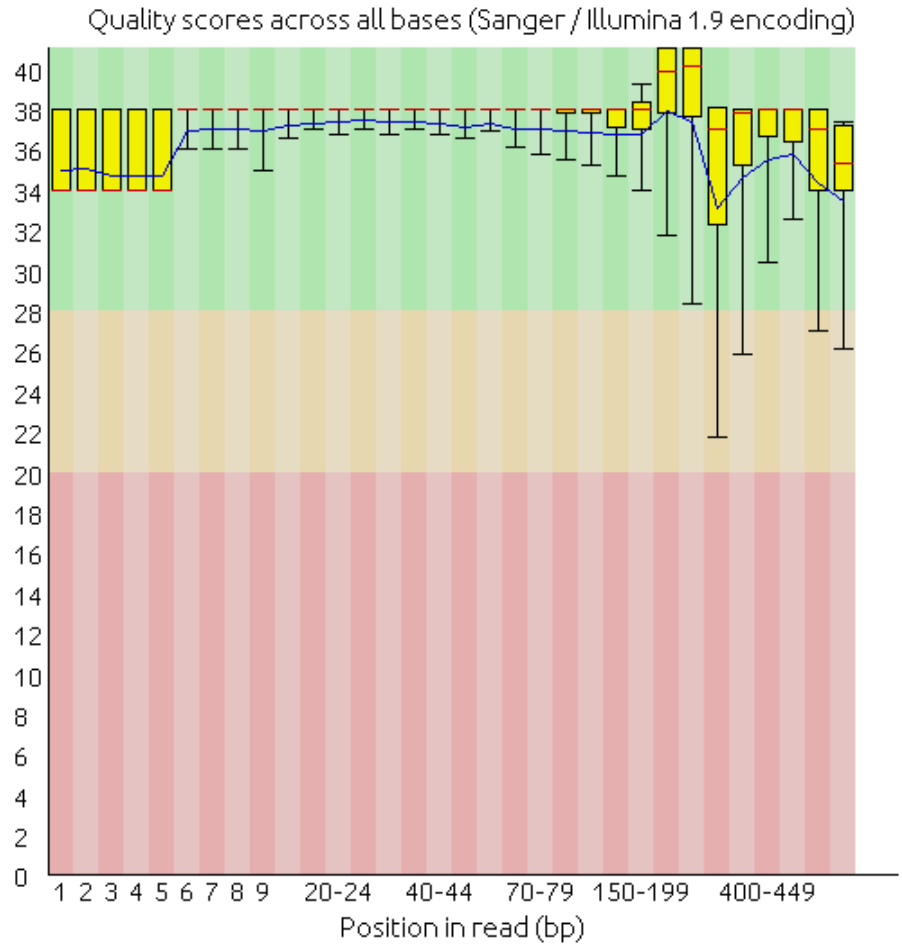


Figure B-12. PHRED scores per base position of the merged paired reads from the Illumina *g23* sequencing runs

