5-12-2014

Marine Viral Diversity and Spatiotemporal Variability

Dawn Goldsmith
University of South Florida, dawn.goldsmith@gmail.com

Marine Viral Diversity and Spatiotemporal Variability

by

Dawn B. Goldsmith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
College of Marine Science
University of South Florida

Major Professor: Mya Breitbart, Ph.D.
John Paul, Ph.D.
Gary Mitchum, Ph.D.
Kathleen Scott, Ph.D.
Craig Carlson, Ph.D.

Date of Approval:
May 12, 2014

Keywords: phage, phoH, Bermuda Atlantic Time-series Study, signature gene

DEDICATION

I dedicate this work to my husband, Warren Firschein. He was willing to pack up, sell our house, move to a new state with our two kids (then ages 1 and 4), and telecommute indefinitely (which he does not enjoy) in order for me to pursue my dream. He did all this with (mostly) good humor, although I have noticed that his consumption of wine and ice cream has increased over the years. During the tougher parts of graduate school, he held me up. He is not a scientist, but that does not deter him from talking through issues with me to figure out the best approach to solve a problem, or the best way to improve a graph. Even before I started graduate school, he thought of creative ways to encourage and assist me. While I was taking one of my first biology classes, and struggling to understand an upcoming lab, he read the lab manual and then brought me a glass of orange Gatorade to help explain spectrophotometry. Overall, Warren put up with lots of stress, numerous 12-day absences while I conducted field work, and foster dogs falsely advertised as housetrained. I could not have done this without him, and I am forever grateful.

ACKNOWLEDGMENTS

My heartfelt thanks go to Mya Breitbart for taking me on, for sticking with me longer than either of us expected it would take, and for being a terrific mentor. She was clearly paying attention in advisor school—except for the day they talked about “among” and “amongst.” Thanks also to my committee for their insight and encouragement: John Paul, Gary Mitchum, KT Scott, and Craig Carlson. I am also grateful to all the members of the Breitbart Lab, for support and camaraderie both in the lab and in the field. Thank you to Ruth-Anne Sandaa, who gave me phoH; Rachel Parsons, for invaluable assistance during my field work at the Bermuda Institute of Ocean Sciences; and David Jones, for training in statistics. My funding came from the University of South Florida Presidential Doctoral Fellowship, the College of Marine Science Von Rosenstiel Endowed Fellowship, and Mya Breitbart. Finally, I thank my family for their endless encouragement: Warren Firschein, Sophie Goldsmith, Elena Firschein, Linda Goldsmith, and Donald Goldsmith, who lives forever in my heart.

TABLE OF CONTENTS

List of Tables iii

List of Figures iv

Abstract vi

Introduction 1
- Abundance of marine viruses 1
- Importance of marine viruses 1
- Significance of viral diversity 3
- Seasonal studies of viral dynamics 6
- Interannual studies of viral dynamics 8
- Depth studies 10
- Study site: Site of the Bermuda Atlantic Time-series Study 13
- Viruses at BATS 17

Chapter 1: Development of phoH as a novel signature gene for assessing marine phage diversity 20

Chapter 2: Deep sequencing of the viral phoH gene reveals seasonal variations, depth-specific composition, and persistent dominance of the same phage phoH genes in the Sargasso Sea 21
- Summary 21
- Introduction 22
- Results 24
- Discussion 30
- Experimental procedures 39
  - Sample collection and DNA extraction 39
  - Amplification of the *phoH* gene 40
  - DNA quantification and sequencing 41
  - Sequence analysis 42
- Acknowledgements 44
- Tables and figures 45

Chapter 3: Depth and seasonal variation in viral diversity in the northwestern Sargasso Sea 54
- Summary 54

Introduction..55
Methods...57
Sample collection and processing...57
Amplification of signature genes...58
Cloning and sequencing of signature genes...59
Data analysis for signature genes...60
RAPD PCR amplification and data analysis..61
Discussion of methods..61
Results and discussion..63
Constraints of present study..68
Conclusion...69
Tables and figures...71

Conclusion 77

References...82

Appendix A: Copyright Permission...94

Appendix B: Development of phoH as a novel signature gene for assessing marine phage diversity...95

LIST OF TABLES

Table 2.1: F values for pairwise comparisons of depths by permutational MANOVA..........48

Table 2.2: Barcodes used to tag phoH amplicons..53

Table 3.1: Sample collection data including viral counts and number of sequences obtained from each sample 71

LIST OF FIGURES

Figure 2.1: Rarefaction curves for phoH sequences from all 85 depths/times; OTUs are defined by sequence identity of 97% or greater 45

Figure 2.2: Heatmap displaying Chao1 minimum richness estimate for all dates and depths except March 2010, 20 m 46

Figure 2.3: Heatmap displaying inverse Simpson index diversity estimate for all dates and depths 46

Figure 2.4: Dendrogram illustrating hierarchical clustering of all 85 depths/times 47

Figure 2.5: OTU composition of total sequences and individual samples 49

Figure 2.6: Percent of sample sequences falling in an OTU versus OTU 50

Figure 2.7: Prevalence of phylogenetically clustered OTUs in each sample, indicated as percent of each sample's sequences that come from each of the 94 top OTUs 51

Figure 2.8: Percent of each sample belonging to phylogenetic Groups 1 through 5................52

Figure 3.1: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on g23 OTUs (98% sequence identity)..72

Figure 3.2: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on phoH OTUs (98% sequence identity)..72

Figure 3.3: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on ssDNA MCP OTUs (98% sequence identity) 73

Figure 3.4: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on gel photos of RAPD PCR products 73

Figure 3.5: Phylogenetic tree showing the relationship among g23 sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes) and g23 sequences from cultured phages and other environmental samples (indicated by names) 74

Figure 3.6: Phylogenetic tree showing the relationship among phoH sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes), phoH sequences from cultured phages and viruses (indicated by names), and placeholder sequences from the Sargasso Sea (indicated by names beginning with “Ph_”) 75

Figure 3.7: Phylogenetic tree showing the relationship among ssDNA MCP sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes) and ssDNA MCP sequences from fully sequenced phage (indicated by names) 76

ABSTRACT

Marine viruses are the most numerous biological entities in the ocean, with an estimated abundance of 4 x 10^{30}. They merit study not only because of their sheer abundance, but also because of the role they play in the Earth's biogeochemical cycles. Viral lysis of bacteria redirects the flow of nutrients among marine microbes, which ultimately affects the efficiency of the biological pump. Viral diversity is important because most viruses are host-specific. In preying on a certain type of bacteria, viruses affect the diversity and structure of the bacterial community, leading to changes in carbon and nutrient flows. In turn, such variations can alter the amount of carbon dioxide in the Earth's atmosphere. However, studying viral diversity presents challenges. Morphological similarities among many types of viruses make it preferable to use genetic methods of investigation, but the absence of a single gene common to all families of viruses hampers the identification of viruses in environmental samples. Nonetheless, some genes are shared within phage families, and those shared ("signature") genes can be used as markers to identify members of a family. In addition, community profiling methods can fingerprint the diversity of a viral community.

Most previous studies of marine viral communities consist of a single glimpse—a representation of the community at a single time and place, or at a few depths sampled at one time. While the resources required to collect marine samples often make broader or repeated sampling impracticable, without studies conducted over greater time and spatial ranges, our knowledge of marine viral dynamics will remain limited. To gain strides in understanding spatial and temporal variability in marine viral diversity, this dissertation focused on a detailed examination of viral diversity at a single site in the Sargasso Sea. Time and depth intervals for sampling were kept as uniform as possible in order to strengthen the conclusions to be drawn from the research.

The Sargasso Sea is a seasonally oligotrophic portion of the North Atlantic Ocean, characterized by deep convective winter mixing and summer stratification of the water column. A tremendous amount of oceanographic research has been conducted in the Sargasso Sea because it is home to the Bermuda Atlantic Time-series Study (BATS), one of the world's longest-running ocean time series studies. Because of the core monthly measurements made at the BATS site and the vast amount of ancillary research that uses BATS as a platform, the site is an excellent place to study viral diversity. Using a variety of techniques, this research aimed to expand our knowledge of viral dynamics by analyzing the viral community of the Sargasso Sea over a several-year period, through different seasons, and at different depths.

The first chapter developed phoH as a new signature gene for assessing marine viral diversity. The phoH gene is disproportionately present in fully-sequenced marine phage, as opposed to phage isolated from non-marine environments, and is widespread in the marine environment. Diversity of the phoH gene was high, and most of the sequences recovered belonged to phylogenetic groups that did not contain any cultured representatives, indicating that cultured phage isolates do not adequately represent the diversity found in marine environments. Composition of the phoH communities at each sampled location and depth was distinguishable according to phylogenetic clustering, although most phoH clusters were recovered from multiple sites. These factors demonstrate that phoH will be useful for studying marine phage diversity worldwide.

Chapter 2 analyzed the viral diversity of a depth profile at BATS by amplifying and deep sequencing the phoH gene. This comprehensive study of the gene's diversity over three different years, several seasons, and a range of depths from the surface to 1000 m revealed that the viruses at BATS contain a large pool of phoH sequences, but that most of those sequences are rare. The phoH sequences were dominated by just a few operational taxonomic units (OTUs). Rarefaction analysis showed that the sequencing was sufficient to capture the diversity of the gene at BATS, and in fact no new phylogenetic clusters were identified that were not seen in the small amount of Sanger sequencing performed for the initial phoH study in Chapter 1. Some of the more abundant phoH OTUs recurred every season and every year, in varying degrees, although similar depths and seasons clustered together. Overall, the phoH gene revealed depth-based, seasonal, and interannual differences in the diversity of the viral community at BATS.

Chapter 3 continued the extensive examination of viral diversity at BATS by using several signature genes and a fingerprinting technique to assess changes between winter and summer viral communities over two depths in three different years. This chapter investigated whether the annually recurring subsurface peak in viral abundance corresponded to recurring changes in composition of the viral community in the vicinity of the peak. Clustering analysis was used to determine which samples were most similar. The results demonstrated that the viral communities at the surface and at 100 m depth were more similar to each other in winter (March), regardless of the year, than they were in summer (September), when the water column is stratified as opposed to well-mixed. These findings may stem from physical factors such as UV irradiation of viral particles during stratification, as well as seasonal and depth-related differences in host communities associated with the depth of the mixed layer.

This dissertation provides substantial advances to the field of microbial ecology. First, the development of phoH as a signature gene is an important addition to the limited set of tools available for studying marine viral diversity. This research also constitutes the first deep sequencing of a signature gene for marine viruses, providing a guide for the depth of sequencing needed to capture the diversity of a marine viral community and a benchmark for the level of viral diversity to expect in an oligotrophic marine system. Finally, the dissertation expands our knowledge of the viral community at BATS by examining the community based on four different measures of composition, rather than abundance. The research presented here also suggests several avenues of future investigation, including redesigning the phoH primers to expand their scope, sampling the viral community at BATS at the precise depth of the peak in abundance, working to identify the hosts of aquatic gokushoviruses, and culturing and sequencing additional marine viruses in order to improve the reflection of natural environmental communities in genomic databases.

INTRODUCTION

Abundance of marine viruses

Marine viruses merit study not only because of their sheer abundance, but also because of the role they play in the Earth's biogeochemical cycles. While bacteria are the most numerous living organisms on the planet—it is estimated that bacteria in the ocean number 1.2 x 10^{29} (Whitman et al., 1998)—viral abundance can exceed the number of bacteria by an order of magnitude. The virus-to-bacteria ratio in the oceans varies, but usually within a fairly narrow range, and often depends on habitat and season (Breitbart, 2012). While values ranging from less than 0.1 to greater than 50 have been reported (encompassing coastal, open ocean, and estuarine areas), the general range for open ocean surface waters is between 2 and 25 (Bratbak et al., 1994; Fuhrman, 1999; Wommack and Colwell, 2000; Suttle, 2007). Through a decade of viral and bacterial counts in the Sargasso Sea, one of the main study sites in this dissertation, the virus-to-bacteria ratio was maintained within a narrow range of 3-20 in the vast majority of the measurements in the upper 300 meters of the water column (Parsons et al., 2012). Based on extrapolation from bacterial abundances, the total number of marine viruses—the oceans' most abundant biological entities—is estimated at 4 x 10^{30} (Fuhrman, 1999; Suttle, 2005).

Importance of marine viruses

Given their vast numbers, it should not be surprising that marine viruses play a significant role in carbon and nutrient cycling. Most of the carbon and nutrients in the ocean are in the dissolved phase, and this pool of dissolved organic matter (DOM) is taken up by bacteria

(Pomeroy, 1974). When bacteria are grazed upon by protists, the DOM moves up the food web to organisms at higher trophic levels (Azam et al., 1983). However, viruses divert that flow, in a process known as the viral shunt (Wilhelm and Suttle, 1999). When viruses infect and then lyse bacteria, not only are new viral particles released, but carbon is also released, and returns to the pool of DOM. Other organic cellular compounds such as proteins and nucleic acids are also emitted into the dissolved phase, where bacteria can incorporate their nitrogen and phosphorus. In this way, viruses redirect the flow of nutrients among marine microbes (Bratbak et al., 1994; Wilhelm and Suttle, 1999). If viral predation represented only a small fraction of bacterial mortality, then the effect of viral lysis on carbon and nutrient cycling would be correspondingly small. There is always some degree of uncertainty in estimates of mortality caused by viruses, because each method of determining such mortality requires some assumptions to be made (Suttle, 2005). However, there is evidence that as much as half of bacterial mortality is caused by viruses (Fuhrman and Noble, 1995), although the percentage of the bacterial community killed by viruses rather than grazers can vary depending on season, location, and habitat (Fuhrman, 1999).

Because viruses are responsible for a substantial proportion of bacterial mortality, the viral shunt has significant effects on oceanic biogeochemistry as well as atmospheric chemistry. When bacteria are grazed upon by protists, which in turn are eaten by zooplankton, the carbon and nutrients that make up the bacteria enter the food web of larger marine organisms (Fenchel, 1988). However, when viruses lyse cells, cellular contents are returned to the pool of DOM, resulting in a greater amount of carbon respiration occurring in the surface ocean, which contributes additional carbon dioxide to the atmosphere (Suttle, 2005). Thus by converting carbon to the dissolved phase, viral lysis can affect the efficiency of the biological pump, depending on the rate at which carbon is supplied relative to the rate of supply of nutrients needed for the growth of primary producers (Suttle, 2007).

Significance of viral diversity

It is essential to understand not only what viruses are doing in the ocean, but also which viruses are present. Viral diversity is important because most viruses are host-specific: while some cyanophage can infect two genera of cyanobacteria (Sullivan et al., 2003), most viruses have a more limited host range, infecting just one species, one subspecies, or a few related species (Bratbak et al., 1994; Fuhrman, 1999; Suttle, 2007; Holmfeldt et al., 2013). Thus as bacterial populations change, so do virus populations (and vice versa). Under the kill-the-winner model of viral dynamics (Thingstad and Lignell, 1997), viruses control the populations of the most active (or would-be active) hosts. When a particular host population proliferates, viruses capable of infecting that host will increase in abundance. As those viruses prey upon the susceptible host, the population of the host declines. When a new host emerges (resistant to the viruses that killed the first host) to fill the now-vacant niche left by the originally dominant host, the first virus population will decline. As viruses evolve to infect the newly dominant host and increase in abundance, the progression continues (Thingstad and Lignell, 1997; Winter et al., 2010). In preying on a certain type of bacteria, viruses affect the diversity and structure of the bacterial community. Changes in the bacterial community can lead to changes in carbon and nutrient flows, resulting in changes in the net heterotrophy of the ocean (Fandino et al., 2001). In turn, such variations can alter the amount of carbon dioxide in the Earth's atmosphere (Raymond et al., 2000). Evidence has shown that the viral community can control the composition of the bacterial community (Mühling et al., 2005). But most experiments designed to determine the degree to which viruses affect bacterial community composition show variable results, likely due to results being obscured by manipulation or containment effects (Schwalbach et al., 2004; Winter et al., 2004; Hewson et al., 2006; Bouvier and Del Giorgio, 2007).

Marine viral diversity is significant and influential; unfortunately, it is a difficult subject to study. Viruses can be counted through nucleic acid staining and flow cytometry (Marie et al., 1999) or epifluorescence microscopy (Noble and Fuhrman, 1998), but individual viruses cannot be distinguished beyond subgroups based on levels of fluorescence and light scattering (Suttle, 2007). Small viruses such as RNA viruses and single-stranded DNA viruses usually escape detection as well due to their small genome size (Suttle, 2007). Transmission electron microscopy reveals the morphology of individual viruses, but since different viruses can have similar morphology, this technique cannot definitively identify a viral species (Proctor, 1997). For these reasons, genetic methods of characterizing viral diversity are preferable. Identification of viruses in environmental samples is hampered by the absence of a gene common to all viruses (Rohwer and Edwards, 2002); however, some genes are shared within phage families, and those shared ("signature") genes can be used as markers to identify members of a family (Suttle, 2005). A variety of signature genes can capture subsets of viral diversity, such as the DNA polymerase gene for podophage (Huang et al., 2010), capsid genes for myophage (Jameson et al., 2011; Chow and Fuhrman, 2012) and single-stranded DNA phage (Hopkins et al., 2014), *psbA* for viruses of photosynthetic bacteria (Chenard and Suttle, 2008), and *phoH*, which was developed as part of this dissertation and can capture viruses in multiple families that infect both heterotrophic and autotrophic hosts (Chapter 1; Goldsmith et al. (2011)). Amplified signature genes can either be sequenced directly or the diversity of specific signature gene amplicons can be profiled using techniques such as terminal restriction fragment length polymorphism (T-RFLP) (Jiang et al., 2003; Wang and Chen, 2004; Needham et al., 2013) or denaturing gel gradient electrophoresis (DGGE) (Short and Suttle, 1999; Mühling et al., 2005). Additionally, techniques for fingerprinting the viral community diversity based on specific sequences or genome sizes exist, using randomly-amplified polymorphic DNA (RAPD) PCR (Comeau et al., 2004; Winget and Wommack, 2008) or pulsed field gel electrophoresis (PFGE) (Wommack et al., 1999; Steward et al., 2000; Larsen et al., 2001; Hewson et al., 2006), respectively.

Most previous studies of marine viral communities consist of a single glimpse—a representation of the community at a single time and place, or at a few depths sampled at one time (Breitbart, 2012). While the resources required to collect marine samples often make broader or repeated sampling impracticable, without studies conducted over greater time and spatial ranges, our knowledge of marine viral dynamics will remain limited. For example, viruses in surface waters are not necessarily representative of deeper viral communities (Zhong et al., 2002; Parsons et al., 2012). Temporal variations can also be substantial; when Bergh et al. (1989) first reported the abundance of marine viruses to be far greater than originally thought, they also found that viral counts in the Barents Sea in the productive part of the year exceeded winter abundance by nearly three orders of magnitude. Abundance is not the only facet of the viral community that can change with time. Viral assemblages sampled in one season may well differ from assemblages collected at a different time of year. For that reason, sampling over multiple years is critical in order to elucidate annually recurring trends. Thus in order to understand viral ecology and viral interactions with hosts, the first step is to understand how viral communities change. Because we know that marine bacterial communities are dynamic rather than static (Giovannoni and Vergin, 2012), analysis of viral dynamics may lead to understanding how each influences the other.

Seasonal studies of viral dynamics

Numerous studies have investigated temporal variations in marine viral community composition throughout the world, using a variety of techniques. For example, Winget and Wommack (2008) used randomly-amplified polymorphic DNA (RAPD) PCR to assess the diversity of the Chesapeake Bay viral community on a seasonal scale. Differences in RAPD PCR banding patterns from samples drawn from the same station six months apart demonstrated that the composition of the viral community had changed. This study confirmed the results of an earlier study in the Chesapeake Bay, in which pulsed field gel electrophoresis (PFGE) was used to assess viral community structure. Wommack et al. (1999) sampled six stations in the Bay four times over a year, and reported variation over time in the frequency distribution of the viral genome sizes.

A similar result was observed for the viral community in coastal California water. Eight samples drawn between May and October 1997 exhibited different PFGE banding patterns over the six-month sampling period (Steward et al., 2000). The number of discrete bands (indicating viral genome sizes) was at its highest in May, and was at its lowest in October. Others have similarly reported that viral richness varies by season. Sandaa and Larsen (2006) used PFGE to examine seasonal changes in viral diversity off the coast of Norway. They reported the least number of average PFGE bands in August, while the highest average number of bands was found in June. Moreover, the composition of the viral community varied depending on the season. Clasen et al. (2013) monitored the diversity and composition of cyanomyophage communities in coastal Southern California for 15 months. They isolated cyanophage from monthly surface water samples and amplified two signature genes, g20 and psbA. Seasonal variation based on those genes was evident: in winter and fall, the cyanomyophage community was dominated by a single OTU of each gene, while in summer, the community appeared to be in flux and the dominant OTU often changed.

Other techniques have also been used to assess seasonal changes in phage diversity. Denaturing gel gradient electrophoresis (DGGE) analysis of the g20 signature gene revealed seasonal variation in phage diversity in the Gulf of Aqaba (Mühling et al., 2005). The greatest number of DGGE types appeared in May, while the lowest numbers appeared in July and October. Using the same gene, other researchers found seasonal variations in the abundance and diversity of phage isolates from coastal Rhode Island waters (Marston and Sallee, 2003). In summer months, they found a greater number of different genotypes than they found during winter months. However, they did not see specific genotypes appear only in summer and disappear in winter, or vice versa. Seasonal variations may not appear if the time frame of the study is too short. A study of viral diversity in coastal Denmark over three months found that despite large changes in viral abundance, the viral community structure (as shown by PFGE) was relatively stable (Riemann and Middelboe, 2002). Similarly, Fuhrman et al. (2002) reported that PFGE fingerprints of the viral community in Southern California waters showed temporal stability between August and October 2000.

A study in the Eastern Mediterranean Sea compared viral communities sampled during mixing season and again while the water column was stratified (Magiopoulos and Pitta, 2012). Viral abundance varied with the season, as did the contribution to the total viral community of three individual groups of viruses, designated by the strength of their fluorescence signal. Another study in the Mediterranean Sea sampled a site nine times over the course of one year. Using flow cytometry, Winter et al. (2009) measured viral abundance, and also distinguished among three viral types based on fluorescence. They reported seasonal changes in viral abundance, and in composition of the viral communities (according to the three types).

Variations in the composition of marine viral communities also appear on shorter time scales. A TRFLP analysis of the g23 gene amplified from viral DNA extracted from samples collected near an island off the coast of Southern California revealed 153 myoviral OTUs over the course of a 78-day time series (Needham et al., 2013). Examination of the relative abundance of each OTU showed that while most of the OTUs appeared in less than 25% of the 45 samples collected during the time series, more than 80% of the viral community consisted of OTUs that appeared in at least 90% of the samples. Some OTUs with the highest average abundance exhibited a wide range of relative abundances during the course of the experiment. For example, one OTU comprising approximately 25% of the total myoviral community on the first day of the study made up less than 5% of the community at the end of the experiment. Overall, the myoviral community was fairly stable on a scale of weeks to months, while the scale of days to weeks showed greater variation in community composition (Needham et al., 2013).

Interannual studies of viral dynamics

In order to determine whether any observed seasonal patterns recur, a small number of studies have begun to examine changes in viral community composition on an interannual scale. Jiang et al. (2003) found by RFLP analysis that phage isolates from the coast of southern California in August 1999 were similar to phage infecting the same hosts one year later. Hewson et al. (2006) used PFGE to examine changes in viral diversity in the Gulf of Mexico from 2001 to 2003. Richness of the viral community varied from 3 to 20 phylotypes over the three years. However, because of a change in the method of sample preparation, the authors could not rule out the possibility that the difference resulted from the change in protocol rather than reflecting a real change in viral diversity.

Using the genome of EhV-86, a virus of coccolithophore *E. huxleyi*, Allen et al. (2007) employed a microarray to analyze the diversity of 14 strains of coccolithovirus isolates. The findings revealed temporal clustering: virus strains isolated from the English Channel in 1999 grouped separately from the strains isolated in the same location in 2001. In a hypersaline system, Emerson et al. (2012) observed that viral assemblages were relatively stable on a time-scale of days, but more dynamic over a period of nearly three years. However, the number of viral populations remained relatively stable over that period (Emerson et al., 2013).

Chen et al. (2009) sampled the cyanopodophage community of the Chesapeake Bay during winter and summer for two years and discovered repeating seasonal differences: winter phage communities sampled in different years grouped more closely with each other than with summer phage communities from the same year. Jamindar et al. (2012) also observed seasonal changes in g23 revealed by PFGE of samples from the Chesapeake Bay and Delaware Bay. Cyanomyophage communities in paddy field soil change over time too, as demonstrated by a study analyzing the g20 gene from samples collected over a three-year span (Wang et al., 2011). Further north, cyanophage from the surface waters of Narragansett Bay also exhibited seasonal patterns of community structure. Marston et al. (2013) sampled every month for six years, then isolated cyanophage from the samples and amplified a myoviral DNA polymerase gene (g43) from the isolates. Similarity analysis of 31 of those isolates showed clustering according to season; composition of the cyanomyophage community was more similar to the composition of samples from the same season of any year than a different season of the same year.

Sampling every month for three years enabled Chow and Fuhrman (2012) to see that T4-like myophage communities at the site of the San Pedro Ocean Time-series displayed seasonally recurring patterns of diversity. Some OTUs (revealed by T-RFLP conducted on g23 amplicons) peaked in spring or summer, while others peaked in fall or winter. Communities 3-7 months apart were negatively correlated, while communities from adjacent months were highly correlated, as were communities from the same month one year apart. For two years, Pagarete et al. (2013) took monthly samples from the water at 5 m depth in Raunefjorden, Norway. Like Chow and Fuhrman, these investigators studied changes in the myoviral community using T-RFLP analysis of the g23 gene. The sampling resolution and length of the study enabled them to distinguish three different viral communities depending on season: summer, fall, and winter/spring each harbored distinct communities (Pagarete et al., 2013).

Depth studies

In a long-term study of the upper 300 m of the water column at the site of the Bermuda Atlantic Time-series Study (BATS) in the Sargasso Sea, Parsons et al. (2012) demonstrated that both time and depth are important for understanding the dynamics of a marine viral community. The results of ten years of monthly sampling revealed annually recurring seasonal patterns, in which viral abundance peaked every summer between 60-100 m depth. However, this study did not analyze the composition of the viral community over depth and time, which is a limitation that this dissertation overcomes.

Several previous studies have examined viral community composition and dynamics over a depth profile, with variable results. In a range of locations, viral community composition varied with depth, and in some cases was related to stability of the water column. A study by Wilson et al. (1999) investigated viral diversity along a meridional Atlantic Ocean transect from the Falkland Islands to the United Kingdom. This study used DGGE to examine the structure of the cyanophage population in two depth profiles. At one station, where the water column was well mixed, samples drawn from six depths in the upper 100 m revealed that the cyanophage population structure was similar throughout the water column. In the other depth profile, the water column was stratified, and the structure of the cyanophage population was variable throughout the profile (Wilson et al., 1999). Zhong et al. (2002) amplified and analyzed g20 sequences and discovered that phage population structure was different in Sargasso Sea surface waters than at the depth of the deep chlorophyll maximum (DCM). This was also true in the Gulf of Mexico, where myophage at the surface were distinguishable from myophage at the DCM. In British Columbia, DGGE analysis of g20 amplicons showed that different cyanomyophage communities resided at different depths in the Straits of Georgia (Frederickson et al., 2003). Differences of only a few meters in depth resulted in shifts in community composition. While some cyanomyophage appeared throughout the water column, others were found at only a few depths. In the Eastern Mediterranean Sea, Magiopoulos and Pitta (2012) sampled a variety of depths encompassing the epi-, meso-, and bathypelagic layers, and characterized three groups of viruses according to fluorescence level (high, medium, and low). Viral community composition varied by depth, and during stratification of the water column, viral abundance increased at all depths. Winter et al. (2009) also explored depth-related differences in viral community composition in the Mediterranean Sea. High-fluorescence viruses preferentially occupied the surface waters versus deeper waters, but low- and medium-fluorescence viruses showed no preference. The epipelagic layer had significantly higher viral abundance than either the meso- or bathypelagic layers.

Other studies show mixed results. In their study of Pacific near-coastal waters, Jiang et al. (2003) observed some differences in marine viral communities with depth. Viral abundance decreased in the top 200 m, while from 200 m to the bottom (890 m), there were relatively little changes in abundance. Bands identified by PFGE were slightly larger at deeper depths. However, similar phage were isolated from both the surface and the bottom of the water column, suggesting that the distribution of viruses did not vary with stratification of the water column (Jiang et al., 2003). Another study of viral diversity in the waters of Southern California used PFGE and found few differences in viral diversity throughout the top 45 m (Fuhrman et al., 2002). In Danish coastal waters, a depth profile revealed that the structure of the viral community showed no significant changes with depth (Riemann and Middelboe, 2002). An assessment of viral diversity in the Chesapeake Bay used RAPD PCR to analyze samples drawn from both the top and the bottom of the water column (1 m below the surface and 2 m above the sediment-water interface) (Winget and Wommack, 2008). Comparison of the banding patterns reflecting the diversity of the viral communities at the surface and the bottom revealed highly similar viral communities.

The above studies were conducted at myriad sites and over a variety of depths and time intervals. These differences prevent direct comparisons of the results and limit our ability to draw general conclusions from the investigations. To gain strides in understanding spatial and temporal variability in marine viral diversity, this dissertation focused on a detailed examination of viral diversity at a single site, the BATS site (described below). Time and depth intervals for sampling were kept as uniform as possible in order to strengthen the conclusions to be drawn from the research.

Study site: Site of the Bermuda Atlantic Time-series Study

The primary sampling site for this project is the BATS site in the northwestern Sargasso Sea. BATS was launched in 1988 as part of the United States Joint Global Ocean Flux Study (JGOFS) (Michaels and Knap, 1996). The goal of BATS is to study seasonal and interannual variations in the biogeochemistry of this region (Michaels and Knap, 1996). The BATS program, along with the Hawaii Ocean Time-series program (HOT), also established by JGOFS in 1988, are designed to study the flux of carbon between the ocean and the atmosphere (Karl and Lukas, 1996). Time-series studies in the Sargasso Sea actually began well before the initiation of the BATS program. Beginning in 1954, biweekly measurements of temperature, salinity, and oxygen have been collected at Hydrostation S, which is 26 km southeast of Bermuda. Since then, other data began to be collected as well, including measurements of nutrients, chlorophyll, primary productivity, particle fluxes, and atmospheric chemistry (Michaels and Knap, 1996).

The BATS site is approximately 85 km southeast of Bermuda. Sampling occurs at least once a month, with additional cruises added during blooms. For the first few years after BATS was established, measurements were taken near a drifting sediment trap array, which was deployed at a set location but often drifted between 25 and 75 km from the deployment site. Beginning in July 1994, the location of trap deployment was moved, and casts from which core measurements are taken are now drawn within a few kilometers of the deployment site (Michaels and Knap, 1996). Among the many core measurements taken at BATS are salinity, macronutrients, fluorescence, oxygen, alkalinity, total CO_2, dissolved and particulate organic carbon and nitrogen, dissolved organic phosphorus, chlorophyll *a*, bacterial counts, bacterial production, primary production, and particle fluxes (Michaels and Knap, 1996). Some variables are measured only in the upper 100-200 m, while other measurements are taken along a depth profile down to 4200 m (Steinberg et al., 2001; Lomas et al., 2013).

In addition to the core monthly measurements, a vast amount of ancillary research has been conducted that takes advantage of the time-series infrastructure at BATS and Hydrostation S. In particular, research at BATS contributes to a better understanding of biogeochemical cycles and the biological pump, leading to improved modeling of oceanic processes and knowledge of how the ocean responds to climate change (Ducklow et al., 2009). Toward this end, studies of bacterial dynamics, nitrogen and phosphate cycling, phytoplankton community structure, optics and remote sensing, and zooplankton dynamics have occurred at BATS and Hydrostation S (Michaels and Knap, 1996; Steinberg et al., 2001; Lomas et al., 2013). There is now a wealth of biogeochemical data concerning the northwestern Sargasso Sea.

