ABSTRACT

Title of Dissertation:	**VIROMICS AND BIOGEOGRAPHY OF ESTUARINE VIRIOPLANKTON**
	Mengqi Sun, Doctor of Philosophy, 2021
Dissertation directed by:	Dr. Feng Chen, Professor, Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science

Viruses are the most abundant biological entity in the ocean, and they can influence microbial mortality, evolution and biogeochemical cycles in marine ecosystems. Virioplankton communities in oceans have been studied extensively using viral metagenomics (viromics), but the estuarine viromes remain relatively unexplored. Estuaries are a complex and dynamic ecosystem. My dissertation is dedicated to understanding the composition and distribution of the virioplankton community in the Delaware Bay and Chesapeake Bay by investigating 16 viromes collected from these two bays. A total of 26,487 viral populations (contigs > 5kb) were identified in the two bays, establishing a high quality viromic dataset.

The vast majority of the dominant viral populations are unclassified viruses. Viral sequences obtained from marine single cell genomes or long read single molecule sequencing comprised 13 of the top 20 most abundant viral populations, suggesting that we are still far from understanding the diversity of viruses in estuaries. Abundant

viral populations (top 5,000) are significantly different between the Delaware Bay and Chesapeake Bay, indicating a strong niche adaptation of the viral community to each estuary. Surprisingly, no clear spatiotemporal patterns were observed for the viral community based on water temperature and salinity.

The composition of known viruses (i.e. phages infecting *Acinetobacter*, *Puniceispirillum*, *Pelagibacter*, *Synechococcus*, *Prochlorococcus*, etc.) appeared to be relatively consistent across a wide range of salinity gradients and different seasons. Overall, the estuarine viral community is distinct from that in the ocean according to the composition of known viruses.

N4-like viruses belong to a newly established viral family and have been isolated from diverse bacterial groups. Marine N4-like viruses were first found in the Chesapeake Bay, but little is known about their biogeographic pattern in the estuarine environment. N4-like viruses were confirmed to be rare in the estuary, and relatively more abundant in the samples from lower water temperature.

Viruses which infect SAR11 bacteria (pelagiphage) are one of most abundant viral groups in the open ocean. We found that the abundance and community profile of pelagiphage in the estuaries is similar to that in the open ocean, and has no correlation with environmental factors.

VIROMICS AND BIOGEOGRAPHY OF ESTUARINE VIRIOPLANKTON

by

Mengqi Sun

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Feng Chen, Chair
Professor Tsvetan Bachvaroff
Professor Barbara Campbell
Professor Russell Hill
Professor Shawn Polson

© Copyright by
Mengqi Sun
2021

Acknowledgements

I have received a great deal of support over the course of my graduate studies and the writing of this dissertation.

First and foremost, I would like to thank my advisor Dr. Feng Chen, for his mentorship and guidance throughout the course of my Ph.D. His expertise and insight on microbial ecology was invaluable in formulating this research project, and I am incredibly grateful for his patience, critical thinking, and fast responses for the past six years.

I would like to thank my committee members for their expertise: Dr. Tsvetan Bachvaroff for his extensive help with bioinformatics, Dr. Barbara Campbell for her help with bacterial community ecology, Dr. Shawn Polson for his assistance with methodology, and Dr. Russell Hill for his scientific and professional guidance.

This dissertation would not have been possible without the community sequencing program from the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) (CSP 1621), and the National Science Foundation (NSF) (OCE 1261359). I also acknowledge the financial support from the 4-year fellowship from China Scholarship Council (CSC), and the 1-year fellowship from the Ratcliffe Environmental Entrepreneurs Fellowship (REEF). In addition, the IMET (Institute of Marine and Environmental Technology) Student Enhancement Fund travel award and the Jacob K. Goldhaber Travel Grant of University of Maryland helped me travel to many conferences and collaborations.

I am grateful for the contributions from Dr. David Paez-Espino, Dr. Yanlin Zhao, Dr. Lanlan Cai, and the scientific guidance from Dr. Simon Roux. I would like to thank Dr. Rita Colwell for being an inspiration for my scientific endeavors.

Special thanks to Dr. Yuanchao Zhan and Dr. David Marsan for attending the cruises and producing the amazingly high-quality samples which laid the groundwork for this research. I also appreciate the help and support from my other lab-mates Ana Sosa, Dr. Daniel Fucich, Dr. Pervaiz Ali, Dr. Hualong Wang, Yufeng Jia, and many more fellow students and scholars.

Finally, I would like to thank my family for their unconditional love and support, especially my mother Dr. Xiao-Hua Zhang for providing such an inspiring role model to follow.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DEV: Delmarva Estuarine Virome

POV: Pacific Ocean Virome

TOV: TARA Ocean Virome

GOV: Global Ocean Virome

TOPC: TARA Ocean Polar Circle

DOM: Dissolved organic matter

GTA: Gene transfer agents

AMG: Auxiliary metabolic genes

FPKM: Fragments per kilobase million

PPT: Parts per thousand

MAGs: Metagenome assembled genomes

SAGs: Single amplified genomes

Chapter 1: Introduction

1.1 Marine viral ecology

1.1.1 Role of viruses in aquatic environments

Viruses are the most abundant biological entities in the world and are a critical part of microbial communities (Suttle, 2007). Since the discovery that viral-like particles exceed 10^6/ml in seawater (Bergh *et al.*, 1989; Proctor and Fuhrman, 1990), the role of viruses in the marine ecosystem has been studied extensively (Fuhrman, 1999; Wommack and Colwell, 2000; Suttle, 2005; Zimmerman *et al.*, 2020). The concentration of virioplankton can be more than 10^8/ml in eutrophic water (i.e. estuarine and coastal waters), and it can be lower than 10^4/ml in the oligotrophic water such as the Sargasso Sea (Wommack and Colwell, 2000). The virus particle to bacterial cell abundance ratio generally falls between 3-10, and this ratio is higher in eutrophic, productive environments (Wommack and Colwell, 2000).

Viruses were found to cause high mortality of marine bacteria and cyanobacteria (Proctor and Fuhrman, 1990), prompting the beginning of the field of marine viral ecology. Although viruses infect all forms of life, the majority of viruses in the ocean are phages, which are responsible for 10-40% of total bacterial morality, and influence the dissolved organic matter (DOM) cycle by lysing bacteria (Fuhrman, 1999; Weinbauer, 2004). In the microbial loop, bacteria utilize dissolved organic matter (DOM) released by phytoplankton and zooplankton, resulting in more DOM being respired by bacteria instead of being transported to the higher trophic levels of the food chain (Azam *et al.*, 1983). Viruses contribute to the microbial loop by lysing their hosts and producing dissolved organic matter, which can then be consumed by other planktonic microbes. This recycling process of organic matter is called the viral shunt, diverting carbon from the classical food chain (Wilhelm and Suttle, 1999; Suttle, 2005) (Fig. 1.1). It is estimated that 25% of primary production in the ocean flows through the viral shunt, releasing 3 gigatons of carbon into seawater every year, indicating the substantial impact of the viral shunt on marine carbon cycling (Breitbart *et al.*, 2018). Viruses have also been shown to play a role in marine nitrogen, iron and phosphorus cycling, and the effect depends on which nutrient is limiting (Pourtois *et al.*, 2020; Zimmerman *et al.*, 2020).

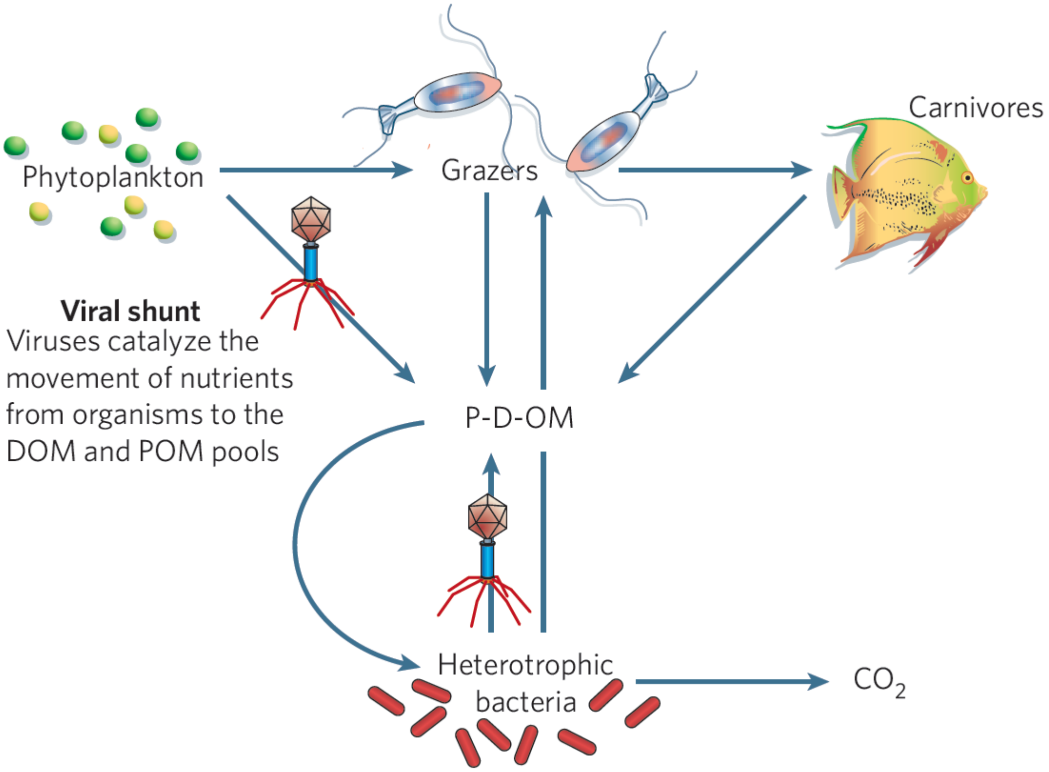
Carnivores
Grazers
Phytoplankton
Viral shunt
Viruses catalyze the movement of nutrients from organisms to the DOM and POM pools
P-D-OM
Heterotrophic bacteria
CO2

Figure 1.1 Viruses are catalysts for biogeochemical cycling. Viruses short circuit the flow of carbon and nutrients from phytoplankton and bacteria to higher trophic levels by causing the lysis of cells and shunting the flux to the pool of dissolved and particulate organic matter (D-P-OM). The result is that more of the carbon is respired, thereby decreasing the trophic transfer efficiency of nutrients and energy through the marine food web (Suttle, 2005). [Image Reprinted with permission from Springer Nature 2005, license no. 5117390600230].

Although viruses have a significant impact on marine ecology by killing their hosts, they do not always lyse their hosts immediately after infection. The lysogenic lifestyle is thought to aid phages through times of low host growth rates by protecting them from decay while "hiding" inside the host (Paul, 2008; Breitbart *et al.*, 2018). It has been estimated that up to 60% of aquatic bacteria contain lysogens, although only a small proportion (< 3%) can be induced to produce free bacteriophage (Ackermann and DuBow, 1987; Ogunseitan *et al.*, 1992). In the open ocean, lytic infections dominate the surface waters, while lysogeny is more common in waters below the deep chlorophyll maximum (DCM), and at deep sea hydrothermal vents (Luo *et al.*, 2020). Marine prophages can dramatically alter the phenotype of their hosts, and help promote host survival by suppressing unnecessary metabolic activities, changing their host's ecological function (Paul, 2008).

Viruses influence the diversity and metabolism of their hosts through a variety of mechanisms. Viruses transduce DNA between hosts via viral particles and gene transfer agents (GTAs) (Breitbart, 2012). GTAs are small particles resembling phages which contain random fragments of host DNA, and they mediate substantial amounts of horizontal gene transfer (Lang and Beatty, 2000). Phages also carry auxiliary metabolic genes (AMGs), genes mostly derived from their hosts to supplement host cell function during infection, in turn ensuring their own success (Breitbart *et al.*, 2007; Zimmerman *et al.*, 2020). In this way, phages act as genetic reservoirs for host evolution, and enhance host diversity (Hurwitz and U'Ren, 2016). Viral AMGs associated with carbon, sulfur and nitrogen cycling have been found in the ocean (Roux *et al.*, 2016). The power of viruses to reprogram host cells indicate significant influence on marine biogeochemical cycles, although the quantitative effects remain unclear due to the difficulty of conducting relevant large-scale experiments that imitate *in situ* behavior (Zimmerman *et al.*, 2020).

1.1.2 Models about virus-host interactions

Interactions between viruses and their hosts in the marine environment are dynamic and complex. A few main models that describe the effect of virus-host interactions on microbial diversity are “Red Queen Hypothesis”, “Kill the Winner”, “Piggyback the Winner” and “Bank Model”.

In general, the co-adaptation of virus and host follows “Red Queen” dynamics. The “Red Queen Hypothesis” proposes that the interaction between virus and host drives molecular co-evolution through natural selection for adaptation to each other (Van Valen, 1973). The impact of lytic viruses on host populations has been described as the “Kill the Winner” action (Thingstad, 2000). Unlike heterotrophic grazers, viruses are highly host-specific, so they can control the population of the most dominant bacterial hosts at a given time, promoting greater bacterial diversity and allowing for more efficient population succession (Rodriguez-Valera *et al.*, 2009). The “Bank Model” states that only the most abundant viruses are actively infecting hosts, while the rest of the viruses are inactive and act like a seed-bank, awaiting their turn to infect their hosts after population succession (Breitbart and Rohwer, 2005) (Fig. 1.2). This model compliments the “Kill the Winner” model and explains the high local community diversity of marine viruses, with most of the viruses at low abundance. Meanwhile, “Piggyback the Winner” refers to the situation in which temperate viruses tend to integrate into their hosts when the host cells are at high densities, exploiting the favorable survivability of the hosts for virus propagation (Knowles *et al.*, 2016). This interaction between temperate viruses and their hosts is particularly important for maintaining the stability of host-associated microbial communities (Silveira and Rohwer, 2016).

These models explain the presence of diverse viral communities from different aspects (i.e. at the population and genomic level), and contribute to resolving the paradox of plankton, which describes the presence of diverse planktonic community in a resource-limited environment (Hutchinson, 1961). Despite being largely non-living organisms, viruses still rely on their hosts for replication, thus require nutrient resources. The constant flux of dynamic virus-host interactions enables the existence of the vast microbial diversity found in aquatic environments.

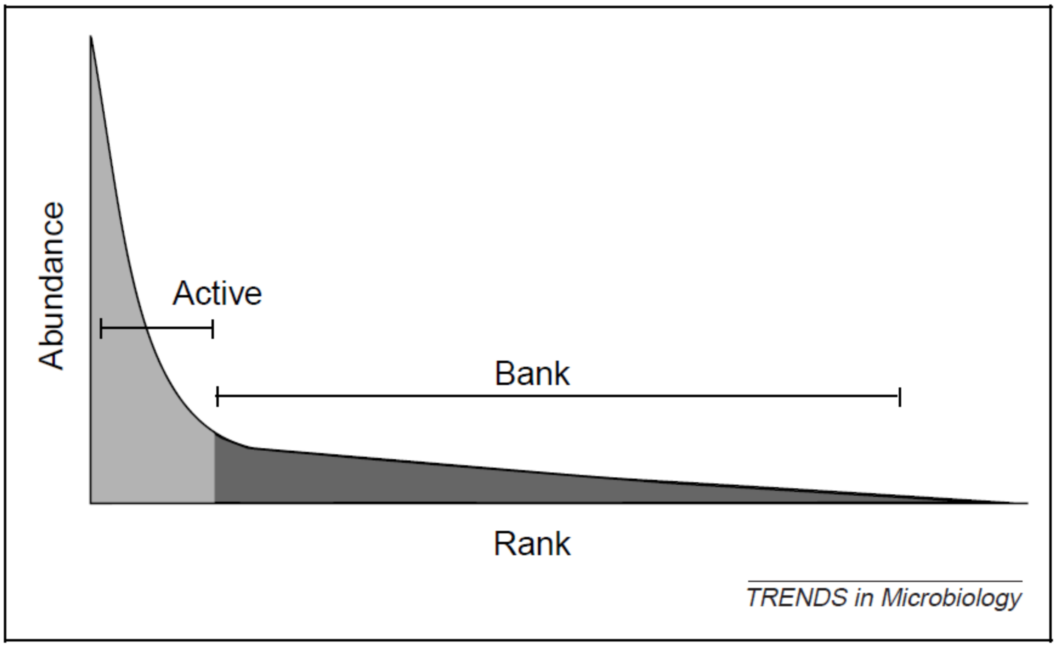

Figure 1.2 Example of a rank-abundance curve. In the Bank model, only a few of the most abundant viral genotypes are in the Active fraction. As new prey items become dominant in response to changing environmental conditions, the viruses that can prey on those hosts also become abundant. The viruses that were previously in the Active fraction begin to decay and in the absence of new production become part of the bank fraction (Breitbart and Rohwer, 2005). [Image Reprinted with permission from Elsevier 2005, license no. 5117391073027].

1.2 Delaware Bay and Chesapeake Bay

1.2.1 Estuarine ecosystems

An estuary is a body of water where fresh water and seawater measurably mix (Pritchard, 1967). Estuaries are vital links between marine and terrestrial ecosystems, and are among the most productive ecosystems on the planet (Field *et al.*, 1998). Estuarine systems encompass a complex spectrum of environmental gradients, creating distinct microbial habitats, and the frequent fluctuation of environmental conditions posts unique selective pressures to be exerted on organisms (Fortunato and Crump, 2011). This study focuses on viral communities in the Delaware Bay and the Chesapeake Bay, which are briefly described below.

1.2.2 The Delaware Bay

As the second largest estuary on the U.S. Atlantic coast, the Delaware Bay is an archetypal, funnel shaped, well-mixed coastal plain estuary (Hermes and Sikes, 2016). The geometry of the bay is simple, with a mean depth of 8 m and a maximum depth of 45 m (Aristizábal and Chant, 2014). It is heavily urbanized in the upper bay, yet it supports important wetlands and fisheries in the lower bay, and its drainage basin is dominated by agricultural activity (Sharp, 1983). It has been characterized as a high nutrient and low biomass growth environment, with very little bottom water hypoxia (Sharp *et al.*, 2009).

1.2.3 The Chesapeake Bay

The Chesapeake Bay is the largest and most productive estuary in the U.S, featuring shallow waters with a mean depth of 6.5 m. It is a partially mixed estuary featuring dynamic patterns of internal transport and a long (~180 days) water residence time (Marshall *et al.*, 2005; Du and Shen, 2016). The modern Chesapeake Bay was formed by the most recent rise in sea level and is less than 10,000 years old (Schubel and Pritchard, 1986). Annual freshwater flow from the Susquehanna River is highly variable, impacting the ecology of the bay (Harding *et al.*, 2016). The Chesapeake Bay is considered a prominent example of nutrient over-enrichment in estuaries, resulting in zones of hypoxia which is enhanced by summer stratification (Sharp *et al.*, 2009).

1.2.4 Comparison of Delaware Bay and Chesapeake Bay

The Delaware and Chesapeake Bays are separated by the Delmarva Peninsula, and they differ in many aspects. The Chesapeake Bay has a huge watershed (about 166,000 km^2) that is about 80 times larger than the Delaware Bay (about 2,000 km^2) (Scudlark and Church, 1993). The Delaware Bay receives enormous tidal flow, with 6 to 7 feet of tidal water, while the Chesapeake Bay has a smaller tidal difference (ca. 2 feet between high and low tide) (Sharp *et al.*, 2009). Salt marshes and mudflats build up the major shoreline of Delaware Bay, but are less prevalent in the Chesapeake Bay (Scudlark and Church, 1993). The Delaware River, the main river input to the Delaware Bay, is among the worst polluted waterways in the nation due to the release of toxic chemicals from the surrounding industries (Seth Augenstein, 2012). On the other hand, the oxygen-depleted zone caused by eutrophication in the stratified summer Chesapeake Bay waters posts a serious threat to many economically important animal species such as blue crabs, oysters, and fish (Boesch *et al.*, 2001). In the Chesapeake Bay, phytoplankton productivity appears to decrease seaward with reduced nutrient levels (Harding *et al.*, 1986; Fisher *et al.*, 1988). In contrast, only slight or no nutrient limitation was found in the Delaware Bay (D'Elia *et al.*, 1986; Fisher *et al.*, 1988). In general, a large portion of the Chesapeake Bay is nutrient-limited, while the Delaware Bay has higher nutrient and turbidity levels (Fisher *et al.*, 1988). It is unknown how these profound abiotic differences in the two different estuarine ecosystems impact the virioplankton communities.

1.3 Bacterioplankton communities in the estuarine environment

1.3.1 Effect of estuarine conditions on bacterioplankton

In a highly dynamic estuarine environment, changes in environmental factors can trigger genetic and ecological shifts in microbial communities (Herbert, 1999). Cell densities and growth rates of bacteria in estuaries are generally higher than those in coastal and river waters, and they tend to be highest in surface waters and turbid regions (Wright and Coffin, 1983).

Bacterioplankton community shifts are highly dependent on the biological and hydrological condition of the estuary. The community structure of bacterioplankton in the Chesapeake Bay showed repeatable and predictable seasonal patterns as revealed by the analysis of denaturing gradient gel electrophoresis (DGGE) of PCR-amplified 16S rRNA genes (Kan *et al.*, 2006). This study provided the first comprehensive understanding of the change of the Chesapeake Bay bacterial community over time and space. Water temperature, Chl a, dissolved oxygen, nutrients, and viral abundance all appear to play important roles in structuring the bacterial communities in the Chesapeake Bay, while Chl a and water temperature are two major factors affecting the shift of bacterial communities (Kan *et al.*, 2006). In a later study involving cloning and sequencing analysis, the Chesapeake Bay bacterial community exhibited much stronger seasonal, rather than spatial variation (Kan *et al.*, 2007). The stronger seasonal than spatial and interannual variations in the Chesapeake Bay were confirmed by a recent study based on the deep sequencing of bacterial community (H. Wang *et al.*, 2020). The study also found repeatable patterns of interannual variation among the estuarine bacterioplankton community (H. Wang *et al.*, 2020). Alternatively, in the Columbia River estuary, a salt wedge estuary, seasonal variability of bacterioplankton was obscured by strong spatial variability (Fortunato *et al.*, 2012). Yet in the Delaware Bay, a well-mixed estuary, both seasonal and spatial variation of bacterioplankton were strong (Campbell *et al.*, 2011; Campbell and Kirchman, 2013). Estuarine ecosystems are complex, and cannot be simplified based on one ecosystem model. Therefore, the impact of environmental factors on bacterioplankton community can be different between the Chesapeake Bay and Delaware Bay.

1.3.2 Taxonomic composition of bacterioplankton in the Delaware Bay and Chesapeake Bay

The taxonomic composition of the bacterioplankton community in the Delaware Bay and Chesapeake Bay share similarities, and also some differences. Along the salinity gradient of the Delaware Bay, the bacterioplankton composition changed from a community dominated by *Actinobacteria, Verrucomircobia* and *Betaproteobacteria* in fresh waters to a typical marine community dominated by SAR11 taxa, *Rhodobacterales, Gammaproteobacteria* and *Bacteroidetes* in the lower bay

(Campbell and Kirchman, 2013). Seasonally, it has been found that about half of the Delaware Bay bacterial community cycles between rare and abundant species, with rare bacteria acting as a ‘seed bank’ waiting for conditions to change (Campbell *et al.*, 2011). On the other hand, the Chesapeake Bay contains certain SAR11, *Roseobacter*, SAR86 and *Actinobacteria* subclades that may be adapted to estuaries with long residence times (Kan *et al.*, 2008). The Chesapeake Bay has stronger temporal than spatial variation, but in terms of spatial variation, *Actinobacteria* and *Betaproteobacteria* still give way to *Gammaproteobacteria* from upper to lower Chesapeake Bay, although the contrast is less stark compared to the Delaware Bay (H. Wang *et al.*, 2020). Meanwhile, *Verrucomicrobia* follows the opposite pattern compared to the Delaware Bay, increasing along the salinity gradient (H. Wang *et al.*, 2020). Metatranscriptomic patterns of the Chesapeake Bay microbial community differ between shallow and deep water, reflecting the effect of its summer stratification (Hewson *et al.*, 2014).

1.4 Virioplankton community in estuaries

1.4.1 Role of virioplankton in estuaries

Marine viruses are the most numerically abundant biological entities in the world and are an important part of microbial communities (Suttle, 2005). Virioplankton are usually one order of magnitude more abundant than bacterioplankton (Wommack and Colwell, 2000). The abundance of virioplankton in the Chesapeake Bay is in the range of 10^6-10^8 viral like particles (VLPs) per milliliter (Bergh *et al.*, 1989; Wommack *et al.*, 1992), which can be 10-1,000 times more abundant than the viral concentration in the open ocean (Wommack and Colwell, 2000). The high abundance of viruses in estuaries seems to be related to high bacterial biomass and productivity in the estuarine environment.

Virioplankton are an active and dynamic component of estuarine microbiomes, and are responsive to environmental changes (Wommack *et al.*, 1999; Bench *et al.*, 2007). Viruses are an important part of the trophic system in estuaries as they are responsible for bacterial mortality at a level similar to protist grazing (Wommack *et al.*, 1992; Fuhrman and Noble, 1995). Also, viruses are sensitive to the mixing of fresh and marine water. Experiments to test the effects of freshwater and seawater mixing on virioplankton and bacterioplankton in a tropical estuary showed viral production to rapidly respond to shifts in the estuarine bacteria community, with virioplankton following the dynamics of bacterioplankton within 24 hours (Cissoko *et al.*, 2008). Production of freshwater bacteria and viruses sharply declined as a result of seawater addition, but marine bacteria and viruses were not significantly affected by freshwater addition, possibly taking advantage of less adaptable freshwater cells bursting from osmotic shock (Cissoko *et al.*, 2008).

1.4.2 History of virioplankton ecology in the Chesapeake Bay

The Chesapeake Bay has a rich history of pioneer studies in virioplankton ecology. Marine viruses were first discovered in 1955, but they were considered to be sparse, hence unimportant until 3 decades later (Spencer, 1955). Numerous viral particles were observed in the Chesapeake Bay as part of the first study to report high abundances of viral particles in aquatic environments, alerting the world to the potential ecological impacts of aquatic viruses (Bergh *et al.*, 1989). There has been a continuous effort to understand the ecological role and diversity of the virioplankton community in the Chesapeake Bay ever since.

The virus-to-bacterium ratio (VBR) is a good indicator of how phages interact with bacteria in the natural environment. In the Chesapeake Bay the ratio varied from 3.2 to 25.6 (Wommack *et al.*, 1992), suggesting that viruses are more abundant than bacteria, and viral infectivity can affect the abundance of bacterioplankton. An inter-annual study of virioplankton ecology, the Microbial Observatory of Virioplankton Ecology (MOVE) project in the Chesapeake Bay was conducted between 2003 and 2007, with a goal to understand how the abundance and community structure of virioplankton in the Chesapeake Bay change over time and space (Wang, 2007). During the MOVE project, viral abundance and production were measured over a 54-month study. Interestingly, viral abundance and viral production did not change greatly from the upper to lower bay, despite strong environmental gradients (i.e. nutrient, light, salinity, etc.) (Winget *et al.*, 2011). The temporal dynamics of viral productivity in the Chesapeake Bay can be affected by the abundance, productivity and composition of bacterioplankton (Winget *et al.*, 2011). It was estimated that viral lysis released 76 μg of organic carbon per L per day in the Chesapeake Bay, and such a viral lysis rate could support about 55% of organic carbon needed for daily bacterioplankton production (Winget *et al.*, 2011). The study suggested that bacterioplankton are subject to frequent infection by virioplankton, and viral activity contributes greatly to microbial carbon cycling in the Chesapeake Bay.

Wommack et al. (1992) applied Pulsed-Field Gel Electrophoresis (PFGE) to investigate the change of viral community structure in the Chesapeake Bay. PFGE is able to separate viruses based on their genome sizes, therefore, the ds DNA dominant viral populations from natural samples can be visualized and compared on PFGE gels. In a survey of Chesapeake Bay samples collected from six stations in August 1995 and May, June, and July 1996, seven distinct genome size groups (23, 23 to 48.5, 48.5 to 97, 97 to 145.5, 145.5 to 194, 194 to 242.5, and 242.5 kb) were identified based on PFGE fingerprints (Wommack *et al.*, 1992). The patterns of PFGE fingerprints were analyzed based on principal-component and clustering analyses, and they concluded that variations of viral communities in the Chesapeake Bay were correlated with time, location and level of stratification. In a later study, a different molecular method, Randomly Amplified Polymorphic DNA (RAPD) PCR, was developed and used to investigate the dynamics of virioplankton communities in the Chesapeake Bay (Winget and Wommack, 2008). Based on the RAPD-PCR banding patterns, they reported that the Chesapeake Bay virioplankton community exhibited stronger temporal than spatial variation, a pattern similar to the spatiotemporal variations seen for the Chesapeake Bay bacterioplankton community (Kan *et al.*, 2007).

The first metagenomics study on estuarine virioplankton was conducted in the Chesapeake Bay by sequence analysis of one sample pooled from nine different locations in the Bay (Bench *et al.*, 2007). The Chesapeake Bay viromics unveiled a high proportion of unknown and novel sequences. Among identified and assigned viral sequences, more than 90% were most similar to Caudovirales. More specifically, 42 and 41% of virus sequences belonged to members of the *Myoviridae* and *Podoviridae* families, respectively, while *Siphoviridae* only account for 6% of the viral homolog sequences. This study provided the first viromic data on an estuarine virioplankton community. Despite the limitation on low sequencing coverage in the early days of viromic studies, it was found that the Chesapeake Bay virome contains a high proportion of unknown and novel sequences. As viral samples were pooled, no information on spatial and temporal patterns can be obtained from this study. Compared to the Chesapeake Bay, less is known about virioplankton in other estuaries. In contrast to the Chesapeake Bay virioplankton, *Siphoviridae* consist as much as one third of Caudovirales in the Jiulong River estuary and also the Pearl River estuary (Cai *et al.*, 2016; C. Zhang *et al.*, 2021). Metagenomic studies have found that estuarine viral communities are heavily influenced by marine waters, with high percentages of typical marine viruses such as *Pelagibacter* phage, *Roseobacter* phage, *Puniceispirillum* phage, and *Ostreococcus* phage (Cai *et al.*, 2016; Hwang *et al.*, 2016).

1.5 Marine viral metagenomics

1.5.1 Findings of marine viral metagenomics

In the past ten years, the development of new sequencing technologies has greatly advanced our understanding on microbial diversity in nature. Using next-generation sequencing (NGS) technologies, a number of large-scale ocean sequencing projects (e.g. Global Ocean Sampling Expedition, Malaspina Expedition, Pacific Ocean Virome, Tara Ocean’s Global Ocean Virome (GOV)) have made viral metagenomic databases increasingly accessible, revealing important findings about the diversity, spatial and temporal distribution of ocean viruses (Williamson *et al.*, 2008; Hurwitz and Sullivan, 2013; Brum *et al.*, 2015; Roux *et al.*, 2016, Gregory *et al*. 2019).

Early metagenomic studies revealed most of the viral community sequence space to be unknown, inciting interest in exploring the vast diversity of marine viruses (Breitbart *et al.*, 2002, 2007). As studies involving a more diverse range of samples were conducted, viruses were found to be globally distributed, but highly diverse on the local scale, likely due to selection pressure of local environmental and biological factors (Breitbart and Rohwer, 2005; Angly *et al.*, 2006; Paez-Espino *et al.*, 2016).

More detailed analysis of viral sequences revealed marine virus communities to be taxonomically and functionally distinct across different seasons, depths and proximity to shore (Hurwitz and Sullivan, 2013). Also, the distribution of viruses is dependent on the distribution of their bacterial hosts, although viruses are also passively dispersed by ocean currents (Brum *et al.*, 2015). In the oligotrophic open ocean, most dsDNA viruses persist over several years, forming a core viral community, but their relative abundance and transcriptional activity fluctuates depending on the population variation of their hosts (Aylward *et al.*, 2017; Luo *et al.*, 2020). A recent large-scale study, Tara Ocean's GOV 2.0, shows that marine viral communities can be separated into five ecological zones, although no estuarine samples were included (Gregory *et al.*, 2019). Meanwhile, many viromic studies have shown that the most abundant viral species in the ocean still remain unknown (Paez-Espino *et al.*, 2016; Roux *et al.*, 2016).

Large-scale sequencing efforts generally only include a few sampling sites at coastal and brackish locations (Bench *et al.*, 2007; Rusch *et al.*, 2007; McDaniel *et al.*, 2008;

Williamson *et al.*, 2008; Cai *et al.*, 2016; Hwang *et al.*, 2016; Zeigler Allen *et al.*, 2017). Samples from different sites were sometimes combined to reduce the cost of sequencing, and different sequencing methods often yielded differing results (Table 1.1). The Delaware Bay and Chesapeake Bay virioplankton community has been studied before using metagenomics, but only using outdated Sanger sequencing technology. As of now, there has not been any systematic study of spatial and temporal variation of virus communities in dynamic estuarine environments using deep sequencing technology.

Table 1.1 Summary of estuarine metagenomic viral datasets to date. Abbreviations: GOS, Global Ocean Sampling; BSV, Baltic Sea Virome; DEV, Delmarva Estuarine Virome.

Publication	**Sample site(s)**	**Salinity**	**Study type**	**Sequencing method**
Bench 2007	Chesapeake Bay (9 stations combined)	NA	Environmental	Sanger
Williamson 2008 (GOS)	Bay of Fundy, Canada	NA	Environmental	Sanger
	Delaware Bay	NA		
	Chesapeake Bay	3.47		
McDaniel 2008	Tampa Bay	NA	Induced virome	454 GS20
Cai 2016	Jiulong Estuary, China	25.50	Environmental	454 GS FLX
Hwang 2016	Goseong Bay, Korea (6 stations combined)	34	Environmental	Illumina Hiseq 2000
Allen 2017 (BSV)	Baltic Sea (10 separate stations)	0 - 34.35 (10 samples)	Environmental	454 GS FLX
Zhang 2021	Pearl River estuary (3 separate stations)	5 – 30 (3 samples)	Environmental	Illumina Miseq
This study (DEV)	Delaware Bay (10 separate stations) Chesapeake Bay (6 separate stations)	0.2 - 30.4 (16 samples)	Environmental	Illumina HiSeq 2500

1.5.2 Tools for viral metagenomic analysis

With the increased interest in viral metagenomics, numerous tools and pipelines have emerged to handle different aspects of viral community data analysis. These tools are in various states of accessibility, with some tools available as simple browser-based web applications, and others that require extensive programming background. It has been noted that the rapid development of viral community analysis software has enabled the viral ecology field to shift from "specialists" to "non-specialists", allowing for more collaboration across different fields (Sullivan *et al.*, 2017).

Here I list some of the tools I have used or attempted to use during the course of my dissertation work, and is by no means an exhaustive list (Table 1.2). Other reviews provide a more detailed evaluation and benchmarking of the functionality of various virus metagenomic tools (Rose *et al.*, 2016; Tangherlini *et al.*, 2016; Roux *et al.*, 2017; Nooij *et al.*, 2018).

Without a universal marker gene like in cellular microorganisms, methods specific for viral community sequence analyses have been developed. Broad scale viral classification generally relies on whole genomes to determine their phylogeny, with methods such as the Phage Proteomic Tree, ViPTree, and VICTOR (Rohwer and Edwards, 2002; Meier-Kolthoff and Göker, 2017; Nishimura *et al.*, 2017). Assigning viral sequences to known taxonomy is more difficult, since typically less than 3% of aquatic virus community sequence data can be assigned to standard ICTV (International Committee on Taxonomy of Viruses) taxonomy, but network analysis based methods have been developed (Bolduc *et al.*, 2017; Bin Jang *et al.*, 2019). New technology such as machine learning has also been deployed to detect viruses from metagenomic sequences (Ponsero and Hurwitz, 2019; Ren *et al.*, 2020).

Table 1.2 Summary of different virus metagenomic tools and pipelines. *Not specifically for viruses.

Function	**Name**	**Research group**	**Citation**	**Availability**
Comprehensive	Metavir2	François Enault	(Roux *et al.*, 2014)	web application
	VIROME	Eric Wommack	(Wommack *et al.*, 2012)	web application
	IMG/VR	David Paez-Espino	(Paez-Espino, Chen, *et al.*, 2017)	IMG website
	MG-RAST*	Folker Meyer	(Keegan *et al.*, 2016)	web application
Find viruses	Virsorter (iVirus)	Matt Sullivan	(Roux *et al.*, 2015)	Cyverse app
	Virfinder	Jed Fuhrman	(Ren *et al.*, 2017)	R package on github
	VirusSeeker	David Wang	(Zhao *et al.*, 2017)	bunch of perl scripts in need of local configuration
	IMG/VR (identification part)	David Paez-Espino	(Paez-Espino, Chen, *et al.*, 2017)	Command line instructions
Classification	vContact 2.0 (iVirus)	Matt Sullivan	(Bin Jang *et al.*, 2019)	Cyverse app
	Kaiju*	Anders Krogh	(Menzel *et al.*, 2016)	web application
Link viruses and hosts	VirhostMatcher	Jed Furhman	(Ahlgren *et al.*, 2017)	script on github
	VirHostMatcher-Net	Jed Furhman	(W. Wang *et al.*, 2020)	Command line application
	IMG/VR	David Paez-Espino	(Paez-Espino, Chen, *et al.*, 2017)	IMG website
	HostPhinder	Mette Voldby Larsen	(Villarroel *et al.*, 2016)	web application

1.5.3 Limitations of viral metagenomics

Unlike bacteria and other cellular microbes, viruses lack a universal conserved gene such as the 16S rRNA and the 18S rRNA, limiting marker gene-based viral studies to certain groups of viruses of which genome information is already reasonably known (Sullivan, 2015). Hence, with the constantly decreasing costs of high-throughput sequencing, viral metagenomics has become a widely used method of relatively unbiased exploration of naturally occurring viral communities (Breitbart *et al.*, 2007). Despite being a powerful and widely used approach, metagenomic methods have a few limitations:

(1) The analysis of community sequences is dependent on known sequences, but the vast majority of metagenomic viral sequences are of unknown origin, for which the term “viral dark matter” has been coined (Youle *et al.*, 2012).

(2) Due to the nature of community-based sequencing, it is unknown whether the sequences are from viable viral particles.

(3) It yields only relative viral abundance counts, and not absolute abundance counts (Coutinho *et al.*, 2018).

(4) It is difficult to connect viruses to the hosts they infect, although there have been various methods to predict the hosts of viruses based on genome data (Table 1.2).

The limitations of viral metagenomics have prompted new methods of discovery based on more precise, uncultivated methods such as single cell, single virus and single molecule sequencing.

1.6 Uncultured virus discovery

1.6.1 Using single-cell and single-molecule technology for virus discovery

With the rise of new methodology, an unprecedented amount of viruses is now being discovered through non-culture based methods (Brum and Sullivan, 2015). The uncultivated microbial community is being explored using shotgun metagenomics, amplicon metagenomics, and single cell/virus genomics (Fig. 1.3). In recent years, single-cell genomics has offered valuable insights into the marine viral community (Labonté *et al.*, 2015), discovering some of the most abundant and ecologically significant viruses in the marine ecosystem (Martinez-Hernandez *et al.*, 2017; Berube *et al.*, 2018). For example, the abundance of the single virus isolate vSAG 37-F6, of which the putative host is *Pelagibacter* (Martinez-Hernandez, Fornas, *et al.*, 2019), is thought to rival or exceed that of *Pelagibacter* phage HTVC010P and *Puniceispirillum* phage HMO-2011, which were previously thought to be the most abundant viruses in the ocean (Kang *et al.*, 2013; Zhao *et al.*, 2013; Martinez-Hernandez *et al.*, 2017). 37-F6 has also been found to be transcriptionally active in the ocean, including in coastal eutrophic waters (Martinez-Hernandez *et al.*, 2020). Despite being ecologically important, 37-F6 was completely missed in viral ecology studies prior to 2017, because the high microdiversity of its genome prevented it from being assembled in traditional shotgun metagenomic datasets (Martinez-Hernandez *et al.*, 2017). Likewise, long-read single-molecule sequencing uses long Nanopore reads (20 – 80 kb) to capture entire viral genomes without assembling, avoiding some of the biases induced by short-read de novo assembly, thus revealing “hidden” viral diversity not covered by conventional metagenomic sequencing methods (Beaulaurier *et al.*, 2020). In summary, single amplified genomes (SAGs) provide a necessary compliment to metagenome-assembled genomes (MAGs), due to limitations of metagenome assemblies (Figure 1.3).

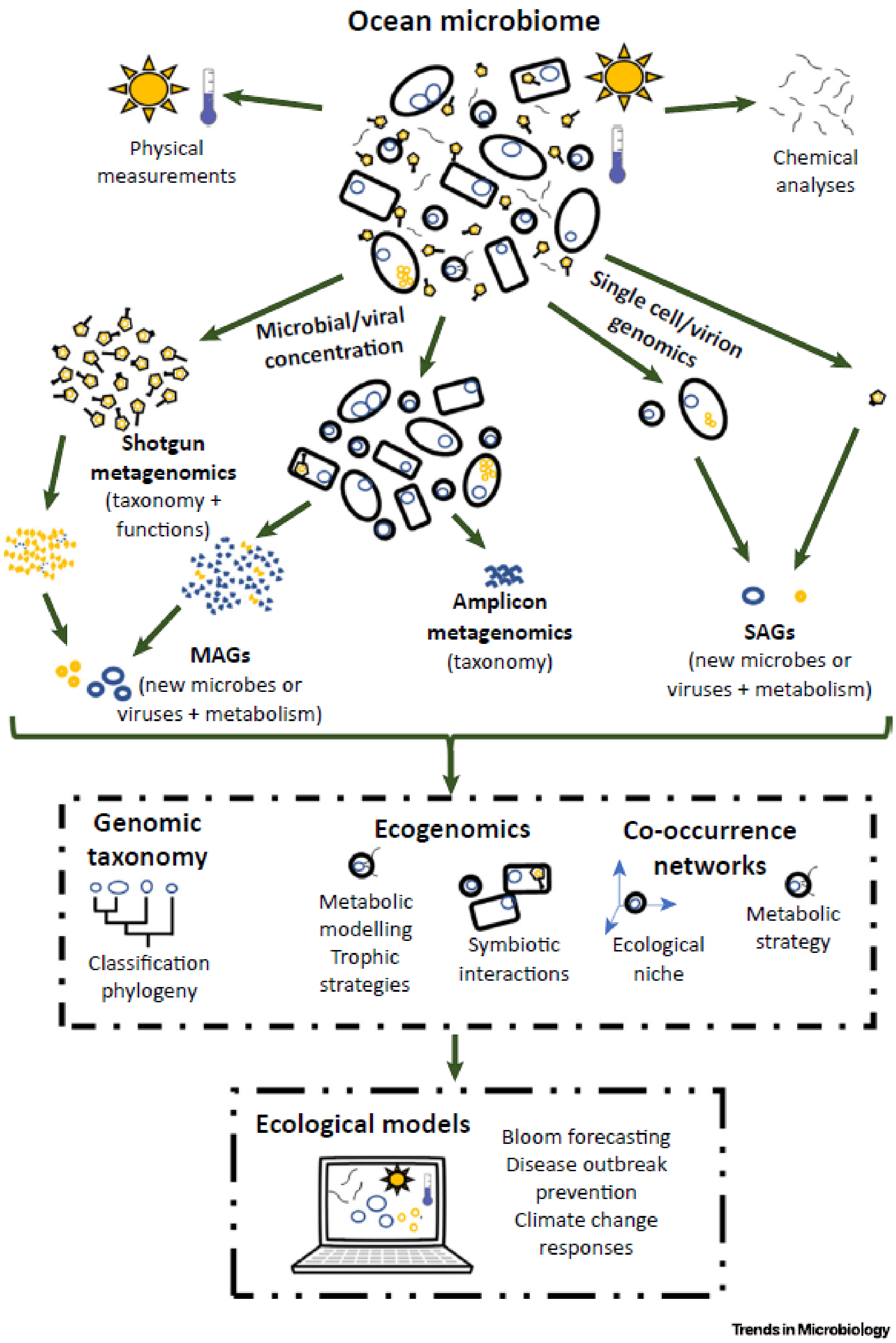
Ocean microbiome
Physical measurements
Chemical analyses
Microbial/viral concentration
Single cell/virion genomics
Shotgun metagenomics (taxonomy + functions)
Amplicon metagenomics (taxonomy)
MAGs (new microbes or viruses + metabolism)
SAGs (new microbes or viruses + metabolism)
Genomic taxonomy
Classification phylogeny
Ecogenomics
Metabolic modelling Trophic strategies
Symbiotic interactions
Co-occurrence networks
Ecological niche
Metabolic strategy
Ecological models
Bloom forecasting
Disease outbreak prevention
Climate change responses
Trends in Microbiology

Figure 1.3 Approaches to study the ocean microbiome through ecogenomics and metagenomics. MAG, metagenome-assembled genome; SAG, single amplified genome (Coutinho *et al.*, 2018). [Image Reprinted with permission from Elsevier 2018, license no. 5117391254725].