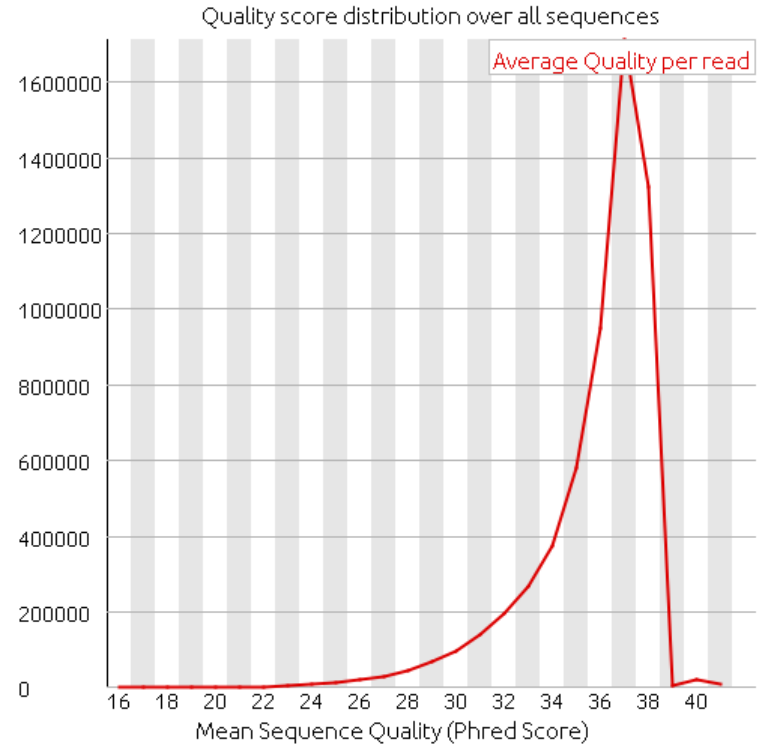


Figure B-13. Abundance of average per read PHRED scores for the merged paired reads from Illumina *g23* dataset.

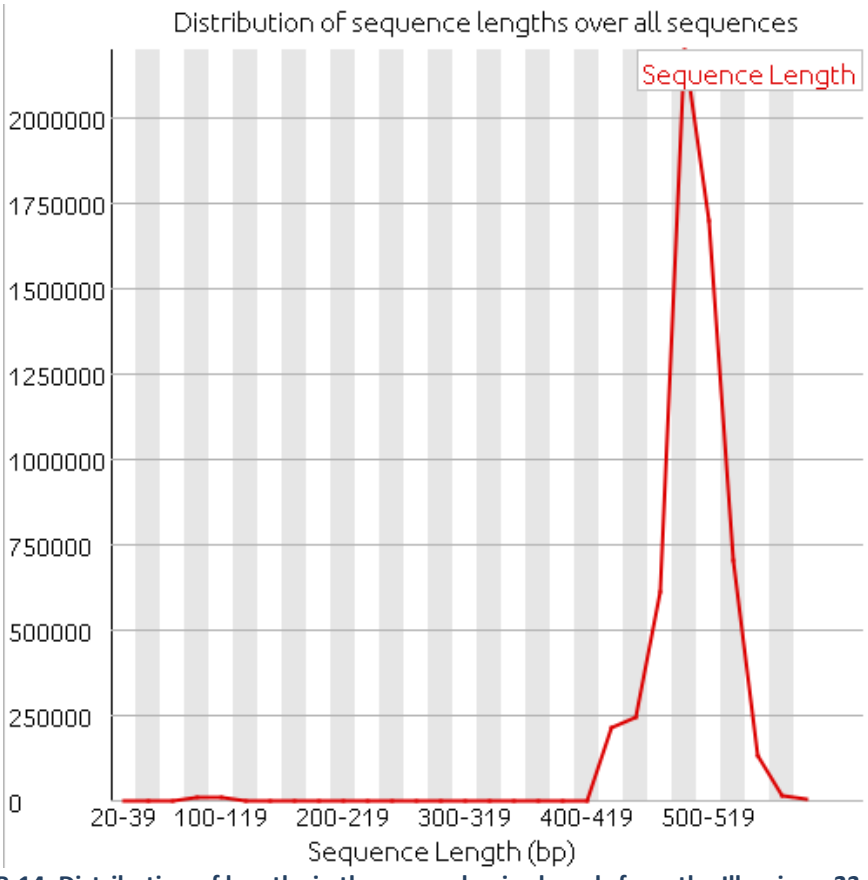


Figure B-14. Distribution of lengths in the merged paired reads from the Illumina *g23* dataset.