Bound by the Gulf Stream to the west and northwest, and the North Atlantic Equatorial Current to the south, the Sargasso Sea is also known as the North Atlantic Subtropical Gyre (Michaels and Knap, 1996; Steinberg et al., 2001). The BATS site is a seasonally oligotrophic ecosystem, which is a transitional area between the more eutrophic region to the north and the more oligotrophic subtropical convergence zone to the south. The area is characterized by mesoscale eddies (discussed further below) and a net direction of flow to the southwest. The BATS site experiences significant seasonal changes, and exhibits a seasonal pattern of stratification. Every winter, convective mixing occurs when surface water becomes denser than the water below it (Siegel et al., 1999; Steinberg et al., 2001). This results in deepening of the mixed layer (Siegel et al., 1999), which reaches its maximum in February at a depth of 160 m to 350 m (Michaels et al., 1994; Michaels and Knap, 1996; Lomas et al., 2013). Convection is thus the mechanism by which nutrients enter the mixed layer in winter (Siegel et al., 1999). The deep mixed layer, now nutrient-rich, leads to a phytoplankton bloom, which is manifested by peaks in pigments, primary production, particle flux, and total biomass (Michaels et al., 1994). Pulses of high production, coincident with deep mixing, occur over a period of approximately three months. A transition to a shallow mixed layer then follows, persisting through summer and fall (Michaels et al., 1994). Eddies—rotating packets of water—are another important physical feature of the Sargasso Sea (Doney, 1996; Sweeney and McGillicuddy, 2003). Through a process called eddy pumping, eddies move nutrient-rich water upward toward the photic zone, where the nutrients can be used for primary production (Siegel et al., 1999). This is the primary mechanism for supplying nutrients to the photic zone in summer. As the fall progresses and temperatures cool, the mixed layer deepens, beginning the cycle again (Sweeney and McGillicuddy, 2003).

Nutrient upwelling such as that caused by eddies at BATS leads to biological changes in the water column (Sweeney and McGillicuddy, 2003). These changes can include greater productivity (Sweeney and McGillicuddy, 2003) and increased particle flux (Buesseler et al., 2008). Changes in productivity have been quantified; one study noted that bacterial production increased by a factor of three as a cyclonic eddy passed through BATS (Ewart et al., 2008). New production—primary production associated with nitrogen input—is principally sustained at BATS by nitrogen introduced through winter mixing and mesoscale eddies (Lipschultz et al., 2002). Bacterial respiration levels also vary with eddy activity at BATS; the regions between cyclonic and anticyclonic eddies are associated with decreased respiration (Mourino-Carballido, 2009).

The microbial community at BATS has been extensively studied. The dominant heterotrophic bacteria at BATS are alphaproteobacteria belonging to the SAR11 group (Morris et al., 2002). SAR11 cells account for 31-41% of total cell counts in the photic zone in the Sargasso Sea, and in one sample drawn from 40 m depth at BATS, constituted as much as 51% of the total bacteria (Morris et al., 2002). The ubiquity of SAR11 in the photic zone is consistent with their status as photoheterotrophs; they grow by assimilating dissolved organic carbon, and their metabolic energy is derived from a light-driven proton pump encoded by proteorhodopsin genes (Giovannoni et al., 2005). A three-year study of the abundance of SAR11 in the upper 300 m at BATS revealed that its population density varies with its location in the water column (abundance is greater in the euphotic zone than in the upper mesopelagic zone) (Carlson et al., 2009). Recently, pyrosequencing has enabled greater resolution of the SAR11 clade, and now nine distinct ecotypes have been identified at BATS; partitioning the water column enables them to diversify (Vergin et al., 2013). The dominant photosynthetic organisms in the oligotrophic ocean are members of the genus *Prochlorococcus* (Zinser et al., 2006). *Synechococcus*, another key cyanobacterial genus, is larger in size than *Prochlorococcus*, and less abundant by a factor of ten in oligotrophic waters (DuRand et al., 2001b; Treusch et al., 2009).

Not surprisingly, the microbial community at BATS undergoes seasonal changes, and the structure of the community is associated with position in the water column. Analysis of changes in the microbial community over more than a decade using terminal restriction fragment length polymorphism has revealed that the most important factors in determining composition of the community are deep winter mixing at BATS, warming of the surface layer, and summer stratification (Treusch et al., 2009). Microbial species richness is greater in the mesopelagic zone, possibly because the euphotic zone is dominated by cyanobacteria and proteobacteria such as SAR11. Over an annual cycle, four distinct microbial communities develop in the upper 300 m at BATS. The spring phytoplankton bloom produces one such community. When the bloom ebbs, three separate communities develop at different levels of the water column: one in the upper euphotic zone, one at the deep chlorophyll maximum, and one in the upper pelagic zone. Different microbial clades dominate in each community, but SAR11 appears in all of them (Treusch et al., 2009). SAR11 abundance changes as the mixing of the water column changes; abundance decreases when the mixed layer deepens in winter, and increases in the spring as the mixed layer begins to shoal again (Carlson et al., 2009).

Not all clades of bacteria respond similarly to the deep winter mixing at BATS. T-RFLP analysis of rRNA genes at the surface and 200 m depth was used to identify the clades of bacteria that exhibited the greatest increases in abundance. SAR11, SAR116 (another clade of alphaproteobacteria), and SAR86 (gammaproteobacteria) showed the greatest increases at the surface following deep winter mixing, while at 200 m, the bacterial groups whose abundance increased the most were SAR11, OCS116 (alphaproteobacteria), and a group of marine Actinobacteria (Morris et al., 2005). Seasonal changes are also observed in the phytoplankton community at BATS. The abundance of *Prochlorococcus* is greatest when the water column is stratified and nutrient-poor (Zinser et al., 2006). The maximum *Prochlorococcus* concentration occurs in the summer and fall, and appears at approximately 60-80 m depth (DuRand et al., 2001b). The concentration of *Prochlorococcus* remains high down to almost 200 m. In contrast, *Synechococcus* concentration reaches its maximum in spring. When *Synechococcus* concentration is at its maximum, *Prochlorococcus* is at its lowest abundance, and vice versa (DuRand et al., 2001a).

Viruses at BATS

Because of their critical role as predators of bacteria, viruses merit significant study at the BATS site. A decade-long study of viral abundance revealed that a subsurface peak in viral abundance recurs every summer at BATS between 60-100 m during maximum stratification of the water column (Parsons et al., 2012). Total bacterial counts of over 1200 samples drawn at the same time as viral abundance samples show that despite the varying viral counts, the virus-to-bacteria ratio at BATS stays within a narrow range, from 3 to 20, in 96% of the measurements. Counts of specific bacterial species indicate that neither *Synechococcus* abundance nor SAR11 abundance correlates with viral abundance; however, *Prochlorococcus* abundance coincides with viral abundance in time and depth, suggesting that a large portion of the viruses at this site are cyanophage (Parsons et al., 2012).

The seasonally recurring subsurface peak in viral abundance prompts several questions which this research aims to answer. Does the viral community in the vicinity of that subsurface peak resemble the viral community of surrounding depths, or is there a distinct composition in the peak? Also, seasonal differences may arise between viral communities at BATS as the depth of the mixed layer changes throughout the year. The surface viral community in particular warrants examination because it is part of the shallow mixed layer in summer. If there are differences between the surface viral community and peak-abundance viral community during summer stratification of the water column, do those differences persist in winter, when the water column is well-mixed? Or do differences diminish in winter, when both the surface viral community and the viruses at the depth of the summer abundance peak are contained within the mixed layer?

Using a variety of approaches and techniques, this research aims to expand our knowledge of viral dynamics by analyzing the viral community of the Sargasso Sea over a several-year period, through different seasons, and at different depth scales. The first chapter addresses whether a newly-developed signature gene (*phoH*) can distinguish among marine viral communities at different depths at BATS, as well as among different locations worldwide. The second chapter presents an analysis of the viral diversity of a depth profile at BATS over several years by amplifying and deep sequencing the *phoH* gene. Finally, in the third chapter, several signature genes and a fingerprinting technique are used to assess changes in the viral community at BATS between seasons and between depths. Methods for each experiment, as well as an introduction to the topic in greater depth, are contained within each chapter.

CHAPTER 1

Development of *phoH* as a novel signature gene for assessing marine phage diversity

Note to Reader: This paper has been previously published. The citation is Goldsmith, D.B., Crosti, G., Dwivedi, B., McDaniel, L.D., Varsani, A., Suttle, C.A. et al. (2011) Development of phoH as a novel signature gene for assessing marine phage diversity. *Applied and environmental microbiology* **77**: 7730-7739. The full text of the paper appears in Appendix B.

CHAPTER 2

Deep sequencing of the viral *phoH* gene reveals seasonal variations, depth-specific composition, and persistent dominance of the same phage *phoH* genes in the Sargasso Sea

Summary

Deep sequencing of the viral *phoH* gene, a host-derived auxiliary metabolic gene, was used to track viral diversity at the Bermuda Atlantic Time-series Study site throughout the water column in two seasons from three years. *PhoH* sequences reveal depth-related, seasonal, and interannual differences in the viral communities. The viral *phoH* gene in the Sargasso Sea is quite diverse, with over 3600 operational taxonomic units (OTUs; 97% sequence identity) identified. Despite high richness, most *phoH* sequences belong to a few large, common OTUs while the majority of the OTUs are small and rare. Viral diversity exhibits clear seasonal patterns, with winter *phoH* viral communities more similar to each other than to the summer *phoH* viral communities, and vice versa. While many OTUs make fleeting appearances at just a few times or depths, a small number of OTUs dominate the community throughout the seasons, depths, and years, in seeming contradiction to kill-the-winner dynamics and the Bank model. This apparent inconsistency is reconciled by the newly proposed "Royal Family" model, in which two microbial compartments—abundant and rare—coexist in the ocean. Fluctuations on the level of interacting viral-host pairs occur rapidly within each compartment, but exchange between the two categories rarely occurs.

Introduction

Viruses are the most abundant organisms on the planet (Breitbart, 2012), an order of magnitude more abundant than bacteria (Fuhrman, 1999). Most ocean viruses prey upon bacteria (Fuhrman, 1999), and as a result, play a critical role in all ecosystems. When these viruses (bacteriophage, or phage) lyse bacterial cells, carbon is converted to its dissolved form, slowing the export of carbon to the deep ocean (Suttle, 2005). Marine viruses thus ultimately influence biogeochemical cycling and can affect the rate of atmospheric warming (Wilhelm and Suttle, 1999; Danovaro et al., 2011). Besides being abundant and fundamental contributors to the Earth's biogeochemical cycle, marine viruses are also extremely diverse, and in fact constitute the largest reservoir of genetic diversity on the planet (Rohwer, 2003; Cesar Ignacio-Espinoza et al., 2013). Moreover, viruses can change the genetic makeup of bacteria through horizontal gene transfer (Lindell et al., 2004; Monier et al., 2009; Hurwitz and Sullivan, 2013). For all of these reasons, understanding the diversity of marine viruses has been a research focus for more than 20 years, since Bergh et al. (1989) observed that the abundance of marine viruses was far greater than previously thought.

Studying viral diversity is challenging because the lack of a single gene common to all viruses precludes PCR-based surveys of total viral diversity (cf. 16S rDNA for bacteria) (Rohwer and Edwards, 2002). However, a variety of signature genes exist that can be used to capture subsets of viral diversity, such as the DNA polymerase for podophage (Huang et al., 2010), capsid genes for myophage (Jameson et al., 2011; Chow and Fuhrman, 2012), *psbA* for viruses of photosynthetic bacteria (Chenard and Suttle, 2008), and more recently introduced, *phoH*, which can capture viruses in multiple families that infect both heterotrophic and autotrophic hosts (Goldsmith et al., 2011). *PhoH* has been successfully used to study the diversity of marine viruses from a variety of geographic locations (Goldsmith et al., 2011).

The diversity of marine viral communities has been examined through numerous snapshots—analyses at a single time and place, or a depth profile studied at a single time. However, analysis of a surface viral community is unlikely to be representative of the viruses throughout the water column and viral communities sampled in one season are likely to differ in composition from viruses at the same site but in a different season. Moreover, multiyear experiments are needed to determine whether seasonal patterns repeat over time. Our understanding of marine viral ecology and viral interactions with their hosts can be improved significantly by exploring viral diversity on a variety of temporal and spatial scales. Insight into marine viral dynamics recently expanded when ten years of monthly sampling at the Bermuda Atlantic Time-series Study (BATS) site in the northwestern Sargasso Sea revealed annually recurring seasonal patterns of viral abundance (Parsons et al., 2012). Examination of the upper 300 m of the water column showed that viral abundance peaked every summer between 60-100 m depth concurrent with stratification of the water column. This subsurface peak in viral abundance was highly correlated with a localized increase in the concentrations of *Prochlorococcus*, the dominant photosynthetic organism at this site. Convective overturn each winter deepened the mixed layer and abolished the subsurface peak in viral abundance, leading to fairly stable viral concentrations in the upper water column (Parsons et al., 2012).

Knowledge of the dynamics of viral abundance at the BATS site makes it the ideal location to conduct a thorough analysis of dynamics in viral diversity. In this study, we performed deep 454 pyrosequencing of the *phoH* gene from the viral community in two different seasons (September = summer; March = winter) in multiple years over a depth profile from the surface to 1000 m. To our knowledge this is the first examination of viral diversity using deep sequencing of a signature gene. Both seasonal and depth-related patterns of *phoH* diversity emerge. In addition, this study reveals that while the viral *phoH* community at BATS is extremely rich, only a few operational taxonomic units (OTUs) dominate many depths and times. The remainder of the viral community comprises OTUs that appear infrequently and have few members. While these results seem to contradict the kill-the-winner theory and the Bank model, which predict a cycling of dominant taxa, we propose a new model to resolve the apparent inconsistency.

Results

Deep 454 pyrosequencing of the *phoH* gene from 85 depth/time samples from the BATS site yielded a total of 313,312 sequences containing the forward primer. The number of sequences per sample ranged from 288 to 12,791, with a median of 3,028 sequences recovered per sample. Based on operational taxonomic units (OTUs) defined by sequence identity greater than or equal to 97%, the total dataset consisted of 3,619 OTUs. Although the shape of the rarefaction curves differs for each of the 85 samples (Fig. 2.1a), the rarefaction curves for all the samples have approached an asymptote (Fig. 2.1b), indicating that this level of sequencing sufficiently captured the diversity of the viral *phoH* gene at BATS.

Calculation of two diversity metrics (Chao1 and the inverse Simpson's index) did not reveal clear trends in viral *phoH* diversity over depth or time. The Chao1 richness estimator predicts the minimum richness of a community (Chao, 1984) and values for this dataset ranged from 89 to 1164 *phoH* OTUs per sample. In September 2008, the richest *phoH* communities were at the 300 m, 140 m, and 180 m depths, in that order (Fig. 2.2). However, in September of 2010 and 2011, the richest *phoH* communities were in the top 100 m of the water column. In

September 2010, the surface community had the highest richness, followed by 80 m, 20 m, and 60 m. In September 2011, the *phoH* communities at 60 m and 100 m were the richest according to Chao1. In March 2010, the richest community was at 20 m depth, followed by 160 m and 250 m. One year later, in March 2011, the 180 m community had the highest richness, followed by the 700 m and surface communities.

Another diversity metric, the inverse Simpson's index, incorporates not only richness but also a measure of evenness (Simpson, 1949); it is influenced by the abundance of the most common species (Magurran, 2004). The inverse Simpson's index thus potentially provides greater insight and is more robust than diversity measures based solely on richness (Magurran, 2004). The inverse Simpson's index ranges from a minimum of 1 (where only one OTU is present) to a maximum of the total number of OTUs (3619 in this study) (Ricklefs and Lovette, 1999). According to the inverse Simpson's index, the surface sample from September 2008 was the most diverse, with a diversity measure of 19.8, while the 700 m sample from September 2011 was the least diverse, with an inverse Simpson's index of 2.5 (Fig. 2.3). The median value of the inverse Simpson's index for all 85 samples was 6.3.

A hierarchical cluster analysis performed after constructing a Bray-Curtis dissimilarity matrix revealed that similar depths and seasons cluster together (Fig. 2.4). For example, 14 of the 15 samples from September at depths shallower than 100 m fall into just two clusters, with no other samples contained in those clusters. In addition, 12 of the 16 samples collected in September 2010 and September 2011, from depths between 120 m and 500 m, form a well-supported cluster, with only two other samples contained in the cluster (March 2010 300 m and September 2008 180 m). Winter samples appear to cluster not only by season and depth, but also by year. The nine March 2010 samples from 0 m to 160 m are all in the same well-supported cluster, joined by only one other sample (September 2010, 40 m). Twelve samples from March 2011 form a well-supported cluster, including all depths from 40 m to 400 m. Regardless of season or year, deep water samples cluster together. Seven of the nine samples drawn from depths greater than or equal to 800 m form a well-supported cluster, which further divides into two subclusters according to season.

Over time, the *phoH* communities are more different between depths than they are within depths, according to a permutational MANOVA (F = 3.095, p = 0.0001). Pairwise comparisons reveal that many of the largest differences are between 1000 m and other depths, especially depths shallower than 500 m (F values range from 4.7 to 14.3; p-values range from 0.015 to 0.03) (Table 2.1). Among the other depths investigated, 0 m is significantly different from every depth below 80 m (F values range from 3.13 to 5.79; p-values range from 0.008 to 0.04). The largest pairwise difference in *phoH* communities is between the 400 m community and the 1000 m community (F = 14.3; p = 0.019). The depths with the fewest significant differences with other depths are 180 m (significantly different only from the 0 m community, F = 3.62, p = 0.037) and the 500 m community (significantly different only from the 0 m community (F = 4.06, p = 0.04) and the 40 m community (F = 3.77, p = 0.04)). Combining all depths and years, the season of sampling also influences the *phoH* viral community structure. The differences between the March and September *phoH* communities are greater than the differences within communities in the same month (F = 2.781, p = 0.011).

Of the 3,619 OTUs recovered in this study, the vast majority of the OTUs were rare (~96% of these OTUs contain <0.01% of the total number of sequences). Only 18 OTUs contain at least 1% of the total number of sequences (Fig. 2.5a). Fifty-one OTUs have at least 0.1% of the sequences, and 150 OTUs contain at least 0.01% of the sequences. Distribution of the sequences among the OTUs is highly skewed, in that together, the two largest OTUs (OTUs 1 and 2) contain more than one third of the sequences. The five largest OTUs (OTUs 1 through 5) contain 52.4% of the sequences, and more than 82% of the sequences are contained in the top 18 OTUs.

Analysis of the five largest OTUs provides significant insight into compositional changes of the *phoH* community at BATS with season, depth, and year. Although the five largest OTUs together contain more than half of the total number of sequences, the degree to which those OTUs contribute to the community of each individual sample varies considerably (Fig. 2.5b). Sequences from these five OTUs comprise up to 77.1% of a sample (March 2011, 160 m) or as little as 0.2% of a sample (September 2011, 1000 m). Although OTU 1 contains the largest proportion of sequences overall, this OTU is virtually absent from each of the three September surface communities. OTU 1 starts to appear in September below the surface, but sequences from OTU 1 do not reach 20% of the community until 100 m (2010), 120 m (2008), or 140 m depth (2011). In March, however, OTU 1 is a more consistent component of the *phoH* community throughout the depth profile; OTU 1 comprises 14% to 32% of the March 2010 community at all sampled depths, and 21% to 30% of the March 2011 community from the surface to 700 m.

Similarly, OTU 2 is a consistent presence in March 2011, from the surface to 500 m, and in March 2010, from the surface to 250 m (except for 60 m). However, OTU 2 constitutes less than 2% of each of the three September surface communities. OTU 2 becomes a larger portion of the September *phoH* communities starting with 20 m in September 2008 and 40 m in September 2010. In September 2011, OTU 2 has a sporadic and varied presence among the sampled depths. No sequences belong to OTU 2 from the communities sampled at 140 m, 600 m, 900 m, and 1000 m. At the other depths in September 2011, the contribution of OTU 2 sequences ranges from 0.01% at 100 m to 24% at 400 m.

OTU 3 has a strong presence in the upper 80 m during September 2010 and September 2011, as well as the upper 160 m of March 2010. OTU 3 appears in smaller percentages during September 2008 and March 2011. Sequences from OTU 3 are not found below 250 m, with a few exceptions where they constitute less than 1% of the *phoH* communities (300 m in March and September 2011, 400 m in September 2010 and March 2011, 500 m in September 2011). OTUs 4 and 5 constitute a smaller percentage of the *phoH* community at BATS; however, OTU 4 makes an especially large contribution to the 400 m community in March 2010 (27.5%) and the 700 m community in September 2011 (61.6%). The 61.6% contribution of OTU4 to the September 2011 700 m community is the single largest contribution by any OTU to any sampled date and depth.

For the ease of data visualization, further analyses consider 94 OTUs: the 51 OTUs that contain at least 0.1% of the total number of sequences, and an additional 43 OTUs that contain at least 1% of the sequences from any individual sample. Figure 2.6 demonstrates the percent of sequences in the top 94 OTUs from each of the samples from 2010 and 2011. Few OTUs constitute a substantial portion of any individual sample. Only one OTU (OTU 4, discussed above) constitutes more than 50% of the sequences recovered from a single sample. Six OTUs constitute more than 40% of an individual sample. As the threshold decreases, more OTUs are included: 8 OTUs constitute at least 30% of a sample; 13 OTUs constitute at least 20%; and 24 constitute at least 10%. However, even at 5%, only 42 OTUs out of 3,619 meet the threshold. Thus only 1.1% of the OTUs constitute at least 5% of any sample, and the vast majority of the OTUs are rare. Figure 2.6 also demonstrates the seasonal nature of some OTUs. Some OTUs appear only in *phoH* communities sampled in March, while other OTUs appear only in September samples.

Figure 2.7 displays a phylogenetic tree of the *phoH* gene containing representatives from each of the top 94 OTUs, as well as the *phoH* gene from several fully-sequenced “reference” viral genomes. The heat map next to the tree displays the percentage that each OTU (rows) comprises in each sample (columns). The groups identified in the phylogenetic tree are the same groups identified in a previous study of marine viral *phoH* diversity (Goldsmith et al., 2011). Despite the greatly increased sequencing depth in the present study, no new phylogenetic groups were identified among those top 94 OTUs as comprising more than 0.1% of the total sequences or more than 1% of the sequences from any individual sample. The five largest OTUs (Fig. 2.5) belong to three phylogenetic groups: OTUs 1 and 4 are in Group 1; OTUs 2 and 5 are in Group 3; and OTU 3 is in Group 2.

Based on the phylogenetic groups to which each of the top 94 OTUs belongs (Fig. 2.7), Figure 2.8 displays the phylogenetic group composition of each sample. Group 1 (containing OTUs 1 and 4) is a dominating presence throughout the dataset, constituting at least 40% of 68 of the 85 samples, and at least 30% of 78 of the samples. Group 3 (containing OTUs 2 and 5) has a strong presence at all depths in March 2011, but is more varied in its abundance throughout the depth profile in March 2010. In September 2008, Group 3 comprises a smaller portion of the three September surface communities than it does of the March surface communities. In September 2011 in particular, Group 3 forms less than 15% of every sample from the surface through 140 m, with the exception of the 60 m community (20.7%).

Overall, Groups 4 and 5 comprise a smaller part of the sampled *phoH* communities, and play a more important role at depth than in the upper water column. In only 3 out of 85 samples did Group 4 comprise at least 1% of the *phoH* community. All three of those samples were from September 2011, at depths of 700 m, 800 m, and 900 m. Group 5 is more prevalent than Group 4, but even so, only five samples contain Group 5 as at least 15% of the community. The maximum contribution Group 5 makes to a sample is in September 2008, 900 m, where it constitutes 31% of the *phoH* community.

Discussion

Numerous studies have examined changes in viral community composition over short time scales, with contradicting results regarding stability of the viral community over time. A study of viral diversity in coastal Denmark over three months found that despite large changes in viral abundance, the viral community structure as shown by pulsed field gel electrophoresis (PFGE) was relatively stable (Riemann and Middelboe, 2002). Similarly, PFGE fingerprints of the viral community in Southern California waters showed temporal stability between August and October 2000 (Fuhrman et al., 2002). In contrast, another study revealed significant changes in viral community composition over a period of 78 days. Terminal restriction fragment length polymorphism (T-RFLP) analysis of the g23 gene amplified from viral communities collected near an island off the coast of Southern California revealed 153 myophage OTUs over the course of a 78-day time series (Needham et al., 2013). One OTU comprising approximately 25% of the total myophage community on the first day of the study made up less than 5% of the community at the end of the experiment. However, the restricted length of these viral diversity studies may not capture seasonal distinctions. Moreover, the low-resolution techniques used in these studies may underestimate viral diversity (for example, phage genomes of the same size may co-migrate on a gel, but represent different phage types).

The present study demonstrates clear seasonal patterns in viral diversity; a permutational MANOVA and a dendrogram based on a Bray-Curtis dissimilarity matrix shows that the winter *phoH* viral communities are more similar to each other than they are to the summer *phoH* viral communities, and vice versa (Fig. 2.4). These data are consistent with numerous previous studies that have demonstrated seasonal variation in marine viral communities. Notably, seasonal differences have been observed using vastly different methods (including signature gene amplification and sequencing, PFGE, randomly amplified polymorphic DNA (RAPD) PCR, denaturing gradient gel electrophoresis (DGGE)). In the Eastern Mediterranean Sea, viral community composition (according to fluorescence intensity detected by flow cytometry) varied depending on whether the water column was stratified or well-mixed (Magiopoulos and Pitta, 2012), and monthly changes in viral community composition over the course of one year at this site have also been documented with that method (Winter et al., 2009). Viral community structure in the Chesapeake Bay also undergoes seasonal changes, as evidenced by RAPD PCR (Winget and Wommack, 2008) and PFGE (Wommack et al., 1999; Jamindar et al., 2012). PFGE has also revealed changes in viral genome banding patterns in coastal California waters over a six-month period (Steward et al., 2000) and in waters off the coast of Norway over an eight-month period (Sandaa and Larsen, 2006), and seasonal variation of phage diversity has been observed in the Gulf of Aqaba through DGGE (Mühling et al., 2005). Another coastal California study, based on more than a year of monthly analyses of two signature genes in cyanomyophage (g20 and psbA), showed clear distinctions between the summer viral community and the winter/fall viral community (Clasen et al., 2013). In coastal Rhode Island waters, analysis of myophage isolates through the g20 signature gene revealed seasonal variations of abundance and diversity (Marston and Sallee, 2003).

Most of these studies have been limited to seasonal analyses within a single year, so the repeatability of these patterns cannot be addressed. Multiyear time-series studies, such as the data presented here, are especially valuable for addressing this issue. The cyanopodophage community of the Chesapeake Bay, analyzed via the DNA polymerase gene during winter and summer for two years, exhibited repeating seasonal differences, and winter phage communities sampled in different years grouped more closely with each other than with summer phage communities from the same year (Chen et al., 2009). For cyanophage isolated from Narragansett Bay, similarity analysis based on the g43 DNA polymerase gene showed clustering according to season: composition of the cyanomyophage community was more similar to the composition of samples from the same season of any year than a different season of the same year (Marston et al., 2013). Viral communities at the site of the San Pedro Ocean Time-series (SPOT) also displayed seasonally recurring patterns of diversity (Chow and Fuhrman, 2012). Communities 3-7 months apart were negatively correlated, while communities from adjacent months were highly correlated, as were communities from the same month one year apart (Chow and Fuhrman, 2012). Using the same type of analysis (TRFLP analysis of the g23 gene), Pagarete et al. (2013) studied changes in the myophage community sampled monthly for two years from water in Raunefjorden, Norway and observed three distinct viral communities depending on the season: summer, fall, and winter/spring.

The present study found that only a few OTUs contained most of the sequences (the 18 largest OTUs combined contain more than 82% of the sequences), while most of the OTUs contained a small fraction of the sequences (3,469 OTUs each contained less than 0.01% of the sequences). Thus most of the sequences are in a few (large) OTUs that were found in a high proportion of sampling dates/depths (i.e. common OTUs). The remaining OTUs (the bulk of the

OTUs) were small and rare. These results are in concordance with the findings of Needham et al. in their 78-day time series conducted in coastal California waters, during which they collected 45 samples. Examination of the relative abundance of each OTU showed that while most of the OTUs appeared in less than 25% of the samples, more than 80% of the viral community consisted of OTUs that appeared in at least 90% of the samples (Needham et al., 2013). A culture-based study by Marston et al. obtained similar results for cyanophage isolates from several locations. One set of isolates, collected at 41 time periods from Narragansett Bay, contained 108 OTUs. However, the 12 most abundant OTUs represented 63.5% of the isolates. Another set of 2,406 isolates, collected at numerous locations throughout southern New England, constituted 162 OTUs, but the five most abundant OTUs represented 58% of the isolates (Marston et al., 2013). Pagarete et al. had similar findings in their recent Raunefjorden study, in which they identified 160 OTUs in the 28 samples they collected monthly over two years. The most commonly observed OTUs had higher average and maximum contributions to the viral community (based on the g23 gene), while the OTUs that appeared less frequently in the samples tended to represent fewer sequences from the viral community (Pagarete et al., 2013). Chow and Fuhrman produced concordant findings using T-RFLP analysis of the g23 gene to study three years of monthly samples. The most common OTUs (those that appeared in at least 30 out of 34 months) made up a higher proportion of the viral community than did the least common OTUs, with a positive relationship between number of times an OTU appeared and its contribution to the viral community at SPOT (Chow and Fuhrman, 2012).

Both the present study and Parsons et al. (2012) underscore the importance of investigating both time and depth in order to understand the dynamics of a marine viral community. In this study, the OTU composition of the upper 250 to 500 m is fairly consistent in

March when the water column at BATS is well-mixed (Fig. 2.5b), while the September samples, drawn during summer stratification at BATS, reflect a much more variable composition of the phoH community in the upper 200 m.

Part of the September variability in OTU composition is due to the presence of OTUs belonging to phylogenetic Group 2 (Fig. 2.8). Group 2 is particularly interesting because its presence is strong but limited to the upper water column. Almost all of the *phoH* sequences from fully-sequenced cyanophage (phage that infect cyanobacteria) genomes fall into Group 2 (Fig. 2.7). The fact that Group 2 *phoH* genes are concentrated in the photic zone and absent from deeper depths supports the idea that this group is dominated by cyanophage *phoH* genes. The 20 m sample from September 2010 is especially noteworthy, because more than 93% of the community is part of the phylogenetic Group 2 (Fig. 2.8). Moreover, based on the top 94 OTUs, none of that community comes from Groups 3, 4, or 5. The communities surrounding the 20 m community in September 2010 are also heavily comprised of Group 2, as are upper water column communities in September 2011 (especially the 0 m and 40 m communities). In winter, though, when the water column is well-mixed, Group 2 exhibits interannual variation. In March 2010, Group 2 forms between 11% and 36% of each sample from 0 m to 160 m. However, in March 2011, Group 2 is virtually absent from the 20-depth profile: it constitutes less than 5% of the surface sample, less than 3% of four other depths, and less than 1% of all the remaining depths.

Assuming that Group 2 does in fact represent cyanophage, these data are consistent with a study by Wilson et al. (1999) that used DGGE to examine the structure of the cyanophage community in two depth profiles along an Atlantic Ocean transect from the Falkland Islands to the United Kingdom. At one station, where the water column was well mixed, samples drawn from six depths in the upper 100 m revealed that the cyanophage population structure was similar throughout the water column. In the other depth profile, the water column was stratified, and the structure of the cyanophage population was variable throughout the profile.

Other studies of viral community composition changes with depth show mixed results, possibly due to the different methods used, different geographical region, or the unknown degree of stratification in the water column at the time of sampling. For example, a depth profile in Danish coastal waters revealed through PFGE that the structure of the viral community showed no significant changes with depth (Riemann and Middelboe, 2002). In contrast, DGGE analysis of g20 amplicons showed that different cyanomyophage communities resided at different depths in the Straits of Georgia (British Columbia) (Frederickson et al., 2003). Differences of only a few meters in depth resulted in shifts in community composition, and while some cyanomyophage appeared throughout the water column, others were found at only a few depths.