1.6.2 Cultured vs. uncultured viruses

Many marine microbes are difficult to culture in the laboratory due to their special nutritional requirements (Joint *et al.*, 2010). Cultivation and maintaining a large collection of microbial cultures is labor-intensive and only practical for a limited amount of microbial species. Since the recovery of infectious viruses is dependent on the cultivation of the hosts they infect, cultured virus isolates represent an extremely limited slice of total viral diversity in nature. Although cultivation techniques have been improved significantly and many important viruses infecting abundant groups of bacterioplankton have been isolated (Kang *et al.*, 2013; Brum and Sullivan, 2015; Buchholz *et al.*, 2021), the progress is slow and still cannot provide a comprehensive picture on viral diversity in nature. As of now, phages infecting abundant but relatively slow-growing and difficult-to-culture marine bacteria still make up a significant portion of marine viruses in the ocean (Z. Zhang *et al.*, 2021).

Due to the advances of single cell, single molecule, and metagenomic methods, starting from around 2016, uncultivated virus genomes have vastly outnumbered cultured virus isolates (Fig. 1.4). As such, the International Committee on Taxonomy of Viruses (ICTV) has recently started accepting viral taxonomy without cultured isolates, using Minimum Information about an Uncultivated Virus Genome (MIUViG) standards (Roux *et al.*, 2019). This development indicates the trend of increased importance of uncultured viruses not only in viral community ecology, but also virology in general.

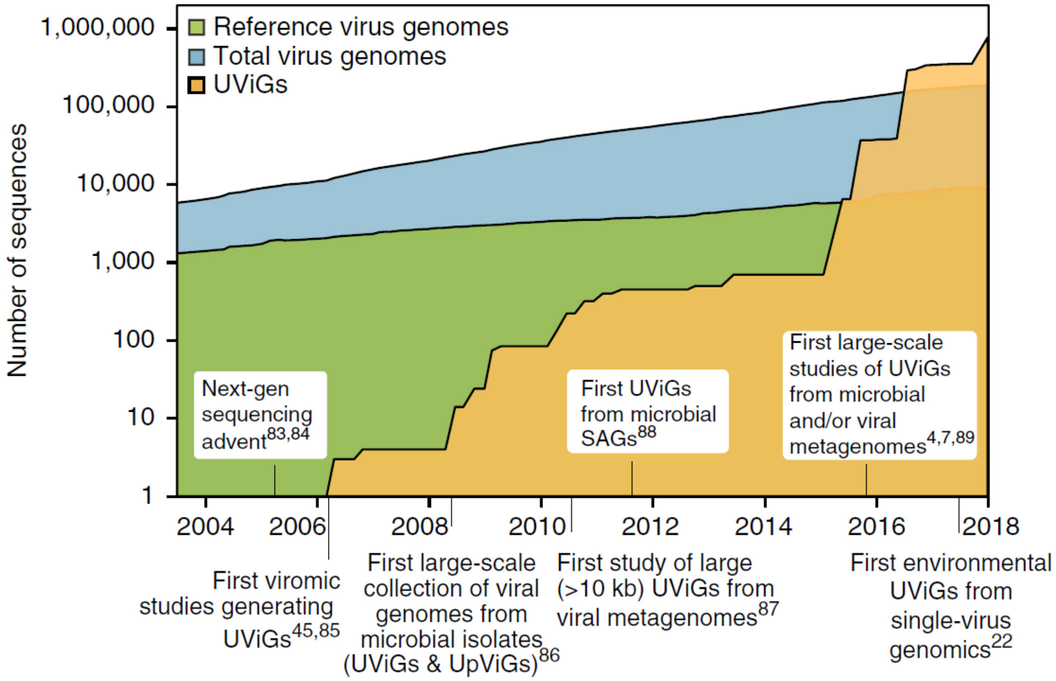

Figure 1.4 Size of virus genome databases over time. Genome sequences from isolates (blue and green) or from UViGs (Uncultivated Virus Genomes) (yellow) are shown. For genomes from isolates, the total number of genomes (blue) and the number of 'reference' genomes (green) are shown (Roux *et al.*, 2019). [Image reprinted under Creative Commons license 4.0].

1.7 Scope of this dissertation

1.7.1 Scope, questions and hypotheses

Among the most productive environments on earth, estuaries are important links that connect terrestrial and marine ecosystems. The field of viral ecology is now 30 years old, and is a maturing field, with changing paradigms in the face of accelerated discovery and new technologies (Brum and Sullivan, 2015; Sullivan *et al.*, 2017). Recent years have seen a marked increase in the study of viral ecology using next-generation sequencing technology, mostly focused on marine and freshwater environments. However, our knowledge of estuarine viral communities is still limited, and there are unanswered questions. How does the estuarine virioplankton community vary across seasons and environmental gradients as seen through viral metagenomics, and is the viral variation correlated with changes in the estuarine bacterioplankton community? How do rare viruses and abundant viruses behave differently in the estuarine ecosystem? How does the estuarine virioplankton community differ from freshwater and oceanic virioplankton communities? The Delaware Bay and Chesapeake Bay have a rich history of discoveries in microbial ecology, but how has the advancement of technology impacted our understanding of microbial ecology paradigms?

The goal of this dissertation is to investigate the diversity and tempo-spatial variation of virioplankton communities in two temperate estuaries, Delaware and Chesapeake Bays, using deep sequencing technology. The estuarine viral communities were compared with coastal and open ocean viral communities which are available in the public domain. We also investigated the presence of one rare marine virus group (N4-

like viruses), and one abundant marine virus group (pelagiphages) in our viromic database. Our laboratory has a long history of studying N4-like viruses, which have unique genomic characteristics, but are of low abundance in the aquatic environment. *Pelagibacter* phages are considered to be the most dominant viruses in the ocean, but little is known about their presence in estuarine and freshwater environments. With regard to these topics of interest, the following four specific hypotheses were tested in my dissertation.

Hypothesis 1. The community structure of viruses in the Delaware Bay and Chesapeake Bay varies greatly over spatial and temporal scales.

Hypothesis 2. The composition of the virioplankton community in the estuary is different from that in coastal and open oceans.

Hypothesis 3. N4-like viruses are more prevalent in cold water based on viral metagenomics.

Hypothesis 4. *Pelagibacter* phage abundance varies greatly along the salinity gradient in the estuaries.

1.7.2 Summary of chapters in this dissertation

This dissertation is split into 6 chapters. Chapter 1 is the introduction, providing background and context to our studies. Chapter 2 describes the Delmarva Estuarine Virome (DEV) dataset in detail, including the viral sequence processing (Sun *et al.*, 2021), providing a baseline for chapters 3, 4 and 5. Chapter 3, the highlight of this dissertation, reveals how the viral community varies on the spatial and temporal scale in the Delaware and Chesapeake estuaries (Sun *et al.*, 2021). Chapter 4 and Chapter 5 investigates specific viral groups, based on the viral populations provided in Chapter 2. Chapter 4 explores the abundance and dynamics of N4-like viruses (a rare virus group), and Chapter 5 explores pelagiphages (an abundant virus group) in the estuary. While N4-like viruses are a group of viruses with highly conserved genomes infecting various hosts, pelagiphages are a set of various phages that infect Pelagibacterales (SAR11). Chapter 6 provides the summary and future directions of this dissertation.

Chapter 2: Data descriptor: Delmarva Estuarine Virome (DEV)

2.1 Introduction

The viral metagenomics study on the Delaware Bay and Chesapeake Bay is part of the research project supported by the Community Science Program of the Department of Energy (DOE) Joint Genome Institute (JGI). The project is entitled "Biogeochemical cycling links between terrestrial and marine systems", and was led by four PIs, Barbara Campbell, David Kirchman, Feng Chen, and Michael Gonsior. The goal of the project is to understand the impact of dissolved organic matter on bacterial community in the estuarine environment. Virioplankton are abundant and diverse in the estuarine ecosystem (Wommack and Colwell, 2000). The lytic activity of viruses releases a significant amount of DOM which will influence the growth, abundance and structure of microbial populations. Although the abundance of virus like particles is generally correlated to the abundance of bacteria and phytoplankton, little is known how viral lysis affects the composition of the microbial community and re-distribution of DOM. To address this question, it is necessary to understand the composition of the viral community and compare that to bacterial community structure. In this project, paired metagenomic and metatranscriptomic analyses were conducted for viral and bacterial communities collected from the same water samples, on different spatial and temporal scales. The viral samples were collected from the bays in 2014 and 2015, with a total of 16 viral community samples selected for genome sequencing.

Thus, we present a set of dsDNA viral metagenomes from the Delaware Bay and Chesapeake Bay, which we refer to as the Delmarva Estuarine Virome (DEV).

2.2 Methods

2.2.1 Water sample collection

Ten water samples were collected from the Delaware Bay in March, August/September, and November, 2014, and six samples were collected from the Chesapeake Bay in April and August 2015, on board RV *Hugh R Sharp*. Samples were collected to reflect different salinity gradient in each estuarine ecosystem (Fig. 2.1). The overall sampling strategy was to collect viral communities across wide spatial and temporal scale in both estuaries. Additional information about environmental conditions can be found in Table 2.1 and Table S1. Samples DB8.2A and DB8.2B are diel samples; samples CB8.2S, CB8.2M and CB8.2D were taken at different depths (~1, 13 and 22 m, respectively).

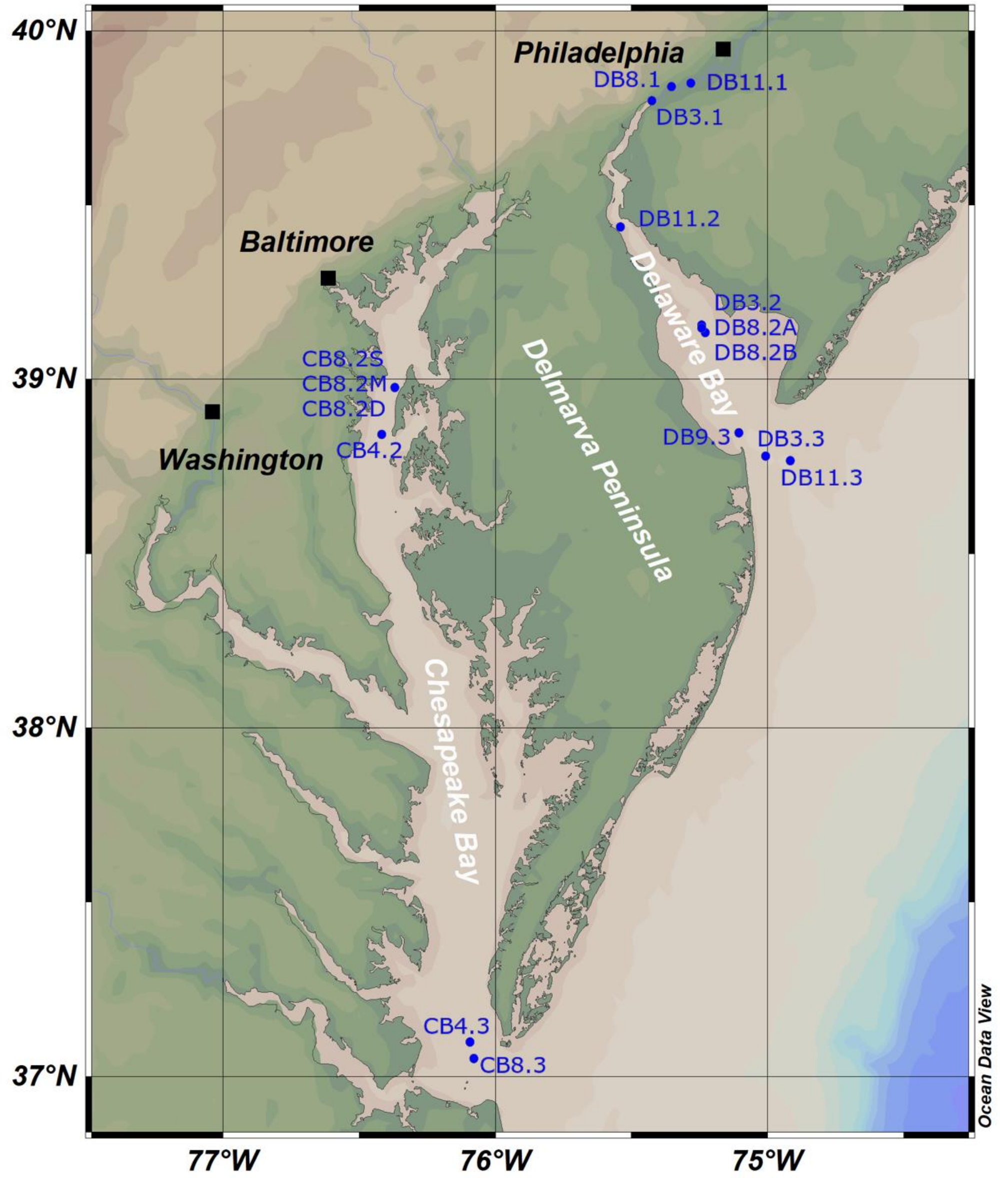

Figure 2.1 Sampling map for Delmarva Estuarine Virome (DEV) on the East coast of North America. The map was created using Ocean Data View (Schlitzer, R., Ocean Data View, https://odv.awi.de, 2019), with the ETOPO1 map (Amante and Eakins, 2009).

At each of the sampling sites, water samples were collected using a Niskin bottle on a Sealogger conductivity-temperature-depth rosette water sampler. For each sample, 10 L of seawater was prefiltered through 0.2 μm pore-size membrane filters (Millipore Corporation, Billerica, MA) to remove bacteria and larger organisms. Viral communities were concentrated from the 0.2μm filtrates following the $FeCl_3$ flocculation method described by John et al. (John *et al.*, 2011). Viral dsDNA was extracted using the phenol/chloroform/isoamylol method (Sambrook and Russell, 2006).

2.2.2 Viral and bacterial counts

For viral and bacterial counts, 2 mL seawater was fixed in a final concentration of 0.5% glutaraldehyde at 4 ° C for 20 min, then stored at 4 ° C. Viral and bacterial abundances were determined by Epics Altra II flow cytometer (Beckman Coulter, Miami, FL, USA) according to Brussaard (Brussaard, 2004). The fixed samples were stained with SYBR Green I (Invitrogen, CA, USA) and enumerated at event rates of 50 - 200 particles/s (bacteria) or 100 - 300 particles/s (viruses). For every sample, 10 μL of 1 mm-diameter fluorescent microspheres (Molecular Probes Inc., OR, USA) was added as reference beads. Each sample was run twice on the flow cytometer, and the average of count values was taken. The data were analyzed by EXPOTM_32 MultiCOMP software (Beckman Coulter, Miami, FL, USA).

2.2.3 DNA sequencing and metagenome assembly

Viral DNA was sequenced using Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) at the Joint Genome Institute, US Department of Energy, generating paired-end

(PE) reads with a read length of 150 bp. The resulting virome collection will be referred to as Delmarva Estuarine Virome (DEV). Known Illumina adapters were removed from sequencing reads and low-quality reads (Phred quality score < 12, containing more than 3 "N"s, or length under 51 bp) were trimmed with BBDuk (Bushnell, 2015). Remaining reads were mapped to a masked version of human HG19 with BBMap, discarding all hits over 93% identity, in order to remove genetic contamination during sample handling (Bushnell, 2015). Trimmed Illumina reads were *de novo* assembled with Megahit using a range of K-mers (D. Li *et al.*, 2016).

2.2.4 Viral contigs identification and annotation

Contigs that are likely to be of viral origin were selected using the method described in Paez-Espino *et al.* (Paez-Espino, Pavlopoulos, *et al.*, 2017). Briefly, contigs smaller than 5 kb were discarded, ORFs were predicted for the remaining contigs and were filtered based on the number of genes that they shared with known viral proteins. The resulting list of contigs were considered to be viral, and were uploaded to MG-RAST and annotated using the Refseq database (Keegan *et al.*, 2016). Rarefaction curves were generated by MG-RAST using data from the M5NR database and visualized using ggplot2 in R (Ginestet, 2011; Keegan *et al.*, 2016).

2.2.5 Viral contig cluster network

Viral contigs were clustered with BLASTN (e-value 1×10^{-50}, ≥90% identity, ≥75% covered length) using single linkage clustering (Paez-Espino, Pavlopoulos, *et al.*, 2017). Contigs not belonging to a cluster were deemed singletons. The clusters and their interaction with the samples they are associated with were visualized using the *prefuse force directed layout* in Cytoscape (Shannon *et al.*, 2003). Singletons were omitted from the cluster visualization for clarity.

2.2.6 Viral populations and detection of circular viral contigs

To reduce redundancy for read mapping analysis, for each viral cluster, the longest sequence within the cluster was deemed the seed sequence and was combined with the singletons to form a non-redundant viral population database. Circular viral contigs were detected using VRCA (ViRal and Circular content from metAgenomes), which finds circular contigs in metagenome assemblies by identifying read overlaps at the start/end of contigs (Crits-Christoph *et al.*, 2016). To examine chosen circular contigs of interest, a complete viral genome was reverse-complimented, annotated using RAST, and visualized using DNAplotter from Artemis (Carver *et al.*, 2009; Brettin *et al.*, 2015).

A summary of the sequence processing pipeline is given in Fig. 2.2.

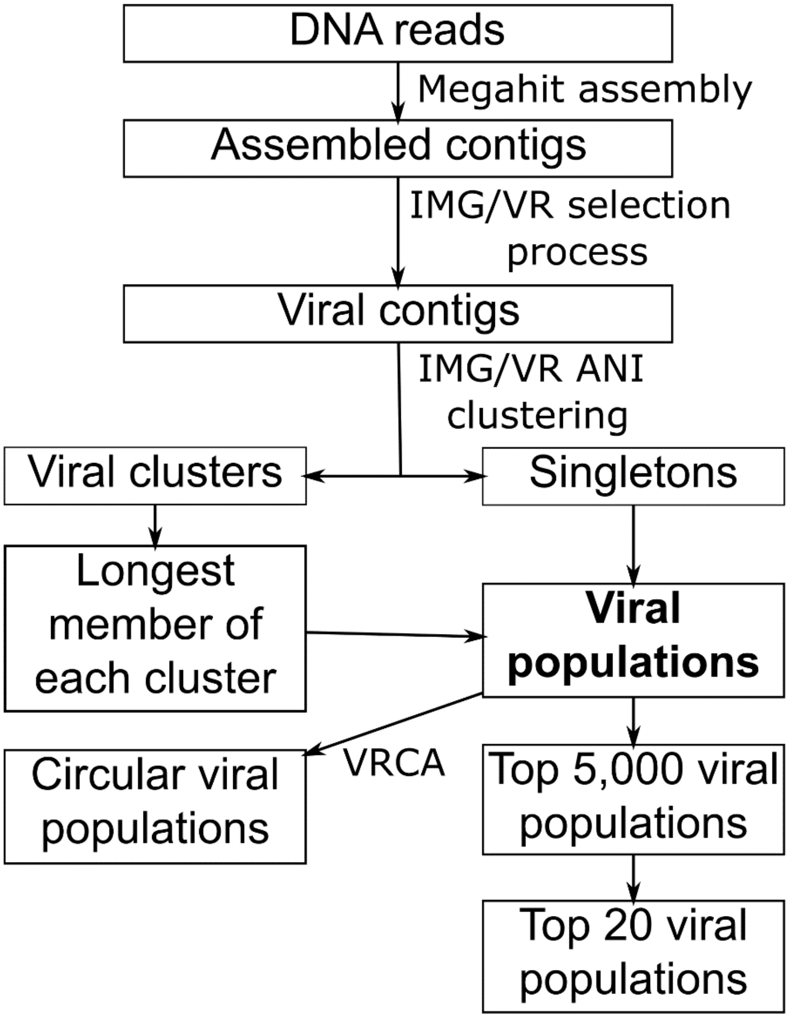

Figure 2.2 Diagram of viral sequence processing pipeline. For the number of reads and scaffolds in each sample, see Table 2.2. For the number of contigs, clusters and singletons in each sample, see Table 2.3.

2.3 Results

2.3.1 Sample collection

Sixteen virioplankton community samples were collected from Delaware and Chesapeake bays under a wide range of environmental conditions, with temperature ranging from 4.0 °C to 27.3 °C, and salinity ranging from 0.2 to 30.0 ppt (Table 2.1). In the Chesapeake Bay, samples from three different sampling depths were taken at station 8.2 in August, stratification in the water column can be seen from the salinity data (Table 2.1). The surface low salinity water contained higher concentrations of nitrate, chlorophyll a and bacterial count compared to the middle (13.3 m) and deep water (22.5 m) (Table S1). Fewer surface samples (n=4) were taken from the Chesapeake Bay than the Delaware Bay (n=10). No November samples were taken in the Chesapeake Bay.

Table 2.1 Sample site information.

Sample	**Year**	**Date**	**Temperature (°C)**	**Salinity (ppt)**
DB3.1	2014	19-Mar	4.4	0.2
DB3.2	2014	21-Mar	4.0	20.0
DB3.3	2014	22-Mar	4.0	30.4
DB8.1	2014	28-Aug	25.3	0.2
DB8.2A	2014	30-Aug	24.3	21.5
DB8.2B	2014	31-Aug	24.5	22.0
DB9.3	2014	1-Sep	24.3	28.8
DB11.1	2014	1-Nov	15.1	0.3
DB11.2	2014	2-Nov	13.8	15.4
DB11.3	2014	3-Nov	13.5	30.0
CB4.2	2015	12-Apr	8.5	9.1
CB4.3	2015	15-Apr	10.8	25.4
CB8.2S	2015	19-Aug	27.3	10.4
CB8.2M	2015	19-Aug	26.3	15.5
CB8.2D	2015	19-Aug	26.3	18.1
CB8.3	2015	22-Aug	26.6	26.7

2.3.2 Viral and bacterial counts

Bacterial cell counts range from 1.4×10^6 to 8.7×10^6 cells per ml, while viral counts range from 1.9×10^5 to 2.3×10^8 per ml, showing a much wider variance compared to bacterial counts. As expected, the viral concentration is lower in winter months compared to that of warmer seasons and is approximately 15-fold higher (ranging from 0.07 to 99.13, average 21.10) than the bacterial concentration (Fig. 2.3). In the Delaware Bay, viral and bacterial abundances remained consistent during the summer, and increased with the salinity gradient during the winter.

The abundance of viruses in the sea is around 15-fold higher than that of bacteria and archaea (Suttle, 2005), which matches our observations (Fig. 2.3). Other studies also found viral counts and cell counts to be positively correlated to temperature in the Chesapeake Bay, and observed stronger seasonal variation compared to spatial variation (Winget *et al.*, 2011).

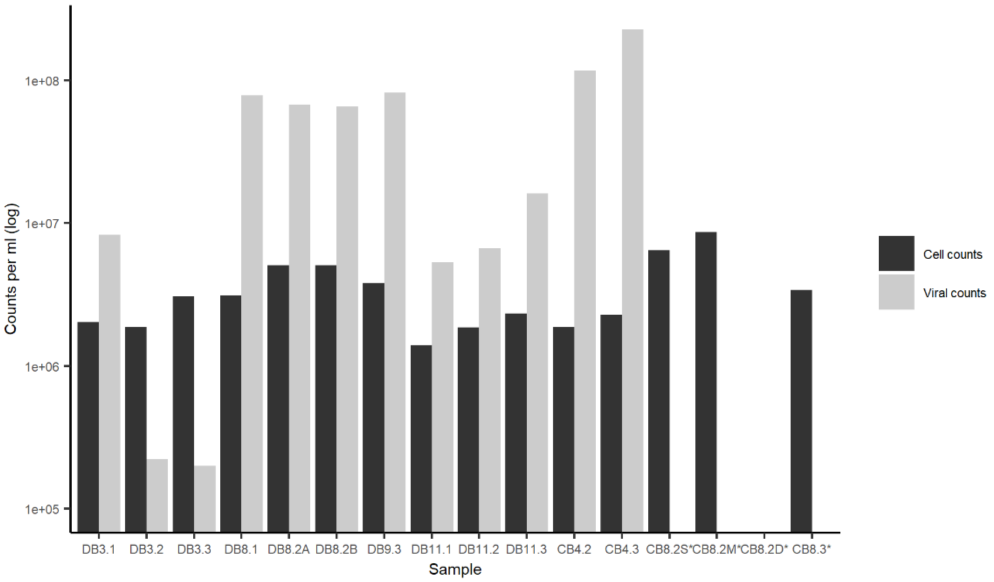

Figure 2.3 Bacterial and viral and count data in Delaware Bay (DB) and Chesapeake Bay (CB) determined by flow cytometric counting. Cells per milliliter and viral particles per milliliter are plotted on a logarithmic scale. *Cell counts for CB8.2D and viral counts for CB8.2S to CB8.3 are missing.

2.3.3 DNA sequencing and metagenome assembly

Illumina HiSeq sequencing of the 16 viral samples produced 1,924 billion reads (150 bp, paired-end) in total, which was named the Delmarva Estuarine Virome (DEV). The Delaware Bay samples yielded over twice as much sequencing depth as the Chesapeake Bay samples, with an average of 151 million reads for the Delaware Bay, and an average of 68 million reads for the Chesapeake Bay. An average of 690 Mbp worth of contigs were assembled per sample. An overview of sequencing and assembly results is shown in Table 2.2.

Table 2.2 Sequencing results for DEV samples.

Sample	# of reads (million)	Percentage of low quality reads (Q<12)	# of scaffolds (thousand)	Scaffold total size (MB)
DB3.1	135	1.20%	954	652
DB3.2	150	1.20%	1276	903
DB3.3	146	1.30%	899	595
DB8.1	124	1%	965	689
DB8.2A	120	1.20%	937	720
DB8.2B	140	1.30%	974	659
DB9.3	210	1.80%	1365	964
DB11.1	131	1.10%	827	590
DB11.2	218	1.50%	1816	1267
DB11.3	135	1%	1150	808
CB4.2	59	1%	658	509
CB4.3	64	0.60%	573	395
CB8.2S	66	0.80%	688	537
CB8.2M	68	0.60%	764	581
CB8.2D	87	1.20%	866	633
CB8.3	62	0.70%	690	536

2.3.4 Viral contigs identification and annotation

An average of 3,012 viral contigs were identified for each sample using the approach described in the IMG/VR database (Table 2.3) (Paez-Espino, Pavlopoulos, *et al.*, 2017; Paez-Espino *et al.*, 2019). Rarefaction curves showed that the sampling of DEV is close to saturation (Fig. 2.4).

Table 2.3 Number of viral clusters and singletons, and percentage of trimmed reads that map to viral populations.

Sample	Viral contigs	Unique clusters	Singletons	Mapped percentage
DB3.1	2666	2065	439	27%
DB3.2	4521	2960	1353	32%
DB3.3	2645	1066	1472	24%
DB8.1	2846	697	1026	24%
DB8.2A	3025	2307	536	26%
DB8.2B	2650	2070	419	27%
DB9.3	3909	2535	1137	27%
DB11.1	2119	1046	1000	18%
DB11.2	5374	2106	3115	25%
DB11.3	3910	2776	900	32%
CB4.2	2309	1608	623	24%
CB4.3	1770	2144	542	26%
CB8.2S	2661	1282	1173	28%
CB8.2M	2842	823	1968	30%
CB8.2D	2548	1102	1491	25%
CB8.3	2395	1647	651	24%

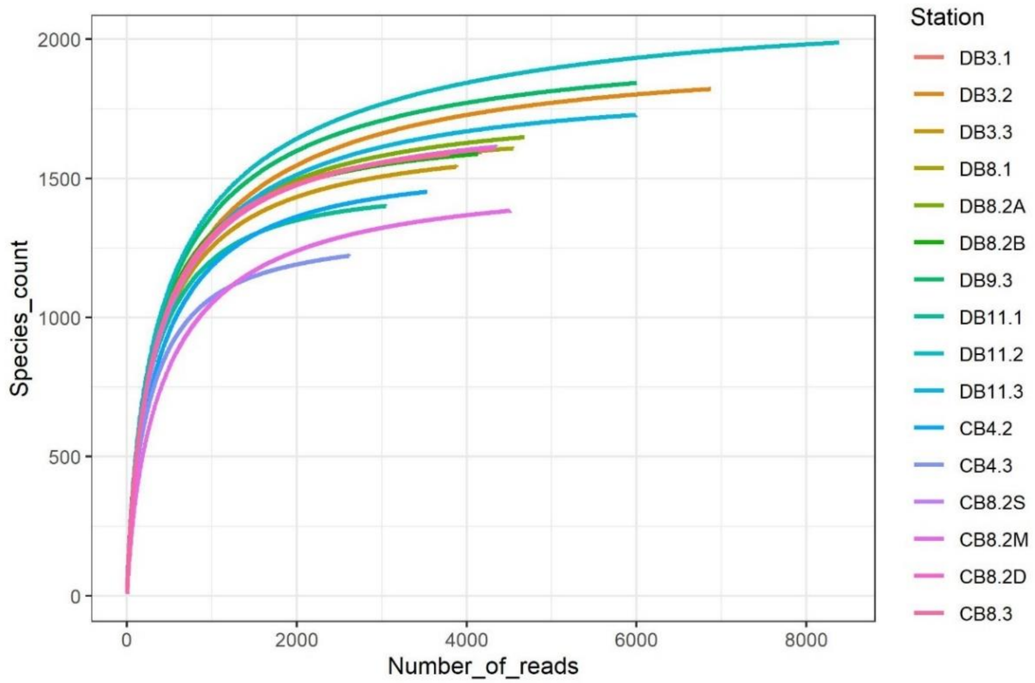

Figure 2.4 Rarefaction curves of each sample. Rarefaction curves were produced using data from the M5NR database, representing species data of taxonomic categories from 16 viral metagenomes. The cutoffs used were: alignment length: 15bp; e-value: e-5; percent identity: 60%.

2.3.5 Viral contig cluster network

To explore the diversity of contigs recovered from the DEV samples, we classified viral contigs into clusters and singletons based on the sequence similarity. A cluster is a group of DEV contigs (at least two contigs) that share high sequence similarity, while a singleton is a contig that does not belong to a cluster. From the 48,190 viral contigs (16 samples combined), 9,204 viral clusters and 17,845 singletons were detected. The number of clusters for each sample ranged from 697 to 2,960, while the number of singletons ranged from 419 to 3,115, reflecting a large number of viral contigs that are unique to their sample (Table 2.3). Sample DB11.2 produced the largest amount of viral contigs (2,106), and also the largest amount of singletons (3,115), suggesting the presence of a rich mid-bay viral diversity not found elsewhere (Table 2.3).

A bipartite network was used to visualize the association between samples and clusters (Fig. 2.5). Delaware Bay summer samples share relatively more clusters with each other. Chesapeake Bay samples cluster distinctly from Delaware Bay samples, and appear to show less similarity between them. Strangely, the two samples DB3.3 and DB11.1 were grouped together and away from the other samples, despite having little in common (Fig. 2.5).

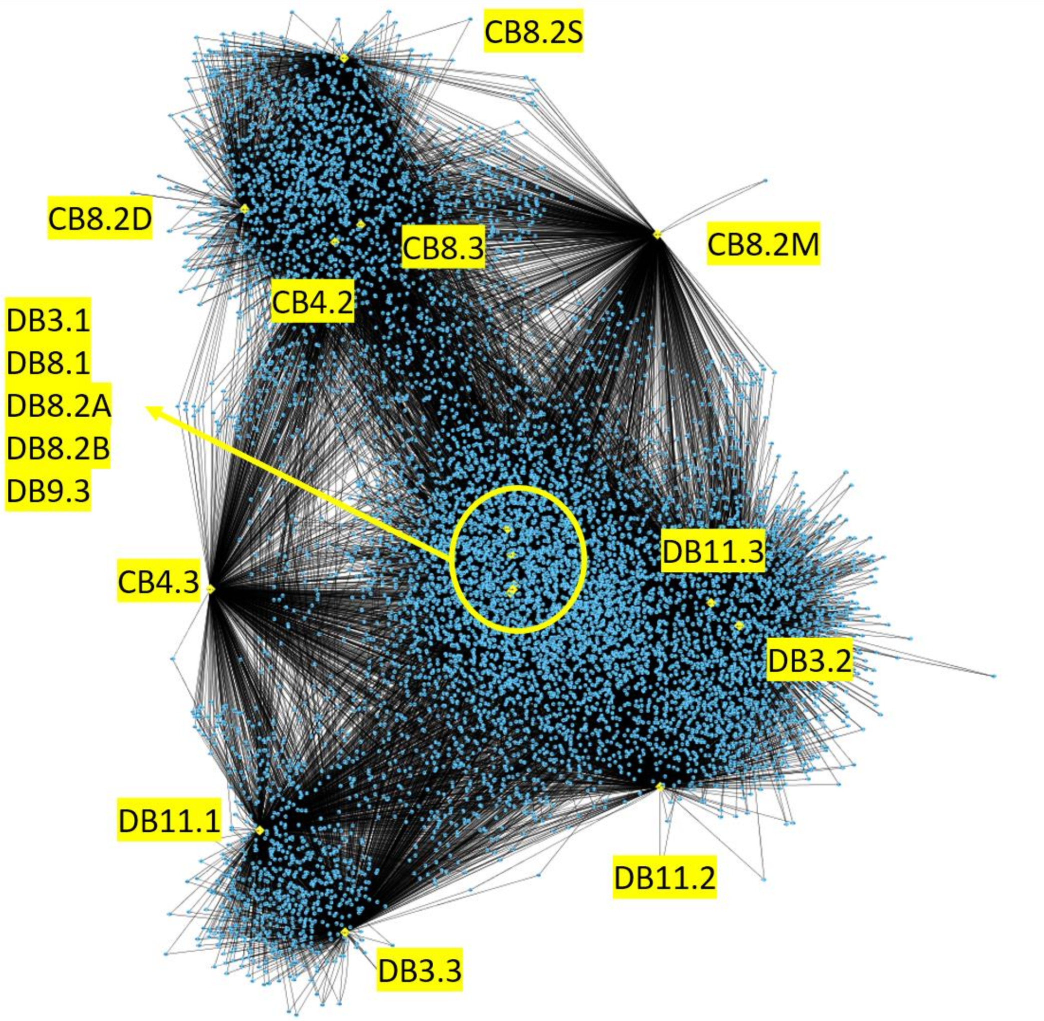

Figure 2.5 Cluster network of viral clusters and samples visualized using Cytoscape. Yellow nodes represent sampling stations; blue nodes represent viral clusters; edges (black lines connecting the nodes) represent their association. Singletons were omitted from the visualization for clarity.

2.3.6 Viral populations and detection of circular viral contigs

By combining the viral cluster and singleton information, a total of 26,487 viral populations were identified in the DEV samples (Table 2.4). The term “viral populations” is a commonly defined term in viral ecology, the equivalent of (Brum *et al.*, 2015). An average of 26.2% trimmed reads mapped to viral populations in each sample (Table 2.3), indicating that nearly three quarters of sequencing reads were not identified as viral at the current setting. Among the viral populations, 319 circular viral genomes were predicted via sequence overlaps. The length of circular viral genomes ranged from 7.5 kb to 161.8 kb, and they were mostly present in low abundance (average fpkm ≤ 20) with one exception (Ga0070751_1000196). Further details about said population can be found in Appendix B.

Since the viral contigs are assembled from short reads (150 bp), there is a limited amount of complete or near-complete viral genomes, so it is likely that the number of singletons is overestimated when different portions of the same viral genome are not clustered together and instead broken into multiple contigs. This would overestimate viral populations, but the same bias applies to all of the samples, so the cross-comparison between DEV samples should not be affected as much. Our clustering method, although not ideal, is nonetheless widely used by viral ecology researchers and for applications including generating the IMG/VR database (Paez-Espino, Chen, *et al.*, 2017). As is mentioned in the IMG/VR paper, the advantage of this method over tools such as Virsorter and vContact, is its non-targeted nature and ability to detect highly divergent viral sequences, which we consider complimentary to the high sequencing depth and quality of our dataset. Therefore, we consider this clustering method to be adequate in our situation.

Table 2.4 Length distribution of viral populations.

Viral populations > 5 kb	26,487
Viral populations > 10 kb	12,531
Viral populations > 25 kb	2,523
Viral populations > 50 kb	353
Total length (bp)	346,627,485
Largest contig (bp)	186,740
N50 (bp)	15,181
N75 (bp)	9,232
L50	6,505
L75	13,896

2.4 Significance of the Delmarva Estuarine Virome (DEV)

This study revealed the diversity of the dsDNA virioplankton community in the Delaware Bay and Chesapeake Bay using high-throughput sequencing. This is the first systematic study of the spatial and temporal variation of the viral community in estuarine habitats using deep metagenomics. Previously, the virioplankton community structure in the Chesapeake Bay has been studied by sequence analysis of one sample pooled from nine different locations in the bay, which was the first metagenomics attempt to study the estuarine virioplankton (Bench *et al.*, 2007). However, the metagenomic sample was sequenced using Sanger technology and thus it could not provide sufficient sequencing coverage for an in-depth assessment of the viral community structure.

The DEV produced 288 Gb of sequencing data, which is a ~74,000 fold increase over the 3.92 Mb produced by the previous Chesapeake Bay virome (Bench *et al.*, 2007), reflecting the vast improvement in sequencing technology over the years. Deep sequencing generated a total of 48,190 assembled viral sequences (>5kb) and 26,487 viral populations (9,204 virus clusters and 17,845 singletons), including 319 circular viral contigs between 7.5 kb to 161.8 kb. Sequencing information of our DEV samples was compared to several major viromic datasets, which include Pacific Ocean Virome, Tara Ocean Virome, Malaspina Virome, and Tara Oceans Polar Circle Virome from GOV 2.0 (Table 2.5). Compared to other recent marine viral metagenomic datasets, the DEV returned similar sequencing quality and its sequence processing methods are in accordance with current standards. It is one of the highest quality viral metagenomic datasets to date, showing remarkably consistent sequencing depth and quality across samples, allowing us to discover the above patterns.

Table 2.5 Comparison of recent marine viral metagenomic datasets. Abbreviations: Pacific Ocean Virome (POV); Tara Ocean Virome (TOV); Tara Oceans polar circle (TOPC); Global Ocean Virome (GOV); Delmarva Estuarine Virome (DEV). GOV 2.0 consists of TOV, Malaspina and TOPC.

	POV (2013)	**TOV (2015)**	**Malaspina (2016)**	**TOPC (2019)**	**DEV (This study)**
# of metagenomes	32	43	14	41	16
Average # of reads per metagenome	188,128	100,706,767	28,334,677	53,500,000	120,278,861
Read length (bp)	310	101	151	101	151
Sequencing platform	454 Titanium	Illumina Hiseq 2000	Illumina Hiseq	Illumina Hiseq 2000	Illumina HiSeq 2500 1TB
Average # of contigs per sample	NA	88,878 (SOAP denovo)	Unknown	Unknown	962,521 (Megahit)
Average # of viral contigs per sample	NA	5,852 (GOV 2.0)			3,012

Chapter 3: Uncultivated viral populations dominate estuarine viromes on the spatiotemporal scale

3.1 Abstract

Viruses are ubiquitous and abundant in the oceans, and viral metagenomes (viromes) have been investigated extensively via several large-scale ocean sequencing projects. However, there has not been any systematic viromic studies in estuaries. Here, we investigated viromes of the Delaware and Chesapeake Bay, two Mid-Atlantic estuaries. Unknown viruses represented the vast majority of the dominant populations, while the composition of known viruses, such as pelagiphage and cyanophage, appeared to be relatively consistent across a wide salinity gradient and in 3 different seasons. A difference between estuarine and ocean viromes was reflected by the proportions of *Myoviridae*, *Podoviridae*, *Siphoviridae*, *Phycodnaviridae*, and a few well-studied virus representatives. The difference between the viral community in the Delaware and Chesapeake Bays is significantly ($P < 0.05$) more pronounced than the differences caused by temperature or salinity, indicating strong local profiles caused by the unique ecology of each estuary. Highly abundant viruses (top 20) in both estuaries have close hits to viral sequences derived from the marine single cell genomes or long read single molecule sequencing, suggesting that important viruses are still waiting to be discovered in the estuarine environment.

3.2 Introduction

Estuaries are vital links between marine and terrestrial ecosystems, and are among the most productive ecosystems on the planet (Field *et al.*, 1998). Estuarine systems encompass a complex spectrum of environmental gradients, creating distinct microbial habitats, and the frequent fluctuation of environmental conditions cause unique selective pressures to be exerted on organisms (Fortunato and Crump, 2011). In a highly dynamic estuarine environment, changes in environmental factors can trigger genetic and ecological shifts in microbial communities (Herbert, 1999). Compared to coastal marine and river waters, bacterial densities and growth rates in estuaries are generally higher, and tend to be highest in surface waters and turbid regions (Wright and Coffin, 1983). The bacterioplankton community in the Chesapeake estuary exhibit a strong and repeatable seasonal pattern, but less variation across the spatial scale (Kan *et al.*, 2006, 2007). Virioplankton are usually one order of magnitude more abundant than bacterioplankton (Wommack and Colwell, 2000). The abundance of virioplankton in the Chesapeake Bay is in the range of 10^6-10^8 viral like particles (VLPs) per milliliter (Bergh *et al.*, 1989; Wommack *et al.*, 1992), which can be 10-1,000 times more abundant than the viral concentration in the open ocean (Wommack and Colwell, 2000). Virioplankton are an active and dynamic component of estuarine microbiomes, and are responsive to changes in environmental factors and the bacterial community (Wommack *et al.*, 1999; Bench *et al.*, 2007; Cissoko *et al.*, 2008). They are an important part of the trophic system in estuaries as they are responsible for bacterial mortality at a level similar to protist grazing (Wommack *et al.*, 1992; Fuhrman and Noble, 1995).

In the past ten years, the development of new sequencing technologies has greatly advanced our understanding on microbial diversity in nature. Using next-generation sequencing (NGS) technologies, a number of large-scale ocean sequencing projects (e.g. Global Ocean Sampling Expedition, Malaspina Expedition, Pacific Ocean Virome, Tara Ocean's Global Ocean Virome (GOV)) have made viral metagenomic databases increasingly accessible, revealing important findings about the diversity, spatial and temporal distribution of ocean viruses (Williamson *et al.*, 2008; Hurwitz and Sullivan, 2013; Brum *et al.*, 2015; Roux *et al.*, 2016, Gregory *et al*. 2019). The most recent study, Tara Ocean's GOV 2.0, shows that marine viral communities can be separated into five ecological zones, although no estuarine samples were included (Gregory *et al.*, 2019). Meanwhile, many viromic studies have shown that the most abundant viral species in the ocean still remain unassigned to known viral taxa (Paez-Espino *et al.*, 2016; Roux *et al.*, 2016). Large-scale sequencing efforts generally only include a few sampling sites at coastal and brackish locations (Bench *et al.*, 2007; Rusch *et al.*, 2007; McDaniel *et al.*, 2008; Williamson *et al.*, 2008; Cai *et al.*, 2016; Hwang *et al.*, 2016; Zeigler Allen *et al.*, 2017), but there has not been any systematic study of spatial and temporal variation of virus communities in dynamic estuarine environments using deep sequencing technology (Table 1.1).

In this study, sixteen virioplankton samples were collected from the Delaware Bay and Chesapeake Bay from low (0.2 - 0.3 ppt), medium (9.1 – 22.0 ppt) and high (25.4 - 30.4 ppt) salinity sites during three different seasons. High throughput sequencing with deep sequencing coverage of these estuarine samples enabled us to analyze the spatiotemporal variation of viral community in the two large estuarine ecosystems.

3.3 Methods

3.3.1 Relative abundance of viral populations and relationship with environmental variables

Quality trimmed DNA reads were mapped to the non-redundant viral populations using BBMap with the mapping parameters as recommended in viromic benchmarking studies (>90% identity, >75% contig length) (Bushnell, 2014; Roux *et al.*, 2017). Reads were counted and normalized to FPKM (Fragments Per Kilobase Million) using SAMtools (Li *et al.*, 2009). FPKM is commonly used as a proxy for relative abundance in viral community studies (Roux *et al.*, 2017). Total FPKM of each sample was added together for each viral population and ranked to find the most abundant viral populations.

To explore similarity of samples based on viral population profiles, a non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarity matrices was plotted using the vegan package in R and visualized using ggplot2 (Ginestet, 2011; Oksanen *et al.*, 2018). Due to computing constraints, only the most abundant 5,000 (out of 26,487) viral populations (mean FPKM > 3.87) were used for this analysis. To further quantify the similarity of viral population profiles across different groups of samples, analysis of variance (ANOSIM) test was performed with the same 5,000 viral populations using the vegan package in R (Oksanen *et al.*, 2018).