B.2.4 FASTQC Report on raw sequencing run of *g23* on Ion Torrent PGM

Filename	**IT_all.fastq**
File type	**Conventional base calls**
Encoding	**Sanger/Illumina 1.9**
Total Sequences	**2,688,165**
Filtered Sequences	**0**
Sequence length	**8 - 747**
%GC	**46**

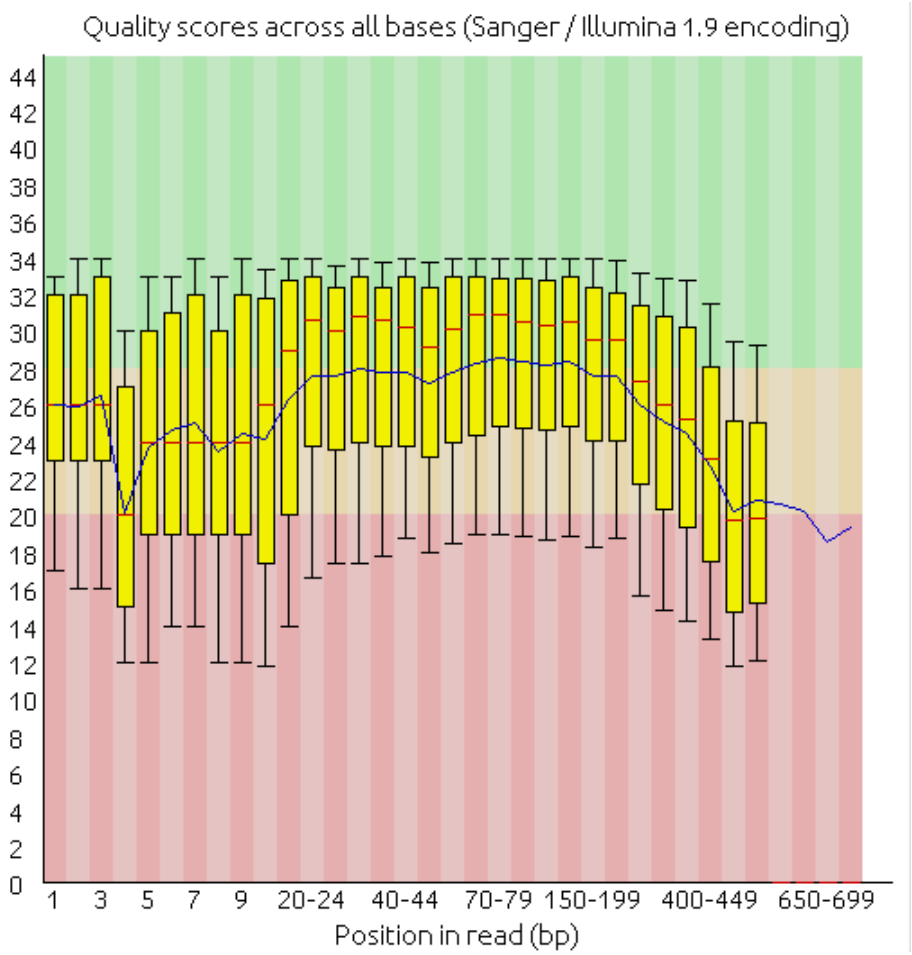


Figure B-15. PHRED scores per base position in the raw data from the Ion Torrent *g23* sequencing run.