A RAPD PCR study of viral diversity in the Chesapeake Bay included samples drawn from both the top and the bottom of the water column (1 m below the surface and 2 m above the sediment-water interface) (Winget and Wommack, 2008). Comparison of the amplicon banding patterns of the viral communities at the surface and the bottom revealed highly similar viral communities. However, Zhong et al. amplified and analyzed g20 sequences and discovered that in early summer, phage population structure was different in Sargasso Sea surface waters than at the depth of the deep chlorophyll maximum (DCM). This was also true in the Gulf of Mexico, where myophage at the surface were distinguishable from myophage at the DCM (Zhong et al., 2002).

Persistence of some OTUs and transience of other OTUs are recurring themes in studies of viral diversity, and the present study is no exception. At SPOT, certain OTUs showed repeating seasonal patterns, but the patterns varied among OTUs: some OTUs persisted throughout the year at moderate levels, while others had peak abundances in a particular season (Chow and Fuhrman, 2012). In a hypersaline lake in Australia, over nearly three years, much of the viral community was dynamic, while at least one assembled viral genome and two other viral genome fragments appeared in 91 to 100% of the samples (Emerson et al., 2012). In Lake Ontario, qPCR was used to track the abundance of three algal virus genes for 13 months (Short and Short, 2009). Two of the genes appeared in nearly every sample, with seasonal variations in abundance, while the other gene appeared in only a few samples but at higher abundance than the other two genes. This study posited that some aquatic viruses persist throughout the year, while others are transient. Rozon and Short expanded upon the results of that study by using qPCR to monitor the abundance of 10 viral genes at three stations in an embayment of Lake Ontario from May to October. The genes (from algal viruses and freshwater cyanophage) exhibited several different patterns of abundance. Some OTUs appeared at all locations and all time points at fairly constant abundances; some taxa appeared at all locations but only sporadically; and other taxa showed patchy distribution (Rozon and Short, 2013). Similarly, in the present study, we find that some OTU persist throughout the seasons, depths, and years, while many other OTUs make fleeting appearances at just one or a few times or depths.

The high dominance of a few viral OTUs, complemented by a large number of rare viruses, is contradictory to community models predicted from metagenomic data which typically predict that even the most abundant viral genotype accounts for <10% of the total community (Angly et al., 2006). This may be because *phoH* genes are present in only a subset of viruses, or because only a single gene is being examined (as opposed to metagenomic assemblies). In addition, persistence of the largest *phoH* OTUs throughout the present study despite changes in seasons, depths, and years initially seems to contradict the expectations of kill-the-winner dynamics and the Bank model. The Bank model predicts that viral taxa will cycle in and out of dominance between an "active" fraction and a "bank" fraction (Breitbart and Rohwer, 2005). A potential result of the kill-the-winner hypothesis is that when a particular host becomes active, a virus that can infect that host will increase in abundance (Winter et al., 2010). Infection of the dominant host by the virus causes the susceptible host population to decline, followed by a decline in the virus population. As the originally dominant host population declines, a new host emerges to fill the now-vacant niche, and this host will be resistant to the virus that infected the first host. Subsequently a new virus emerges that is able to infect the newly-dominant host, and the cycling continues (Thingstad and Lignell, 1997).

The question is, what do the "new" hosts and viruses in this progression look like? One possibility, the "Equal Opportunity Model," is that all hosts (and their respective viruses) have an equal opportunity to step in and replace the empty niche, moving from the bank fraction to the active fraction. Another possibility, which we dub the "Royal Family Model," is that these active bacteria are active because they are optimized to that specific niche, and therefore the "new" host is much more likely to be a variant of a previously active host—perhaps an adapted strain that has acquired resistance to its active viral predator. The "new" virus may have adapted to overcome this resistance. This evolutionary arms race (Comeau and Krisch, 2005; Stern and Sorek, 2011) has been demonstrated in chemostat experiments (Mizoguchi et al., 2003; Middelboe et al., 2009; Marston et al., 2012) and proposed to occur in natural marine systems (Bidle and Vardi, 2011; Marston et al., 2012).

In the Equal Opportunity Model, where microbes move between the bank and active fractions, it would be highly unlikely that the same host and virus sequences would be abundant at two different times, at least on a broadly measured scale. Indeed, this was predicted in modeling experiments by Heinz Hoffman et al. (2007). In the present study, the Equal Opportunity Model would predict that completely different *phoH* sequences should be dominant at different time points, since the odds of two randomly chosen viruses from the bank containing identical *phoH* genes are extraordinarily small. The results therefore contradict this model, since only a few *phoH* OTUs were abundant and common throughout the time and depth series. In addition, the vast majority of the OTUs were rare throughout the samples, as opposed to seeing movement of OTUs from the rare to active fractions.

In contrast, in the Royal Family Model, the hosts and viruses that dominate over time are likely to be closely related variants. For example, in evolving to overcome the resistance developed by the previously-dominant host, it is unlikely that the *phoH* gene will be distinct enough from the *phoH* gene of the previous virus to fall into a different OTU. Although the function of *phoH* in viruses is unknown, in *E. coli* the gene's protein product is an ATPase (Koonin and Rudd, 1996; Makino et al., 1998; Hsieh and Wanner, 2010; Sullivan et al., 2010b). Thus the part of the viral genome that needs to evolve in order to overcome host resistance is unlikely to occur within the *phoH* part of the genome. The present data support this model, where, with respect to the *phoH* gene, the newly-dominant virus may closely resemble the formerly-dominant virus, leading to a relatively steady population of the same *phoH* OTU. It is more likely that changes will evolve in viral genes whose products are implicated in host specificity or attachment, such as genes encoding a tail fiber protein. Indeed, these genes are found to be highly diverse in sequenced viral genomes (Letarov et al., 2005; Comeau et al., 2007; Sullivan et al., 2010a), and it would be challenging to design degenerate PCR primers capable of capturing their diversity for use as a signature gene (Dwivedi et al., 2012).

The Royal Family Model presents a mechanism for resolving the seeming contradiction to the kill-the-winner predictions and Bank model presented by the *phoH* deep sequencing data. Moreover, this model is supported by the findings of Rodriguez-Brito et al. (2010), who studied virus and host dynamics in four aquatic environments. Their data demonstrated that persistence of broad viral and host taxa occurred simultaneously with kill-the-winner-type fluctuations at the level of host strains and viral genotypes. Complementing these results with the *in situ* viral *phoH* diversity data generated in the present study supports the idea that the oceans contain two general microbial categories: an abundant component containing a small number of dominant, niche-optimized bacterial types and their viruses, and a rare component containing a large number of rare bacterial and viral types. While rapid fluctuations on the level of virus-host interacting pairs (bacterial strains and viral genotypes) are predicted within each compartment as a result of an evolutionary arms race, these data suggest that exchange between the rare and dominant compartments rarely occurs.

Experimental procedures

Sample collection and DNA extraction. Samples were collected from throughout a depth profile on September 2-3, 2008, March 9 and September 5, 2010, and March 28 and September 17, 2011. All samples were collected in the vicinity of the Bermuda Atlantic Time-series Study (BATS) site (31º40′ N, 64º10′ W) from 0, 20, 40, 60, 80, 100, 120, 140, 160, 180 (2008 and 2011 only), 200, 250 (2010 and 2011 only), 300, and 400 m depth. In 2008 and 2011, samples were also collected from 500, 600, 700, 800, 900, and 1000 m. Metadata associated with these sampling dates and depths are available at the BATS website (bats.bios.edu). Whole seawater samples (100 mL) were filtered through a 0.22 µm Sterivex filter (Millipore, Billerica, MA) and then onto a 0.02 µm Anotop filter (Whatman, Piscataway, NJ). Anotop filters were stored at -80ºC until DNA was extracted with a MasterPure complete DNA and RNA purification kit (Epicentre Biotechnologies, Madison, WI) following the protocol of Culley and Steward (2007). Briefly, filters were defrosted, and all liquid was purged from the filter by pushing air through with a sterile syringe. A flame-sealed pipette tip was used to temporarily seal the filter outlet, and a mixture of 400 µL of 2X T&C lysis buffer (from the MasterPure kit) and 50 µg proteinase K was forced onto the filter. The filter was then incubated for 10 min in the air at 65ºC before the lysate was expelled into a microcentrifuge tube and immediately placed on ice. Then 150 µL of MPC protein precipitation reagent (from the MasterPure kit) was added to the lysate and vortexed vigorously for 10 s. The debris was pelleted by centrifugation at 10,000 x*g* for 10 min. Isopropanol was added to the recovered supernatant, and the tube was inverted 30 to 40 times. The DNA was then pelleted by centrifugation at 20,000 x*g* at 4ºC for 10 min and washed twice with 75% ethanol. Extracted DNA was resuspended in sterile water and stored at -20ºC.

Amplification of the *phoH* gene. The extracted DNA was amplified in triplicate reactions using the strand displacement method of the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare, Piscataway, NJ) according to the manufacturer's instructions and then pooled. Next a first-stage PCR was conducted for amplification of the *phoH* gene, using viral *phoH* primers vPhoHf (5′-TGCRGGWACAGGTAARACAT-3′) and vPhoHr (5′-TCRCCRCAGAAAAYMATTTT-3′) (Goldsmith et al., 2011). Four replicates of the PCR reaction were conducted for each sample, and the products were pooled after a reconditioning PCR and cleaning (see below). The 50-µL reaction mixture contained 1 U Apex *Taq* DNA polymerase (Genesee Scientific, San Diego, CA), 1X Apex *Taq* reaction buffer, 1.5mM Apex $MgCl_2$, a 0.5 µM concentration of each primer, 0.2 mM deoxynucleoside triphosphates, 0.08%

bovine serum albumin, and 1 µL of template DNA (pooled Genomiphi product). The reaction conditions were: (i) 5 min of initial denaturation at 95ºC; (ii) 35 cycles of 1 min of denaturation (95ºC), 1 min of annealing (53ºC), and 1 min of extension (72ºC); and (iii) 10 min of final extension at 72ºC.

Next, each PCR product underwent a reconditioning step as recommended by Berry et al. (Berry et al., 2011), in order to minimize variation that can accompany different barcoded primers. The reaction mixture was the same as in the first-stage PCR, except that 10-bp barcodes were attached to the viral *phoH* primers (see Table 2.2). The template DNA consisted of 1 µL of product from the first-stage PCR reaction and the same reaction conditions were used, except that only 10 amplification cycles were run. After the reconditioning PCR, the four replicates for each sample were individually cleaned with a DNA Clean & Concentrator-25 kit (Zymo Research Corp., Irvine, CA) following the manufacturer's instructions and resuspended in 45 µL of sterile water. The four replicates for each sample were pooled for quantification and downstream processing.

DNA quantification and sequencing. The amount of DNA recovered for each sample was quantified using a real-time PCR measurement of fluorescence as suggested by Blotta et al. (2005), with Quant-iT PicoGreen as the detector (Life Technologies, Grand Island, NY). Each sample was run in duplicate, with the real-time PCR machine set to obtain a fluorescence reading during each of three 75-second cycles. The six fluorescence readings were averaged to obtain a mean fluorescence reading for each sample. After quantitation based on a standard curve, equal amounts of each sample (~1600 ng) were placed into one of four pools for sequencing. Sequencing of the *phoH* amplicon was performed on the 454 GS FLX Titanium platform by Beckman Coulter Genomics (Danvers, MA). Before sequencing, Beckman ligated sequencing adaptors to each of the four pools, multiplexing them onto half of a picotiter plate. After sequencing, the four pools were de-multiplexed before the sequences were returned for analysis. The FASTA, .qual, and .sff files for each sample have been submitted to GenBank's Sequence Read Archive as accession SRP039081. The BioProject Accession Number is PRJNA239691, and individual sample accession numbers are SAMN02670781 to SAMN02670865.

Sequence analysis. After the barcodes were removed, the sequences were searched for the forward primer, and the downstream analyses proceeded with those sequences containing the forward primer. The sequences have been deposited in METAVIR (http://metavir-meb.univ-bpclermont.fr) under the project name "Viral phoH at BATS – Goldsmith et al. 2014", virome name "All phoH sequences, forward primer." The sequences were analyzed using mothur (Schloss et al., 2009). After mothur was used to align the sequences, trim them to include only the aligned space, filter out columns of the alignment that do not contain data, pre-cluster the sequences to merge sequences that are with two base pairs of a more abundant sequence, and remove chimeras, the number of sequences remaining was 313,312. Using mothur, the sequences were grouped into operational taxonomic units (OTUs) defined by sequence identity of 97% or greater. Rarefaction curve data, Chao1 richness estimates, and inverse Simpson diversity estimates were also calculated using mothur, and plotted in R (R Development Core Team, 2013). In particular, the heatmap reflecting the inverse Simpson diversity estimates was plotted using the gplots (Warnes et al., 2009) and RColorBrewer (Neuwirth) packages in R. The heatmap reflecting the Chao1 richness estimates was plotting with the fossil package (Vavrek, 2011). Hierarchical clustering was performed from a Bray-Curtis dissimilarity matrix using the picante package (Kembel et al., 2010). In order to bootstrap the dendrogram, Jaccard stability means were computed using the fpc package (Hennig, 2013). The dot plot (Fig. 2.6) was constructed with the lattice package (Sarkar, 2008).

PhoH sequences representative of each of 94 OTUs were selected for the phylogenetic tree: the 51 OTUs that contain at least 0.1% of the total number of sequences, and an additional 43 OTUs that contain at least 1% of the sequences from any individual sample. These 94 representative sequences have been deposited in GenBank's Sequence Read Archive under accession SRP039081. The representative sequences are also in METAVIR (http://metavir-meb.univ-bpclermont.fr) under the public project name "Viral phoH at BATS – Goldsmith et al. 2014", virome name "phoH OTU representatives." Next, the HAXAT program (Lysholm, 2012) was applied to the sequences (against a custom-built database of viral *phoH* sequences) in order to correct homopolymer sequence errors (using default parameters, except that both strands were queried and a minimum score of 200 was used). *PhoH* sequences from several fully-sequenced phage genomes were added, and then an amino acid alignment was built from the sequences using MUSCLE (Edgar, 2004) (with default parameters) as implemented by TranslatorX (Abascal et al., 2010). The alignment was then back-translated into nucleotides, and FastTree (Price et al., 2010) was used to build an approximate maximum likelihood phylogenetic tree, with the Jukes-Cantor model of nucleotide evolution and the CAT approximation of a single rate of evolution across all sites. In R, the tree was prepared for aligning with the heatmap using the ape (Paradis et al., 2004) and phangorn (Schliep, 2011) packages. The heatmap was constructed and aligned with the tree (Fig. 2.7) using the gplots (Warnes et al., 2009), RColorBrewer (Neuwirth, 2011), and colorRamps (Keitt, 2012) packages. Permutational MANOVA analyses were conducted in PAST, version 3.01 (Hammer et al., 2001).

Acknowledgements

Many thanks to Federico Abascal, Damitu Beyene, Craig Carlson, Bhakti Dwivedi, Karoline Faust, Julia Gustavsen, Christian Hennig, David Jones, Kazutaka Katoh, Fredrik Lysholm, Vincent Ranwez, Peter Salamon, Pat Schloss, and Palle Villesen for bioinformatic, coding, and other assistance, and to the crew of the R/V Atlantic Explorer for logistical support. This research was funded by a grant from the National Science Foundation to MB (Microbial Interactions and Processes MCB-0701984). DBG was supported by a Presidential Doctoral Fellowship from the University of South Florida and the Von Rosenstiel Endowed Fellowship.

Tables and figures

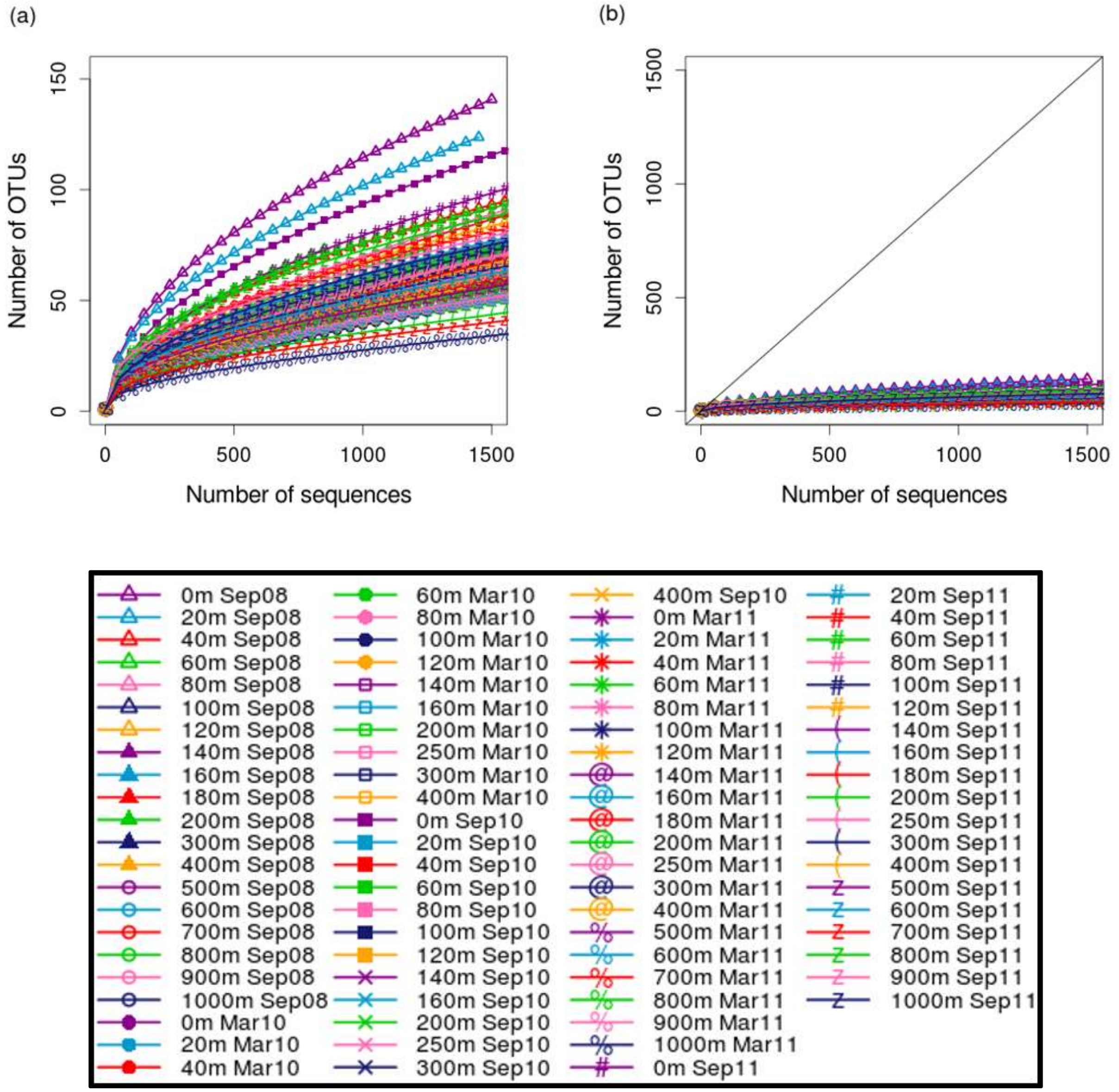


Fig. 2.1: Rarefaction curves for phoH sequences from all 85 depths/times; OTUs are defined by sequence identity of 97% or greater. (a) Plotting with different scales on x- and y-axes demonstrates some separation of the curves. Curve with greatest slope is September 2008, 0 m, while curve with least slope is March 2011, 1000 m. (b) Plotting in relation to the 1:1 line demonstrates flattening of all rarefaction curves.

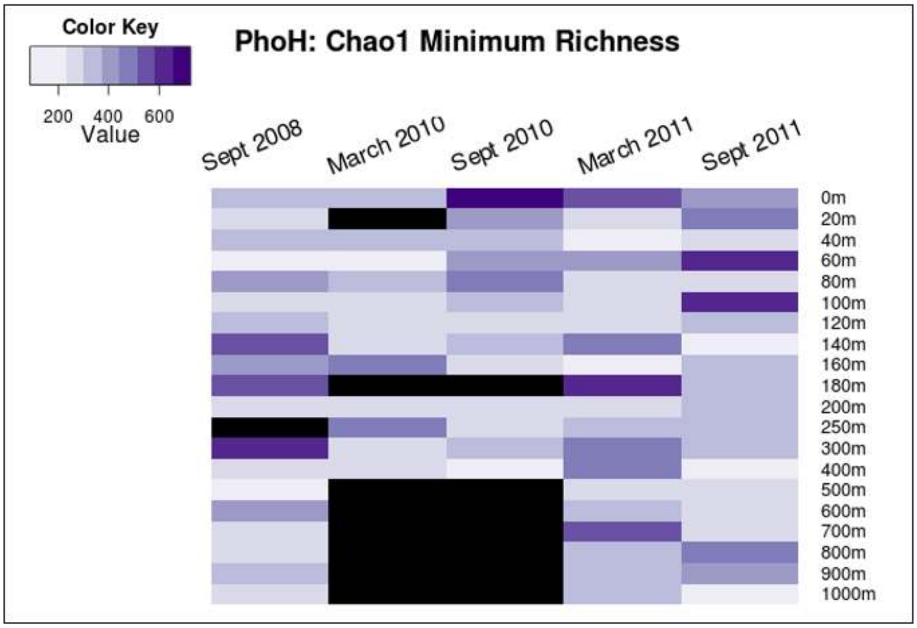


Fig. 2.2: Heatmap displaying Chao1 minimum richness estimate for all dates and depths except March 2010, 20 m. A black bar indicates absence of sample for that date/depth, except that March 2010, 20 m is represented by a black bar because its Chao1 minimum richness estimate (1164) obscured differences in the estimates for the other dates and depths.

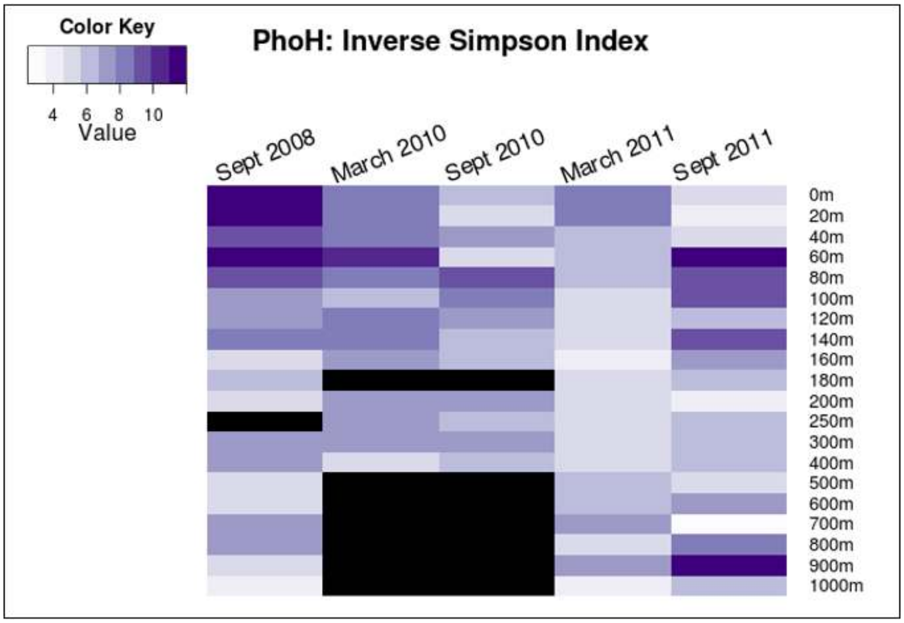


Fig. 2.3: Heatmap displaying inverse Simpson index diversity estimate for all dates and depths. A black bar indicates absence of sample for that date/depth.

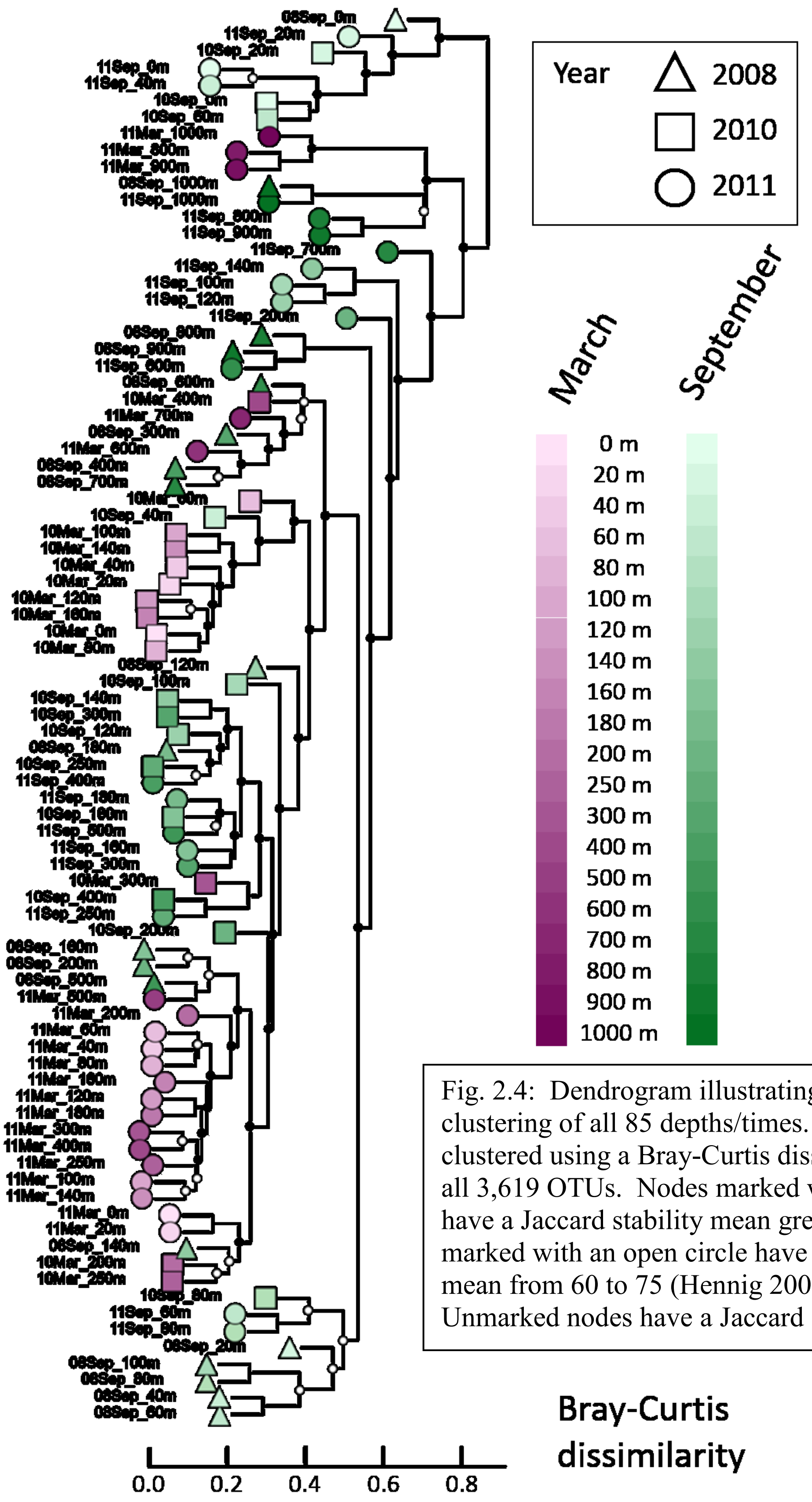


Fig. 2.4: Dendrogram illustrating hierarchical clustering of all 85 depths/times. Samples are clustered using a Bray-Curtis dissimilarity matrix for all 3,619 OTUs. Nodes marked with a filled circle have a Jaccard stability mean greater than 75; nodes marked with an open circle have a Jaccard stability mean from 60 to 75 (Hennig 2007; Hennig 2008). Unmarked nodes have a Jaccard stability below 60.

Table 2.1: F values for pairwise comparisons of depths by permutational MANOVA. Input consisted of normalized abundance data for all depths/times and all 3,619 OTUs. Highlighted cells reflect F values for which the p-value is < 0.05.

Table 2.1: Permutational MANOVA, F values for pairwise comparisons of depths

	0m	20m	40m	60m	80m	100m	120m	140m	160m	180m	200m	250m	300m	400m	500m	600m	700m	800m	900m
20m	0.38																		
40m	1.15	0.54																	
60m	1.27	0.79	0.40																
80m	2.67	1.53	1.02	0.56															
100m	3.51	2.30	1.92	1.17	0.55														
120m	3.54	2.37	2.02	1.68	1.20	0.20													
140m	4.11	2.74	2.71	2.07	1.66	0.56	0.43												
160m	4.76	3.21	3.09	2.55	2.57	1.28	0.98	0.64											
180m	3.62	2.62	2.98	2.49	2.44	1.25	0.99	0.67	0.40										
200m	4.30	3.01	3.07	2.53	2.80	1.32	0.99	0.77	0.42	0.45									
250m	4.63	3.18	3.65	2.92	3.15	1.67	1.40	0.76	0.29	0.57	0.40								
300m	5.50	3.98	4.99	3.68	4.06	2.27	2.15	1.17	1.19	1.00	1.07	0.59							
400m	5.79	4.18	5.43	4.00	4.76	2.64	2.50	1.46	1.48	1.35	1.24	0.63	0.31						
500m	4.06	2.90	3.77	3.11	3.94	2.42	1.87	1.54	1.14	1.37	0.72	0.91	0.85	0.92					
600m	4.03	3.37	5.42	3.95	5.24	3.55	3.35	3.01	3.64	3.22	2.49	3.32	1.90	1.91	1.99				
700m	3.13	2.47	3.74	2.77	3.47	2.34	2.48	1.90	2.69	2.06	2.11	2.22	1.52	1.49	1.68	0.77			
800m	3.14	2.50	3.55	2.82	3.63	3.14	2.93	2.95	3.53	2.81	2.84	3.50	2.61	3.25	1.98	1.59	0.91		
900m	3.43	3.02	4.94	3.88	5.02	4.36	4.30	4.16	5.13	3.64	4.03	4.67	3.59	4.13	2.82	1.45	1.06	0.51	
1000m	5.25	4.71	8.19	7.52	9.35	9.44	9.18	10.04	12.87	9.78	9.92	13.61	11.83	14.30	10.00	7.80	4.90	2.93	2.70

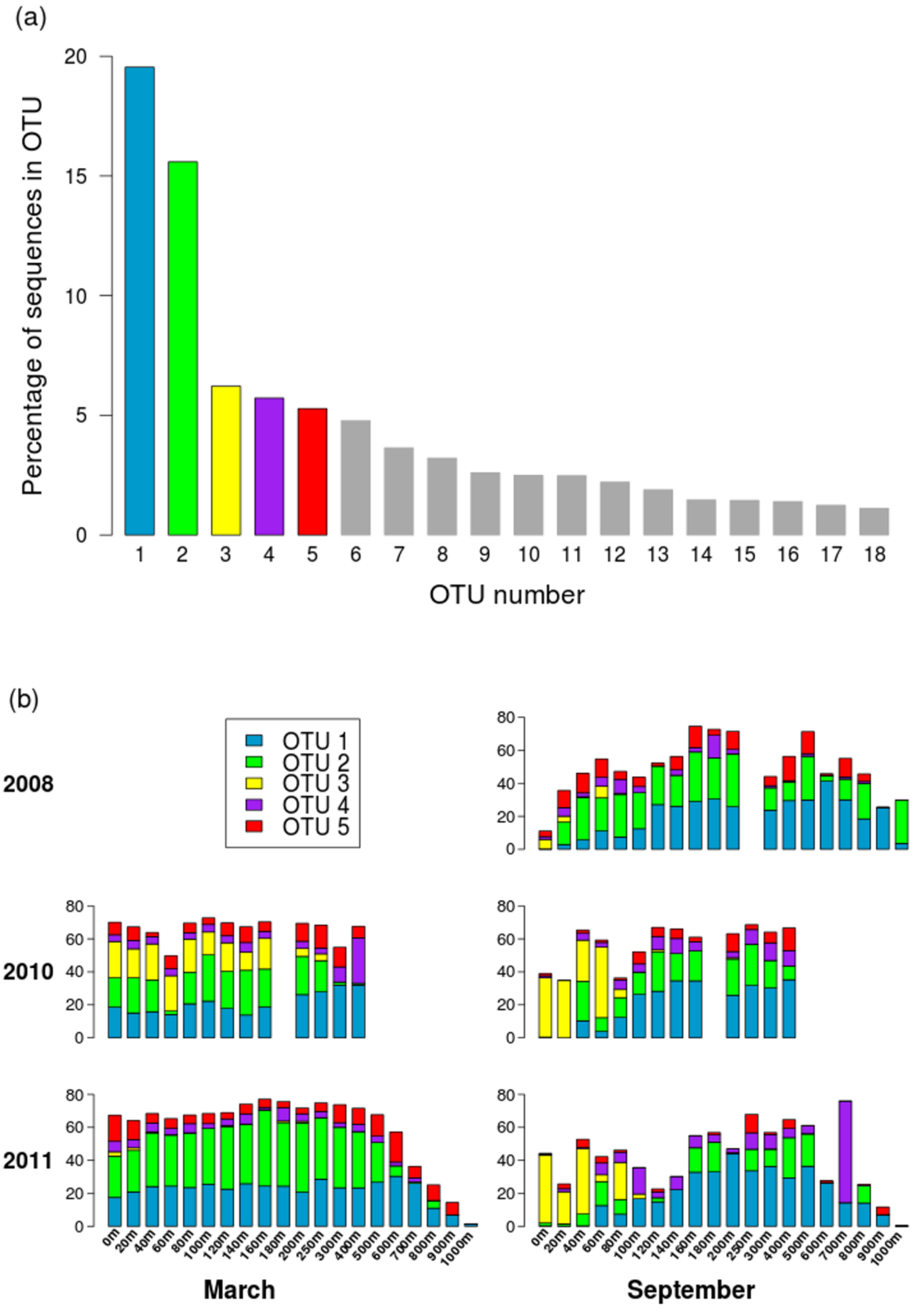


Fig. 2.5: OTU composition of total sequences and individual samples. (a) Percent of total sequences belonging to the 18 OTUs that contain at least 1% of the total sequences (n = 313,312). (b) Percent of each sample belonging to OTUs 1 through 5. An empty spot indicates absence of sample for that date/depth.