The top 20 most abundant viral populations were chosen to represent the dominant viruses in the estuaries, and their abundance was plotted using ggplot2 in R (Ginestet, 2011). To identify the top 20 viral populations, they were searched against the NCBI-nr database with BLASTN (Altschul *et al.*, 1990). To further explain the relationship between the abundance of dominant viruses and environmental variables, redundancy analysis (RDA) was plotted for the top 20 viruses using the vegan package in R, and visualized using type I scaling in ggplot2 (Ginestet, 2011; Oksanen *et al.*, 2018).

3.3.2 Host prediction

Putative hosts were predicted *in silico* by comparison of viral populations to known CRISPR (clustered regularly interspaced short palindromic repeats) spacers. The collection of CRISPR spacers from the Microbial Isolate Genomes from the IMG/M database was used as a blastn query against all of the viral populations, and hits were used if they were 100% length, allowing a max of 1 mismatch (Altschul *et al.*, 1990). The resulting virus-host pairings were sorted according to the total relative abundance (FPKM) of the viral populations.

3.3.3 Viral taxonomy of DEV reads and relationship with environmental variables

The analysis of known viral taxonomy was handled separately from that of abundant viral populations, in order to get a comprehensive picture of both the classified viruses and the viral “dark matter” in the estuaries. To acquire the taxonomy of known viruses, trimmed reads were classified using Kaiju (Menzel *et al.*, 2016), and taxonomy was assigned via comparison with Kaiju’s built-in “viruses” database (as of June 2019), using the default greedy mode parameters. A classification summary was created using the kaiju2table program, and percentages of reads for each taxon were used as a proxy for species relative abundance. The abundance of species with percentage greater than 0.1% in DEV were plotted using ggplot2 in R (Ginestet, 2011). These species were categorized according to the host they are presumed to infect, derived from the species name, and may not reflect their ability to infect other potential hosts. The category “Cyanophage” may include *Prochlorococcus* and *Synechococcus* phages. All species were categorized according to family, and the top 4 most abundant viral families were plotted.

To explain the relationship between abundant species and environmental variables, redundancy analysis (RDA) was plotted for species in DEV with greater percentage than 0.1% in DEV using the vegan package in R, and visualized using type I scaling in ggplot2 (Ginestet, 2011; Oksanen *et al.*, 2018).

3.3.4 Comparison of viral taxonomy with oceanic samples

To compare the viral composition of estuarine and open ocean waters, the metagenomic reads of 4 publicly available oceanic surface water samples were downloaded and assigned taxonomy with Kaiju, using the above methods (Aylward *et al.*, 2017; Gregory *et al.*, 2019). The viral metagenomic samples (from TARA Oceans, Hawaii Ocean Experiment) were chosen due to their similar sequencing technology and depth, and their wide global distribution (Fig. 3.1, Table 3.1).

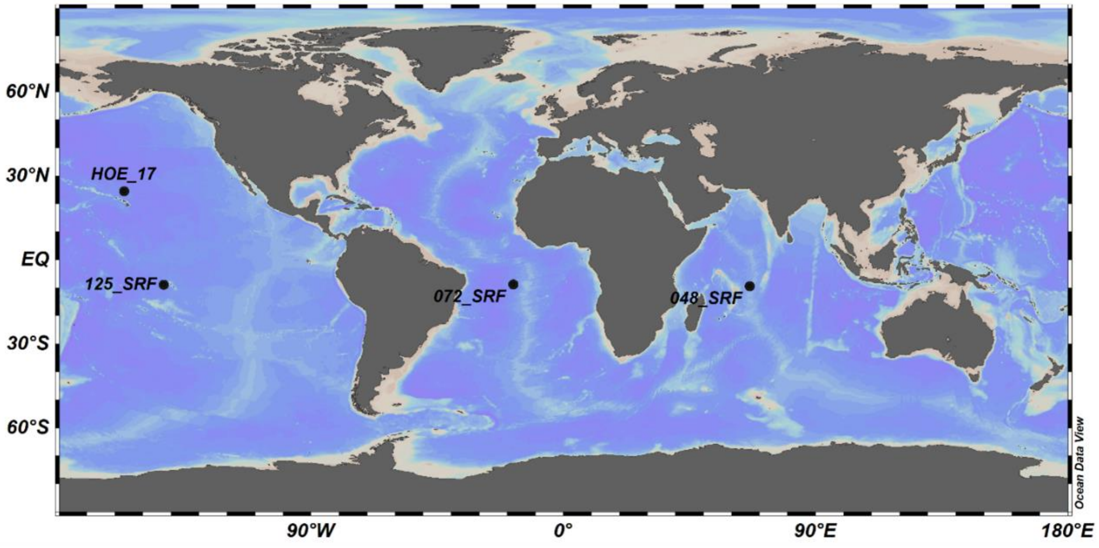

Figure 3.1 Map of oceanic samples used in viral taxonomy analysis. The map was created using Ocean Data View (Schlitzer, R., Ocean Data View, https://odv.awi.de, 2019).

Table 3.1 Sampling Conditions of oceanic samples used in viral taxonomy analysis.

Dataset	Sample	Date	Local Time	Region	Temp (°C)	Salinity (ppt)
GOV 2.0	125_SRF	2011-08-08	17:33	Mid-Pacific	27	35
GOV 2.0	072_SRF	2010-10-05	08:00	Mid-Atlantic	25	36
GOV 2.0	048_SRF	2010-04-19	07:56	Indian Ocean	30	34
HOE	HOE_17	2015-07-27	10:05	Pacific (Hawaii)	27	35

3.4 Results

3.4.1 Spatiotemporal distribution of abundant viral populations

The relative distribution frequency of the top 20 most abundant viral populations (recruiting 4.6% of all reads) in these 16 estuarine samples were compared (Fig. 3.2). In the Delaware Bay, abundance variation in summer samples appears to be more consistent across the salinity gradient than that of spring or fall samples (Fig. 3.2). The relative abundances of these top 20 viral populations seem to be more variable in the Delaware Bay than in the Chesapeake Bay (Fig. 3.2).

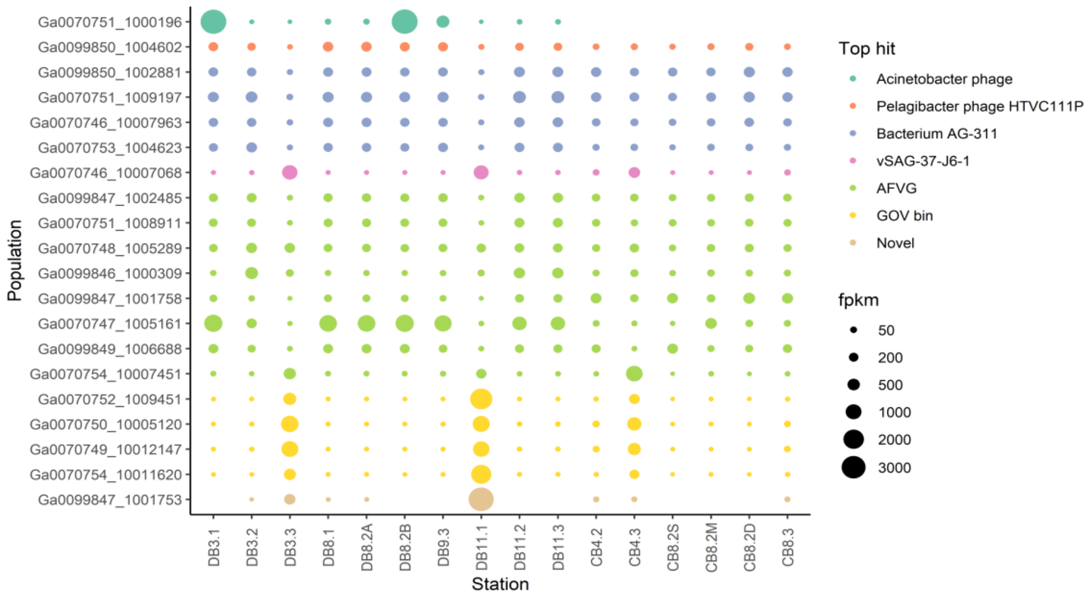

Figure 3.2 Relative abundance bubble plot of the top 20 most abundant viral populations for all 16 samples. Size of the bubbles correspond to the FPKM (Fragments Per Kilobase Million) for each sample, colors correspond to top BLAST hit of the viral population.

When comparing the most abundant viruses against other known sequences using BLASTN search against the NCBI-nr database, they were mostly found to share the closest similarity to other viral metagenomic sequences, or prokaryotes discovered using non culture-based methods such as single cell genomics and single-molecule sequencing (Table 3.2). Of the top 20 abundant virus populations, four shared the closest similarity to *Bacterium* AG-311-K16, a marine cyanobacteria isolated using single cell technology (Berube *et al.*, 2018); one shared the closest similarity with vSAG 37-J6, a virus discovered using single-virus genomics (Martinez-Hernandez *et al.*, 2017); eight matched viral sequences derived from assembly-free single-molecule sequencing (Beaulaurier *et al.*, 2020); four matched uncultured viral populations from GOV (Gregory *et al.*, 2019); and one was completely novel. The only two readily identifiable cultured virus isolates in the top 20 were the putative *Acinetobacter phage* (Ga0070751_1000196) and *Pelagibacter phage* HTVC111P (Ga0099850_1004602). The putative *Acinetobacter* phage was found to be highly abundant in several Delaware Bay samples (the most abundant population in DB3.1 and DB8.2B) but was not present in Chesapeake Bay samples. More information about this population is provided in Appendix B. In addition, a diel variation was noticed in DB8.2A and DB8.2B samples.

Table 3.2 Nucleotide BLAST results of top 20 abundant viral populations against nr database. FPKM: fragments per kilobase million.

Viral population	Length	Total fpkm	Top hit	Query cover	E value	Id
Ga0070747_1005161	5953	9474	Marine virus AFVG_25M393	4%	3.00E-35	75%
Ga0070751_1000196	42033	7894	Acinetobacter phage vB_AbaP_Acibel007	47%	0.00E+00	73%
Ga0070751_1009197	5120	4862	Bacterium AG-311-K16 Ga0172223_11	90%	0.00E+00	80%
Ga0099847_1001753	7593	3814	None			
Ga0099850_1002881	8091	3508	Bacterium AG-311-K16 Ga0172223_11	90%	0.00E+00	77%
Ga0070750_10005120	7119	3497	Prokaryotic dsDNA virus sp. isolate GOV_bin_15	54%	0.00E+00	73%
Ga0070752_1009451	5331	3343	Prokaryotic dsDNA virus sp. isolate Tp1_138_SUR_25606_1	65%	6.00E-164	71%
Ga0070749_10012147	5544	3042	Prokaryotic dsDNA virus sp. isolate GOV_bin_3107	3%	3.00E-29	76%
Ga0070748_1005289	5790	2875	Marine virus AFVG_117M37	86%	0.00E+00	77%
Ga0070746_10007963	6108	2797	Bacterium AG-311-K16 Ga0172223_11	58%	0.00E+00	80%
Ga0070754_10011620	5489	2618	Prokaryotic dsDNA virus sp. isolate GOV_bin_2950	39%	3.00E-123	70%
Ga0099847_1001758	7589	2580	Marine virus AFVG_117M42	97%	0.00E+00	75%
Ga0099847_1002485	6383	2551	Marine virus AFVG_117M61	39%	0.00E+00	74%
Ga0099849_1006688	5235	2485	Marine virus AFVG_25M322	100%	0.00E+00	80%
Ga0070746_10007068	6491	2343	Uncultured virus clone vSAG-37-J6-1	57%	0.00E+00	70%
Ga0070753_1004623	6993	2269	Bacterium AG-311-K16 Ga0172223_13	39%	0.00E+00	80%
Ga0070751_1008911	5219	2166	Marine virus AFVG_117M42	56%	0.00E+00	78%
Ga0099846_1000309	20226	2129	Marine virus AFVG_25M87	43%	0.00E+00	83%
Ga0099850_1004602	6449	2127	Pelagibacter phage HTVC111P	86%	0.00E+00	78%
Ga0070754_10007451	7156	2077	Marine virus AFVG_25M13	52%	0.00E+00	71%

The top 5,000 most abundant virus populations were used to evaluate the similarity between different samples. NMDS ordination shows that the 16 viromes were clustered according to their bay of origin (Fig. 3.3). Delaware Bay summer samples clustered together, but otherwise, samples generally did not cluster according to season or salinity (Fig. 3.3). This is further confirmed by an ANOSIM test; dissimilarity between groups was only significant when grouping samples by bay of origin (Table 3.3). Inexplicably, samples DB3.1 and DB11.1 clustered together and away from other samples, the two of them showing significant dissimilarity against other samples (Fig. 3.3, Table 3.3).

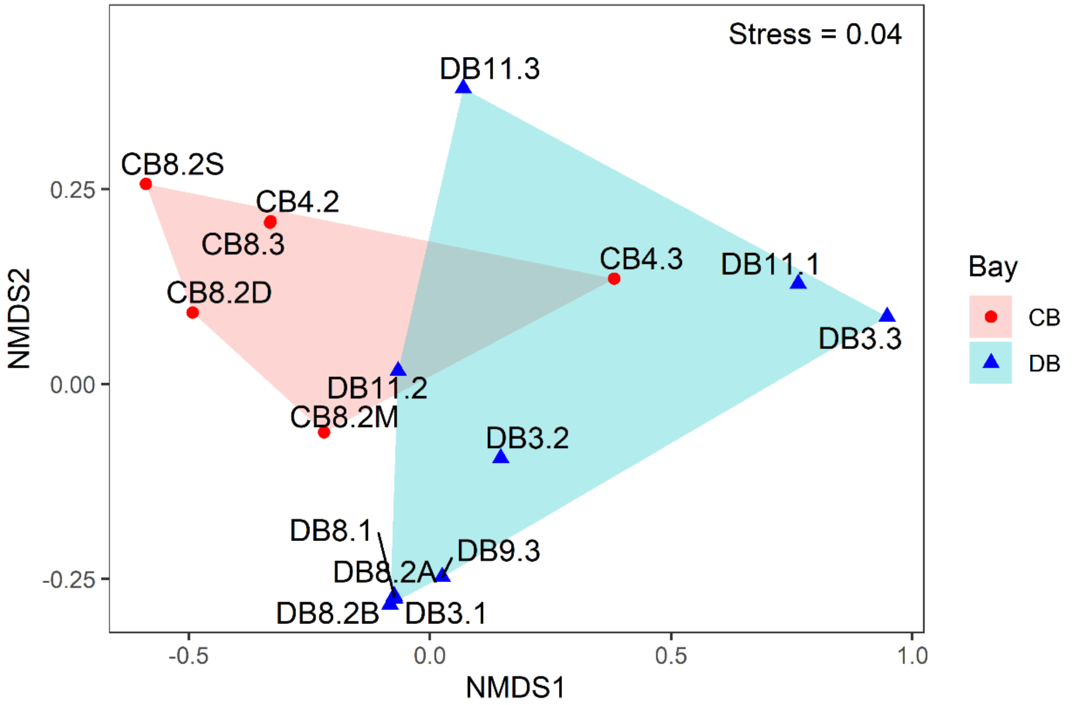

Figure 3.3 Non-metric multidimensional scaling (NMDS) plot made from top 5,000 most abundant viral populations. Stress level is indicated. DB: Delaware Bay; CB: Chesapeake Bay. Convex hulls are plotted around samples of each bay.

Table 3.3 Analysis of similarity (ANOSIM) test based on top 5,000 most abundant viral populations (*P < 0.05).

Grouping by	***R***	***P***
Delaware vs. Chesapeake Bay	0.35	0.0168*
Delaware vs. Chesapeake Bay (without CB8.2S and CB8.2M)	0.3422	0.0348*
Seasons	0.1905	0.087
Temperature (>20°C vs. <20°C)	0.1624	0.0566
Location (upper vs. mid vs. lower bay)	0.1667	0.1059
DB3.3 & DB11.1 vs. other samples	0.9519	8e-04*

Redundancy analysis (RDA) revealed the putative *Acinetobacter* phage (Ga0070751_1000196) and the most abundant viral population (Ga0070747_1005161) to be outliers with regard to their relationship with environmental parameters (Fig. 3.4). Their variance is not significantly ($P < 0.05$) correlated with Chl.a concentrations, despite what the RDA figure may suggest.

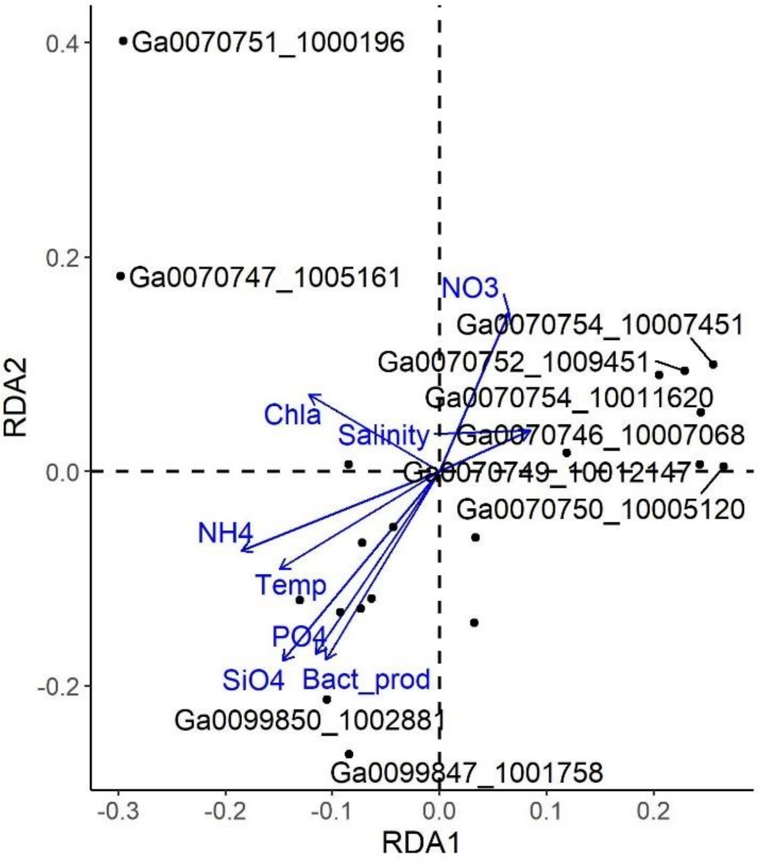

Figure 3.4 Redundancy analysis (RDA) ordination diagram (biplot) of top 20 viral populations (black) and environmental variables (blue). RDA1 explains 9.2% of variance, while RDA2 explains 6.5% of variance. Labels of data points below 0.15 have been omitted for clarity. The angles between populations and environmental factors denote their degree of correlation.

3.4.2 Host prediction

Putative hosts were able to be predicted for 102 out of 26,487 viral populations based on shared CRISPR spacers (Table S2). The relative abundances of these viral populations are low, all ranking below the top 3,000, consisting only 0.1% of the total FPKM of viral populations. Their predicted hosts also tend to be prokaryotes of low abundance.

3.4.3 Read-based viral taxonomy of DEV

Since the majority of sequences are unable to be connected to known viral taxa, separate analyses were conducted for reads assigned to known viruses, and viral contigs in general. The following 4 figures (Fig. 3.5 to Fig. 3.8) show the results of read level classification by Kaiju. Kaiju assigned ca. 10% (7.2% to 16.9%) of trimmed reads to known viruses in all the DEV samples except for CB8.2M (Fig. 3.5, Fig. 3.6). The proportion of reads matching representative viral groups (*Acinetobacter* phage, *Puniceispirillum* phage, *Pelagibacter* phage, *Synechococcus* phage, *Prochlorococcus* phage, unknown cyanophage) is markedly lower in samples DB3.3 and DB11.1 (Fig. 3.6). Viruses infecting other hosts were omitted from Fig. 3.6 due to low abundance (< 0.05%). When the overall abundance pattern of Kaiju-determined known viruses (Fig. 3.7) is compared to the top 20 viral populations (Fig. 3.2), the variation of known viruses seem to be less dramatic.

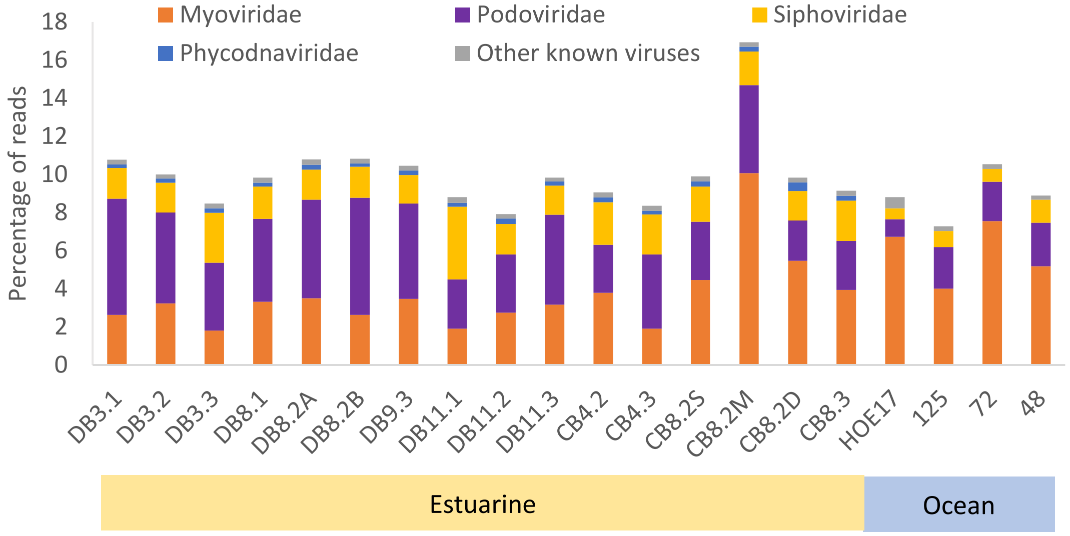

Figure 3.5 Relative abundance of main viral families, from categorization of known viruses by Kaiju read classification. The last four samples are oceanic; sample information can be found in Fig. 3.1 and Table 3.1.

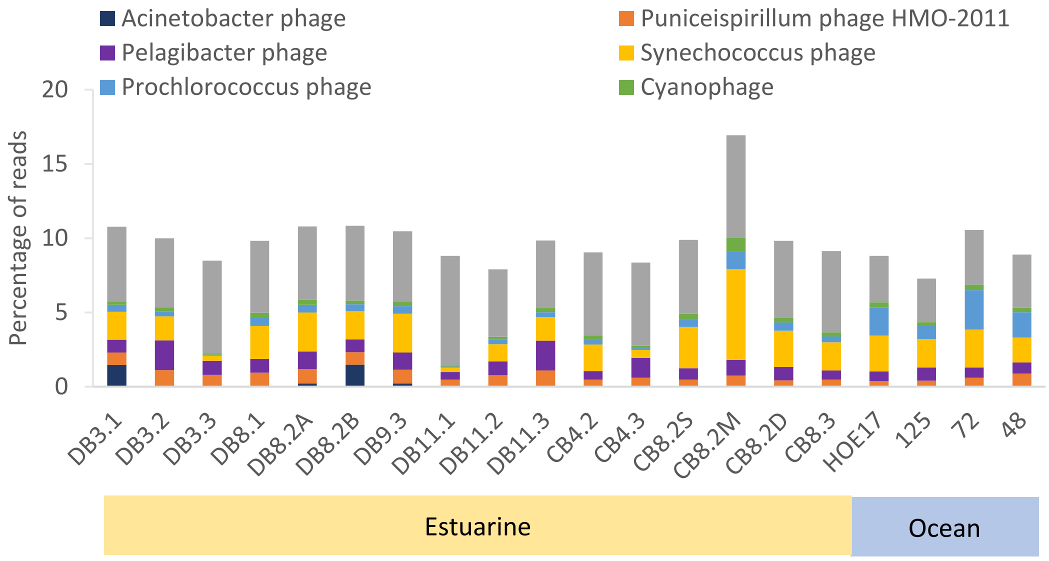

Figure 3.6. Relative abundance of viral species categorized by presumed host, from categorization of known viruses by Kaiju read classification. "Cyanophage" may include *Prochlorococcus* and *Synechococcus* phages. Groups of viral species assigned a certain host with low abundance (< 0.05%) were omitted. The last four samples are oceanic; sample information can be found in Fig. 3.1 and Table 3.1.

At the family level, the majority of reads (93% ~ 98%) were assigned to *Caudovirales*, with a lower proportion of *Siphoviridae* compared to the other two families (Fig. 3.5). Viral taxonomy at the family level is relatively stable across different samples, although the Chesapeake Bay appears to have a higher relative abundance of *Myoviridae* compared to the Delaware Bay. Sample CB8.2M showed an especially high proportion of myoviruses, and DB11.1 showed a relatively higher proportion of *Siphoviridae* (Fig. 3.5).

When categorizing the viruses by the host they are presumed to infect, cyanophages were found to be prevalent in the estuaries and more abundant during warmer seasons (Fig. 3.7, Fig. 3.5). The CB8.2M sample shows a large number of *Synechococcus* phages (Fig. 3.6). The most abundant cyanophages in the DEV tend to be related to those isolated from the North Atlantic Ocean or the Chesapeake Bay (Fig. 3.7). A small fraction (<1%) of *Prochlorococcus* phage sequences were present in almost all estuarine samples (Fig. 3.6). *Pelagibacter* phages and *Puniceispirillum* phages consist a large proportion of reads (up to 3%) (Fig. 3.6), but do not show strong variation patterns throughout different samples, despite strong salinity gradients (Table 2.1, Fig. 3.7).

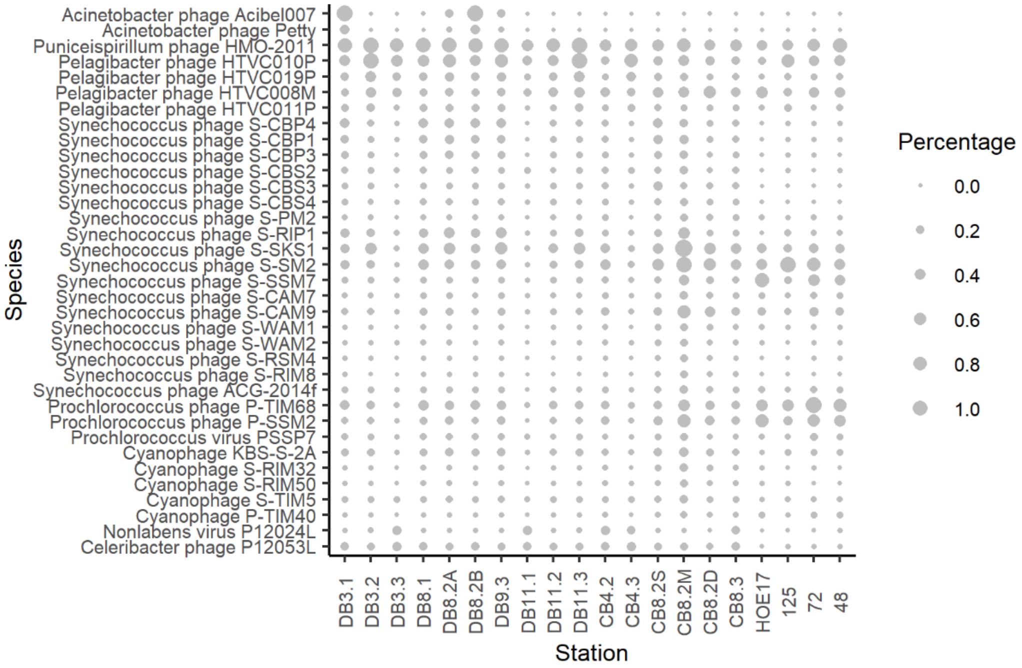

Figure 3.7 Bubble plot of most known abundant viral species (greater than 0.1% reads) in DEV, derived from taxonomy of known viruses by Kaiju read classification. Size of bubbles corresponds to the percentage of reads that are binned to the virus species. The last four samples are oceanic; sample information can be found in Fig. 3.1 and Table 3.1.

Redundancy analysis (RDA) indicated the degree of correlation between abundant viral species and environmental factors. As expected, viruses are generally grouped according to their putative hosts, with all cyanophages, pelagiphages and *Acinetobacter* phages clustered near each other on the biplot (Fig. 3.8). *Acinetobacter* phages are outliers compared to other abundant species in terms of their relationship with environmental variables, and are positively correlated with chlorophyll *a* concentration. *Pelagibacter* phages and *Puniceispirillum* phages exhibited a positive correlation with salinity, while cyanophage presented a positive correlation with temperature, NH_4^+, SiO_4^- and PO_4^{3-} concentrations, and a negative correlation with NO_3^- concentrations (Fig. 3.8).

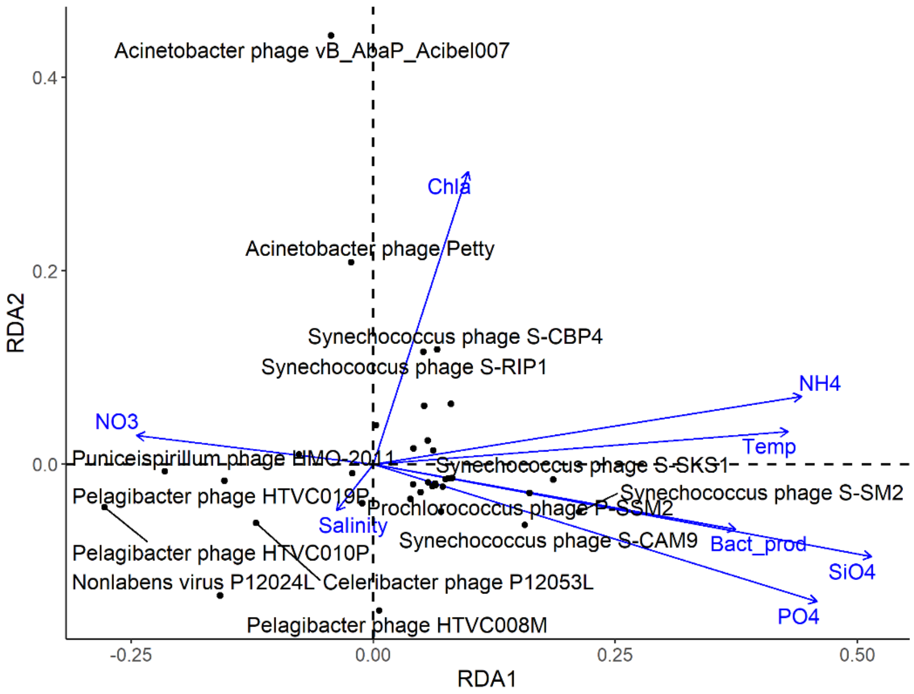

Figure 3.8 Redundancy analysis (RDA) ordination diagram (biplot) of abundant known viral species (black) in DEV and environmental variables (blue), from taxonomy of known viruses by Kaiju read classification. RDA1 explains 33% of variance, while RDA2 explains 28% of variance. Labels of data points below 0.1 have been omitted for clarity. The angles between virus species and environmental factors denote their degree of correlation.

3.4.4 Viral taxonomy of estuarine viromes vs. open ocean viromes

The percentage of known viruses (ca. 10%) were similar between the DEV samples and the four ocean samples (Fig. 3.5, Fig. 3.6). On the family level, a higher proportion of *Myoviridae* were found in oceanic samples; *Phycodnaviridae* were found in all estuarine samples (ranging from 1.5% to 4.6% of all viral reads), but were not detected in oceanic samples (Fig. 3.5). Oceanic samples contained significantly more *Prochlorococcus* phage than the estuarine environments (Fig. 3.6). Compared to open ocean, *Puniceispirillum* phage and *Pelagibacter* phage appear to more abundant in the estuarine environment (Fig. 3.6). Despite differences in sampling methods across different cruises, the viral taxonomy results were comparable due to the similar sequencing technology employed, lending reasonable legitimacy to the viral taxonomy methods used in this study.

3.5 Discussion

3.5.1 Known and unknown viruses in the DEV

Due to the large proportion of unknown viruses in metagenomic datasets, the analysis of known viruses and abundant viruses were handled separately. In accordance with other viral metagenomic studies, the majority of trimmed reads remain unclassified; only 10% of reads were assigned to known viral taxa, while this value for other viromes range from 0.74% to 21% (Cai *et al.*, 2016; Hwang *et al.*, 2016). When looking at viral contigs identified using the IMG/VR process, approximately 26% of reads were mapped to the viral populations (Table 2.3), indicating that the viral populations encompass significantly more of the sequence data than known RefSeq viruses. This proportion echoes a global viromic study where only 25% of predicted proteins were found to have similarity with any known viral proteins (Paez-Espino *et al.*, 2016), suggesting that the majority of viral sequences are still unknown. Compared to the dramatically changing unknown viral populations, the composition of the known viral community is relatively more stable throughout different seasons and locations in the estuaries (Fig. 3.7, Fig. 3.2). Attempts to identify the most abundant viral populations in the DEV found them to be mostly novel and not matched to cultured viral isolates (Table 3.2). This implies that the most dynamic and abundant viral species in the estuaries have not yet been characterized. Indeed, the failure of known CRISPR spacers to predict hosts of more abundant (FPKM > 100) viral populations further indicates the novelty of the most prolific species in the DEV (Table S2). The spatiotemporal pattern of these abundant but uncultivated viruses is more variable compared to that of cultured viruses.

3.5.2 Spatiotemporal pattern of estuarine virioplankton

The relative abundance of viral populations varied greatly throughout different seasons in the Delaware Bay (Fig. 3.2), supporting the “seed bank model” which states that most viruses exist in an inactive status throughout the year while only the most abundant viruses are active in a given community (Breitbart and Rohwer, 2005). It has been found that about half of the Delaware Bay bacterial community cycles between rare and abundant species, with rare bacteria acting as a “seed bank” waiting for conditions to change (Campbell *et al.*, 2011). Our results showed that the Delaware Bay viral community displays a similar pattern to its bacterial community, which is also consistent with a previous viromics study (Angly *et al.*, 2006).

It was difficult to discern a variation pattern in the Chesapeake Bay due to the low number of samples, and the lack of upper bay sites. CB8.2M showed a significantly higher proportion of known viral reads compared to other samples (Fig. 3.5, Fig 3.6), but did not show high amounts of reads mapping to the most abundant viruses (Fig. 3.2), further indicating that known viruses follow different patterns than abundant viruses.

In general, the bacterioplankton community in the Delaware Bay varies drastically along the salinity gradient, the dominant bacteria changing from *Actinobacteria* and *Verrucomircobia* in the upper estuary, to *Pelagibacter* and *Rhodobacterales* in the lower estuary, the community showing a clear shift from a “freshwater” profile to an “oceanic” profile (Campbell and Kirchman, 2013). In contrast, although also variable, the virioplankton community does not show such a distinct transition from upper to lower estuary (Fig. 3.7, Fig.3.6, Fig. 3.2). This is supported by the finding that location in the estuary is not a significant factor in community similarity (Fig. 3.3, Table 3.3). This is perplexing given that viruses are dependent on their hosts for replication, but our identification of viruses may be skewed since freshwater viruses are poorly characterized compared to marine viruses, while bacteria in both environments are better characterized than viruses in general (Kavagutti *et al.*, 2019). Despite the geographic proximity of the two estuaries, the viral community in the Delaware Bay is significantly different from that in the Chesapeake Bay (Fig. 3.3, Table 3.3). The viral population differences between the two bays is more distinct than that caused by similar temperature or salinity (Table 3.3). This distinction may be a result of the various abiotic differences between the two estuaries, including the larger watershed and nutrient limitation in the Chesapeake Bay (Fisher *et al.*, 1988). In the Delaware Bay, abundance patterns of both known and unknown viruses appear to be variable along the salinity gradient in the spring and fall, but relatively consistent from the upper to lower bay in the summer (Fig. 3.7, Fig. 3.2). This spatial and seasonal pattern is more pronounced in the unknown viruses, which display more dramatic changes (Fig. 3.2). The primary source of freshwater in the Delaware Bay is the Delaware River, and high levels of river discharge during the spring causes stratification in the estuary, impacting the spatial variation of phytoplankton production, leading to variation in the microbial community along the salinity gradient (Sharp *et al.*, 1986). While in the summer, lower levels of discharge allow for better mixing and more consistent phytoplankton production levels along the Delaware estuary, leading to a more stable microbial community. In contrast to the Delaware Bay, such spatial and seasonal abundance patterns are obscured for the partially-mixed Chesapeake Bay due to the amount of tributaries along its length and its relatively long water residence time (~180 d) (Du and Shen, 2016). An inter-annual study found that viral abundance and viral production did not change greatly from the upper to lower Chesapeake Bay, despite strong environmental gradients (Winget *et al.*, 2011). The DEV relative abundance data concurs by showing little influence from salinity gradients in the Chesapeake Bay, although this may be due to the lack of upper bay samples in this study (Fig. 3.2). The inclusion of different sampling depths in the Chesapeake Bay but not the Delaware Bay is also a contributor to the statistical dissimilarity between the two bays (Table 2.1, Fig. 3.3, Table 3.3). The spatial and temporal variations have allowed us to reveal the above patterns in the estuarine virome.

In several of the analyses conducted in this study, samples DB3.3 and DB11.1 show a similar community structure that is distinct from the other DEV samples. A lower percentage of known viruses were identified in these two samples (Fig. 3.5, Fig. 3.6), correspondingly, higher abundances of unknown viruses were observed (Fig. 3.2). These two samples were grouped together and away from the other samples, both in the qualitative cluster network plot of viral contigs (Fig. 2.3), and the NMDS plot of abundant viral populations (Fig. 3.3). Analysis of variance (ANOSIM) testing showed significant dissimilarity when grouping these two samples vs. other samples (Table 3.3). The different community structure of these two samples may be indicative of some episodic event in the Delaware Bay, the cause of which is not documented in the environmental factors we currently have access to (see Table S1). It is also possible that DB11.1 may have been switched with DB11.2 or DB11.3 at some point during the sample or sequencing processing, but not enough evidence was found regarding the nature of the switch, so the current sample organization will be retained until further supporting evidence is uncovered.

3.5.3 Comparison of the DEV with other estuarine and oceanic viromes

On the family level, members of the viral family *Myoviridae* are generally found to be most abundant in the open ocean, followed by those from the *Podoviridae*, while *Siphoviridae* family viruses are less common (Aylward *et al.*, 2017). Estuaries appear to follow an overall similar trend. The higher proportion of *Siphoviridae* in DB11.1 may be influenced by terrestrial runoff at its high, riverine position (Fig. 2.1, Fig.

3.5). Estuarine samples from the GOS viral metagenomic study found that the Chesapeake Bay has a higher relative abundance of *Myoviridae* compared to the Delaware Bay (Williamson *et al.*, 2008), which concurs with our results (Fig. 3.5). Since then, a viral community study involving both the Delaware Bay and the Chesapeake Bay has not been conducted. The early study on the Chesapeake Bay found that the proportion of *Siphoviridae* is much lower than that of *Myoviridae* and *Podoviridae*, and rare occurrence of viruses with eukaryotic hosts (Bench *et al.*, 2007), which is consistent with this study (Fig. 3.5). Other estuarine viromes in Korea and the Baltic Sea also showed high proportions of *Myoviridae* and *Podoviridae* members (Hwang *et al.*, 2016; Zeigler Allen *et al.*, 2017; Garin-Fernandez *et al.*, 2018), although a study in China found higher proportions of *Siphoviridae* than *Myoviridae* in the estuary (Cai *et al.*, 2016). This shows that virioplankton in estuaries around the world have a similar structure on the family level. In this study, a higher proportion of *Myoviridae* was found in oceanic samples compared to estuarine samples; the relatively higher proportion of *Myoviridae* in CB8.2M and CB8.2D may be due to the influence of oceanic water from vertical stratification, as is evidenced by their higher salinity compared to the surface water sample (Fig. 3.5, Table 3.1). Cyanomyovirus are more abundant relative to cyanopodovirus in coastal and open ocean viral metagenomes compared to those in estuaries (Huang *et al.*, 2015). Since a large portion of known viruses in the DEV are cyanophage (Fig. 3.6), this supports our current findings. *Phycodnaviridae* are abundant and ubiquitous in the oceans, but this study did not find *Phycodnaviridae* in oceanic sites (Endo *et al.*, 2020). The absence of *Phycodnaviridae* in oceanic sites in this study may be due to differing bioinformatic methods used. Since *Phycodnaviridae* are larger than *Caudovirales* with capsid size ranging from 100-220 nm (Wilson *et al.*, 2009), it may also be due to the difference in viral sampling techniques on different cruises.

Cyanophages and pelagiphages are thought to be the most abundant known viruses in marine environments (Sieradzki *et al.*, 2019). The higher prevalence of cyanophage in the summer and large proportions of *Pelagibacter* phage and *Puniceispirillum* phage is consistent with other estuarine viromic studies (Fig. 3.6) (Cai *et al.*, 2016; Hwang *et al.*, 2016). Pelagibacter consist 40-60% of the bacterioplankton community in mid to lower Delaware Bay, and are significantly less abundant in the upper bay, consisting 0-5% of metagenomic reads (B. Campbell unpubl.); meanwhile, known pelagiphage only make up 1-2% of total reads and about 10% of known viral reads, and do not show a clear transition from upper to lower bay, displaying completely different patterns compared to their presumed hosts (Fig. 3.6). Since isolation of pelagiphage is difficult and sometimes require methods such as single-cell genomics (Zhao *et al.*, 2013; Martinez-Hernandez, Fornas, *et al.*, 2019), our current ability to identify pelagiphages from metagenomic sequences is highly limited and may be causing this discrepancy between phage and host. Cyanophage play an important role in the regulation of cyanobacterial abundance in the Chesapeake Bay (Wang and Chen, 2004; Wang *et al.*, 2011). The most abundant cyanophage species in DEV matched with some *Synechococcus* phages isolated from Chesapeake Bay, including podoviruses *Synechococcus* phage S-CBP1, S-CBP3 and S-CBP4, and siphoviruses *Synechococcus* phage S-CBS2, S-CBS3 and S-CBS4 (Wang and Chen, 2008) (Fig. 3.7). Based on lab studies, all of these cyanophages are highly host-specific, infecting locally isolated *Synechococcus* species CB0101, CB0204, and CB0202 (Wang and Chen, 2008). Unlike for pelagiphage, the extensive cyanophage isolation work conducted in the geographic vicinity allows us to make more connections between phage and host. We anticipate similar findings for *Pelagibacter* phage-host relationships with the isolation and documentation of more pelagiphage strains. In contrast with the broad distribution of *Synechococcus*, *Prochlorococcus* is rarely found in coastal eutrophic systems, while abundant in warm oligotrophic waters (Partensky and Garczarek, 2010). The significant presence of *Prochlorococcus* phage in oceanic samples compared to estuarine samples (Fig. 3.6) supports this paradigm, and is consistent with previous studies (Huang *et al.*, 2015). The small fraction (<1%) of *Prochlorococcus* phage sequences found in estuarine samples (Fig. 3.6) may be due to the fact that certain cyanophages such as cyanomyoviruses tend to cross-infect *Synechococcus* and *Prochlorococcus* (Sullivan *et al.*, 2003). The host range of current phage isolates were explored to differing degrees, so a cyanophage isolated using *Prochlorococcus*, does not indicate that it does not also infect *Synechococcus*.

The most abundant viral populations in the DEV tend to be very novel, which concurs with other contig-level virome studies (Paez-Espino *et al.*, 2016; Aylward *et al.*, 2017). Abundant marine viral populations have been found to be both variable and persistent across seasons (Aylward *et al.*, 2017) and locations (Brum *et al.*, 2015; Roux *et al.*, 2016). Similarly, abundant viral populations in the DEV were found to have varying patterns across samples (Fig. 3.2). Despite most of these populations being unknown, their dominancy in the estuarine environment suggests they may infect some abundant bacterial populations which have not yet been identified. Since unknown viral populations account for a large portion of these estuarine viromes, and their potential hosts and ecological role still remain largely unknown, it is necessary to understand more about these cryptic viral groups.