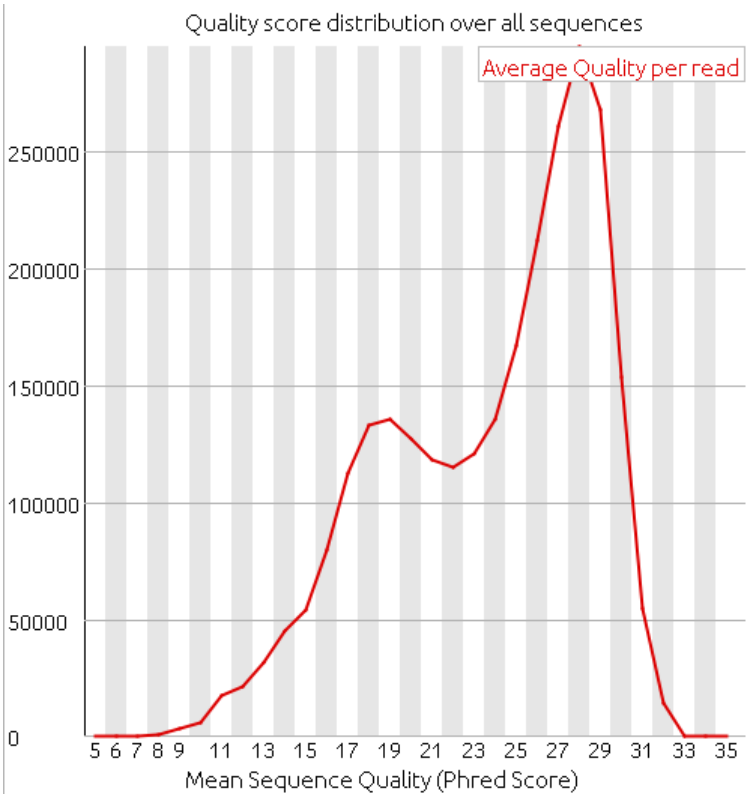


Figure B-16. Abundance of average per read PHRED scores in the raw Ion Torrent *g23* dataset.

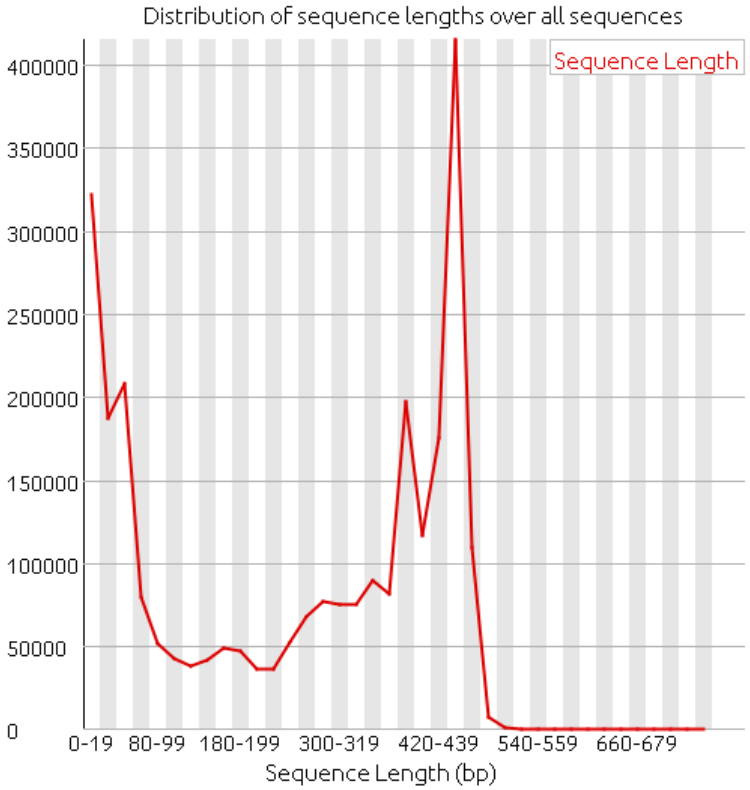


Figure B-17. Distribution of lengths in the raw Ion Torrent *g23* dataset. A large number of short reads (under 50 bp) is revealed by the distribution. Likely this owed to primer that was not entirely cleaned from the sample.

B.3 Alpha diversity measures

Table B-1. Numbers of total observed OTUs (richness) from each sample at 97% sequence identity after removal of singleton OTUs from the Roche/454 sequencing runs. Coloration of rows indicate samples within each water mass.

Sample Name	Number of *g23* OTUs	Number of *phoH* OTUs	Number of *MCP* OTUs
B16.surface	792	94	79
B16.20m	1126	85	84
B16.500m	805	67	59
B16.1000m	1025	59	40
B8.surface	730	84	39
B8.20m	766	70	60
B8.500m	869	72	53
B8.1000m	812	57	42

Table B-2. Pielou's evenness for rarefied *g23*, *phoH*, and *MCP* samples from Roche/454 datasets, expressed as a fraction of 1 (1 being the greatest evenness possible).