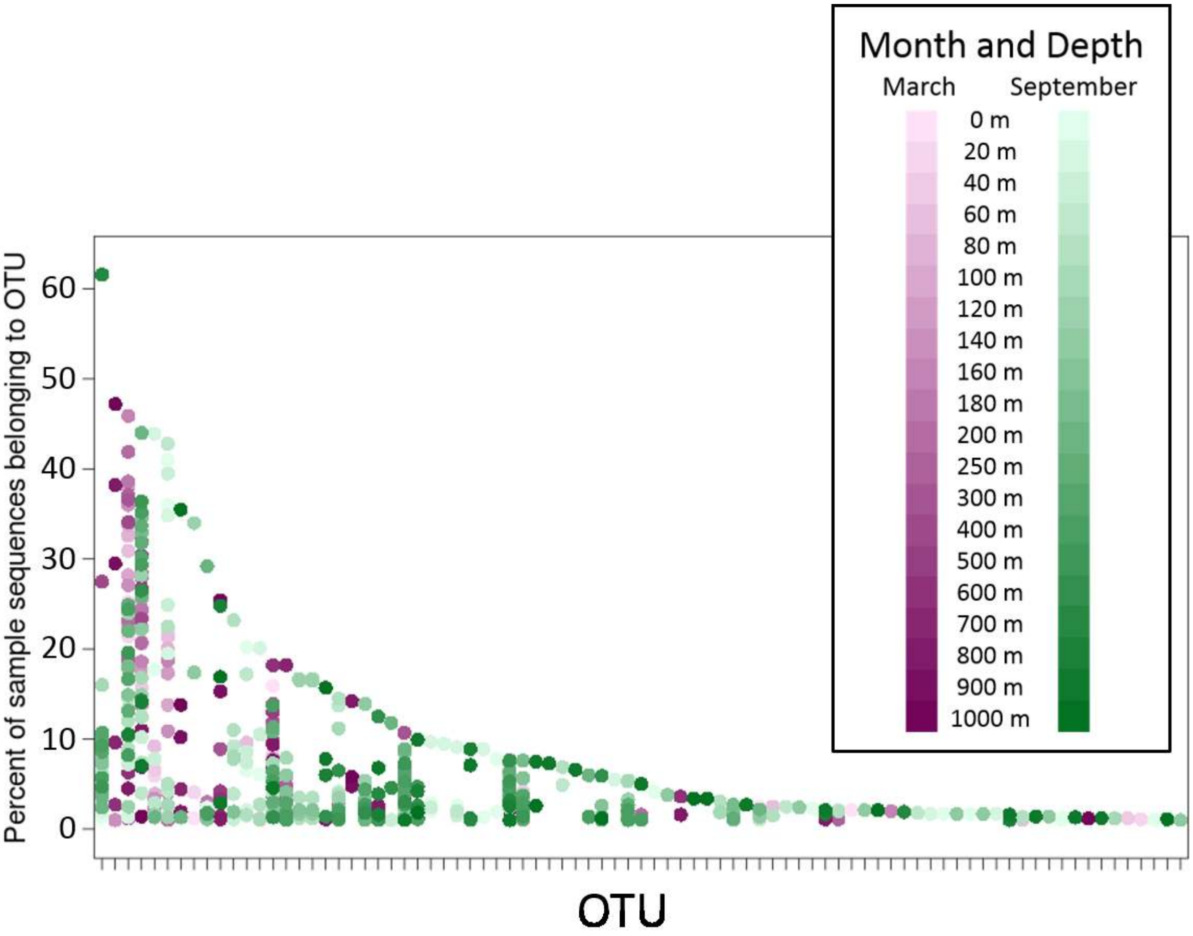


Fig. 2.6: Percent of sample sequences falling in an OTU versus OTU. The dots represent samples and are color-coded to indicate month and depth. Samples from March 2010, September 2010, March 2011, and September 2011 are displayed in this plot; the September 2008 samples are not included. The plot considers the 83 OTUs that contain ≥1% of the sequences from at least one sample from 2010 or 2011. No dots are displayed in OTU columns for samples in which less than 1% of the sample's sequences belong to that OTU. OTUs are arranged along the x-axis in descending order of largest contribution to any single sample.

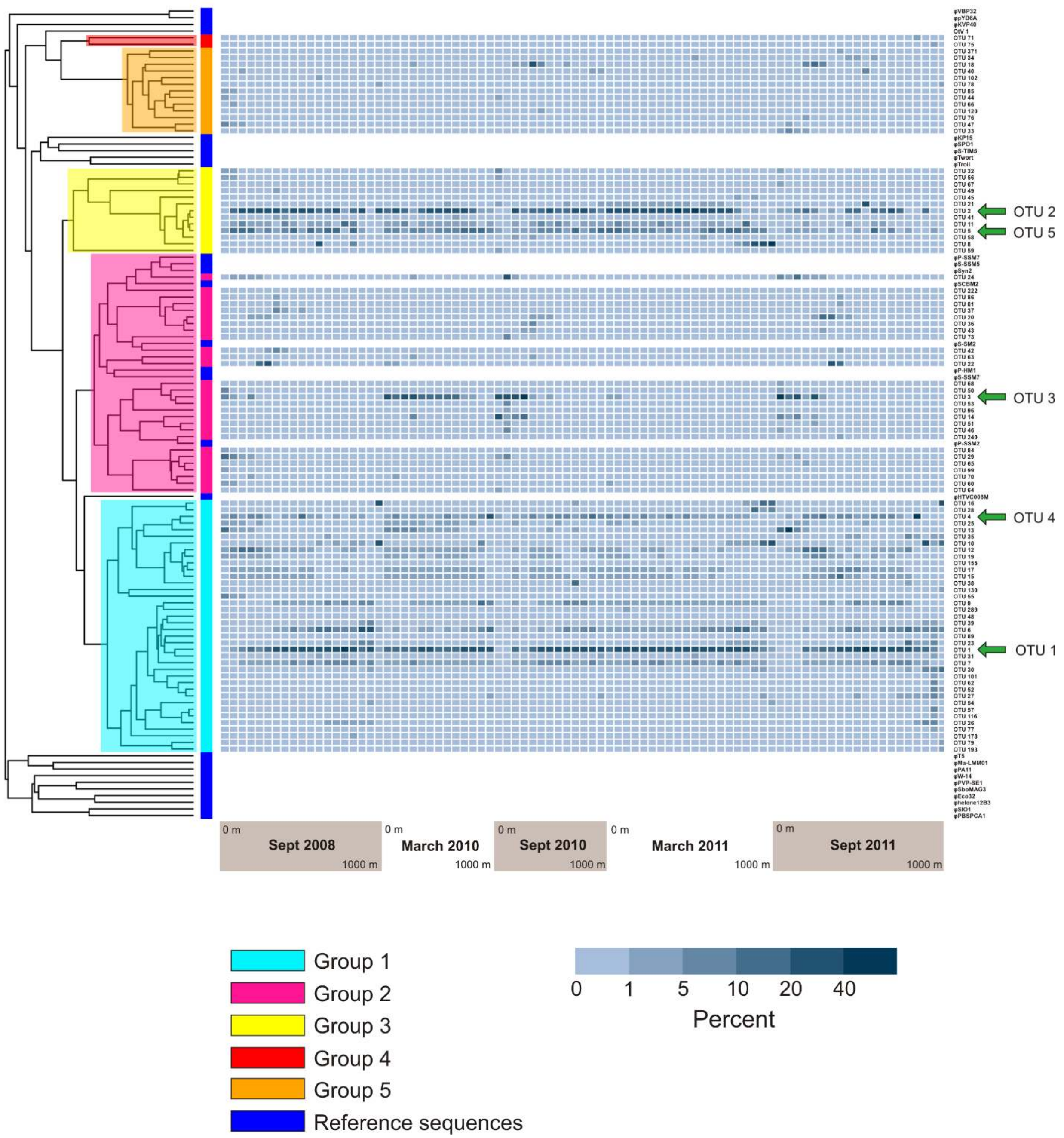


Fig. 2.7: Prevalence of phylogenetically clustered OTUs in each sample, indicated as percent of each sample's sequences that come from each of the 94 top OTUs. Reference phoH sequences from fully-sequenced phage genomes (and one eukaryotic virus) are indicated with dark blue in the vertical color bar along the left side of the heatmap, between the heatmap and tree. The phylogenetic Groups 1-5 indicated in the tree are the same groups designated in Goldsmith et al. (2011).

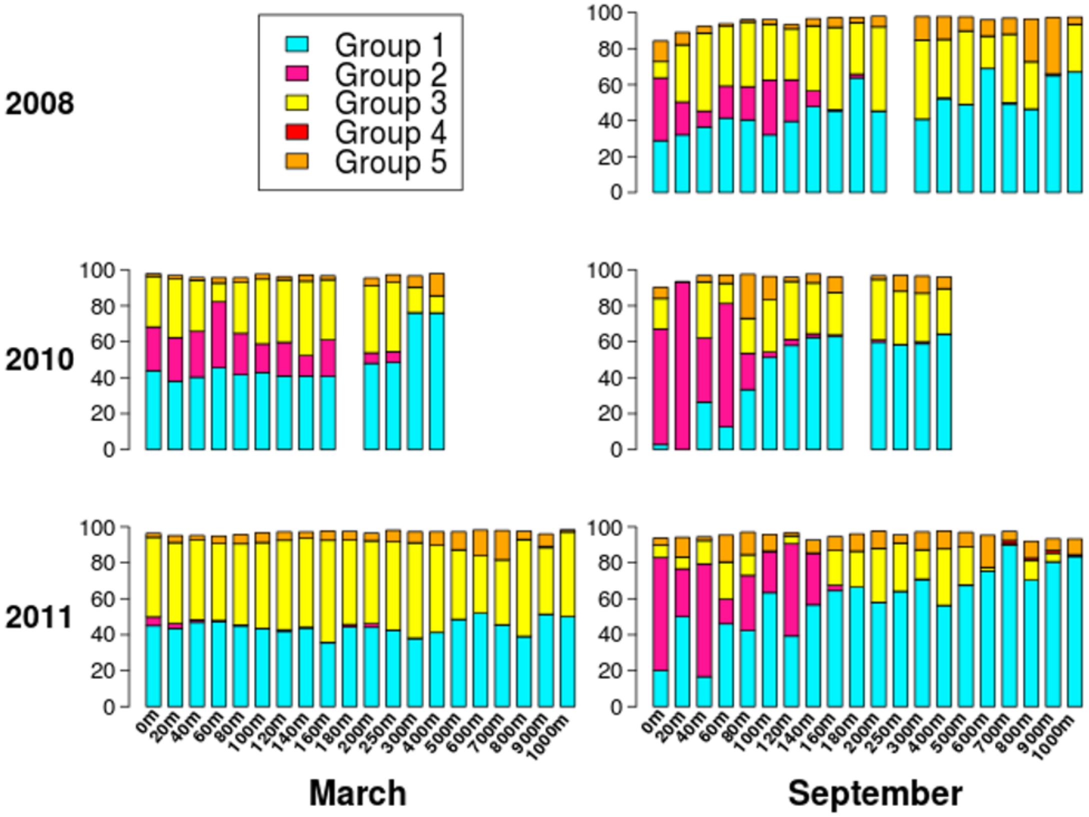


Fig. 2.8: Percent of each sample belonging to phylogenetic Groups 1 through 5. An empty spot indicates absence of sample for that date/depth.

Table 2.2: Barcodes used to tag phoH amplicons.

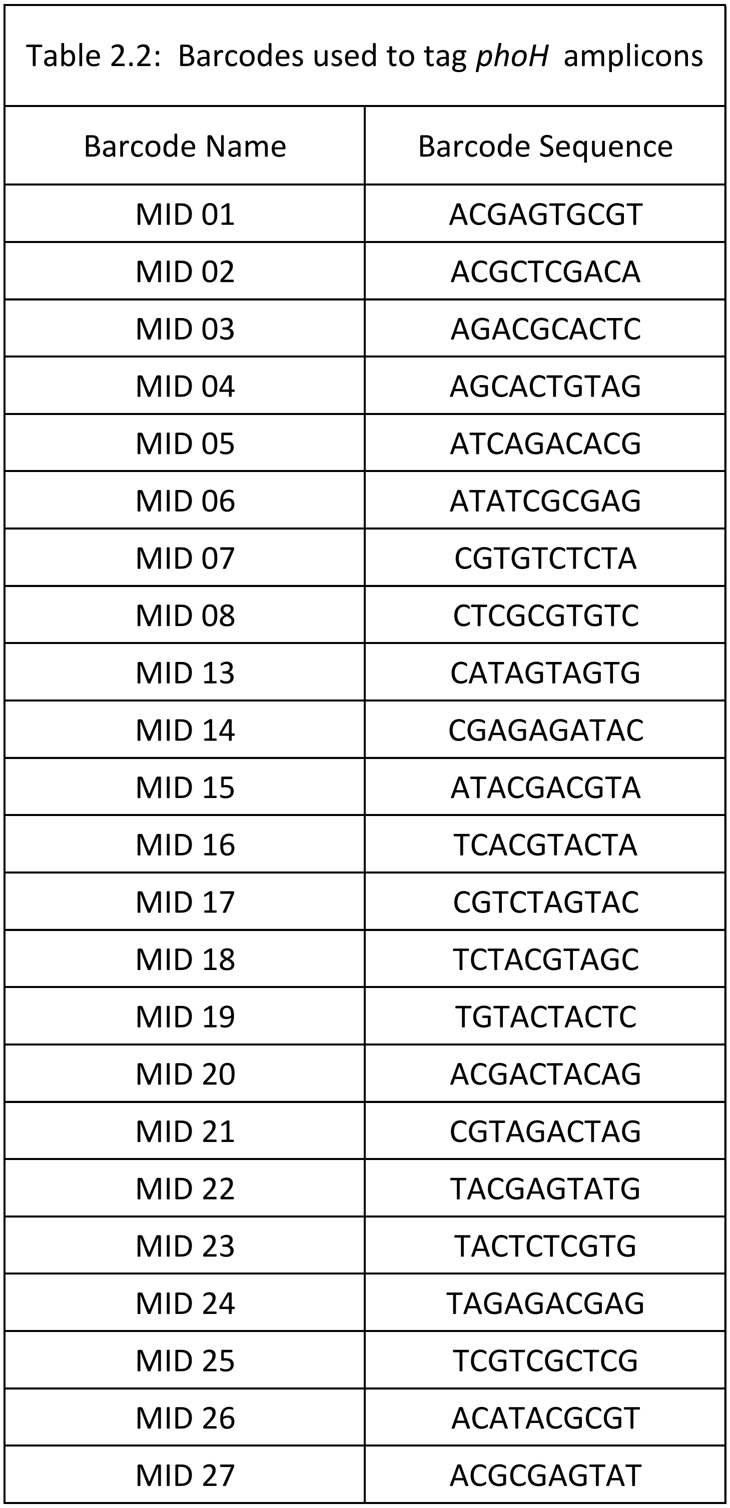


Table 2.2: Barcodes used to tag *phoH* amplicons

Barcode Name	Barcode Sequence
MID 01	ACGAGTGCGT
MID 02	ACGCTCGACA
MID 03	AGACGCACTC
MID 04	AGCACTGTAG
MID 05	ATCAGACACG
MID 06	ATATCGCGAG
MID 07	CGTGTCTCTA
MID 08	CTCGCGTGTC
MID 13	CATAGTAGTG
MID 14	CGAGAGATAC
MID 15	ATACGACGTA
MID 16	TCACGTACTA
MID 17	CGTCTAGTAC
MID 18	TCTACGTAGC
MID 19	TGTACTACTC
MID 20	ACGACTACAG
MID 21	CGTAGACTAG
MID 22	TACGAGTATG
MID 23	TACTCTCGTG
MID 24	TAGAGACGAG
MID 25	TCGTCGCTCG
MID 26	ACATACGCGT
MID 27	ACGCGAGTAT

CHAPTER 3

Depth and seasonal variation in viral diversity in the northwestern Sargasso Sea

Summary

The Sargasso Sea is an excellent place to study viral diversity because it is home to the Bermuda Atlantic Time-series Study (BATS), one of the world's longest-running ocean time series studies. Knowledge of viral dynamics at BATS expanded greatly when a long-term study that counted viral abundance at 12-13 depths every month for ten years revealed an annually recurring subsurface peak in viral abundance between 60-100 m every summer (Parsons et al., 2012). The present study was designed to determine whether in summer (September), when the mixed layer is shallow, the surface viral community differs from the viral assemblage occupying the depth of the abundance peak, which lies below the mixed layer. This study also examined the composition of the viral communities at those two depths in winter (March), when the water column is well-mixed, and hypothesized that in winter, the viral communities at the surface and 100 m depth would resemble each other, while during summer stratification, differences in the viruses occupying the two depths would emerge. Four techniques were employed to examine composition of the viral community. Statistical analysis of signature gene sequences (each targeting a different family or subset of phage) and RAPD gel banding patterns were used to determine degree of similarity among the viral communities. The results of this study supported the hypotheses. All analyses showed that in winter, the surface and 100 m viral communities almost always clustered together, while this did not occur in summer. Morever, interannual

comparisons revealed that summer surface viral communities often clustered with same assemblages from other years, but did not resemble the winter assemblages, and usually grouped separately from most of the other viral communities. Possible explanations for these findings include the greater effect of ultraviolet radiation on viruses residing in the surface during summer, as well as differences in bacterial communities between the surface and 100 m while the water column is stratified.

Introduction

The Sargasso Sea is a seasonally oligotrophic portion of the Northern Atlantic Ocean. The Sargasso Sea is the site of the Bermuda Atlantic Time-series Study (BATS), one of the world's longest-running oceanographic time series. The primary seasonal characteristics of the BATS site are the annual deep winter convective mixing, followed by strong summer thermal stratification (Steinberg et al., 2001). Specifically, in winter, convective mixing leads to deepening of the mixed layer down to a maximum of 150 m to more than 300 m and results in nutrient enrichment of the surface layer (Michaels et al., 1994; Michaels and Knap, 1996; Lomas et al., 2013). As summer approaches, high pressure systems and thermal stratification decrease the amount of mixing because they prevent fronts from passing and result in lower wind stress. These factors lead to stratification of the water column, resulting in a warm mixed layer that can be as shallow as 10-20 m. As fall progresses, temperatures cool, winds increase, the mixed layer deepens, and the cycle begins again (Steinberg et al., 2001).

Core monthly measurements at the BATS site include salinity, dissolved oxygen, total CO_2, chlorophyll *a*, alkalinity, and nutrient measurements such as nitrate, nitrite, phosphate, silicate, dissolved organic carbon and nitrogen, and particulate organic carbon and nitrogen (Steinberg et al., 2001). Bacteria are counted, and the rates of primary production, particle flux, and bacterial production are measured (data available from the BATS web site, www.bios.edu/research/projects/bats). Moreover, because of the foundation in place for obtaining monthly measurements, a wealth of ancillary research covering a wide range of topics stems from the BATS site. One such project explored dynamics in viral abundance at BATS in great detail. In a 10-year study, Parsons et al. (2012) counted viral abundance every month at 11-12 depths from the surface to 300 m. They discovered that a subsurface maximum in viral abundance recurs every year in late summer, at approximately 80 m to 100 m depth. The Parsons study combined investigation of viral abundance with an examination of the abundance of several bacterial lineages in order to determine whether viral dynamics were correlated with the dynamics of bacteria at the BATS site. While viral dynamics did not correlate with total bacterial counts, viral abundance was negatively correlated with the abundance of SAR11, the dominant heterotrophic bacterial lineage, and positively correlated with *Prochlorococcus* populations, the most abundant cyanobacteria at BATS.

The Parsons et al. (2012) study raises numerous questions. Given the striking difference in viral abundance at BATS between winter (March) and summer (September) at the depth of the subsurface peak, does the composition of the viral community also vary seasonally and with depth? In winter, when viral abundance is fairly constant throughout the upper 100 m, do the surface and 100 m viral communities resemble each other? Both of those communities are within the mixed layer in winter, as the depth of the mixed layer descends to well below 100 m. However, during summer stratification of the water column the subsurface peak of viral abundance is below the mixed layer. In summer, then, when the surface communities and the 100 m viral communities vary greatly in abundance, are they composed of different viruses, or are they comprised of the same types of viruses, just concentrated more heavily at the subsurface peak?

This study was designed to address these questions. Based on the abundance and dynamics data demonstrated by Parsons et al. (Parsons et al., 2012), we hypothesized that the composition of the viral communities at the surface and 100 m at BATS would vary seasonally and display annually recurring patterns. In winter (March), when the mixed layer is deeper, we hypothesized that the surface viral community would resemble the 100 m viral community more closely than in summer. In summer (September), when the mixed layer is shallower, we expected to see greater differences between the two viral communities. In order to test these hypotheses, this study investigated the composition of the viral community at BATS over a multiyear period. Several different methods were utilized to determine viral community composition: randomly amplified polymorphic DNA (RAPD) PCR was used for viral community profiling, and amplification and sequencing of three different signature genes were used to compare specific subsets of the viral community. Both depths (surface and 100 m) were examined in both seasons (winter and summer) in three different years. By combining RAPD PCR with the analysis of several signature genes, we were able to obtain a broad picture of the spatial and temporal variability in viral diversity at this well-studied marine site.

Methods

Sample collection and processing. Samples were collected from the surface and 100 m depth in the Sargasso Sea in March and September in 2008, 2010, and 2011 (see Table 3.1 for sampling dates and viral counts). All samples were collected in the vicinity of the BATS site (31º40′ N, 64º10′ W). Approximately 200 liters for each sample (see Table 3.1; volumes ranged from 90 L to 383 L, median volume 245 L) were concentrated by tangential flow filtration with

100-kDa filters (GE Healthcare, Piscataway, NJ) to a volume of approximately 50 mL. The viral concentrates were filtered through 0.22-µm Sterivex filters to remove bacteria and stored at 4°C until further processing. Viruses were further concentrated and purified from the Sargasso Sea concentrates by polyethylene glycol precipitation followed by cesium chloride density-dependent centrifugation (Thurber et al., 2009). Solid polyethylene glycol 8000 (PEG 8000) was added to the concentrates at a final concentration of 10% (wt/vol), and the concentrates were stored at 4°C overnight. The concentrates were then centrifuged for 40 min at 11,000 x*g* and 4°C to pellet the viruses. The pelleted viruses were resuspended in 0.02-µm-filtered seawater and further purified through ultracentrifugation in a cesium chloride density gradient with layers of 1.2 g/mL, 1.5 g/mL, and 1.7 g/mL (22,000 rpm on a Beckman SW40 Ti rotor for 3 h at 4°C). The viral fractions from the September 2008 samples were further concentrated with a Microcon centrifugal filter device (Millipore, Billerica, MA). Viral DNA was extracted from all samples using the formamide method as described by Sambrook et al. (1989).

Amplification of signature genes. For signature gene analysis, the extracted DNA was amplified using the strand displacement method of the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare, Piscataway, NJ) according to the manufacturer's instructions.

The *phoH* gene was amplified by PCR using viral *phoH* primers vPhoHf (5′-TGCRGGWACAGGTAARACAT-3′) and vPhoHr (5′-TCRCCRCAGAAAAYMATTTT-3′) (Goldsmith et al., 2011). The 50-µL reaction mixture contained 1 U Apex Taq DNA polymerase (Genesee Scientific, San Diego, CA), 1X Apex reaction buffer, 1.5 mM Apex $MgCl_2$, 0.5 µM concentration of each primer, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.08% bovine serum albumin, and 1 µL of template DNA (Genomiphi product). The reaction conditions were:

(i) 5 min of initial denaturation at 95ºC; (ii) 35 cycles of 1 min of denaturation (95ºC), 1 min of annealing (53ºC), and 1 min of extension (72ºC); and (iii) 10 min of final extension at 72ºC.

The g23 gene was amplified using primers T4superF (5′-GAYHTIKSIGGIGTICARCCIATG-3′) and T4superR (5′-GCIYKIARRTCYTGIGCIARYTC-3′) (designed by Andre Comeau; published in Chow and Fuhrman (2012)). The 50-μL PCR mixture contained 1 U Apex Taq DNA polymerase, 1X Apex reaction buffer, 1.5 mM Apex $MgCl_2$, 1.0 μM concentration of each primer, 0.2 mM dNTPs, and 1 μL of template DNA (Genomiphi product). The reaction conditions were: (i) 5 min of initial denaturation at 94ºC; (ii) 35 cycles of 1 min of denaturation (94ºC), 1 min of annealing (58ºC), and 1 min of extension (72ºC); and (iii) 10 min of final extension at 72ºC.

The ssDNA virus major capsid gene was amplified by Max Hopkins using primers MCPf (5′-CCYKGKYYNCARAAAGG-3′) and MCPr (5′-AHCKYTCYTGRTADCC-3′) (Hopkins et al., 2014). The 50-μL PCR mixture contained 1 U Apex Taq DNA polymerase, 1X Apex Taq reaction buffer, 0.5 μM of each primer, 0.2 mM dNTPs and 1 μL of template DNA (Genomiphi product). The touchdown PCR conditions were (i) 3 min of initial denaturation at 94ºC; (ii) 32 cycles of 60 s of denaturation (95ºC), 45 s of annealing (47ºC with a 0.11ºC decrease/cycle), and 90 s of extension (72ºC); and (iii) 10 min of final extension at 72ºC.

Cloning and sequencing of signature genes. PCR products were cleaned using the MO BIO UltraClean PCR Clean-Up Kit (MO BIO Laboratories, Inc., Carlsbad, CA) following the manufacturer’s instructions. After tailing with Sigma-Aldrich REDTaq DNA polymerase (Sigma-Aldrich, St. Louis, MO), PCR products were cloned into vectors using the TOPO TA cloning kit for sequencing (Invitrogen, Carlsbad, CA) and were then used to transform competent cells. The cells were then screened, and the inserts in positive transformants were sequenced with the M13F primer by Beckman Genomics (Danvers, MA).

Data analysis for signature genes. Vector and low-quality sequences were trimmed with Sequencher 4.7 (Gene Codes, Ann Arbor, MI). The sequences were dereplicated into operational taxonomic units (OTUs) using FastGroup II at a level of 98% sequence identity with gaps (Yu et al., 2006). FastGroup representative sequences and reference sequences were aligned at the amino acid level with Muscle (Edgar, 2004) using the default parameters as implemented by TranslatorX (Abascal et al., 2010). The alignments were cleaned with Gblocks (as implemented by TranslatorX) using the options for a less stringent selection (Castresana, 2000; Talavera and Castresana, 2007). Back-translated nucleotide alignments were used to build maximum-likelihood phylogenetic trees with FastTree version 2.1 (Price et al., 2010). Branch supports in FastTree were calculated using the Shimodaira-Hasegawa-like approximate likelihood ratio test on 1000 resamplings. Branches with support below 50 were collapsed using TreeCollapseCL 4 (Hodcroft, 2013). Hierarchical clustering was performed in R (R Development Core Team, 2013) from a Bray-Curtis dissimilarity matrix based on OTU abundance data using the picante package (Kembel et al., 2010). In order to bootstrap the dendrograms, Jaccard stability means were computed using the fpc package (Hennig, 2007, 2008, 2013). The Jaccard similarity value, which represents the stability of the cluster, is averaged for every bootstrapping of the clustering (1000 times), resulting in a Jaccard stability mean for each cluster. Clusters with Jaccard stability means of 75 and greater can be considered valid, stable clusters. Clusters supported by Jaccard stability means between 60 and 75 indicate patterns in the data (Hennig, 2007, 2008, 2013).

RAPD PCR amplification and data analysis. Randomly-amplified polymorphic DNA (RAPD) PCR was conducted on the extracted DNA using primer SP2 (5′-CGCAACAGGG-3′). This primer was designed by Shawn Polson by identifying the most common decamers in a viral metagenome from the Sargasso Sea (Srinivasiah et al., 2013). Each 50-µL reaction contained 2 U of Apex *Taq* DNA polymerase (Genesee Scientific, San Diego, CA), 1X Apex *Taq* reaction buffer, 1.5 mM Apex $MgCl_2$, 4 µM of primer, 0.2 mM dNTPs, and 1 µL of template DNA. Reaction conditions for the RAPD PCR were: (i) 10 min of initial denaturation at 94ºC; (ii) 35 cycles of 3 min of annealing at 35ºC, 1 min of extension at 72ºC, and 30 s of denaturation at 94ºC; (iii) 3 min of annealing at 35ºC; and (iv) 10 min of final extension at 72ºC. PCR products were cleaned using the MO BIO UltraClean PCR Clean-Up Kit (MO BIO Laboratories, Inc., Carlsbad, CA) following the manufacturer's instructions. The RAPD PCR viral community fingerprints were visualized on the Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) following the manufacturer's instructions. Bioanalyzer results were analyzed using GelCompar II, version 6.5 (Applied Maths, Austin, TX), which was also used to build a similarity matrix and clustering dendrogram.

