DEFINING THE ROLE OF BALLAST WATER IN THE TRANSPORT OF VIRUSES IN AQUATIC ENVIRONMENTS THROUGH METAGENOMIC APPROACHES

By

Yiseul Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics – Doctor of Philosophy

2015

ABSTRACT

DEFINING THE ROLE OF BALLAST WATER IN THE TRANSPORT OF VIRUSES IN AQUATIC ENVIRONMENTS THROUGH METAGENOMIC APPROACHES

By

Yiseul Kim

Global shipping activities transport 12 billion tons of water across regions each year. This so called ballast water contains a variety of biological materials and has been considered to transfer non−native species between biomes. Despite the large amount of ballast water transported around the globe and its negative impact on native ecosystems, relatively little attention has been paid to viral invasions via ballast water due to technical challenges in detecting the wide range of viruses. The limitations of virus discovery using traditional approaches can now be overcome with the emergence of metagenomics, which enables unprecedented views of viral diversity and functions. This dissertation integrated environmental virology, metagenomics, and bioinformatics for the first time in order to examine composition and diversity of viruses in ballast and harbor waters collected from a freshwater system, and to investigate global transport of viruses through ballast water and effect of engineered, management, and environmental parameters on ocean viruses.

Viral communities in ballast water in the Great Lakes were examined due to the long history of non−native species invasions in this region of the world. Five ballast and three harbor waters were collected from the Port of Duluth−Superior. Bioinformatics analyses of over 550 million Illumina reads showed that the viral sequences had mostly no homologs in the public database. Among the sequences homologous to known viruses (22.3 ± 6.2%), ballast and harbor waters contained a diversity of viruses, which were

largely dominated by double–stranded (ds) DNA phages. Along with these phage families, viruses that could infect a broad range of hosts, some of which are highly pathogenic to fish and shrimp, were present at different levels in the viral metagenomes (viromes). Comparative virome analyses showed that viromes were distinct among the Great Lakes and formed a specific group of temperate freshwater viromes, separate from viromes associated with marine environments and engineered freshwater systems.

Sixteen ballast and eight harbor waters were further collected from the Port of Los Angeles/Long Beach and the Port of Singapore. Bioinformatics analyses of 3.8 billion Illumina reads revealed that taxonomic profile of the sequences homologous to known viruses (30.6 ± 0.03%) was similar to that observed in the Great Lakes viromes, which were largely dominated by dsDNA phages. Moreover, this research was able to detect sequences most similar to viruses infecting human, fish, and shrimp, which are related to significant public health problems or direct economic impact. Variations in virome composition were found between geographic locations, suggesting that the movement of ballast water across the global shipping network transports the ocean viromes. Importantly, this research showed that virus richness in ballast water was governed by conditions of local environment showing associations with latittude.

Outcomes of the present research represent the most detailed characterization to date of viruses in ballast water, defining the role of ballast water in the transport of freshwater and ocean viromes and an increased risk of exposure of aquatic fauna and flora to viruses. The present findings emphasize the need for implementing ballast water discharge limits for viruses and treatment. More research is needed on host population structure to better understand the impact of the transport of viruses between biomes.

This thesis is dedicated to my parents, Youngseok Kim and Youngae Kim and my husband Sangho Jeon.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Joan B. Rose for her guidance, encouragement, and motivation throughout my Ph.D. research. From her, I have learned how to think critically toward scientific questions and problems. Despite her busy travel schedule, she always found time for almost real–time online/offline discussions that my research could progress favorably. Thanks for introducing me to magnificent virus world and showing me a fantastic example of a scientist that I ever want to be. I would like to extend my thanks to all my research committee members, Dr. C. Titus Brown, Dr. Christopher M. Waters, Dr. Kristin N. Parents, and Dr. Volodymyr V. Tarabara for their insightful comments and guidance.

My special thanks go to the entire laboratory members, especially, Dr. Tiong Gim Aw and Ms. Rebecca L. Ives. During the past years working with Dr. Aw, I learned not only about virology techniques but also generosity, decisiveness, and positive attitude as a scientist. I was really fortunate to have someone like him that I could talk to whenever I ran into any kinds of problems. He always had an answer for me as my mentor in the laboratory and as my friend. I will remember our ballast water sampling adventure. Ms. Ives has been always supportive and dedicated in the laboratory that I could solely focus on my research for the past years. Becca is like a walking laboratory SOP that I could always follow without question.

I would like to thank all the ballast water inspectors from the Wisconsin Department of Natural Resources and the California State Lands Commission and personnel from an anonymous shipping company for organizing ballast water sampling

and helping with collection of ballast water. I wish to thank laboratory members of Dr. Randall E. Hicks at University of Minnesota Duluth and of Dr. Karina Gin Yew−Hoong at National University of Singapore and personnel from the Danish Hydraulic Institute, Singapore for assistance with ballast water sampling. My thank also goes to numerous ship crew members. This dissertation could not have been started without their support and interest toward my research.

I wish to thank my friends that I met in Michigan State University. Without their support and encouragement in every way, I doubt about completion of my dissertation. Adina Howe and Mike Howe provided unconditional support for the past years. They are the most lovely and generous couple I have ever seen. I am thankful to Qingpeng Zhang for entertaining me with the funniest stories and jokes. I am also grateful to him for fruitful discussion about bioinformatics analyses. My thanks go to Tamara Cole and Fan Yang for listening to my frustrations on my research and life. I will remember the wonderful moments I shared with all of you.

Most importantly, I owe my deepest gratitude to my parents and parents−in−low. Thanks for their unconditional love and support, and understanding for my decision to obtain my degree. I wish to thank Sangho Jeon, who chose to spend his life with me and has supported me with his tireless patience and tolerance toward my work and me for the past years. He is an incredible mentor in my life who has shown me generosity and patience. He is an excellent scientist, who always inspires me with creativity and enthusiasm for research. He has truly made this possible. Deep thanks also go to my other family members for their love and support.

It is gratefully acknowledged that the financial support of this research was provided by grants OISE−0530174 from the National Science Foundation, Partnerships for International Research and Education. I am also particularly grateful to Dr. Joan B. Rose and Dr. Volodymyr V. Tarabara for generous funding for my stipend, travel, and research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Bioinvasion through ballast water

Human activities such as agriculture, aquaculture, global transportation, and recreational activities have promoted spread of species across their natural geographic barriers (Cox, 2004). These alien, exotic, non–indigenous, and non–native species introduced into a new region outside of their historic range are generally referred as invasive species (Ricciardi and Cohen, 2007). Broadly, the steps in the invasion process consist of transport, introduction, establishment, and spread of invasive species as well as their ecological impact (Kolar and Lodge, 2001; Sakai et al., 2001). Among these, understanding the first transition (transport and introduction) of invasive species is critical because reducing the number of invasive species is the most practical step to prevent further invasions (Kolar and Lodge, 2001).

Ballast water is one of the most important vectors for transporting biological species within its region of origin to a new ecosystem (Mills et al., 1993; Drake and Lodge, 2007). Ships' ballast water has been used to increase the draft, change the trim, regulate the stability, or maintain stress loads within acceptable limits beginning in the 1870's (National Research Council, 1996; Gollasch et al., 2000). Ballast water is taken on board when a ship is traveling without cargo to compensate for the lack of weight. Once the ship has reached the next port and is loading new cargo on board, the ballast water, typically containing a variety of biological materials, including animals, bacteria, plants, and viruses, is discharged.

In the past several decades, research on aquatic invasive species and ballast water management has centered on metazoans with the exception of those on dinoflagellates (Drake et al., 2002; Litchman, 2010). Comparatively little attention has been paid to

invasions by organisms from microbial domains of life, such as archaea, bacteria, fungi, and viruses. Among the microbial invasions via ships' ballast water, most studies have focused on detecting the presence of pathogenic bacteria, such as *Escherichia coli* O157:H7, *Listeria monocytogenes*, *Pseudomonas aeruginosa*, or *Vibrio parahaemolyticus* (DePaola, 2003; Burkholder et al., 2007; Altug et al., 2012). Increasing attention has been directed to invasions by viruses since the emergence and impact of deadly viruses such as viral hemorrhagic septicemia virus (VHSV) (Elsayed et al., 2006; Lumsden et al., 2007; Bain et al., 2010) and koi herpesvirus (KHV) (Grimmett et al, 2006; Garver et al., 2010). Theses studies were, however, limited to measure species abundance or to detect and characterize the presence of known species. Surprisingly, composition and diversity of introduced species in ballast water remain largely unknown and potential ecological impacts and public health risks are not well understood. Consequently, native biodiversity and ecosystem functions are at risk to the impacts of invasive species, which are novel and previously uncharacterized.

It has been reported that over 10 billion tons of ballast water is moved worldwide annually (Engineering Center Transzvuk, 2012) and 50,000 species are introduced in the United States (U.S.), causing $137 billion in extensive ecological and economic damages to aquatic ecosystems each year (Ferrate Treatment Technologies, 2011).

1.2. Status of ballast water management guidelines and standards

In order to restrict ballast water–mediated invasions, the International Maritime Organization (IMO) established a mid–ocean ballast water exchange guideline (IMO, 2004). Under the guideline, a vessel should conduct ballast water exchange as far from

the nearest land as possible, but at least 200 nautical miles (1 nautical mile = 1.852 kilometers) and in water depths of at least 200 m. If this is impossible, ballast water exchange should be carried out at least 50 nautical miles from the nearest land and in water depths of at least 200 m.

In the U.S., ballast water management is addressed by the federal agencies (U.S. Coast Guard (U.S.C.G.) and U.S. Environmental Protection Agency (U.S. EPA)) and at the state level (David and Gollasch et al., 2015). In 2012, due to limited ecological protection afforded by ballast water exchange practice, U.S.C.G. issued a Final Rule (Phase 1), which requires ballast water discharges to meet the IMO ballast water management Convention D–2 standard (Department of Homeland Security, 2012). Recently, U.S.C.G. is moving forward with its examination of a Phase 2 standard, which is 1,000 times more stringent than the Phase 1 standard. Table 1.1 summarized numeric ballast water discharge standards implemented and proposed by U.S.C.G. and the State of California.

Table 1.1. Numeric ballast water discharge standards at the federal and state level

	Federal level (U.S.C.G.)		State level (California)	
Organism size class	Phase 1 standard	Phase 2 standard	Interim standard	Final standard
Organisms ≥ 50 μm	< 10 organisms / m^3	< 1 organism / 100 m^3	Zero detectable	Zero detectable
10 μm ≤ Organisms < 50 μm	< 10 organisms / mL	< 1 organism / 100 mL	< 0.01 organisms / 1 mL	Zero detectable
Organisms < 10 μm	Not addressed	Not addressed	< 10^4 viruses / 100 mL	Zero detectable
Escherichia coli	< 250 CFU / 100 mL	< 126 CFU / 100 mL	< 10^3 CFU / 100 mL	Not addressed
Intestinal enterococci	< 100 CFU < 100 mL	< 33 CFU < 100 mL	< 33 CFU < 100 mL	Not addressed
Vibrio cholera (O1 & O139)	< 1 CFU / 100 mL	< 1 CFU / 100 mL	< 1 CFU / 100 mL	Not addressed

Abbreviations: U.S.C.G., United States Coast Guard; CFU, colony–forming unit.

At the state level, several states have developed regulations for numeric ballast water discharge standards and California is considered to have the most stringent requirements (David and Gollasch et al., 2015). The implementation schedule for the California's final discharge standard of zero detectable organisms has been delayed due to the lack of available treatment technologies. It is currently undergoing a review process and is proposed to go into effect January 2020.

1.3. Knowledge gaps in the current understanding of viruses in ballast water

While the introduction of viruses through ballast water is one of the challenges facing the coastal environment, taxonomic composition and diversity of introduced viruses in ballast water remain largely unknown. Although previous findings showed a high level of virus–like particles (VLPs) in discharged ballast water (Ruiz et al., 2000; Drake et al., 2007), viral community structure, which is critical in assessing and preventing impact of viral invasion has not been studied. Lack of information on viral invasion has subsequently hindered the implementation of ballast water discharge limits for viruses, which are currently being considered only by the State of California. Key questions remain, such as what viruses are prevalent in ballast water and which viruses should be monitored for and by what method? These issues have emphasized the need for an in–depth investigation of viral communities in ballast water.

1.4. Dissertation outline

This dissertation is organized into seven chapters. Chapter 2 is a comprehensive literature review of the viruses in ballast water, methods used for the detection of viruses

in aquatic environments, and the application of metagenomics in aquatic viral ecology. Chapter 3 provides research questions and objectives addressed in this dissertation. Chapter 4 describes the materials and methods used for the research presented. Chapter 5 provides results and detailed discussion on the taxonomic composition and diversity of viruses in ballast water collected from a freshwater system (the Great Lakes). Chapter 6 focuses on the global transport of viruses through ballast water collected from marine environments. The potential impact of various engineered, management, and environmental parameters on ocean viromes is also presented. Chapter 7 summarizes key findings and outlines limitations and recommendations for future research as well as policy implications.

REFERENCES

REFERENCES

1. Altug, G.; Gurun, S.; Cardak, M.; Ciftci, P.; Kalkan, S., The occurrence of pathogenic bacteria in some ships' ballast water incoming from various marine regions to the Sea of Marmara, Turkey. *Marine Environmental Research* **2012,** *81*, 35-42.

2. Bain, M. B.; Cornwell, E. R.; Hope, K. M.; Eckerlin, G. E.; Casey, R. N.; Groocock, G. H.; Getchell, R. G.; Bowser, P. R.; Winton, J. R.; Batts, W. N.; Cangelosi, A.; Casey, J. W., Distribution of an Invasive Aquatic Pathogen (Viral Hemorrhagic Septicemia Virus) in the Great Lakes and Its Relationship to Shipping. *Plos One* **2010,** *5* (4).

3. Burkholder, J.; Hallegraeff, G.; Melia, G.; Cohen, A.; Bowers, H.; Oldach, D.; Parrow, M.; Sullivan, M.; Zimba, P.; Allen, E.; Kinder, C.; Mallin, M., Phytoplankton and bacterial assemblages in ballast water of US military ships as a function of port of origin, voyage time, and ocean exchange practices. *Harmful Algae* **2007,** *6* (4), 486-518.

4. Cox, G. W., Alien species and evolution: the evolutionary ecology of exotic plants, animals, microbes, and interacting native species. Island Press, Washington, DC, **2004**.

5. David, M.; Gollasch, S., Global Maritime Transport and Ballast Water Management Issues and Solutions. In Invading Nature - Springer Series in Invasion Ecology 8, Springer, Dordrecht, The Netherlands, **2015**.

6. DePaola, A.; Ulaszek, J.; Kaysner, C.; Tenge, B.; Nordstrom, J.; Wells, J.; Puhr, N.; Gendel, S., Molecular, serological, and virulence characteristics of Vibrio parahaemolyticus isolated from environmental, food, and clinical sources in north America and Asia. *Applied and Environmental Microbiology* **2003,** *69* (7), 3999-4005.

7. Department of Homeland Security, **2012,** Available at http://www.gpo.gov/fdsys/pkg/FR-2012-03-23/pdf/2012-6579.pdf.

8. Drake, J.; Lodge, D., Rate of species introductions in the Great Lakes via ships' ballast water and sediments. *Canadian Journal of Fisheries and Aquatic Sciences* **2007,** *64* (3), 530-538.

9. Drake, L.; Doblin, M.; Dobbs, F., Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. *Marine Pollution Bulletin* **2007,** *55* (7-9), 333-341.

10. Drake, L.; Ruiz, G.; Galil, B.; Mullady, T.; Friedman, D.; Dobbs, F., Microbial ecology of ballast water during a transoceanic voyage and the effects of open-ocean exchange. *Marine Ecology Progress Series* **2002,** *233*, 13-20.

11. Elsayed, E.; Faisal, M.; Thomas, M.; Whelan, G.; Batts, W.; Winton, J., Isolation of viral haemorrhagic septicaemia virus from muskellunge, Esox masquinongy (Mitchill), in Lake St Clair, Michigan, USA reveals a new sublineage of the North American genotype. *Journal of Fish Diseases* **2006,** *29* (10), 611-619.

12. Engineering Center Transzvuk, Hydrodynamic Ballast Water Treatment Technology, **2012,** Available at http://www.ballastwater.org/PDF/Ballast_Water.pdf.

13. Ferrate Treatment Technologies, Ballast Water, **2011,** Available at http://www.ferratetreatment.com/.

14. Garver, K. A.; Al-Hussinee, L.; Hawley, L. M.; Schroeder, T.; Edes, S.; LePage, V.; Contador, E.; Russell, S.; Lord, S.; Stevenson, R. M.; Souter, B.; Wright, E.; Lumsden, J. S., Mass mortality associated with koi herpesvirus in wild common carp in Canada. *Journal of Wildlife Diseases* **2010,** *46* (4), 1242-51.