Sample	Pielou'sEvenness *g23*	Pielou'sEvenness *phoH*	Pielou'sEvenness *MCP*
B16. surface	0.735	0.621	0.437
B16. 20m	0.749	0.620	0.449
B16.500m	0.752	0.573	0.608
B16.1000m	0.820	0.606	0.659
B8.surface	0.729	0.589	0.339
B8.20m	0.732	0.593	0.446
B8.500m	0.731	0.611	0.381
B8.1000m	0.742	0.665	0.771

B.4 Bray-Curtis distance matrices

Table B-3. Bray-Curtis distance matrix of *g23* (Roche/454) samples derived from OTU table

	B8.surface	B8.20m	B8.500m	B8.1000m	B16.surface	B16.20m	B16.500m	B16.1000
B8.surface	0	0.07	0.11	0.13	0.51	0.51	0.6	0.69
B8.20m	0.07	0	0.11	0.13	0.49	0.49	0.61	0.69
B8.500m	0.11	0.11	0	0.14	0.53	0.51	0.59	0.68
B8.1000m	0.13	0.13	0.14	0	0.5	0.48	0.55	0.65
B16.surface	0.51	0.49	0.53	0.5	0	0.11	0.66	0.7
B16.20m	0.51	0.49	0.51	0.48	0.11	0	0.61	0.64
B16.500m	0.6	0.61	0.59	0.55	0.66	0.61	0	0.43
B16.1000m	0.69	0.69	0.68	0.65	0.7	0.64	0.43	0

Table B-4. Bray-Curtis distance matrix of *phoH* samples.

	B8.surface	B8.20m	B8.500m	B8.1000m	B16.surface	B16.20m	B16.500m	B16.1000
B8.surface	0	0.1	0.16	0.18	0.22	0.27	0.37	0.22
B8.20m	0.1	0	0.1	0.16	0.23	0.24	0.37	0.24
B8.500m	0.16	0.1	0	0.14	0.25	0.2	0.37	0.26
B8.1000m	0.18	0.16	0.14	0	0.23	0.2	0.32	0.21
B16.surface	0.22	0.23	0.25	0.23	0	0.18	0.4	0.32
B16.20m	0.27	0.24	0.2	0.2	0.18	0	0.38	0.33
B16.500m	0.37	0.37	0.37	0.32	0.4	0.38	0	0.33
B16.1000m	0.22	0.24	0.26	0.21	0.32	0.33	0.33	0

Table B-5. Bray-Curtis distance matrix of *MCP* samples.

	B8.surface	B8.20m	B8.500m	B8.1000m	B16.surface	B16.20m	B16.500m	B16.1000m
B8.surface	0	0.2	0.13	0.92	0.41	0.39	0.91	1
B8.20m	0.2	0	0.19	0.87	0.38	0.35	0.87	1
B8.500m	0.13	0.19	0	0.85	0.42	0.4	0.87	1
B8.1000m	0.92	0.87	0.85	0	0.92	0.88	0.78	0.74
B16.surface	0.41	0.38	0.42	0.92	0	0.11	0.85	1
B16.20m	0.39	0.35	0.4	0.88	0.11	0	0.84	1
B16.500m	0.91	0.87	0.87	0.78	0.85	0.84	0	0.78
B16.1000m	1	1	1	0.74	1	1	0.78	0