Discussion of methods

The signature genes chosen for this study each target a different subset of marine phage. The g23 gene is structural, encoding the major capsid protein for T4-like myophage (Filée et al., 2005; Comeau and Krisch, 2008), and the primers used in this study were designed by André Comeau in order to capture a broader group of myophage than the primers previously used for g23 (Filée et al., 2005; Chow and Fuhrman, 2012). In contrast, the phoH gene is found in more than one morphological type of phage. While the function of phoH in phage is unknown, it is similar to a gene in the Pho regulon of *E. coli* that is upregulated during phosphate stress

(Wanner, 1996; Sullivan et al., 2010a). The phoH gene is present in myoviruses, podoviruses, and siphoviruses and occurs in viruses that infect heterotrophs as well as viruses whose hosts are autotrophic (Goldsmith et al., 2011). In addition, this gene is more prevalent among marine phage than in the genomes of phage isolated from other environments. While both g23 and phoH target double-stranded DNA (dsDNA) viruses, the third signature gene used in this study targets the major capsid protein (MCP) gene of single-stranded DNA (ssDNA) phage belonging to the *Gokushovirinae* subfamily (family *Microviridae*) (Hopkins et al., 2014), which were recently shown to be abundant in the Sargasso Sea (Angly et al., 2006; Tucker et al., 2011). The combined analysis of these three signature genes, each of which targets a different type of phage, provides a broader picture of the viral community than analysis of any single signature gene can offer. Statistical analysis of each signature gene sequence allows us to determine the degree to which viral communities from different depths, seasons, or years are similar. For each signature gene, after assigning the sequences to OTUs based on 98% sequence identity, the abundance of each OTU in each sample was calculated. The abundance matrices were then used to calculate the Bray-Curtis dissimilarity for each of the samples. Hierarchical clustering based on the dissimilarity was computed for each gene and displayed in a dendrogram (see Figs. 3.1-3.3).

Analysis of signature genes provides insight into the diversity of a viral community based on an individual gene present in a specific subset of the viral community. Analyzing several signature genes, each representing a different group of viruses, yields a more complete reflection of the overall viral community composition. To further bolster the analysis of viral diversity, this study also incorporated an analysis of the community through RAPD PCR. In contrast to the signature genes, RAPD PCR provides a view of the whole viral community by using a random decamer primer rather than primers designed to capture a specific gene. Thus while signature genes illuminate a subset of the viral community, the gel banding pattern of RAPD PCR products can be viewed as a fingerprint of the viral community at the time of sampling (Winter and Weinbauer, 2010). In order to visualize the community profile for each sample and compare those profiles, a similarity matrix was produced by comparing the gels resulting from the Bioanalyzer runs for the RAPD PCR products for each date and depth. The samples were then clustered using the unweighted pair group method with arithmetic mean (Fig. 3.4).

Results and discussion

In winter, the surface and 100 m viral communities at BATS resemble each other. Hierarchical clustering based on the Bray-Curtis dissimilarity for all three sets of signature gene sequences (g23 (Fig. 3.1), phoH (Fig. 3.2), ssDNA MCP (Fig. 3.3)) revealed that in both 2008 and 2011, the March viral communities from the surface and 100 m of a given year clustered together. For g23 and phoH, March 2010 was an exception; the winter surface and 100 m communities did not cluster together for that year. In addition, for phoH and g23, the March samples (both depths) from 2008 and 2011, as well as one of the March 2010 samples, clustered together. The March 100 m phoH community from 2010 was an exception; in the dendrogram it appears in a strongly-supported split, separate from the large cluster that contains all of the other winter samples. The March surface g23 community from 2010 also appears in a strongly-supported split, separate from the large cluster containing all of the other winter samples. No PCR products could be obtained for the ssDNA MCP for the March 2010 0 m community, so we are unable to assess its ssDNA community structure; however, the inability to obtain PCR amplification suggests that this community is distinct from the other recovered communities. The RAPD viral community fingerprints from all three years supported the hypothesis that the surface and 100 m samples would be similar in the well-mixed water column in March: within each year, the March 0 m and March 100 m samples were most similar to each other. Overall, the data suggest that the March communities (0 m and 100 m) from a given year generally cluster more closely with each other than with any other date or depth.

In contrast to the winter viral assemblages, the surface and 100 m summer viral communities at BATS have distinctly different compositions within a given year. The summer surface populations from the three years often cluster with each other, but separately from 100 m summer communities and the winter communities. For example, in the g23 analysis (Fig. 3.1) and the RAPD analysis (Fig. 3.4), all three summer surface samples group together, apart from nearly all of the other samples. In the phoH and ssDNA MCP communities, two of the summer surface samples are in the same cluster (2008 and 2010 for phoH; 2010 and 2011 for ssDNA MCP), while the third sample splits strongly from all other samples. However, the 100 m summer virus populations behave differently, grouping more closely with the winter viruses than the surface summer viruses do. For all three signature genes, two of the summer 100 m samples group together, and then form a larger cluster with the third summer 100 m samples and all of the winter samples (Figs. 3.1-3.3). According to the RAPD analysis, the summer 100 m samples group most closely with the winter samples from the same year for two out of the three years (2008 and 2011), while the 2010 summer 100 m sample is separate from all other samples (Fig. 3.4). All methods of analysis therefore support the hypothesis that in summer, the BATS viral communities from the surface and from 100 m are distinct from each other.

The results of this study indicate that it is often the summer surface samples that are the most different from the other communities. While the summer surface samples from the three years are sometimes related to each other, they generally group separately from both sets of winter samples (surface and 100 m) and the 100 m summer samples. A potential reason for differences between the surface viral communities in summer and all other viral communities is the greater influence of ultraviolet light at the surface during that season. Ultraviolet radiation degrades viral particles (Wommack et al., 1996; Weinbauer and Suttle, 1999; Araújo and Godinho, 2009), and would thereby create greater differences between the summer viral communities than between the winter communities. Another explanation for the September 0 m viral communities' difference from the other viral assemblages may stem from the changes in the stability of the water column in the summer at BATS. As the mixed layer begins to shoal after the winter deep mixing, the bacterial community residing near the surface begins to change (Morris et al., 2005; Carlson et al., 2009; Treusch et al., 2009). Under this model, as the surface bacterial community changes, so would its associated viral community.

To understand why the summer viral assemblages are different between the surface and 100 m depth, future studies should examine the summer bacterial community composition in the upper water column. Maximum *Prochlorococcus* concentrations (on the order of 10^5 cells/mL) occur near 60-80 m depth in summer and fall, and high concentrations persist to nearly 200 m (DuRand et al., 2001a). Viruses that prey upon *Prochlorococcus* likely assemble in the vicinity of the host peak. If a significant portion of the phage community consists of phage that infect *Prochlorococcus*, the differential in *Prochlorococcus* abundance between the surface and 100 m would account for some of the difference between the surface and 100 m viral assemblages, since surface *Prochlorococcus* concentrations are lower than they are at 100 m at that time of year (DuRand et al., 2001a; Malmstrom et al., 2010; Parsons et al., 2012). Research has shown that cyanomyophage, whose hosts include *Prochlorococcus*, are an important component of the viral community in the Sargasso Sea; they have been found to constitute up to 25% of the viral community at the surface during a June transect (Matteson et al., 2013). Moreover, based on the phylogenetic trees for g23 and phoH (Figs. 3.5 and 3.6), we can reasonably infer that cyanophage are a component of our sampled viral communities, because the sequences from our sampled communities cluster in the groups with fully-sequenced cyanophage.

In contrast to *Prochlorococcus*, for which phage infection has been well-studied, and numerous phage genomes have been sequenced, much less is known about phage infecting other members of the BATS microbial community. The most abundant heterotrophic bacteria at BATS belong to the SAR11 clade (Morris et al., 2002), for which no phage were known until 2013. In fact, previous studies had even suggested that SAR11 may be resistant to phage infection (Suttle, 2007). In 2013, "pelagiphage" were cultured on SAR11 and their genomes were sequenced (Zhao et al., 2013). BLAST searches of oceanic viromes (including the Sargasso Sea) with these phage genome sequences suggest that pelagiphage comprise a large proportion of oceanic viral communities. All four of the cultured pelagiphage are double-stranded DNA phage, and only one of them (HTVC008M) contains one of the signature genes used in this study (phoH), so it is possible that pelagiphage were not covered in the analyses of this study. Thus additional tools are needed for investigating the phage of the dominant heterotrophs at this site. However, there is likely a high diversity of uncultured pelagiphage at this site, so some of the sequences generated in this study may in fact represent yet undescribed phage infecting SAR11 or other dominant heterotrophs. Even less well-described than the pelagiphage are marine ssDNA phage, for which no cultured representatives exist. Although these phage were first identified at BATS and have recently been shown to be widespread in the oceans (Angly et al., 2006; Tucker et al., 2011; Labonté and Suttle, 2013), the hosts for these phage remain unknown, preventing study of their interactions with hosts (Hopkins et al., 2014). In addition, examining the joint dynamics of phage and specific bacterial lineages at BATS would benefit from better representation of environmental phage genomes in sequence databases. The sequencing and phylogenetic analyses performed in this study illustrate the gap that remains between cultured phage genomes and phage in the environment. The phylogenetic trees for each of the signature genes (Figs. 3.5-3.7) reveal that most of the environmental phage fall in clusters without any cultured or sequenced representatives. Therefore the phage that are sequenced so far are insufficient for accurately characterizing environmental phage communities. Substantial additional sequencing is required before we will have a database of phage genomes that better describe natural populations of viruses.

While the dynamics of specific host bacterial lineages is one factor influencing patterns in viral diversity, the data presented here suggest that the physical mixing of the water column in winter, and stratification in summer, contribute to the structure and differences in phage communities. The three signature genes encompassing varied groups of phage, as well as the RAPD fingerprinting analysis, all support the hypothesis that the winter viral communities are closely related to each other, while the summer communities are divergent. Not all genes, or the phage which encode them, should necessarily behave in the same way. Some genes may preferentially be found in phage closer to the surface (such as phage-encoded photosynthesis genes), and thus would not be expected to follow the winter pattern of being distributed equally at the surface and at 100 m depth. Nonetheless, the analyses of each of the three genes studied here, along with the insight provided by a random decamer primer known to occur frequently at this site, all point to seasonal variation in viral community composition, supporting the robustness of this pattern. This is striking considering that for all three signature genes, the dendrograms show that the dissimilarity among the samples is relatively high, ranging from approximately 55% to nearly 100% (see Figs. 3.1-3.3). Moreover, the results suggest that physical forces causing water column stratification in summer and mixing in winter are the dominant forces in controlling the distribution of viral communities at the BATS site.

Constraints of present study

Because this study's goal was to obtain an overview of the viral communities over two depths, two seasons, and several years, we sequenced broadly rather than deeply. Therefore, the data presented in this study should be viewed as a representation of the dominant members of the viral community, as opposed to an exhaustive analysis. Note that the dendrograms were built from abundance data, rather than presence/absence of OTUs. This method was chosen because of the small amount of sequencing performed, because rare communities' members would be "absent" at this level of coverage. While the samples might have exhibited greater similarity had we sequenced more deeply, most of the groupings of the samples are quite strong as measured by the Jaccard stability means, with nearly all falling within ranges that merit confidence in the patterns exhibited by the data.

Signature gene analysis and RAPD PCR capture a portion of the viral community, but cannot portray the diversity of the entire community. With RAPD PCR, only viral genomes that contain the decamer sequence used as the primer will be amplified. In that sense, RAPD analysis is similar to signature gene analysis, in that not all phage in the community will contain the sequence of interest. However, phage genomes containing the RAPD primer should be distributed more widely across phage genomes than phage bearing the signature genes, which tend to be shared by a specific family or phage type (although phoH is an exception). Sequencing the entire community through metagenomics is one possible avenue for overcoming these limitations (Angly et al., 2006).

The depth of 100 m was chosen for sampling in order to approximate as closely as possible the depth of the subsurface peak in viral abundance. From 2002 through 2006, the peak in abundance extended to 100 m. Sampling for this project began in March 2008, and the remaining sampling continued at 100 m for consistency. This depth was also chosen in order ensure sampling in summer below the mixed layer. Interannual variation in the depth of the viral abundance peak could explain variation in the clustering of the September 100 m viral communities. The clustering patterns of the September 100 m viral communities might be more similar from year to year if the center of the peak range had been examined. This can be accomplished by determining viral abundance while on board during the cruise, and then sampling from the depth revealed to have the highest abundance.

Conclusion

Four techniques employed to examine composition of the viral community at BATS over three years, two seasons, and two depths per season supported the hypotheses that winter viral communities at the surface and 100 m depth resemble each other, while the compositions of summer viral communities at those depths diverge. All analyses demonstrated that in winter, the surface and 100 m viral communities frequently clustered together. During summer stratification of the water column, however, the surface and 100 m viral communities were distinct. Between years, winter viral communities were generally similar, forming large groups containing nearly all of the winter samples. The summer communities behaved differently: interannual comparison of summer surface viral communities revealed that they often clustered with the summer surface communities of other years. However, the surface summer viral communities did not resemble the winter assemblages, and usually grouped separately from most of the other viral communities. In contrast, the summer 100 m viral communities, collected from the vicinity of the subsurface peak in viral abundance, occasionally grouped with other summer 100 m communities, and tended to resemble the winter communities more closely than the summer surface communities did. Physical factors such as UV irradiation of viral particles, as well as seasonal and depth-related differences in host communities related to the depth of the mixed layer may explain these findings. Future work should sample precisely at the peak depth by determining the depth of maximum viral abundance at the time of sampling, and should track the joint dynamics of viruses and hosts in order to better resolve the factors connected with the recurring seasonal patterns in viral community composition revealed here.

Tables and figures

Table 3.1: Sample collection data including viral counts and number of sequences obtained from each sample.

Year	Month	Date	Depth	Volume of water concen-trated	Number of sequences obtained (g23)	Number of sequences obtained (phoH)	Number of sequences obtained (ssDNA MCP)*	Viral conc in whole water (viruses/mL)
2008	March	24	0 m	144 L	43	40	82	4.00 x 10E6
2008	March	24	100 m	125 L	45	38	74	3.67 x 10E6
2008	September	2-3	0 m	245 L	40	44	38	2.64 x 10E6
2008	September	2-3	100 m	245 L	46	42	95	2.86 x 10E6
2010	March	8	0 m	383 L	50	37	0	2.75 x 10E6
2010	March	8	100 m	288 L	34	34	27	2.51 x 10E6
2010	September	5	0 m	245 L	43	48	91	5.25 x 10E6
2010	September	7	100 m	90 L	36	48	90	4.25 x 10E6
2011	March	27	0 m	180 L	40	48	69	4.58 x 10E6
2011	March	27	100 m	180 L	43	47	85	4.61 x 10E6
2011	September	13	0 m	280 L	44	48	96	2.46 x 10E6
2011	September	13	100 m	280 L	45	46	96	5.74 x 10E6

*The ssDNA MCP PCR products were cloned and screened by Max Hopkins.

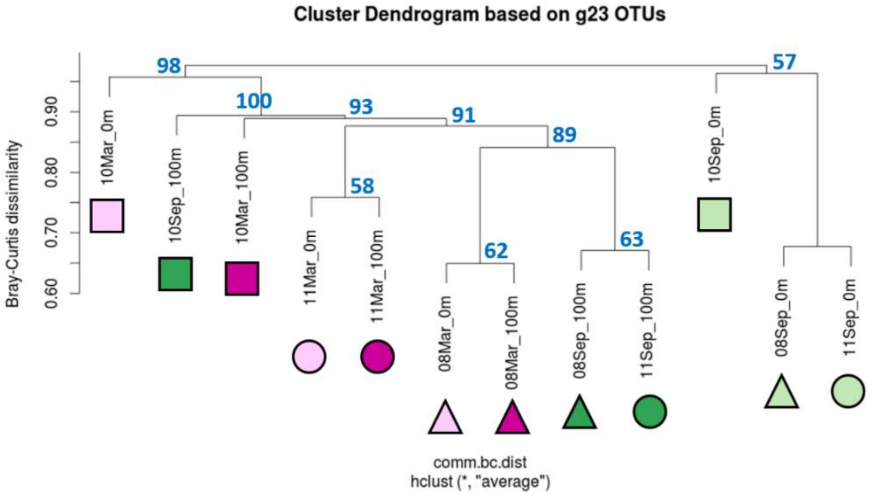


Fig. 3.1: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on g23 OTUs (98% sequence identity). Clustering is calculated from Bray-Curtis dissimilarity of the samples. Branch supports are shown where support is greater than 50 and represent Jaccard stability means. Jaccard stability means > 75 represent valid, stable clusters. Jaccard stability means from 60 to 75 indicate the presence of patterns in the data.

Month/Depth

⬭	March, 0 m
⬬	March, 100 m
⬭	September, 0 m
⬬	September, 100 m

Year		
	△	2008
	□	2010
	○	2011

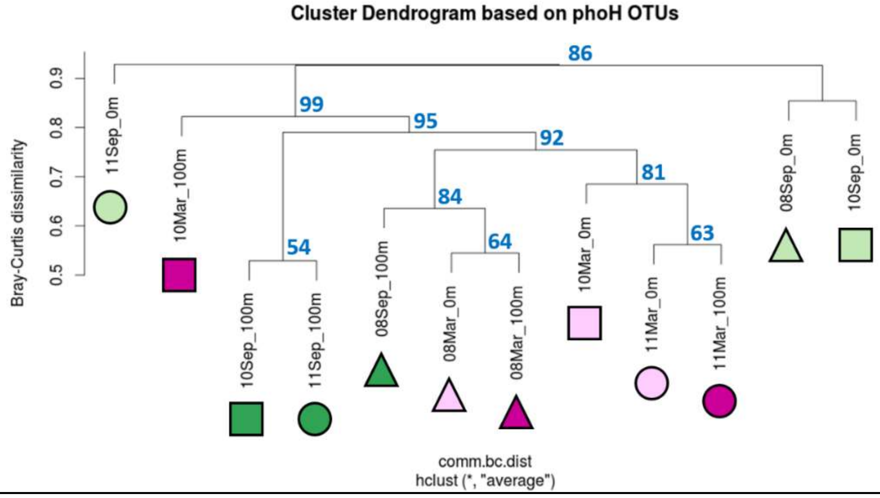


Fig. 3.2: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on phoH OTUs (98% sequence identity). Clustering is calculated from Bray-Curtis dissimilarity of the samples. Branch supports are shown where support is greater than 50 and represent Jaccard stability means. Jaccard stability means > 75 represent valid, stable clusters. Jaccard stability means from 60 to 75 indicate the presence of patterns in the data.

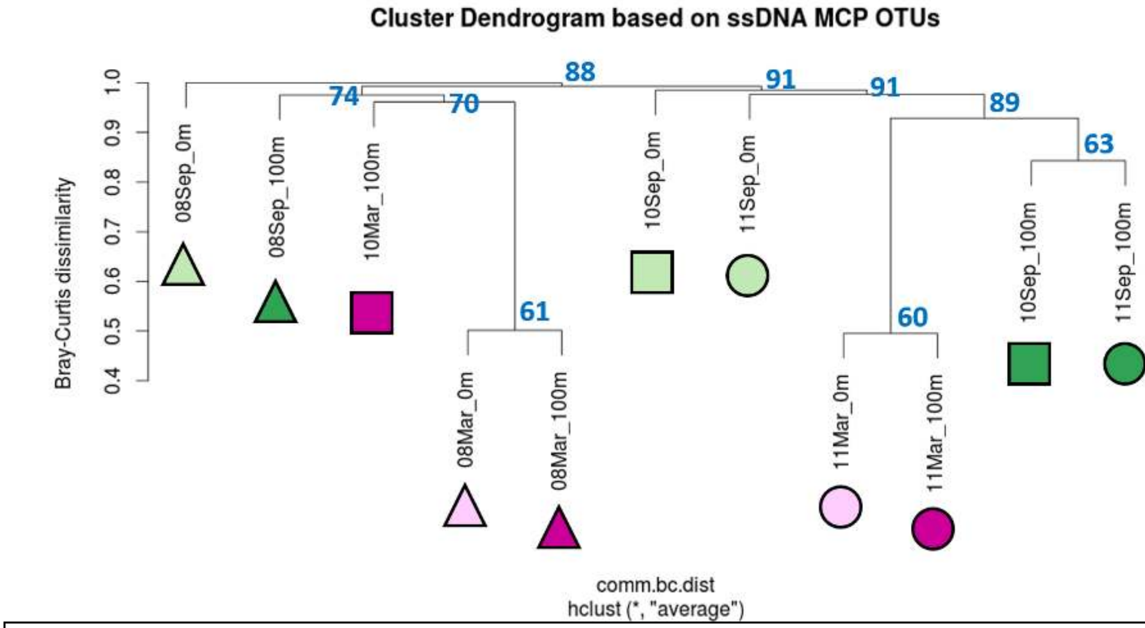
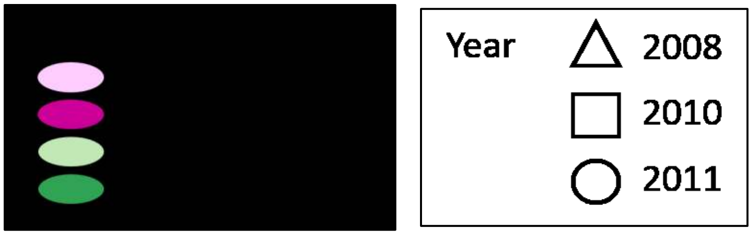


Fig. 3.3: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on ssDNA MCP OTUs (98% sequence identity). Clustering is calculated from Bray-Curtis dissimilarity of the samples. Branch supports are shown where support is greater than 50 and represent Jaccard stability means. Jaccard stability means > 75 represent valid, stable clusters. Jaccard stability means from 60 to 75 indicate the presence of patterns in the data.



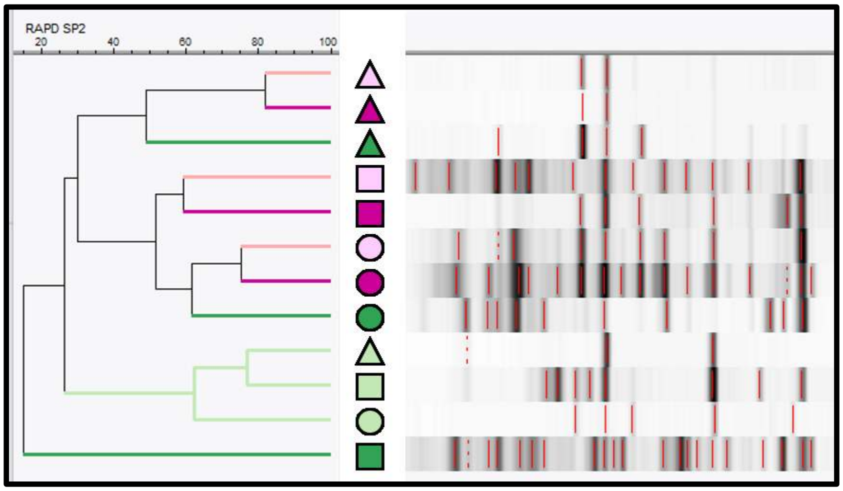


Fig. 3.4: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on gel photos of RAPD PCR

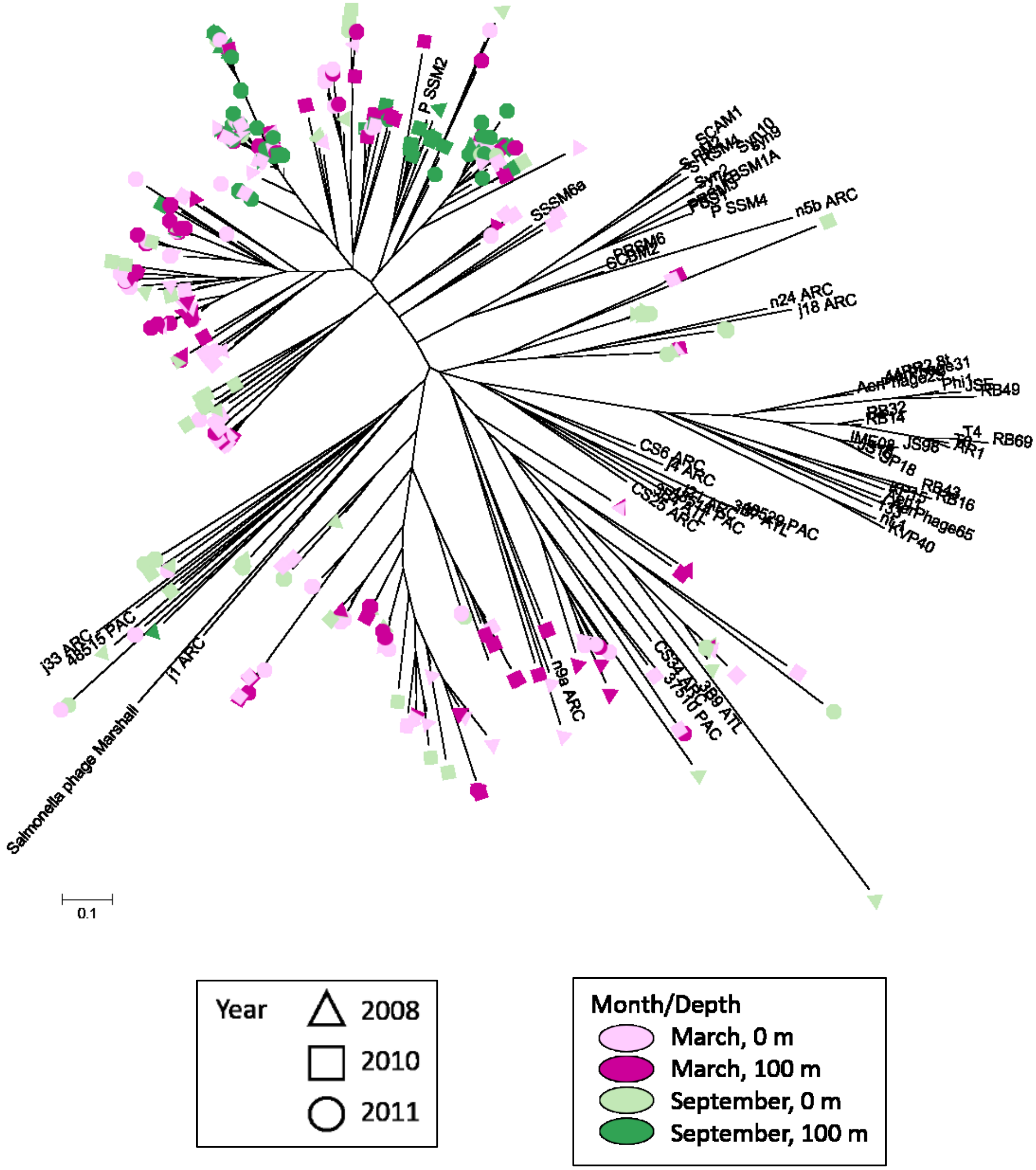


Fig. 3.5: Phylogenetic tree showing the relationship among g23 sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes) and g23 sequences from cultured phages and other environmental samples (indicated by names). The scale bar represents substitutions per site.

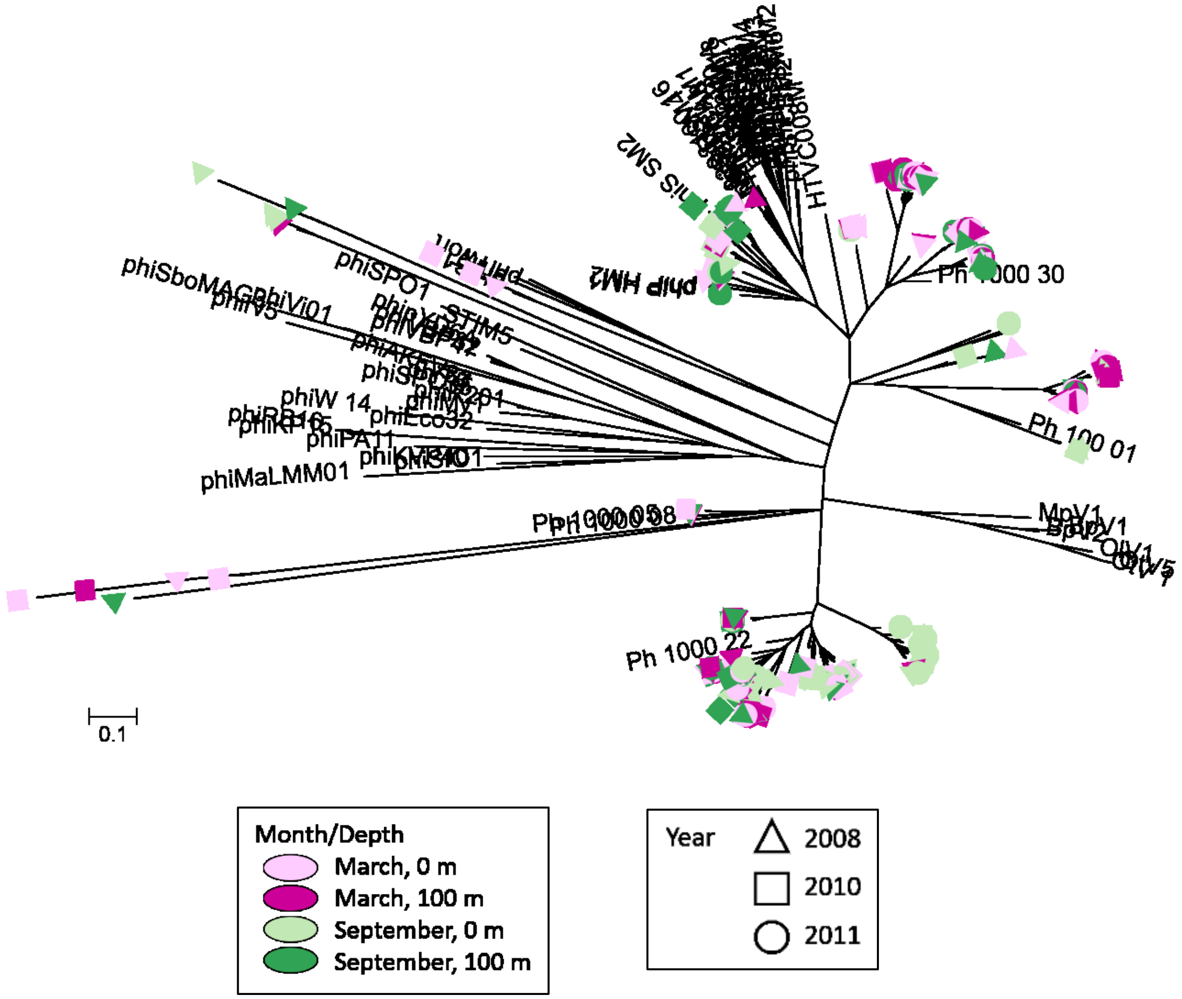


Fig. 3.6: Phylogenetic tree showing the relationship among phoH sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes), phoH sequences from cultured phages and viruses (indicated by names), and placeholder sequences from the Sargasso Sea (indicated by names beginning with "Ph_"). The scale bar represents substitutions per site.