15. Gollasch, S.; Rosenthal, H.; Botnen, H.; Hamer, J.; Laing, I.; Leppakoski, E.; Macdonald, E.; Minchin, D.; Nauke, M.; Olenin, S.; Utting, S.; Voigt, M.; Wallentinus, I., Fluctuations of zooplankton taxa in ballast water during short-term and long-term ocean-going voyages. *International Review of Hydrobiology* **2000,** *85* (5-6), 597-608.

16. Grimmett, S. G.; Warg, J. V.; Getchell, R. G.; Johnson, D. J.; Bowser, P. R., An unusual koi herpesvirus associated with a mortality event of common carp Cyprinus carpio in New York State, USA. *Journal of Wildlife Diseases* **2006,** *42* (3), 658-62.

17. International Maritime Organization, International convention for the control and management of ships' ballast water and sediments, **2004,** Available at http://www.uscg.mil/hq/cg5/cg522/cg5224/docs/BWM-Treaty.pdf.

18. Kolar, C.; Lodge, D., Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution* **2001,** *16* (4), 199-204.

19. Litchman, E., Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. *Ecology Letters* **2010,** *13* (12), 1560-1572.

20. Lumsden, J. S.; Morrison, B.; Yason, C.; Russell, S.; Young, K.; Yazdanpanah, A.; Huber, P.; Al-Hussinee, L.; Stone, D.; Way, K., Mortality event in freshwater drum Aplodinotus grunniens from Lake Ontario, Canada, associated with viral haemorrhagic septicemia virus, Type IV. *Diseases of Aquatic Organisms* **2007,** *76* (2), 99-111.

21. Mills, E.; Leach, J.; Carlton, J.; Secor, C., Exotic species in the Great Lakes - a history of biotic crises and anthropogenic introductions. *Journal of Great Lakes Research* **1993,** *19* (1), 1-54.

22. National Research Council (U.S.), Committee on Ships' Ballast Operations., Stemming the tide : controlling introductions of nonindigenous species by ships'

ballast water. National Academy Press, Washington, D.C., **1996**.

23. Ricciardi, A.; Cohen, J., The invasiveness of an introduced species does not predict its impact. *Biological Invasions* **2007,** *9* (3), 309-315.

24. Ruiz, G.; Rawlings, T.; Dobbs, F.; Drake, L.; Mullady, T.; Huq, A.; Colwell, R., Global spread of microorganisms by ships - Ballast water discharged from vessels harbours a cocktail of potential pathogens. *Nature* **2000,** *408* (6808), 49-50.

25. Sakai, A.; Allendorf, F.; Holt, J.; Lodge, D.; Molofsky, J.; With, K.; Baughman, S.; Cabin, R.; Cohen, J.; Ellstrand, N.; McCauley, D.; O'Neil, P.; Parker, I.; Thompson, J.; Weller, S., The population biology of invasive species. *Annual Review of Ecology and Systematics* **2001,** *32*, 305-332.

CHAPTER 2

LITERATURE REVIEW

2.1. Overview

This chapter is organized into four sections: (1) an overview of virus occurrence in ballast water, (2) a description of current methodologies used for the detection of viruses in aquatic environments, (3) a review of metagenomics applications in aquatic viral ecology, and (4) a review of methodologies used to prepare viral metagenomes (viromes) from aquatic environments.

2.2. Occurrence of viruses in ballast water

Invasions of archaea, bacteria, fungi, and viruses due to ballast water discharge have not received much attention because such invasions are much difficult to detect than invasions by macroorganisms, such as Zebra Mussels (Litchman, 2010). However, increased attention and research are needed on invasive microorganisms due to their capacity to invade and cause detrimental effects in new environments. Such attributes, which enhance the potential for invasion include high densities in natural water, ability to form resting stages, and potential pathogenicity and toxicity (Drake et al., 2007). Viral invasion, in particular, needs special attention not only because viruses are the most abundant biological entities in the sea (estimated 10^{30} viruses; Suttle, 2007) but also affect and control the abundance and diversity of algal and bacterial host populations by infection associated with outcomes such as lysis and gene transfer (e.g., antibiotic resistance; Wommack and Colwell, 2000).

Currently, there are a limited number of studies on viral dynamics within ballast tanks, all estimating the number of virus–like particles (VLPs; Table 2.1). In general, these studies revealed a high level of VLPs (7×10^9 to 3×10^{11} VLPs/liter) in ballast

water. Even higher numbers of VLPs were found in biofilms in ballast tanks than that found in ballast water (Drake et al., 2005; Drake et al., 2007). Studies also showed that there was no significant difference in viral levels between exchanged and unexchanged ballast water (Drake et al., 2002; Leichsenring and Lawrence, 2011). Drake et al. (2002) stressed that future research needs to examine viral community composition not solely in terms of viral level, as species composition in the discharged ballast water is a critical element in evaluating the risk of microbial invasion.

Table 2.1. Viral levels found in ballast water

Reference	Major finding
Ruiz et al., 2000	• Ballast waters of vessels arriving to Chesapeake Bay from foreign ports contained 7.4×10^9 VLPs L^{-1} (n = 7).
Drake et al., 2001	• Ballast waters of vessels arriving to Chesapeake Bay from foreign ports in 1996–2000 contained 1.4×10^{10} VLPs L^{-1} (n = 12). • In some cases, ballast water from the bottom of the tank had higher VLPs than ballast water at the surface.
Drake et al., 2002	• Average VLPs densities in ballast water varied from 0.7 to 3.8×10^{10} L^{-1} (n = 5) throughout transit. • No significant differences existed between exchanged and unexchanged ballast water on the final day of sampling. • The efficacy of open–ocean ballast water exchange to reduce invasion by non–indigenous microorganisms could not be determined solely on viral levels.
Drake et al., 2005	• Surface ballast waters of bulk carriers arriving from foreign ports to Chesapeake Bay contained 3×10^{11} VLPs L^{-1} (n = 4). • Biofilms in ballast tanks contained 6.3×10^{11} VLPs L^{-1} (n = 5).
Soto et al., 2005	• Ballast waters of vessels arriving to different ports in Chile contained 1.8 to 2.0×10^7 VLPs L^{-1}.
Wilhelm et al., 2006	• One cargo vessel sampled at five Great Lakes' ports contained 3.3×10^{11} VLPs L^{-1} (n = 5). • Pervasive distribution of cyanophages that infect the marine cyanobacterial isolate *Synechococcus* sp. was observed throughout the western basin of Lake Erie, as well as in locations within the central and eastern basins.
Drake et al., 2007	• Ballast waters of vessels arriving to Chesapeake Bay contained 1.39×10^{10} VLPs L^{-1} (n = 31). • 6.8×10^{19} VLPs, assuming a survival rate of 56% and applying estimates of ship traffic, were discharged annually to the lower Chesapeake Bay. • The potential delivery of viruses was greatest in ballast water > sediment and water residuals > biofilms.

Table 2.1 (cont'd)

Reference	Major finding
Leichsenring and Lawrence, 2011	• Ballast water exchange did not significantly reduce viral level during voyages. • Ballast tanks were highly variable with respect to total viral level, and the efficacy of exchange requires investigation into the dynamics of specific viruses.

Abbreviation: VLPs, virus–like particles.

2.3. Methods for the detection of viruses in aquatic environments

The methods used for the detection of viruses in aquatic environments can be divided into four categories: (1) direct visualization of viral particles using electron microscope (EM), (2) virus infectivity using cell culture, and molecular methods either (3) for polymerase chain reaction (PCR) when sequence information exists or (4) metagenomics requiring no prior knowledge of gene sequences.

2.3.1. Electron microscopy

Since most viruses cannot be seen under a light microscope, direct visualization of viral particles requires the use of EM. The EM has long been used in the discovery and description of viruses, including marine viruses (Goldsmith and Miller, 2009). As EM can be a rapid procedure, it is essential in identifying unknown agents of emerging diseases. Taking a visual look can also elucidate mechanisms of virus attachment and replication.

However, since EM is not a high–throughput technique and is labor intensive, it is not appropriate for processing multiple samples. The acquisition and maintenance of equipment is expensive and the operation requires an experienced observer. Another limitation is the relatively low sensitivity, which results in a high detection limit (10^5–10^6 viral particles per mL; Schramlová et al., 2010).

2.3.2. Cell culture

The use of host–virus systems has long served as the ‘golden standard’ for virus detection (Hamza et al., 2011). Cell lines that are susceptible to virus infection are used

for specifically to propagate viruses, which may produce cytopathic effects (CPE) observable under a light microscope. The advantages of using cell culture for virus detection include good specificity and sensitivity, use of large sample volume, direct indication of viral infectivity, and the ability to isolate viruses of interest for further characterization (Leland and Ginocchio, 2007).

On the other hand, one of the main limitations of cell culture techniques is that most of the viruses, such as norovirus, cannot be cultivated in a conventional cell culture system (Winner and Hugenholtz, 2013). In addition, all viruses cannot be propagated in one cell line, requiring different cell lines to detect different viruses. Other disadvantages of conventional cell culture include long incubation times (e.g., days to weeks) from virus inoculation to the time when CPE become visible by light microscopy, inability to detect noncytopathic viruses, the intensive labor needed, and the expense. Lastly, cell culture is susceptible to toxic substances in the environmental samples, which leads to cell die–off and potentially false–positive results (Leland and Ginocchio, 2007; Rodríguez et al., 2009). These limitations make cell culture a more challenging method for routine monitoring of viruses in environmental water.

2.3.3. Polymerase chain reaction (PCR)

The advent of the PCR overcomes some of the limitations of conventional cell culture technique and has greatly enhanced the ability to detect viruses in the environment. This is especially useful for nonculturable or noncytopathic viruses. As PCR drastically reduces time needed for virus detection and has higher specificity, it has been widely used to monitor viruses, especially those causing diseases in aquatic environments.

However, one important limitation of PCR is that it does not provide any information about the infectious state of detected viruses. The detection of viral nucleic acid does not necessarily represent an infectious virus, which is often essential in addressing water and food safety (Sobsey et al., 1988) as well as in understanding virus persistence in the environment. PCR is susceptible to inhibitory compounds such as humic acids found in environmental samples, leading to false–negative results. Moreover, PCR requires a prior knowledge of sequence information, limiting its application only on known viruses. Considering viruses lack universally conserved phylogenetic marker, such as the 16S rRNA gene, shared by all bacteria and archaea (Rohwer and Edwards, 2002), PCR is less useful for investigating an array of viruses in environmental water. Currently, marker genes are limited to specific viral groups, such as the T4–like myoviruses (e.g., major capsid and portal proteins), T7–like podoviruses (e.g., DNA polymerase), or *Phycodnaviridae* (e.g., DNA polymerase; Sullivan, 2015).

2.3.4. Metagenomics

During the past decade, metagenomics with dramatic evolution of high–throughput sequencing technologies have revolutionized microbiological studies and provided new insights into the diversity and dynamics of microbial communities. Metagenomics is sequence–based analysis of the whole collection of genomes directly isolated from a sample (Handelsman et al., 1998). Metagenomics overcomes the principal limitations of the classical tools for virus detection and can provide a comprehensive view of the microbial communities. It does not require virus isolation used in cell culture

techniques nor does it rely on target genomic sequences that are expected to be present used in PCR techniques.

Viruses are well suited to be studied by metagenomic approaches because they are genetically diverse and most of them are unculturable. The small size of viral genomes is also advantageous for bioinformatics analysis (Thurber et al., 2009). Metagenomic approaches have been used to explore viral communities in a wide range of environments, including oceans and freshwater (described in the next section), soil (Fierer et al., 2007), wastewater (Cantalupo et al., 2011; Aw et al., 2014), acidic hot springs (Rice et al., 2001), human feces (Zhang et al., 2006; Breitbart et al., 2008; Finkbeiner et al., 2008), and human respiratory tract (Willner et al., 2009).

While some characteristics of viruses such as genetic abundance, unculturability, and small genome size make them suitable for metagenomic approaches, other aspects of viruses complicate metagenomic approaches, including: (1) the wide range of viral particle sizes, shapes, densities, and sensitivities, (2) variation in viral genome type (DNA vs. RNA and single–stranded (ss) vs. double–stranded (ds)) and length (Thurber, 2011), (3) high mutation rate and divergence, (4) existence in a proviral form, (5) incompleteness of public viral genome database, and (6) current bioinformatics tools designed mainly on the analysis of bacterial communities (Fancello et al., 2012). The aspects of (4), (5), and (6) as described above result in a large part of sequencing reads (average 40% to 50%, occasionally up to 90% of sequencing reads) classified as "unknown" (Rosario and Breitbart, 2011), which limits extracting meaningful information from virome data sets.

2.4. Application of metagenomics to study viruses in aquatic environments

The number of viral metagenomic studies has increased gradually since the first virome study by Breitbart et al. (2002). Several studies have demonstrated the feasibility of metagenomic approaches to examine viral communities in various complex environmental systems from freshwater and marine to host–associated environments, as described in the previous section. To identify key findings and research gaps associated with virome studies in aquatic environments, a review of 16 published studies using high–throughput sequencing technologies was undertaken and summarized in Table 2.2 (studies using cloning followed by Sanger–based sequencing were not included here). The general findings are as follows:

(1) Most viral metagenomic investigations have focused on marine environments. Recently, studies have started describing freshwater viromes and adding new virome data sets associated with freshwater environments to the currently limited virome databases. Among the freshwater viromes, most of the studies investigated viromes from non–natural (reclaimed and potable waters by Rosario et al., 2009; aquaculture ponds by Rodriguez–Brito et al., 2010) or extreme environments (an Antarctic lake by Lopez–Bueno et al., 2009; desert ponds by Fancello et al., 2013).

(2) Current studies of aquatic viromes have focused on either DNA or RNA viruses (but not both). Although the genetic material of viruses consists of dsDNA, ssDNA, dsRNA, or ssRNA, most studies have examined DNA viruses, most of which were DNA phages. To date, only one study looked into freshwater RNA viruses where the authors suggested that the freshwater lake ecosystems might serve repositories of pathogenic and non–pathogenic RNA viruses because of their direct contact with

humans, and domestic and wild animals (Djikeng et al., 2009). Lack of research on RNA viruses limits our understanding of their genetic diversity and impact on host populations in aquatic environments.

(3) The majority of sequences in virome data sets does not have significant sequence similarity to current databases or have higher homology. This demonstrates the limited knowledge about the genetic diversity of viruses as suggested by Mokili et al. (2012) where only less than 1% of the extant viral diversity has been currently explored.

(4) Bioinformatics analyses of viral communities revealed a high degree of genetic diversity of known viruses. Indeed, viral metagenomics enabled the in–depth characterization of viral communities that would not have been possible with traditional methods. Previous virome studies showed that aquatic environments harbor viruses of many different viral families, which infect a wide range of hosts, including bacteria as well as plants/fungi, invertebrates, and vertebrates (including humans).

(5) Among known viruses, viromes were largely dominated by dsDNA phages belonging to the order *Caudovirales* (i.e., *Myoviridae*, *Podoviridae*, and *Siphoviridae*). In contrast, a few virome studies showed that the viral communities were dominated by ssDNA phage (López Bueno et al., 2009; Roux et al., 2012; de Cárcer et al., 2015). However, these studies applied multiple displacement amplification (MDA) to obtain enough DNA prior to metagenomic sequencing, which is known to preferentially amplify ssDNA viruses (Kim and Bae, 2011). This might have produced a biased estimate of true relative abundances of viral communities, resulting in dominant ssDNA phages in the viral communities.

(6) Lastly, as opposed to the Bass Becking's view of microbial distributions (everything is everywhere; de Wit & Bouvier, 2006), recent virome studies (Angly et al., 2006; Brum et al., 2015) have shown that viruses exhibited geographical patterns. Moreover, these studies have revealed the significant impact of local environmental conditions on structuring viral diversity. In addition to geographic variation, temporal variation of diversity and relative abundance of viral communities has also been observed in aquatic environments (López–Bueno et al., 2009; Tseng et al., 2013).

Table 2.2. Metagenomic studies in aquatic viral ecology

Reference	Name	Major finding
Angly et al., 2006	Marine viromes	• Composition of viromes varied in different geographic regions. • Global viral diversity was high but regional diversity could be almost as high due to viral migration. • Some viral species were endemic and others were ubiquitous, the vast majority was widespread and shared between several oceanic regions.
Dinsdale et al., 2008	Viromes	• 42 viromes showed strongly discriminatory metabolic profiles across environments. • Most of the functional diversity was maintained in all of the communities, but the relative occurrence of metabolisms varied, and the differences between metagenomes predicted the biogeochemical conditions of each environment.
Djikeng et al., 2009	Freshwater lake RNA viruses	• The majority of sequences did not show any significant similarity to known sequences. • The known sequences were mainly from viral types with significant similarity to approximately 30 viral families. • Viral sequences closely related to Banna Virus and distantly to Israeli Acute Paralysis virus were found.
López–Bueno et al., 2009	Antarctic Lake viral community	• Antarctic virome had a large proportion of sequences related to eukaryotic viruses, including phycodnaviruses and ssDNA viruses. • Transition from an ice–covered lake in spring to an open–water lake in summer led to a change from a ssDNA virus– to a dsDNA virus–dominated assemblage.