3.5.4 Importance of single-cell and single-molecule methods

Phages infecting abundant but relatively slow-growing and difficult-to-culture marine bacteria make up a significant portion of marine viruses in the ocean (Zhang *et al.*, 2019). Since 2017, uncultivated virus genomes have outnumbered virus genomes sequenced from isolates (Roux *et al.*, 2019), but identification of metagenomic sequences still relies primarily on culture-dependent microbial discovery. In recent years, single-cell genomics have offered valuable insights into the marine viral community (Labonté *et al.*, 2015), discovering some of the most abundant and ecologically significant viruses in the marine ecosystem (Martinez-Hernandez *et al.*, 2017; Berube *et al.*, 2018). In particular, the abundance of the single virus isolate 37-F6, of which the putative host is *Pelagibacter* (Martinez-Hernandez, Fornas, *et al.*, 2019), is thought to rival or exceed that of *Pelagibacter phage* HTVC010P and *Puniceispirillum phage* HMO-2011, which were previously thought to be the most abundant viruses in the ocean (Kang *et al.*, 2013; Zhao *et al.*, 2013; Martinez-Hernandez *et al.*, 2017). Likewise, long read single molecule sequencing uses long nanopore reads (20-80 kb) to capture entire viral genomes without assembling, avoiding some of the biases induced by short-read de novo assembly, thus revealing “hidden” viral diversity not covered by conventional metagenomic sequencing methods (Beaulaurier *et al.*, 2020). Several of the most abundant viral populations in the DEV have the closest match to prokaryotes discovered using non-conventional methods such as single cell genomics, single virus genomics and long read single molecule sequencing (Fig. 3.2, Table 3.2), demonstrating the importance of non-cultivation dependent virus characterization methods for revealing viral diversity. These results indicate that discoveries using the above methods may be important for revealing the most abundant and ecologically relevant viral species in the marine and estuarine environment, improving our understanding of viral dark matter.

3.6 Conclusion

We were surprised to find that the virioplankton community does not show a distinct transition from upper to lower estuary, or across different seasons despite strong environmental gradients, compared to their prokaryotic hosts. In contrast, Delaware Bay and Chesapeake Bay viral populations were found to be significantly different from each other, despite their geographical proximity. We found that the most abundant viral populations in estuaries (top 20) are not the usually dominant viral groups such as pelagiphage and cyanophage, but viruses which have not yet been cultivated, related to uncultured viral sequences discovered via single cell and assembly-free long read single molecule methods, highlighting the importance of these unconventional methods for viral discovery. Comparison with other aquatic environments showed that estuarine virioplankton around the world have a similar structure on the family level (*Siphoviridae*, *Myoviridae*, *Podoviridae*); while open ocean virioplankton have a higher proportion of *Myoviridae* and *Prochlorococcus* phage. We anticipate the further isolation of novel viral species will enhance our understanding of the estuarine virome.

Chapter 4: Distribution of N4-like viruses in estuarine viromes

4.1 Abstract

N4-like viruses are interesting due to their conserved genetic features, a large RNA polymerase gene (~10Kb), distinct taxonomy and widespread occurrence in nature. Currently, 115 N4-like viruses which infect different bacterial species have been isolated, and this group has been proposed to be a new viral family "*Schitoviridae*". An earlier study based on the PCR detection of N4 viruses led to a hypothesis that N4-like viruses are more prevalent in cold season or cold waters, however, this has not been confirmed based on qPCR or metagenomic analysis. We obtained 16 viromes from the Chesapeake Bay and Delaware Bay from different seasons between 2014 and 2015. This dataset allowed us to evaluate the relative abundance of N4-like viruses in different seasons of the estuarine environment. The relative abundance of N4-like viruses in two temperate estuaries was assessed using four different methods: 1) read mapping to known N4-like virus isolates, 2) read mapping to native viral contigs, 3) reciprocal blast search based on core genes, and 4) read taxonomy classification using Kaiju. A total of 11 N4 contigs were identified based on de novo assembly. Discrepancies existed between these different methods. Overall, N4-like viruses were found to be much less abundant compared to pelagiphage and cyanophage in the estuarine viromes. At the read level, high occurrences of N4-like viruses infecting *Roseobacter* and *Vibrio* were found, and their distribution patterns seemed closely connected with their hosts. When using contig-based methods and

Kaiju classification, N4-like viruses were found to be more abundant in winter, and less abundant or not detectable in summer. This result explains the failure of PCR detection of N4-like viruses in summer in the earlier study, suggesting a strong seasonal variation of N4-like viruses in the estuarine ecosystem. A core gene based analysis also provided guidance on the choice of genes when using marker gene based approaches for future studies of N4-like virus in the environment. Our study indicates that N4-like viruses are rare in the marine environment, and also provides insights into how to evaluate the ecology of rare viruses such as N4-like viruses.

4.2 Introduction

Virioplankton are known to be abundant and diverse in nature, and they influence nutrient dynamics and biogeochemical cycles in the aquatic environment by interacting with living organisms (Coutinho *et al.*, 2018). Isolation of viruses, viral metagenomics and new technologies such as single-virus isolation have continued to uncover novel viruses and new viral lineages (Dion *et al.*, 2020). One of these interesting viral groups is N4-like viruses. Bacteriophage N4 was first isolated from sewage water in Italy using *Escherichia coli* (Schito *et al.*, 1966). Phage N4 is unique in the viral world due to its giant RNA polymerase, which takes up about a seventh of its genome and makes it the only known bacteriophage that is not reliant on host RNA polymerase in early transcription (Choi *et al.*, 2008). Also, N4 causes delayed lysis in its host and a resulting large burst size of around 3,000 viral particles per cell (Stojković and Rothman-Denes, 2007).

N4 remained a genetic orphan for 40 years until the isolation of two N4-like viruses infecting *Roseobacter* from the Chesapeake Bay (Zhao *et al.*, 2009). Since then, more

N4-like viruses have been isolated from coastal waters (Huang *et al.*, 2011; Chan *et al.*, 2014; Ji *et al.*, 2014; Cai *et al.*, 2015; B. Li *et al.*, 2016), soil (Born *et al.*, 2011), and farms (Nho *et al.*, 2012; Moreno Switt *et al.*, 2013), using a variety of bacterial hosts. The genomes of these N4-like viruses are highly conserved, with a distinctive set of core genes (Chan *et al.*, 2014; Wittmann *et al.*, 2015; B. Li *et al.*, 2016). As of 2020, 115 N4-like viruses have been discovered, sharing 17 core genes, with total number of genes ranging from 76 to 92 per virus (Wittmann *et al.*, 2020). Since the International Committee on Taxonomy of Viruses (ICTV) has moved to classifying viruses based on their genomic features rather than morphology, N4-like viruses were recently reassessed and new viral family, "*Schitoviridae*" was proposed (Wittmann *et al.*, 2020). Despite these developments, *Escherichia virus* N4 remains the only officially recognized member of the "*N4virus*" genus (Lefkowitz *et al.*, 2017). A few N4-like viruses are classified under 10 other genera by the ICTV.

The first two N4-like viruses were isolated using *Roseobacter* (Zhao *et al.*, 2009). A high number of phages infecting marine *Roseobacter* are N4-like viruses, although the reason for this association is unknown (Zhan and Chen, 2019). *Roseobacter* are an abundant and extensively studied lineage of marine bacteria, consisting up to 25% of bacteria in coastal waters, and up to 10% of bacteria in the open ocean (DeLong, 2005; Moran *et al.*, 2007). They are especially abundant in coastal and polar waters (Wagner-Döbler and Biebl, 2006). They have a wide range of metabolic capabilities, playing an active role in marine nitrogen, sulfur and carbon cycles, and interacting closely with phytoplankton (Luo and Moran, 2014).

Early studies based on the two isolates of N4-like phage found them to be more prevalent in coastal waters compared to open ocean water, using BLASTP of the N4-like DNA polymerase gene against the GOS database (Zhao *et al.*, 2009). Later, N4-like viruses were found to be widespread in both coastal and open ocean environments, using reciprocal BLAST of one N4-like phage against CAMERA or EBI metagenomes (Chan *et al.*, 2014). Using a DNA polymerase-based PCR clone library approach in the Chesapeake Bay, N4-like viruses were detected in the Chesapeake Bay. Interestingly, among 56 viral communities, only 12 samples collected in winter were PCR-positive, and none of the samples collected in spring, summer and fall were PCR-positive (Zhan *et al.*, 2015). This result suggests that N4-like viruses are more prevalent in the winter season. N4-like viruses were also found to be prevalent in colder waters when their core genes were used for recruitment in global metagenomic databases (Chan *et al.*, 2014; Zhan *et al.*, 2015). However, there was not sufficient environmental data to correlate N4-like virus abundance with environmental parameters using the metagenomic databases available at the time. In a comparative metagenomic study, N4-like viruses were found to be more dominant in waste water treatment plants, compared to other aquatic environments (Parmar *et al.*, 2018); which is not surprising given a large number of N4-like viruses infect human-associated bacteria (Wittmann *et al.*, 2020). These early studies were only able to assess the prevalence of N4-like viruses in different environments. However, there is no systematic study to compare relative abundance of N4-like viruses in different seasons and across a wide range of salinity gradients based on viromic analysis.

The development of next generation sequencing technologies (Parmar *et al.*, 2017) and the increase in viral metagenomic data (Suttle, 2016) have prompted revision of major paradigms regarding viral ecology (Sullivan *et al.*, 2017). Although N4-like viruses are widespread in aquatic environments, they seem to be present in low abundance in the natural environment (Chan *et al.*, 2014; Zhan *et al.*, 2015). Non-targeted virus taxonomy assignment confirmed that N4-like viruses are rare in estuarine virioplankton compared to other viruses (Sun *et al.*, 2021). Rare viruses can be difficult to quantify in metagenomic datasets, due to the lower chance that they will assemble into longer metagenome-assembled genomes (MAGs). In order to recover the sequences of rare viral groups such as N4-like viruses, deep sequencing of viromes is required. A high quality viromic database (Delmarva Estuarine Virome) using deep sequencing is available for the Chesapeake Bay and Delaware Bay (Sun *et al.*, 2021). In this study, we assess the spatial and temporal distribution of N4-like viruses using the DEV dataset.

4.3 Methods

4.3.1. Viromes from Delaware bay and Chesapeake Bay

Sixteen deeply sequenced viromes (DEV) from the Chesapeake Bay and Delaware Bay were used in this study (Sun *et al.*, 2021). They were collected from low, medium and high salinity sites along each bay. Ten samples were from the Delaware Bay, collected in March, August/September, and November, 2014; six samples were from the Chesapeake Bay, collected in April and August, 2015. These 16 DEV viromes were used to explore the distribution of N4-like viruses in both estuarine bays using different recruitment methods.

4.3.2 Relative abundance of N4-like sequences using read recruitment

To detect the relative abundance of N4-like viral sequences in different aquatic environments, metagenomic reads from 16 estuarine DEV samples and 11 publicly available offshore water samples were mapped to the 115 N4-like virus genomes described in 2020 (Wittmann *et al.*, 2020), using BBMap with the mapping parameters as recommended in viromic benchmarking studies (>90% identity, >75% contig length) (Bushnell, 2014; Roux *et al.*, 2017). Ambigous reads (reads that map equally well to multiple sites) are set to map to the first best possible. The offshore viral sequences were taken from Global Ocean Virome (GOV) 2.0, and Hawaiian Ocean Experiment (HOE) (Table 4.1, Fig. 4.1) (Aylward *et al.*, 2017; Gregory *et al.*, 2019). All chosen samples were taken from surface water. These viral metagenomes were chosen because they were obtained using similar sampling and sequencing technology, which reduces the reduce bias that may be present when evaluating relative abundance across sequence datasets of different origin. Trimmed reads were counted and normalized to FPKM (Fragments Per Kilobase Million) using SAMtools (Li *et al.*, 2009). FPKM is used as a proxy for relative abundance (Roux *et al.*, 2017). Read mapping results were visualized with Integrative Genomics Viewer (IGV) using the default parameters (Thorvaldsdóttir *et al.*, 2013).

Table 4.1 Sampling information of oceanic viromes used for N4 read recruitment.

Dataset	**Sample**	**Date**	**Time**	**Region**	**Longhurst biome**	**Temp (°C)**	**Salinity (ppt)**
GOV 2.0	109_SRF	5/12/2011	14:00	Pacific equatorial	Coastal	27	33
GOV 2.0	067_SRF	9/7/2010	06:19	African cape	Coastal	13	35
GOV 2.0	036_SRF	3/12/2010	06:06	Arabic sea	Coastal	26	37
GOV 2.0	201_SRF	2013-09-30	15:02	Arctic (Canada)	Polar	-1	31
GOV 2.0	173_SRF	2013-07-08	04:12	Arctic (Russia)	Polar	0	34
GOV 2.0	155_SRF	2013-05-24	05:36	North- Atlantic (Ireland)	Westerlies	11	35
GOV 2.0	125_SRF	2011-08-08	17:33	Mid-Pacific	Trades	27	35
HOE	HOE_17	2015-07-27	10:05	Pacific (Hawaii)	Trades	27	35
GOV 2.0	072_SRF	2010-10-05	08:00	Mid-Atlantic	Trades	25	36
GOV 2.0	048_SRF	2010-04-19	07:56	Indian Ocean	Trades	30	34
GOV 2.0	085_SRF	2011-01-06	10:38	Antarctic	Polar	1	34

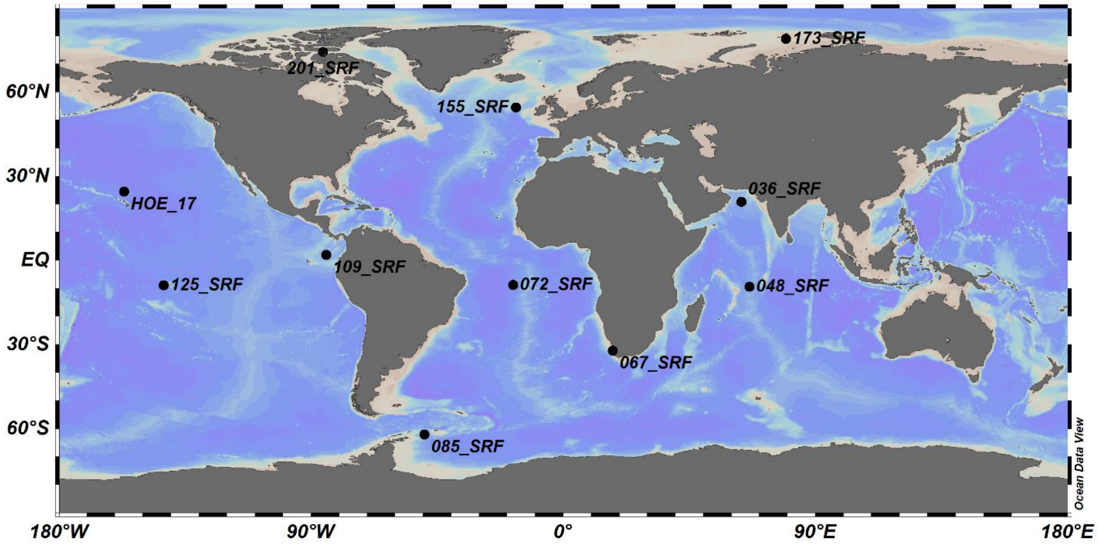

Figure 4.1 Sampling map of publicly available offshore viromes used for N4-like viral read recruitment. The map was created using Ocean Data View (Schlitzer, R., Ocean Data View, https://odv.awi.de, 2019), with the ETOPO1 map (Amante and Eakins, 2009).

4.3.3 Comparison of N4-like virus, cyanophage, and pelagiphage relative abundance

To compare N4-like viruses against viral groups that are known to be abundant in the marine environment, four representative N4-like viruses, one pelagiphage, one *Puniceispirillum* phage, and two cyanophages were combined and used as a reference for read recruitment. The four reference phages are known to be abundant in the estuary and ocean based on non-targeted metagenomic studies (Sun *et al.*, 2021). The metagenomic reads of the 27 water samples were mapped to the eight reference genomes, and processed using the methods described above.

4.3.4 Identification and abundance of N4-like contigs

To identify N4-like sequences among estuarine viromic contigs, the unique feature of N4-like viruses, the N4 vRNA polymerase gene was searched against DEV (Delmarva Estuarine Virome) viral populations (Sun *et al.*, 2021) using TBLASTN with an e-value of 10^{-5} (Altschul *et al.*, 1990). Only matches with a length above 1,000 bp were retained. Contigs were annotated by comparison to GenBank using MG-RAST (Keegan *et al.*, 2016).

Relative abundance of the N4-like contigs were derived via read mapping, using the methods in (Sun *et al.*, 2021). Briefly, trimmed reads from DEV were mapped to all viral contigs, and the FPKM of N4-like contigs were used as a proxy for their relative abundance. The FPKM values of the N4-like contigs in the 16 estuarine samples were plotted in R using ggplot2 (Ginestet, 2011).

4.3.5 Phylogenetic analysis of N4-like contigs

The vRNA polymerase sequences from the N4-like contigs were extracted using TBLASTN and aligned using MEGA X using MUSCLE, along with the vRNA polymerase of six representative N4-like viruses from different subfamilies as reference sequences (Altschul *et al.*, 1990; Kumar *et al.*, 2018). Due to the long length of the vRNA polymerase gene (~3,500 aa), the full length of the gene was unable to be extracted from every contig, so the alignment was trimmed down to the last 1,203 aa of the gene for generating the phylogenetic tree. The phylogenetic tree was calculated with the maximum likelihood algorithm, and replicate trees were assessed with the bootstrap test (100) in MEGA X (Kumar *et al.*, 2018). The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model.

4.3.6 Detection of N4-like virus using reciprocal best hits

Relative abundance of certain virus groups of interest was determined using a reciprocal best-hit BLAST strategy as in (Zhao *et al.*, 2013). Due to the genomic mosaicism of viruses and limited knowledge of their genomes, a core gene-based approach was chosen (Table 4.2). After discarding contigs under 600 bp, each estuarine virome was made into a BLAST database and queried with all of the N4-like core genes described elsewhere (Chan *et al.*, 2014). The matching portion of each putative hit was then extracted and queried against a protein database containing: (a) the core genes of N4-like viruses and (b) the non-redundant proteins of all bacterial and viral genomes from Refseq, excluding those with over 98% BLAST identity with any of the core genes. Reciprocal best-hits were extracted, counted and normalized against the size of the metagenomic database and the size of the core gene.

Over the course of our investigation, the size of the viral metagenomes and reference databases became significantly larger, exponentially increasing the computational load of the reciprocal best hits method. Thus, this analysis was performed using the estuarine samples against an earlier version of the NCBI Refseq database (release of 1-12-2018).

Table 4.2 N4 virus core genes used in reciprocal best hit analysis (Chan *et al.*, 2014).

Gene in N4	Gene name	Category
15	RNAP1	Transcriptional control
16	RNAP2	Transcriptional control
24	Unknown	N/A
25	vWFA domain	N/A
39	DNA polymerase	DNA metabolism/replication
45	SSB	DNA metabolism/replication
50	vRNAP	DNA metabolism/replication
53	Unknown	N/A
54	Structural protein	Structural
55	Unknown	N/A
56	Major coat protein	Structural
59	94 kDa portal protein	Structural
68	Terminase, large subunit	Virus assembly
69	Unknown	N/A

4.3.7 Correlation between N4-like virus and environmental factors

To explain the effect of temperature and salinity on the abundance of N4-like viral sequences, redundancy analysis (RDA) was plotted for the N4-like virus contigs using the vegan package in R, and visualized using type I scaling in ggplot2 (Ginestet, 2011; Oksanen *et al.*, 2018). Environmental factors were normalized by making sum of squares equal to one, and the species abundance was normalized using the Hellinger method. The RDA biplot explains the variation of the abundance of the N4-like virus contig samples using temperature and salinity.

A summary of the different methods used to evaluate the abundance of N4-like viruses is provided in Fig. 4.2.

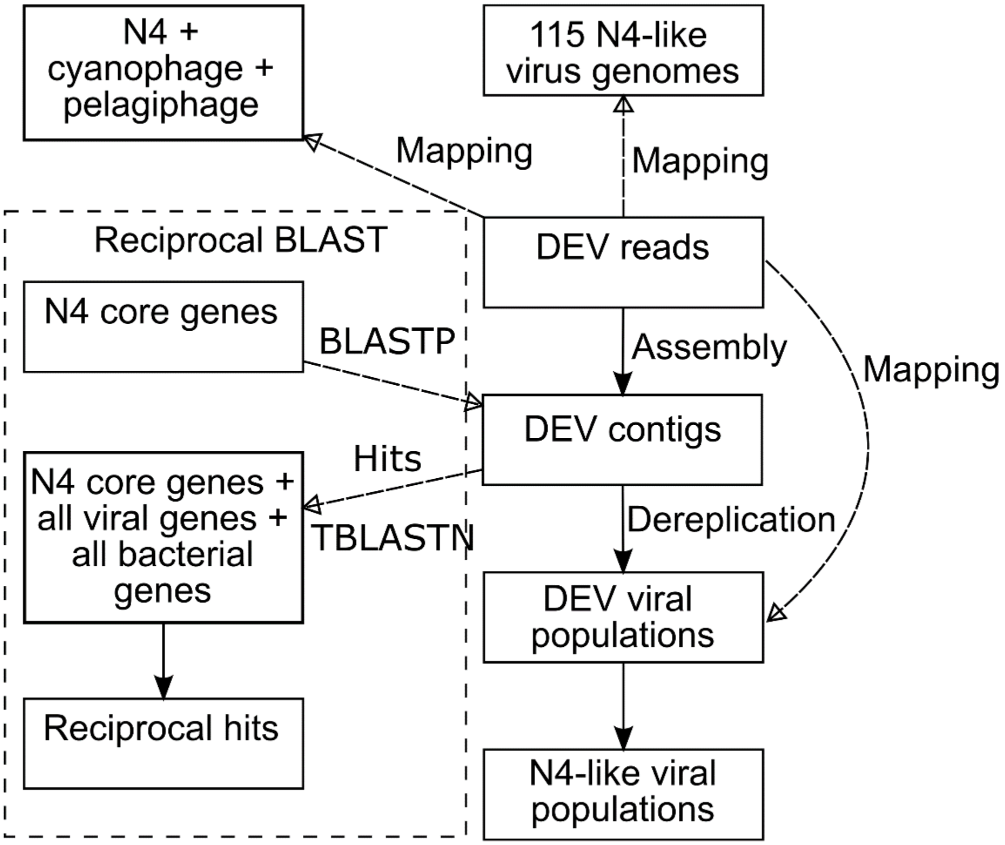

Figure 4.2 Schematic diagram of the different methods used to evaluate abundance of N4-like viruses. Solid arrows indicate sequences generated from the prior box, dashed arrows indicate sequence alignment processes such as BLAST and read mapping.

4.4 Results and discussion

4.4.1 Distinct composition of N4-like viruses between estuaries and other marine environments

The composition of N4-like viruses in the Chesapeake Bay and Delaware Bay is distinct from that of coastal and open ocean, except for samples 048 and 085 which were collected from the Indian Ocean and the Antarctic Drake Passage (Table 4.1, Fig 4.1, Fig. 4.3). The most striking feature of the N4-like viral diversity is the high relative occurrences of sequences similar to *Roseobacter* N4-like viruses in most estuarine samples and some ocean samples (Fig. 4.3). *Roseobacter* N4-like viruses predominate the N4-like virus populations in the Delaware Bay and samples 048 and 085 (Fig. 4.3). In the Chesapeake Bay, *Roseobacter* N4-like viruses dominated the most samples, but the N4-like viruses infecting *Vibrio, Pseudomanas, and Pseudoaltermonas* were also abundant, especially in summer (August). Interestingly, *Roseobacter* N4-like viruses were not prevalent in most of coastal and open ocean samples (except for samples 048 and 085). Instead, Enterobacterial N4-like viruses were common in the coastal and open ocean. The difference in host community may explain the distinct pattern of N4-like viruses seen between estuarine and coastal/open environments. The Chesapeake Bay is known to contain its own unique bacterial populations including *Roseobacter*, as a consequence of adaptation to the estuarine environment (Jinjun and Jun, 2011). The dominance of *Roseobacter* N4-like viruses in both bays could be related to the fact that most available N4-like *Roseobacter* viruses were isolated from the Baltimore Inner Harbor, which is in close geographic proximity to the Delaware Bay and Chesapeake Bay (Zhan and Chen, 2019). The composition of N4-like viruses also varied with depth in the Chesapeake Bay. The N4-like viruses infecting *Roseobacter* made up the majority of N4 populations in the surface water (CB8.2S), while N4-like viruses infecting *Vibrio*, *Pseudoaltermonas* and *Pseudomonas* became abundant in the deeper water (CB8.2M and CB8.2D) (Fig. 4.3). Such a shift in N4-like virus populations along the vertical profile has not been seen before, suggesting that the viral community in deeper and saltier water in the Chesapeake Bay may differ from that in the surface or upper water. Coastal water enters the Chesapeake Bay from the bottom, forming a strong stratification in summer because surface water in the bay is warmer and has lower salinity. The two-layer circulation is a well known feature for the Chesapeake Bay (Goodrich and Blumberg, 1991). This stratification is confirmed by the increasing salinity with depth in these three samples (Sun *et al.*, 2021).

We did not detect any N4-like viruses in sample 155 or the HOE sample (Fig. 4.3). The absence of N4-like viruses in sample 155 may be due to sampling or sequencing problems, since other viruses were not detected either (Sun *et al.*, 2021). The overall viral profile of the HOE sample was comparable to other marine samples, therefore, it is not clear why N4-like viruses are missing in these two samples. The read recruitment method showed remarkably consistent total relative abundances of N4-like viral reads across the 27 marine samples we examined (Fig. 4.3). To check whether this is a false positive caused by mapping artifacts, manual examination of read mapping visualization confirmed that reads were indeed mapping to abundant N4-like genomes with high identity and coverage (Fig. 4.4).

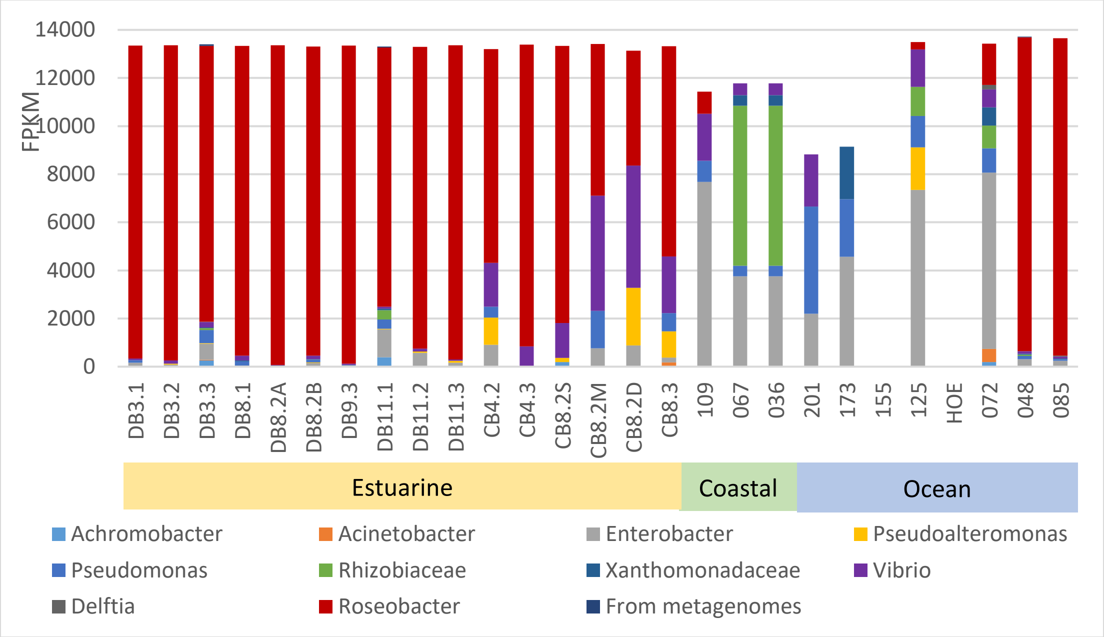

Figure 4.3 Relative abundance of N4-like viruses using read recruitment, grouped by bacterial families they infect.

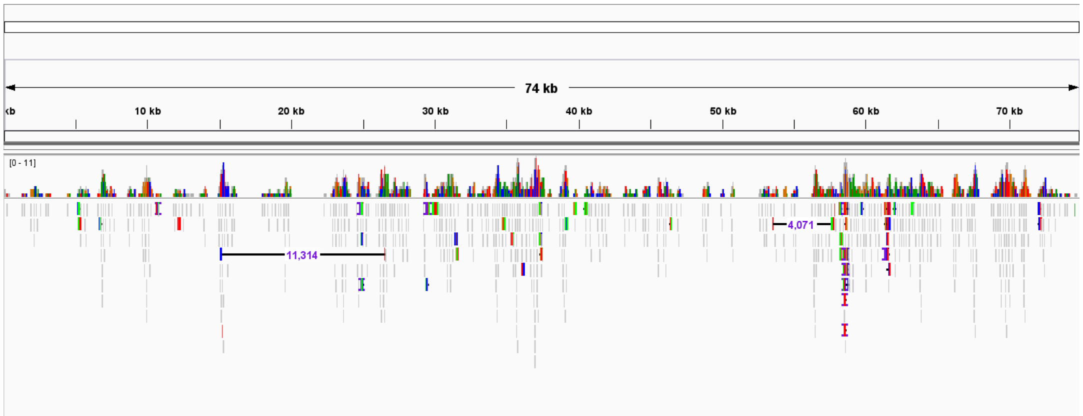

Figure 4.4 Read mapping visualization of DB3.1 to *Ruegeria phage* vB_RpoP-V13 using Integrative Genomics Viewer (IGV) with the default parameters. Mapping data is derived from Fig 4.3.

4.4.2 Composition of major N4-like viruses in estuaries and other marine environments.

The composition of *Roseobacter* N4-like viruses varied greatly between DEV samples. In the Delaware Bay, *Ruegeria phage* vB_RpoP-V13 made up the vast majority of *Roseobacter* N4-like phage community in samples DB3.1, DB8.1, DB8.2A, DB8.2B, and DB9.3 (Fig. 4.5). In the Chesapeake Bay, the *Roseobacter* N4-like virus community was represented by several known *Roseobacter* N4-like viruses. This is also the case for a few samples in the Delaware Bay (DB3.3, DB11.1, DB11. 2, and DB11.3). These results suggest that different *Roseobacter* N4-like viruses are present in these two estuaries, with no clear spatiotemporal pattern observed.

The most abundant *Roseobacter* N4-like viruses observed (*Ruegeria phage* vB_RpoP-V13, *Roseophage* DSS3P2) were previously isolated from the Baltimore Inner Harbor (Fig. 4.5). Roseophage DSS3P2 was the first marine N4-like virus to be discovered, isolated in 2009 using *Roseobacter pomeroyi* DSS-3; while *Ruegeria phage* vB_RpoP-V13 was isolated in 2019 from the same host (Zhao *et al.*, 2009; Zhan and Chen, 2019).

It is noteworthy that within the *Roseobacter* N4-like virus community, one particular type of N4 tends to dominate the N4 community at a particular time (Fig. 4.5). These results suggest that individual N4-like viruses can be very dynamic in nature, and may reflect Red Queen-like virus-host succession dynamics (Ignacio-Espinoza *et al.*, 2020).

Despite *Roseobacter* being abundant in coastal waters, *Roseobacter* N4-like viruses were found to be at low abundance or non-existent in coastal waters (Fig. 4.5). It is possible that *Roseobacter* possess advantages that make them less susceptible to viral infection in coastal waters. The *Roseobacter* clade was found to be dominant in Arctic and Antarctic sea ice (Brinkmeyer *et al.*, 2003). The association of *Roseobacter* with polar environments is consistent with our observation of dominant *Rosoebacter* N4-like viruses in Antarctic oceans (Fig. 4.3). The ocean sites 048 and 085 were predominated by *Roseophage* RD-1410W1-01, which is at low abundance in the Delaware Bay and Chesapeake Bay (Fig. 4.5). *Roseophage* RD-1410W1-01 is a N4-like virus isolated from South China Sea, which infects *Roseobacter denitrificans* OCh114 isolated from coastal Australia, a model organism for the study of aerobic anoxygenic photosynthesis in bacteria (Tang *et al.*, 2009; B. Li *et al.*, 2016).

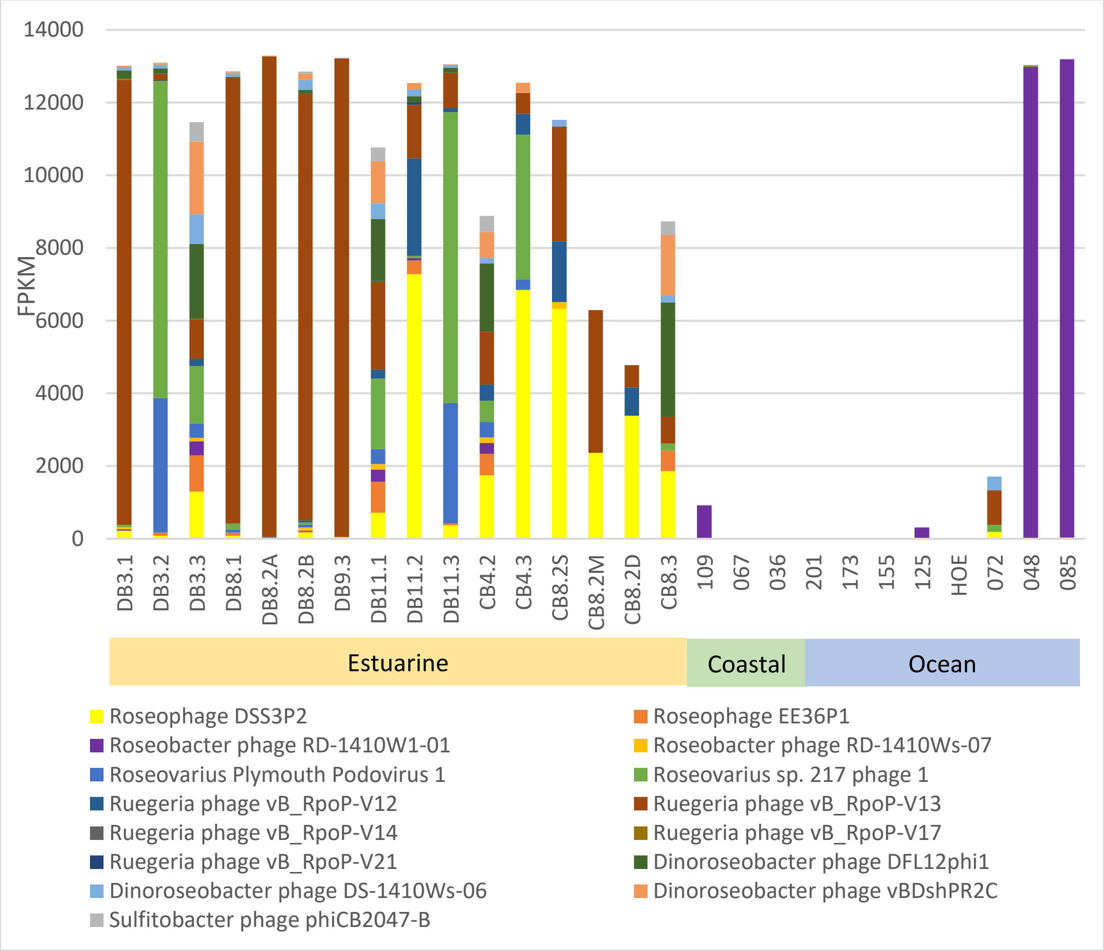

Figure 4.5 Relative abundance of N4-like viruses infecting *Roseobacter* in the Delaware Bay, Chesapeake Bay and reference sites (coastal and open ocean viromes). Expanded diversity from the "Roseobacter" section in Fig. 4.3

N4-like viruses infecting *Vibrio* in the Delaware Bay were much less abundant compared to those in the Chesapeake Bay (Fig. 4.6). This may be due to bias arising from the extensive *Vibrio* strains isolated from the Chesapeake Bay (Colwell *et al.*, 1977; Ceccarelli and Colwell, 2014). The *Vibrio* N4-like virus community in the Chesapeake Bay contained mixed populations represented by different *Vibrio* N4 virus isolates (Fig. 4.6). Two deeper water samples (CB8.2M and CB8.2D) in the Chesapeake Bay contain relatively more abundant *Vibrio* N4-like viruses. These two water samples were likely affected more by the oceanic water as described above. The high occurrence of *Vibrio* N4-like viruses can also be related to a relatively higher abundance of *Vibrio* in the deeper water. *Vibrio* are known to reside in sediments in estuaries (Kaneko and Colwell, 1973; Froelich *et al.*, 2013). Indeed, the CB8.2S, CB8.2M, and CB8.2D samples show increasing levels of *Vibrio* N4-like viruses with depth in the estuary (Fig. 4.3).

Warmer temperature in summer can also contributes to more *Vibrio* N4-like viruses in the Chesapeake Bay as the abundance of marine *Vibrio* has a strong positive correlation with seawater temperature (Thompson *et al.*, 2004; Tout *et al.*, 2015). A prevalent N4-like virus in the Chesapeake Bay is *Vibrio phage* VBP47, which infects *Vibrio parahaemolyticus*, a well-known pathogen associated with brackish water (Colwell *et al.*, 1977; Wittmann *et al.*, 2020). Since the replication of viruses is dependent on the viability of their hosts, the prevalence of N4-like viruses infecting pathogenic hosts may be indicative of elevated pathogenic activity in the region.

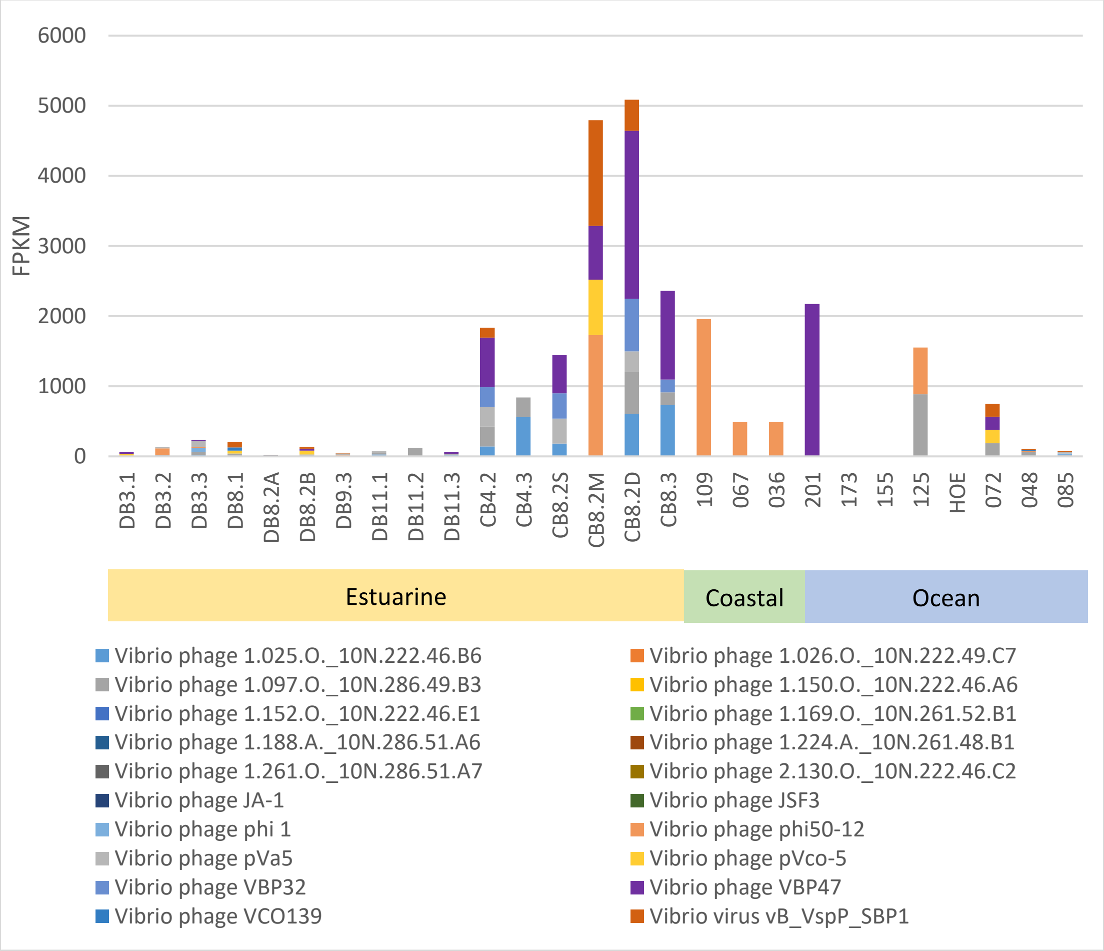

Figure 4.6 Relative abundance of N4-like viruses infecting *Vibrio* in the Delaware Bay, Chesapeake Bay and reference sites (coastal and open ocean viromes). Expanded diversity from the “Vibrio” section in Fig. 4.3.

It is also intriguing that the N4-like virus infecting *Rhizobiaceae*, *Sinorhizobium phage* ort11, was detected in high abundances in certain coastal samples (Fig. 4.3). *Sinorhizobium* phage ort11 infects *Sinorhizobium meliloti*, a symbiotic soil bacterium associated with legume root nodules, where nitrogen is fixed (Cubo *et al.*, 2020).

4.4.3 Comparison of N4-like viruses with other abundant marine viruses

The read mappings described above were conducted using known N4-like viruses as references. We included more well-studied abundant marine viruses as references in order to understand the relative abundance of these viruses in the same samples. Four reference phages (*Pelagibacter phage* HTVC010P, *Puniceispirillum phage* HMO-2011, *Synechococcus phage* S-CBP4 and S-SK1) were chosen because these strains are known to be among the most abundant viruses in marine and estuarine environments (Sun *et al.*, 2021). When these four reference viruses were mixed with four N4-like viruses for read mapping, read recruitment to representative virus genomes showed that the N4-like viruses are much less abundant (range of 0-200 FPKM) compared to the four reference phages (range of 0-27,500 FPKM) in the estuarine and other marine samples (Fig. 4.7). When observing only the four N4-like viruses in this situation, it is evident that few reads map to N4-like viruses, although *Roseobacter phage* RD-1410W1-01 is still relatively abundant in Indian Ocean and the Southern Ocean (Fig. 4.7b). This result suggests that N4-like viruses are much less abundant compared to pelagiphages, cyanophages and phage HMO-2011 which are known abundant viral groups in the marine environment.

This result is striking considering a large amount of reads mapped to the 115 N4-like virus genomes with high identity (>90%) and coverage (up to 70% of the genome) when these four abundant phages were not included in read mapping (Fig. 4.3, Fig. 4.4). However, many fewer reads mapped to N4-like viruses when the abundant phages were included in the reference database (ambigous reads are set to map to the first best possible site) (Fig. 4.7). This indicates that the commonly used 90% identity threshold for viral metagenomic read mapping does not provide enough resolution to differentiate between different viral groups, and viral metagenomic read mapping results are highly dependent on the reference database provided. Our results suggest that the effects may be especially prominent for low abundance viruses such as N4-like viruses.

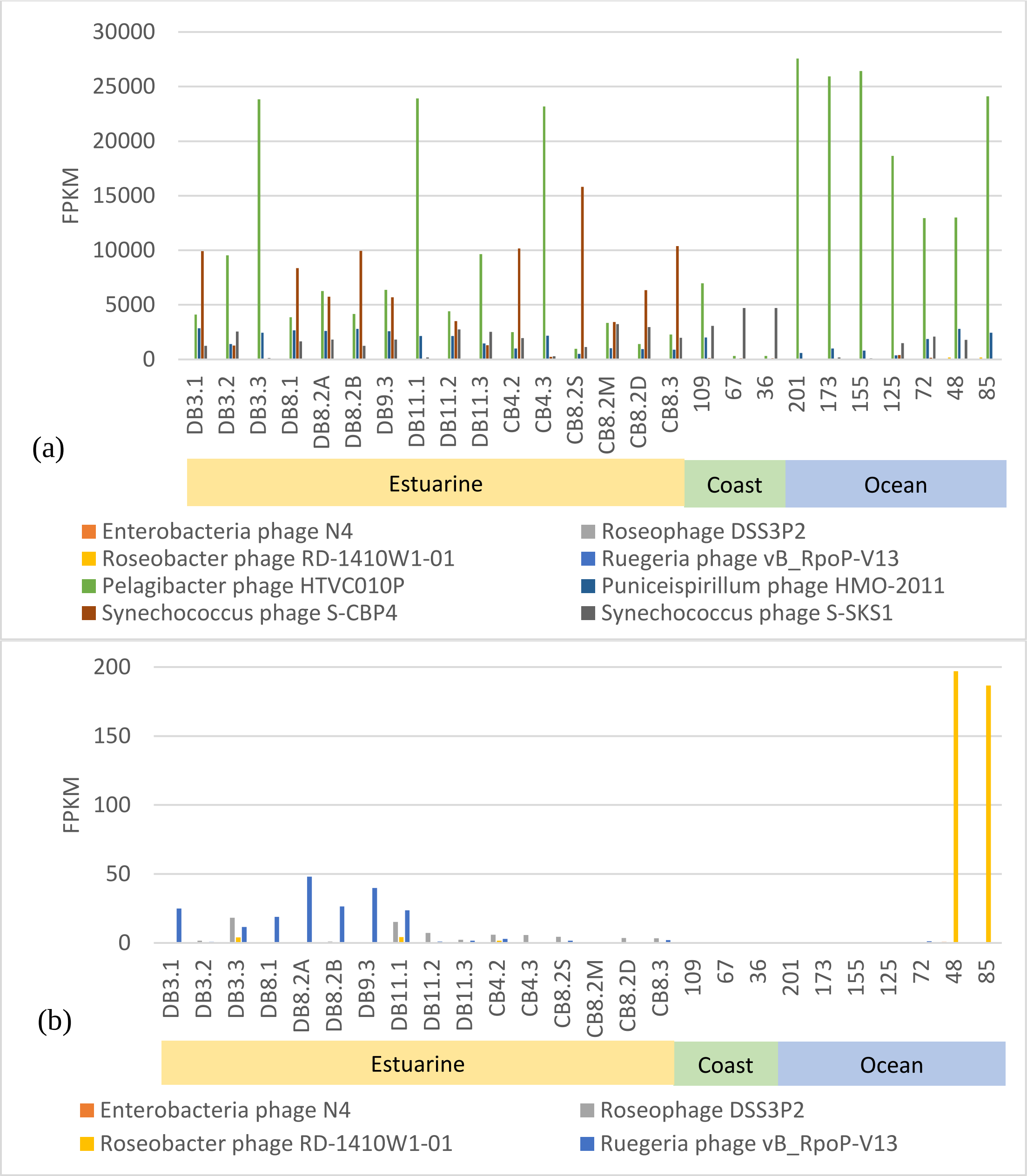

Figure 4.7 Comparison of select N4-like viruses, cyanophage, and pelagiphage in the Delaware Bay, Chesapeake Bay and reference sites (coastal and open ocean viromes). (a) Relative abundance of all eight viruses. (b) Zoomed in visualization of (a), showing only the four N4-like viruses.