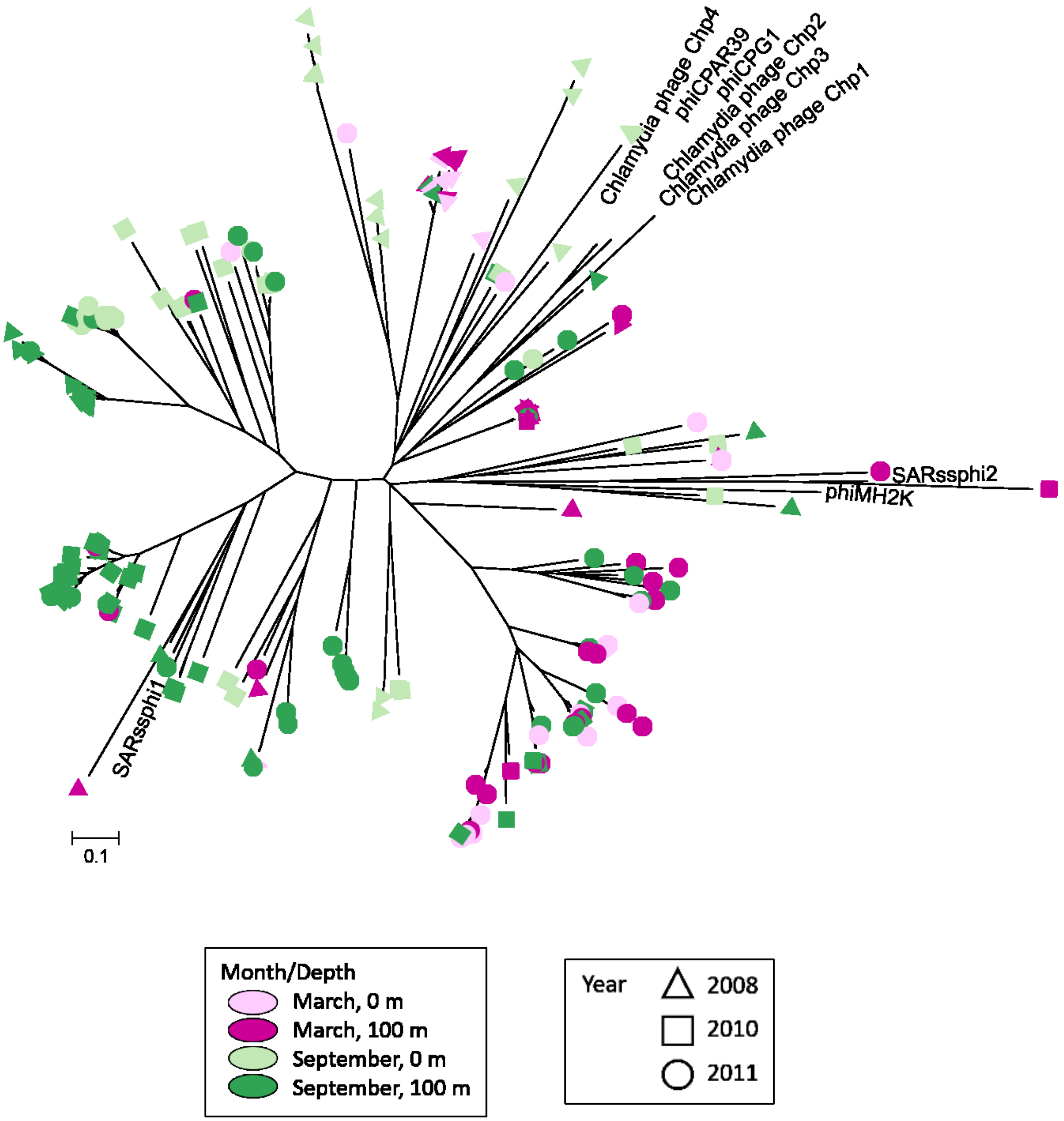


Fig. 3.7: Phylogenetic tree showing the relationship among ssDNA MCP sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes) and ssDNA MCP sequences from fully sequenced phage (indicated by names). The scale bar represents substitutions per site.

CONCLUSION

The research described here provides a significant advance in our understanding of the spatial and temporal variability in the diversity of marine viruses. Chapter 1 developed phoH as a new signature gene for assessing marine viral diversity. The phoH gene is disproportionately present in fully-sequenced marine phage, as opposed to phage isolated from non-marine environments. Moreover, the gene is widespread in the marine environment; phoH was recovered from every viral community collected (from every depth and time point) in the Sargasso Sea, as well as from all six locations in other parts of the world. Diversity of the gene was high, and most of the sequences recovered belonged to phylogenetic groups that did not contain any cultured representatives, indicating that cultured phage isolates do not adequately represent the diversity found in marine environments. Composition of the phoH communities at each sampled location and depth were distinguishable according to phylogenetic clustering, although most phoH clusters were recovered from multiple sites. These factors demonstrate that phoH will be useful for studying marine phage diversity worldwide.

Chapters 2 and 3 performed extensive examination of viral diversity at the site of the Bermuda Atlantic Time-series Study (BATS) over depth and time using the newly developed phoH marker as well as other techniques. Chapter 2 described the use of phoH to conduct a comprehensive study of the gene's diversity in the marine viral community at BATS over three different years, several seasons, and a range of depths from the surface to 1000 m. Deep sequencing performed using next-generation pyrosequencing revealed that the viruses at BATS contain a large pool of phoH sequences, but that most of those sequences are rare. The phoH sequences were dominated by just a few operational taxonomic units (OTUs). Only 1% of the >3600 OTUs recovered comprised at least 5% of any sample. Rarefaction analysis showed that the sequencing was sufficient to capture the diversity of the gene at BATS, and in fact no new phylogenetic clusters were identified that were not seen in the small amount of Sanger sequencing performed for the initial phoH study in Chapter 1. Some of the more abundant OTUs recurred every season and every year, in varying degrees, although similar depths and seasons clustered together. Overall, the phoH gene revealed depth-based, seasonal, and interannual differences in the diversity of the viral community at BATS.

Chapter 3 used several methods to study changes in diversity of the viral community at BATS between winter and summer over two depths in three different years. A ten-year study previously revealed that in late summer, a subsurface peak in viral abundance recurs annually, and this chapter investigated whether that peak in abundance corresponded to recurring changes in composition of the viral community in the vicinity of the peak. Three different signature genes were examined, each targeting a different subset of marine viruses, and a community fingerprinting method was used to complement the signature gene analyses. Clustering analysis was then used to determine which samples were most similar. Together these techniques demonstrated that the viral communities at the surface and at 100 m depth were more similar to each other in winter (March), regardless of the year, than they were in summer (September), when the water column is stratified as opposed to well-mixed. In summer, the surface viral communities clustered together, while the 100 m viral community was less predictable and often seemed more similar to the March communities. These findings may stem from physical factors such as UV irradiation of viral particles, as well as seasonal and depth-related differences in host communities associated with the depth of the mixed layer.

Overall, this dissertation provides substantial advances to the field of microbial ecology. First, the development of phoH as a signature gene is an important addition to the limited set of tools available for studying marine viral diversity. Unlike other signature genes, phoH encompasses several morphological types of phage, and is found in phage that infect both autotrophic and heterotrophic bacteria. It is also widely distributed geographically and over depth profiles. Thus phoH has the potential to enable study of a broader range of viruses than other signature genes. The publication describing this work (Goldsmith et al., 2011) has already been cited 13 times as of April 2014. In addition, phoH has now been incorporated into the publicly available Metavir tool (Roux et al., 2011), to allow others to incorporate phoH as a signature gene in analyses of their metagenomic data.

This research also constitutes the first deep sequencing of a signature gene for marine viruses. This study revealed the great diversity of the marine viral community at BATS, over a multi-year time series and depth profile, and also showed that sequencing at such a deep level may not be necessary in order to fully capture the diversity of a the community at this site. These findings will guide other researchers as they determine how best to deploy their resources in studying marine viral diversity, and will also provide a benchmark for what level of viral diversity to expect in an oligotrophic marine system.

Finally, this study expands our knowledge of the viral community at BATS by examining the community based on four different measures of composition, rather than abundance. Studying composition provides a more detailed picture of the viral community, as it gives the first hints as to which viruses are present. Composition of the viral community is critical because of viruses' role in structuring bacterial communities; seasonal and depth-related changes in viruses correspondingly affect the bacteria on which they prey, and thus can cause potentially large changes in carbon and nutrient flow. These biogeochemical effects can ultimately have global consequences, such as altering the amount of carbon dioxide in Earth's atmosphere.

Not surprisingly, this research raises numerous questions, and suggests several paths for future examination. Although we know that the phoH gene occurs preferentially in marine phage, we do not know why, or what function this host-derived auxiliary metabolic gene serves in phage. The gene may represent an adaptation for viruses in oligotrophic environments, where phosphate is limited. In addition, although the viral phoH primers used in this research were designed from phage including the phage of a heterotrophic bacterium, few of the environmental sequences obtained in this research occurred in phylogenetic clusters with the phoH sequences from fully-sequenced heterotrophic phage. A redesign of the phoH primers to expand their scope and capture more of the marine viral community would provide greater insight into the prevalence and diversity of the gene in heterotrophic phage.

Further assessment of the seasonal differences in the viral communities at BATS, and of the significance of the annually recurring subsurface peak in viral abundance, could benefit from an attempt to sample the viral community at the precise depth of the peak in abundance. This would be possible if the viral communities are counted on board during the cruise, followed by sample collection from the depth revealed to have the highest abundance. In addition, the 16S ribosomal DNA of host bacterial communities from the same samples could be sequenced to determine correlations between the dominant host and viral types.

Another challenge for future work, stemming from the abundance of SAR11 at BATS, is to study the ecology of the viruses that infect this dominant clade. Now that four phage infecting SAR11 have recently been isolated and sequenced (Kang et al., 2013; Zhao et al., 2013), research into the lifestyle of these phages should expand. For single-stranded DNA phage, however, even more remains to be done. Identifying the hosts of the aquatic gokushoviruses (whose major capsid gene was investigated in this research) would enable us to learn more about the ecology of the viruses, and could potentially lead to culture-based experiments. Methods such as size fractionation of host communities, followed by PCR with the gokushovirus primers, single-cell sorting, and 16S rDNA PCR to establish bacterial identity, could prove promising.

Lastly, prospective research should continue to work toward culturing and sequencing marine viruses. The phylogenetic analyses conducted here demonstrate that many of the viruses sampled for this research belong to novel clusters, and that their sequences exhibit little similarity to the genomes of fully-sequenced viruses. The present state of the genomic databases contributes to the difficulties of studying viral diversity; however, as culturing and sequencing of environmental viruses proceed, our insights into the diversity of the most abundant biological entities on the planet will accordingly expand.

REFERENCES

Abascal, F., Zardoya, R., and Telford, M. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38**: W7-13.

Allen, M.J., Martinez-Martinez, J., Schroeder, D.C., Somerfield, P.J., and Wilson, W.H. (2007) Use of microarrays to assess viral diversity: from genotype to phenotype. *Environmental Microbiology* **9**: 971-982.

Angly, F., Felts, B., Breitbart, M., Salamon, P., Edwards, R., Carlson, C. et al. (2006) The marine viromes of four oceanic regions. *PLoS Biology* **4**: 2121-2131.

Araújo, M.F.F.d., and Godinho, M.J. (2009) Short-term variations of virus-like particles in a tropical lake: relationship with microbial communities (bacteria, ciliates and flagellates). *Microbiological research* **164**: 411-419.

Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983) The ecological role of water-column microbes in the sea. *Marine ecology progress series* **10**: 257-263.

Bergh, Ø., BØrsheim, K., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found in aquatic environments.

Berry, D., Mahfoudh, K.B., Wagner, M., and Loy, A. (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and environmental microbiology* **77**: 7846-7849.

Bidle, K.D., and Vardi, A. (2011) A chemical arms race at sea mediates algal host–virus interactions. *Current opinion in microbiology* **14**: 449-457.

Blotta, I., Prestinaci, F., Mirante, S., and Cantafora, A. (2005) Quantitative assay of total dsDNA with PicoGreen reagent and real-time fluorescent detection. *ANNALI-ISTITUTO SUPERIORE DI SANITA* **41**: 119.

Bouvier, T., and Del Giorgio, P. (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environmental microbiology* **9**: 287-297.

Bratbak, G., Thingstad, F., and Heldal, M. (1994) Viruses and the microbial loop. *Microbial Ecology* **28**: 209-221.

Breitbart, M. (2012) Marine viruses: truth or dare. *Marine Science* **4**.

Breitbart, M., and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* **13**: 278-284.

Buesseler, K.O., Lamborg, C., Cai, P., Escoube, R., Johnson, R., Pike, S. et al. (2008) Particle fluxes associated with mesoscale eddies in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **55**: 1426-1444.

Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *The ISME Journal* **3**: 283-295.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**: 540.

Cesar Ignacio-Espinoza, J., Solonenko, S.A., and Sullivan, M.B. (2013) The global virome: not as big as we thought? *Current opinion in virology* **3**: 566-571.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*: 265-270.

Chen, F., Wang, K., Huang, S., Cai, H., Zhao, M., Jiao, N., and Wommack, K.E. (2009) Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environmental Microbiology* **11**: 2884-2892.

Chenard, C., and Suttle, C. (2008) Phylogenetic diversity of sequences of cyanophage photosynthetic gene *psbA* in marine and freshwaters. *Applied and Environmental Microbiology* **74**: 5317-5324.

Chow, C.E.T., and Fuhrman, J.A. (2012) Seasonality and monthly dynamics of marine myovirus communities. *Environmental Microbiology* **14**: 2171-2183.

Clasen, J.L., Hanson, C.A., Ibrahim, Y., Weihe, C., Marston, M.F., and Martiny, J.B. (2013) Diversity and temporal dynamics of Southern California coastal marine cyanophage isolates. *Aquatic Microbial Ecology* **69**: 17-31.

Comeau, A.M., and Krisch, H.M. (2005) War is peace—dispatches from the bacterial and phage killing fields. *Current opinion in microbiology* **8**: 488-494.

Comeau, A.M., and Krisch, H.M. (2008) The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Molecular biology and evolution* **25**: 1321-1332.

Comeau, A.M., Short, S., and Suttle, C.A. (2004) The use of degenerate-primed random amplification of polymorphic DNA (DP-RAPD) for strain-typing and inferring the genetic similarity among closely related viruses. *Journal of virological methods* **118**: 95-100.

Comeau, A.M., Bertrand, C., Letarov, A., Tétart, F., and Krisch, H. (2007) Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**: 384-396.

Culley, A.I., and Steward, G.F. (2007) New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Applied and Environmental Microbiology* **73**: 5937-5944.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J.A., Middelburg, J.J., Noble, R.T., and Suttle, C.A. (2011) Marine viruses and global climate change. *FEMS microbiology reviews* **35**: 993-1034.

Doney, S.C. (1996) A synoptic atmospheric surface forcing data set and physical upper ocean model for the US JGOFS Bermuda Atlantic Time-Series Study site. *Journal of Geophysical Research* **101**: 25615.

Ducklow, H.W., Doney, S.C., and Steinberg, D.K. (2009) Contributions of long-term research and time-series observations to marine ecology and biogeochemistry. *Annual Review of Marine Science* **1**: 279-302.

DuRand, M., Olson, R., and Chisholm, S. (2001a) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 1983-2003.

DuRand, M.D., Olson, R.J., and Chisholm, S.W. (2001b) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 1983-2003.

Dwivedi, B., Schmieder, R., Goldsmith, D.B., Edwards, R.A., and Breitbart, M. (2012) PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. *BMC bioinformatics* **13**: 37.

Edgar, R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792.

Emerson, J.B., Thomas, B.C., Andrade, K., Heidelberg, K.B., and Banfield, J.F. (2013) New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. *Applied and environmental microbiology* **79**: 6755-6764.

Emerson, J.B., Thomas, B.C., Andrade, K., Allen, E.E., Heidelberg, K.B., and Banfield, J.F. (2012) Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Applied and environmental microbiology* **78**: 6309-6320.

Ewart, C., Meyers, M., Wallner, E., McGillicuddy Jr, D., and Carlson, C. (2008) Microbial dynamics in cyclonic and anticyclonic mode-water eddies in the northwestern Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **55**: 1334-1347.

Fandino, L.B., Riemann, L., Steward, G.F., Long, R.A., and Azam, F. (2001) Variations in bacterial community structure during a dinoflagellate bloom analyzed by DGGE and 16S rDNA sequencing. *Aquatic Microbial Ecology* **23**: 119-130.

Fenchel, T. (1988) Marine plankton food chains. *Annual Review of Ecology and Systematics*: 19-38.

Filée, J., Tétart, F., Suttle, C., and Krisch, H. (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**: 12471-12476.

Frederickson, C., Short, S., and Suttle, C. (2003) The physical environment affects cyanophage communities in British Columbia inlets. *Microbial ecology* **46**: 348-357.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Fuhrman, J.A., and Noble, R.T. (1995) Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnology and Oceanography* **40**: 1236-1242.

Fuhrman, J.A., Griffith, J.F., and Schwalbach, M.S. (2002) Prokaryotic and viral diversity patterns in marine plankton. *Ecological Research* **17**: 183-194.

Giovannoni, S., Tripp, H., Givan, S., Podar, M., Vergin, K., Baptista, D. et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242.

Giovannoni, S.J., and Vergin, K.L. (2012) Seasonality in ocean microbial communities. *Science* **335**: 671-676.

Goldsmith, D.B., Crosti, G., Dwivedi, B., McDaniel, L.D., Varsani, A., Suttle, C.A. et al. (2011) Development of phoH as a novel signature gene for assessing marine phage diversity. *Applied and environmental microbiology* **77**: 7730-7739.

Hammer, Ø., Harper, D., and Ryan, P. (2001) PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**: 9.

Hennig, C. (2007) Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* **52**: 258-271.

Hennig, C. (2008) Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of multivariate analysis* **99**: 1154-1176.

Hennig, C. (2013) fpc: Flexible procedures for clustering. *R package version*: 2.1-6.

Hewson, I., Winget, D.M., Williamson, K.E., Fuhrman, J.A., and Wommack, K.E. (2006) Viral and bacterial assemblage covariance in oligotrophic waters of the West Florida Shelf (Gulf of Mexico). *JMBA-Journal of the Marine Biological Association of the United Kingdom* **86**: 591-604.

Hodcroft, E. (2013). TreeCollapserCL 4. URL http://emmahodcroft.com/TreeCollapseCL.html

Hoffmann, K.H., Rodriguez-Brito, B., Breitbart, M., Bangor, D., Angly, F., Felts, B. et al. (2007) Power law rank–abundance models for marine phage communities. *FEMS microbiology letters* **273**: 224-228.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., VerBerkmoes, N.C., and Sullivan, M.B. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proceedings of the National Academy of Sciences* **110**: 12798-12803.

Hopkins, M., Kailasan, S., Cohen, A., Roux, S., Tucker, K., Shevenell, A. et al. (2014) Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *The ISME Journal* **8**.

Hsieh, Y.-J., and Wanner, B.L. (2010) Global regulation by the seven-component Pi signaling system. *Current Opinion in Microbiology* **13**: 198-203.

Huang, S., Wilhelm, S.W., Jiao, N., and Chen, F. (2010) Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences. *The ISME Journal* **4**: 1243-1251.

Hurwitz, B.L., and Sullivan, M.B. (2013) The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PloS one* **8**: e57355.

Jameson, E., Mann, N.H., Joint, I., Sambles, C., and Mühling, M. (2011) The diversity of cyanomyovirus populations along a North–South Atlantic Ocean transect. *The ISME journal* **5**: 1713-1721.

Jamindar, S., Polson, S.W., Srinivasiah, S., Waidner, L., and Wommack, K.E. (2012) Evaluation of two approaches for assessing the genetic similarity of virioplankton populations as defined by genome size. *Applied and environmental microbiology* **78**: 8773-8783.

Jiang, S., Fu, W., Chu, W., and Fuhrman, J. (2003) The vertical distribution and diversity of marine bacteriophage at a station off Southern California. *Microbial ecology* **45**: 399-410.

Kang, I., Oh, H.-M., Kang, D., and Cho, J.-C. (2013) Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proceedings of the National Academy of Sciences* **110**: 12343-12348.

Karl, D.M., and Lukas, R. (1996) The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* **43**: 129-156.

Keitt, T. (2012) colorRamps: Builds color tables. *R package version*: 2.3.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D. et al. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463-1464.

Koonin, E., and Rudd, K. (1996) Two domains of superfamily I helicases may exist as separate proteins. *Protein Science* **5**: 178-180.

Labonté, J.M., and Suttle, C.A. (2013) Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**: 2169-2177.

Larsen, A., Castberg, T., Sandaa, R., Brussaard, C., Egge, J., Heldal, M. et al. (2001) Population dynamics and diversity of phytoplankton, bacteria and viruses in a seawater enclosure. *Marine Ecology Progress Series* **221**: 47-57.

Letarov, A., Manival, X., Desplats, C., and Krisch, H. (2005) gpwac of the T4-type bacteriophages: structure, function, and evolution of a segmented coiled-coil protein that controls viral infectivity. *Journal of bacteriology* **187**: 1055-1066.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 11013-11018.

Lipschultz, F., Bates, N., Carlson, C., and Hansell, D. (2002) New production in the Sargasso Sea: History and current status. *Global Biogeochem Cycles* **16**: 1001.

Lomas, M., Bates, N., Johnson, R., Knap, A., Steinberg, D., and Carlson, C. (2013) Two decades and counting: 24-years of sustained open ocean biogeochemical measurements in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **93**: 16-32.

Lysholm, F. (2012) Highly improved homopolymer aware nucleotide-protein alignments with 454 data. *BMC bioinformatics* **13**: 230.

Magiopoulos, I., and Pitta, P. (2012) Viruses in a deep oligotrophic sea: Seasonal distribution of marine viruses in the epi-, meso-and bathypelagic waters of the Eastern Mediterranean Sea. *Deep Sea Research Part I: Oceanographic Research Papers* **66**: 1-10.

Magurran, A.E. (2004) Measuring biological diversity.

Makino, K., Amemura, M., Kim, S., Yokoyama, K., and Kimura, S. (1998) Mechanism of transcriptional activation of the phosphate regulon in *Escherichia coli*. *The Journal of Microbiology*: 231-238.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010) Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *The ISME journal* **4**: 1252-1264.

Marie, D., Brussaard, C.P.D., Thyrhaug, R., Bratbak, G., and Vaulot, D. (1999) Enumeration of marine viruses in culture and natural samples by flow cytometry. *Applied and Environmental Microbiology* **65**: 45.

Marston, M.F., and Sallee, J.L. (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters. *Applied and Environmental Microbiology* **69**: 4639-4647.

Marston, M.F., Taylor, S., Sme, N., Parsons, R.J., Noyes, T.J., and Martiny, J.B. (2013) Marine cyanophages exhibit local and regional biogeography. *Environmental microbiology* **15**: 1452-1463.

Marston, M.F., Pierciey, F.J., Shepard, A., Gearin, G., Qi, J., Yandava, C. et al. (2012) Rapid diversification of coevolving marine Synechococcus and a virus. *Proceedings of the National Academy of Sciences* **109**: 4544-4549.

Matteson, A.R., Rowe, J.M., Ponsero, A.J., Pimentel, T.M., Boyd, P.W., and Wilhelm, S.W. (2013) High abundances of cyanomyoviruses in marine ecosystems demonstrate ecological relevance. *FEMS microbiology ecology* **84**: 223-234.

Michaels, A., and Knap, A. (1996) Overview of the US JGOFS Bermuda Atlantic Time-series Study and the Hydrostation S program. *Deep Sea Research Part II: Topical Studies in Oceanography* **43**: 157-198.

Michaels, A.F., Knap, A.H., Dow, R.L., Gundersen, K., Johnson, R.J., Sorensen, J. et al. (1994) Seasonal patterns of ocean biogeochemistry at the US JGOFS Bermuda Atlantic Time-series Study site. *Deep Sea Research Part I: Oceanographic Research Papers* **41**: 1013-1038.

Middelboe, M., Holmfeldt, K., Riemann, L., Nybroe, O., and Haaber, J. (2009) Bacteriophages drive strain diversification in a marine Flavobacterium: implications for phage resistance and physiological properties. *Environmental microbiology* **11**: 1971-1982.

Mizoguchi, K., Morita, M., Fischer, C.R., Yoichi, M., Tanji, Y., and Unno, H. (2003) Coevolution of bacteriophage PP01 and Escherichia coli O157: H7 in continuous culture. *Applied and Environmental Microbiology* **69**: 170-176.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Claverie, J.-M., and Ogata, H. (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome research* **19**: 1441-1449.

Morris, R., Vergin, K., Cho, J., Rappé, M., Carlson, C., and Giovannoni, S. (2005) Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnology and Oceanography* **50**: 1687-1696.

Morris, R.M., Rappé, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806-810.

Mourino-Carballido, B. (2009) Eddy-driven pulses of respiration in the Sargasso Sea. *Deep Sea Research Part I: Oceanographic Research Papers* **56**: 1242-1250.

Mühling, M., Fuller, N., Millard, A., Somerfield, P., Marie, D., Wilson, W. et al. (2005) Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environmental Microbiology* **7**: 499-508.

Needham, D.M., Chow, C.-E.T., Cram, J.A., Sachdeva, R., Parada, A., and Fuhrman, J.A. (2013) Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *The ISME journal*.

Neuwirth, E. (2011) RColorBrewer: ColorBrewer palettes. *R package version*: 1.0-5.

Noble, R.T., and Fuhrman, J.A. (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology* **14**: 113-118.

Pagarete, A., Chow, C.-E., Johannessen, T., Fuhrman, J., Thingstad, T., and Sandaa, R. (2013) Strong Seasonality and Interannual Recurrence in Marine Myovirus Communities. *Applied and environmental microbiology* **79**: 6253-6259.

Paradis, E., Claude, J., and Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289-290.

Parsons, R.J., Breitbart, M., Lomas, M.W., and Carlson, C.A. (2012) Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *The ISME journal* **6**: 273-284.

Pomeroy, L.R. (1974) The ocean's food web, a changing paradigm. *Bioscience* **24**: 499-504.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.

Proctor, L.M. (1997) Advances in the study of marine viruses. *Microscopy research and technique* **37**: 136-161.

R Development Core Team (2013) R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Raymond, P.A., Bauer, J.E., and Cole, J.J. (2000) Atmospheric CO2 evasion, dissolved inorganic carbon production, and net heterotrophy in the York River estuary. *Limnology and Oceanography*: 1707-1717.

Ricklefs, R.E., and Lovette, I.J. (1999) The roles of island area per se and habitat diversity in the species–area relationships of four Lesser Antillean faunal groups. *Journal of Animal Ecology* **68**: 1142-1160.

Riemann, L., and Middelboe, M. (2002) Stability of bacterial and viral community compositions in Danish coastal waters as depicted by DNA fingerprinting techniques. *Aquatic Microbial Ecology* **27**: 219-232.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M. et al. (2010) Viral and microbial community dynamics in four aquatic environments. *The ISME journal* **4**: 739-751.

Rohwer, F. (2003) Global phage diversity. *Cell* **113**: 141.

Rohwer, F., and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology* **184**: 4529-4535.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., and Enault, F. (2011) Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**: 3074-3075.

Rozon, R., and Short, S. (2013) Complex seasonality observed amongst diverse phytoplankton viruses in the Bay of Quinte, an embayment of Lake Ontario. *Freshwater Biology* **58**: 2648-2663.

Sambrook, J., Fritsch, E., and Maniatis, T. (1989) *Molecular cloning: a laboratory manual. Vol. 1*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Sandaa, R.-A., and Larsen, A. (2006) Seasonal variations in virus-host populations in Norwegian coastal waters: focusing on the cyanophage community infecting marine Synechococcus spp. *Applied and environmental microbiology* **72**: 4610-4618.

Sarkar, D. (2008) *Lattice: multivariate data visualization with R*. New York: Springer.

Schliep, K.P. (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**: 592-593.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**: 7537-7541.

Schwalbach, M., Hewson, I., and Fuhrman, J. (2004) Viral effects on bacterial community composition in marine plankton microcosms. *Aquatic Microbial Ecology* **34**: 117-127.

Short, S., and Suttle, C. (1999) Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. In *Molecular Ecology of Aquatic Communities*: Springer, pp. 19-32.

Short, S.M., and Short, C.M. (2009) Quantitative PCR reveals transient and persistent algal viruses in Lake Ontario, Canada. *Environmental microbiology* **11**: 2639-2648.

Siegel, D.A., McGillicuddy, D.J., and Fields, E.A. (1999) Mesoscale eddies, satellite altimetry, and new production in the Sargasso Sea. *Journal of Geophysical Research* **104**: 359.

Simpson, E.H. (1949) Measurement of diversity. *Nature*.

Srinivasiah, S., Lovett, J., Polson, S., Bhavsar, J., Ghosh, D., Roy, K. et al. (2013) Direct Assessment of Viral Diversity in Soils by Random PCR Amplification of Polymorphic DNA. *Applied and environmental microbiology* **79**: 5450-5457.

Steinberg, D.K., Carlson, C.A., Bates, N.R., Johnson, R.J., Michaels, A.F., and Knap, A.H. (2001) Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 1405-1447.

Stern, A., and Sorek, R. (2011) The phage-host arms race: Shaping the evolution of microbes. *Bioessays* **33**: 43-51.

Steward, G.F., Montiel, J.L., and Azam, F. (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnology and Oceanography* **45**: 1697-1706.

Sullivan, M., Huang, K., Ignacio Espinoza, J., Berlin, A., Kelly, L., Weigele, P. et al. (2010a) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4 like myoviruses from diverse hosts and environments. *Environmental Microbiology* **12**: 3035-3056.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424**: 1047-1051.

Sullivan, M.B., Huang, K.H., Ignacio Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R. et al. (2010b) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4 like myoviruses from diverse hosts and environments. *Environmental Microbiology*.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suttle, C.A. (2007) Marine viruses—major players in the global ecosystem. *Nature reviews microbiology* **5**: 801-812.

Sweeney, E.N., and McGillicuddy, D.J. (2003) Biogeochemical impacts due to mesoscale eddy activity in the Sargasso Sea as measured at the Bermuda Atlantic Time-series Study (BATS). *Deep Sea Research Part II: Topical Studies in Oceanography* **50**: 3017-3039.

Talavera, G., and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**: 564.

Thingstad, T., and Lignell, R. (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13**: 19-27.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009) Laboratory procedures to generate viral metagenomes. *Nature protocols* **4**: 470-483.

Treusch, A.H., Vergin, K.L., Finlay, L.A., Donatz, M.G., Burton, R.M., Carlson, C.A., and Giovannoni, S.J. (2009) Seasonality and vertical structure of microbial communities in an ocean gyre. *The ISME Journal* **3**: 1148-1163.

Tucker, K.P., Parsons, R., Symonds, E.M., and Breitbart, M. (2011) Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *The ISME journal* **5**: 822-830.

Vavrek, M.J. (2011) fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica* **14**: 16.

Vergin, K.L., Beszteri, B., Monier, A., Thrash, J.C., Temperton, B., Treusch, A.H. et al. (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *The ISME journal* **7**: 1322-1332.

Wang, G., Asakawa, S., and Kimura, M. (2011) Spatial and temporal changes of cyanophage communities in paddy field soils as revealed by the capsid assembly protein gene g20. *Fems Microbiology Ecology*.

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

Wanner, B. (1996) Phosphorus assimilation and control of the phosphate regulon. In *Escherichia coli and Salmonella: cellular and molecular biology*. Washington, D.C.: ASM Press, pp. 1357–1381.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A. et al. (2009) gplots: Various R programming tools for plotting data. *R package version* **2.12.1**.

Weinbauer, M., and Suttle, C. (1999) Lysogeny and prophage induction in coastal and offshore bacterial communities. *Aquatic Microbial Ecology* **18**: 217-225.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 6578.

Wilhelm, S.W., and Suttle, C.A. (1999) Viruses and nutrient cycles in the sea. *Bioscience* **49**: 781-788.

Wilson, W.H., Fuller, N., Joint, I., and Mann, N. (1999) Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. *Bulletin de l'Institut océanographique*: 209-216.

Winget, D., and Wommack, K. (2008) Randomly amplified polymorphic DNA PCR as a tool for assessment of marine viral richness. *Applied and Environmental Microbiology* **74**: 2612.

Winter, C., and Weinbauer, M.G. (2010) Randomly Amplified Polymorphic DNA Reveals Tight Links between Viruses and Microbes in the Bathypelagic Zone of the Northwestern Mediterranean Sea. *Applied and Environmental Microbiology* **76**: 6724.

Winter, C., Kerros, M.-E., and Weinbauer, M.G. (2009) Seasonal and depth-related dynamics of prokaryotes and viruses in surface and deep waters of the northwestern Mediterranean Sea. *Deep Sea Research Part I: Oceanographic Research Papers* **56**: 1972-1982.

Winter, C., Smit, A., Herndl, G., and Weinbauer, M. (2004) Impact of virioplankton on archaeal and bacterial community richness as assessed in seawater batch cultures. *Applied and Environmental Microbiology* **70**: 804.

Winter, C., Bouvier, T., Weinbauer, M.G., and Thingstad, T.F. (2010) Trade-Offs between Competition and Defense Specialists among Unicellular Planktonic Organisms: the "Killing the Winner" Hypothesis Revisited. *Microbiology and Molecular Biology Reviews* **74**: 42-57.

Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiology and molecular biology reviews* **64**: 69-114.

Wommack, K.E., Hill, R.T., Muller, T.A., and Colwell, R.R. (1996) Effects of sunlight on bacteriophage viability and structure. *Applied and environmental microbiology* **62**: 1336-1341.

Wommack, K.E., Ravel, J., Hill, R.T., Chun, J., and Colwell, R.R. (1999) Population dynamics of Chesapeake Bay virioplankton: total-community analysis by pulsed-field gel electrophoresis. *Applied and Environmental Microbiology* **65**: 231.