Table 2.2 (cont'd)

Reference	Name	Major finding
Rosario et al., 2009	Reclaimed water viral community	• Most of the viruses in reclaimed and potable water were novel. • Phages dominated the DNA viral community in reclaimed and potable water, but reclaimed water had a distinct phage community. • Eukaryotic viruses similar to plant pathogens and invertebrate picornaviruses dominated RNA viromes.
Rodriguez-Brito et al., 2010	Aquatic viral community	• Viromes from human–controlled aquatic environments at various time points showed continuous variation of viruses and their relative abundances at the genotype level.
Rooks et al., 2010	Freshwater viruses	• The most abundant viral genotypes in the pond on a cattle farm were phages. • The predominant viral genotypes infecting higher life forms found in association with the farm were pathogens that cause disease in cattle and humans (*Herpesviridae*).
Roux et al., 2012	Freshwater viral communities	• Viral species richness in a mesotrophic lake was greater than the one in an oligotrophic lake. • Freshwater viral communities appeared genetically distinct from other aquatic ecosystems.
Williamson et al., 2012	Indian Ocean viruses	• Size fractionation of marine microbial communities enriched for specific groups of viruses within the different size classes. • A relative enrichment for metabolic proteins of viral origin that potentially reflected the physiological condition of host cells was found.
Fancello et al., 2013	Perennial ponds viral communities	• Sequences belonging to tailed phages were the most abundant in four perennial ponds. • A decrease in the local viral biodiversity was observed in a pond with sustained human activities.

Table 2.2 (cont'd)

Reference	Name	Major finding
Hurwitz and Sullivan, 2013	Pacific Ocean virome	• Quantitative data set and protein clusters organization have a potential to provide an invaluable mapping resource for future comparative viral metagenomic research.
Tseng et al., 2013	Subtropical freshwater reservoir viromes	• Viral community regularly showed higher relative abundances and diversity during summer in comparison to winter, with major variations happening in several viral families, including *Siphoviridae*, *Myoviridae*, *Podoviridae*, and *Microviridae*.
Martínez et al., 2014	Marine viruses	• Fluorescence–activated sorting approach was an effective way to target and investigate specific virus groups.
Winter et al., 2014	Deep–water viromes	• The identifiable relative abundance in viromes from the Atlantic Ocean (5200 m depth) and the Mediterranean Sea (2400 m depth) were dominated by archaeal and bacterial viruses. • Contrasting deep–sea environments of the Atlantic Ocean and the Mediterranean Sea shared a common core set of virus types constituting the majority of both viral communities.
Brum et al., 2015	Ocean viral communities	• Viral communities were passively transported on oceanic currents and locally structured by environmental conditions that affected host community structure.
de Cárcer et al., 2015	Polar freshwater DNA viruses	• Arctic viromes were dominated by unknown and ssDNA viruses. • Arctic viromes presented some minor genetic overlap with an Antarctic Ocean virome.

Abbreviations: virome, viral metagenome; ds, doube–stranded; ss, single–stranded.

2.5. Methods used to prepare viromes in aquatic environments

Preparing high quality viromes is critical as it has a great influence upon the sequence data and the interpretation of the results. However, despite a wide use of metagenomic approaches for studying viral communities, the sample collection to metagenomic sequencing workflow is still experimentally challenging at each step. A few methodological studies have been published to describe and evaluate the virome preparation workflow for environmental water samples (Thurber et al., 2009; Thurber, 2011; Duhaime and Sullivan, 2012; Hurwitz et al., 2013). To identify general patterns and challenges associated with virome preparation, a review of 16 published studies was undertaken. Table 2.3 summarized methods used to prepare viromes from environmental water samples (studies using cloning followed by sanger–based sequencing were not included here).

Table 2.3. Methods used in published studies to prepare viromes from aquatic environments

Reference	Sample type	Virus concentration	Virus purification	Nucleic acid	Nucleic acid extraction	Nucleic acid amplification	Sequencing platform
Angly et al., 2006	Marine water	TFF	0.22 μm + DNase & RNase + CsCl	DNA	Formamide/CTAB	MDA	454 GS20
Dinsdale et al., 2008	Marine and lake water	TFF	CsCl	DNA	Phenol/chloroform + CTAB	MDA	454 GS20
McDaniel et al., 2008	Marine water	TFF + PEG	0.22 μm + CsCl	DNA	Formamide/CTAB	MDA	454 GS20
Djikeng et al., 2009	Lake water	TFF	Ultracent + DNase & RNase	RNA	Qiagen viral RNA preparation kit	RP-SISPA	454 GS–FLX
López Bueno et al., 2009	Antarctic lake water	TFF	0.45 μm + DNase	DNA	Phenol/chloroform	MDA	454 GS–FLX
Rosario et al., 2009	Reclaimed water	TFF + PEG	0.22 μm + CsCl + DNase	DNA	QIAmp MinElute Virus Spin Kit	MDA	454 GS–FLX
Rodriguez-Brito et al., 2010	Pond and solar saltern	TFF + PEG	CsCl	DNA	Formamide/CTAB	MDA	454 GS20
Rooks et al., 2010	Farm pond	$NaCl_2$	DNase & RNase + PEG	DNA	Phenol/chloroform	–	454 GS–FLX
Roux et al., 2012	Lake water	TFF + PEG	0.22 μm + DNase	DNA	QIAamp DNA mini kit	MDA	454 GS–FLX

Table 2.3 (cont'd)

Reference	Sample type	Virus concentration	Virus purification	Nucleic acid	Nucleic acid extraction	Nucleic acid amplification	Sequencing platform
Williamson et al., 2012	Marine water	TFF	DNase + Sucrose	DNA	Phenol/chloroform	LASL	454 GS–FLX Titanium
Fancello et al., 2013	Pond	PEG	CsCl + DNase	DNA	Formamide/CTAB	MDA	454 GS20
Hurwitz and Sullivan, 2013	Marine water	$FeCl_3$	CsCl + DNase	DNA	PCR DNA purification	LA	454 GS–FLX Titanium
Tseng et al., 2013	Freshwater reservoir	TFF	CsCl + DNase	DNA	Formamide/CTAB	MDA	454 GS–FLX
Winter et al., 2014	Marine water	TFF	–	DNA	QIAmp MinElute Virus Spin Kit	–	454 GS–FLX Titanium
Brum et al., 2015	Marine water	$FeCl_3$	DNase	DNA	Phenol/chloroform	LA	Illumina HiSeq 2000
de Cárcer et al., 2015	Arctic lake water	TFF	DNase & RNase + Sucrose	DNA	Phenol/chloroform	MDA	454 GS–FLX

Table 2.3 (cont'd)

Abbreviations: TFF, tangential flow filtration; PEG, polyethylene glycol precipitation; $FeCl_3$, iron chloride precipitation; 0.22 μm, 0.22 μm filtration; 0.45 μm, 0.45 μm filtration; Ultracent, ultracentrifugation; CsCl, cesium chloride density gradient; Sucrose, sucrose density gradient; CTAB, cetyltrimethyl ammonium bromide; MDA, multiple displacement amplification; RP–SISPA, random priming mediated sequence independent single primer amplification; LA, linker amplification; LASL, linker amplified shotgun library.

Concentration of large volumes of water is needed to recover sufficient quantity of viral particles for more efficient nucleic acid extraction (Edwards and Rohwer, 2005; Delwart, 2007; Thurber et al., 2009; Duhaime and Sullivan, 2012; Mokili et al., 2012). In general, tangential flow filtration (TFF) is used as the first step to concentrate VLPs from large volumes of environmental water samples. TFF is a technique used for concentrating diverse microorganisms in water samples based on size–exclusion (Morales–Morales et al., 2003). In some virome studies (López-Bueno et al., 2009; Roux et al., 2012), water samples were pre–filtered (e.g., 25 μm, 1.2 μm, or 0.22 μm filtration) before TFF to minimize potential changes in viral communities derived from cellular organisms and to prevent a filter clogging problem.

Following VLPs concentration, at least one of the purification methods, such as 0.45 μm and/or 0.22 μm filtrations, DNase and/or RNase treatments, cesium chloride (CsCl) density gradient ultracentrifugation, chloroform treatment, and sucrose gradient, is generally used to purify VLPs in the TFF concentrates. These steps serve to reduce contamination by non–viral cells and increase the levels of viral nucleic acids that results in a maximum amount of viral sequences and thus higher sequencing coverage of viruses (Bibby, 2013). Hurwitz et al. (2013) reported that the choice of purification method had much less impact on the resulting virome sequence data set than that of VLPs concentration or amplification method. Therefore, the purification method should be chosen depending on research question or sample type.

However, careful consideration is needed when choosing a purification method as it may affect the composition of metagenomes. For example, some virome studies (Angly et al., 2006; de Cárcer et al., 2015) used both DNase and RNase for exonuclease

digestion to reduce free cellular nucleic acids. However, some RNA viruses contain RNA into their coat structure, and thus treatment with RNase may destroy those particular RNAs (Thurber et al., 2009). Another example is the use of 0.22 μm filtration to reduce microbial contamination. The discovery of giant viruses larger than bacteria (in amoebae from the water of a cooling tower by La Scola et al., 2003; from coast and freshwater pond by Philippe et al., 2013) indicates that the common use of 0.22 μm filtration may induce under–sampling of these giant viruses. To avoid this problem, Lopez–Bueno et al (2009) used only 0.45 μm filters for preparing viromes and found no bacterial genomic contamination.

Methods to extract and amplify viral nucleic acids are chosen depending on the target nucleic acid. While there have been few studies investigating both DNA and RNA viral communities in wastewater treatment systems (Cantalupo et al., 2011; Alhamlan et al., 2013; Bibby and Peccia, 2013; Aw et al., 2014), all studies as shown in Table 2.3 targeted either DNA or RNA viruses (but not both) in natural aquatic environments. As it is now well known that RNA viruses are also present in aquatic environments (e.g., freshwater lake by Djikeng et al., 2009), preparation of both DNA and RNA viromes will provide a more comprehensive view of aquatic viromes.

Unlike clinical samples, low density of viruses in environmental water samples makes nucleic acid amplification steps inevitable prior to metagenomic sequencing. In addition, various viral genome types in environmental samples complicate the viral genome amplification step. Common methods used to amplify viral nucleic acids are restricted to certain types of viral genomes. For example, only dsDNA viruses are detected using the linker amplified shotgun library (LASL) method (Breitbart et el.,

2002) and circular viral genomes are selectively amplified using the MDA technique (Kim and Bae, 2011; Gilbert et al., 2010).

2.6. Summary of literature review

Despite methodological challenges in viral metagenomic studies, metagenomics is currently believed to be the most suitable approach to overcome limitations of classical viral detection methods and allows for an in-depth examination of viral communities. Previous virome studies have added a wealth of knowledge to viral ecology with the maturation of metagenomics approaches and high–throughput sequencing technologies. They demonstrated that highly diverse viruses are present in a wide range of aquatic environments with geographic variation. Consequently, this non–cosmopolitan distribution pattern creates the potential for invasive species to arise, when non–native species invade and spread into new habitats. These significant discoveries were the motivation for this dissertation to characterize viral communities in ballast water and to investigate their transport across different geographical locations as a result of global commerce. Chapter 3 provides research questions and objectives addressed in this dissertation.

REFERENCES

REFERENCES

1. Alhamlan, F.; Ederer, M.; Brown, C.; Coats, E.; Crawford, R., Metagenomics-based analysis of viral communities in dairy lagoon wastewater. *Journal of Microbiological Methods* **2013,** *92* (2), 183-188.

2. Angly, F.; Felts, B.; Breitbart, M.; Salamon, P.; Edwards, R.; Carlson, C.; Chan, A.; Haynes, M.; Kelley, S.; Liu, H.; Mahaffy, J.; Mueller, J.; Nulton, J.; Olson, R.; Parsons, R.; Rayhawk, S.; Suttle, C.; Rohwer, F., The marine viromes of four oceanic regions. *Plos Biology* **2006,** *4* (11), 2121-2131.

3. Aw, T. G.; Howe, A.; Rose, J. B., Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. *Journal of Virological Methods* **2014,** *210C*, 15-21.

4. Bibby, K., Metagenomic identification of viral pathogens. *Trends in Biotechnology* **2013,** *31* (5), 275-9.

5. Bibby, K.; Peccia, J., Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Journal of Environmental Science and Technology* **2013,** *47* (4), 1945-51.

6. Breitbart, M.; Haynes, M.; Kelley, S.; Angly, F.; Edwards, R.; Felts, B.; Mahaffy, J.; Mueller, J.; Nulton, J.; Rayhawk, S.; Rodriguez-Brito, B.; Salamon, P.; Rohwer, F., Viral diversity and dynamics in an infant gut. *Research in Microbiology* **2008,** *159* (5), 367-373.

7. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.; Segall, A.; Mead, D.; Azam, F.; Rohwer, F., Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99* (22), 14250-14255.

8. Brum, J.; Ignacio-Espinoza, J.; Roux, S.; Doulcier, G.; Acinas, S.; Alberti, A.; Chaffron, S.; Cruaud, C.; de Vargas, C.; Gasol, J.; Gorsky, G.; Gregory, A.; Guidi, L.; Hingamp, P.; Iudicone, D.; Not, F.; Ogata, H.; Pesant, S.; Poulos, B.; Schwenck, S.; Speich, S.; Dimier, C.; Kandels-Lewis, S.; Picheral, M.; Searson, S.; Bork, P.; Bowler, C.; Sunagawa, S.; Wincker, P.; Karsenti, E.; Sullivan, M.; Coordinators, T. O.; Coordinators, T. O., Patterns and ecological drivers of ocean viral communities. *Science* **2015,** *348* (6237).

9. Cantalupo, P.; Calgua, B.; Zhao, G.; Hundesa, A.; Wier, A.; Katz, J.; Grabe, M.; Hendrix, R.; Girones, R.; Wang, D.; Pipas, J., Raw Sewage Harbors Diverse Viral Populations. *Mbio* **2011,** *2* (5).

10. de Cárcer, D.; López-Bueno, A.; Pearce, D.; Alcamí, A., Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances* **2015,** 1, e1400127.

11. de Wit, R.; Bouvier, T., 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology* **2006,** *8* (4), 755-758.

12. Delwart, E., Viral metagenomics. *Reviews in Medical Virology* **2007,** *17* (2), 115-131.

13. Dinsdale, E. A.; Edwards, R. A.; Hall, D.; Angly, F.; Breitbart, M.; Brulc, J. M.; Furlan, M.; Desnues, C.; Haynes, M.; Li, L.; McDaniel, L.; Moran, M. A.; Nelson, K. E.; Nilsson, C.; Olson, R.; Paul, J.; Brito, B. R.; Ruan, Y.; Swan, B. K.; Stevens, R.; Valentine, D. L.; Thurber, R. V.; Wegley, L.; White, B. A.; Rohwer, F., Functional metagenomic profiling of nine biomes. *Nature* **2008,** *452* (7187), 629-32.

14. Djikeng, A.; Kuzmickas, R.; Anderson, N.; Spiro, D., Metagenomic Analysis of RNA Viruses in a Fresh Water Lake. *Plos One* **2009,** *4* (9).

15. Drake, L.; Doblin, M.; Dobbs, F., Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. *Marine Pollution Bulletin* **2007,** *55* (7-9), 333-341.

16. Drake, L.; Meyer, A.; Forsberg, R.; Baier, R.; Doblin, M.; Heinemann, S.; Johnson, W.; Koch, M.; Rublee, P.; Dobbs, F., Potential invasion of microorganisms and pathogens via 'interior hull fouling': biofilms inside ballast water tanks. *Biological Invasions* **2005,** *7* (6), 969-982.

17. Drake, L.; Ruiz, G.; Galil, B.; Mullady, T.; Friedman, D.; Dobbs, F., Microbial ecology of ballast water during a transoceanic voyage and the effects of open-ocean exchange. *Marine Ecology Progress Series* **2002,** *233*, 13-20.

18. Drake, L. A.; Choi, K.-H.; Ruiz, G. M.; Dobbs, F. C., Global redistribution of bacterioplankton and virioplankton communities. *Biological Invasions* **2001,** *3* (2), 193-199.

19. Duhaime, M.; Sullivan, M., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **2012,** *434* (2), 181-186.

20. Edwards, R.; Rohwer, F., Viral metagenomics. *Nature Reviews Microbiology* **2005,** *3* (6), 504-510.

21. Fancello, L.; Trape, S.; Robert, C.; Boyer, M.; Popgeorgiev, N.; Raoult, D.; Desnues, C., Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *Isme Journal* **2013,** *7* (2), 359-369.