B.5 Rank-abundance curves

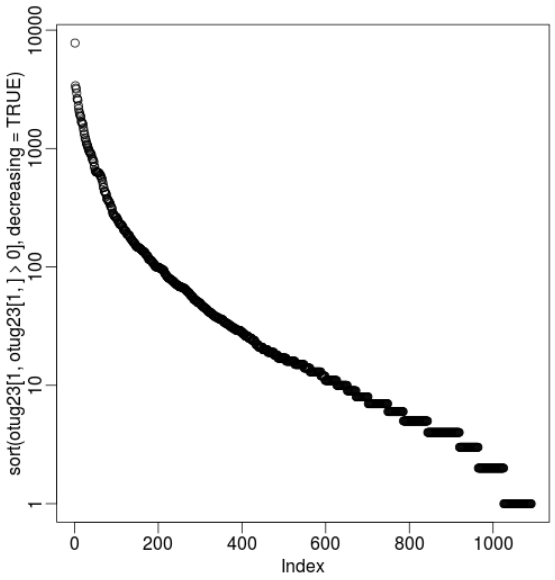


Figure B-18. Rank (x-axis) abundance (y-axis) of *g23* OTUs based on number of sequences per OTU prior to removal of singletons.

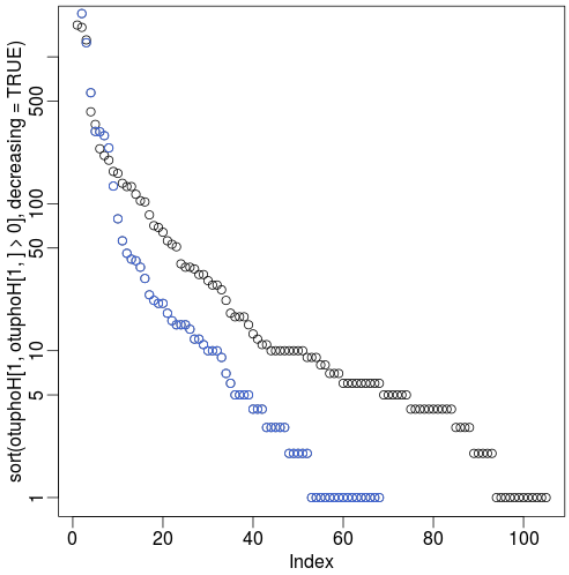


Figure B-19. Rank (x-axis) abundance (y-axis) of *phoH* (black) and *MCP* (blue) OTUs based on number of sequences per OTU prior to removal of singletons.

B.6 OTU distribution among samples and abundance

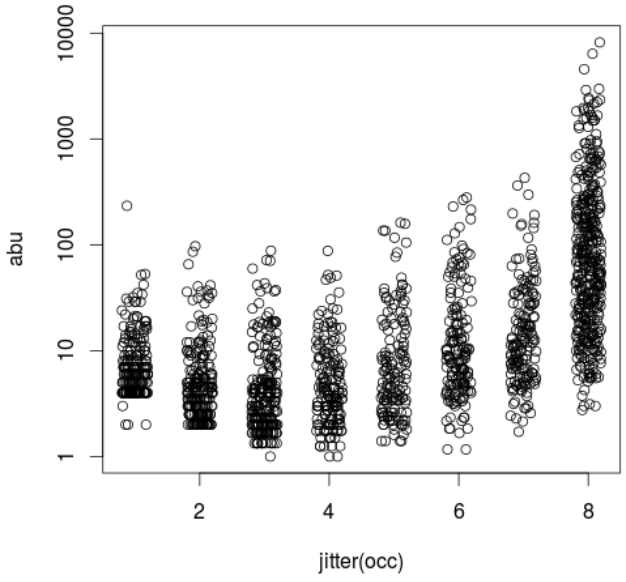


Figure B-20. Distribution of frequencies of sequences (y-axis) within *g23* OTUs which occur within all (8), several (7-2), or are unique to a sample (1) (x-axis).