Yu, Y., Breitbart, M., McNairnie, P., and Rohwer, F. (2006) FastGroupII: A web-based bioinformatics platform for analyses of large 16 S rDNA libraries. *BMC Bioinformatics* **7**: 57.

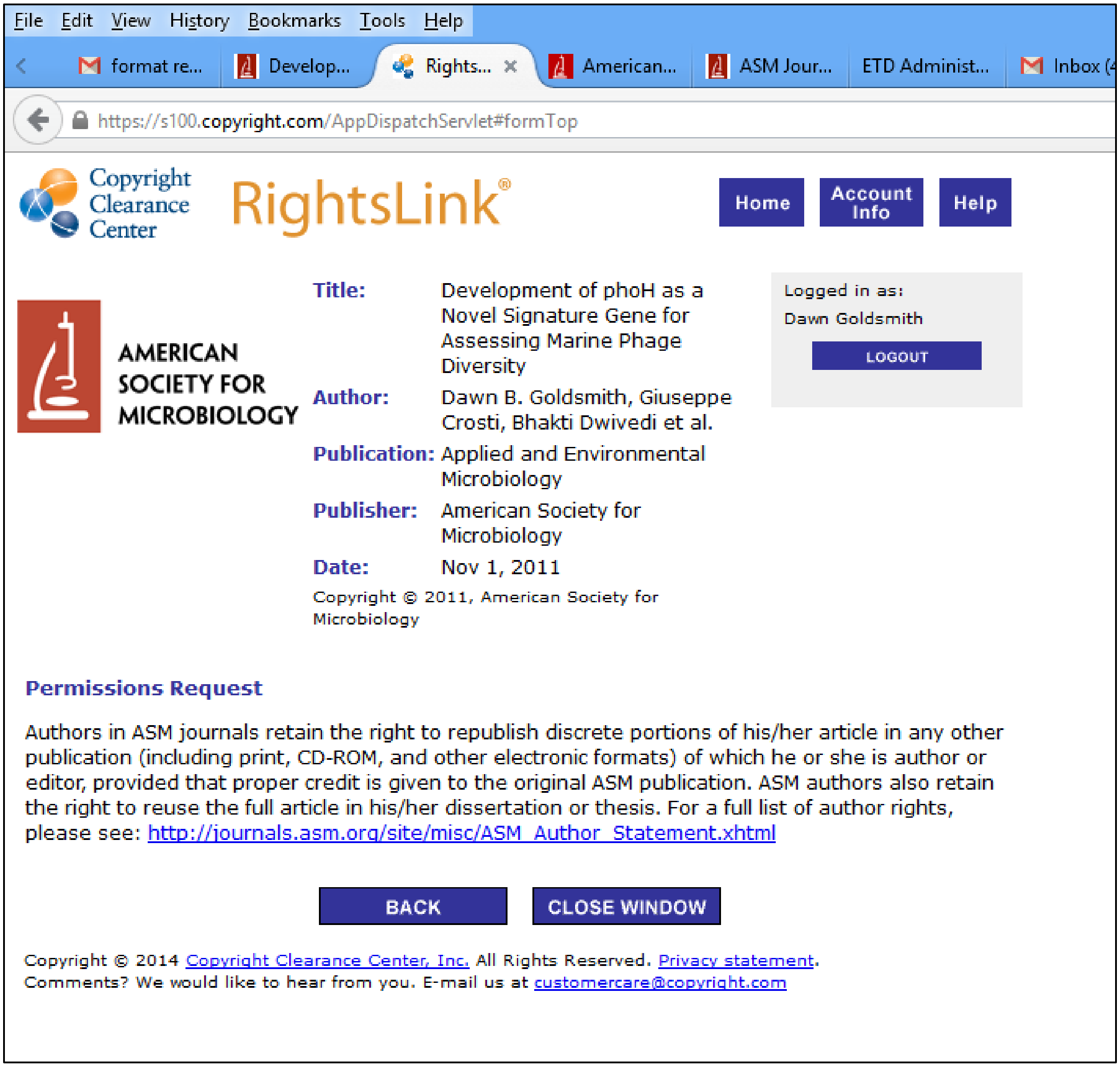
Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C. et al. (2013) Abundant SAR11 viruses in the ocean. *Nature* **494**: 357-360.

Zhong, Y., Chen, F., Wilhelm, S., Poorvin, L., and Hodson, R. (2002) Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Applied and Environmental Microbiology* **68**: 1576-1584.

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006) Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Applied and Environmental Microbiology* **72**: 723.

APPENDIX A

Copyright Permission


File Edit View History Bookmarks Tools Help
format re... Develop... Rights... American... ASM Jour... ETD Administ... Inbox (4
https://s100.copyright.com/AppDispatchServlet#formTop
Copyright Clearance Center
RightsLink®
Home
Account Info
Help
AMERICAN SOCIETY FOR MICROBIOLOGY

Title:	Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity
Author:	Dawn B. Goldsmith, Giuseppe Crosti, Bhakti Dwivedi et al.
Publication:	Applied and Environmental Microbiology
Publisher:	American Society for Microbiology
Date:	Nov 1, 2011

Copyright © 2011, American Society for Microbiology

Logged in as:
Dawn Goldsmith
LOGOUT

Permissions Request

Authors in ASM journals retain the right to republish discrete portions of his/her article in any other publication (including print, CD-ROM, and other electronic formats) of which he or she is author or editor, provided that proper credit is given to the original ASM publication. ASM authors also retain the right to reuse the full article in his/her dissertation or thesis. For a full list of author rights, please see: http://journals.asm.org/site/misc/ASM_Author_Statement.xhtml

BACK CLOSE WINDOW

Copyright © 2014 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

APPENDIX B

Development of phoH as a novel signature gene for assessing marine phage diversity

Applied and Environmental
Microbiology

Development of *phoH* as a Novel Signature Gene for Assessing Marine Phage Diversity

Dawn B. Goldsmith, Giuseppe Crosti, Bhakti Dwivedi,
Lauren D. McDaniel, Arvind Varsani, Curtis A. Suttle,
Markus G. Weinbauer, Ruth-Anne Sandaa and Mya Breitbart

Appl. Environ. Microbiol. 2011, 77(21):7730. DOI:
10.1128/AEM.05531-11.
Published Ahead of Print 16 September 2011.

Applied and Environmental Microbiology, Nov. 2011, p. 7730–7739 Vol. 77, No. 21
0099-2240/11/$12.00 doi:10.1128/AEM.05531-11

Development of *phoH* as a Novel Signature Gene for Assessing Marine Phage Diversity[▽]

Dawn B. Goldsmith,[1] Giuseppe Crosti,[1] Bhakti Dwivedi,[1] Lauren D. McDaniel,[1] Arvind Varsani,[2,3,4]
Curtis A. Suttle,[5] Markus G. Weinbauer,[6] Ruth-Anne Sandaa,[7] and Mya Breitbart[1]*

College of Marine Science, University of South Florida, St. Petersburg, Florida[1]; School of Biological Sciences, University of Canterbury, Christchurch, New Zealand[2]; Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand[3]; Department of Molecular and Cell Biology, University of Cape Town, Cape Town, South Africa[4]; Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, Canada[5]; Laboratoire d'Océanographie de Villefranche, Université Pierre et Marie Curie, and CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France[6]; and Department of Biology, University of Bergen, Bergen, Norway[7]

Received 27 May 2011/Accepted 3 September 2011

Phages play a key role in the marine environment by regulating the transfer of energy between trophic levels and influencing global carbon and nutrient cycles. The diversity of marine phage communities remains difficult to characterize because of the lack of a signature gene common to all phages. Recent studies have demonstrated the presence of host-derived auxiliary metabolic genes in phage genomes, such as those belonging to the Pho regulon, which regulates phosphate uptake and metabolism under low-phosphate conditions. Among the completely sequenced phage genomes in GenBank, this study identified Pho regulon genes in nearly 40% of the marine phage genomes, while only 4% of nonmarine phage genomes contained these genes. While several Pho regulon genes were identified, *phoH* was the most prevalent, appearing in 42 out of 602 completely sequenced phage genomes. Phylogenetic analysis demonstrated that phage *phoH* sequences formed a cluster distinct from those of their bacterial hosts. PCR primers designed to amplify a region of the *phoH* gene were used to determine the diversity of phage *phoH* sequences throughout a depth profile in the Sargasso Sea and at six locations worldwide. *phoH* was present at all sites examined, and a high diversity of *phoH* sequences was recovered. Most *phoH* sequences belonged to clusters without any cultured representatives. Each depth and geographic location had a distinct *phoH* composition, although most *phoH* clusters were recovered from multiple sites. Overall, *phoH* is an effective signature gene for examining phage diversity in the marine environment.

Marine viruses merit study not only because of their sheer abundance but also because of the critical roles they play in the Earth's biogeochemical cycles (11). The majority of these viruses are phages (viruses that infect bacteria). Because phages are host-specific predators that influence the composition of the bacterial community (9, 47), it is essential to understand the diversity of marine phages. Microscopy-based methods have only limited resolution for analyzing marine phage diversity, and therefore genetic methods are preferable. However, identification of phages in environmental samples is hampered by the lack of a single gene found in all phages (50). Nonetheless, some genes are shared within groups of phages, and these "signature genes" can be used as markers to examine the diversity of a phage group of interest (70). Several signature genes have been developed to examine the diversity of phages in the marine environment, including structural genes (61, 64, 86), replication genes (10, 33), and auxiliary metabolic genes (14, 54, 60, 68, 80).

Auxiliary metabolic genes (AMGs) are phage-borne metabolic genes that were typically thought to be restricted to cellular genomes yet have been identified in phage genomes through sequencing (11). Numerous AMGs involved in photosynthesis, carbon metabolism, and nucleotide metabolism have been identified in marine phages (14, 35, 36, 42, 43, 65, 68, 78, 80). In addition, marine phages carry AMGs involved in nutrient limitation (51, 65, 67, 78), such as those belonging to the Pho regulon, which regulates phosphate uptake and metabolism under low-phosphate conditions (24, 77). Here we examined the presence of genes belonging to the Pho regulon in completely sequenced phage genomes and demonstrated the utility of *phoH* as a new signature gene for the study of marine phage diversity. Newly described PCR primers were used to amplify *phoH* from viral samples collected throughout the world's oceans. A high diversity of *phoH* genes was found in marine viral communities, with the types of *phoH* identified varying with depth and location.

MATERIALS AND METHODS

Prevalence of Pho regulon genes in phages. To determine the presence of Pho regulon genes in completely sequenced phage genomes, a pool of bacterial Pho regulon genes was collected from three bacterial strains. First, the nucleotide sequences of the 35 genes of the Pho regulon (*amn*, *eda*, *phnCDEFGHIJKLMNOP*, *phoABEHRU*, *psiEF*, *pstABCS*, *ugpABCEQ*, *yibD*, and *ytfK*) (24) from *Escherichia coli* strain K-12 (substrain MG1655; accession number U00096) were retrieved from GenBank. Next, potential Pho regulon genes from *Prochlorococcus marinus* strain NATL1A (accession number NC_008819) were collected by using the 35 *E. coli* Pho regulon genes as the query in a TBLASTX (3) search

* Corresponding author. Mailing address: College of Marine Science, University of South Florida, 140 7th Ave. South, St. Petersburg, FL 33701. Phone: (727) 553-3520. Fax: (727) 553-1189. E-mail: mya@marine.usf.edu.

[▽] Published ahead of print on 16 September 2011.

against the genome of NATL1A. Twelve of the 35 queries produced hits with E values of <0.001. Those hits in the genome of NATL1A (genes annotated as *eda*, *phoB*, *phoH*, *phoR*, *pstA*, *pstB*, *pstC*, *salX*, and *potA* and genes with the locus tags NATL1_02681, NATL1_11521, and NATL1_07881) were added to the Pho regulon genes from *E. coli*. One additional NATL1A gene, locus tag NATL1_20941, was included because although it was the second-best hit (when *E. coli*'s *phnL* was used as the query), it is annotated as a phosphate transporter in the NATL1A genome. Finally, genes from *P. marinus* strain NATL2A (accession number CP000095) that were predicted to be part of that cyanobacterium's Pho regulon (63) were included in the pool. This step added 20 genes, with the locus tags PMN2A_0440, PMN2A_0439, PMN2A_0438, PMN2A_0249, PMN2A_0435, PMN2A_0436, PMN2A_0437, PMN2A_0549, PMN2A_0496, PMN2A_0959, PMN2A_0559, PMN2A_0742, PMN2A_1499, PMN2A_1369, PMN2A_0714, PMN2A_0311, PMN2A_0310, PMN2A_0309, PMN2A_0308, and PMN2A_0307. This combined pool of bacterial Pho regulon genes from *E. coli* and *P. marinus* contained 68 sequences. To identify Pho regulon genes in phage genomes, each sequence was compared by BLASTX (3) against the GenBank nonredundant (nr) database (using default parameters), limiting the subject organisms to viruses (taxonomy identification no. [taxid] 10239). All significant hits (E value < 0.001) were confirmed through reciprocal BLASTP analysis against the GenBank nr database.

Collection and processing of depth profile samples. To examine the difference in *phoH* composition of the phage community present at different depths, small-scale samples were collected from throughout a depth profile (0, 200, 500, and 1,000 m) at the Bermuda Atlantic Time-series Study site (31°40'N, 64°10'W) in September 2008. Whole seawater samples (100 ml) were filtered through a 0.22-μm Sterivex filter (Millipore, Billerica, MA) and then onto a 0.02-μm Anotop filter (Whatman, Piscataway, NJ). Anotop filters were stored at −80°C until DNA was extracted with a MasterPure complete DNA and RNA purification kit (Epicentre Biotechnologies, Madison, WI) following the protocol of Culley and Steward (17). Briefly, filters were defrosted, and all liquid was purged from the filter by pushing air through with a sterile syringe. A flame-sealed pipette tip was used to temporarily seal the filter outlet, and a mixture of 400 μl of 2× T&C lysis buffer (from the MasterPure kit) and 50 μg proteinase K was forced onto the filter. The filter was then incubated for 10 min in the air at 65°C before the lysate was expelled into a microcentrifuge tube and immediately placed on ice. Then 150 μl of MPC protein precipitation reagent (from the MasterPure kit) was added to the lysate and vortexed vigorously for 10 s. The debris was pelleted by centrifugation at 10,000 × *g* for 10 min. Isopropanol was added to the recovered supernatant, and the tube was inverted 30 to 40 times. The DNA was then pelleted by centrifugation at 20,000 × *g* at 4°C for 10 min and washed twice with 75% ethanol. Extracted DNA was resuspended in sterile water and stored at −20°C.

Collection and processing of geographic samples. To examine the biogeography of phage *phoH* sequences, samples were collected from the Sargasso Sea, British Columbia coastal waters, the Gulf of Mexico, Raunefjorden, Kongsfjorden, and the Mediterranean Sea. Large-scale samples (approximately 250 liters) from 0 m and 100 m from the Sargasso Sea (31°40'N, 64°10'W) were concentrated by tangential flow filtration with 100-kDa filters (GE Healthcare, Piscataway, NJ) to a volume of approximately 50 ml. These viral concentrates were filtered through 0.22-μm Sterivex filters to remove bacteria and stored at 4°C until further processing. Viruses were further concentrated and purified from the Sargasso Sea concentrates by polyethylene glycol precipitation followed by cesium chloride density-dependent centrifugation. Solid polyethylene glycol 8000 (PEG 8000) was added to the concentrates at a final concentration of 10% (wt/vol), and the concentrates were stored at 4°C overnight. The concentrates were then centrifuged for 40 min at 11,000 × *g* and 4°C to pellet the viruses. The pelleted viruses were resuspended in 0.02-μm-filtered seawater and further purified through ultracentrifugation in a cesium chloride density gradient with layers of 1.2 g/ml, 1.5 g/ml, and 1.7 g/ml (22,000 rpm on a Beckman SW40 Ti rotor for 3 h at 4°C). The viral fractions were further concentrated with a Microcon centrifugal filter device (Millipore), and viral DNA was extracted using the formamide method as described by Sambrook et al. (53). The Raunefjorden (60°16.2'N, 5°12.5'E) and Kongsfjorden (79°00'N, 11°40'E) samples were prefiltered through 0.45-μm-pore-size low-protein-binding Durapore membrane filters 142 mm in diameter (Millipore) in order to remove cellular organisms. The filtrate was then concentrated to approximately 45 ml using a QuixStand benchtop system with 100-kDa hollow fiber cartridges (GE Healthcare Bio-Sciences AB, Uppsala, Sweden). The samples from the Gulf of Mexico (pool of 41 samples collected between 1994 and 2001 from the surface to 164 m) and British Columbia coastal waters (pool of 85 samples collected between 1996 and 2004 from the surface to 245 m) were collected as described by Angly et al. (4) and processed as outlined by Suttle et al. (71). Briefly, the samples were prefiltered through 142-mm-diameter glass fiber filters with a 1.2-μm pore size (Advantec MFS, Dublin, CA) or a 0.7-μm pore size (Whatman, Clifton, NJ), followed by filtration through 0.45-μm or 0.2-μm-pore-size Durapore membrane filters (Millipore, Bedford, MA). Concentration of virus-sized particles from the filtrate was completed with 10-kDa or 30-kDa spiral-wound cartridges (Amicon/Millipore, Billerica, MA). Concentrates were stored in the dark at 4°C until further processing. Mediterranean samples (43°41'N, 7°19'E) were collected and concentrated as described by Bonilla-Findji et al. (8). The samples were prefiltered through 0.8-μm polycarbonate filters (142-mm diameter) (Osmonics, Inc., Minnetonka, MN), followed by tangential flow filtration through 0.2-μm Durapore polycarbonate filters (Millipore) and concentration on 100-kDa spiral polyethersulfone cartridges (Millipore). For all locations except the Sargasso Sea, viral DNA was obtained by incubating 500 μl of viral concentrate at 90°C twice for 2 min, placing the concentrate on ice between incubations. Then, 20 μl of 0.5 M EDTA (pH 8.0) and 5 μl of freshly made proteinase K (10 mg/ml) were added, and the mixture was incubated for 10 min at 55°C. After the addition of 25 μl of 10% sodium dodecyl sulfate, the mixture was further incubated for 1 h at 55°C. The DNA was cleaned with a DNA Clean and Concentrator kit (Zymo Research Corp., Irvine, CA) following the manufacturer's instructions and resuspended in 20 μl of sterile water.

Primer design and DNA amplification. *phoH* primers were designed based on a CLUSTALX (73) alignment of the full-length *phoH* gene from *Synechococcus* phage S-PM2, *Prochlorococcus* phages P-SSM2 and P-SSM4, and *Vibrio* phage KVP40. PCR primers vPhoHf (5'-TGCRGGWACAGGTAARACAT-3') and vPhoHr (5'-TCRCCRCAGAAAAYMATTTT-3') were used to amplify a product of approximately 420 bp. The 50-μl reaction mixture for PCR amplification of the *phoH* gene contained 1 U Apex *Taq* DNA polymerase (Genesee Scientific, San Diego, CA), 1× Apex *Taq* reaction buffer, 1.5 mM Apex MgCl$_2$, a 0.5 μM concentration of each primer, 0.2 mM deoxynucleoside triphosphates, and 0.04% bovine serum albumin. The reaction conditions were (i) 5 min of initial denaturation at 95°C; (ii) 35 cycles of 1 min of denaturation (95°C), 1 min of annealing (53°C), and 1 min of extension (72°C); and (iii) 10 min of final extension at 72°C. Before amplification of the *phoH* gene, DNA from the Sargasso Sea samples was amplified by the strand displacement method of the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare, Piscataway, NJ) according to the manufacturer's instructions.

Cloning and sequencing. *phoH* PCR products were cloned into vectors and used to transform competent cells. After screening, the inserts in positive transformants were sequenced. PCR products from the Sargasso Sea were cloned using the TOPO TA cloning kit for sequencing (Invitrogen, Carlsbad, CA) and were sequenced by Beckman Coulter Genomics (Danvers, MA). PCR products from the remaining samples were cloned with the StrataClone PCR cloning kit (Stratagene, La Jolla, CA) and sequenced by LGC Genomics (Berlin, Germany). PCR products from the cyanophage isolates were directly sequenced (without cloning) by the University of Florida (Gainesville, FL).

Phylogenetic analysis. Vector and low-quality sequences were trimmed with Sequencher 4.7 (Gene Codes, Ann Arbor, MI). The Sargasso Sea samples were dereplicated using FastGroup II at a level of 99% sequence identity with gaps (84). Reference sequences from cultured phages were obtained from GenBank and through amplification of *phoH* from cyanophages isolated from the Gulf of Mexico on *Synechococcus* WH7803 (41). All sequences were aligned at the amino acid level using CLUSTALW (using default parameters) as implemented in TranslatorX (1). The amino acid alignment (see Fig. 1) or back-translated nucleotide alignments (see Fig. 2 and 4) were then used to build maximum-likelihood phylogenetic trees with PhyML 3.0 (21). Protein-coding sequences such as *phoH* are more conserved at the amino acid level than they are at the nucleotide level (1), and thus alignments are more accurate when conducted at the amino acid level. The back-translated nucleotide sequences obtained from the amino acid alignments were used to build the trees in order to better reflect the diversity of the *phoH* sequences in the environment. Nonparametric branch supports were determined by an approximate likelihood ratio test (5). Nodes with branch support values of ≤50 were collapsed using Mesquite (version 2.74) (37). Phylogenetic trees were edited with MEGA 5 (31).

Nucleotide sequence accession numbers. The nucleotide sequences reported in this paper have been submitted to GenBank and assigned accession numbers JF963974 through JF964262.

RESULTS AND DISCUSSION

Pho regulon genes in phages. The Pho regulon contains a group of genes whose products control the uptake and metabolism of phosphate by the cell in response to phosphate limi-

TABLE 1. Genes of the Pho regulon found in the genomes of fully sequenced phages

Phage	Presence of:					Host	Host trophic status	Marine origin
	phoH	*pstS*	*phoA*	*phoE* (*nmpC*)	*ugpQ*			
SPO1	X					*Bacillus subtilis*	Heterotroph	
CP220	X					*Campylobacter*	Heterotroph	
D-1873	X					*Clostridium botulinum*	Heterotroph	
phiW-14	X					*Delftia acidovorans*	Heterotroph	
P7				X		Enterobacteria	Heterotroph	
RB43	X					Enterobacteria	Heterotroph	
phiEco32	X					*Escherichia coli*	Heterotroph	
RB16	X					*Escherichia coli*	Heterotroph	
rv5	X					*Escherichia coli*	Heterotroph	
T5	X					*Escherichia coli*	Heterotroph	
KP15	X					*Klebsiella pneumoniae*	Heterotroph	
949	X					*Lactococcus lactis*	Heterotroph	
Ma-LMM01	X					*Microcystis aeruginosa*	Autotroph	
P-HM1	X					*Prochlorococcus*	Autotroph	X
P-HM2	X					*Prochlorococcus*	Autotroph	X
P-RSM4	X	X				*Prochlorococcus*	Autotroph	X
P-SSM2	X	X				*Prochlorococcus*	Autotroph	X
P-SSM4	X	X				*Prochlorococcus*	Autotroph	X
P-SSM7	X	X				*Prochlorococcus*	Autotroph	X
PA11	X					*Pseudomonas aeruginosa*	Heterotroph	
SIO1	X					*Roseobacter* SIO67	Heterotroph	X
SPC35	X					*Salmonella enterica* and *Escherichia coli*	Heterotroph	
EPS7	X					*Salmonella enterica* serovar Typhimurium	Heterotroph	
Vi01	X					*Salmonella enterica* serovar Typhi Vi	Heterotroph	
phiSboM-AG3	X					*Shigella boydii*	Heterotroph	
A5W	X				X	*Staphylococcus aureus*	Heterotroph	
G1	X				X	*Staphylococcus aureus*	Heterotroph	
K	X				X	*Staphylococcus aureus*	Heterotroph	
Twort	X					*Staphylococcus aureus*	Heterotroph	
S-CRM01	X					*Synechococcus*	Autotroph	
S-PM2	X					*Synechococcus*	Autotroph	X
S-RSM4	X					*Synechococcus*	Autotroph	X
S-SM1	X	X	X			*Synechococcus*	Autotroph	X
S-SM2	X	X	X			*Synechococcus*	Autotroph	X
S-SSM5	X	X				*Synechococcus*	Autotroph	X
S-SSM7	X	X				*Synechococcus*	Autotroph	X
Syn1	X					*Synechococcus*	Autotroph	X
S-ShM2	X					*Synechococcus* and *Prochlorococcus*	Autotroph	X
Syn19	X	X				*Synechococcus* and *Prochlorococcus*	Autotroph	X
Syn33	X					*Synechococcus* and *Prochlorococcus*	Autotroph	X
Syn9	X					*Synechococcus* and *Prochlorococcus*	Autotroph	X
ICP1	X					*Vibrio cholerae*	Heterotroph	
KVP40	X					*Vibrio parahaemolyticus*	Heterotroph	X

tation (24, 77). Phosphorus is essential for cell survival due to its presence in membrane lipids and nucleic acids, as well as its roles in posttranslational protein modification and energy transfer (6, 79). In *E. coli*, expression of the Pho regulon is activated when phosphate is limited (77). There is direct evidence that at least 31 genes are part of the Pho regulon, and indirect evidence of several more (24).

Genes involved in phosphate limitation (i.e., *phoH*, *pstS*, and *phoA*) have been previously identified in the genomes of marine phages (15, 39, 43, 44, 51, 65–67, 69, 78), as well as in marine metagenomes (52, 58, 60, 80). To determine the prevalence of these and other Pho regulon genes in phage genomes, BLAST similarity searches (3) were performed using Pho regulon genes from the genomes of *E. coli* strain K-12 substrain MG1655, *P. marinus* strain NATL1A, and *P. marinus* strain NATL2A against the virus subset of the nr database. Of the 35 Pho regulon genes examined, only five (*phoH*, *pstS*, *phoA*, *phoE*, and *ugpQ*) were found in phage genomes (Table 1). *phoH* was the gene most commonly found in phages, occurring in 42 of the 602 completely sequenced phage genomes in the GenBank database (as of 26 May 2011). A phosphate transporter subunit gene, *pstS*, occurred in nine phages whose genomes are completely sequenced (66). These relative frequencies support prior analyses of the Global Ocean Sampling (GOS) metagenome showing that scaffolds containing *phoH* genes included a much higher percentage of viral open readings frames than scaffolds containing *pstS* genes (80). *phoA*, a gene of the Pho regulon that encodes bacterial alkaline phosphatase (24, 77), was found in two fully sequenced phages, located next to *pstS* in the genomes (S-SM1 and S-SM2 [66]). The metagenomic GOS data revealed that uncultured cyanophages contained *phoA* as well (26). A phage that contained neither *phoH* nor *pstS* nonetheless possessed a different Pho regulon gene; the enterobacterial phage P7 contained *nmpC*, a gene encoding an outer membrane porin precursor homologous to porins of the *phoE* family (77). In addition, three

Staphylococcus phages (G1, K, and A5W) contained *ugpQ*, which encodes a glycerophosphoryl diester phosphodiesterase (74). Interestingly, the genomes of marine phages appeared to be enriched in Pho regulon genes compared to the genomes of phages from other environments. Forty-four percent of the phage genomes containing Pho regulon genes were isolated from the marine environment (19 out of 43), while marine phages comprised only a small proportion (8%) of the 602 completely sequenced phage genomes in GenBank. Among the completely sequenced phage genomes in GenBank, nearly 40% of the marine phages contained Pho regulon genes, while only 4% of nonmarine phage genomes contained these genes. Thus, the data from this study show that it is not equally likely for sequenced marine and nonmarine phages to contain Pho regulon genes, although this result could be biased by the representation of phage genomes in GenBank. These data support previous assertions that it may be advantageous for marine phages to encode genes involved in phosphate regulation because phosphate is often a limiting nutrient in the oceans (26, 36, 51, 65, 66).

Given that *phoH* is much more abundant in phage genomes than any of the other Pho regulon genes, it is possible that PhoH in phages serves a role unrelated to phosphate uptake. The study that first identified and characterized the *phoH* gene noted that PhoH could bind ATP, and that it was probably a cytoplasmic protein involved in the uptake of phosphate under conditions of phosphate starvation (29, 38). However, despite the well-studied nature of the Pho regulon and the presence of *phoH* genes in a wide array of phage genomes, the function of PhoH remains unknown. The gene product is likely an ATPase, given its conserved nucleoside triphosphate hydrolase domain (24, 30, 38, 66). If PhoH hydrolyzes ATP, the resulting reaction would release energy to drive another reaction, presumably to assist in the uptake of phosphate by the cell. Kazakov et al. (27) examined homologs of the *phoH* genes from *E. coli* and *Bacillus subtilis*; they found that the homologs clustered into three groups. The positions of the homologs and their presence in two different clusters of orthologous groups suggested several different potential functions for PhoH, including fatty acid beta oxidation, phospholipid metabolism, and metal-dependent RNA modification (27).

Not only does the role of PhoH remain unclear, but the expression of *phoH* under conditions of phosphate stress has also been shown to vary among species. In *E. coli*, *phoH* is upregulated when the cell is subjected to phosphate stress (66, 77). Similarly, levels of *phoH* mRNA transcripts in *Corynebacterium glutamicum* are 4.6 times higher when phosphate is limited than when the cell has sufficient phosphate (25). In contrast, *phoH* is downregulated in *Synechococcus* sp. WH8102 under conditions of phosphate limitation (72). In two strains of *Prochlorococcus* (MED4 and MIT9313), there is no change in the expression of *phoH* when phosphate is limited (40). The only study to examine the link between phage infection and *phoH* expression demonstrated an increase in the *phoH* transcript level in *Prochlorococcus* MED4 upon infection with phage P-SSP7 (34). At 4 h postinfection, *phoH* is upregulated by a factor of 1.8. It has been hypothesized that the increased expression of the gene could represent a response by the host to the stress of phage infection (34). Upregulation of phosphate-uptake genes, whether carried by the host or by the phage, may work to the advantage of the phage, since phosphorus is a key limiting nutrient in the marine environment. Thus, the existence of Pho regulon genes in phage genomes may constitute a selective advantage to the phage, enabling phosphate uptake during infection and allowing further phage replication despite phosphate limitation (34, 36, 55, 65, 79).

***phoH* as a signature gene for phage identification.** Several signature genes are currently being used to study phage diversity, but each of these marker genes has limitations. For example, primers available for amplifying the DNA polymerase gene of T7-like podophages are restricted to only a subset of that phage group (10, 33). Structural genes such as *g20*, which encodes a portal protein (14, 64, 86), and *g23*, which encodes a major capsid protein (20), are also commonly used as genetic markers in phages. However, the available primers for these genes are restricted to myophages, with the *g20* primers specifically targeting cyanomyophages (20, 47, 86). Although primers for genes homologous to *psbA* and *psbD* (encoding photosystem II reaction center proteins D1 and D2) have proven useful for phage identification (14, 35, 36, 68), the ability of the *psb* primers to characterize phage diversity is limited to cyanophages.

The presence of *phoH* in phages that infect both heterotrophic and autotrophic hosts suggests that it could potentially capture a broad range of phages and therefore be used to analyze phage diversity. *phoH* genes have been found in many phages infecting autotrophic bacteria (Table 1), such as the cyanophages P-SSM2 and P-SSM4, which infect *Prochlorococcus* (65), cyanophage Syn9, which infects *Synechococcus* (78), and cyanophage Ma-LMM01, which infects *Microcystis aeruginosa* (82). In addition, *phoH* genes have been detected in a range of phages infecting heterotrophic bacteria, such as roseophage SIO1, a phage of *Roseobacter* (51), PA11, a phage of *Pseudomonas aeruginosa* (32), and KVP40, a broad-host-range vibriophage (44). Another advantage of *phoH* as a signature gene for examining phage diversity is that this gene is not restricted to one morphological type of phage. The *phoH* gene has been found in the genomes of podophages, such as the enterobacterial phage phiEco32 (57), in siphophages, such as enterobacterial phage EPS7 (23) and enterobacterial phage T5 (76), and also in myophages, such as *Bacillus* phage SPO1 (62). Among heterotrophic marine phages, *phoH* has been detected in both podophages (such as *Roseobacter* phage SIO1 [51]) and myophages (such as vibriophage KVP40 [44]); however, among sequenced cyanophages, *phoH* has so far been identified only in myophages (65–67). Finally, *phoH* genes are not restricted to phages and have also been detected in viruses that infect autotrophic eukaryotes. For example, several viruses of unicellular photosynthetic marine green algae of the *Ostreococcus* genus, as well as viruses infecting *Micromonas* and *Bathycoccus*, have been shown to contain *phoH* (18, 45, 79).

phoH has been found in phages and viruses isolated from a wide variety of geographic areas, including the coast of Japan (KVP40 [44]), coastal lagoons in the northwestern Mediterranean Sea (OlV1 and MpV1 [45]), the coast of southern California (SIO1 [51]), the Red Sea (S-RSM4 [43]), the Sargasso Sea (P-SSM4 [69]), the English Channel (S-PM2 [39, 43, 81]), the Pacific Ocean near Hawaii (P-HM1 and P-HM2 [66]), and the coast of Massachusetts by Woods Hole (Syn9 [43, 78]).