22. Fancello, L.; Raoult, D.; Desnues, C., Computational tools for viral metagenomics and their application in clinical research. *Virology* **2012,** *434* (2), 162-174.

23. Fierer, N.; Breitbart, M.; Nulton, J.; Salamon, P.; Lozupone, C.; Jones, R.; Robeson,

M.; Edwards, R.; Felts, B.; Rayhawk, S.; Knight, R.; Rohwer, F.; Jackson, R., Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology* **2007,** *73* (21), 7059-7066.

24. Finkbeiner, S.; Allred, A.; Tarr, P.; Klein, E.; Kirkwood, C.; Wang, D., Metagenomic analysis of human diarrhea: Viral detection and discovery. *Plos Pathogens* **2008,** *4* (2).

25. Goldsmith, C.; Miller, S., Modern Uses of Electron Microscopy for Detection of Viruses. *Clinical Microbiology Reviews* **2009,** *22* (4), 552-+.

26. Gilbert J, Zhang K, Neufeld J., Multiple Displacement Amplification, Handbook of Hydrocarbon and Lipid Microbiology, **2010,** 4255-4263.

27. Hamza, I.; Jurzik, L.; Uberla, K.; Wilhelm, M., Methods to detect infectious human enteric viruses in environmental water samples. *International Journal of Hygiene and Environmental Health* **2011,** *214* (6), 424-436.

28. Handelsman, J.; Rondon, M.; Brady, S.; Clardy, J.; Goodman, R., Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology* **1998,** *5* (10), R245-R249.

29. Hurwitz, B.; Deng, L.; Poulos, B.; Sullivan, M., Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology* **2013,** *15* (5), 1428-1440.

30. Hurwitz, B.; Sullivan, M., The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *Plos One* **2013,** *8* (2).

31. Kim, K.; Bae, J., Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Applied and Environmental Microbiology* **2011,** *77* (21), 7663-7668.

32. La Scola, B.; Audic, S.; Robert, C.; Jungang, L.; de Lamballerie, X.; Drancourt, M.; Birtles, R.; Claverie, J.; Raoult, D., A giant virus in amoebae. *Science* **2003,** *299* (5615), 2033-2033.

33. Leichsenring, J.; Lawrence, J., Effect of mid-oceanic ballast water exchange on virus-like particle abundance during two trans-Pacific voyages. *Marine Pollution Bulletin* **2011,** *62* (5), 1103-1108.

34. Leland, D.; Ginocchio, C., Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews* **2007,** *20* (1), 49-+.

35. Litchman, E., Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. *Ecology Letters* **2010,** *13* (12), 1560-1572.

36. Lopez-Bueno, A.; Tamames, J.; Velazquez, D.; Moya, A.; Quesada, A.; Alcami, A., High Diversity of the Viral Community from an Antarctic Lake. *Science* **2009,** *326* (5954), 858-861.

37. Martinez, J.; Swan, B.; Wilson, W., Marine viruses, a genetic reservoir revealed by targeted viromics. *Isme Journal* **2014,** *8* (5), 1079-1088.

38. McDaniel, L.; Breitbart, M.; Mobberley, J.; Long, A.; Haynes, M.; Rohwer, F.; Paul, J., Metagenomic Analysis of Lysogeny in Tampa Bay: Implications for Prophage Gene Expression. *Plos One* **2008,** *3* (9).

39. Mokili, J.; Rohwer, F.; Dutilh, B., Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2012,** *2* (1), 63-77.

40. Morales-Morales, H.; Vidal, G.; Olszewski, J.; Rock, C.; Dasgupta, D.; Oshima, K.; Smith, G., Optimization of a reusable hollow-fiber ultrafilter for simultaneous concentration of enteric bacteria, protozoa, and viruses from water. *Applied and Environmental Microbiology* **2003,** *69* (7), 4098-4102.

41. Philippe, N.; Legendre, M.; Doutre, G.; Coute, Y.; Poirot, O.; Lescot, M.; Arslan, D.; Seltzer, V.; Bertaux, L.; Bruley, C.; Garin, J.; Claverie, J.; Abergel, C., Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **2013,** *341* (6143), 281-286.

42. Rice, G.; Stedman, K.; Snyder, J.; Wiedenheft, B.; Willits, D.; Brumfield, S.; McDermott, T.; Young, M., Viruses from extreme thermal environments. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (23), 13341-13345.

43. Rodriguez, R.; Pepper, I.; Gerba, C., Application of PCR-Based Methods To Assess the Infectivity of Enteric Viruses in Environmental Samples. *Applied and Environmental Microbiology* **2009,** *75* (2), 297-307.

44. Rodriguez-Brito, B.; Li, L.; Wegley, L.; Furlan, M.; Angly, F.; Breitbart, M.; Buchanan, J.; Desnues, C.; Dinsdale, E.; Edwards, R.; Felts, B.; Haynes, M.; Liu, H.; Lipson, D.; Mahaffy, J.; Martin-Cuadrado, A.; Mira, A.; Nulton, J.; Pasic, L.; Rayhawk, S.; Rodriguez-Mueller, J.; Rodriguez-Valera, F.; Salamon, P.; Srinagesh, S.; Thingstad, T.; Tran, T.; Thurber, R.; Willner, D.; Youle, M.; Rohwer, F., Viral and microbial community dynamics in four aquatic environments. *Isme Journal* **2010,** *4* (6), 739-751.

45. Rohwer, F.; Edwards, R., The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology* **2002,** *184* (16), 4529-4535.

46. Rooks, D. J.; Smith, D. L.; McDonald, J. E.; Woodward, M. J.; McCarthy, A. J.; Allison, H. E., 454-pyrosequencing: a molecular battiscope for freshwater viral ecology. *Genes* **2010,** *1* (2), 210-26.

47. Rosario, K.; Breitbart, M., Exploring the viral world through metagenomics. *Current Opinion in Virology* **2011,** *1* (4), 289-297.

48. Rosario, K.; Nilsson, C.; Lim, Y.; Ruan, Y.; Breitbart, M., Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **2009,** *11* (11), 2806-2820.

49. Roux, S.; Enault, F.; Robin, A.; Ravet, V.; Personnic, S.; Theil, S.; Colombet, J.; Sime-Ngando, T.; Debroas, D., Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *Plos One* **2012,** *7* (3).

50. Ruiz, G.; Rawlings, T.; Dobbs, F.; Drake, L.; Mullady, T.; Huq, A.; Colwell, R., Global spread of microorganisms by ships - Ballast water discharged from vessels harbours a cocktail of potential pathogens. *Nature* **2000,** *408* (6808), 49-50.

51. Schramlová J.; Arientová S.; Hulínská D, The role of electron microscopy in the rapid diagnosis of viral infections. *Folia Microbiologica* **2010,** 55 (1), 88-101.

52. Sobsey, M.; Battigelli, D.; Shin, G.; Newland, S., RT-PCR amplification detects inactivated viruses in water and wastewater. *Water Science and Technology* **1998,** *38* (12), 91-94.

53. Soto, K.; Durán, R.; Kuznar, J., Rapid examination of microorganisms in ballast waters. *Revista de Biología Marina y Oceanografía* **2005,** 40 (1), 77–82.

54. Sullivan, M., Viromes, Not gene markers, for studying double-stranded DNA virus communities. *Journal of Virology* **2015,** 89, 2459–2461.

55. Suttle, C., Marine viruses - major players in the global ecosystem. *Nature Reviews Microbiology* **2007,** *5* (10), 801-812.

56. Thurber R., Methods in Viral Metagenomics, Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats, **2011,** 15–24.

57. Thurber, R.; Haynes, M.; Breitbart, M.; Wegley, L.; Rohwer, F., Laboratory procedures to generate viral metagenomes. *Nature Protocols* **2009,** *4* (4), 470-483.

58. Tseng, C.; Chiang, P.; Shiah, F.; Chen, Y.; Liou, J.; Hsu, T.; Maheswararajah, S.; Saeed, I.; Halgamuge, S.; Tang, S., Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *Isme Journal* **2013,** *7* (12), 2374-2386.

59. Wilhelm, S. W.; Carberry, M. J.; Eldridge, M. L.; Poorvin, L.; Saxton, M. A.; Doblin, M. A., Marine and freshwater cyanophages in a Laurentian Great Lake: Evidence from infectivity assays and molecular analyses of g20 genes. *Applied and Environmental Microbiology* **2006,** *72* (7), 4957-4963.

60. Williamson, S.; Allen, L.; Lorenzi, H.; Fadrosh, D.; Brami, D.; Thiagarajan, M.; McCrow, J.; Tovchigrechko, A.; Yooseph, S.; Venter, J., Metagenomic Exploration

of Viruses throughout the Indian Ocean. *Plos One* **2012,** *7* (10).

61. Willner, D.; Furlan, M.; Haynes, M.; Schmieder, R.; Angly, F.; Silva, J.; Tammadoni, S.; Nosrat, B.; Conrad, D.; Rohwer, F., Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. *Plos One* **2009,** *4* (10).

62. Willner, D.; Hugenholtz, P., From deep sequencing to viral tagging: Recent advances in viral metagenomics. *Bioessays* **2013,** *35* (5), 436-442.

63. Winter, C.; Garcia, J.; Weinbauer, M.; DuBow, M.; Herndl, G., Comparison of Deep-Water Viromes from the Atlantic Ocean and the Mediterranean Sea. *Plos One* **2014,** *9* (6).

64. Wommack, K. E.; Colwell, R. R., Virioplankton: Viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* **2000,** *64* (1), 69-+.

65. Zhang, T.; Breitbart, M.; Lee, W.; Run, J.; Wei, C.; Soh, S.; Hibberd, M.; Liu, E.; Rohwer, F.; Ruan, Y., RNA viral community in human feces: Prevalence of plant pathogenic viruses. *Plos Biology* **2006,** *4* (1), 108-118.

CHAPTER 3

RESEARCH QUESTIONS AND OBJECTIVES

Previous viral metagenomic studies have provided a tremendous amount of valuable knowledge about the diversity of viruses and their important roles in mediating bacterial and eukaryotic diversity and biogeochemical cycling. Despite the fundamental importance of viruses in aquatic ecosystems and the capability of metagenomics for exploring uncultured viruses, metagenomic approaches have not yet been applied to fill existing knowledge gaps in global transport of viruses via ships' ballast water. This research is the first to investigate a metagenomic profile of viral communities in ballast water and to examine the influence of ships' ballast water on the transport of viral metagenome (virome) at a global scale. The insights provided by this research will add much needed knowledge to the area, which currently lags behind the understanding of invasive macro–organisms.

In this dissertation, the following research questions are addressed:

1. a) Does ballast water contain diverse viruses infecting a wide range of hosts?

   b) Does ballast water have a characteristic virome signature that is not found in harbor water?

   c) Does freshwater have a distinct virome signature from other aquatic ecosystems?

2. a) What is the influence of global shipping on transport of the ocean virome?

b) How do engineered, management, and environmental variables affect the differences in the ocean virome?

c) What is the potential for invasion by rare viral pathogens?

Based on these research questions, the objectives of this research are as follows:

1. To evaluate and optimize a workflow from sample collection to metagenomic sequencing to prepare ballast and harbor water viromes for a comprehensive view of viral communities;

2. To investigate taxonomic composition and diversity of viruses in ballast and harbor waters collected from a freshwater system (using the Great Lakes as a model system) and to understand the Great Lakes virome signatures;

3. To apply viral metagenomics for a large–scale research of marine ballast water viromes to investigate global transport of viruses and to examine the effects of engineered, management, and environmental parameters associated with ballast water on ocean viromes.

CHAPTER 4

MATERIALS AND METHODS

4.1. Overview

Sample preparation is a critical step in viral metagenomic studies, as it has a significant impact on downstream data analysis. Yet, a general workflow from sample collection to metagenomic sequencing is experimentally challenging at each step and requires a systematic evaluation. In this regard, this chapter is organized into two sections: (1) an evaluation and optimization of sample preparation methods and (2) a description of materials and methods used in the present research for preparing and analyzing viral metagenomes (viromes). Establishing a workflow from sample collection to metagenomic sequencing with a systematic evaluation will begin to minimize any biases on the resulting metagenomic sequence data and provide a comprehensive view of viral communities.

4.2. Evaluation and optimization of sample preparation methods

4.2.1. Evaluation of the tangential flow filtration method using groundwater and surface water

4.2.1.1. Tangential flow filtration setup and procedure

The effectiveness of a low–cost tangential flow filtration (TFF) system using disposable hollow fiber ultrafilters was evaluated for concentrating three types of viruses from groundwater and surface water samples. Three, 20 L samples of groundwater were collected in Lansing, Michigan and three, 20 L samples of surface water from the Red Cedar River in East Lansing, Michigan. Phages MS2 and PhiX174 were chosen for the

TFF evaluation, as they are well–characterized surrogates for human enteric viruses and P22 based on its bigger particle size (60 nm) than MS2 (27 nm) and PhiX174 (30 nm). Each 20 L of water sample was seeded with phages MS2, PhiX174, and P22 at levels of 10^5 plaque–forming units (PFUs; approximately 5×10^3 PFU/L) and mixed for 30 min at room temperature.

The TFF system was set up as described previously with a few modifications (Hill et al., 2005; Figure 4.1). Briefly, a peristaltic pump (model 7554–90; Cole–Parmer Instrument Co., Vernon Hills, IL, USA) and a pump head (model 77800–52; Cole–Parmer Instrument Co., Vernon Hills, IL, USA) were used with L/S 36 and L/S 24 silicone tubing (Masterflex; Cole–Parmer Instrument Co., Vernon Hills, IL, USA). The seeded groundwater and surface water samples were passed through single–use Fresenius Optiflux F200NR (2.0 m^2 surface area, 30,000 Dalton molecular weight cutoff (MWCO); Fresenius Medical Care, Lexington, MA, USA) and Asahi Kasei REXEED 25S ultrafilters (2.5 m^2 surface area, 30,000 Dalton MWCO; Asahi Kasei Medical Co., Ltd., Tokyo, Japan), respectively. The ultrafilters were blocked immediately prior to filtration by recirculating 500 mL of sterile 0.01% NaPP solution through the ultrafilters for 15 min with the filtrate port closed. Filtration was performed at a filtration rate of approximately 1,000 mL/min until approximately 250 mL of concentrated sample remained in the TFF system. Elution was performed by the recirculation of 500 mL of sterile surfactant solution (0.001% Antifoam A, 0.01% NaPP, and 0.5% Tween 80) through the system for 5 min. The eluent was then added to the retentate and produced the final volume of approximately 500 mL. A procedure for virus concentration from the groundwater and surface water samples is depicted in Figure 4.2.

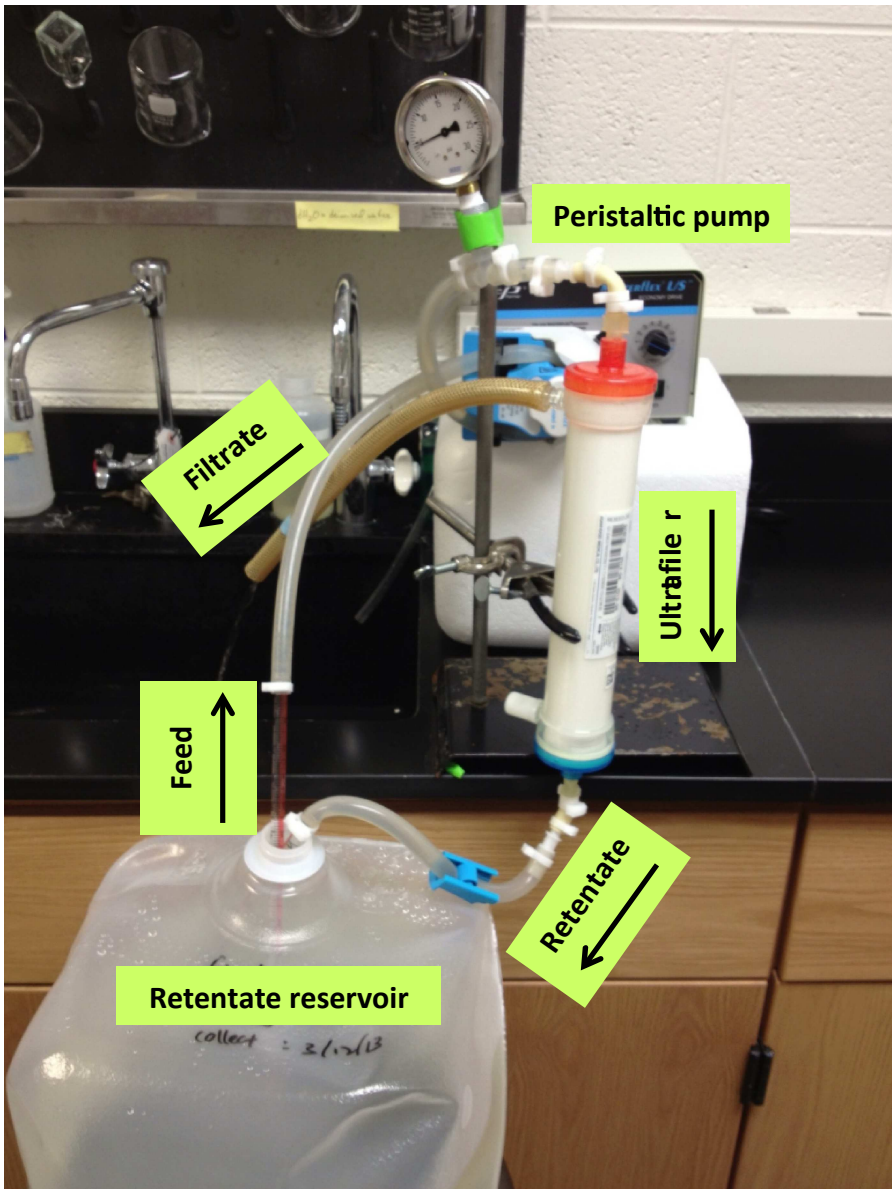


Figure 4.1. Experimental setup for the tangential flow filtration using ultrafilters with 30,000 Dalton molecular weight cutoff. The retentate is recirculated until the volume remained is less than 500 mL.