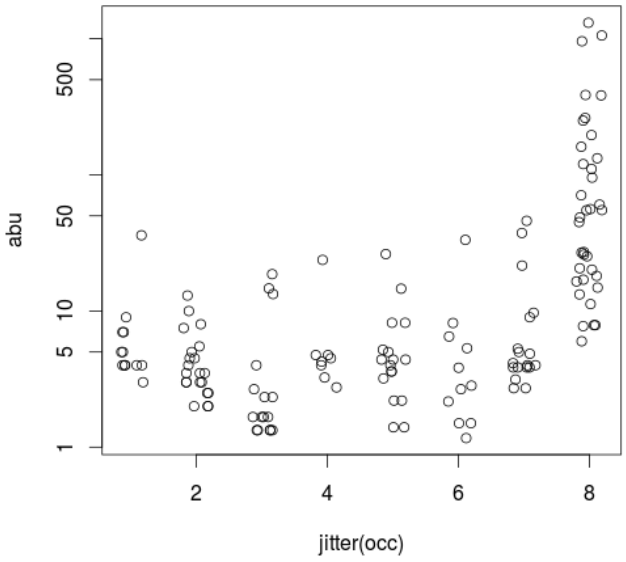


Figure B-21. Distribution of frequencies of sequences (y-axis) within *phoH* OTUs which occur within all (8), several (7-2), or are unique to a sample (1) (x-axis).

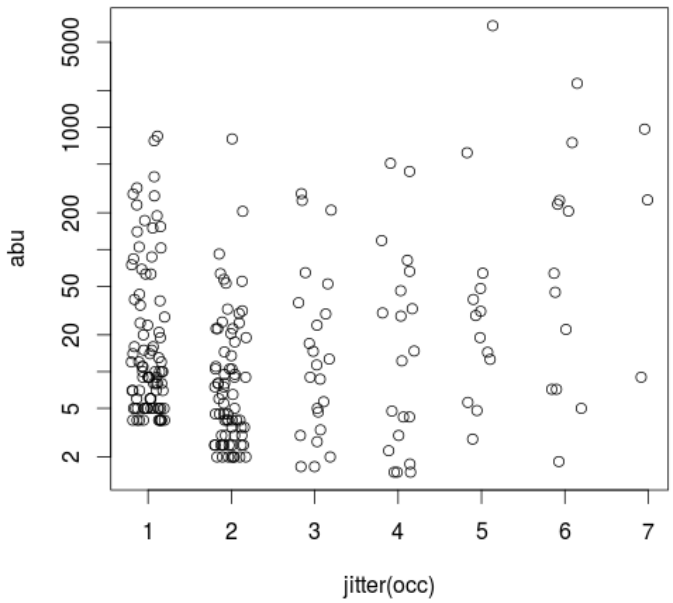


Figure B-22. (to the left) Distribution of frequencies of sequences (y-axis) within *MCP* OTUs which occur within several (7-2), or are unique to a sample (1) (x-axis).

B.7. ANOSIM outputs from QIIME script compare_categories.py

Table B-6. Test of platform comparison data based on grouping by sequencing platform

method name	ANOSIM
test statistic name	R
sample size	24
number of groups	3
test statistic	-0.026909722222222359
p-value	0.55300000000000005
number of permutations	999

Table B-7. Test of platform comparison datasets based on grouping by water masses

method name	ANOSIM
test statistic name	R
sample size	24
number of groups	4
test statistic	0.68788069733249313
p-value	0.001
number of permutations	999

Table B-8. Test of Roche/454 *g23* dataset based on grouping by water masses

method name	ANOSIM
test statistic name	R
sample size	8
number of groups	4
test statistic	0.91304347826086973
p-value	0.0030000000000000001
number of permutations	999

Table B-9. Test of Roche/454 *phoH* dataset based on grouping by water masses

method name	ANOSIM
test statistic name	R
sample size	8
number of groups	4
test statistic	0.095652173913043495
p-value	0.313
number of permutations	999

Table B-10. Test of Roche/454 *MCP* dataset based on grouping by water masses

method name	ANOSIM
test statistic name	R
sample size	8
number of groups	4
test statistic	0.30434782608695649
p-value	0.14499999999999999
number of permutations	999

B.8. Mantel Test output of correlation between *g23* and *16S* datasets

Table B-11. Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis =bact.dist, ydis= *g23*.dist)

Mantel statistic r: 0.2744
Significance: 0.155

Upper quantiles of permutations (null model):

90%	95%	97.5%	99%
0.343	0.479	0.571	0.599

Permutation: free
Number of permutations: 999