The four non N4-like viruses are present in 15 out of 16 of the estuarine samples. Phage HTVC010P infects *Pelagibacter ubique* HTCC1062, and is often considered the most abundant virus in the marine environment (Zhao *et al.*, 2013; Martinez-Hernandez *et al.*, 2017). Indeed, HTVC010P was found to be the most abundant virus in most of the estuarine and coastal samples, and overwhelmingly abundant in the ocean samples (Fig. 4.7a). Phage HMO-2011 infects *Puniceispirillum marinum* of the abundant SAR116 clade, and is also thought to be among the most abundant viruses in the ocean (Kang *et al.*, 2013). HMO-2011 is less abundant than HTVC010P, and has less variation across different samples (Fig, 4.6a). *Synechococcus phage* S-CBP4 is abundant in most of the estuarine samples (11 out of 16), but not in the coastal and open ocean samples (Fig. 4.7). S-CBP4 is a podovirus which infects Chesapeake Bay *Synechococcus* strain CB0101 and was isolated from the middle Chesapeake Bay (Wang and Chen, 2008). This result suggests that the distribution of S-CBP1 is likely more confined to the estuarine ecosystem. In contrast, *Synechococcus* phage S-SKS1 which infects marine *Synechococcus* WH7803 is present in most estuarine, coastal and open ocean samples (Fig. 4.7a). S-SK1 is not present in sample 201 likely due to the very low temperature (-1 °C) in this sample. Abundance of *Synechococcus* can be lower than 1,000 cells per ml at this temperature, and the abundance of cyanophage co-varies with cell density of *Synechococcus* in the natural environment (Wang *et al.*, 2011).

4.4.4 N4-like contigs in the DEV

We also tried to use local N4-like contigs identified in the DEV viromes to explore their relative abundance in the DEV viromes. Eleven N4-like virus contigs from the DEV were identified using the vRNA polymerase gene of N4-like viruses, and the lengths of these contigs range from 8.8 kb to 73.7 kb (Table 4.3). Annotation using MG-RAST showed that the majority of genes in each of the contigs has a closest match to N4-like viruses, indicating that these contigs are indeed N4-like viruses or closely related to N4 (Table S3). Relative abundance of these N4-like contigs was derived from the results of mapping all DEV trimmed reads to all DEV viral contigs, then selecting the FPKM values assigned to the N4-like contigs. The overall relative abundance of these 11 N4-like contigs was relatively low (0-20 FPKM), with higher abundances in samples DB3.3 and DB11.1 (Fig. 4.8). Lower abundances of N4-like contigs were seen in the Chesapeake Bay compared to the Delaware Bay (Fig. 4.8). The low relative abundance of N4-like contigs confirms that N4-like viruses are rare in the estuary compared to the abundant viral groups such as cyanophage and pelagiphage.

Table 4.3 N4-like viral contigs.

Contig	**Length**	**Total fpkm (all viral contigs)**
Ga0070748_1000030	70,734	35.0472
Ga0070748_1000124	40,467	6.4641
Ga0070747_1000707	16,180	3.4784
Ga0070748_1001286	11,681	4.7523
Ga0070748_1000026	73,662	6.6385
Ga0070748_1000068	51,723	8.9789
Ga0070746_10000011	73,167	3.2182
Ga0070754_10005210	8,785	2.444
Ga0070748_1000096	44,402	7.5398
Ga0070748_1000074	50,797	18.0558
Ga0070747_1000830	15,197	2.8914

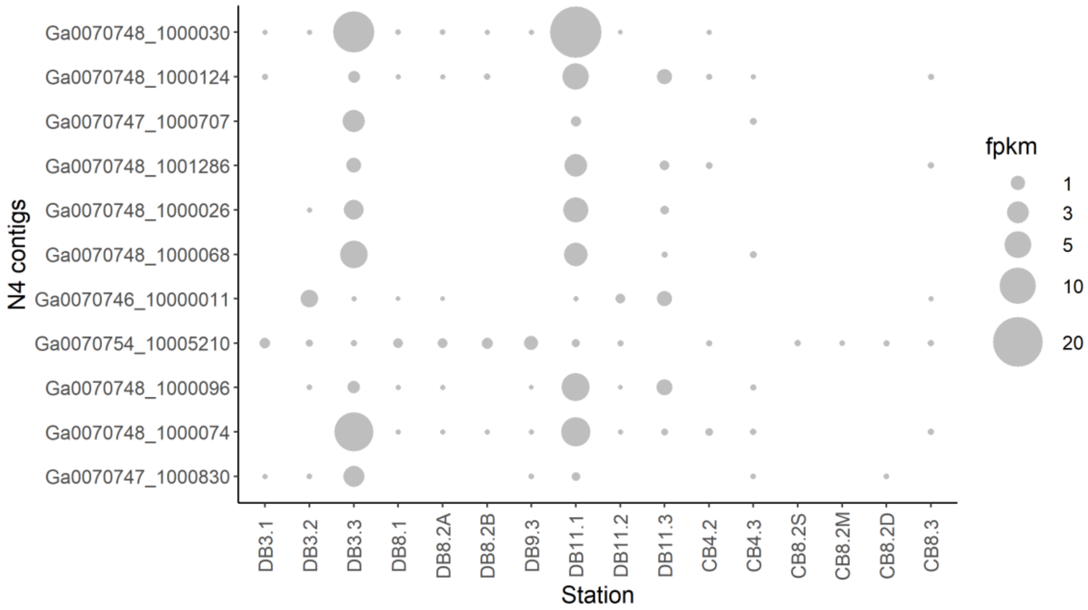

Figure 4.8 Relative abundance of N4-like contigs in DEV. Size of the bubbles correspond to the FPKM (Fragments Per Kilobase Million) for each sample.

Phylogenetic analysis using either terminase large subunit or vRNA polymerase is a good indicator of N4-like virus diversity, as confirmed by VIRIDIC analysis (Wittmann *et al.*, 2020). In this study, phylogenetic analysis based on the partial vRNA polymerase gene shows that the N4-like contigs in the DEV have a similar diversity to known N4-like viruses overall (Fig. 4.9) (Wittmann *et al.*, 2020). The most abundant N4-like contig (Ga0070748_1000030) found in the Delaware Bay did not cluster with other N4-like viruses on the phylogenetic tree, indicating that this type of N4-like virus may be relatively novel in the estuary (Fig. 4.8, Table 4.3, Fig. 4.9).

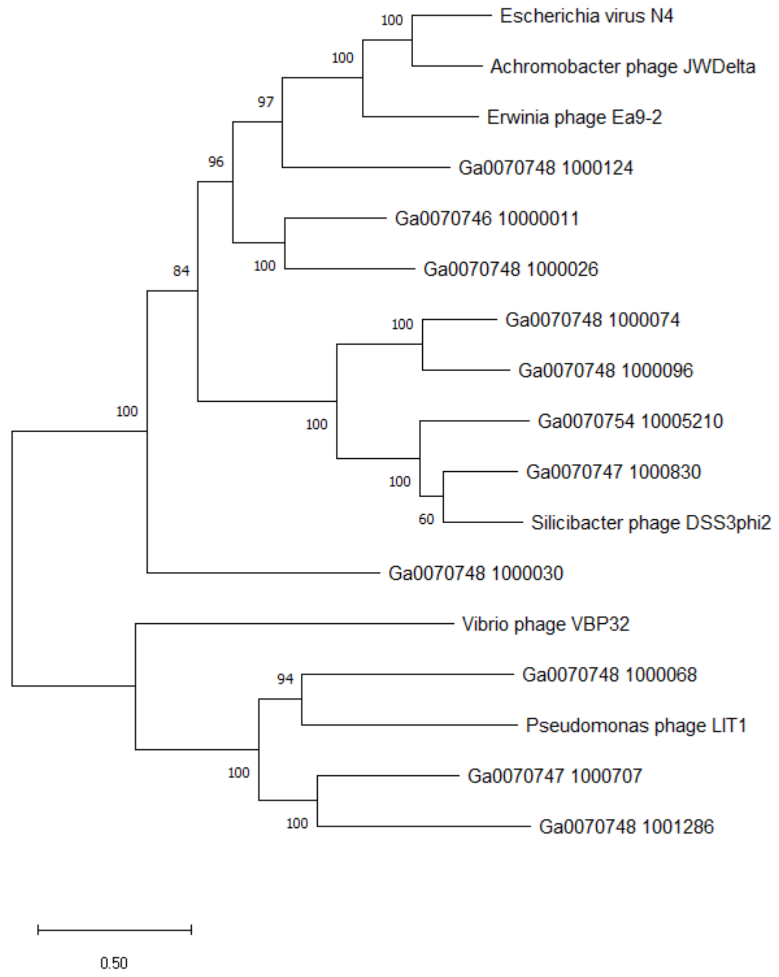

Figure 4.9 N4-like contigs evolutionary analysis using partial vRNA polymerase gene. The tree with the highest log likelihood is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. There were a total of 1,203 aa in the dataset.

4.4.5 Relative abundance of N4-like virus using reciprocal best hit BLAST

Reciprocal best hit BLAST using N4-like virus core genes revealed that N4-like viruses are present in low abundance in both estuaries, with the raw number of reciprocal hits ranging from 0 to 290 (Fig. 4.10). The distribution pattern of N4-like viruses based on core genes echoes that based on contig identification (Fig. 4.8). Overall, N4-like viruses were more prevalent in the Delaware Bay compared to the Chesapeake Bay. N4-like viruses see more patchy distribution in the Chesapeake Bay (Fig. 4.10). The Chesapeake Bay and Delaware Bay virome samples were taken in different years and different months. In addition, the average number of reads recovered per Delaware Bay sample (151 million) is over twice (2.2-fold) the amount of reads per average Chesapeake Bay sample (68 million) (Sun *et al.*, 2021). In the current study, the sequences were obtained from non-amplified viral DNA, in contrast to the previous study that used methods specifically targeted to recovering N4-like sequences (Zhan *et al.,* 2015). Since N4-like viruses are not abundant, the lower number of reads recovered per Chesapeake Bay sample may have caused the patchiness of N4-like viruses there, and suggest that the current sequencing depth may be near the detection limits for this rare virus group.

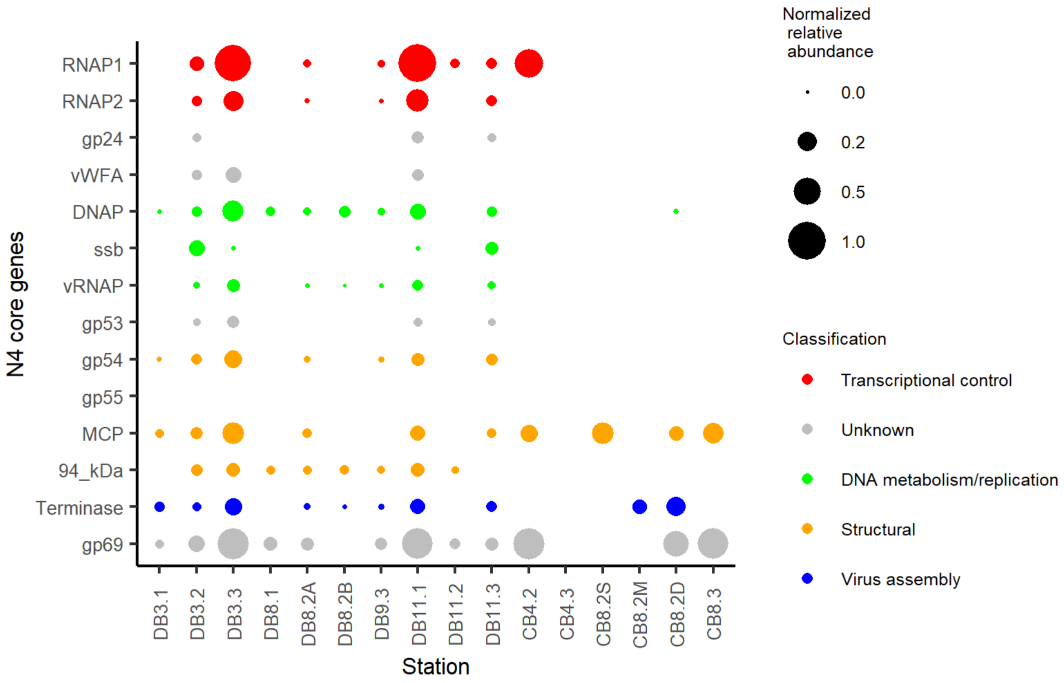

Figure 4.10 Relative abundance of N4-like viruses based on reciprocal best BLAST hits of core genes. Names of core genes are given in Table 4.2.

The results of reciprocal best hit using core genes show that in the Delaware Bay, almost all N4 core genes are more abundant in the March and November samples, although there is variation between different locations in the estuary (Fig. 4.10). The water temperature of March samples was about 4°C, while the water temperature of November samples varied between 13 and 15°C (Table 2.1). Much lower abundance of N4-like virus core genes was detected in the summer, when the water temperature was between 24 and 25°C. In the Chesapeake Bay, the water temperature was 8-10°C in April, and 26-27°C in August (Table 2.1).

Using reciprocal best hit BLAST methods, N4-like viruses were found to be more abundant in spring and fall, and less abundant in summer in the Delaware Bay (Fig. 4.10). The same pattern was observed using read mapping against N4-like contigs in the estuary (Fig. 4.8). The six Chesapeake Bay viromes were collected in April and August, and N4-like contigs were barely detected in August at all depths (Fig. 4.8). This matches our results using reciprocal best hit methods, showing that the majority of N4-like core genes have no matches in the Chesapeake Bay in April and August (Fig. 4.10). In an earlier study, N4-like viruses were not detected in the spring, summer and fall seasons, but were detectable in the winter in the Chesapeake Bay using PCR targeting the N4 DNAP gene (Zhan *et al.*, 2015). All the 12 samples in which N4-like viruses were detected were from the samples collected in February when surface water temperature was below 4°C (Zhan *et al.*, 2015). Given the limited amount of winter samples in our viromes, we are not able to confirm that N4-like viruses are indeed more prevalent in winter using reciprocal BLAST with core genes. However, the low abundance, or lack of N4-like virus core genes during the summer in both estuaries seems to explain the difficulty of detecting N4-like viruses in summer based on PCR (Zhan *et al.*, 2015). The prevalence of N4-like phage in cold water may be due to the unique advantages their vRNA polymerase offers in early transcription, and their large burst size (Zhan *et al.*, 2015). Since most of the Chesapeake Bay samples were taken in the summer, this may be why N4-like viruses are less likely to be found in the Chesapeake Bay in the DEV dataset (Fig. 4.10). The correlations between N4-like virus contigs, temperature and salinity were statistically evaluated. RDA showed that other than one outlier, the abundance of the majority of N4-like virus contigs are negatively correlated with temperature (Fig. 4.11). This supports previous findings that N4-like viruses are more prevalent in colder waters (Zhan *et al.*, 2015). The same study also found evidence that N4-like viruses are prevalent in saline environments such as Antarctic salt lakes where temperature is below -10℃ (Zhan *et al.*, 2015).

Core gene RNAP1 and a gene of unknown function, gp69 were found to be the most abundant compared to other genes (Table 4.3, Fig. 4.10). Despite the highly conserved nature of N4-like virus genomes, unknown gene gp55 was not found reciprocally in any of the samples (Table 4.3, Fig. 4.10). The reason for this is unclear, nevertheless, it indicates that gp55 may be a poor candidate for marker gene-based analyses of N4-like virus community. On the other hand, most of the other N4 core genes of known function (RNAP2, DNAP, vRNAP etc.) follow a similar abundance pattern to N4-like contigs, so marker gene-based investigations of N4-like viruses using these genes will likely yield accurate results, further validating the previous studies based on the N4 DNAP gene (Fig. 4.8, Fig. 4.10) (Zhao *et al.*, 2009; Zhan *et al.*, 2015).

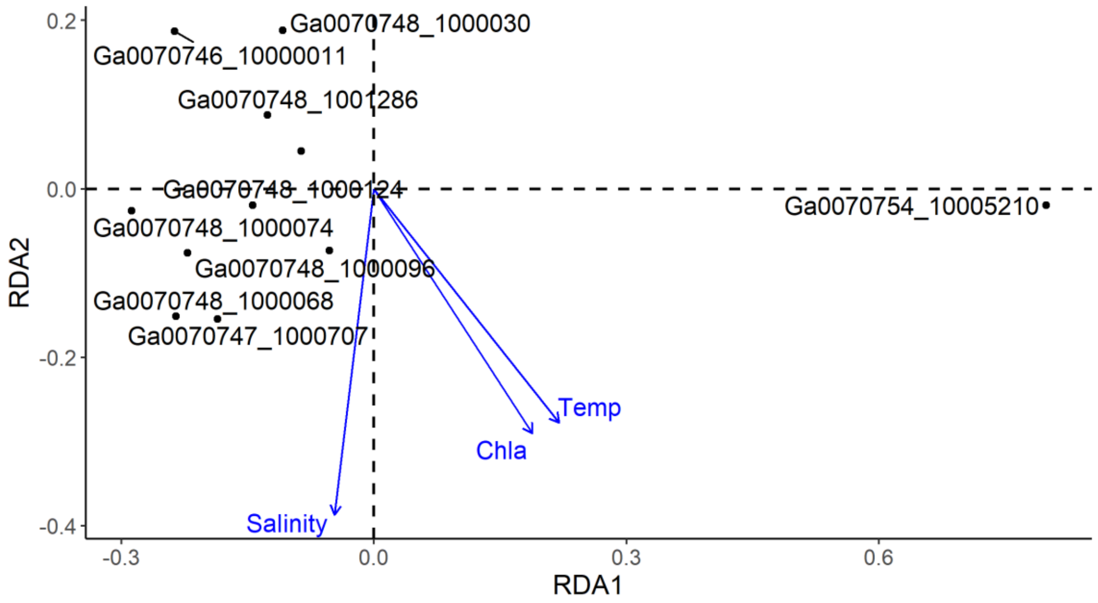

Figure 4.11 Redundancy analysis (RDA) ordination diagram (biplot) of N4-like virus contigs and environmental variables. RDA1 explains 18.9% of variance, while RDA2 explains 2.7% of variance. Total constrained variance is 22.2%, while unconstrained variance is 34.4%. Each black dot represents the variation of the relative abundance of said N4-like virus contig in the 16 samples. The angles between populations and environmental factors denote their degree of correlation.

4.4.6 N4-like viruses are rare in the aquatic environment

Previous studies only evaluated the prevalence of N4-like viruses, but not their relative abundance in natural environments (Zhao *et al.*, 2009; Chan *et al.*, 2014; Zhan *et al.*, 2015). Our results show that N4-like viruses are present in low abundance compared to pelagiphage and cyanophage in the Chesapeake Bay and Delaware Bay in March, April, August and November (Fig. 4.7a). Read mapping to local viral contigs show that the abundance of N4-like contigs was relatively low (0-20 FPKM) compared to the top viral contig in DEV (0-3600 FPKM) (Fig. 4.8) (Sun *et al.*, 2021). The low amount of reciprocal BLAST hits (0-290) also supports the rareness of N4-like viruses in the estuary (Fig. 4.10).

Low abundance of N4-like viruses was also observed when compared to other viral taxa in the DEV. Classification of DEV reads was made with Kaiju using viruses as the reference database (Menzel *et al.*, 2016). The detailed Kaiju methods can be found in (Sun *et al.*, 2021). Two N4-like virus strains were found among the classified species, and their percentage is visualized in Fig. 4.12. Around 10% of reads were identified as viral by Kaiju in DEV (Sun *et al.*, 2021), and N4-like virus percentages only go up to 0.0034%, indicating that only 0.00034% of reads are classified as N4-like viruses (Fig. 4.12). This method underestimates the abundance of N4-like viruses, since only two strains of N4-like virus were available in the Kaiju database at the time. Nonetheless, it provides support to the notion that N4-like viruses are rare in the environment.

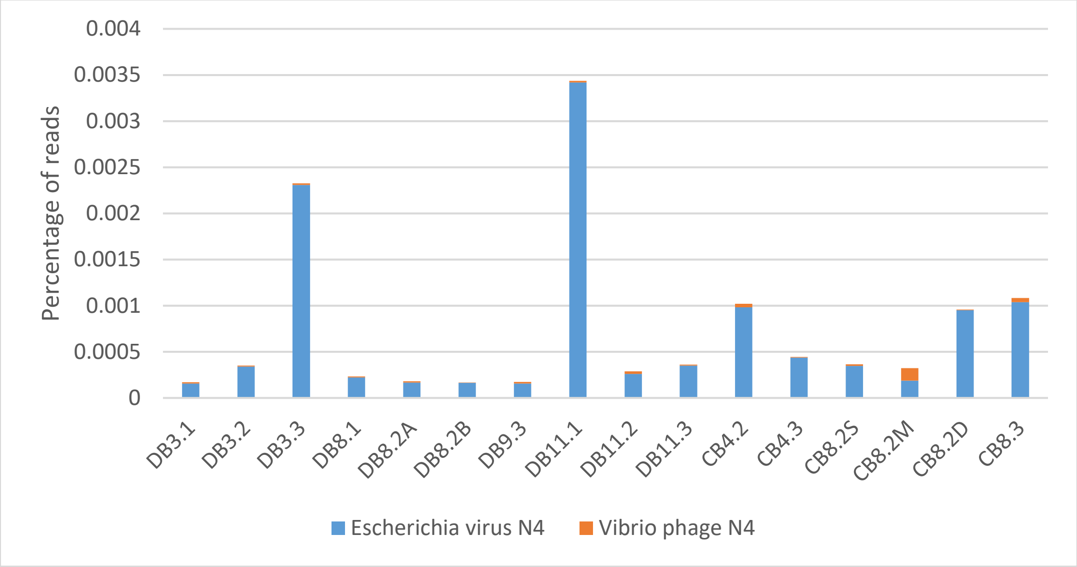

Figure 4.12 The percentages of DEV reads that are binned to N4-like viruses by Kaiju classification (Sun *et al.*, 2021).

4.4.7 Effect of methodology on understanding N4-like virus abundance

Searches of N4-like viruses in global metagenomic datasets over the years have yielded various conclusions regarding their environmental distribution. One study found them to be more prevalent in coastal waters compared to open ocean water (Zhao *et al.*, 2009), another found them to be widespread in both coastal and open ocean environments (Chan *et al.*, 2014), another study found them to be more prevalent in the winter season (Zhan *et al.*, 2015), and another study found them to be more dominant in waste water treatment plants compared to other aquatic environments (Parmar *et al.*, 2018). The rapidly increasing numbers of genomes of N4-like viruses over the past decade, and the expanding metagenomic databases may have affected the conclusions. Our results also suggest that the selection of reference genomes is important when evaluating the relative abundance of a given viral group (Fig. 4.7). Previous studies mostly searched for N4-like viruses in bacterial metagenomes from the GOS and the CAMERA database (Zhao *et al.*, 2009; Chan *et al.*, 2014; Zhan *et al.*, 2015). Meanwhile, this study searched for N4-like viruses exclusively in high throughput viral metagenomes, which contain mostly free-living viruses with deep sequencing coverage (~125 fold increase in bp yielded per sample compared to GOS) (Rusch *et al.*, 2007; Williamson *et al.*, 2008). In addition, the previous sequencing technologies used in GOS and CAMERA gave sequences with longer read length, which may also affect read recruitment results. Thus, the increased isolation of N4-like viruses and the shifting sequencing technology of available metagenomic datasets may have an impact on the conclusions.

In this study, different methods were used to evaluate the relative abundance of N4-like viruses (Fig. 4.2). Read mapping may cause overestimation of viruses due to the short read length (150 bp), even if mapping identity is high. The vastly different FPKM values resulting from mapping to all viral contigs vs. only known N4-like virus genomes indicates that a substantial number of ambiguous reads were produced, despite a 90% mapping identity cutoff and manual verification of high mapping coverage (Fig. 4.3, Fig. 4.7, Fig. 4.10). Since the read recruitment method to viral genomes is often used to evaluate the prevalence of various viruses in the environment (Bischoff *et al.*, 2019; Zaragoza-Solas *et al.*, 2020; Buchholz *et al.*, 2021; Z. Zhang *et al.*, 2021), caution should be taken when interpreting such results. Our results show that when provided with abundant viral genomes in the reference, most reads mapped to the dominant viruses instead (Fig. 4.7). This effect may be particularly pronounced when the virus group of interest is rare. We recommend including a few major groups of marine viruses when the read mapping is used for newly found viruses or contigs, especially when the virus is of low abundance in the environment.

The PCR method only detected N4-like viruses in winter (February), but not in in spring, summer and fall in the Chesapeake Bay (Zhan *et al.*, 2015), suggesting that N4-like viruses are relatively more abundant in winter compared to the rest of seasons. Furthermore, N4-like viruses in the Chesapeake Bay are least abundant (<1 FPKM, Fig. 4.8) compared to all the DEV samples (Fig. 4.8, Fig. 4.10). Unfortunately, DEV does not have samples from January or February which usually has the lowest water temperature throughout the year. The 12 positive PCR detections of N4-like viruses were all from samples collected in February (Zhan *et al.*, 2015). While the DEV viromes allow us to detect some N4-like viruses in warmer seasons, the limited number of samples from both bays and the rareness of N4-like viruses make the spatiotemporal pattern of N4-like viruses less conclusive. Unlike cyanophages or SAR11 phages, N4-like viruses are much less abundant in nature (Fig 4.6a). The inconsistent results based on the four different methods used in this study are likely due to the low abundance of N4-like viruses.

The reciprocal best hit BLAST method uses the collection of core genes, thus does not consider the non-conserved portion of the N4-like virus genome, lending it to show similar biases as PCR-based methods targeting marker genes. Meanwhile, read recruitment methods consider the entire genome, potentially covering a wider microdiversity compared to methods that only utilize conserved regions for identification. Although reciprocal best hit BLAST is a powerful and accurate method, it is computationally expensive when used on large scale datasets, and cannot be adapted to the very large file sizes of deep sequencing used today. The FPKM abundance of local N4-like contigs in the estuary follow a similar pattern to that derived from reciprocal BLAST based on core genes, with significantly higher abundance in samples DB3.3 and DB11.1, and lower abundance in the Delaware Bay summer and the Chesapeake Bay (Fig. 4.8, Fig. 4.10). The fact that similar population dynamics of N4-like viruses were observed using different approaches lends credibility to both methods used, and also emphasizes the importance of having local sequences. Since the mapping to local viral contigs yields similar results to reciprocal

BLAST, but is much less computationally intensive, we recommend this method for evaluating the relative abundance of rare viral groups.

In recent years, the accelerated isolation of novel N4-like viruses has greatly expanded their sequence space and provided more insight into their diversity and taxonomy, which is summarized well in Wittmann *et al.*, 2020. Most of the N4-like virus strains available so far were isolated from bacteria that are easy to culture in the laboratory, such as *Roseobacter*, *Enterobacter* and *Vibrio* (Wittmann *et al.*, 2020). Many novel uncultured N4-like viruses are present in nature (Zhan et al. 2015); to gain a better understanding of viruses in an environmental context, it is necessary to tackle the isolation and characterization of viruses from bacterial groups that are less well established in the lab. Deep sequencing of viromes has enabled us to identify rare viruses like N4-like viruses and retrieve their sequences. Since N4-like viruses have highly conserved genomes and a unique vRNA polymerase gene, they are good candidates for benchmarking detection methods for low abundance viruses in metagenomic datasets. Further investigation of environmental N4-like viruses using a combination of different methods is needed to piece together the mystery of N4-like virus abundance and their potential ecological role.

4.5 Concluding remarks

Read mapping based on the known N4-like virus isolates recruited substantial numbers of reads matching *Roseobacter* N4-like viruses with high identity and coverage in both the the Delaware Bay and Chesapeake Bay. A distinct distribution pattern of N4-like viruses was observed between the two estuaries and most of the marine reference sites. When a few well-studied viruses such as *Pelagibacter* phage, HMO-2011 and cyanophages were included for read mapping, the majority of reads mapped to these abundant viruses instead of N4-like viruses at high identity, indicating that N4-like viruses are of low abundance in estuarine, coastal and oceanic waters. The reasons that might cause different read recruitments on N4-like viruses with and without other abundant viral groups were discussed. Although the read mapping method based on known viruses is a common way to investigate the distribution of reference viruses, extra caution should be taken when searching for rare viral groups in viromes.

To better understand the spatiotemporal distribution of N4-like viruses in both estuaries, a contig-based recruitment method was used. Eleven large contigs of N4-like viruses were recovered among the DEV viral contigs, and their relative abundance patterns match the results of reciprocal best hit BLAST based on 14 N4-like virus core genes, showing relatively high abundance in colder seasons, and relatively low abundance in summer. Our analysis suggests that N4-like viruses are indeed more abundant in colder water in the natural environment, and confirms that N4-like viruses are indeed rare in the environment overall. RDA confirmed a negative correlation between the abundance of N4-like virus contigs and temperature. Phylogenetic analysis suggested that the most abundant N4-like viruses in the DEV viromes may be relatively novel compared to the existing N4-like viruses. The assessment using core genes provides insight on choice of genes when performing marker gene-based analyses of N4-like virus communities.

Chapter 5: Prevalence of SAR11 phage (pelagiphage) in estuarine environments

5.1 Abstract

Pelagibacterales (SAR11) is one of the most abundant bacterial orders in marine and freshwater environments. Viruses infecting members of Pelagibacterales (pelagiphages) dominate the global oceans, and play an important role in marine biogeochemical cycling. Pelagiphages of both freshwater and marine SAR11 have been reported. However, little is known about pelagiphage in estuaries and how they are distributed along the wide estuarine salinity gradient. In this study, we investigated the diversity and distribution of pelagiphage in two estuaries, the Chesapeake Bay and Delaware Bay. A total of 78 reference pelagiphage genomes were divided into eight distinct groups based on shared gene content. All eight groups of pelagiphage are present in both bays, and the pelagiphage community composition appears to be stable in the estuaries. Podoviruses, myoviruses and siphoviruses make up 96.2%, 3.6% and 0.2% of the estuarine pelagiphage community, respectively, a distribution pattern similar to their oceanic counterparts. Viruses related to uncultured phage vSAG-37-F6 was the most abundant (up to 34% of pelagiphages) in both estuaries. Freshwater pelagiphage were rare (< 2.4%) in the estuary, even in samples with low salinity. No clear transition between freshwater and oceanic pelagiphage ecotypes was seen in the estuaries. Despite the strong environmental gradients, no correlation was found between pelagiphage abundance and environmental factors in the estuarine environment. This is the first study to evaluate the pelagiphage

community in estuaries, and they were abundant and represent all 8 pelagiphage genome types. Our results are consistent with the presence of abundant and diverse SAR11 bacteria in the estuarine environment, suggesting that phage infecting SAR11 bacteria in estuaries are as important as those in open oceans.

5.2 Introduction

SAR11 bacteria (Pelagibacterales) are the most abundant cellular organisms in the world, and dominate global ocean waters (Morris *et al.*, 2002; Giovannoni, 2017). They are small, slow-growing, free-living bacteria with streamlined genomes, with genome size at around 1.3 Mb (Giovannoni, 2017). SAR11 bacteria are difficult to culture in the lab using conventional methods. Using the dilution-to-extinction method, different SAR11 genotypes in the marine environments have been cultivated and sequenced (Haro-Moreno *et al.*, 2020). The SAR11 clade typically makes up 20-50% of the ocean's bacterioplankton community (Giovannoni, 2017), and they play active roles in biogeochemical cycling in the ocean (Malmstrom *et al.*, 2004). In recent years, SAR11 bacteria from non-marine environments has also been cultivated; a freshwater SAR11 (*Fonsibacter* LD12) was isolated from the coastal lagoon of Lake Borgne and has the smallest genome size (1.16Mb) ever reported for a Pelagibacterales strain (Henson *et al.*, 2018).

5.2.1 Pelagibacter in the estuary

SAR11 bacteria are also abundant in estuarine environments such as the Chesapeake Bay, Delaware Bay, and Baltic Sea. SAR11 contributes up to 18% of the bacterioplankton community in the Chesapeake Bay (Kan *et al.*, 2007). A more recent amplicon-based study found SAR11 to comprise up to 16% of the bacterial community in the Chesapeake Bay across various seasons (Fig. 5.2) (H. Wang *et al.*, 2020). The dominance of SAR11 in the Delaware Bay was also reported (Kirchman *et al.*, 2005; Campbell and Kirchman, 2013). In the Delaware Bay, SAR11 makes up <10% of the total bacterioplankton community in the river sites, and >20% in the higher salinity sites (Kirchman *et al.*, 2005). A subsequent study found that SAR11 accounts for up to 70% of the bacterial community in the Delaware Bay, with lower abundance in the upper estuary and higher abundance in the mid-and lower estuary (Campbell and Kirchman, 2013). The transcriptional activity of SAR11 bacteria increased from the upper bay to lower bay, although their overall growth rate and transcriptional activity is low (Campbell *et al.*, 2011; Campbell and Kirchman, 2013). The SAR11 type varied along the salinity gradient in estuaries (Campbell and Kirchman, 2013; H. Wang *et al.*, 2020). SAR11 is also abundant in the Baltic Sea, where SAR11-IIIa can make up 35% of the bacterial community in the oligohaline–mesohaline region where salinity ranges from 2.7–13.3 ppt (Herlemann *et al.*, 2014). Other SAR11 types (SAR11-I/II) were more abundant (27% of total bacteria) in the marine parts of the Baltic Sea, while the freshwater lineage LD12 was not detected in any stations (Herlemann *et al.*, 2014). On a global scale, SAR11 has a strong marine profile, being significantly more abundant in marine and estuarine environments compared to freshwater environments (Hugerth *et al.*, 2015). However, in the Delaware Bay, there is no significant correlation between overall SAR11 abundance and environmental factors including salinity, despite the strong environmental gradients present in the estuary (Campbell and Kirchman, 2013).

5.2.2 Pelagiphages

The abundance of SAR11 bacteria and their important role in the aquatic environment has sparked interest in the viruses that infect members of SAR11, or pelagiphages. SAR11 bacteria were initially thought to be immune to viral predation, and SAR11 phage were entirely missed in viromic datasets, until pelagiphages were first discovered in 2013, and were also found to be highly abundant in the ocean (Zhao *et al.*, 2013). The discovery of pelagiphage and their abundance in the ocean indicate that SAR11 bacteria are susceptible to viral infection, and suggest that pelagiphage contribute to the success of SAR11 through co-evolution (Zhao *et al.*, 2013). *Pelagibacter phage* HTVC010P, one of the first four pelagiphages discovered, is often cited as the most abundant viral species in the marine environment (Zhao *et al.*, 2013; Wu *et al.*, 2020). HTVC010P can reach up to absolute abundances of up to 10^5/ml ddPCR (droplet digital PCR) copies in seawater, and can consist up to 50% of identified viral species in surface water (Eggleston and Hewson, 2016; Martinez-Hernandez *et al.*, 2017; Martinez-Hernandez, Garcia-Heredia, *et al.*, 2019). As of now, 44 pelagiphages have been isolated, and over a hundred complete assembled genomes have been identified in microbial community databases (Table 5.1). Prophages have also been identified in two marine *Pelagibacter* strains, and they were suggested to contribute to the evolutionary success of SAR11 in oligotrophic waters (Morris *et al.*, 2020).

Since bacteria in the SAR11 clade are slow-growing bacteria and difficult to culture, phages that infect SAR11 bacteria have been difficult to study (Z. Zhang *et al.*, 2021). Isolation of *Pelagibacter* phage typically involves detection of host lysis using flow cytometry, and purification of phage using dilution-to-extinction (Zhao *et al.*, 2013; Buchholz *et al.*, 2021). Non-culture based methods have also been proven to be important in pelagiphage discovery. One of the most abundant virus species in the world, vSAG 37-F6, was discovered using cultivation-independent single-virus genomics technology (Martinez-Hernandez *et al.*, 2017). Its host was found to be pelagibacter using cultivation-independent single-cell genomics, confirming vSAG 37-F6 to be a pelagiphage (Martinez-Hernandez, Fornas, *et al.*, 2019). The abundance of vSAG 37-F6 is generally on par with HTVC010P, consisting up to 10^5/ml ddPCR copies in surface water (Martinez-Hernandez, Garcia-Heredia, *et al.*, 2019).

5.2.3 Pelagiphage in freshwater and estuaries

In addition to being abundant in the marine environment, a SAR11 bacterium has been cultivated in freshwater (*Fonsibacter* LD12), and the prevalence of LD12 has a negative correlation with salinity on a global scale (Henson *et al.*, 2018). A few prophage genomes have also been identified and reconstructed from *Fonsibacter* genomes assembled via metagenomic analysis (Chen *et al.*, 2019). These freshwater pelagiphages were found to be associated with freshwater habitats, although some were found in estuaries such as the Delaware Bay, the San Francisco Bay, and the Columbia River estuary (Chen *et al.*, 2019). Freshwater pelagiphage were also found among reconstructed myophage genomes (Zaragoza-Solas *et al.*, 2020). Although no lytic phages have been isolated from freshwater SAR11, these studies indicate that SAR11 viruses are present in the freshwater environment. Despite the fact that SAR11 and pelagiphages have been found in both freshwater and marine water, little is known about the diversity and ecological distribution of pelagiphage in the estuarine environment where the transition of freshwater to seawater forms a strong salinity gradient.

In this study, we collected 78 reference pelagiphage genomes from various studies, categorized them based on their genomic similarities, and searched for their presence and distribution in the viromic datasets (DEV) from the Delaware Bay and Chesapeake Bay (Sun *et al.*, 2021). The reference genomes include 33 genomes derived from uncultured MAGs or SAGs, in order to gain a comprehensive view on the diversity of pelagiphage. Representative viromes from freshwater and marine systems were also included in this study. We found that pelagiphage in estuarine environments are as abundant and diverse as those in the open ocean.

5.3 Methods

5.3.1 Relative abundance of SAR11 in the Chesapeake Bay

The relative abundance of SAR11 was derived from 16S rDNA data from the Chesapeake Bay in multiple locations and years, and visualized using ggplot2 in R (Ginestet, 2011; H. Wang *et al.*, 2020). Multiple linear regression was performed between the relative abundance of SAR11 and environmental factors (temperature and salinity).

5.3.2 Shared gene analysis of known pelagiphages

To evaluate the diversity of known pelagiphages, 78 existing pelagiphage genomes were chosen (Table 5.1) from various studies (Zhao *et al.*, 2013, 2019; Chen *et al.*, 2019; Martinez-Hernandez, Fornas, *et al.*, 2019; Morris *et al.*, 2020; Zaragoza-Solas *et al.*, 2020; Buchholz *et al.*, 2021; Z. Zhang *et al.*, 2021). The shared genes of these

78 pelagiphages were used to determine their relationship. Prodigal (Hyatt *et al.*, 2010) and GeneMark (Borodovsky and McIninch, 1993) were used for phage ORF prediction. Genes with ⩾25% amino acid identity, ⩾50% alignment coverage of the shortest protein, and an E-value cutoff of ⩽1E-3 were considered putative homologues. The percentage of shared genes between the pelagiphage genomes was calculated from BLASTP comparison. The heatmap of the shared genes between all the isolated pelagiphages was plotted using pheatmap package in R. The comparative genome map and connections between homologous genes were visualized using Easyfig (Sullivan *et al.*, 2011).

Table 5.1 The 78 reference pelagiphages used in this study. Pelagiphages that have no host listed are from uncultured assembled sequences.