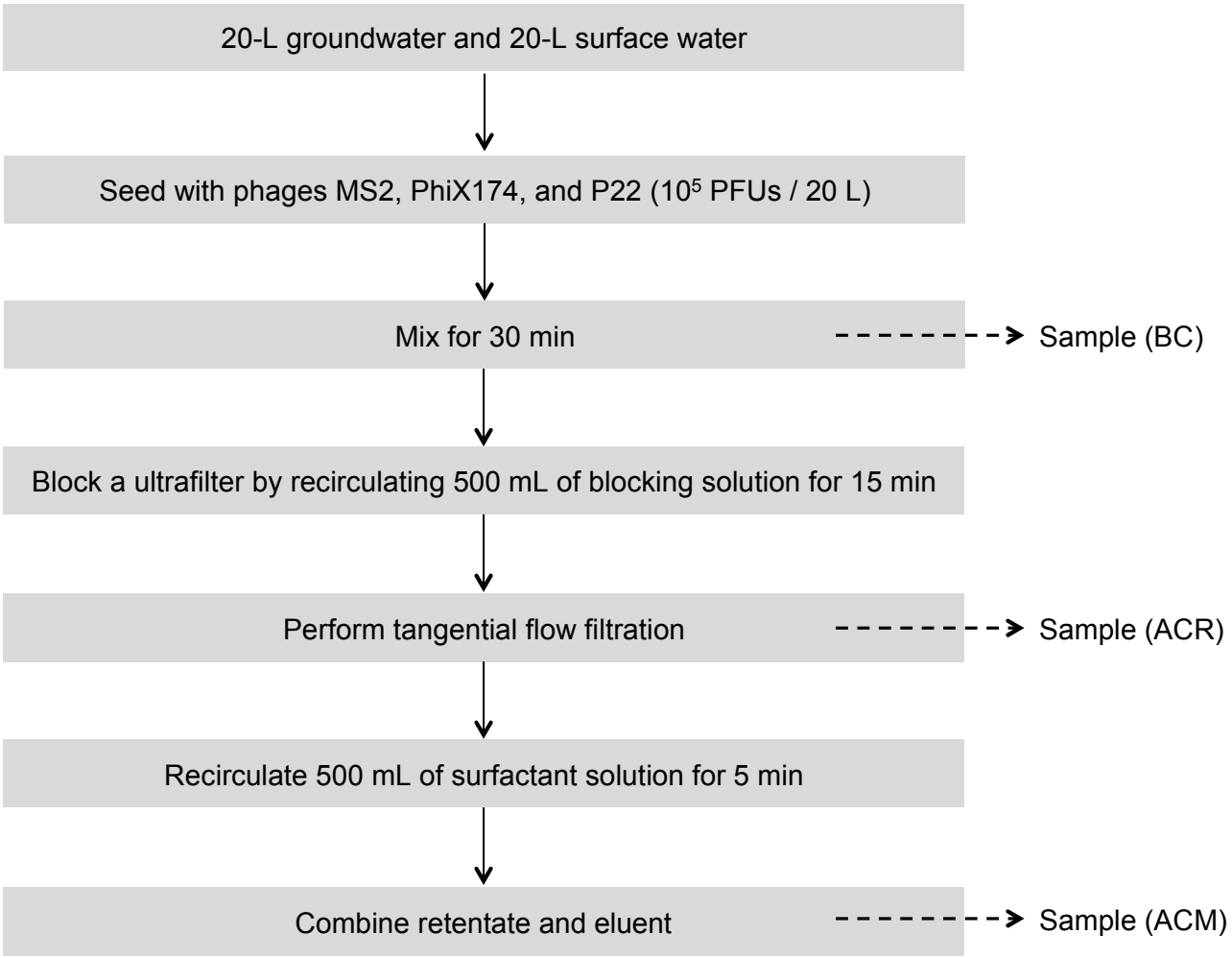


Figure 4.2. A procedure for virus concentration from the groundwater and surface water samples.

Abbreviation: PFUs, plaque–forming units.

4.2.1.2. Data analysis and statistics

Phages MS2, PhiX174, and P22 were enumerated using single agar overlay plaque assay (U.S. EPA method 1602) before and after filtration and with and without the elution procedure to calculate their recovery efficiencies. The number of PFU per mL in the phage–seeded groundwater and surface water samples was calculated based on the following equation (U.S. EPA method 1602).

$$PFU / mL = (PFU_1 + PFU_2 + \ldots + PFU_N) / (V_1 + V_2 + \ldots + V_N)$$

Where:

- PFU_N = Number of PFU from plates of all countable sample dilutions, excluding dilutions with too numerous to count (TNTC) or zeros
- V_N = Volume of undiluted sample in all plates with countable plaques
- N = Number of useable counts

Recovery efficiencies, expressed as percentages, were then calculated based on the following equations.

$$R_{NE} (\%) = 100 \times \{PFU_{ACR} / (PFU_{BC} \times CF_{NE})\}$$

$$R_{WE} (\%) = 100 \times \{PFU_{ACM} / (PFU_{BC} \times CF_{WE})\}$$

Where:

- R_{NE} = Recovery efficiency without the elution procedure

- R_{WE} = Recovery efficiency with the elution procedure
- PFU_{BC} = Number of PFU / mL before filtration
- PFU_{ACR} = Number of PFU / mL in the retentate after filtration
- PFU_{ACM} = Number of PFU / mL in the mixture of retentate and eluent after filtration
- CF_{NE} = Concentration factor without the elution procedure (e.g., starting volume / retentate volume)
- CF_{WE} = Concentration factor with the elution procedure (e.g., starting volume / retentate and eluent volume)

The recovery efficiencies between before and after the elution procedure were compared using the Student's t–test. To determine whether TFF recovery efficiency varied by phages, a one–way fixed effects analysis of variance (ANOVA) was used. The least significant difference (LSD) multiple comparison tests was then used to perform a pairwise comparison between mean recovery efficiency of three phages. Statistical analysis was performed using agricolae package (de Mendiburu, 2013) in the R environment (R Core Team, 2013).

4.2.1.3. Virus recovery by tangential flow filtration

The TFF system in this research was found to be capable of achieving high recovery efficiencies for three different viruses in 20–L groundwater and 20–L surface water samples. The average recovery efficiencies without filter elution for MS2, PhiX174, and P22 were 74.9%, 75.9%, and 77.8%, respectively from 20–L groundwater

samples and 85.5%, 81.0%, and 58.7%, respectively from 20–L surface water samples (Table 4.1). The use of the elution procedure was found to significantly increase average recovery efficiencies of MS2 in groundwater but not in surface water samples ($p = 0.05$). The elution procedure exhibited no significant increase in the average recovery of phages PhiX174 and P22 in both types of water samples ($p = 0.05$). A significant difference in recovery efficiencies between phages was found only in groundwater with the elution procedure, where phage MS2 had a significantly higher recovery than P22 ($p = 0.01$). However, the recovery of PhiX174 showed no significant difference with that of MS2 or P22. It is assumed that P22 might have remained in the ultrafilter membranes even after the elution procedure. The bigger particle size of P22 (60 nm) than MS2 (27 nm) is not believed as a main reason of its lower recovery as TFF is based on size exclusion method and any particles having bigger MWCOs of ultrafilters (30,000 Dalton in this research) remained in the system. Further study is needed to evaluate the effect of different viral morphology on recovery performance of the ultrafilters.

Table 4.1. Percent recovery efficiencies (average ± standard deviation) of phages in 20–L groundwater and 20–L surface water samples

Groundwater using Fresenius F200NR ultrafilter

Phage	N	No elution	With elution
MS2	3	74.9 ± 5.0	109.4 ± 2.4
PhiX174	3	75.9 ± 2.6	94.4 ± 13.2
P22	3	77.8 ± 5.1	76.8 ± 12.6

Surface water using Asahi Kasei REXEED 25S ultrafilter

Phage	N	No elution	With elution
MS2	3	85.5 ± 13.0	106.4 ± 26.6
PhiX174	3	81.0 ± 25.3	92.8 ± 26.5
P22	3	58.7 ± 6.4	66.6 ± 13.6

The recoveries of MS2 using the Fresenius Optiflux F200NR and the Asahi Kasei REXEED 25S ultrafilters in this research were comparable to previous studies using 10 L (Hill et al., 2005) and 100 L (Mull and Hill, 2012) of water samples, respectively. A direct comparison of recovery efficiencies between two ultrafilters was not possible as different types of water samples were used for each ultrafilter. However, Asahi Kasei ultrafilter (2.5 m^2 surface area) was chosen for concentrating viral particles for this research as it has a larger surface area and cheaper than the Fresenius Optiflux F200NR ultrafilter (2.0 m^2 surface area).

4.2.2. Evaluation of nucleic acid extraction methods for simultaneous recovery of viral DNA and RNA

4.2.2.1. Standard curve generation for quantitative PCR assay

Four commercially available viral nucleic acid extraction kits were compared via MS2– and PhiX174–specific quantitative PCR (qPCR) using LightCycler 480 Instrument (Roche Diagnostics, Indianapolis, IN, USA) for their ability to extract and purify MS2 RNA and PhiX174 DNA. To generate standard curves for qPCR, approximately 10^9 PFUs of MS2 and PhiX174 were prepared. MS2 RNA and PhiX174 DNA were then extracted using QIAamp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA). Reverse transcription (RT) was performed on 5 µL of MS2 nucleic acid extracts, using 20 µM of random hexamers and GoScript™ Reverse Transcriptase (Promega, Madison, WI, USA) under the following conditions: 70 °C for 5 min, 25 °C for 5 min, 42 °C for 60 min, and 70 °C for 15 min. qPCR was then performed on MS2 reverse transcribed RNA (cDNA)

and PhiX174 DNA following published protocols optimized for the detection of MS2 (O'Connell et al., 2006) and PhiX174 (Verreault et al., 2010) with a few modifications. The qPCR reaction mixture for MS2 contained 10 µL of LightCycler 480 Probe Master mix (Roche Diagnostics, Indianapolis, IN, USA), 8 µL of water, 0.8 µL of each primer (0.4 µM as final concentration), 0.4 µL of probe (0.2 µM as final concentration), and 5 µL of cDNA in a 25 µL of final volume. The qPCR reaction mixture for PhiX174 contained 10 µL of LightCycler 480 Probe Master mix (Roche Diagnostics, Indianapolis, IN, USA), 7.6 µL of water, 1 µL of each primer (0.5 µM as final concentration), 0.4 µL of probe (0.2 µM as final concentration), and 5 µL of DNA in a 25 µL of final volume. Sequence of the primers and probes and PCR protocol used for qPCR are summarized in Table 4.2. The resulting standard curves (slope = – 3.637, Y intercept = 50.31, error = 0.0118, and efficiency = 1.884 for MS2, Figure 4.3A; slope = – 3.513, Y intercept = 40.49, error = 0.00802, and efficiency = 1.926 for PhiX174, Figure 4.3B) were used for the quantification of MS2 cDNA and PhiX174 DNA.

Table 4.2. Sequence of the primers and probes and PCR protocol used for quantitative PCR assay

Phage	Primer and probe	Position	Sequence (5'–3')	PCR protocol	Reference
MS2	Forward primer	632–648	GTCGCGGTAATTGGCGC	50 cycles	O'Connell et al., 2006
	Reverse primer	690–708	GGCCACGTGTTTTGATCGA	(95 °C, 10 sec; 55 °C,	
	Probe	650–671	AGGCGCTCCGCTACCTTGCCCT	1 min; 72 °C, 1 sec)	
PhiX174	Forward primer	508–531	ACAAAGTTTGGATTGCTACTGACC	40 cycles	Verreault et al., 2010
	Reverse primer	609–630	CGGCAGCAATAAACTCAACAGG	(95 °C, 10 sec; 55 °C,	
	Probe	533–556	CTCTCGTGCTCGTCGCTGCGTTGA	1 min; 72 °C, 1 sec)	