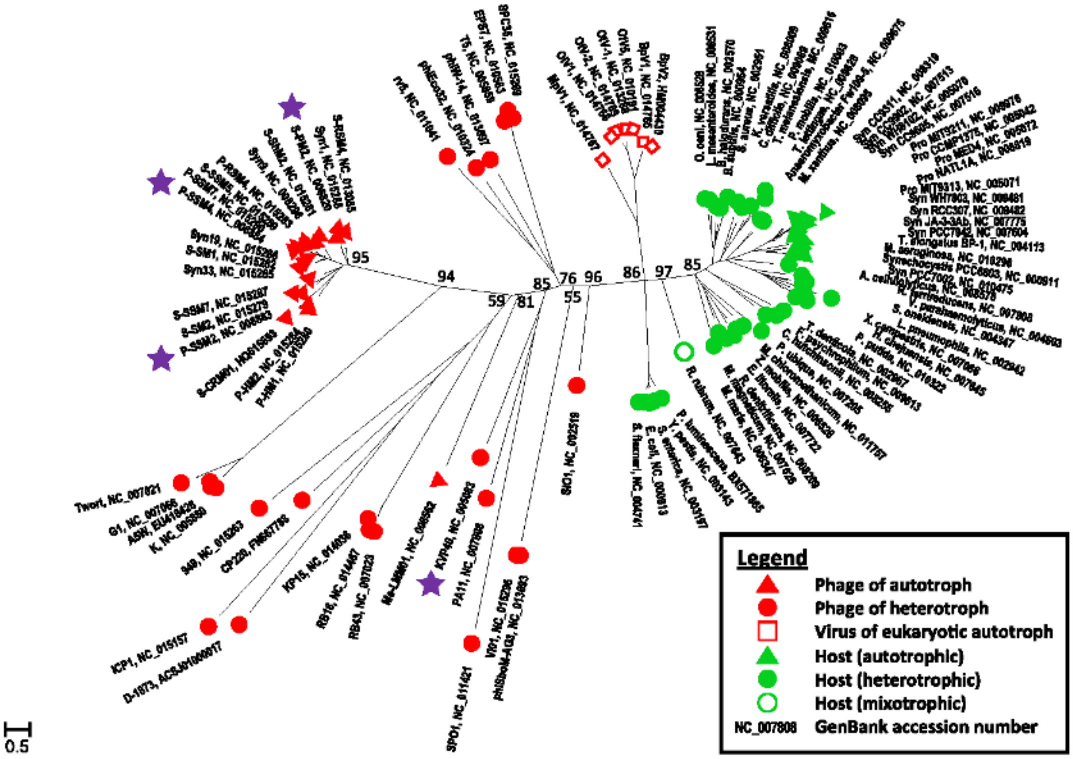


FIG. 1. Phylogenetic tree (from an amino acid alignment) showing the relationship among the *phoH* genes of completely sequenced bacteria, phages, and eukaryotic viruses. The scale bar shows substitutions per site. *phoH* primers were designed from sequences marked with a star.

While most of the cultured phages containing the *phoH* gene originated from marine waters, some were isolated from other habitats. For example, SPO1 was isolated from soil in Japan (62); EPS7 was isolated from Korean sewage samples (23); phiEco32 was found in a river in Tbilisi, Georgia (57); Ma-LMM01 was isolated from a lake in Japan (83); Vi01 came from human stool samples from Canadian patients with typhoid fever (49); and phage KP15 was obtained from sewage samples from Warsaw, Poland (48).

Comparison of phage and host *phoH*. Comparison of other AMGs in phages and the hosts they infect has demonstrated that many of these genes are evolving differently from their host counterparts. Phylogenetic analysis reveals that phage and host versions of the photosynthesis gene *psbA* tend to cluster separately, though not completely (14, 54). However, the phage genes group next to the genes from their hosts: *psbA* from phages that infect *Synechococcus* form a sister clade to *Synechococcus psbA* genes, and *psbA* genes from phages that infect *Prochlorococcus* form a sister clade to *Prochlorococcus psbA* genes (22, 36, 68, 80, 85). A similar pattern exists for *psbD* genes, which are involved in photosynthesis (36, 54, 68, 80), and PTOX genes (encoding plastoquinol terminal oxidase) (43). Recent research reveals that phage-borne PSI genes are also evolving separately from the host versions of those genes (2, 59). Finally, analysis of GOS data shows that NAD(P)H dehydrogenase genes in phages mainly cluster separately from bacterial versions (59), and *mazG* genes from cyanophages cluster separately from host *Prochlorococcus* and *Synechococcus* versions of the gene (12). These phylogenetic patterns suggest that after the host genes have been incorporated into phage genomes, the selective pressure on those genes changes in such a way that it becomes possible to distinguish between host and phage versions.

Phylogenetic analysis of the *phoH* gene from the genomes of fully sequenced phages and bacteria revealed that phages clustered separately from hosts (Fig. 1), thereby demonstrating that *phoH* can be used as a signature gene to discriminate between host and phages when phage diversity is being investigated. Within that primary division, there was further resolution by trophic strategy. Cyanobacteria formed their own well-supported clade, while the heterotrophic bacteria formed several separate *phoH* clusters. Similarly, phages clustered according to the nutrition mode of their hosts; there was a well-supported clade of cyanophages, while the heterotrophic phages fell into other groups. *phoH* of phages infecting heterotrophs displayed a greater diversity than *phoH* of those

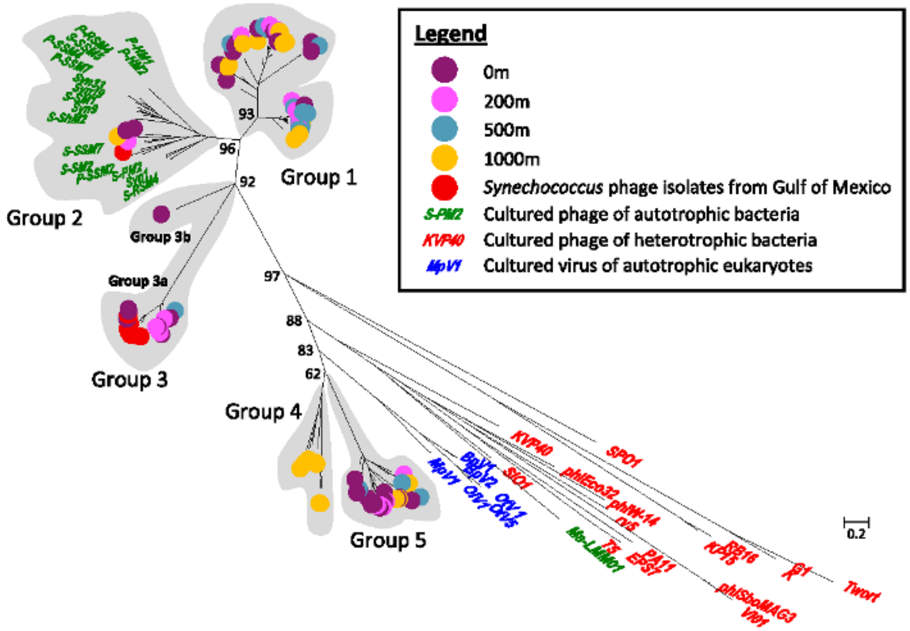


FIG. 2. Phylogenetic tree (from a nucleotide alignment) showing the relationship among *phoH* sequences from environmental virus samples throughout a depth profile in the Sargasso Sea and *phoH* sequences from cultured phages and viruses. Group classifications for environmental sequences are indicated. The scale bar shows substitutions per site.

infecting autotrophs. Viruses of eukaryotes also formed their own well-supported cluster.

Phage *phoH* diversity throughout the water column in the Sargasso Sea. The diversity of *phoH* throughout a depth profile at the Bermuda Atlantic Time-series Study site in the Sargasso Sea was examined to determine whether distinct phage types were present throughout the water column. A significant diversity of *phoH* sequences was identified along the depth profile containing samples from 0, 200, 500, and 1,000 m. The depth profile *phoH* sequences formed five distinct clusters, identified as groups 1 through 5, with the majority of the Sargasso Sea sequences belonging to clusters without any cultured isolates (Fig. 2). This is similar to the situation observed for several other signature genes in the marine environment (10, 20, 33) and demonstrates that environmental phages are not well represented by the phage isolates currently available in culture. Since many of the environmental *phoH* groups do not contain cultured isolates, it is possible that some of the sequences obtained in this study are not viral in origin. However, several steps were taken during sample processing to ensure removal of host DNA, including filtration of all samples and density-dependent centrifugation of some samples. Phylogenetic trees containing environmental *phoH* sequences alongside cultured phages, viruses, and hosts revealed that none of the environmental sequences clustered with those of hosts (data not shown), which is not surprising given that the primers were designed specifically based on phage sequences. Nonetheless, although it is extremely likely that the environmental *phoH* sequences are viral in origin, the possibility that the samples contain host-derived DNA, such as that contained within gene transfer agents or transducing particles, cannot be excluded.

The *phoH* composition of the phage community varied throughout the water column (Fig. 3). Changes in the composition of the phage community are likely driven by differences in the composition of the host community, which has been studied in great detail at this site (13, 19, 46, 75). All depths were dominated by sequences belonging to group 1, which did not contain any cultured representatives. Group 2, which contained most of the *phoH* sequences of cultured cyanophages that have been fully sequenced, comprised only a minor component of the sequences recovered at any depth. Each depth contained sequences from multiple groups, and the proportion of sequences represented by each group varied among depths (Fig. 3). For instance, although group 1 sequences were found at all four depths, over 80% of the 500-m sequences belonged to group 1, while just over 40% of the 0-m and 1,000-m sequences belonged to group 1. Group 3 sequences were more abundant in the photic zone, decreasing with depth and not detected in the 1,000-m sample. In contrast, group 4 comprised 35% of the *phoH* sequences recovered from 1,000 m and was not detected at any of the other depths. It is not surprising that the 1,000-m sequences were distinct from those of the other depths, because the 1,000-m phage community would not be expected to contain the cyanophages that populate the photic zone.

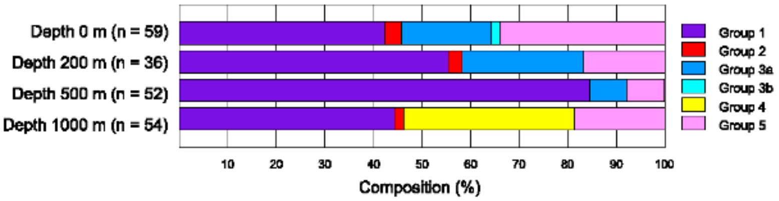


FIG. 3. Composition of *phoH* sequences found at each depth in the Sargasso Sea, based on the groups defined in the phylogenetic tree in Fig. 2.

Biogeography of phage *phoH* sequences. In addition to the depth profile of the *phoH* gene, the biogeography of *phoH* was studied in viral concentrates from six locations around the world (the Gulf of Mexico, the Arctic Ocean, British Columbia coastal waters, the Mediterranean Sea, the Sargasso Sea, and a site near the coast of Norway). Along with the five groups identified in the depth profile phylogenetic tree, the global study also revealed a sixth cluster, group 6, which did not appear in the sequences from the Sargasso Sea (Fig. 4). *phoH* composition differed for the phage community from each location, with no single group found at all sites (Fig. 5). Different *phoH* groups dominated at different locations. For example, group 1 represented over 80% of the sequences from Raunefjorden but only approximately 10% of the sequences from Kongsfjorden. Group 3 represented less than 20% of the sequences from Kongsfjorden but nearly 80% of the British Columbia sequences and was not present at all in the Raunefjorden sequences. Group 5 was also absent from the Raunefjorden profile and varied from approximately 5% in Kongsfjorden to over 40% in the Sargasso Sea. While the Raunefjorden, Kongsfjorden, and Mediterranean samples were all drawn from the surface, samples from the other three locations were pooled from the surface down to 100 m (Sargasso Sea), 164 m (Gulf of Mexico), and 245 m (British Columbia). Given that the depth profile drawn from the Sargasso Sea (Fig. 3) showed that each depth exhibited a distinct *phoH* composition, further work is required in order to better resolve biogeographical differences in *phoH* sequences.

In light of these different profiles, it is apparent that the *phoH* gene can distinguish phage communities from different locations and serve as a useful biogeographical marker. However, this analysis also points out gaps in our knowledge. It is somewhat surprising that group 2, which contained almost all of the completely sequenced cyanophage isolates in GenBank, was found at only two of the studied locations: the Sargasso Sea and the Gulf of Mexico. In contrast, the cultured cya-

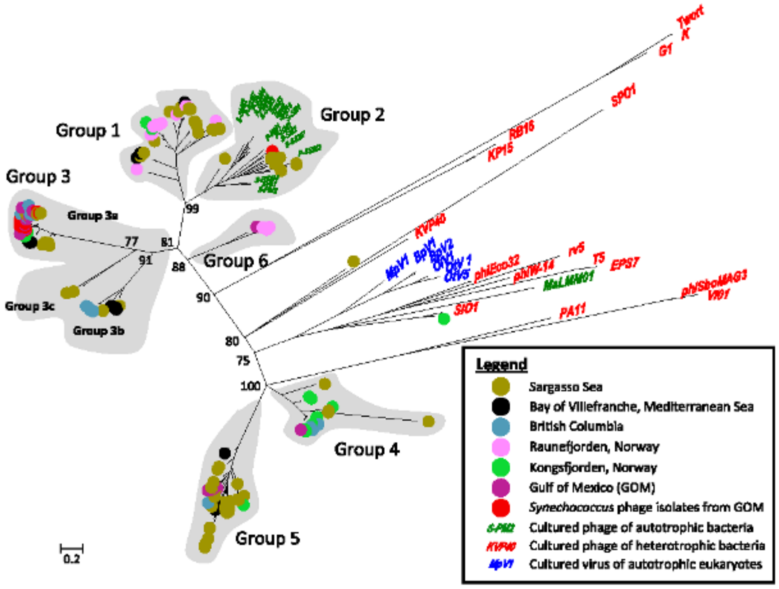


FIG. 4. Phylogenetic tree (from a nucleotide alignment) showing the biogeography of *phoH* sequences from environmental virus samples from six locations. *phoH* sequences from cultured phages and viruses are also shown. Group classifications for environmental sequences are indicated; groups 1 through 5 are the same as groups 1 through 5 in Fig. 2. The scale bar shows substitutions per site.

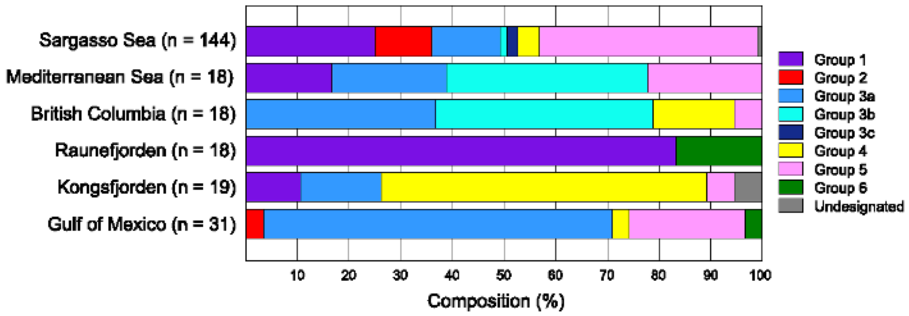


FIG. 5. Composition of *phoH* sequences detected at each location from the biogeographical survey, based on the groups defined in the phylogenetic tree in Fig. 4.

nophages from the Gulf of Mexico whose *phoH* genes were sequenced in this study belonged almost entirely to group 3a, which was represented in every studied location except one. Consistent with group 3a originating from cyanophages, in the Sargasso Sea depth profile, group 3a was most abundant in the photic zone and was not present at 1,000 m. This supports the idea that the sequenced cyanophages currently in the database do not adequately represent total cyanophage diversity and that more cyanophages need to be cultured and sequenced. In addition, although a great deal of *phoH* diversity was elucidated using these primers, it is notable that only two of the 248 environmental *phoH* sequences in the global study appeared in the group with the cultured heterotrophic phages. This suggests that the cultured heterotrophic phages are not well represented in the marine environment, or that the *phoH* primers used in this study do not amplify the *phoH* gene of many of the cultured heterotrophic phages. Designing additional *phoH* primers to capture more of the cultured heterotrophic phages, as well as the viruses infecting photosynthetic eukaryotes, will enable a broader analysis of *phoH* diversity in the marine viral community. Since many of the major *phoH* groups identified in the environmental samples did not contain cultured representatives, it is unknown whether these groups consisted of cyanophages or heterotrophic phages. As additional phage-host systems are isolated, insight into the identity of the phages in the *phoH* environmental clusters will be gained.

Despite the different *phoH* compositions identified at the disparate locations, most of the *phoH* groups were found at multiple sites. This supports previous research suggesting that phages are not limited by geography. Sano et al. (56) examined phages from four different environments (soil, marine sediment, freshwater, and seawater) and discovered that soil, freshwater, and sediment phages can propagate on hosts from the marine environment. That study also showed that marine phages from one location can infect hosts from a different marine location (56). Signature gene studies using both DNA polymerase and structural genes have detected identical phage sequences from widely separated geographical locations, as well as from different habitats (10, 33, 61). Studies of phages infecting *Vibrio* species also demonstrated that genetically related vibriophages can be found throughout the water column, as well as in marine locations separated by up to 4,500 miles (16, 28). Metagenomic sequencing of viral communities from throughout the world's oceans confirmed that the majority of the viral genotypes are shared between locations, with differences between sites being driven by changes in the relative abundance of specific viruses (4). A more recent analysis of the GOS expedition supported these results, finding differential distribution of myophages, podophages, and siphophages by location while further establishing that many AMGs occur in phages worldwide (80). These studies, in combination with the *phoH* data presented here, support the idea that "everything is everywhere, but the environment selects" (7) and suggest that the selection may be driven not only by the composition of the host community but also by auxiliary metabolic genes present in the phage genomes.

ACKNOWLEDGMENTS

Thanks go to Hilde Kristiansen, Rachel Parsons, Craig Carlson, and many past and present members of the Suttle lab for collecting and processing the virus communities, as well as the crews of the R/V Atlantic Explorer, R/V Longhorn and CCGS Vector for logistical support.

This research was funded by grants from the National Science Foundation to MB (Microbial Interactions and Processes MCB-0701984 and Division of Biological Infrastructure DBI-0850206). The University of Bergen received financial support from the Norwegian Research Council for research program "Viral lysis and programmed cell death in marine phytoplankton" (VIPMAP, 186142/V40) and the project ERC Advanced Grant "MIcrobial Network OrganiSation" (MINOS). Collection of the virus communities from British Columbia and the Gulf of Mexico was supported by grants to C.A.S. from the Natural Science and Engineering Council of Canada, the National Science Foundation and the Office of Naval Research. D.B.G. was supported by a Presidential Doctoral Fellowship from the University of South Florida and the Von Rosenstiel Endowed Fellowship. A.V. was supported by an Erskine Fellowship (University of Canterbury, New Zealand).

REFERENCES

1. **Abascal, F., R. Zardoya, and M. J. Telford.** 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. **38:**W7–W13.
2. **Alperovitch-Lavy, A., et al.** 2011. Reconstructing a puzzle: existence of cyanophages containing both photosystem-I and photosystem-II gene suites inferred from oceanic metagenomic datasets. Environ. Microbiol. **13:**24–32.
3. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.

4. **Angly, F. E., et al.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:**2121–2131.
5. **Anisimova, M., and O. Gascuel.** 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. **55:**539–552.
6. **Baek, J. H., and S. Y. Lee.** 2006. Novel gene members in the Pho regulon of *Escherichia coli*. FEMS Microbiol. Lett. **264:**104–109.
7. **Becking, L. G. M. B.** 1934. Geobiologie of inleiding tot de milieukunde, vol. 18. WP Van Stockum & Zoon NV, The Hague, The Netherlands.
8. **Bonilla-Findji, O., et al.** 2008. Viral effects on bacterial respiration, production and growth efficiency: consistent trends in the Southern Ocean and the Mediterranean Sea. Deep Sea Res. Part 2 Top. Stud. Oceanogr. **55:**790–800.
9. **Bratbak, G., F. Thingstad, and M. Heldal.** 1994. Viruses and the microbial loop. Microb. Ecol. **28:**209–221.
10. **Breitbart, M., J. H. Miyake, and F. Rohwer.** 2004. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol. Lett. **236:**249–256.
11. **Breitbart, M., L. R. Thompson, C. A. Suttle, and M. B. Sullivan.** 2007. Exploring the vast diversity of marine viruses. Oceanography (Wash. D C) **20:**135–139.
12. **Bryan, M. J., et al.** 2008. Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. PLoS One **3:**17–34.
13. **Carlson, C. A., et al.** 2009. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. **3:**283–295.
14. **Chénard, C., and C. A. Suttle.** 2008. Phylogenetic diversity of sequences of cyanophage photosynthetic gene *psbA* in marine and freshwaters. Appl. Environ. Microbiol. **74:**5317–5324.
15. **Coleman, M. L., et al.** 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. Science **311:**1768–1770.
16. **Comeau, A. M., A. M. Chan, and C. A. Suttle.** 2006. Genetic richness of vibriophages isolated in a coastal environment. Environ. Microbiol. **8:**1164–1176.
17. **Culley, A. I., and G. F. Steward.** 2007. New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. Appl. Environ. Microbiol. **73:**5937–5944.
18. **Derelle, E., et al.** 2008. Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. PLoS One **3:**e2250.
19. **DuRand, M. D., R. J. Olson, and S. W. Chisholm.** 2001. Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. Deep Sea Res. Part 2 Top. Stud. Oceanogr. **48:**1983–2003.
20. **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. Proc. Natl. Acad. Sci. U. S. A. **102:**12471–12476.
21. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.
22. **Hambly, E., and C. A. Suttle.** 2005. The viriosphere, diversity, and genetic exchange within phage communities. Curr. Opin. Microbiol. **8:**444–450.
23. **Hong, J., et al.** 2008. Identification of host receptor and receptor binding module of a newly sequenced T5-like phage EPS7. FEMS Microbiol. Lett. **289:**202–209.
24. **Hsieh, Y.-J., and B. L. Wanner.** 2010. Global regulation by the seven-component P_i signaling system. Curr. Opin. Microbiol. **13:**198–203.
25. **Ishige, T., M. Krause, M. Bott, V. F. Wendisch, and H. Sahm.** 2003. The phosphate starvation stimulon of *Corynebacterium glutamicum* determined by DNA microarray analyses. J. Bacteriol. **185:**4519–4529.
26. **Kathuria, S., and A. C. Martiny.** 2011. Prevalence of a calcium-based alkaline phosphatase associated with the marine cyanobacterium *Prochlorococcus* and other ocean bacteria. Environ. Microbiol. **13:**74–83.
27. **Kazakov, A. E., O. Vassieva, M. S. Gelfand, A. Osterman, and R. Overbeek.** 2003. Bioinformatics classification and functional analysis of PhoH homologs. In Silico Biol. **3:**3–15.
28. **Kellogg, C. A., J. B. Rose, S. C. Jiang, J. M. Thurmond, and J. H. Paul.** 1995. Genetic diversity of related vibriophages isolated from marine environments around Florida and Hawaii, U. S. A. Mar. Ecol. Prog. Ser. **120:**89–98.
29. **Kim, S.-K., K. Makino, M. Amemura, H. Shinagawa, and A. Nakata.** 1993. Molecular analysis of the *phoH* gene, belonging to the phosphate regulon in *Escherichia coli*. J. Bacteriol. **175:**1316–1324.
30. **Koonin, E. V., and K. E. Rudd.** 1996. Two domains of superfamily I helicases may exist as separate proteins. Protein Sci. **5:**178–180.
31. **Kumar, S., M. Nei, J. Dudley, and K. Tamura.** 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. **9:**299–306.
32. **Kwan, T., J. Liu, M. DuBow, P. Gros, and J. Pelletier.** 2006. Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. J. Bacteriol. **188:**1184–1187.
33. **Labonté, J. M., K. E. Reid, and C. A. Suttle.** 2009. Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. Appl. Environ. Microbiol. **75:**3634–3640.
34. **Lindell, D., et al.** 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature **449:**83–86.
35. **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:**86–89.
36. **Lindell, D., et al.** 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. U. S. A. **101:**11013–11018.
37. **Maddison, W. P., and D. R. Maddison.** 2010. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org.
38. **Makino, K., M. Amemura, S.-K. Kim, K. Yokoyama, and S. Kimura.** 1998. Mechanism of transcriptional activation of the phosphate regulon in *Escherichia coli*. J. Microbiol. **36:**231–238.
39. **Mann, N. H., et al.** 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. J. Bacteriol. **187:**3188–3200.
40. **Martiny, A. C., M. L. Coleman, and S. W. Chisholm.** 2006. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. Proc. Natl. Acad. Sci. U. S. A. **103:**12552–12557.
41. **McDaniel, L. D., M. Delarosa, and J. H. Paul.** 2006. Temperate and lytic cyanophages from the Gulf of Mexico. J. Mar. Biol. Assoc. **86:**517–527.
42. **Millard, A. D., M. R. J. Clokie, D. A. Shub, and N. H. Mann.** 2004. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. Proc. Natl. Acad. Sci. U. S. A. **101:**11007–11012.
43. **Millard, A. D., K. Zwirglmaier, M. J. Downey, N. H. Mann, and D. J. Scanlan.** 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ. Microbiol. **11:**2370–2387.
44. **Miller, E. S., et al.** 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. J. Bacteriol. **185:**5220–5233.
45. **Moreau, H., et al.** 2010. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J. Virol. **84:**12555–12563.
46. **Morris, R. M., et al.** 2005. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. Limnol. Oceanogr. **50:**1687–1696.
47. **Mühling, M., et al.** 2005. Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. Environ. Microbiol. **7:**499–508.
48. **Petrov, V. M., S. Ratnayaka, J. M. Nolan, E. S. Miller, and J. D. Karam.** 2010. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol. J. **7:**292.
49. **Pickard, D., et al.** 2010. A conserved acetyl esterase domain targets diverse bacteriophages to the Vi capsular receptor of *Salmonella enterica* serovar Typhi. J. Bacteriol. **192:**5746–5754.
50. **Rohwer, F., and R. Edwards.** 2002. The phage proteomic tree: a genome-based taxonomy for phage. J. Bacteriol. **184:**4529–4535.
51. **Rohwer, F., et al.** 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. Limnol. Oceanogr. **45:**408–418.
52. **Rusch, D. B., et al.** 2007. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**e77.
53. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed., vol. 1. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
54. **Sandaa, R.-A., M. Clokie, and N. H. Mann.** 2008. Photosynthetic genes in viral populations with a large genomic size range from Norwegian coastal waters. FEMS Microbiol. Ecol. **63:**2–11.
55. **Sandaa, R.-A.** 2008. Burden or benefit? Virus-host interactions in the marine environment. Res. Microbiol. **159:**374–381.
56. **Sano, E., S. Carlson, L. Wegley, and F. Rohwer.** 2004. Movement of viruses between biomes. Appl. Environ. Microbiol. **70:**5842–5846.
57. **Savalia, D., et al.** 2008. Genomic and proteomic analysis of phiEco32, a novel *Escherichia coli* bacteriophage. J. Mol. Biol. **377:**774–789.
58. **Sebastian, M., and J. W. Ammerman.** 2009. The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. ISME J. **3:**563–572.
59. **Sharon, I., et al.** 2011. Comparative metagenomics of microbial traits within oceanic viral communities. ISME J. **5:**1178–1190.
60. **Sharon, I., et al.** 2007. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J. **1:**492–501.
61. **Short, C. M., and C. A. Suttle.** 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl. Environ. Microbiol. **71:**480–486.
62. **Stewart, C. R., et al.** 2009. The genome of *Bacillus subtilis* bacteriophage SPO1. J. Mol. Biol. **388:**48–70.
63. **Su, Z., V. Olman, and Y. Xu.** 2007. Computational prediction of Pho regulons in cyanobacteria. BMC Genomics **8:**156.
64. **Sullivan, M. B., et al.** 2008. Portal protein diversity and phage ecology. Environ. Microbiol. **10:**2810–2823.
65. **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 Biol. **3:**e144.

66. **Sullivan, M. B., et al.** 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ. Microbiol. **12:**3035–3056.
67. **Sullivan, M. B., et al.** 2009. The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome.' Environ. Microbiol. **11:**2935–2951.
68. **Sullivan, M. B., et al.** 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. **4:**e234.
69. **Sullivan, M. B., J. B. Waterbury, and S. W. Chisholm.** 2003. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature **424:**1047–1051.
70. **Suttle, C. A.** 2005. Viruses in the sea. Nature **437:**356–361.
71. **Suttle, C. A., A. M. Chan, and M. T. Cottrell.** 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Appl. Environ. Microbiol. **57:**721–726.
72. **Tetu, S. G., et al.** 2009. Microarray analysis of phosphate regulation in the marine cyanobacterium *Synechococcus* sp. WH8102. ISME J. **3:**835–849.
73. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins.** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:**4876–4882.
74. **Tommassen, J., et al.** 1991. Characterization of two genes, *glpQ* and *ugpQ*, encoding glycerophosphoryl diester phosphodiesterases of *Escherichia coli*. Mol. Gen. Genet. **226:**321–327.
75. **Treusch, A. H., et al.** 2009. Seasonality and vertical structure of microbial communities in an ocean gyre. ISME J. **3:**1148–1163.
76. **Wang, J., et al.** 2005. Complete genome sequence of bacteriophage T5. Virology **332:**45–65.
77. **Wanner, B. L.** 1996. Phosphorus assimilation and control of the phosphate regulon, p. 1357–1381. *In* F. C. Neidhardt and R. Curtiss (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC.
78. **Weigele, P. R., et al.** 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. Environ. Microbiol. **9:**1675–1695.
79. **Weynberg, K. D., M. J. Allen, K. Ashelford, D. J. Scanlan, and W. H. Wilson.** 2009. From small hosts come big viruses: the complete genome of a second *Ostreococcus tauri* virus, OtV-1. Environ. Microbiol. **11:**2821–2839.
80. **Williamson, S. J., et al.** 2008. The Sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One **3:**e1456.
81. **Wilson, W. H., I. R. Joint, N. G. Carr, and N. H. Mann.** 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. Appl. Environ. Microbiol. **59:**3736–3743.
82. **Yoshida, T., et al.** 2008. Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. J. Bacteriol. **190:**1762–1772.
83. **Yoshida, T., et al.** 2006. Isolation and characterization of a cyanophage infecting the toxic cyanobacterium *Microcystis aeruginosa*. Appl. Environ. Microbiol. **72:**1239–1247.
84. **Yu, Y., M. Breitbart, P. McNairnie, and F. Rohwer.** 2006. FastGroupII: a web-based bioinformatics platform for analyses of large 16S rDNA libraries. BMC Bioinformatics **7:**57.
85. **Zeidner, G., et al.** 2005. Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. Environ. Microbiol. **7:**1505–1513.
86. **Zhong, Y., F. Chen, S. W. Wilhelm, L. Poorvin, and R. E. Hodson.** 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. Appl. Environ. Microbiol. **68:**1576–1584.