Name	Host	Group	Alternative group name	Genome size (bp)	G + C %	Reference
EXVC010P	H2P3α	A	019P-type	41069	33.5	Buchholz 2021
EXVC011P	H2P3α	A	019P-type	41069	33.5	Buchholz 2021
EXVC012P	H2P3α	A	019P-type	41529	34.1	Buchholz 2021
EXVC014P	H2P3α	A	019P-type	41529	34.1	Buchholz 2021
EXVC015P	H2P3α	A	019P-type	41069	33.5	Buchholz 2021
EXVC018P	HTCC1062	A	019P-type	38005	32.6	Buchholz 2021
EXVC019P	H2P3α	A	019P-type	41069	33.5	Buchholz 2021
EXVC020P	HTCC1062	A	019P-type	37857	32.5	Buchholz 2021
EXVC025P	HTCC1062	A	019P-type	39638	32.7	Buchholz 2021
HTVC021P	HTCC1062	A	019P-type	39921	32	Zhang 2020
HTVC022P	HTCC1062	A	019P-type	42102	34	Zhang 2020
HTVC025P	HTCC1062	A	019P-type	42809	33.5	Zhao 2019
HTVC031P	HTCC1062	A	019P-type	42010	34.2	Zhao 2019
HTVC201P	FZCC0015	A	019P-type	37251	32.5	Zhao 2019
HTVC200P	FZCC0015	A	019P-type	41046	33.4	Zhao 2019
HTVC121P	HTCC7211	A	019P-type	41415	33.1	Zhao 2019
HTVC105P	HTCC7211	A	019P-type	42221	33.2	Zhao 2019
HTVC109P	HTCC7211	A	019P-type	42600	33.5	Zhao 2019
HTVC119P	HTCC7211	A	019P-type	42835	33.5	Zhao 2019
HTVC120P	HTCC7211	A	019P-type	41323	35.5	Zhao 2019
HTVC019P	HTCC1062	A	019P-type	38357	32	Zhao 2013
HTVC011P	HTCC1062	A	019P-type	42622	33.3	Zhao 2013
uv-Fonsiphage-EPL		A	019P-type	39413	32.1	Chen 2019
HTVC008M	HTCC1062	B	Myoviridae	147284	33.5	Zhao 2013
PMP-MAVG-1		B	Myoviridae	118124	33.71	Zaragoza 2020
PMP-MAVG-10		B	Myoviridae	127706	32.6	Zaragoza 2020
PMP-MAVG-11		B	Myoviridae	141312	34.54	Zaragoza 2020
PMP-MAVG-12		B	Myoviridae	104791	33.36	Zaragoza 2020
PMP-MAVG-13		B	Myoviridae	155847	34.2	Zaragoza 2020
PMP-MAVG-14		B	Myoviridae	136460	32.92	Zaragoza 2020
PMP-MAVG-15		B	Myoviridae	144833	31.3	Zaragoza 2020
PMP-MAVG-16		B	Myoviridae	132453	32.99	Zaragoza 2020
PMP-MAVG-17		B	Myoviridae	149073	34.51	Zaragoza 2020
PMP-MAVG-18		B	Myoviridae	153977	32.58	Zaragoza 2020
PMP-MAVG-19		B	Myoviridae	149077	34.83	Zaragoza 2020
PMP-MAVG-2		B	Myoviridae	139426	32.4	Zaragoza 2020

PMP-MAVG-20		B	Myoviridae	122912	31.08	Zaragoza 2020
PMP-MAVG-21		B	Myoviridae	135163	31.59	Zaragoza 2020
PMP-MAVG-22		B	Myoviridae	103989	34.17	Zaragoza 2020
PMP-MAVG-23		B	Myoviridae	110977	34.96	Zaragoza 2020
PMP-MAVG-24		B	Myoviridae	116502	34.74	Zaragoza 2020
PMP-MAVG-25		B	Myoviridae	142712	31.7	Zaragoza 2020
PMP-MAVG-26		B	Myoviridae	142788	32.48	Zaragoza 2020
PMP-MAVG-3		B	Myoviridae	147773	32.66	Zaragoza 2020
PMP-MAVG-4		B	Myoviridae	179730	32.04	Zaragoza 2020
PMP-MAVG-5		B	Myoviridae	149934	33.6	Zaragoza 2020
PMP-MAVG-6		B	Myoviridae	135833	33.58	Zaragoza 2020
PMP-MAVG-7		B	Myoviridae	135598	33.82	Zaragoza 2020
PMP-MAVG-8		B	Myoviridae	118694	31.91	Zaragoza 2020
PMP-MAVG-9		B	Myoviridae	124621	33.95	Zaragoza 2020
BMMRE_07242016_10_scaffold_124		C	010P-type	27140	31.6	Chen 2019
EPL_06132017_6.25m_HTVC010P-related_33_76		C	010P-type	35816	32.5	Chen 2019
EPL_08022017_1.5m_HTVC010P-related_32_16		C	010P-type	36457	31.9	Chen 2019
EXVC021P	HTCC1062	C	010P-type	34916	31.5	Buchholz 2021
HTVC010P	HTCC1062	C	010P-type	34892	29.7	Zhao 2013
HTVC028P	HTCC1062	C	010P-type	36388	33.1	Du 2021
HTVC203P	FZCC0015	C	010P-type	34938	32.1	Du 2021
HTVC034P	HTCC1062	C	010P-type	35450	32.6	Du 2021
HTVC035P	HTCC1062	C	010P-type	36066	31.9	Du 2021
HTVC024P	HTCC1062	C	010P-type	35448	31.5	Du 2021
HTVC204P	FZCC0015	C	010P-type	34069	31	Du 2021
HTVC100P	HTCC7211	C	010P-type	34605	31.8	Du 2021
I-EPL_09192017_0.5m_HTVC010P-related_33_10		C	010P-type	36507	32.5	Chen 2019
Lake_Mendota_HTVC010P-related_phage		C	010P-type	35984	31.83359	Chen 2019
PNP1	NP1	C	010P-type	35831	32.58352	Morris 2020
HTVC106P	HTCC7211	D	HTVC106P-type	36945	32.1	Zhang 2020
HTVC023P	HTCC1062	E	HTVC023P-type	60878	35	Zhang 2020
HTVC027P	HTCC1062	E	HTVC023P-type	57595	34.8	Zhang 2020
vSAG-37-F6		E	HTVC023P-type	13783	37.64057	Martinez-Hernandez 2017
HTVC111P	HTCC7211	F	HTVC111P-type	31577	30.5	Zhang 2020

HTVC112P	HTCC7211	F	HTVC111P-type	32478	30.4	Zhang 2020
HTVC026P	HTCC1062	F	HTVC111P-type	32480	31.3	Zhang 2020
HTVC202P	FZCC0015	F	HTVC111P-type	32226	31.3	Zhang 2020
HTVC103P	HTCC7211	G	HTVC103P-type	54103	31	Zhang 2020
HTVC104P	HTCC7211	G	HTVC103P-type	54359	30.9	Zhang 2020
HTVC115P	HTCC7211	G	HTVC103P-type	54819	33.2	Zhang 2020
EXVC013S	H2P3α	H	Siphoviridae	18297	31.1	Buchholz 2021
EXVC016S	H2P3α	H	Siphoviridae	48659	30.5	Buchholz 2021

5.3.3 Phylogenomic analysis of known pelagiphages

The whole genome phylogenetic tree of the 78 known pelagiphages based on amino acid sequences was constructed using the Virus Classification and Tree Building Online Resource (VICTOR) (Meier-Kolthoff and Göker, 2017) with the Genome-BLAST Distance Phylogeny (GBDP) method (Meier-Kolthoff *et al.*, 2013) under recommended settings for prokaryotic viruses, 100 pseudo-bootstrap replicates. The phylogenetic trees were visualized using iTOL (Letunic and Bork, 2019).

5.3.4 Relative abundance of known pelagiphages

To estimate the relative abundance of pelagiphage sequences in different aquatic environments, metagenomic reads from 16 estuarine DEV samples and 10 publicly available offshore water samples (Table 4.1, Fig. 4.1 in Chapter 4) were mapped to the 78 pelagiphage genomes, using BBMap with the mapping parameters as recommended in viromic benchmarking studies (>90% identity, >75% contig length) (Bushnell, 2014; Roux *et al.*, 2017). Ambiguous reads (reads that map equally well to multiple sites) were set to map to the first best possible site. The offshore viral sequences were taken from Global Ocean Virome (GOV) 2.0 (Gregory *et al.*, 2019). All chosen samples were taken from surface water. These viral metagenomes were chosen because they were obtained using similar sampling and sequencing technology. Trimmed reads were counted and normalized to FPKM (Fragments Per Kilobase Million) using SAMtools (Li *et al.*, 2009). FPKM is used as a proxy for relative abundance (Roux *et al.*, 2017). The abundance values of each group were combined, and visualized using ggplot2 in R (Ginestet, 2011).

5.3.5 Identification of pelagiphage contigs

To identify pelagiphage contigs in DEV, the 78 known pelagiphage genomes were used as reference. The pelagiphage genomes were used as query and aligned against DEV viral populations using BLASTN with evalue= 1E-10 and perc_identity=80 (Altschul *et al.*, 1990; Sun *et al.*, 2021). Contigs were considered to be pelagiphages if they match a known pelagiphage with least 90% identity across ⩾50% of the sequence length, and at least one hit over 1 kb in length. BLASTN results were parsed with the "Parse_BLAST" script with modified parameters from (Paez-Espino, Pavlopoulos, *et al.*, 2017). The identity of the resulting contigs was assigned to their closest match to a known pelagiphage genome.

5.3.6 Relative abundance of pelagiphage contigs

Relative abundance of the pelagiphage contigs were derived via read mapping, using the methods described in a previous study (Sun *et al.*, 2021). Briefly, trimmed reads from DEV were mapped to all viral contigs, and the FPKM values of pelagiphage contigs were used as a proxy for their relative abundance. The FPKM values of the pelagiphage contigs in the 16 estuarine samples were plotted in R using ggplot2 (Ginestet, 2011).

5.3.7 Correlation between pelagiphage contigs and environmental factors

To explain the effect of environmental and biological factors (temperature, salinity, Chlorophyll a, NO_3^-, NH_4^+, PO_4^{3-}, SiO_4^{2-}) on the abundance of pelagiphage contigs, redundancy analysis (RDA) was plotted for the pelagiphage using the vegan package in R, and visualized using type I scaling in ggplot2 (Ginestet, 2011; Oksanen *et al.*, 2018). The RDA biplot explains the variation of the abundance of the pelagiphage contigs samples using temperature and salinity.

5.4 Results and discussion

5.4.1 Pelagiphage can be classified into eight groups

To understand the diversity of existing pelagiphages, 78 known pelagiphage genomes were compared to each other using shared genes and whole genome phylogeny. The 78 known pelagiphages include 27 myoviruses, 49 podoviruses, and 2 siphoviruses. Forty-four of these 78 pelagiphages are cultured isolates, while the remaining 34 are complete genomes assembled from environmental sequences. The majority of these pelagiphages (70 of 78) were obtained from oceans, and eight of them are from freshwater (Table 5.1). The shared gene analysis clustered the 78 pelagiphages into eight distinct groups, with one group of myoviruses, one of siphoviruses, and six groups of podoviruses (Fig. 5.1, Table 5.2). The whole genome phylogenomic analysis further supported the division of these eight pelagiphage groups (Fig. 5.2, Fig. 5.3, Table 5.2). Pelagimyoviruses were placed in a separate tree from siphoviruses and podoviruses, because pelagimyoviruses have much a larger genome size, resulting in low similarity with other pelagiphages. Only one genome of pelagimyovirus is from an isolated phage (strain HTVC008M), the other 26 pelagimyovirus genomes are cross-assembled from various sequence databases (Table 5.2) (Zhao *et al.*, 2013; Zaragoza-Solas *et al.*, 2020). Other than group H (pelagisiphoviruses), the seven other groups have been reported before, albeit with much fewer members (Z. Zhang *et al.*, 2021). This study further extends our understanding of the diversity of the pelagiphage groups.

The majority of these 78 pelagiphages belong to group A, B or C. The detailed diversity within these three large groups of pelagiphages have been reported in recent studies (Du *et al.* 2021, Zaragoza-Solas *et al.*, 2020; Zhang *et al.*, 2020). Group A share more homologous genes compared to group B and C (Fig. 5.1). Uncultured single virus isolate vSAG 37-F6, one of the most abundant viruses in the marine environment, was shown to be related to HTVC023P and HTVC027P (Fig. 5.2). The *Pelagibacter* phages HTCC1062, HTVC023P and HTVC027P were isolated from South China Sea and South Pole, respectively. The relationship between vSAG 37-F6 and HTVC023P/HTVC027P has been reported in a recent study through comparative genomics (Z. Zhang *et al.*, 2021), and the our results based on the phylogenomic analysis concur with the kinship of these three pelagiphages in group E.

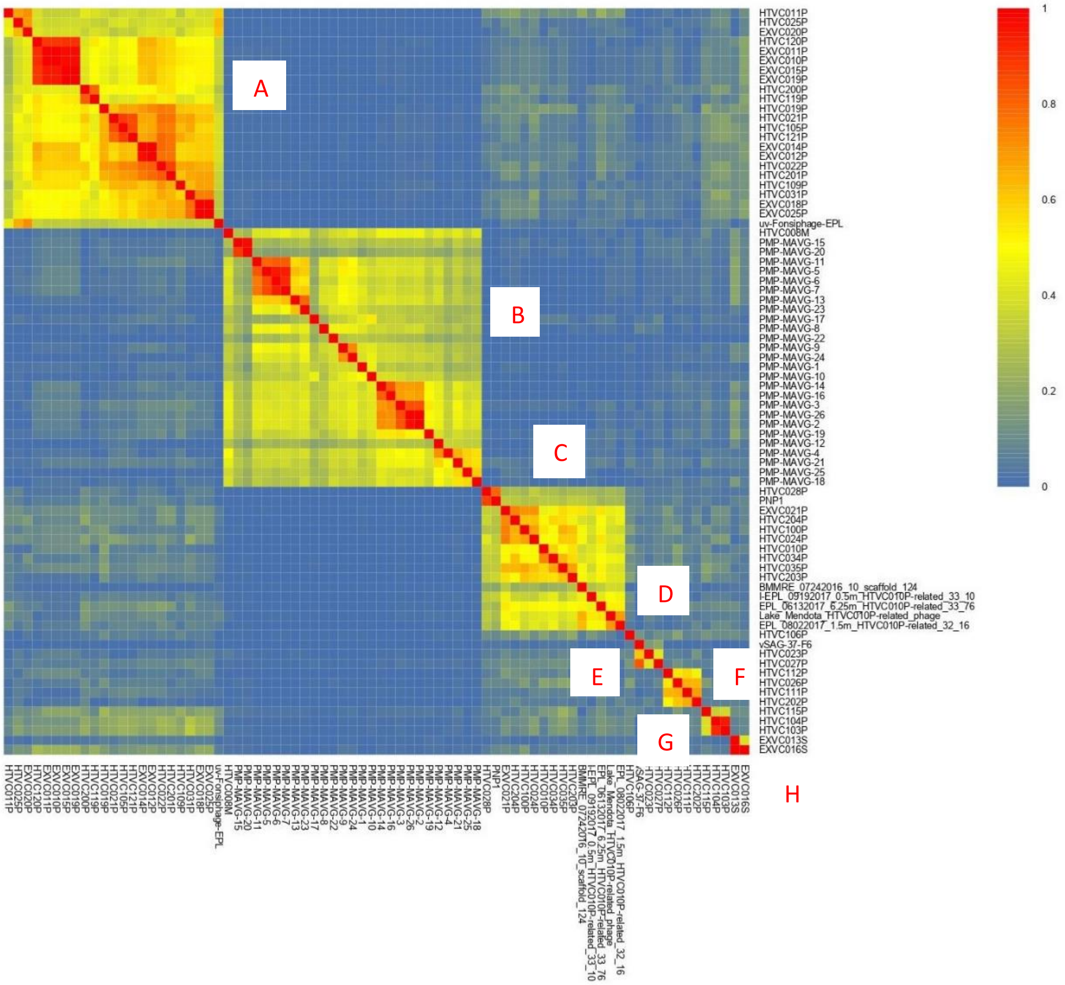

Figure 5.1 Heatmap of the shared genes of 78 pelagiphages. The color scale is the percentage of shared genes between genomes, normalized on a 0-1 scale. They are categorized into eight groups named A-H.

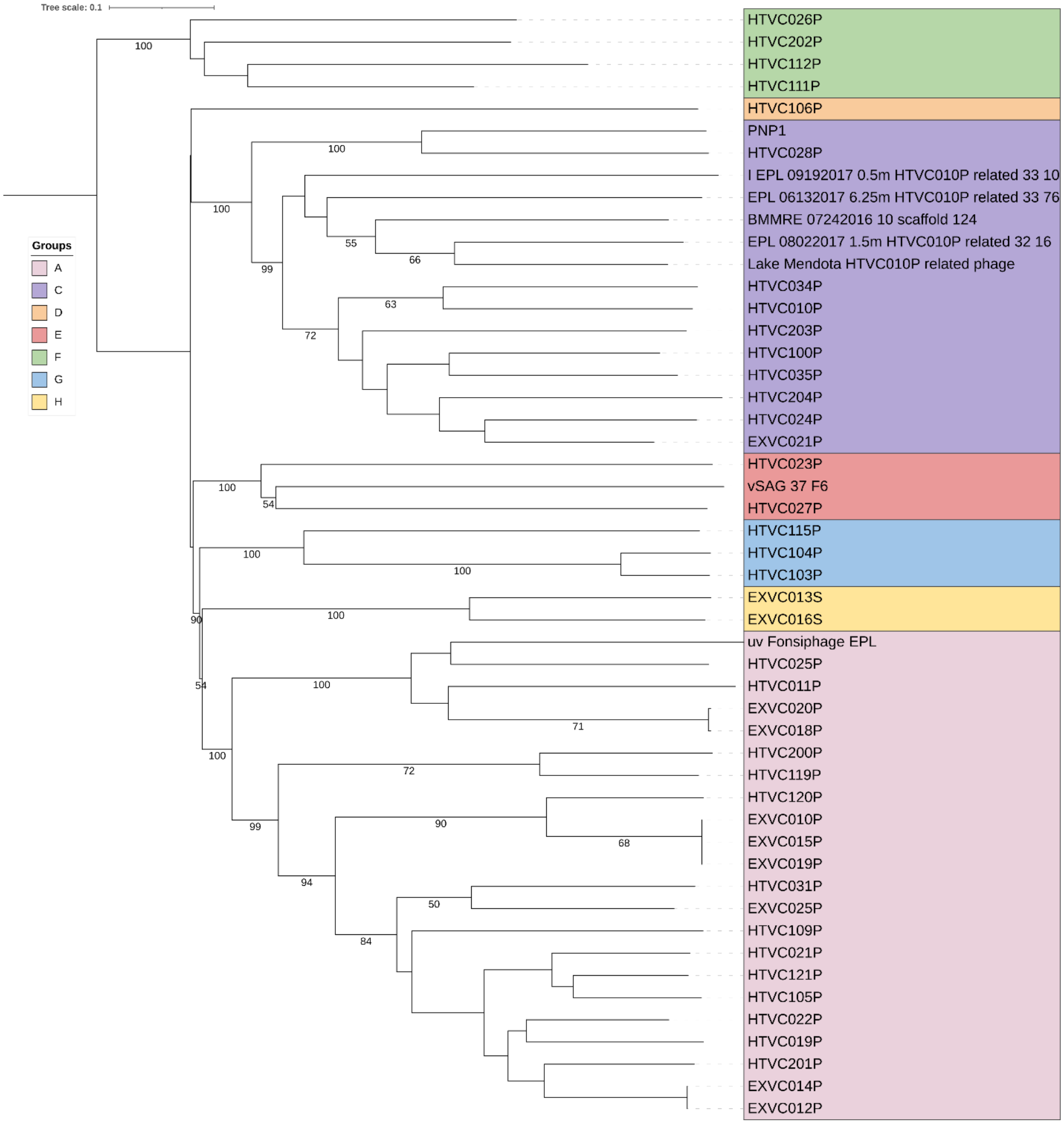

Figure 5.2 Whole-genome-based phylogenetic tree of pelagipodoviruses and pelagisiphoviruses constructed by VICTOR with the formula D6 (Meier-Kolthoff and Göker, 2017). Bootstrap values under 50 were omitted. Seven of the eight groups are indicated.

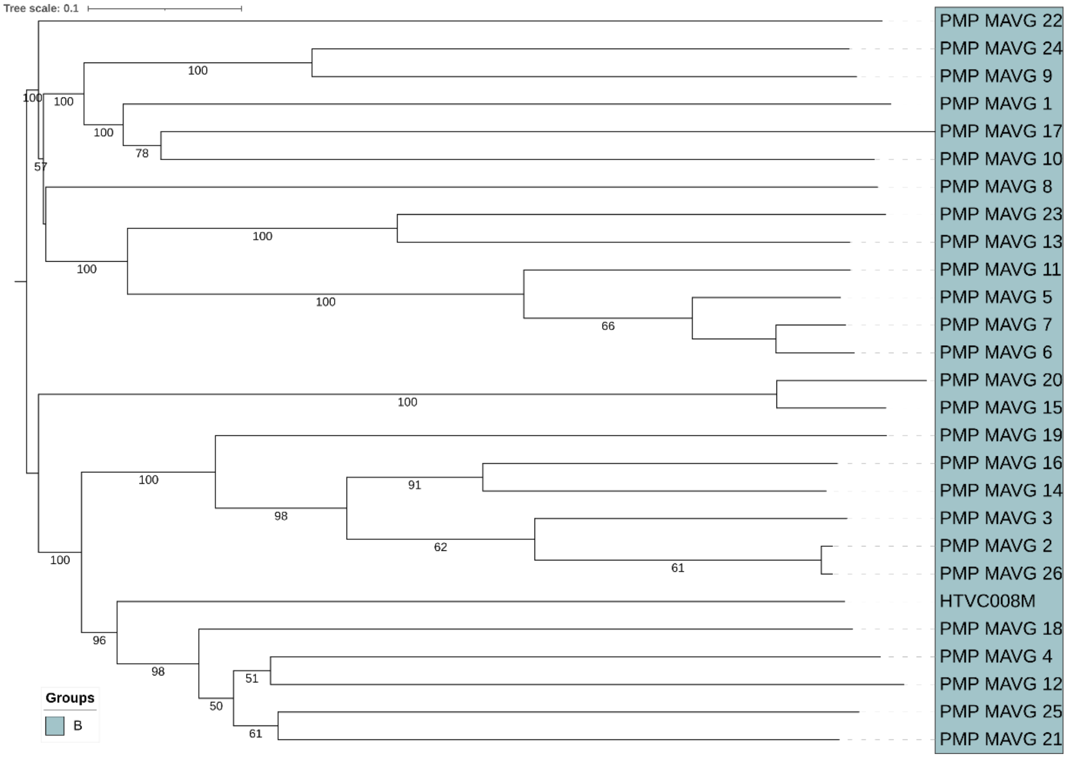

Figure 5.3 Whole-genome-based phylogenetic tree of group B (pelagimyoviruses) constructed by VICTOR with the formula D6 (Meier-Kolthoff and Göker, 2017). Bootstrap values under 50 were omitted. Pelagimyoviruses consist one of the eight groups.

Table 5.2 Information about the eight pelagiphage groups.

Group	**Group name**	**No. of members**
A	HTVC109P-type	23
B	Myovirus	27
C	HTVC010P-type	15
D	HTVC106P-type	1
E	HTVC023P-type	3
F	HTVC111P-type	4
G	HTVC103P-type	3
H	Siphovirus	2

5.4.2 Relative abundance of pelagiphage groups

Relative abundances of the 78 pelagiphages in the DEV and ocean samples were catergorized into the eight groups and shown in Figure 5.4. Groups C, D, E, and F (all podoviruses) are more abundant than group A, B, G, and H in estuaries, while group C and E dominate the coastal and oceanic water. Group E includes vSAG 37-F6, the most abundant pelagiphage in the coastal and open ocean, which is also abundant in the Chesapeake Bay and Delaware Bay (Fig. 5.4). This result suggests that the group E pelagiphage is prevalent and widely distributed in the estuarine, coastal and oceanic environment. Group D, consisting of 1 pelagiphage HTVC106P, is more abundant in the estuarine water compared to offshore water (Fig. 5.4). HTVC106P does not have a clear relationship to any known virus isolates, and its genome organization differs from other pelagiphage, lacking DNA replication genes, suggesting its dependence on host replication systems (Z. Zhang *et al.*, 2021). The increased bacterial productivity in estuaries compared to ocean environments may favor the survival of this particular type of pelagiphage. Group F (the HTVC111P-type) is also relatively more abundant in the estuaries than oceans. In general, estuaries contain diverse and abundant pelagiphage, and group A, D, F in estuaries appear to be more abundant in offshore water. Whether unique genotypes of pelagiphage are present in the estuarine ecosystem will require further investigation on the micro-diversity of pelagiphage.

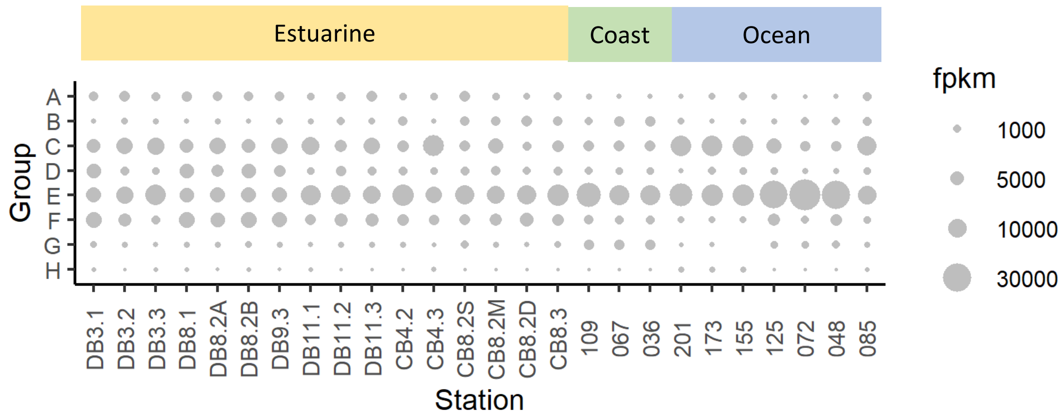

Figure 5.4 Bubble plot of the eight pelagiphage groups in estuarine and other environments. Size of the bubbles correspond to FPKM values.

The eight pelagiphage groups do not show a clear spatiotemporal pattern. The abundance of pelagiphage also does not have a strong correlation with temperature or salinity ($p > 0.1$, data not shown) in both bays. We also did not observe a clear seasonal and spatial pattern for SAR11 bacteria in the Chesapeake Bay. The SAR11 bacteria from the Chesapeake Bay 16S rDNA data, extracted from a previous publication, was visualized as a bubble plot (Fig. 5.5, Fig. 5.6) (H. Wang *et al.*, 2020). The relative abundance of the SAR11 bacteria in the Chesapeake Bay did not see a significant correlation with temperature and salinity ($P > 0.1$, data not shown). The lack of correlation between SAR11 bacteria and environmental factors concurs with previous studies on the Delaware Bay (Campbell and Kirchman, 2013). Compared to the Delaware Bay study (salinity ranging from 0 to 30) (Campbell and Kirchman, 2013), the Chesapeake Bay study had a narrower range of salinity (2.4 – 24.1) excluding freshwater and oceanic sites (H. Wang *et al.*, 2020), which may cause abundance patterns to be less evident.

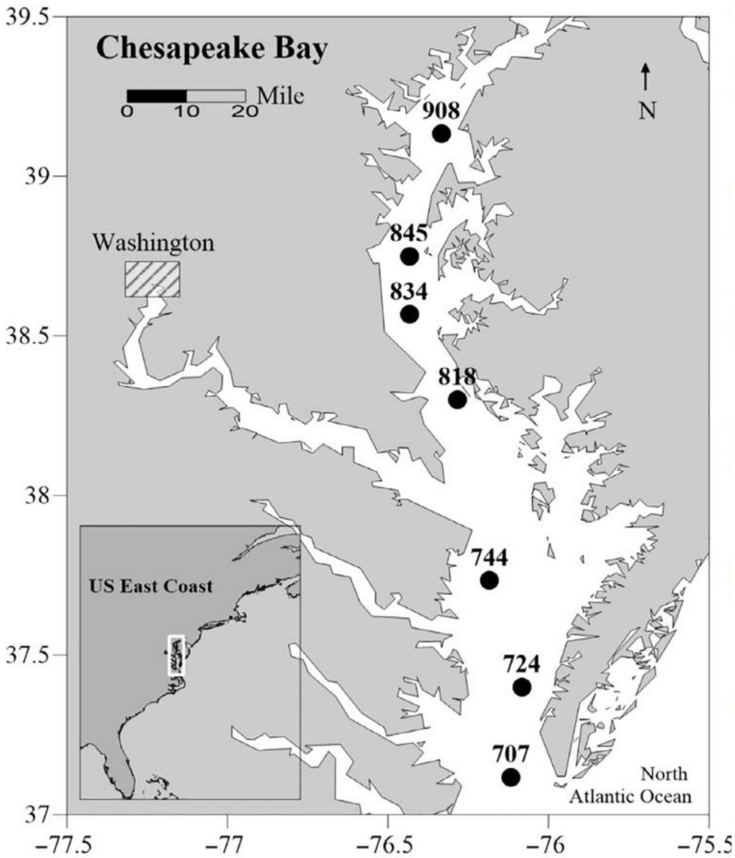

Figure 5.5 Sampling map of sites for bacteria in Chesapeake Bay (H. Wang *et al.*, 2020). Salinity ranges from 2.4 – 24.1. [Image Reprinted with permission from John Wiley and Sons 2020, license no. 5117400143825].

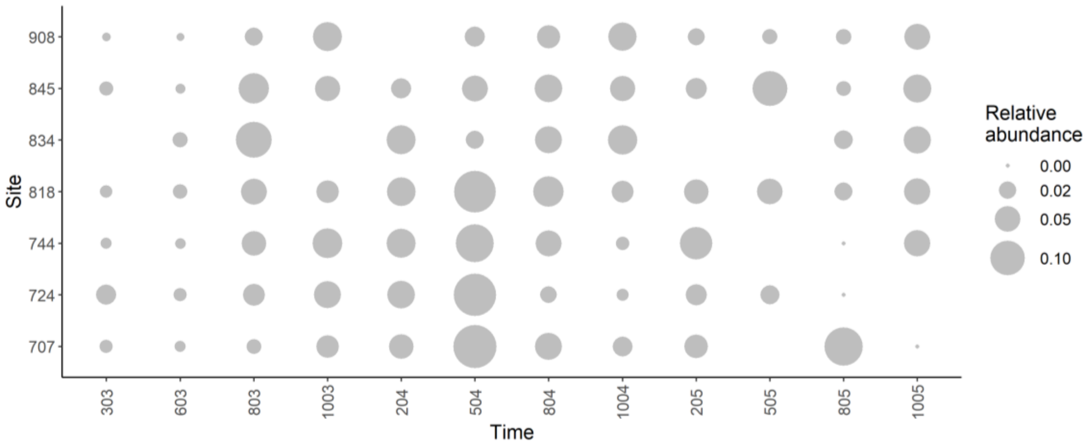

Figure 5.6 Relative abundance of SAR11 bacteria in the Chesapeake Bay based on 16S rDNA gene sequences (H. Wang *et al.*, 2020). Figure was generated from supplementary material of H. Wang *et al.* 2020. Y axis indicates sampling stations (908, 845, 834, 818, 744, 724 and 707) from the upper to lower Chesapeake Bay (see Fig. 5.1), X axis indicates time (month and year); for instance, 603 represents a sample from June 2003.

On the family level, the vast majority of pelagiphage in the estuaries are podoviruses. Podoviruses, myoviruses, and siphoviruses make up 96.2, 3.6 and 0.2% of estuarine pelagiphage community, and this composition is similar to that in the ocean (Fig 5.7). Meanwhile, pelagisiphoviruses (group H) are rare across both estuarine and oceanic environments. As of now, only two pelagisiphoviruses and one pelagimyovirus have been cultured and isolated (Table 5.1). More isolation of pelagimyoviruses and pelagisiphoviruses is needed to better evaluate their composition in the environment. Freshwater pelagiphage only make up a small proportion of total pelagiphage in the estuarine, coastal and oceanic water, ranging from 0.04% to 2.6% (Fig. 5.8) but these values may be underestimated since only eight of the 78 pelagiphage genomes are from freshwater (Table 5.1). The distribution of freshwater pelagiphage does not show a clear pattern along the estuarine salinity gradient. The proportion of freshwater pelagiphage is higher in the Chesapeake Bay than the Delaware Bay (Fig. 5.8). Freshwater pelagiphage in the oceanic sites are more abundant than those in the coastal sites (Fig. 5.8). Freshwater pelagiphage are generally found to be widespread in freshwater habitats, although occasionally found in brackish and oceanic environments (Chen *et al.*, 2019; Zaragoza-Solas *et al.*, 2020). Despite several of our upper estuarine samples being close to freshwater (salinity = 0.2 ppt), freshwater pelagiphage were not more abundant at these sites. Since freshwater pelagiphage are under sampled compared to marine pelagiphage, more studies on freshwater pelagiphage will allow us to get a more balanced view of their transition.

When categorized according to cultured isolates (44 species) vs. assembled environmental genomes (34 species), the oceanic sites appear to have a higher proportion of uncultured pelagiphages compared to the estuarine sites (Fig. 5.9). Most of the uncultured pelagiphage consist of vSAG 37-F6 (Fig. 5.9). This suggests that uncultured pelagiphages may be more representative of the diversity in oceanic environments, and reaffirms the importance of 37-F6 in the open ocean.

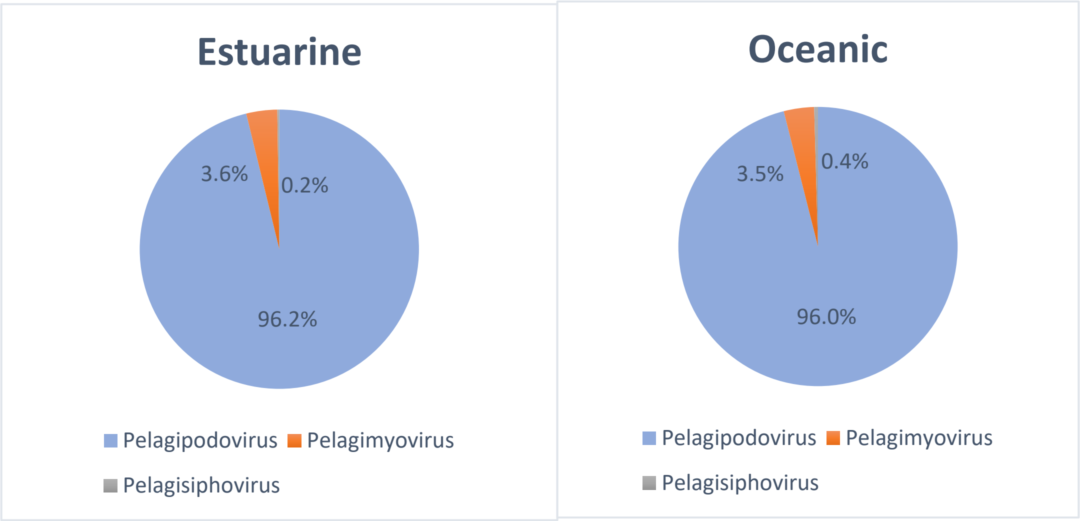

Figure 5.7 Compositions of three major phage families (*Myoviridae*, *Podoviridae* and *Siphoviridae*) of pelagiphage in estuarine (16 DEV samples) and oceanic (10 GOV 2.0 samples) environments. Data were derived from Figure 5.4.

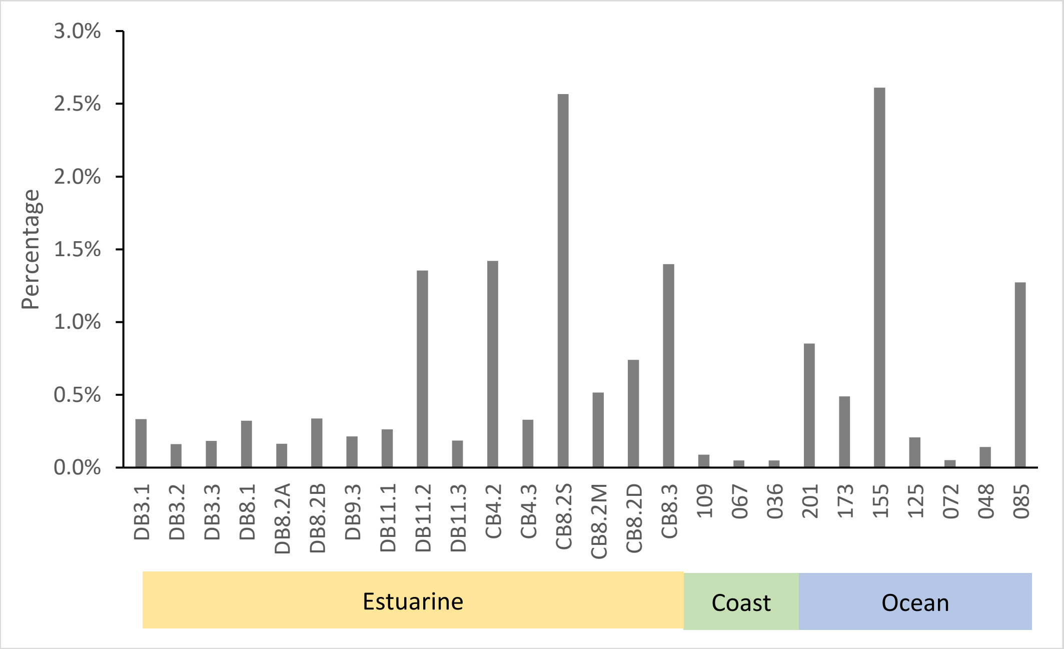

Figure 5.8 Percentage of freshwater pelagiphage as a portion of total pelagiphage (based on FPKM). Data were derived from Figure 5.4.

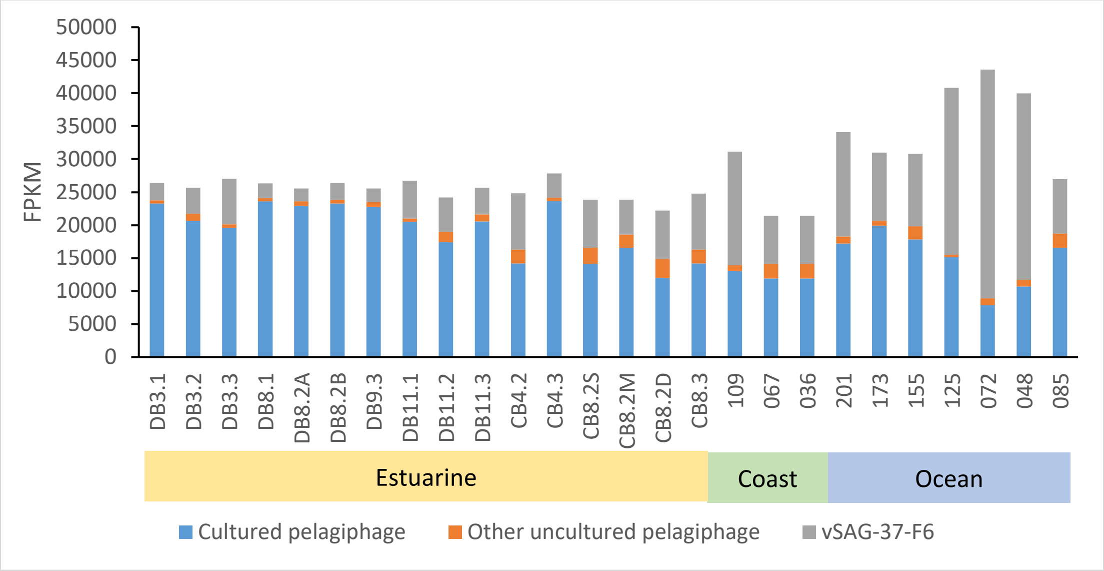

Figure 5.9 Relative abundance of cultured pelagiphages and uncultured pelagiphage MAGs. Data were derived from Figure 5.4.

5.4.3 Pelagiphage contigs

In order to get an understanding of the local pelagiphage contigs in the estuary, pelagiphage-like sequences were found within the de novo assembled contigs of DEV. In the DEV, 19 contigs were identified as pelagiphages, belonging to groups A, B and C (Table 5.3). Of the 19 contigs, seven belonged to myoviruses (group B), and 12 belonged to podoviruses (group A and C). Their relative abundance in each DEV sample ranged from 0 to 1035 FPKM (Fig. 5.10a). Most of the contigs identified as pelagiphage were relatively short (< 15 kb), except for the three contigs identified as uncultured pelagimyovirus (52 - 90 kb). Despite the large size of pelagimyovirus genomes (~150 kb), the contigs matching to the only cultured pelagimyovirus (HTVC008M) were short (5-8 kb) (Table 5.3). The rest of the pelagimyoviruses were cross-assembled from a wide range of viral metagenomic sequences, including the DEV (Zaragoza-Solas *et al.*, 2020). Thus, this is likely a result of the 3 pelagimyovirus contigs aligning back to the original DEV contigs from which they were assembled. No pelagisiphovirus contigs were detected in the DEV (Table 5.3). In another study, only six pelagisiphovirus contigs were identified from the entire GOV database, indicating that this newly discovered viral group may be rare in the environment, or difficult to identify on the contig level (Buchholz *et al.*, 2021). Also, no freshwater pelagiphage contigs were identified in DEV, despite the DEV containing low salinity samples that can be considered to be freshwater (salinity = 0.2 ppt).

Table 5.3 The 19 pelagiphage contigs identified in DEV.

Contig	Length	Closest match	Group	Family
Ga0070745_1001261	14173	HTVC119P	A	Podoviridae
Ga0070746_10002083	11891	HTVC109P	A	Podoviridae
Ga0070746_10007146	6449	HTVC025P	A	Podoviridae
Ga0070748_1001577	10613	HTVC031P	A	Podoviridae
Ga0070749_10002709	11861	HTVC019P	A	Podoviridae
Ga0070749_10003560	10410	HTVC025P	A	Podoviridae
Ga0070749_10007456	7149	HTVC025P	A	Podoviridae
Ga0070753_1001672	12308	HTVC031P	A	Podoviridae
Ga0099847_1001272	8729	HTVC201P	A	Podoviridae
Ga0070749_10001642	15165	Jormungand_EXVC012P	A	Podoviridae
Ga0070746_10009847	5445	HTVC008M	B	Myoviridae
Ga0070746_10004768	7955	HTVC008M	B	Myoviridae
Ga0070746_10006292	6917	HTVC008M	B	Myoviridae
Ga0070746_10006702	6664	HTVC008M	B	Myoviridae
Ga0070751_1000116	64504	PMP-MAVG-18	B	Myoviridae
Ga0070754_10000264	52626	PMP-MAVG-18	B	Myoviridae
Ga0070750_10000006	90121	PMP-MAVG-22	B	Myoviridae
Ga0070749_10006924	7414	Greip_EXVC021P	C	Podoviridae
Ga0070747_1005852	5525	HTVC204P	C	Podoviridae

Pelagiphage contig Ga0070747_1005852 was overwhelmingly abundant in CB4.3 (up to 1000 FPKM) (Fig. 5.10a), skewing the visualization of the other pelagiphage contigs (range 0-100 FPKM), so another plot omitting Ga0070747_1005852 is provided (Fig. 5.10b). Contig Ga0070747_1005852 is most closely matched to HTVC204P (Table 5.3). High abundance of HTVC204P was also observed in CB4.3 when recruiting the reads to only known pelagiphage genomes (FPKM=8426). This indicates the consensus of the two different methods (one method is mapping the reads against only known pelagiphage genomes, the other method is mapping the reads against all DEV viral populations, and selecting the populations that are pelagiphage). It is not clear why this viral species is so abundant in sample CB4.3 in particular.

In contrast to the known pelagiphage read recruitment results where group A were found to be rare, contigs belonging to group A are more abundant among the local DEV virome sequences (Fig. 5.4, Fig. 5.10b). In group B, contigs matching uncultured pelagimyovirus (Ga0070751_1000116, Ga0070746_10006292, Ga0070750_10000006) are more abundant than contigs matching cultured pelagimyoviruses (Table 5.3, Fig. 5.10b). Uncultured pelagimyoviruses are also more abundant than HTVC008M when recruiting to known pelagiphage genomes (data not shown). Currently, HTVC008M is the only isolated pelagimyovirus. Assembled pelagimyoviruses may represent broader community of pelagimyoviruses compared to HTVC008M.

Overall, pelagiphage contigs are more abundant in the Delaware Bay than the Chesapeake Bay (Fig. 5.10). During March and November in the Delaware Bay, pelagiphage contigs are more abundant in the mid bay compared to the upper and lower bay (Fig. 5.10b). This is surprising given that pelagibacter are associated with marine environments, and no freshwater pelagiphage contigs were identified (Table 5.3). Samples DB3.2 and DB9.3 have the highest pelagiphage contig abundance in DEV (Fig. 5.10b).

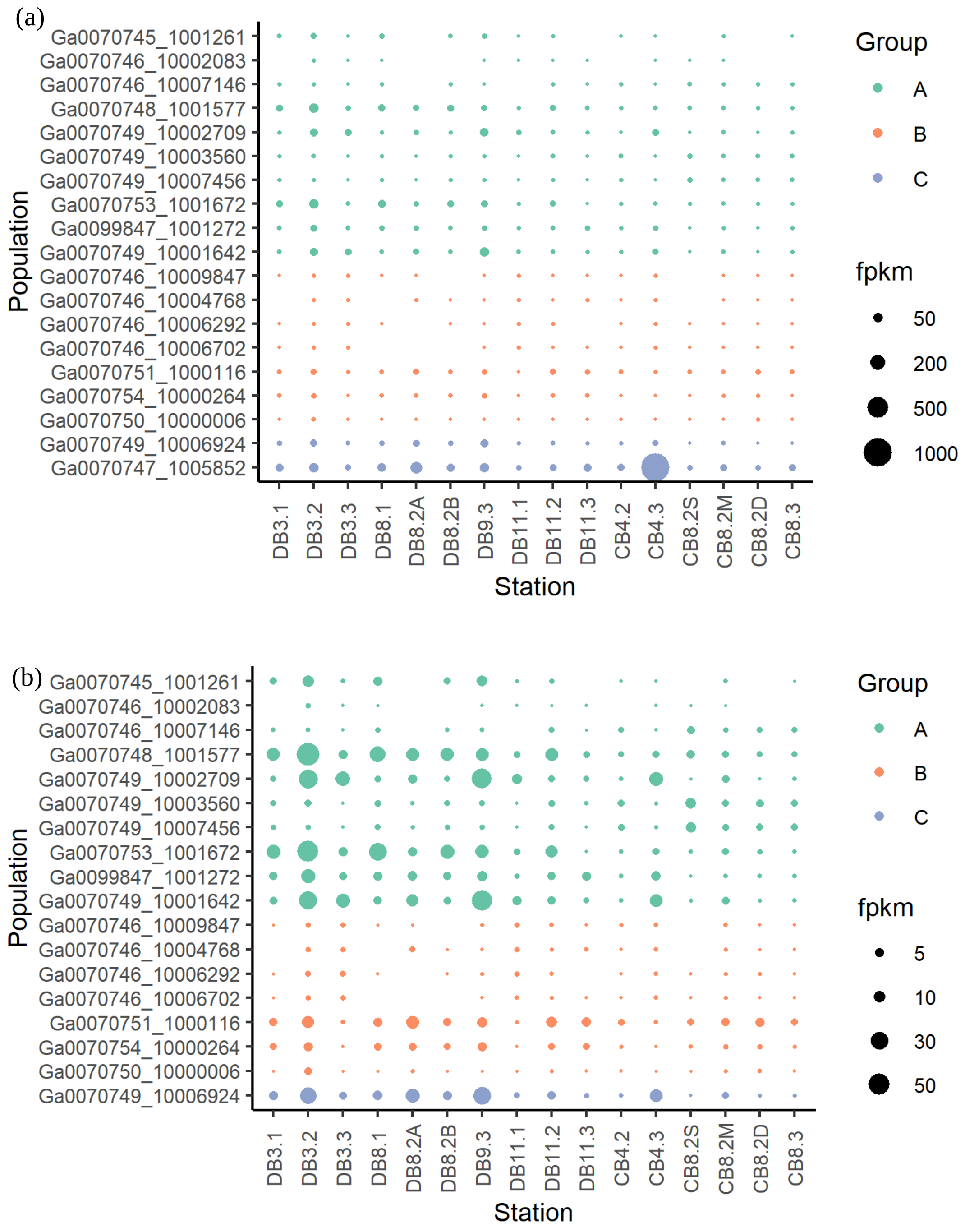

Figure 5.10 Relative abundance of 19 pelagiphage contigs in the Delaware Bay and Chesapeake Bay. (a) Full figure; (b) Figure without outlier Ga0070747_1005852.