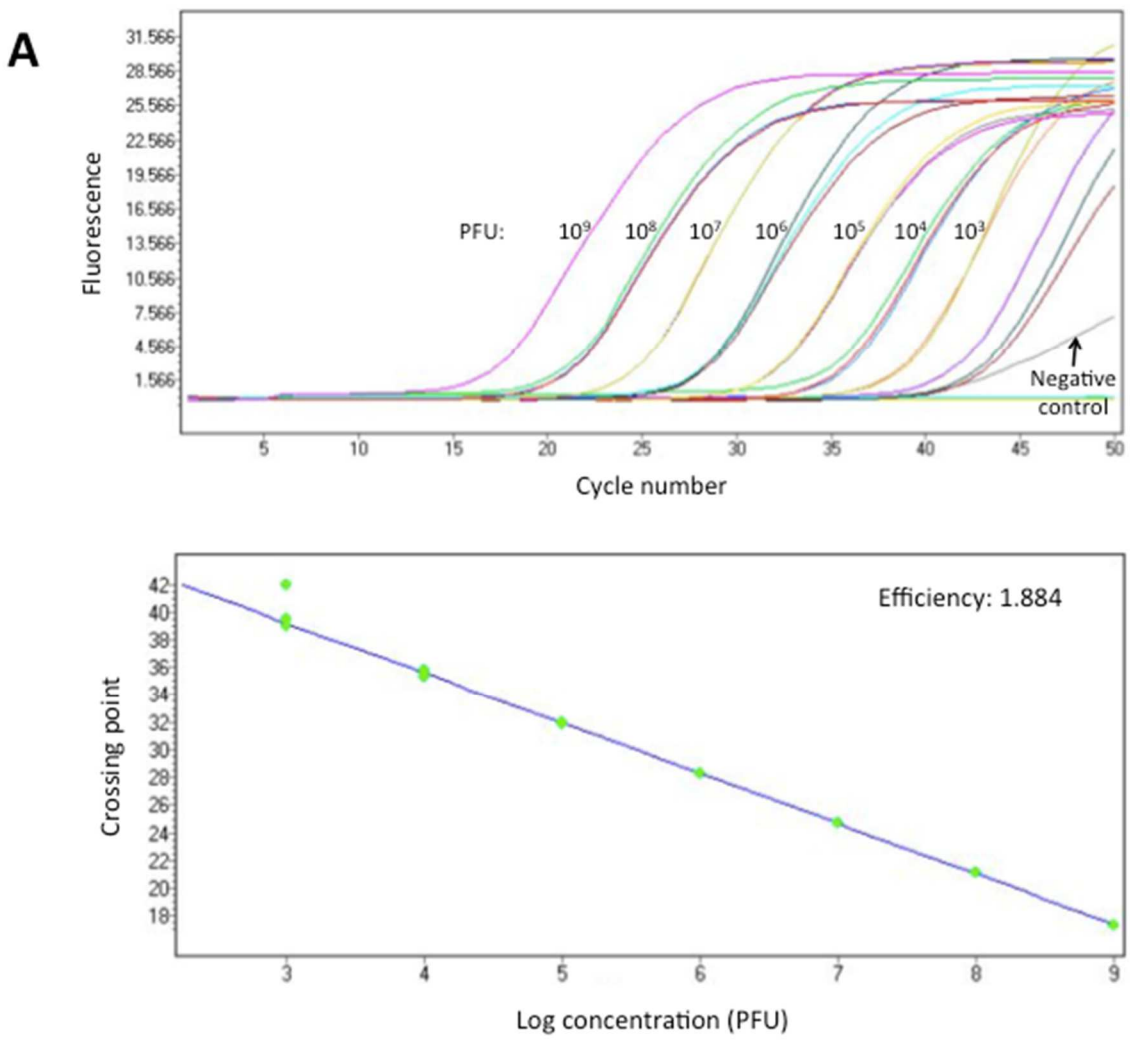
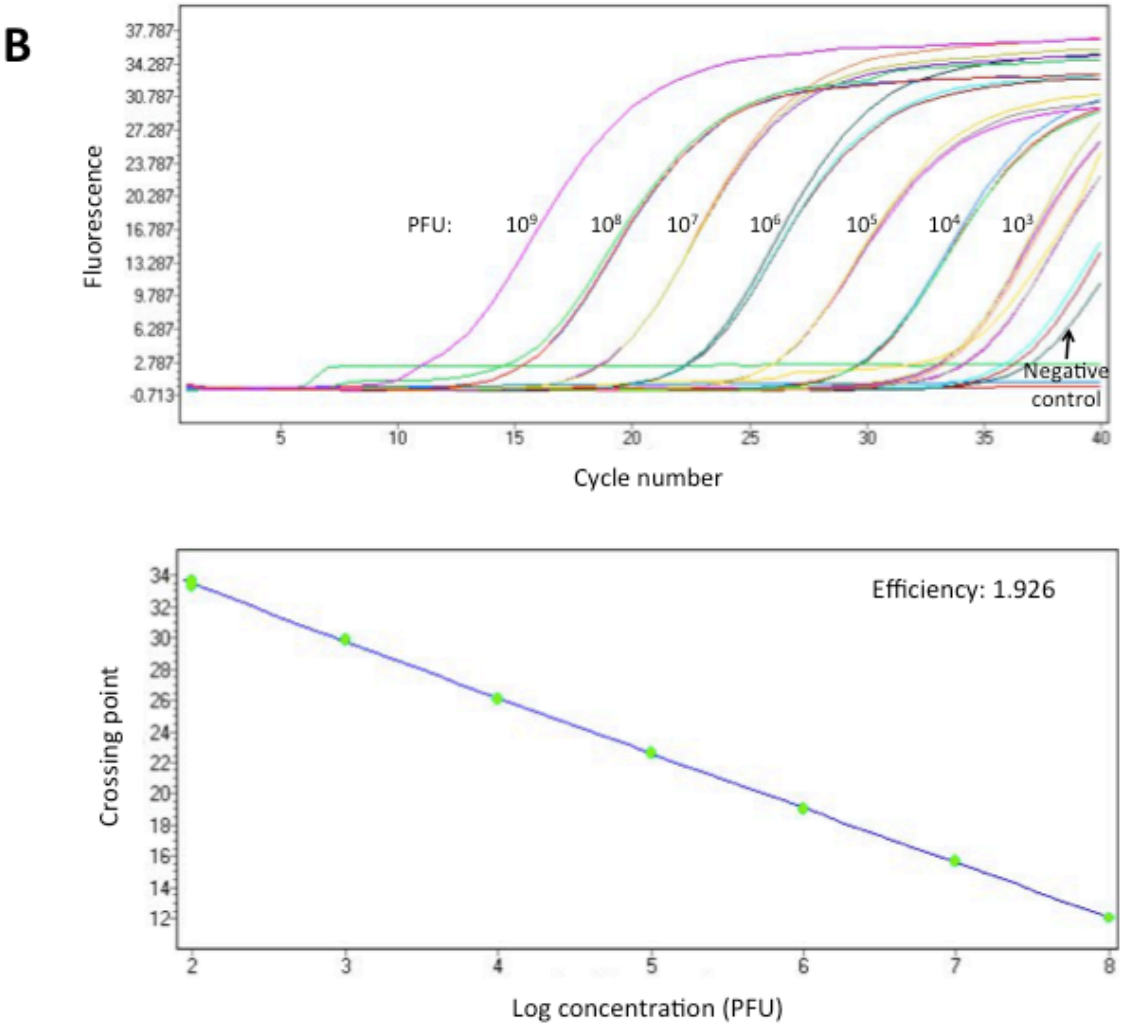


Figure 4.3. Standard curves for the quantification of MS2 cDNA (A) and PhiX174 DNA (B) using the LightCycler Instrument and the LightCycler 480 Probe Master mix.

Abbreviation: PFU, plaque–forming unit.

Figure 4.3 (cont'd)

B



4.2.2.2. Viral nucleic acid extraction and quantitative PCR detection

After qPCR standard curve generation, two sets of sterile phosphate buffered saline (PBS, 4.3 mM Na_2HPO_4, 1.4 mM KH_2HPO_4, 137 mM NaCl, and 2.7 mM KCl in nanopure water, pH 7.2) were seeded with phages MS2 and PhiX174 at two different levels, 10^5 and 10^6 PFUs per mL of PBS. The starting sample volumes for viral DNA and RNA extraction were increased from the recommended starting volume, 140 µL up to 1,000 µL (reagents needed for the extraction were increased proportionally with increased starting sample volume). 1,000 µL of PBS seeded with MS2 and PhiX174 were used for viral nucleic acid extraction with QIAamp MinElute Virus Spin Kit, QIAamp UltraSens Virus Kit, and RNeasy Plus Mini Kit, while 560 µL of seeded PBS was used with QIAamp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA). 1,000 µL of seeded PBS was not applied to the QIAamp Viral RNA Mini Kit because only up to 560 µL of starting volume could be processed for this kit according to the manufacturer's instructions. MS2 nucleic acid extracts were reverse transcribed and the extraction of MS2 RNA and PhiX174 DNA was detected via qPCR using previously described conditions.

4.2.2.3. Data analysis and statistics

The qPCR result was expressed as cycle threshold (C_T) value, which is defined as the number of cycles required for the fluorescent signal to cross the threshold. The statistical significance of difference in C_T values between the extraction kits was evaluated by one–way fixed effects ANOVA and post hoc testing with Bonferroni LSD multiple comparison tests. A p–value < 0.05 was considered statistically significant.

Statistical analysis was performed using agricolae package (de Mendiburu, 2013) in the R environment (R Core Team, 2013).

4.2.2.4. Comparison of viral nucleic acid extraction methods

In order to compare the ability of commercial kits to extract both viral DNA and RNA and to handle increased starting sample volume, PBS solutions were seeded with known concentrations of phages MS2 and PhiX174 and evaluated by qPCR. For both phages MS2 and PhiX174 at both seeding concentrations of 10^5 and 10^6 PFUs per mL, the use of QIAamp MinElute Virus Spin Kit yielded the lowest C_T values for qPCR with significant difference among the kits used ($p < 0.05$; Figure 4.4). This suggested that QIAamp MinElute Virus Spin Kit was best suited for simultaneously extracting viral DNA and RNA because C_T levels are inversely proportional to the amount of target nucleic acid in a sample. On the other hand, the use of RNeasy Plus Mini Kit yielded the highest C_T values for both phages MS2 and PhiX174 at both seeding concentrations, suggesting its low efficiency in recovering viral nucleic acids. C_T values generated from the use of QIAamp Viral RNA Mini Kit were also included in the efficiency comparison. However, it should be noted that lower starting sample volume (560 μL) was used for extracting viral nucleic acids with QIAamp Viral RNA Mini Kit.

Based on the lowest C_T value among the kits evaluated, the QIAamp MinElute Virus Spin Kit was used for simultaneously extracting viral DNA and RNA for this research.

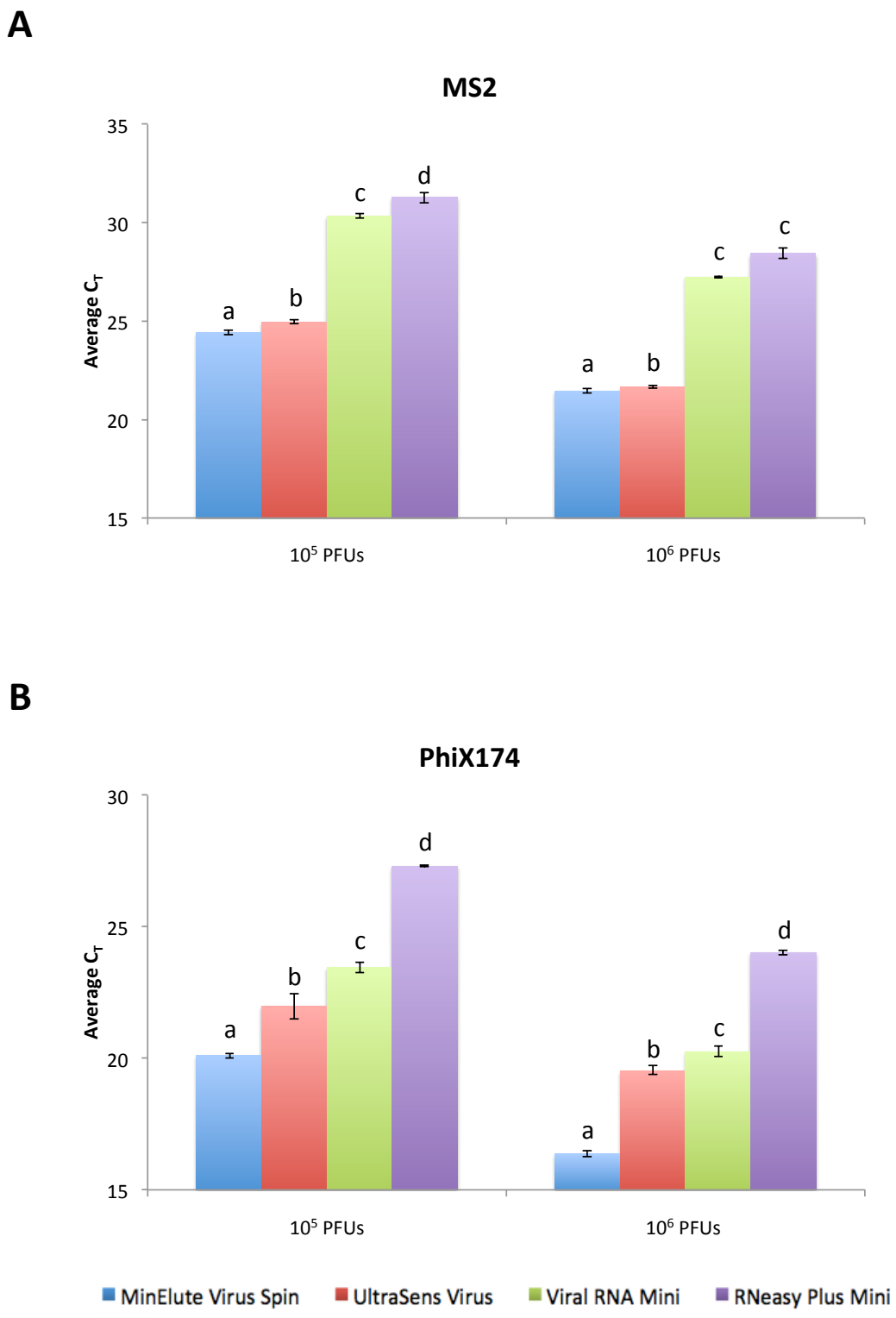


Figure 4.4. Comparison of qPCR C_T value for viral nucleic acid extracts from seeded water with 10^5 and 10^6 PFUs/mL of MS2 (A) and PhiX174 (B). The error bars represent

Figure 4.4 (cont'd)

standard deviation of the C_T value calculated from six replicates. Different letters in the figure indicate significant differences according to Bonferroni LSD multiple comparison tests ($p < 0.05$).

Abbreviations: qPCR, quantitative PCR; C_T, cycle threshold; PFUs, plaque–forming units; LSD; least significant difference.

4.3. Description of materials and methods used for preparing and analyzing viromes

Generally, a viral metagenomic study is composed of three main steps: (1) virome preparation, (2) high–throughput sequencing, and (3) bioinformatics analyses. A pipeline for the viral metagenomic study with high–throughput sequencing used in this research is depicted in Figure 4.5. Detailed description of each of these steps was provided below.

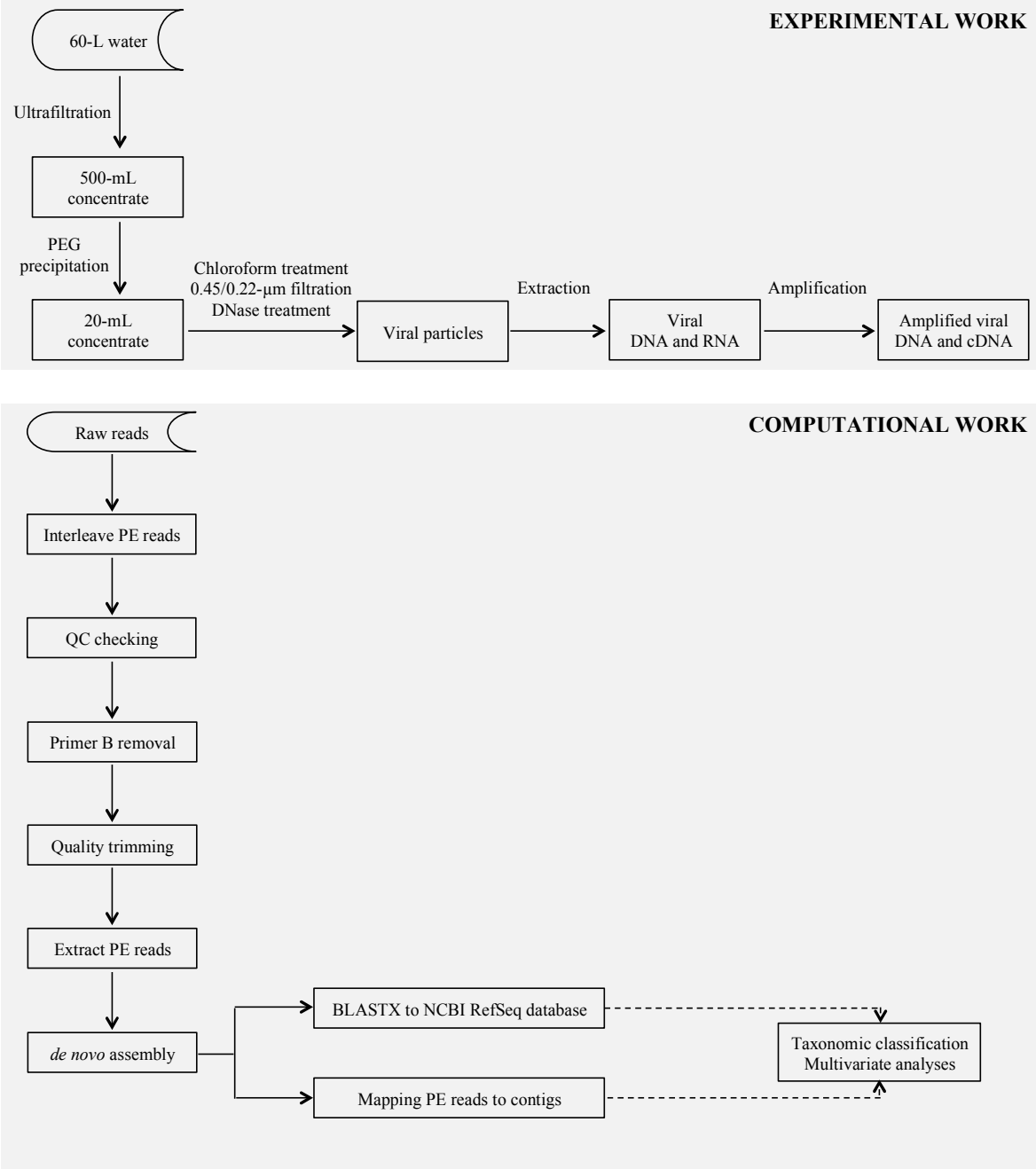


Figure 4.5. A pipeline for the viral metagenomic study with high–throughput sequencing used in this research.

Abbreviation: PE, paired–end.

4.3.1. Virome preparation

4.3.1.1. Sample collection

Ballast and harbor water samples were collected from the Port of Duluth–Superior, the Port of Los Angeles/Long Beach (LA/LB), and the Port of Singapore from May 2013 to May 2014. Description about identification and collection information of the ballast and harbor waters from the Port of Duluth–Superior and the Port of LA/LB and the Port of Singapore was provided in Chapter 5 and Chapter 6, respectively.

4.3.1.2. Variable measurement and estimation

Environmental parameters, including pH, salinity, and temperature of ballast and harbor waters were measured on site using a hand−held meter (model 63; Yellow Springs Instruments, Yellow Springs, OH, USA) and turbidity using a portable meter (model 2020we; LaMotte Company, Chestertown, MD, USA).

Ballast water storage duration was calculated based on the difference in days the ballast water was held in the tanks before sample collection. Surface harbor waters were considered to have storage duration of zero−day. Location of ballast water exchange of vessels who conducted ballast water management practice was used as geographic origin of ballast water. Otherwise, location of last port of vessels carrying unexchanged ballast water was used as geographic origin of ballast water. Coordinates of ballast water exchange location were retrieved from ballast water reporting form under the permission of captains of vessels. Distance in nautical miles (1 nautical mile = 1.852 kilometers) between where ballast water exchange took place and nearest shoreline was calculated

using a data set (http://oceancolor.gsfc.nasa.gov/DOCS/DistFromCoast/) generated by National Aeronautics and Space Administration Ocean Color Group.

4.3.1.3. Primary concentration of viral particles

Viral particles in the ballast and harbor waters were concentrated using TFF with Asahi Kasei ultrafilter within 24 hours of sample collection using the previously described procedure.

4.3.1.4. Secondary concentration of viral particles

As the volume of the TFF concentrates (approximately 500 mL) is still large for direct viral nucleic acid extraction, viral particles were further concentrated by polyethylene glycol (PEG) precipitation as described previously with a few modifications (Jaykus et al., 1996). The TFF concentrates were adjusted to pH 7.2 and supplemented with 0.3 M NaCl and 10% (weight/volume) PEG 8000 (Promega, Madison, WI, USA). The mixture was incubated for 18 hours at 4 °C and centrifuged at 11,300 × g for 30 min at 4 °C. The precipitated viral particles were dissolved in 20 mL of sterile PBS (pH 7.2).

4.3.1.5. Purification of viral particles

Prior to viral nucleic acid extraction, viral particles were purified with a combination of methods to remove free DNA and any cellular materials. An equal volume of chloroform was added to the PEG concentrates (20 mL) and mixed by vortexing for 30 sec to disrupt cell membranes. The mixture was centrifuged at 1,600 × g for 30 min at 4 °C and the aqueous layer was recovered. The chloroform–purified viral

particles were further purified using a series of 0.45 μm and 0.22 μm sterile syringe filtrations (EMD Millipore Corp., Billerica, MA USA) to further remove any remaining cellular microorganisms. A one mL aliquot of the 0.22 μm filtrates was incubated with 100 U of DNase I (Invitrogen, Carlsbad, CA, USA) at room temperature for 2 hours. DNase I was inactivated by adding 1 μL of 25 mM EDTA (Sigma–Aldrich, St. Louis, MO, USA) and incubating at 65 °C for 15 min.

4.3.1.6. Viral nucleic acid extraction

After the final concentration and purification of viral particles from the ballast and harbor waters, the viral DNA and RNA were simultaneously extracted using the QIAamp MinElute Virus Spin Kit (Qiagen, Valencia, CA, USA). To confirm the absence of microbial contamination, extracted viral nucleic acids were screened by 16S rDNA PCR with 27F/1492R universal primers (Lane, 1991). Samples were again passed through a 0.22 μm filter and treated with DNase I if microbial contamination was detected.

4.3.1.7. Random transcription/amplification of viral nucleic acid extracts

To obtain a sufficient quantity of DNA and cDNA for metagenomic sequencing, the extracted viral nucleic acids were reverse transcribed and amplified as previously described (Wang et al, 2002; Wang et al., 2003).

Briefly, the extracted viral nucleic acids were incubated at 65 °C for 5 min with Primer A (5'–GTTTCCCAGTCACGATCNNNNNNNNN–3') followed by cooling at room temperature for 5 min to encourage primer annealing and inactivate any native RNases. The reaction mixture contained 1 μL of Primer A (40 pmol/μL), 5 μL of water,

and 4 µL of extracted viral nucleic acids in a 10 µL of final volume. Primer A contained a 17–nucleotide specific sequence followed by nine random nucleotides. The degenerate N nucleotides of the primer form a random priming site and anneal to the viral RNA, while the remaining primer bases create an artificial primer site for PCR amplification. First strand of cDNA was synthesized using SuperScript III Reverse transcriptase (Invitrogen, Carlsbad, CA, USA). A 10 µL of RT mixture, containing 4 µL of RT buffer (5×), 1 µL of dNTP (10 mM), 1 µL of RNAseOUT, 0.5 µL of water, 1.5 µL of DTT (0.1 M), and 2 µL of SSIII RT was added to the viral nucleic acid and incubated at 42 °C for 60 min. Second strand synthesis was performed using Sequenase Version 2.0 DNA Polymerase (USB/Affymetrix, Cleveland, OH, USA) under the following condition: 94 °C for 2 min, 10 °C for 5 min, 37 °C for 8 min, 94 °C for 2 min, 10 °C for 5 min, 37 °C for 8 min, 94 °C for 8 min, and 10 °C final holding temperature. The reaction mixture, containing 2 µL of Sequenase buffer (5×), 0.3 µL of Sequenase, and 7.7 µL of water was added to the samples after they were heated at 94 °C for 2 min for the first time and cooled down to 10 °C. Additional 1.2 µL of Sequenase (1:4 diluted) was added to the samples after samples were heated at 94 °C for 2 min for the second time and cooled down to 10 °C.

Primer B (5'–GTTTCCCAGTCACGATC–3'), complementary to the 17–nucleotide sequence of the Primer A, was used to amplify the templates previously generated. Three PCR reactions were performed for each sample using HotStarTaq Master Mix Kit (Qiagen, Valencia, CA, USA) under the following condition: 40 cycles of 94 °C for 30 sec, 40 °C for 30 sec, 50 °C for 30 sec, and 72 °C for 60 sec. The PCR mixture contained 10 µL of PCR buffer (10×), 2 µL of dNTP (10 mM), 1 µL of Primer B (100 pmol/µL), 1 µL of Taq DNA polymerase, 80 µL of water, and 6 µL of previously

generated PCR product in a 100 µL of final volume. A positive control was excluded to avoid cross–contamination while a negative control was included in every random transcription/amplification run. Absence of contaminating DNA and presence of amplified products were confirmed on a 1% agarose gel (between 500 bp to 1 Kbp). The resulting PCR products were combined and purified using a Wizard® SV Gel and PCR Clean–Up System (Promega, Madison, WI, USA).

4.3.2. High−throughput sequencing

The sequencing libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit (Illumina, San Diego, CA, USA) with a few modifications at the Research Technology Support Facility (RTSF) at Michigan State University (MSU). The resulting libraries (200−base pair (bp) insert + 120−bp adapters) were loaded on Illumina HiSeq 2500 Rapid Run flow cells and sequencing was performed in a 2 × 100 bp paired−end (PE) format.

4.3.3. Bioinformatics analyses

Bioinformatics analyses of the sequencing data sets obtained from the ballast and harbor waters were performed using high performance computing resources provided by MSU.

4.3.3.1. Preprocessing and quality control of raw sequence reads

FastQC was used to check the quality of the data sets (Andrews, 2010). Each virome was screened for the 17–bp 'Primer B' sequence (5'–

GTTTCCCAGTCACGATC–3') used for random amplification and any reads homologous to the 'Primer B' sequence at their 5' ends were removed (allowing up to 3 mismatches per read) using HOMER version 4.4 (Heinz et al., 2010). Reads were then filtered requiring that 50% of the bases must have a Phred quality score of 30 or higher (parameters –q 30, –p 50) using FASTX–Toolkit version 0.0.13 (Goecks et al., 2010). Following trimming and filtering of raw reads, reads shorter than 30 bp or containing 'N's were excluded prior to further analyses.

4.3.3.2. *De novo* assembly of sequence reads

To generate contiguous reads (contigs), velvet (Zerbino and Birney, 2008) was previously used to conduct a *de novo* assembly of the cleaned sequence reads using default settings. Despite a wide range of k values, a low number of contigs with small size was generated. Thus, PE reads from each virome were *de novo* assembled using IDBA–UD, which handles data set with highly uneven sequencing depth and generates longer contigs with higher accuracy (Peng et al., 2012).

4.3.3.3. Taxonomic classification

Assembled contigs were then blasted against the National Center for Biotechnology Information (NCBI) Viral Reference Sequence (RefSeq) database (downloaded from ftp://ftp.ncbi.nih.gov/refseq/release/viral in September 2014) for taxonomic assignment using BLASTX with an E–value cutoff of 10^{-5} (Altschul et al., 1990). The BLASTX output was parsed using the MEtaGenome Analyzer (MEGAN) version 5.6.6 (Min Score = 50.0, Max Expected = $1.0E^{-5}$, Top Percent = 10.0, Min

Support Percent = 0.1, Min Support = 1, and LCA Percent = 100.0 for the freshwater virome data set; Min Score = 50.0, Max Expected = $1.0E^{-5}$, Top Percent = 10.0, Min Support Percent = 0.0, Min Support = 1, and LCA Percent = 100.0 for the marine water virome data set) (Huson et al., 2007). Contigs that were assigned to viral taxa but did not meet the selected parameters were placed under 'Not assigned' and contigs that did not have any hits to known sequences in the databases were placed under 'No hits'.

To assess relative abundance of a phylogenetic group in the viromes, read mapping to contigs was performed using default settings in Bowtie 2 version 2.1.0 (Langmead and Salzberg, 2012). To determine a relative abundance for each contig, the number of reads aligned to a contig divided by the contig length (kbp) was calculated. The relative abundances of each phylogenetic group in the viromes were calculated based on the following equation where the abundance for each contig classified in a particular group was summed. Its percentage was then used to compare a particular group in a virome to the rest of the viromes.

$$Ri = \sum (Ni/Li)$$

Where:

- Ri = Relative abundance of a phylogenetic group i
- Ni = Number of reads aligned to a contig in a phylogenetic group i
- Li = Length (kbp) of a contig in a phylogenetic group i

Identification of viral pathogens, annotation–independent virome comparison, and

multivariate analyses of viromes are described in Chapter 5 and Chapter 6.

REFERENCES

REFERENCES

1. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D., Basic Local Alignment Search Tool. *Journal of Molecular Biology* **1990,** *215* (3), 403-410.

2. Andrews S. FastQC: a quality control tool for high throughput sequence data. **2010,** Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

3. de Mendiburu, F., Statistical procedures for agricultural research: package 'agricolae' in R, **2013,** Available at http://CRAN.R-project.org/package=agricolae.

4. Goecks, J.; Nekrutenko, A.; Taylor, J.; Team, G., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **2010,** *11* (8).

5. Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y.; Laslo, P.; Cheng, J.; Murre, C.; Singh, H.; Glass, C., Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **2010,** *38* (4), 576-589.

6. Hill, V.; Polaczyk, A.; Hahn, D.; Narayanan, J.; Cromeans, T.; Roberts, J.; Amburgey, J., Development of a rapid method for simultaneous recovery of diverse microbes in drinking water by ultrafiltration with sodium polyphosphate and surfactants. *Applied and Environmental Microbiology* **2005,** *71* (11), 6878-6884.

7. Huson, D.; Auch, A.; Qi, J.; Schuster, S., MEGAN analysis of metagenomic data. *Genome Research* **2007,** *17* (3), 377-386.

8. Jaykus, L.; DeLeon, R.; Sobsey, M., A virion concentration method for detection of human enteric viruses in oysters by PCR and oligoprobe hybridization. *Applied and Environmental Microbiology* **1996,** *62* (6), 2074-2080.

9. Lane J, 16S/23S rRNA sequencing. In E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics, John Wiley and Sons Ltd., New York, NY, **1991**.

10. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012,** *9* (4), 357-U54.

11. Mull, B.; Hill, V., Recovery of diverse microbes in high turbidity surface water samples using dead-end ultrafiltration. *Journal of Microbiological Methods* **2012,** *91* (3), 429-433.

12. O'Connell, K.; Bucher, J.; Anderson, P.; Cao, C.; Khan, A.; Gostomski, M.; Valdes, J., Real-time fluorogenic reverse transcription-PCR assays for detection of

bacteriophage MS2. *Applied and Environmental Microbiology* **2006,** *72* (1), 478-483.

13. Peng, Y.; Leung, H.; Yiu, S.; Chin, F., IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012,** *28* (11), 1420-1428.

14. R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing Vienna, Austria, **2010**.

15. U.S. Environmental Protection Agency, Method 1602: Detection of Male-specific (F+) and Somatic Coliphage in Water by Single Agar Layer (SAL) Procedure, Washington, DC, **2001**.

16. Verreault, D.; Rousseau, G.; Gendron, L.; Masse, D.; Moineau, S.; Duchaine, C., Comparison of Polycarbonate and Polytetrafluoroethylene Filters for Sampling of Airborne Bacteriophages. *Aerosol Science and Technology* **2010,** *44* (3), 197-201.

17. Wang, D.; Coscoy, L.; Zylberberg, M.; Avila, P.; Boushey, H.; Ganem, D.; DeRisi, J., Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99* (24), 15687-15692.

18. Wang, D.; Urisman, A.; Liu, Y.; Springer, M.; Ksiazek, T.; Erdman, D.; Mardis, E.; Hickenbotham, M.; Magrini, V.; Eldred, J.; Latreille, J.; Wilson, R.; Ganem, D.; DeRisi, J., Viral discovery and sequence recovery using DNA microarrays. *Plos Biology* **2003,** *1* (2), 257-260.

19. Zerbino, D.; Birney, E., Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **2008,** *18* (5), 821-829.

CHAPTER 5

METAGENOMIC INVESTIGATION OF BALLAST WATER VIRAL COMMUNITIES IN THE GREAT LAKES

This chapter has been published in *Kim Y., Aw T.G., Teal T.K., and Rose J.B. 2015. Metagenomic investigation of viral communities in ballast water. J. Environ. Sci. Technol. 49 (14), 8396–8407.*

Abstract

Ballast water is one of the most important vectors for the transport of non–native species to new aquatic environments. Due to the development of new ballast water quality standards for viruses, this research aimed to determine the taxonomic diversity and composition of viral communities (viromes) in ballast and harbor waters using metagenomics approaches. Ballast waters from different sources within the North America Great Lakes and paired harbor waters were collected around the Port of Duluth–Superior. Bioinformatics analysis of over 550 million sequences showed that a majority of the viral sequences could not be assigned to any taxa associated with reference sequences, indicating the lack of knowledge on viruses in ballast and harbor waters. However, the assigned viruses were dominated by double–stranded DNA phages, and sequences associated with potentially emerging viral pathogens of fish and shrimp were detected with low amino acid similarity in both ballast and harbor waters. Annotation–independent comparisons showed that viromes were distinct among the Great Lakes, and the Great Lakes viromes were closely related to viromes of other cold natural freshwater systems but distant from viromes of marine and human designed/managed freshwater systems. These results represent the most detailed characterization to date of viruses in ballast water, demonstrating their diversity and the potential significance of the ship– mediated spread of viruses.

5.1. Introduction

Ballast water has been used as an essential component of efficient and safe shipping operations dating back to the 1870s (National Research Council, 1996). Globally, as high as 12 billion tons of ballast water are transported and exchanged by more than 45,000 ocean–going vessels each year (Fredricks, 2002). In the United States (U.S.) alone, an estimated 79 million tons of ballast water are annually discharged into coastal areas from international ports (Carlton et al., 1995). A significant volume, 6.6 million tons of ballast water per year, is also discharged into the freshwaters of the North America Great Lakes (Clark, 2009). It is known that biological materials are discharged and exchanged with ballast water, and therefore this global and widely used practice brings with it potential ecological, economic, and public health problems including invasive species and the disruption of native ecosystems in major ports worldwide (Tsolaki and Diamadopoulos, 2010; Ferrate Treatment Technologies, 2015).

The U.S. Environmental Protection Agency (U.S. EPA) has established ballast water discharge standards that align with the International Maritime Organization (IMO) and U.S. Coast Guard (U.S.C.G.) rules based on organism size classes (David and Gollasch, 2015). At the state level, California has the most stringent ballast water management criteria with a state–specific standard of zero detectable living organisms for all organism size classes including virus–like particles (VLPs) in the final discharge going into effect January 2020 (ABS, 2014). Efforts to comply with increasingly stringent regulatory demands will, however, be one of the most significant challenges for the shipping industry over the next few years as technologies for ballast water treatment are still in the research and development phase (Tsolaki and Diamadopoulos, 2010).

A few studies have shown that bacteria and VLPs numerically dominate ballast water biota and are transferred globally in greater numbers than any other size classes of organisms (Ruiz et al., 2000; Drake et al., 2001; Drake et al., 2007; Sun et al., 2010; Seiden et al., 2011). For example, the ballast water of vessels arriving at Chesapeake Bay on the U.S. East Coast contained an average of 8.3×10^8 bacteria/L and 7.4×10^9 VPLs/L (Ruiz et al., 2000). An interesting follow–up study showed that an estimated 3.9×10^{18} bacteria and 6.8×10^{19} VLPs (assuming a survival rate of 56% and applying estimates of ship traffic) in ballast water were annually discharged to and survived in the lower Chesapeake Bay (Drake et al., 2007).