Redundancy analysis showed that environmental and biological factors had little correlation with pelagiphage contig abundance, because the analysis only explained 10% of variance (Fig. 5.11). This concurs with the observation that SAR11 bacteria have little correlation with environmental factors in the estuary (Campbell and Kirchman, 2013) (Fig. 5.5, Fig. 5.6).

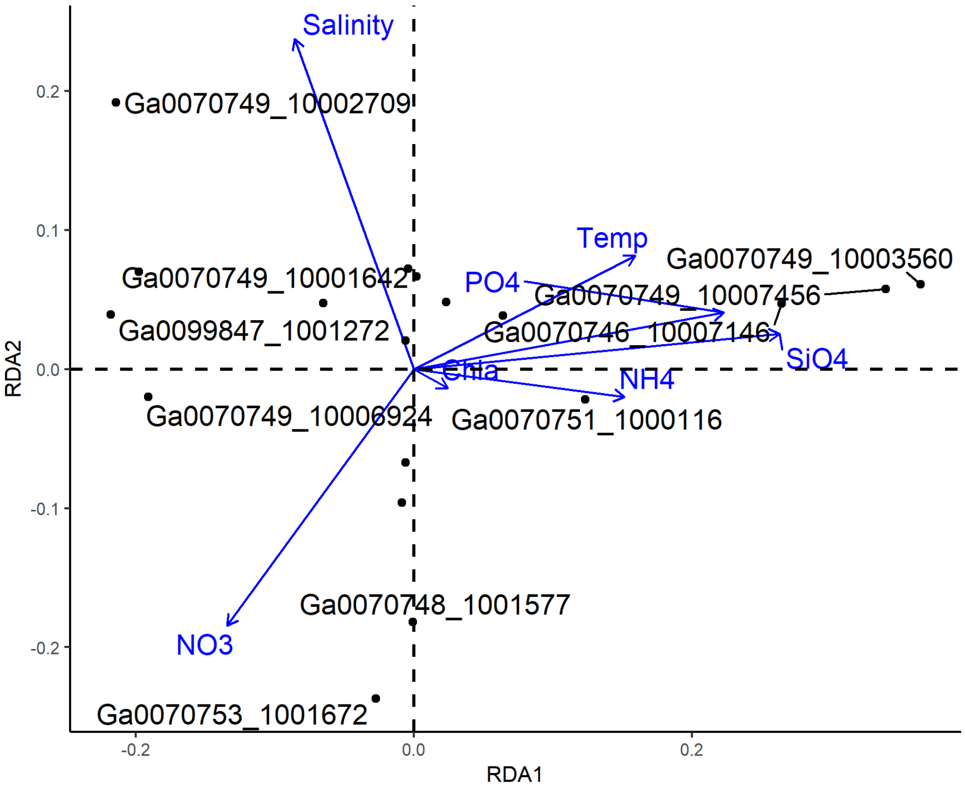

Figure 5.11 Redundancy analysis (RDA) ordination diagram (biplot) of pelagiphage contigs and environmental variables. RDA1 explains 6.2% of variance, while RDA2 explains 2.1% of variance. Total constrained variance is 10.4%, while unconstrained variance is 12.0%. Each black dot represents the variation of the relative abundance of said pelagiphage contig in the 16 samples. The angles between populations and environmental factors denote their degree of correlation.

5.5 Conclusions

We analyzed the phylogeny of pelagiphages using 78 known pelagiphage genomes, and classified them into eight groups according to their phylogeny and gene content. These known pelagiphages were used as references to evaluate the abundance of pelagiphage in the Delaware and Chesapeake estuaries, using two different methods, read recruitment directly to the 78 known pelagiphage genomes, and read recruitment to local pelagiphage contigs de novo assembled from the DEV data (19 contigs were identified). Both methods show that pelagiphage have little correlation to environmental factors including temperature and salinity in the estuary ($P > 0.1$). Despite the freshwater to seawater gradient present in the estuary, freshwater pelagiphage are rare in both bays, and do not show a clear transition along the salinity gradient. Pelagipodoviruses are overwhelmingly more abundant in the estuary compared to pelagimyoviruses and pelagisiphoviruses. The ocean contained a higher proportion of uncultured pelagiphages compared to the estuary. Uncultured virus vSAG-37-F6 was confirmed to be the most abundant virus in the estuary, as well the most abundant virus in the ocean. Local pelagiphage contigs were found to be more abundant in the Delaware Bay than the Chesapeake Bay. There is no significant differences seen between the pelagiphage communities of the estuary and the ocean. Overall, the pelagiphage community does not appear to change much along the estuarine environmental gradients.

Chapter 6: Conclusions and future directions

6.1 Major findings

6.1.1 Virioplankton community in the Chesapeake Bay and Delaware Bay does not show a clear spatial and temporal pattern.

Estuaries are a dynamic ecosystem due to the mixing of freshwater and seawater. Strong environmental gradients (i.e. salinity, nutrient, and light) are present in the estuary. As a temperate estuary, water temperature in the Chesapeake Bay and Delaware Bay varies greatly between winter and summer. It is expected that the abundance and community structure of microorganisms living in these estuaries should vary between seasons and along the environmental gradient. The composition of bacterioplankton community in the Delaware Bay and Chesapeake Bay exhibits a clear spatiotemporal change (Campbell and Kirchman, 2013; H. Wang *et al.*, 2020). Based on this, we hypothesized that the virioplankton community in the Chesapeake Bay and Delaware Bay is highly variable over spatial and temporal scales. Although some variation is seen, the community structure of virioplankton in both bays is more consistent across the estuary than we anticipated (Chapter 3). This conclusion applies to both known cultured and assembled virus populations. In Chapter 3, we used Kaiju to classify reads according to known Refseq virus taxonomy, and did not see any clear overall trends along seasonal, salinity or depth gradients, other than cyanophage being positively correlated with temperature. Targeted investigation into pelagiphages also found their abundance to be less variable than expected in both bays (Chapter 5).

While more variation was seen in the top 20 most abundant virus populations, the variation does not align with any known environmental or biological factors, and RDA analysis explained very little variance (Chapter 3). NMDS analysis of the top 5,000 most abundant populations also did not cluster samples according to temperature or salinity. It seems that in the estuarine environment virioplankton community is not as tightly connected to spatiotemporal variables compared to bacterioplankton community.

6.1.2 Estuarine virioplankton community is different from their counterparts in the open ocean.

We also predicted that the composition of the virioplankton community in the estuary is different from that in open oceans, due to different selection pressures. We observed various characteristics that distinguish the estuarine virus community from the oceanic virus community. Oceanic samples have a higher proportion of myoviruses, and also have more *Prochlorococcus* phage than estuarine samples (Chapter 3). When evaluating the N4-like virus community using read mapping to known N4-like virus genomes, their composition in the estuarine community was clearly distinct from that of the oceanic community (Chapter 4, Fig. 4.3, Fig. 4.5, Fig. 4.6). Admittedly, part of this may be due to the different sampling and processing methods used in the DEV vs. GOV cruises. Meanwhile, pelagiphages in the estuary are more closely related to oceanic viruses than freshwater viruses (Chapter 5). However, this may be in part due to the poorer characterization of freshwater viruses compared to marine viruses.

6.1.3 Viral communities in the Delaware Bay and Chesapeake Bay are different.

Although the Delaware Bay and Chesapeake Bay are in close proximity to each other, their viral communities are significantly different. ANOSIM of the top 5,000 most abundant viruses proved that the difference between these two bays is much larger than the difference caused by season or salinity (Chapter 3). Although the difference is distinct, the majority of the viruses that contribute to this distinction have not been characterized (Chapter 3). The N4-like virus population also showed marked differences between the two bays, with more abundant N4-like contigs in the Delaware Bay, and more *Vibrio* N4-like virus sequences in the Chesapeake Bay (Chapter 4). Also, more pelagiphage were seen in the Delaware Bay than the Chesapeake Bay (Chapter 5).

6.1.4 Uncultured viruses and assembled viruses contribute greatly to the understanding of viral diversity.

In recent years, it has been noted that viruses discovered using single-cell, single-virus, or single-molecule sequencing methods have a significant presence in the natural environment. Our study found that viruses discovered using these methods are abundant in the estuary as well (Chapter 3). Of the top 20 most abundant viral contigs in the DEV, four shared the closest similarity to a marine cyanobacterial cell obtained using single cell technology; one shared the closest similarity with a virus discovered using single-virus genomics; eight matched viral sequences derived from assembly-free single-molecule sequencing; and four matched uncultured viral populations from GOV (Chapter 3). Uncultured pelagiphage such as vSAG 37-F6 consist up to half of the pelagiphage community in the estuary and ocean samples (Chapter 5). These results indicate that yet uncultured viruses are also abundant in the estuarine environment.

6.1.5 N4-like viruses are rare, but relatively more abundant in colder water in the estuary

In chapter 4, we estimated the abundance of N4-like viruses in the estuary and ocean by mapping reads to known N4-like viruses, recruiting high amounts of reads at high identity and coverage. But then we revealed that this method generates large amounts of false positives. We also confirmed N4-like viruses to be rare in the environment. To further investigate the abundance and distribution of N4-like viruses, three alternative methods were employed: (a) Read mapping to local N4-like virus contigs in the DEV, (b) Reciprocal best hit BLAST to N4-like virus core genes, (c) Binning of reads to virus RefSeq using Kaiju. All three methods confirmed that N4-like viruses are rare in the estuarine environment. These three methods also show that N4-like viruses are more abundant where water temperature is low, which supports our laboratory's previous PCR study that N4-like viruses were only detected in winter in the estuary (Zhan *et al.*, 2015).

6.1.6 Pelagiphage community in the estuary does not change much along environmental gradients.

Since SAR11 bacteria (Pelagibacterales) have distinct freshwater and oceanic ecotypes, we hypothesized that the pelagiphage community should vary greatly along the salinity gradient in estuaries. We took 78 reference pelagiphage genomes and categorized them into eight groups according to genome similarity. We found that podoviruses make up the majority of the pelagiphage community in the estuary.

Surprisingly, the pelagiphage composition was fairly consistent in different seasons and locations, and freshwater pelagiphage were rare across the estuary, including the upper bay samples that are almost freshwater. The proportion of different pelagiphages in the estuarine communities was found to be similar to that in oceanic samples, suggesting that estuarine pelagiphages are as important as they are in the ocean.

6.2 Future directions

My thesis research has yielded several interesting findings on virioplankton in the estuarine environment. There are several related questions can be addressed in the future.

6.2.1 Virus-host relationships in the estuary

Viral community analysis in my thesis mostly focused on characterizing the Delaware Bay and Chesapeake Bay viromes, with some discussion about comparison to the bacterial community. However, the relationship between viruses and their hosts still remained largely unexplored. Viruses present in the < 0.8 µm fraction can be characterized and compared to those in the < 0.2 µm fraction. Preliminary data found around 1,800 virus contigs (Virsorter categories 1-3) in one < 0.8 µm fraction sample, and reads from a < 0.8 µm fraction metagenome reached up to 200 FPKM when mapping to DEV viral populations (data not shown). This shows that there are large amounts of viruses present in the bacterial fraction, and they are reasonably abundant. Since estuaries are highly dynamic environments, another direction is to evaluate the evolutionary selection pressure on virus and host community in different seasons and different parts of the estuary. This can be done using the ratio of the number of nonsynonymous to synonymous mutations at different sites (Coutinho *et al.*, 2019). This may help to explain why the estuarine microbial community has evolved into the way it is today. Also, the functional genes of both viruses and hosts can be characterized to understand their metabolic interactions in the estuary. The transcripts of these genes can be derived from the transcriptomic data, as mentioned in the section below.

6.2.2 Virus metatranscriptomics

There is transcriptomic data available for the Delaware Bay and Chesapeake Bay (Dr. Barbara Campbell, personal communication). The transcriptomic dataset can be used to study the expression of viral genes inside bacterial cells and connect viral activity with viromic data. I did not continue this work partly because it is difficult to separate transcripts of viral and host origin, since viruses harbor AMGs (auxiliary metabolic genes) that are closely related to host genes. This kind of complex metatranscriptomics also poses challenges to data normalization. Preliminary data showed that the top five transcriptionally active viruses (10-200 FPKM) in CB3.1 and DB3.1 are cyanophages, of which their *PsbA* gene shows most transcriptional activity (data not shown). I presented this idea of characterizing the viral transcriptomic origin of photosynthetic genes in my proposal in 2018, but another research group published the similar observation that over half of all prokaryotic *PsbA* expression originates from viruses (Sieradzki *et al.*, 2019). However, the expression of other AMGs may still be of interest. Once the transcriptome is well characterized, it can be combined with the functional genomics of both viruses and hosts, to evaluate their ecological interplay in estuaries.

6.2.3 Targeted assembly of viral MAGs in DEV

In recent years, uncultured virus genomes assembled from metagenomic sequences are playing an increasingly important role in viral discovery (Chapter 3). Out of the 27 pelagimyophage genomes available, 26 are derived from various metagenomic sequences via cross-assembly, demonstrating the power of manual assembly using extensive sources of uncultured sequences (Chapter 5). This is especially impressive given the long length of pelagimyophage genomes (110 – 180 kb). In this thesis, I did not further characterize individual viral contigs, mainly because their lengths tend to be far shorter than a full genome. We were not able to conduct comparative genomics or whole genome viral phylogeny (Chapters 3, 4, 5). This is partly due to the fact that all DEV viral contigs were assembled *de novo* within their own samples. Cross-assembly between different DEV samples would likely yield longer contigs, and would not lose information compared to existing methods, since all of the viral contigs are pooled together during clustering and dereplication anyway (Chapter 2). Also, if one were interested in a particular group of viruses (such as N4-like viruses or pelagiphage), known genomes of a specific virus could be used as reference while assembling, further improving the length and quality of the assembly. The improvement may be especially significant for rare viruses, since they are less likely to be fully covered in a single sequencing run. The work based on a putative *Acinetobacter* phage MAG (Appendix B) can also be expanded on. Recently, we found out that a global survey of high quality N4-like virus contigs has been done at the same time as our N4-like virus work (Yantao Liang, personal communication). This shows that MAG assembly is a useful approach to investigate the genomic diversity of certain virus groups.

6.2.4 Evaluation of methodology for abundance estimates

When estimating the relative abundance of N4-like viruses, a rare virus group, we saw an approximately 650-fold difference in FPKM when mapping to known N4-like viruses, compared to mapping to all local virus populations (Chapter 4). When using the same approach for pelagiphages (an abundant virus group), the difference was only 30-fold while the mapping parameters were the same (Chapter 5). Also, the false positives generated during the mapping to N4-like viruses completely obscured the spatiotemporal pattern seen using other methods, whereas the read mapping to pelagiphages still showed some of the patterns seen in other methods (Chapter 4 and 5). This showed that the bias of read recruitment methods is more prominent for rare viruses compared to abundant viruses (see 6.1.4).

This kind of bias has not been reported in other studies and can be challenging when one is trying to interpret the biogeographic distribution of viruses. The effect of different mapping parameters on different viruses etc. would require extensive benchmarking tests. Since read recruitment is widely used for virus abundance estimates, and our studies uncovered significant false positives using the method, the scientific community would benefit from a more accurate evaluation of abundance estimation methods.

6.3 Closing remarks

6.3.1 The Delmarva Estuarine Virome (DEV) project

The DEV project provided me a unique opportunity to explore viral diversity in the two Mid-Atlantic estuaries, the Delaware Bay and Chesapeake Bay. All the 16 estuarine viromes were collected from concentrated viral communities, deeply sequenced by JGI, and have often been utilized by other research groups (Bischoff *et al.*, 2019; Paez-Espino *et al.*, 2019). The ten viromes from the Delaware Bay and the six viromes from the Chesapeake Bay include samples from different seasons, locations and depths, and together they represent the most systematic survey done to date on estuarine viromes. Through my dissertation study, I assembled the raw sequences into viral contigs, analyzed viral diversity based on both the reads and the assembled contigs, and used known viral reference genomes to understand the distribution pattern of viral communities as well as targeted groups of viruses. While my research focus was on the two estuaries, I also compared the viral composition between estuaries, coastal waters and open oceans. In some cases, freshwater viruses (i.e. freshwater pelagiphages) were also included in my analysis.

6.3.2 Methodology for viral community ecology

Extensive community bioinformatic tools, such as IMG/VR, Kaiju, MG-RAST, Vegan in R etc., have been widely used in different chapters of my thesis. There are pros and cons to different community analysis methods.

Identifying viral sequences on the read level is faster, and preserves the relative abundance information better compared to the contig-based method, since there is a relatively equal chance for each read to be generated, while the formation of contigs is more complicated depending on assembly methods. However, since Illumina reads are short, and viruses have many genes shared with other viruses and prokaryotes, many of the reads can yield ambiguous results during binning and mapping. This can be remedied by longer read sequencing such as PacBio, which has less sequencing depth and lower accuracy. A hybrid approach where long read sequencing is used to assemble longer contigs/genomes while short read sequencing is used for transcriptome or abundance estimates is ideal.

Contig-based analyses of viral community is more accurate and gives us a better idea of genomic diversity. Many recent studies have been focused on the network of viral contigs/populations in different samples, revealing patterns about viral community succession, but most of this data is “anonymous”, since most of the populations are not classified as any known virus taxa (Brum *et al.*, 2015; Arkhipova *et al.*, 2017; Aylward *et al.*, 2017). While identifying whether a contig is viral or not is relatively easy (there are numerous tools for this purpose), identifying the actual taxonomy based on contigs is still difficult. This is partly due to the mosaicism of virus genomes, and the fact that not enough viruses have been isolated and characterized in general. While non-reference based methods are unbiased and take advantage of the full sequence data, reference-based methods help to connect viromic data to known virus taxonomy, taking advantage of all the prior knowledge we know about those taxa.

In this thesis, I also explored different methods of relative abundance estimation of viral species, which I discussed extensively in section 4.4.7. A comparison of the effects of read mapping methods between Chapter 4 and Chapter 5 is given in 6.2.4. Briefly, different methods should be used when evaluating rare viruses vs. abundant viruses, and special caution should be used when the virus group in question is rare. It should be noted that the designation of “N4-like viruses” is based on shared morphological and genomic characteristics of the viruses themselves, which infect various hosts; while the group “pelagiphages” is defined by the host they mutually infect, and includes a diverse set of virus types. Thus, more ambiguity may be present when mapping to N4-like virus contigs due to higher sequence similarity. While this study is not sufficiently benchmarked so as to indicate which is the “best” method to estimate viral relative abundance, the current data still provides some insight into the choice of methods when handling different viral groups.

In summary, every method has limitations, and these different methods should be used in conjunction with each other for a comprehensive understanding of the diversity and ecological patterns of a viral community.

6.3.3 Advancement of viral ecology from 2015 to 2021

When I started my Ph.D. study in 2015, there were a limited number of laboratories around the world studying aquatic viral communities using metagenomics; only people who are particularly interested in viral community ecology typically invest in viral metagenomics. Much has changed between now and then, and now in 2021, viral metagenomics is often considered as an integral part of microbial metagenomics studies. This is due to: (a) Increased awareness of the importance of viruses in the environment, (b) Lower cost of sequencing, (c) Improved accessibility of viral metagenomic tools and pipelines. Nowadays, it is possible to conduct a multi-faceted analysis of viral metagenomic data without knowledge of command line tools, thanks to web-based pipelines such as the iVirus system in Cyverse and KBase, the IMG/VR system, and VIROME, in addition to numerous smaller webtools such as PHASTER, HostPhinder, and ViPTree. I was also able to witness some important breakthroughs that helped explain many more of my unknown viral contigs than before: single virus genomics (Martinez-Hernandez *et al.*, 2017), assembly-free single-molecule sequencing virus genomes (Beaulaurier *et al.*, 2020), and GOV 2.0 (Gregory *et al.*, 2019). These three studies alone reduced the number of orphan contigs in the top 20 DEV viral populations from 14 to one. It is an exciting time to study viral ecology as the importance of viruses in the environment becomes more recognized. I anticipate more breakthroughs in the field that could enable a deeper understanding of environmental viral communities.

6.4 Significance of this work

This study represents the first comprehensive viromic analysis of virioplankton community in the two Mid-Atlantic estuarine ecosystems, the Chesapeake Bay and Delaware Bay. Strikingly, the viral contig-based analysis shows that the most abundant viral populations in both bays (top 20) are not those well-studied and abundant viruses, such as cyanophages and pelagiphages. We found that a large pool of abundant and diverse viruses in the estuaries are related to uncultured viruses in viral metagenomic databases, and some of them are related to uncultured viral sequences discovered via single cell genomics. Given the fact that estuaries are a dynamic ecosystem, the abundance and community structure of virioplankton in these two bays are relatively more stable than we expected. The number of infectious viruses may behave differently than the quantification of viruses based on viral like particles and viral DNA. We also showed that the virioplankton community in the Delaware Bay is significantly different from that in the Chesapeake Bay, indicating different niche partitioning of viruses in two neighboring estuaries. Estuaries are complex and diverse ecosystems, and their unique hydrological conditions may post a strong selection force for living organisms in the estuarine environment. This is the only study where the viral community from two adjacent estuaries are characterized, enabled by the close proximity of two large estuaries on the U.S. east coast. Our work also revealed that biases are present when the relative abundance of viruses is evaluated using different recruitment approaches, and the bias is more severe when searching for rare viruses. The rareness of N4-like viruses and the prevalence of SAR11 viruses in both estuaries set the examples of using viromic data to explore the biogeographic pattern of specific viral groups. Although viromes from different marine environments have been sequenced, there are very limited viromes from estuaries. The DEV is a highly valuable database to the research community and will continue to generate new findings in the future.

Appendices

Appendix A: Supplementary tables

Table S1

Table S1. Environmental conditions of DEV samples. Detailed information can be found at (http://dmoserv3.bco-dmo.org/jg/serv/BCO-DMO/Coast_Bact_Growth/newACT_cruises_rs.html0%7Bdir=dmoserv3.whoi.edu/jg/dir/BCO-DMO/Coast_Bact_Growth/,info=dmoserv3.bco-dmo.org/jg/info/BCO-DMO/Coast_Bact_Growth/new_ACT_cruises%7D).

Sample	**Year**	**Date**	**Time**	**Latitude**	**Longitude**	**Depth (m)**	**Temperature (°C)**	**Salinity (ppt)**	**NO3 (mM/L)**	**NH4 (mM/L)**	**PO4 (mM/L)**	**SiO4 (mM/L)**	**Bact_prod (ng of C/L/h)**	**Chla (mg/L)**
DB3.1	2014	19-Mar	7:15	39.7998	-75.4263	1.82	4.4	0.2	123	51.74	0.26	57.41	20.7	4.9
DB3.2	2014	21-Mar	7:00	39.134	-75.2287	1.5	4	20	16.13	36.22	0.01	2.69	32.1	16.1
DB3.3	2014	22-Mar	7:00	38.7788	-75.009	1.35	4	30.4	0.46	6.4	0.03	3.29	7.7	3.5
DB8.1	2014	28-Aug	11:04	39.8385	-75.3542	1.27	25.3	0.2	176.4	7.22	0.43	15	61.58	10.27
DB8.2A	2014	30-Aug	23:01	39.1465	-75.2433	1.49	24.3	21.5	9.63	21.18	0.24	15.17	116.76	11.35
DB8.2B	2014	31-Aug	11:02	39.1572	-75.2425	1.52	24.5	22	5.38	7.06	0.18	10.68	0	10.78
DB9.3	2014	1-Sep	11:00	38.8473	-75.1072	1.78	24.3	28.8	0.39	6.6	0.15	5.64	64.67	8.87
DB11.1	2014	1-Nov	10:58	39.8493	-75.2823	1.2	15.1	0.3	191.235	4.755	0.47	32.785	16.24	6.72
DB11.2	2014	2-Nov	10:53	39.4387	-75.5387	1.53	13.8	15.4	38.36	4.33	0.29	18.995	5.54	2.69
DB11.3	2014	3-Nov	11:00	38.7667	-74.9167	1.8	13.5	30	5.95	7.33	0.14	6.34	11.28	3.54
CB4.2	2015	12-Apr	7:00	38.8407	-76.4182	1.6	8.5	9.1	47.075	3.4325	-0.0005	20.3115	44.8	26.8
CB4.3	2015	15-Apr	6:57	37.0998	-76.0917	1.67	10.8	25.4	2.883	0.41	-0.1145	0.3945	22.2	4.1
CB8.2S	2015	19-Aug	15:01	38.9762	-76.3697	1.29	27.3	10.4	1.4925	0.9135	0.0255	37.49	125.9	19.65
CB8.2M	2015	19-Aug	15:01	38.9762	-76.3697	13.26	26.3	15.5	0.6995	13.65	1.771	40.02	33.9	2.13
CB8.2D	2015	19-Aug	15:01	38.9762	-76.3697	22.46	26.3	18.1	0.6585	21.0705	2.9265	50.43	21.7	1.91
CB8.3	2015	22-Aug	12:05	37.051	-76.0773	1.81	26.6	26.7	6.5755	5.248	0.183	6.01	64.3	3.87

Table S2

Table S2 Predicted hosts using CRISPR. Total FPKM is the FPKM of all 16 samples added together.

Virus population	**Total FPKM**	**Host IMG/M TaxonID**	**Predicted host name**
Ga0070749_10000095	102.7876	2778261322	Pelagibaca sp. ARS5
Ga0099846_1000971	89.0876	2781126070	Ruminococcaceae sp. SRS475613_71
Ga0070750_10006719	86.4714	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070745_1002756	82.3957	2786546125	Rhodococcus sp. ABRD_24
Ga0099846_1000802	77.7044	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0070745_1000312	63.5887	2515154180	Salinispora arenicola CNQ748
Ga0070752_1000258	57.4237	2756170262	Pseudomonas sp. LE5F10
Ga0099847_1000522	42.0789	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0070751_1000114	38.1596	2537561881	Lachnospiraceae bacterium sp. ICM7
Ga0070753_1000107	36.9271	2837149368	Marinomonas rhizomae IVIA-Po-145
Ga0070748_1000090	36.2309	2832081228	Lacimicrobium alkaliphilum X13M-12
Ga0070749_10002521	33.5639	2744055030	Gammaproteobacteria OM60/NOR5 clade bacterium IMCC8485
Ga0070746_10000168	31.6358	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0070754_10003497	29.4975	2744054922	Eikenella sp. NML130454
Ga0070753_1000520	28.9983	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070746_10006842	28.6593	2617271238	Saprospiraceae bacterium KD52
Ga0070749_10000109	27.0321	2778261636	Pseudomonas sp. SP133
Ga0070754_10007498	26.4656	2786546125	Rhodococcus sp. ABRD_24
Ga0070751_1000882	26.4315	2617271238	Saprospiraceae bacterium KD52
Ga0070747_1000614	25.6246	2781126306	Holdemanella sp. ERS235507_4
Ga0070750_10000985	25.2945	2744054922	Eikenella sp. NML130454
Ga0070746_10000249	24.1683	2515154180	Salinispora arenicola CNQ748
Ga0070745_1001519	23.5273	2556793009	Cylindrospermopsis raciborskii N3
Ga0070748_1000858	23.3872	2832175522	Pelagicola sp. LXJ1103
Ga0099846_1004370	22.8096	2675903227	Pedobacter ruber DSM 24536
Ga0099850_1000089	22.1637	2515154180	Salinispora arenicola CNQ748
Ga0070751_1000928	22.1238	2821405612	Jeongeupia sp. S16_009
Ga0099848_1002021	21.5155	2744054922	Eikenella sp. NML130454
Ga0070751_1002469	21.4705	2786546125	Rhodococcus sp. ABRD_24
Ga0099849_1000075	21.4488	2515154180	Salinispora arenicola CNQ748
Ga0070750_10000729	21.2699	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0070751_1000120	19.4631	2806311029	Pseudomonas alcaliphila JAB1
Ga0099851_1001215	18.3688	2751185763	Leptospirillum rubarum

Ga0099849_1001954	17.837	2617271238	Saprospiraceae bacterium KD52
Ga0070750_10004948	17.2834	2781126070	Ruminococcaceae sp. SRS475613_71
Ga0099846_1000004	16.6784	2775506815	Clostridiales bacterium mt11
Ga0070749_10000766	16.2661	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0099850_1000019	15.4762	2521172529	Rubellimicrobium thermophilum DSM 16684
Ga0070747_1001172	14.8323	2515154180	Salinispora arenicola CNQ748
Ga0099848_1000812	14.4858	2786546125	Rhodococcus sp. ABRD_24
Ga0099848_1005836	13.6391	2832220722	Alicycliphilus denitrificans CD02
Ga0070748_1004537	13.2605	2654587666	Chromobacterium sp. LK1
Ga0099848_1000335	12.0927	2627853571	Guam_bin1_Bacteroidetes
Ga0070748_1004703	12.0216	2747843217	Nitrosomonas sp. HKU-PRO10
Ga0070748_1000335	11.8735	2556793009	Cylindrospermopsis raciborskii N3
Ga0070745_1000472	11.6667	2617271238	Saprospiraceae bacterium KD52
Ga0070749_10006763	11.5694	2754412676	Verrucomicrobia bacterium JGI_MCM16ME040 (unscreened)
Ga0070749_10000246	11.0669	2832220722	Alicycliphilus denitrificans CD02
Ga0099850_1000002	10.8018	2617271238	Saprospiraceae bacterium KD52
Ga0099849_1000001	9.2533	2576861799	Peptococcaceae bacterium SCADC1_2_3 (unscreened)
Ga0070749_10004215	8.3767	2821405612	Jeongeupia sp. S16_009
Ga0070752_1001617	7.7976	2585427587	Rhodanobacter sp. FW510-R12
Ga0070754_10004564	7.5533	2556793009	Cylindrospermopsis raciborskii N3
Ga0099849_1000056	6.8429	2597489935	Thalassolituus oleivorans MIL-1
Ga0070749_10001839	6.4961	2834855223	Dickeya sp. FVG10-MFV-A16
Ga0070752_1006308	6.4829	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0099849_1001148	6.3972	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0099849_1000404	6.3553	2721755241	Spirochaetes bacterium GWE1_60_18
Ga0070749_10003065	6.3144	2826279758	Burkholderia gladioli MSMB1756
Ga0070746_10002391	6.2315	2590828656	alpha proteobacterium RS24
Ga0070752_1000339	5.9578	2556793009	Cylindrospermopsis raciborskii N3
Ga0070750_10007788	5.8235	2832646971	Tyzzerella nexilis AM23-10LB
Ga0070747_1000196	5.7326	2597489935	Thalassolituus oleivorans MIL-1
Ga0070747_1000043	5.6828	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070748_1001957	5.5894	2597489935	Thalassolituus oleivorans MIL-1
Ga0099847_1000066	5.4066	2830822581	Shewanella sp. BF02_Schw v.2
Ga0070747_1001413	5.331	2841893499	Photorhabdus luminescens BA1
Ga0070749_10000379	5.2105	2551306195	Salmonella enterica enterica sv. Mississippi 2010K-1406
Ga0070750_10000035	5.166	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070750_10000005	5.1345	2627854041	Pectobacterium carotovorum brasiliense BC D6

Ga0070754_10000405	4.9407	2811995170	Pseudomonas oleovorans oleovorans NBRC 13583
Ga0070748_1000172	4.9362	2556793009	Cylindrospermopsis raciborskii N3
Ga0070745_1000883	4.5121	2778261584	Acinetobacter sp. WCHA39
Ga0070747_1003710	4.5089	2814123109	Cedecea neteri FDAARGOS_392
Ga0070749_10002920	4.4083	2786546812	Rhodobacterales bacterium CG_4_10_14_0_8_um_filter_70_9
Ga0070745_1000697	4.357	2827407764	Alteromonas sp. 154
Ga0070748_1000215	4.1714	2734482271	Delftia sp. bin1_M6
Ga0070752_1000336	4.1513	2667527223	Pseudomonas guangdongensis CCTCC AB 2012022
Ga0070754_10000744	4.0974	2744054922	Eikenella sp. NML130454
Ga0070749_10000790	4.0623	2751185644	Clostridiales bacterium Firm_12
Ga0070749_10000459	4.0531	2556793009	Cylindrospermopsis raciborskii N3
Ga0070749_10000445	3.9732	2821405612	Jeongeupia sp. S16_009
Ga0070749_10003768	3.9454	2828337626	Sphingobium wenxiniae DSM 21828
Ga0070749_10000240	3.9389	2834855223	Dickeya sp. FVG10-MFV-A16
Ga0070746_10000068	3.8197	2521172529	Rubellimicrobium thermophilum DSM 16684
Ga0099847_1002422	3.6202	2597489935	Thalassolituus oleivorans MIL-1
Ga0070749_10002063	3.441	637000238	Rhodopseudomonas palustris BisB5
Ga0070754_10001265	3.3232	2617271238	Saprospiraceae bacterium KD52
Ga0070754_10005102	3.2662	2731957832	Acinetobacter venetianus LUH13518
Ga0070750_10002344	3.1478	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070751_1002830	3.0575	2828730651	Xanthomonas arboricola F22
Ga0099847_1000883	2.9873	2597489935	Thalassolituus oleivorans MIL-1
Ga0070749_10001859	2.9068	2751185644	Clostridiales bacterium Firm_12
Ga0070746_10002181	2.6938	2788500144	Nioella nitratireducens SSW136
Ga0070749_10000267	2.6612	2698537061	Microgenomates bacterium JGI CrystG Apr3-4-D4 (unscreened)
Ga0070747_1000062	2.5695	2684622593	Pseudoalteromonas sp. 10-33
Ga0070749_10005760	2.0327	2585427721	Luteimonas huabeiensis HB2
Ga0070752_1000908	1.8959	2708743121	Pseudomonadales bacterium RIFCSPHIGHO2_01_FULL_64_12
Ga0070746_10000067	1.8303	2627854041	Pectobacterium carotovorum brasiliense BC D6
Ga0070754_10000733	1.6786	2708743121	Pseudomonadales bacterium RIFCSPHIGHO2_01_FULL_64_12
Ga0070747_1000068	0.3793	2832081228	Lacimicrobium alkaliphilum X13M-12
Ga0099848_1000083	0.0759	2515154180	Salinispora arenicola CNQ748

Table S3

Table S3. Species annotation of 11 N4-like contigs from DEV via comparison to GenBank.

1. Ga0070748_1000030

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
1140_9678_-	64.52	62	22		2764	2825	76	137	2.1E-13	74	Roseobacter sp. CCS2
13327_14122_+	70.59	51	15		72	122	51	101	4E-12	69	Ochrobactrum anthropi ATCC 49188
13327_14122_+	68.49	73	23		71	143	69	141	1E-23	108	Enterobacteria phage N4
15626_16562_-	65	40	14		79	118	80	119	1.3E-06	51	Bartonella bacilliformis KC583
16718_22520_+	67.4	181	59		1234	1414	453	633	1.8E-67	253	Enterobacteria phage N4
22536_26359_+	60.17	118	47		1045	1162	492	609	1.9E-39	160	Enterobacteria phage N4
22536_26359_+	74.29	35	9		472	506	245	279	5.1E-08	56	Enterobacteria phage N4
26423_27163_+	59.21	76	31		104	179	107	182	4.9E-19	92	Enterobacteria phage N4
40470_46063_-	57.81	64	27		288	351	363	426	4.7E-13	72	Sulfitobacter phage EE36phi1
46136_46738_-	62.26	53	20		79	131	125	177	7.9E-11	65	Sulfitobacter phage EE36phi1
46795_49240_-	68.85	183	57		427	609	13	195	2.1E-68	256	Enterobacteria phage N4
49709_60692_-	64.4	250	89		740	989	71	320	8.9E-97	351	Sulfitobacter phage EE36phi1
64002_67275_+	78.79	33	7		395	427	100	132	1.2E-09	61	Deftia phage phiW-14
64002_67275_+	65.79	38	13		254	291	229	266	2E-09	60	Enterobacteria phage N4
64002_67275_+	69.7	33	10		395	427	101	133	7.6E-07	52	Methanobrevibacter ruminantium M1

68042_69283_+	81.58	38	7		349	386	337	374	3.5E-11	66	Silicibacter phage DSS3phi2
9721_12115_+	76.39	72	17		199	270	98	169	5.9E-26	115	Sulfitobacter phage EE36phi1

2. Ga0070748_1000124

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
16913_17803_-	72.13	61	17		227	287	232	292	3.7E-19	93	Enterobacteria phage N4
17854_18561_-	74.63	67	17		121	187	93	159	4.4E-24	109	Enterobacteria phage N4
18656_19843_-	81.08	185	35		11	195	13	197	1.2E-83	307	Enterobacteria phage N4
19882_21511_-	60	40	16		271	310	110	149	3.9E-06	49	Enterobacter phage EcP1
21618_23819_-	70.51	234	69		317	550	316	549	3.4E-94	342	Enterobacteria phage N4
26462_30477_-	61.02	59	23		1094	1152	82	140	4.6E-11	66	Sinorhizobium meliloti BL225C
											Sinorhizobium meliloti AK83
26462_30477_-	65.77	111	38		1078	1188	61	171	2.1E-34	143	Brucella sp. NF 2653
26462_30477_-	62.39	109	41		1077	1185	60	168	9.2E-31	131	Methylobacterium nodulans ORS 2060
30668_33615_-	61.22	49	19		85	133	89	137	6.8E-10	62	Enterobacteria phage N4
33693_36141_-	71.91	413	116		289	701	9	421	8.9E-184	640	Enterobacteria phage N4
38032_38265_-	69.7	33	10		44	76	50	82	0.000019	47	Bacillus thuringiensis serovar thuringiensis str. T01001
											Bacillus thuringiensis Bt407
											Bacillus thuringiensis serovar berliner ATCC 10792
38032_38265_-	74.07	27	7		47	73	17	43	0.00029	43	Geobacter bemidjiensis Bem

3. Ga0070747_1000707

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
13830_16180_-	68.29	41	13		17	57	62	102	4E-10	63	Candidatus Puniceispirillum marinum IMCC1322
6580_13545_-	59.79	97	39		1385	1481	151	247	6.8E-27	118	Enterobacteria phage N4

4. Ga0070748_1001286

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
3287_4417_+	50	68	34		157	224	151	218	1.6E-13	74	Enterobacter phage EcP1
48_3189_+	60.84	143	56		446	588	114	256	7.8E-45	178	Enterobacter phage EcP1

5. Ga0070748_1000026

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
26498_29078_-	60.48	167	66		72	238	79	245	2.5E-54	210	Enterobacteria phage N4
29198_32330_-	68.33	240	76		700	939	374	613	5.9E-96	348	Enterobacteria phage N4
32380_37707_-	64.85	202	71		1427	1628	634	835	5E-74	275	Enterobacteria phage N4
37882_43036_-	61.29	124	48		1087	1210	125	248	6.5E-38	155	Rhodobacter sphaeroides KD131
37882_43036_-	61.29	124	48		1087	1210	125	248	2.5E-37	153	Rhodobacter sphaeroides ATCC 17029
411_5908_+	64.68	218	76	1	882	1099	354	570	4E-82	302	Sulfitobacter phage EE36phi1
411_5908_+	66.06	218	73	1	882	1099	354	570	4.7E-83	305	Silicibacter phage DSS3phi2
43057_44826_-	62.65	83	31		116	198	1	83	3E-25	113	Sphingopyxis alaskensis RB2256
43057_44826_-	64.42	104	37		116	219	1	104	3.7E-36	149	Dinoroseobacter shibae DFL 12
45379_45753_-	75.76	33	8		90	122	1006	1038	1.5E-07	54	Campylobacterales bacterium GD 1
47727_50564_-	67.19	64	21		419	482	205	268	8.4E-19	92	Enterobacteria phage N4
51235_51750_-	59.57	47	19		2	48	5	51	1.2E-07	55	Sulfitobacter phage EE36phi1
51235_51750_-	60	45	18		104	148	108	152	4.5E-09	59	Novosphingobium aromaticivorans DSM 12444
51235_51750_-	56.06	66	29		87	152	91	156	2.1E-15	80	Acinetobacter phage Acj61

5931_7340_+	65.75	219	74	1	159	377	88	305	5.1E-78	288	Enterobacteria phage N4
60683_65415_+	66.28	258	87		1205	1462	163	420	4.8E-104	375	Sulfitobacter phage EE36phi1
7410_8012_+	76	50	12		87	136	95	144	2.1E-15	80	Enterobacteria phage N4

6. Ga0070748_1000068

query sequence id	**percentage identity**	**alignment length**	**number of mismatches**	**number of gap openings**	**query start**	**query end**	**hit start**	**hit end**	**e-value**	**bit score**	**annotations**
33210_35353_-	65.71	105	36		185	289	88	192	9.2E-31	131	Enterobacteria phage N4
35369_43560_-	62.44	213	80		350	562	71	283	7.5E-75	278	Enterobacter phage EcP1
35369_43560_-	58.06	155	65		441	595	161	315	8.1E-49	191	Enterobacteria phage N4
47408_50117_+	64.1	39	14		487	525	1	39	2.9E-06	50	Enterobacteria phage N4
8722_11680_+	58.64	162	67		821	982	84	245	2.1E-51	200	Enterobacter phage EcP1
8722_11680_+	59.52	84	34		886	969	151	234	7.6E-24	108	Enterobacteria phage N4

7. Ga0070746_10000011

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
1_6132_-	79.64	221	45		1733	1953	576	796	4E-99	358	Silicibacter phage DSS3phi2
15992_18322_-	70.07	137	41		502	638	20	156	1.6E-53	207	Nitratifractor salsuginis DSM 16511
15992_18322_-	65.91	88	30		506	593	25	112	1E-29	128	Silicibacter sp. TrichCH4B
15992_18322_-	69.01	71	22		29	99	23	93	1.3E-23	108	Silicibacter phage DSS3phi2
18948_22067_-	67.31	52	17		678	729	329	380	6.2E-15	79	Enterobacteria phage N4
22077_25653_-	70.27	74	22		365	438	174	247	2.6E-24	110	Volvox carteri f. nagariensis
22077_25653_-	73.47	49	13		634	682	444	492	3.6E-13	73	Herpetosiphon aurantiacus ATCC 23779
22077_25653_-	69.23	52	16		999	1050	2	53	1.6E-13	74	Enterobacteria phage N4
22077_25653_-	61.02	59	23		1002	1060	7	65	4.1E-14	76	Starkeya novella DSM 506
22077_25653_-	68.75	128	40		486	613	178	305	5.3E-48	189	Acanthocystis turfacea Chlorella virus 1
31063_32307_+	60	60	24		340	399	25	84	1.6E-15	81	Methylobacterium radiotolerans JCM 2831
31063_32307_+	63.75	80	29		264	343	259	338	1.5E-24	111	Xanthobacter autotrophicus Py2
32890_35191_+	76.05	263	63		247	509	10	272	1.4E-120	430	Enterobacteria phage N4
40847_43613_+	63.04	46	17		303	348	199	244	5.9E-09	59	Acidovorax citrulli AAC00-1

43626_45851_+	65.67	201	69		437	637	446	646	4.9E-72	268	Enterobacteria phage N4
46232_47533_+	60.98	41	16		200	240	181	221	3.9E-06	49	Enterobacteria phage N4
47563_48744_+	70.31	229	68		18	246	22	250	3.8E-91	332	Enterobacteria phage N4
48814_50482_+	64.44	90	32		465	554	363	452	1.6E-28	124	Sulfitobacter phage EE36phi1
6136_13714_-	67.2	125	41		1105	1229	12	136	2E-43	173	Rickettsia bellii RML369-C
6136_13714_-	66.93	127	42		1102	1228	25	151	2.7E-45	180	Roseovarius nubinhibens ISM
6136_13714_-	64.23	137	49		1105	1241	28	164	3.5E-47	186	Roseomonas cervicalis ATCC 49957
6136_13714_-	68.55	124	39		1105	1228	28	151	1.3E-44	177	Granulibacter bethesdensis CGDNIH1
6136_13714_-	68.8	125	39		1107	1231	30	154	7.8E-45	178	Rhodobacter capsulatus SB 1003
68875_69603_-	72.13	61	17		180	240	185	245	2.8E-19	93	Enterobacter phage EcP1
68875_69603_-	72.88	59	16		182	240	189	247	8.4E-19	92	Enterobacteria phage N4
69646_73167_-	66.35	315	106		747	1061	289	603	1.3E-125	446	Enterobacteria phage N4

8. Ga0070754_10005210

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
1_2529_+	75.82	91	22		2	92	221	311	7.2E-35	145	Silicibacter phage DSS3phi2
1_2529_+	84.32	421	66		229	649	118	538	4.3E-211	730	Silicibacter phage DSS3phi2
2539_5500_+	81.25	240	45		5	244	5	244	1.6E-117	419	Sulfitobacter phage EE36phi1
5663_8785_-	67.08	161	53		879	1039	3623	3783	1.9E-56	217	Sulfitobacter phage EE36phi1

9. Ga0070748_1000096

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
10872_13015_+	60.87	115	45		400	514	23	137	2.5E-37	153	Wolbachia endosymbiont of Drosophila melanogaster
10872_13015_+	62.61	115	43		400	514	26	140	6.5E-38	155	Wolbachia endosymbiont strain TRS of Brugia malayi
17861_21213_+	67.57	37	12		225	261	35	71	0.000011	48	Delftia acidovorans SPH-1
25401_31471_+	82.05	156	28		632	787	59	214	8.4E-72	268	Jannaschia sp. CCS1
2577_6393_+	60	65	26		663	727	138	202	2.4E-16	83	Sulfitobacter phage EE36phi1
31481_37493_+	71.36	199	57		946	1144	3	201	5.3E-82	301	Silicibacter phage DSS3phi2
31481_37493_+	71.26	414	119		1400	1813	118	531	8.3E-174	606	Silicibacter phage DSS3phi2

37613_40552_+	65.7	207	71		265	471	9	215	2.6E-79	293	Sulfitobacter phage EE36phi1
40624_44402_-	60.92	87	34		1059	1145	3432	3518	5.1E-25	112	Silicibacter phage DSS3phi2
6885_8069_+	70.11	87	26		124	210	132	218	1.6E-32	137	Silicibacter phage DSS3phi2
8136_10859_+	83.54	79	13		190	268	99	177	5.4E-31	132	Sulfitobacter phage EE36phi1

10. Ga0070748_1000074

query sequence id	percentage identity	alignment length	number of mismatches	number of gap openings	query start	query end	hit start	hit end	e-value	bit score	annotations
10320_11561_+	74.42	43	11		144	186	169	211	3.1E-12	70	Silicibacter phage DSS3phi2
11599_12789_+	77.98	277	61		371	647	146	422	1.5E-128	456	Silicibacter phage DSS3phi2
11599_12789_+	72.28	303	84		2036	2338	347	649	3E-129	458	Silicibacter phage DSS3phi2
13625_14632_+	70.31	64	19		247	310	363	426	2.5E-18	90	Sulfitobacter phage EE36phi1
20025_31010_+	65.05	103	36		3513	3615	3636	3738	4.7E-32	136	Sulfitobacter phage EE36phi1
2453_5474_+	73.36	289	77		256	544	26	314	1.7E-127	453	Enterobacteria phage N4
31232_33728_-	66.51	212	71		269	480	9	220	1E-80	297	Sulfitobacter phage EE36phi1
33744_39956_-	69.83	179	54		779	957	639	817	1.6E-70	263	Enterobacteria phage N4
33744_39956_-	72.92	421	114		1457	1877	118	538	7.7E-183	636	Silicibacter phage DSS3phi2

40029_43860_-	81.25	160	30		72	231	55	214	2.2E-72	270	Rhodobacterales bacterium HTCC2150
40029_43860_-	80.12	161	32		71	231	54	214	5.6E-71	265	Roseobacter denitrificans OCh 114
44245_46235_-	68.75	48	15		288	335	38	85	1.1E-14	78	Agrobacterium vitis S4
44245_46235_-	73.68	76	20		276	351	35	110	9.2E-31	131	Mesorhizobium opportunistum WSM2075
44245_46235_-	89.33	75	8		277	351	26	100	6.4E-36	148	Sulfitobacter phage EE36phi1
5475_9873_+	73.93	303	79		1028	1330	347	649	6.1E-134	474	Sulfitobacter phage EE36phi1

11. Ga0070747_1000830

query sequence id	**percentage identity**	**alignment length**	**number of mismatches**	**number of gap openings**	**query start**	**query end**	**hit start**	**hit end**	**e-value**	**bit score**	**annotations**
1_7638_+	63.73	295	107		668	962	1909	2203	1.9E-107	386	Sulfitobacter phage EE36phi1
10832_15197_-	85.6	368	53		851	1218	118	485	3.4E-183	638	Silicibacter phage DSS3phi2
10832_15197_-	80.66	305	59		356	660	8	312	7.8E-149	523	Sulfitobacter phage EE36phi1
7941_10798_-	77.12	236	53	1	291	525	22	257	1.4E-105	380	Sulfitobacter phage EE36phi1

Appendix B: Putative Acinetobacter phage

Introduction

A viral contig similar to phages infecting *Acinetobacter baumannii* (Iraqibacter) was found to be highly abundant in the Delaware Bay, but was not in the Chesapeake Bay. The host is yet to be identified.

Nicknamed “Iraqibacter” due to origin in military hospitals in Iraq, *A. baumannii* is a multidrug-resistant pathogen that is a problem in hospitals around the world, although its natural habitat remains unknown (A. Evans *et al.*, 2012; Howard *et al.*, 2012; Hrenovic *et al.*, 2014). The clinical concern of antibiotic-resistant *A. baumanii* is driving phage isolation in hope of discovering potential viral strains for phage therapy, since antibiotic-resistant *A. baumannii* was found to be more susceptible to phage infection (Mumm *et al.*, 2013; Merabishvili *et al.*, 2014; Chen *et al.*, 2017). As of 2018, 42 Acinetobacter phages have been isolated, and over half of their encoded proteins are of unknown function (Turner *et al.*, 2018).

Methods

To examine chosen circular populations of interest, a complete viral genome was reverse complimented, annotated using RAST, and visualized using DNAplotter from Artemis (Carver *et al.*, 2009; Brettin *et al.*, 2015).

Results

A circular viral genome (accession number: Ga0070751_1000196) was found to be highly abundant in several Delaware Bay samples (the most abundant population in DB3.1 and DB8.2B), but was not present in Chesapeake Bay samples (Fig. 3.2) (Sun *et al.*, 2021). Annotation by RAST showed that this genome had a total of 52 ORFs, of which only 8 proteins are known (Brettin *et al.*, 2015) (Fig. B.1). BLASTN search against NCBI-nr database showed the closest hit to podovirus *Acinetobacter baumannii phage* vB_AbaP_Acibel007, with query cover of 47%, while the top 50 hits are various other Acinetobacter phages. Its host could not be predicted by the IMG/VR method (Paez-Espino *et al.*, 2019). A search against TOV and IMG/VR databases returned no results other than hits to its own sequence. The presence of a unique, novel and abundant viral population in the Delaware Bay remains to be further explored.

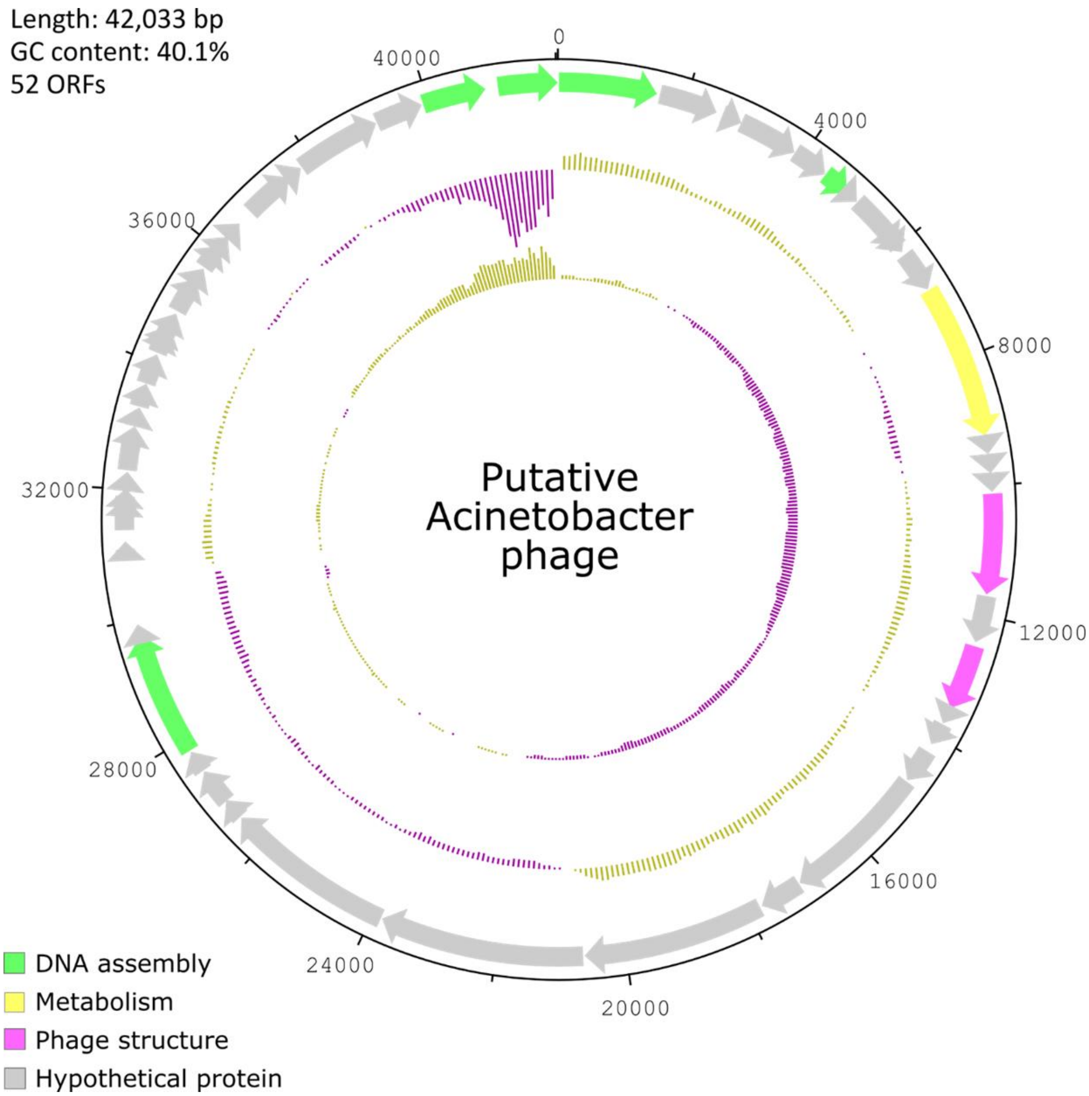

Figure B.1 Whole genome of putative *Acinetobacter* ("Iraqibacter") phage (accession number Ga0070751_1000196). Middle circle is GC content, inner circle is GC skew.

Discussion

A highly abundant viral population was found in the Delaware Bay, which had the closest match to *Acinetobacter baumannii* phages.

Since the information is derived from a MAG (metagenome-assembled genome), it is possible that the genome may be misassembled or inaccurately annotated due to it being a novel virus (Roux *et al.*, 2019). Nevertheless, the discovery of a putative *A. baumannii* phage and the fact that it appears to be exclusive to Delaware Bay suggests a contamination event of hospital origin in the Delaware Bay, likely stemming from the highly polluted Delaware river. Further work is needed to characterize and explore the distribution of this novel viral population.

Bibliography

A. Evans, B., Hamouda, A., and G.B. Amyes, S. (2012) The Rise of Carbapenem-Resistant *Acinetobacter baumannii*. *Curr Pharm Des* **19**: 223–238.

Ackermann, H.-W. and DuBow, M.S. (1987) Viruses of prokaryotes: General properties of bacteriophages, Boca Raton: CRC Press Inc.

Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017) Alignment-free d2∗ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* **45**: 39–53.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Amante, C. and Eakins, B.W. (2009) ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Tech Memo NESDIS NGDC-24*.

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

Aristizábal, M. and Chant, R. (2014) Mechanisms driving stratification in Delaware Bay estuary. *Ocean Dyn* **64**: 1615–1629.

Arkhipova, K., Skvortsov, T., Quinn, J.P., Mcgrath, J.W., Allen, C.C.R., Dutilh, B.E., et al. (2017) Temporal dynamics of uncultured viruses : a new dimension in viral diversity. **12**: 199–211.

Aylward, F.O., Boeuf, D., Mende, D.R., Wood-Charlson, E.M., Vislova, A., Eppley, J.M., et al. (2017) Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc Natl Acad Sci U S A* **114**: 11446–11451.

Azam, F., Fenchel, T., Field, J.G., Gray, J.., Meyer-Reil, L.A., and Thingstad, F. (1983) The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* **10**: 257–263.

Beaulaurier, J., Luo, E., Eppley, J.M., Uyl, P. Den, Dai, X., Burger, A., et al. (2020) Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res* **30**: 437–446.

Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol* **73**: 7629–7641.

Bergh, Ø., BØrsheim, K.Y., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467–468.

Berube, P.M., Biller, S.J., Hackl, T., Hogle, S.L., Satinsky, B.M., Becker, J.W., et al. (2018) Single cell genomes of *Prochlorococcus*, *Synechococcus*, and sympatric microbes from diverse marine environments. *Sci Data* **5**: 180154.

Bischoff, V., Bunk, B., Meier-Kolthoff, J.P., Spröer, C., Poehlein, A., Dogs, M., et al. (2019) Cobaviruses – a new globally distributed phage group infecting Rhodobacteraceae in marine ecosystems. *ISME J* **13**: 1404–1421.

Boesch, D.F., Brinsfield, R.B., and Magnien, R.E. (2001) Chesapeake Bay eutrophication: Scientific understanding, ecosystem restoration, and challenges for agriculture. *J Environ Qual* **30**: 303–320.

Bolduc, B., Jang, H. Bin, Doulcier, G., You, Z.Q., Roux, S., and Sullivan, M.B. (2017) vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **2017**:.

Born, Y., Fieseler, L., Marazzi, J., Lurz, R., Duffy, B., and Loessner, M.J. (2011) Novel virulent and broad-host-range *Erwinia amylovora* bacteriophages reveal a high degree of mosaicism and a relationship to Enterobacteriaceae Phages. *Appl Environ Microbiol* **77**: 5945–5954.

Borodovsky, M. and McIninch, J. (1993) GENMARK: Parallel gene recognition for both DNA strands. *Comput Chem* **17**: 123–133.

Breitbart, M. (2012) Marine viruses: Truth or dare. *Ann Rev Mar Sci* **4**: 425–448.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018) Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**: 754–766.

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

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250–14255.

Breitbart, M., Thompson, L., Suttle, C., and Sullivan, M. (2007) Exploring the vast diversity of marine viruses. *Oceanography* **20**: 135–139.

Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., et al. (2015) RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5**: 8365.

Brinkmeyer, R., Knittel, K., Jürgens, J., Weyland, H., Amann, R., and Helmke, E. (2003) Diversity and structure of bacterial communities in Arctic versus

Antarctic pack ice. *Appl Environ Microbiol* **69**: 6610–6619.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., et al. (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498–1.

Brum, J.R. and Sullivan, M.B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* **13**: 147–159.

Brussaard, C.P.D. (2004) Optimization of procedures for counting viruses by flow cytometry. *Appl Environ Microbiol* **70**: 1506–1513.

Buchholz, H.H., Michelsen, M.L., Bolaños, L.M., Browne, E., Allen, M.J., and Temperton, B. (2021) Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. *ISME J* **15**: 1585–1598.

Bushnell, B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. Berkeley, CA: Joint Genome Institute.

Bushnell, B. (2015) BBMap (version 35.14) [Software]. *Available at https://sourceforge.net/projects/bbmap/.*

Cai, L., Yang, Y., Jiao, N., and Zhang, R. (2015) Complete genome sequence of vB_DshP-R2C, a N4-like lytic roseophage. *Mar Genomics* **22**: 15–17.

Cai, L., Zhang, R., He, Y., Feng, X., and Jiao, N. (2016) Metagenomic analysis of virioplankton of the subtropical Jiulong River estuary, China. *Viruses* **8**: 35.

Campbell, B.J. and Kirchman, D.L. (2013) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME J* **7**: 210–20.

Campbell, B.J., Yu, L., Heidelberg, J.F., and Kirchman, D.L. (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci U S A* **108**: 12776–12781.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25**: 119–120.

Ceccarelli, D. and Colwell, R.R. (2014) Vibrio ecology, pathogenesis, and evolution. *Front Microbiol* **5**:.

Chan, J.Z.-M., Millard, A.D., Mann, N.H., and SchÃ¤fer, H. (2014) Comparative genomics defines the core genome of the growing N4-like phage genus and identifies N4-like Roseophage specific genes. *Front Microbiol* **5**:.

Chen, L.-X., Zhao, Y., McMahon, K.D., Mori, J.F., Jessen, G.L., Nelson, T.C., et al.

(2019) Wide distribution of phage that infect freshwater SAR11 bacteria. *mSystems* **4**: 1–16.

Chen, L.K., Kuo, S.C., Chang, K.C., Cheng, C.C., Yu, P.Y., Chang, C.H., et al. (2017) Clinical antibiotic-resistant Acinetobacter baumannii strains with higher susceptibility to environmental phages than antibiotic-sensitive strains. *Sci Rep* **7**: 1–10.

Choi, K.H., McPartland, J., Kaganman, I., Bowman, V.D., Rothman-Denes, L.B., and Rossmann, M.G. (2008) Insight into DNA and protein transport in double-stranded DNA viruses: The structure of bacteriophage N4. *J Mol Biol* **378**: 726–736.

Cissoko, M., Desnues, A., Bouvy, M., Sime-Ngando, T., Verling, E., and Bettarel, Y. (2008) Effects of freshwater and seawater mixing on virio- and bacterioplankton in a tropical estuary. *Freshw Biol* **53**: 1154–1162.

Colwell, R.R., Kaper, J., and Joseph, S.W. (1977) *Vibrio cholerae*, *Vibrio parahaemolyticus*, and other Vibrios: occurrence and distribution in Chesapeake Bay. *Science (80-)* **198**: 394–396.

Coutinho, F.H., Gregoracci, G.B., Walter, J.M., Thompson, C.C., and Thompson, F.L. (2018) Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends Microbiol* **26**: 955–965.

Coutinho, F.H., Rosselli, R., and Rodríguez-Valera, F. (2019) Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the mediterranean. *mSystems* **4**: 1–17.

Crits-Christoph, A., Gelsinger, D.R., Ma, B., Wierzchos, J., Ravel, J., Davila, A., et al. (2016) Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. *Environ Microbiol* **18**: 2064–2077.

Cubo, M.T., Alías-Villegas, C., Balsanelli, E., Mesa, D., de Souza, E., and Espuny, M.R. (2020) Diversity of Sinorhizobium (Ensifer) meliloti bacteriophages in the rhizosphere of Medicago marina: myoviruses, filamentous and N4-Like podovirus. *Front Microbiol* **11**: 22.

D’Elia, C.F., Sanders, J.G., and Boynton, W.R. (1986) Nutrient enrichment studies in a coastal plain estuary: Phytoplankton growth in large-scale, continuous cultures. *Can J Fish Aquat Sci* **43**: 397–406.

DeLong, E.F. (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459–469.

Dion, M.B., Oechslin, F., and Moineau, S. (2020) Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* **18**: 125–138.

Du, J. and Shen, J. (2016) Water residence time in Chesapeake Bay for 1980–2012. *J Mar Syst* **164**: 101–111.

Eggleston, E.M. and Hewson, I. (2016) Abundance of two Pelagibacter ubique bacteriophage genotypes along a latitudinal transect in the north and south Atlantic Oceans. *Front Microbiol* **7**: 1–9.

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., et al. (2020) Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**: 1639–1649.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237–40.

Fisher, T.R., Harding, L.W., Stanley, D.W., and Ward, L.G. (1988) Phytoplankton, nutrients and turbidity in the Chesapeake, Delaware and Hudson Estuaries. *Estuar Coast Shelf Sci* **27**: 61–93.

Fortunato, C.S. and Crump, B.C. (2011) Bacterioplankton community variation across river to ocean environmental gradients. *Microb Ecol* **62**: 374–382.

Fortunato, C.S., Herfort, L., Zuber, P., Baptista, A.M., and Crump, B.C. (2012) Spatial variability overwhelms seasonal patterns in bacterioplankton communities across a river to ocean gradient. *ISME J* **6**: 554–563.

Froelich, B., Bowen, J., Gonzalez, R., Snedeker, A., and Noble, R. (2013) Mechanistic and statistical models of total Vibrio abundance in the Neuse River Estuary. *Water Res* **47**: 5783–5793.

Fuhrman, J. a (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–8.

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

Garin-Fernandez, A., Pereira-Flores, E., Glöckner, F.O., and Wichels, A. (2018) The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages. *Mar Genomics* **41**: 31–41.

Ginestet, C. (2011) ggplot2: Elegant Graphics for Data Analysis. *J R Stat Soc Ser A* **174**: 245–246.

Giovannoni, S.J. (2017) SAR11 bacteria: The most abundant plankton in the oceans. *Ann Rev Mar Sci* **9**: 231–255.

Goodrich, D.M. and Blumberg, A.F. (1991) The fortnightly mean circulation of Chesapeake Bay. *Estuar Coast Shelf Sci* **32**: 451–462.

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al. (2019) Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**: 1109–1123.

Harding, L.W., Mallonee, M.E., Perry, E.S., Miller, W.D., Adolf, J.E., Gallegos, C.L., and Paerl, H.W. (2016) Variable climatic conditions dominate recent phytoplankton dynamics in Chesapeake Bay. *Sci Rep* **6**: 1–16.

Harding, L.W., Meeson, B.W., and Fisher, T.R. (1986) Phytoplankton production in two east coast estuaries: Photosynthesis-light functions and patterns of carbon assimilation in Chesapeake and Delaware Bays. *Estuar Coast Shelf Sci* **23**: 773–806.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., et al. (2020) Ecogenomics of the SAR11 clade. *Environ Microbiol* **22**: 1748–1763.

Henson, M.W., Lanclos, V.C., Faircloth, B.C., and Thrash, J.C. (2018) Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *ISME J* **12**: 1846–1860.

Herbert, R.A. (1999) Nitrogen cycling in coastal marine ecosystems. *FEMS Microbiol Rev* **23**: 563–590.

Herlemann, D.P.R., Woelk, J., Labrenz, M., and Jürgens, K. (2014) Diversity and abundance of "Pelagibacterales" (SAR11) in the Baltic Sea salinity gradient. *Syst Appl Microbiol* **37**: 601–604.

Hermes, A.L. and Sikes, E.L. (2016) Particulate organic matter higher concentrations, terrestrial sources and losses in bottom waters of the turbidity maximum, Delaware Estuary, U.S.A. *Estuar Coast Shelf Sci* **180**: 179–189.

Hewson, I., Eggleston, E.M., Doherty, M., Lee, D.Y., Owens, M., Shapleigh, J.P., et al. (2014) Metatranscriptomic analyses of plankton communities inhabiting surface and subpycnocline waters of the Chesapeake Bay during oxic-anoxic-oxic transitions. *Appl Environ Microbiol* **80**: 328–338.

Howard, A., O'Donoghue, M., Feeney, A., and Sleator, R.D. (2012) Acinetobacter baumannii: an emerging opportunistic pathogen. *Virulence* **3**: 243–250.

Hrenovic, J., Durn, G., Goic-Barisic, I., and Kovacic, A. (2014) Occurrence of an environmental Acinetobacter baumannii strain similar to a clinical isolate in paleosol from Croatia. *Appl Environ Microbiol* **80**: 2860–2866.

Huang, S., Zhang, S., Jiao, N., and Chen, F. (2015) Marine cyanophages demonstrate biogeographic patterns throughout the global ocean. *Appl Environ Microbiol* **81**: 441–452.

Huang, S., Zhang, Y., Chen, F., and Jiao, N. (2011) Complete genome sequence of a marine roseophage provides evidence into the evolution of gene transfer agents in alphaproteobacteria. *Virol J* **8**: 124.

Hugerth, L.W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., and Andersson, A.F. (2015) Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol* **16**: 1–18.

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.

Hurwitz, B.L. and U'Ren, J.M. (2016) Viral metabolic reprogramming in marine ecosystems. *Curr Opin Microbiol* **31**: 161–168.

Hutchinson, G.E. (1961) The Paradox of the Plankton. *Am Nat* **95**: 137–145.

Hwang, J., Park, S.Y., Park, M., Lee, S., Jo, Y., Cho, W.K., and Lee, T.K. (2016) Metagenomic characterization of viral communities in Goseong Bay, Korea. *Ocean Sci J* **51**: 599–612.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 1–11.

Ignacio-Espinoza, J.C., Ahlgren, N.A., and Fuhrman, J.A. (2020) Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat Microbiol* **5**: 265–271.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., et al. (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**: 632–639.

Ji, J., Zhang, R., and Jiao, N. (2014) Complete genome sequence of Roseophage vB_DshP-R1, which infects *Dinoroseobacter shibae* DFL12. *Stand Genomic Sci* **9**: 31.

Jinjun, K. and Jun, S. (2011) Bacterial community biodiversity in estuaries and its controlling factors: a case study in Chesapeake Bay. *Biodivers Sci* **19**: 770–778.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K.M., Kern, S., et al. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3**: 195–202.

Joint, I., Mühling, M., and Querellou, J. (2010) Culturing marine bacteria - an essential prerequisite for biodiscovery. *Microb Biotechnol* **3**: 564–575.

Kan, J., Crump, B.C., Wang, K., and Chen, F. (2006) Bacterioplankton community in

Chesapeake Bay: Predictable or random assemblages. *Limnol Oceanogr* **51**: 2157–2169.

Kan, J., Evans, S.E., Chen, F., and Suzuki, M.T. (2008) Novel estuarine bacterioplankton in rRNA operon libraries from the Chesapeake Bay. *Aquat Microb Ecol* **51**: 55–66.

Kan, J., Suzuki, M.T., Wang, K., Evans, S.E., and Chen, F. (2007) High temporal but low spatial heterogeneity of bacterioplankton in the Chesapeake Bay. *Appl Environ Microbiol* **73**: 6776–6789.

Kaneko, T. and Colwell, R.R. (1973) Ecology of *Vibrio parahaemolyticus* in Chesapeake Bay. *J Bacteriol* **113**: 24–32.

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. *Proc Natl Acad Sci* **110**: 12343–12348.

Kavagutti, V.S., Andrei, A.-Ş., Mehrshad, M., Salcher, M.M., and Ghai, R. (2019) Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome* **7**: 135.

Keegan, K.P., Glass, E.M., and Meyer, F. (2016) MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In *Methods in Molecular Biology*. New York: Humana Press, pp. 207–233.

Kirchman, D.L., Dittel, A.I., Malmstrom, R.R., and Cottrell, M.T. (2005) Biogeography of major bacterial groups in the Delaware Estuary. *Limnol Oceanogr* **50**: 1697–1706.

Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobian-Guëmes, A.G., et al. (2016) Lytic to temperate switching of viral communities. *Nature* **531**: 466–470.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* **35**: 1547–1549.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J* **9**: 2386–2399.

Lang, A.S. and Beatty, J.T. (2000) Genetic analysis of a bacterial genetic exchange element: The gene transfer agent of *Rhodobacter capsulatus*. *Proc Natl Acad Sci* **97**: 859–864.

Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., and Smith, D.B. (2017) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* **1977**: 1–10.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**: W256–W259.

Li, B., Zhang, S., Long, L., and Huang, S. (2016) Characterization and complete genome sequences of three N4-Like *Roseobacter* phages isolated from the South China Sea. *Curr Microbiol* **73**: 409–418.

Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., et al. (2016) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3–11.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Luo, E., Eppley, J.M., Romano, A.E., Mende, D.R., and DeLong, E.F. (2020) Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J* **14**: 1304–1315.

Luo, H. and Moran, M.A. (2014) Evolutionary ecology of the marine *Roseobacter* clade. *Microbiol Mol Biol Rev* **78**: 573–587.

Malmstrom, R.R., Kiene, R.P., Cottrell, M.T., and Kirchman, D.L. (2004) Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic Ocean. *Appl Environ Microbiol* **70**: 4129–4135.

Marshall, H.G., Burchardt, L., and Lacouture, R. (2005) A review of phytoplankton composition within Chesapeake Bay and its tidal estuaries. *J Plankton Res* **27**: 1083–1102.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., De La Cruz Peña, M.J., Martínez, J.M., et al. (2017) Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat Commun* **8**: 15892.

Martinez-Hernandez, F., Fornas, Ò., Lluesma Gomez, M., Garcia-Heredia, I., Maestre-Carballa, L., López-Pérez, M., et al. (2019) Single-cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. *ISME J* **13**: 232–236.

Martinez-Hernandez, F., Garcia-Heredia, I., Lluesma Gomez, M., Maestre-Carballa, L., Martínez Martínez, J., and Martinez-Garcia, M. (2019) Droplet Digital PCR for estimating absolute abundances of widespread *Pelagibacter* viruses. *Front Microbiol* **10**: 1–13.

Martinez-Hernandez, F., Luo, E., Tominaga, K., Ogata, H., Yoshida, T., DeLong, E.F., and Martinez-Garcia, M. (2020) Diel cycling of the cosmopolitan abundant Pelagibacter virus 37-F6: one of the most abundant viruses on earth. *Environ Microbiol Rep* **12**: 214–219.

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., and Paul, J.H. (2008) Metagenomic analysis of lysogeny in Tampa Bay: Implications for prophage gene expression. *PLoS One* **3**: e3263.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**: 60.

Meier-Kolthoff, J.P. and Göker, M. (2017) VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* **33**: 3396–3404.

Menzel, P., Ng, K.L., and Krogh, A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**: 11257.

Merabishvili, M., Vandenheuvel, D., Kropinski, A.M., Mast, J., De Vos, D., Verbeken, G., et al. (2014) Characterization of newly isolated lytic bacteriophages active against *Acinetobacter baumannii*. *PLoS One* **9**: e104853.

Moran, M.A., Belas, R., Schell, M.A., González, J.M., Sun, F., Sun, S., et al. (2007) Ecological genomics of marine roseobacters. *Appl Environ Microbiol* **73**: 4559–4569.

Moreno Switt, A.I., Orsi, R.H., den Bakker, H.C., Vongkamjan, K., Altier, C., and Wiedmann, M. (2013) Genomic characterization provides new insight into Salmonella phage diversity. *BMC Genomics* **14**: 481.

Morris, R.M., Cain, K.R., Hvorecny, K.L., and Kollman, J.M. (2020) Lysogenic host–virus interactions in SAR11 marine bacteria. *Nat Microbiol* **5**: 1011–1015.

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.

Mumm, I.P., Wood, T.L., Chamakura, K.R., and Kuty Everett, G.F. (2013) Complete genome of *Acinetobacter baumannii* podophage Petty. *Genome Announc* **1**: 6–7.

Nho, S.-W., Ha, M.-A., Kim, K.-S., Kim, T.-H., Jang, H.-B., Cha, I.-S., et al. (2012) Complete genome sequence of the bacteriophages ECBP1 and ECBP2 isolated from two different Escherichia coli strains. *J Virol* **86**: 12439–12440.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017) ViPTree: the viral proteomic tree server. *Bioinformatics* **33**: 2379–2380.

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M.P.G. (2018)

Overview of virus metagenomic classification methods and their biological applications. *Front Microbiol* **9**: 749.

Ogunseitan, O.A., Sayler, G.S., and Miller, R. V (1992) Application of DNA probes to analysis of bacteriophage distribution patterns in the environment. *Appl Environ Microbiol* **58**: 2046–52.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2018) vegan: Community ecology package. R package version 2.5-2. *CRAN R*.

Paez-Espino, D., Chen, I.M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., et al. (2017) IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res* **45**: D457–D465.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., et al. (2016) Uncovering Earth's virome. *Nature* **536**: 425–430.

Paez-Espino, D., Pavlopoulos, G.A., Ivanova, N.N., and Kyrpides, N.C. (2017) Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat Protoc* **12**: 1673–1682.

Paez-Espino, D., Roux, S., Chen, I.M.A., Palaniappan, K., Ratner, A., Chu, K., et al. (2019) IMG/VR v.2.0: An integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res* **47**: D678–D686.

Parmar, K., Dafale, N., Pal, R., Tikariha, H., and Purohit, H. (2018) An insight into phage diversity at environmental habitats using comparative metagenomics approach. *Curr Microbiol* **75**: 132–141.

Parmar, K.M., Gaikwad, S.L., Dhakephalkar, P.K., Kothari, R., and Singh, R.P. (2017) Intriguing interaction of bacteriophage-host association: An understanding in the era of omics. *Front Microbiol* **8**:.

Partensky, F. and Garczarek, L. (2010) *Prochlorococcus* : Advantages and limits of minimalism. *Ann Rev Mar Sci* **2**: 305–331.

Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J* **2**: 579–589.

Ponsero, A.J. and Hurwitz, B.L. (2019) The promises and pitfalls of machine learning for detecting viruses in aquatic metagenomes. *Front Microbiol* **10**: 1–6.

Pourtois, J., Tarnita, C.E., and Bonachela, J.A. (2020) Impact of lytic phages on phosphorus- vs. nitrogen-limited marine microbes. *Front Microbiol* **11**: 221.

Pritchard, D.W. (1967) What is an estuary: Physical Viewpoint. *Am Assoc Adv Sci* 3–5.

Proctor, L.M. and Fuhrman, J. a. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60–62.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**: 69.

Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., et al. (2020) Identifying viruses from metagenomic data using deep learning. *Quant Biol* **8**: 64–77.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009) Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7**: 828–836.

Rohwer, F. and Edwards, R. (2002) The phage proteomic tree: A genome-based taxonomy for phage. *J Bacteriol* **184**: 4529–4535.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., and Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus Evol* **2**: 1–11.

Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E. V., Kropinski, A.M., Krupovic, M., et al. (2019) Minimum information about an uncultivated virus genome (MIUVIG). *Nat Biotechnol* **37**: 29–37.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., et al. (2016) Ecogenomics and biogeochemical impacts of uncultivated globally abundant ocean viruses. *Nature* **537**: 689–693.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017) Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**: e3817.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**: 1–20.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**: 76.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: 0398–0431.

Sambrook, J. and Russell, D.W. (2006) Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harb Protoc* **2006**: pdb.prot4455.

Schito, G.C., Rialdi, G., and Pesce, A. (1966) Biophysical properties of N4 coliphage.

Biochim Biophys Acta - Nucleic Acids Protein Synth **129**: 482–490.

Schubel, J.R. and Pritchard, D.W. (1986) Responses of upper Chesapeake Bay to variations in discharge of the Susquehanna River. *Estuaries and Coasts* **9**: 236–249.

Scudlark, J.R. and Church, T.M. (1993) Atmospheric input of inorganic nitrogen to Delaware Bay. *Estuaries* **16**: 747–759.

Seth Augenstein (2012) Delaware River is 5th most polluted river in U.S., environmental group says. *NJ Adv Media*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Sharp, J.H. (1983) The Delaware estuary: research as background for estuarine management and development., Lewes, Delaware: University of Delaware College of Marine Studies.

Sharp, J.H., Cifuentes, L.A., Coffin, R.B., Pennock, J.R., and Wong, K.C. (1986) The influence of river variability on the circulation, chemistry, and microbiology of the Delaware Estuary. *Estuaries* **9**: 261–269.

Sharp, J.H., Yoshiyama, K., Parker, A.E., Schwartz, M.C., Curless, S.E., Beauregard, A.Y., et al. (2009) A biogeochemical view of estuarine eutrophication: Seasonal and spatial trends and correlations in the Delaware Estuary. *Estuaries and Coasts* **32**: 1023–1043.

Sieradzki, E.T., Ignacio-Espinoza, J.C., Needham, D.M., Fichot, E.B., and Fuhrman, J.A. (2019) Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**: 1169.

Silveira, C.B. and Rohwer, F.L. (2016) Piggyback-the-Winner in host-associated microbial communities. *npj Biofilms Microbiomes* **2**: 16010.

Spencer, R. (1955) A marine bacteriophage. *Nature* **175**: 690–691.

Stojković, E.A. and Rothman-Denes, L.B. (2007) Coliphage N4 N-acetylmuramidase defines a new family of murein hydrolases. *J Mol Biol* **366**: 406–419.

Sullivan, M.B. (2015) Viromes, not gene markers, for studying double-stranded DNA virus communities. *J Virol* **89**: 2459–2461.

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

Sullivan, M.B., Weitz, J.S., and Wilhelm, S. (2017) Viral ecology comes of age. *Environ Microbiol Rep* **9**: 33–35.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011) Easyfig: A genome comparison visualizer. *Bioinformatics* **27**: 1009–1010.

Sun, M., Zhan, Y., Marsan, D., Páez-Espino, D., Cai, L., and Chen, F. (2021) Uncultivated viral populations dominate estuarine viromes on the spatiotemporal scale. *mSystems* **6**.

Suttle, C. a (2007) Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Suttle, C.A. (2016) Environmental microbiology: Viral diversity on the global stage. *Nat Microbiol* **1**: 16205.

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

Tang, K.-H., Feng, X., Tang, Y.J., and Blankenship, R.E. (2009) Carbohydrate metabolism and carbon fixation in *Roseobacter denitrificans* OCh114. *PLoS One* **4**: e7233.

Tangherlini, M., Dell’Anno, a, Zeigler Allen, L., Riccioni, G., and Corinaldesi, C. (2016) Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Sci Rep* **6**: 28428.

Thingstad, T.F. (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.

Thompson, J.R., Randa, M.A., Marcelino, L.A., Tomita-Mitchell, A., Lim, E., and Polz, M.F. (2004) Diversity and dynamics of a North Atlantic coastal *Vibrio* community. *Appl Environ Microbiol* **70**: 4103–4110.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

Tout, J., Siboni, N., Messer, L.F., Garren, M., Stocker, R., Webster, N.S., et al. (2015) Increased seawater temperature increases the abundance and alters the structure of natural Vibrio populations associated with the coral *Pocillopora damicornis*. *Front Microbiol* **6**:.

Turner, D., Ackermann, H.W., Kropinski, A.M., Lavigne, R., Sutton, J.M., and Reynolds, D.M. (2018) Comparative analysis of 37 *Acinetobacter* bacteriophages. *Viruses* **10**: 1–25.

Van Valen, L. (1973) A new evolutionary law. *Evol Theory* **1**: 1–30.

Villarroel, J., Kleinheinz, K., Jurtz, V., Zschach, H., Lund, O., Nielsen, M., and Larsen, M. (2016) HostPhinder: A Phage Host Prediction Tool. *Viruses* **8**: 116.

Wagner-Döbler, I. and Biebl, H. (2006) Environmental biology of the marine Roseobacter lineage. *Annu Rev Microbiol* **60**: 255–280.

Wang, H., Zhang, C., Chen, F., and Kan, J. (2020) Spatial and temporal variations of bacterioplankton in the Chesapeake Bay: A re-examination with high-throughput sequencing analysis. *Limnol Oceanogr* **65**: 3032–3045.

Wang, K. (2007) Biology and Ecology of *Synechococcus* and their viruses in the Chesapeake Bay. Ph.D thesis, University of Maryland College Park.

Wang, K. and Chen, F. (2004) Genetic diversity and population dynamics of cyanophage communities in the Chesapeake Bay. *Aquat Microb Ecol* **34**: 105–116.

Wang, K. and Chen, F. (2008) Prevalence of highly host-specific cyanophages in the estuarine environment. *Environ Microbiol* **10**: 300–312.

Wang, K., Wommack, K.E., and Chen, F. (2011) Abundance and distribution of *Synechococcus* spp. and cyanophages in the Chesapeake Bay. *Appl Environ Microbiol* **77**: 7459–7468.

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J.C., Fuhrman, J.A., et al. (2020) A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics Bioinforma* **2**: 1–19.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127–181.

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

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**: e1456.

Wilson, W.H., Van Etten, J.L., and Allen, M.J. (2009) The Phycodnaviridae: The story of how tiny giants rule the world. *Curr Top Microbiol Immunol* **328**: 1–42.

Winget, D.M., Helton, R.R., Williamson, K.E., Bench, S.R., Williamson, S.J., and Wommack, K.E. (2011) Repeating patterns of virioplankton production within an estuarine ecosystem. *Proc Natl Acad Sci* **108**: 11506–11511.

Winget, D.M. and Wommack, K.E. (2008) Randomly amplified polymorphic DNA PCR as a tool for assessment of marine viral richness. *Appl Environ Microbiol* **74**: 2612–2618.

Wittmann, J., Klumpp, J., Moreno Switt, A.I., Yagubi, A., Ackermann, H.W., Wiedmann, M., et al. (2015) Taxonomic reassessment of N4-like viruses using comparative genomics and proteomics suggests a new subfamily - “Enquartavirinae.” *Arch Virol* **160**: 3053–3062.

Wittmann, J., Turner, D., Millard, A.D., Mahadevan, P., Kropinski, A.M., and Adriaenssens, E.M. (2020) From orphan phage to a proposed new family–the diversity of N4-like viruses. *Antibiotics* **9**: 1–12.

Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., et al. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**: 427–439.

Wommack, K.E. and Colwell, R.R. (2000) Virioplankton: Viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**: 69–114.

Wommack, K.E., Hill, R.T., Kessel, M., Russek-Cohen, E., and Colwell, R.R. (1992) Distribution of viruses in the Chesapeake Bay. *Appl Environ Microbiol* **58**: 2965–2970.

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. *Appl Environ Microbiol* **65**: 231–240.

Wright, R.T. and Coffin, R.B. (1983) Planktonic bacteria in estuaries and coastal waters of northern Massachusetts: spatial and temporal distribution. *Mar Ecol Prog Ser* **11**: 205–216.

Wu, S., Zhou, L., Zhou, Y., Wang, H., Xiao, J., Yan, S., et al. (2020) Diverse and unique viruses discovered in the surface water of the East China Sea. *BMC Genomics* **21**: 1–15.

Youle, M., Haynes, M., and Rohwer, F. (2012) Scratching the surface of biology’s dark matter. In *Viruses: Essential Agents of Life*. Dordrecht: Springer Netherlands, pp. 61–81.

Zaragoza-Solas, A., Rodriguez-Valera, F., and López-Pérez, M. (2020) Metagenome mining reveals hidden genomic diversity of pelagimyophages in aquatic environments. *mSystems* **5**: 1–16.

Zeigler Allen, L., McCrow, J.P., Ininbergs, K., Dupont, C.L., Badger, J.H., Hoffman, J.M., et al. (2017) The Baltic Sea Virome: diversity and transcriptional activity of DNA and RNA viruses. *mSystems* **2**: e00125-16.

Zhan, Y., Buchan, A., and Chen, F. (2015) Novel N4 bacteriophages prevail in the cold biosphere. *Appl Environ Microbiol* **81**: 5196–5202.

Zhan, Y. and Chen, F. (2019) Bacteriophages that infect marine roseobacters: genomics and ecology. *Environ Microbiol* **21**: 1885–1895.

Zhang, C., Du, X.P., Zeng, Y.H., Zhu, J.M., Zhang, S.J., Cai, Z.H., and Zhou, J. (2021) The communities and functional profiles of virioplankton along a salinity gradient in a subtropical estuary. *Sci Total Environ* **759**: 143499.

Zhang, Z., Chen, F., Chu, X., Zhang, H., Luo, H., Qin, F., et al. (2019) Diverse, abundant, and novel viruses infecting the marine Roseobacter RCA lineage. *mSystems* **4**: 1–17.

Zhang, Z., Qin, F., Chen, F., Chu, X., Luo, H., Zhang, R., et al. (2021) Culturing novel and abundant pelagiphages in the ocean. *Environ Microbiol* **23**: 1145–1161.

Zhao, G., Wu, G., Lim, E.S., Droit, L., Krishnamurthy, S., Barouch, D.H., et al. (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**: 21–30.

Zhao, Y., Qin, F., Zhang, R., Giovannoni, S.J., Zhang, Z., Sun, J., et al. (2019) Pelagiphages in the Podoviridae family integrate into host genomes. *Environ Microbiol* **21**: 1989–2001.

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.

Zhao, Y., Wang, K., Jiao, N., and Chen, F. (2009) Genome sequences of two novel phages infecting marine roseobacters. *Environ Microbiol* **11**: 2055–2064.

Zimmerman, A.E., Howard-Varona, C., Needham, D.M., John, S.G., Worden, A.Z., Sullivan, M.B., et al. (2020) Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol* **18**: 21–34.