Viruses are small infectious agents that exist through parasitic relationships with a wide range of hosts, including humans, animals, bacteria, fungi, and plants – some of which are highly host specific. Viruses are of special interest because they are thought to be the most abundant and diverse biological entities on Earth with as many as 10^{10} VLPs/L of seawater, approximately 10 times more than the number of bacteria (Bergh et al., 1989; Mokili et al., 2012). Moreover, viruses influence the structure and diversity of microbial communities by infection and lysis of host communities (Wommack and Colwell, 2000). Examples of specific viruses that have been identified as invasive species introduced via ships' ballast water are marine cyanophage and Viral Hemorrhagic Septicemia Virus (VHSV) in the Great Lakes and Infectious Salmon Anemia Virus (ISAV) in Chile, Europe, and Northwest Atlantic (Elsayed et al., 2006; Wilhelm et al., 2006; European Food Safety Authority, 2007; Lumsden et al., 2007).

While the introduction of a complex assemblage of microorganisms through ballast water is a growing concern globally, the microbial diversity of ballast water

remains largely unknown. Moreover, the potential ecological impacts and public health risk are not well understood. The primary reasons for the lack of knowledge about the microbial communities, particularly viruses, in the ballast water system is related to the difficulty in collection of ballast water for virus analysis, the specificity of the viral–host systems used for identification, and the lack of universal genetic markers for viruses such as the 16S rRNA gene used for prokaryotes (Willner and Hugenholtz, 2013; Rohwer and Edwards, 2002). However, the development of modern genomic tools and the emergence of high–throughput sequencing (HTS) technologies have overcome some of the limitations of classical methods for virus detection and characterization. Increasing capacity in HTS and improvements in bioinformatics analyses have had a major impact on the expansion of virus detection and characterization and discovery of novel viruses, including zooplankton and phytoplankton viruses (Fischer et al., 2010; Nissimov et al., 2011).

These advances have allowed us to learn more about taxonomic diversity and composition of viral communities (viromes) in water. Our ability to characterize the ballast water virome of ships in the Great Lakes is of particular interest as this is an economically important shipping area that is also very susceptible to external influences on its native freshwater communities and therefore a good model system for the study of viral transport. The Great Lakes has a unique shipping system in which ships can move through the St. Lawrence Seaway linking North America with ports throughout the world. Consequently, this large freshwater basin is particularly vulnerable to invasive species and has been invaded by more than 180 non–native species within the past two centuries (Pagnucco et al., 2015). To the best of our knowledge, no studies have been published to

date on viromes in the Great Lakes ballast water. Therefore, the objectives of this research were (i) to investigate the composition and taxonomic diversity of viruses in ballast water collected around the Great Lakes and (ii) to understand the Great Lakes ballast water virome signatures by comparative virome analyses.

5.2. Materials and methods

5.2.1. Sample collection

Ballast waters were collected from five bulk carriers coming from different parts (Lakes Erie, Huron, Michigan, and Ontario) of the North America Great Lakes but all arriving in three terminals of the Port of Duluth–Superior over a one–week period on May 2013 (Figure 5.1, Table 5.1). The ballast water sampling in the Port of Duluth–Superior was approved by the Wisconsin Department of Natural Resources (WDNR) and by captains of vessels whose ballast waters were sampled. The sampling was conducted under the guidance of a ballast water inspector from the WDNR for safety purposes. Names of vessels and port terminals were designated as random letters as part of the sample confidentiality agreement. Ballast waters (60 L) were collected from one ballast tank per vessel either through a ballast water pipeline or sounding pipe. Surface harbor waters from different port terminals were also collected with a bucket near the vicinity of only three vessels whose ballast waters were sampled. While the ballast waters were from different lakes, all harbor waters were from Lake Superior. The ballast (B) waters were designated BB, IB, and MB and their matching harbor (H) waters as BH, IH, and MH. The identification, description, and collection information for the ballast and harbor

waters are summarized in Table 5.1.

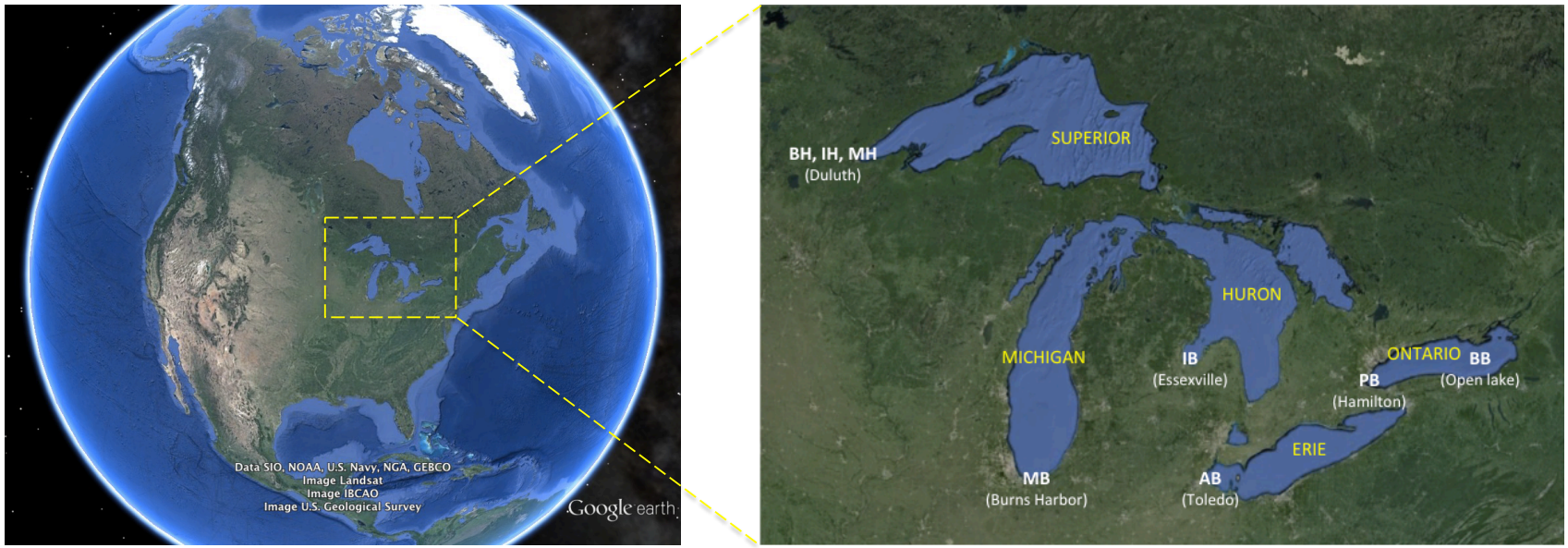


Figure 5.1. Google Earth maps showing the Great Lakes (left) and ballast water source and discharge ports of the five vessels (right). Ballast waters (AB, BB, IB, MB, and PB) were originated from different parts (Lakes Erie, Huron, Michigan, and Ontario) of the Great Lakes but all discharged in three terminals (BH, IH, and MH) of the Port of Duluth–Superior.

Table 5.1. Identification, description, and collection information for the ballast and harbor water samples

Sample[a]	Sample type	Source port[b]	Discharge port (sampling location)	Voyage duration[c] (sampling date)	Sampling method	Ballast water treatment
AB	Ballast water	Toledo (Lake Erie)	Duluth, Terminal C (Lake Superior)	3 days (5/15/13)	Ballast water pipeline	Untreated
BB	Ballast water	Open lake (Lake Ontario)	Duluth, Terminal A (Lake Superior)	7 days (5/9/13)	Ballast water pipeline	Untreated
BH	Harbor water	–	Duluth, Terminal A (Lake Superior)	- (5/9/13)	Bucket with rope	–
IB	Ballast water	Essexville (Lake Huron)	Duluth, Terminal C (Lake Superior)	2 days (5/10/13)	Ballast water pipeline	Untreated
IH	Harbor water	–	Duluth, Terminal C (Lake Superior)	– (5/10/13)	Bucket with rope	–
MB	Ballast water	Burns Harbor (Lake Michigan)	Duluth, Terminal B (Lake Superior)	4-8 days (5/10/13)	Sounding pipe	Untreated
MH	Harbor water	–	Duluth, Terminal B (Lake Superior)	– (5/10/13)	Bucket with rope	–
PB	Ballast water	Hamilton (Lake Ontario)	Duluth, Terminal C (Lake Superior)	4 days (5/14/13)	Ballast water pipeline	Untreated

[a] The ballast water was designated BB, IB, and MB and their matching harbor water as BH, IH, and MH.

[b] Ballast water source ports were where the vessels sampled for ballast water had undergone ballast water exchange prior to their arrivals at the Port of Duluth–Superior.

Table 5.1 (cont'd)

[c] Voyage duration was the difference in days between date of ballast water uptake from the ballast water source port and date of ballast water sampling from the Port of Duluth–Superior.

5.2.2. Preparation and sequencing of viromes

Ballast and harbor water samples were stored at 4 °C and processed for tangential flow filtration (TFF) within 24 hours of sample collection in a laboratory of the University of Minnesota–Duluth. The filtration concentrates (approximately 500 mL) were transported overnight to Michigan State University (MSU) at 4 °C and stored immediately at –80 °C upon arrival for further processing. Details on virome preparation, including concentration and purification of viral particles and extraction and amplification of viral nucleic acids, and metagenomic sequencing of ballast and harbor waters were described in the Chapter 4.

5.2.3. Analysis of viromes

Preprocessing of raw sequence reads, *de novo* assembly of sequence reads and taxonomic classification of contiguous reads (contigs) were performed as previously described in the Chapter 4.

To investigate emerging viral pathogens of fish and shrimp, a list of viral pathogens listed as notifiable by the World Organization for Animal Health (OIE) was retrieved from Walker and Winton (2010) and examined among the ballast and harbor water viromes. Genetic and taxonomic information of viral pathogens of fish and shrimp were obtained from ViralZone database (http://viralzone.expasy.org) (Hulo et al., 2011).

Annotation–free approaches, which are independent from taxonomic assignment of contigs were used to compare individual viromes with each other. Analysis of shared homologs of contigs present in each virome was performed by pairwise comparisons using TBLASTX (E–value < 10^{-3}). The percentage of shared number of contigs in each

direction was used to represent the similarity between viromes (Rodriguez-Brito et al., 2010). Furthermore, pairwise comparisons were performed by determining contig coverage between viromes using QUAST version 2.3 (Gurevich et al., 2013). An alignment of all contigs in a virome to all contigs in the other virome (used as a reference) was performed and the ratio of aligned bases to the number of all bases in the reference was used to represent the contig coverage (Gurevich et al., 2013). Similarities of the eight viromes were visualized with principal coordinates analysis (PCoA) using PAST statistical package version 3 (Hammer et al., 2001). Viromes of two temperate freshwater lakes in France were downloaded from MetaVir version 2 (http://metavir-meb.univ-bpclermont.fr/) and included in the PCoA analysis as an outlier group (Roux et al., 2012; Roux et al., 2014).

Lastly, the Great Lakes ballast and harbor water viromes were compared with a set of previously published viromes from other aquatic environments using MetaVir based on sequence similarity with a cross–TBLASTX search (Roux et al., 2014). The MetaVir workflow requires data sets containing at least 50,000 input sequences for the TBLASTX–based comparison, thus the analysis was not available for the two viromes in this research (35,819 and 15,887 contigs for BB and MB viromes, respectively).

5.2.4. Data access

All Illumina sequencing reads from viromes of the ballast and harbor water collected from the Port of Duluth–Superior are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (www. ncbi.nlm.nih.gov/sra) under accession number SRP048255.

5.3. Results

5.3.1. General water quality

The pH, dissolved oxygen, salinity, and turbidity of the ballast and harbor waters are summarized in Table 5.2. As expected, all samples were low in salinity. The harbor waters collected from Lake Superior were potentially influenced by shipping activities and had an average pH of 6.9, while the ballast water from vessels originating from Lakes Erie, Huron, Michigan, and Ontario had an average pH of 6.7. Ballast waters from Lakes Erie (AB) and Ontario (PB) had low dissolved oxygen (DO, 4.9 mg/L) resulting in a lower average of DO (6.8 mg/L) in comparison to harbor waters from Lake Superior (8.6 mg/L). The harbor waters generally had higher turbidity levels (average 9.7 NTU) compared with the ballast waters (average 3.9 NTU) with the exception of the ballast water from Lake Erie (AB, 10.0 NTU).

Table 5.2. Water quality of the ballast and harbor water samples

Sample (Lake)	Sample type	pH	Dissolved oxygen (mg/L)	Salinity (ppt)	Turbidity (NTU)
AB (Erie)	Ballast water	6.9	4.9	0.0	10.0
BB (Ontario)	Ballast water	6.3	7.3	0.0	4.5
IB (Huron)	Ballast water	6.7	8.4	0.1	3.3
MB (Michigan)	Ballast water	6.8	8.4	0.0	0.4
PB (Ontario)	Ballast water	6.6	4.9	0.0	1.4
Average for the ballast water		6.7	6.8	0.0	3.9
BH (Superior)	Harbor water	7.1	9.3	0.0	10.1
IH (Superior)	Harbor water	7.0	8.0	0.0	8.4
MH (Superior)	Harbor water	6.7	8.4	0.0	10.5
Average for the harbor water		6.9	8.6	0.0	9.7

Abbreviation: NTU, Nephelometric Turbidity Units.

5.3.2. Overview of the virome data sets

A pipeline for metagenomics analysis of the ballast and harbor waters was developed to address the identification of both DNA and RNA viruses. The use of Illumina Hiseq 2500 resulted in a total of 551,022,890 raw sequence reads with a read length of 100 base pair (bp). After quality trimming and filtering of reads, 501,015,363 reads (90.3% of raw reads) remained. Remaining reads were then split into 470,931,386 pair–end (PE; 94% of remaining reads) and 30,083,977 single–end (SE; 6% of remaining reads) reads prior to *de novo* assembly (Table 5.3). The PE reads were used for assembling reads into contigs, producing a total of 867,050 assembled contigs. The mean contig length among the contigs across eight viromes was 590 bp, and the mean N50 was 638 bp. A significant increase in contig length generated from *de novo* assembly improves annotation through homology searches against a reference database. Mapping PE reads to contigs showed that overall alignment rates ranged from 26.8 to 91.2% depending on the viromes (Table 5.4).

Table 5.3. Summary of virome datasets

	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Raw sequence reads								
# Reads	81,819,466	55,195,294	48,435,448	62,910,326	36,241,288	107,604,262	106,295,946	52,520,860
File size (Gb)	20.3	13.8	12.1	15.7	9.0	26.8	26.5	13.1
GC content (%)	47	48	50	46	48	50	49	43
Trimmed sequence reads								
# Reads	76,504,414	50,189,945	43,716,322	53,427,691	31,739,763	98,252,537	99,807,660	47,377,031
# PE reads	73,173,590	47,039,092	41,022,174	48,552,062	29,362,850	91,752,862	95,550,042	44,478,714
# SE reads[a]	3,330,824	3,150,853	2,694,148	4,875,629	2,376,913	6,499,675	4,257,618	2,898,317
Assembled contigs								
# Contigs	159,031	35,819	137,701	64,111	112,519	15,887	244,092	93,890
Total length (bp)	95,474,732	21,608,387	66,966,312	39,669,138	52,267,891	10,148,153	127,453,428	77,579,762
Mean length (bp)	600	603	486	618	464	638	522	792
Maximum length (bp)	14,631	11,142	8,376	42,020	9,291	13,151	5,637	45,918
Minimum length (bp)	200	201	200	200	201	200	200	200
N50 (bp)	646	636	492	660	454	703	547	962

Abbreviations: PE, paired–end; SE, single–end.

[a] The SE reads were not used for downstream analyses.

Table 5.4. Sample–specific PE reads mapped to individual assemblies

	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
PE reads unassembled	73,173,590	47,039,092	41,022,174	48,552,062	29,362,850	91,752,862	95,550,042	44,478,714
Concordant alignment	45,959,868	15,319,200	9,598,184	9,587,298	5,056,950	41,653,272	61,593,574	18,319,482
	(62.81%)	(32.57%)	(23.40%)	(19.75%)	(17.22%)	(45.40%)	(64.46%)	(41.19%)
Discordant alignment	2,292,586	512,246	706,116	2,463,412	349,206	1,079,128	2,468,740	3,274,350
	(3.13%)	(1.09%)	(1.72%)	(5.07%)	(1.19%)	(1.18%)	(2.58%)	(7.36%)
The rest	12,282,818	15,556,252	8,496,063	6,544,750	2,455,376	40,953,824	14,189,021	9,876,566
	(16.79%)	(33.07%)	(20.71%)	(13.48%)	(8.36%)	(44.63%)	(14.85%)	(22.21%)
Overall alignment rate (%)	**82.73**	**66.73**	**45.83**	**38.30**	**26.77**	**91.21**	**81.90**	**70.75**

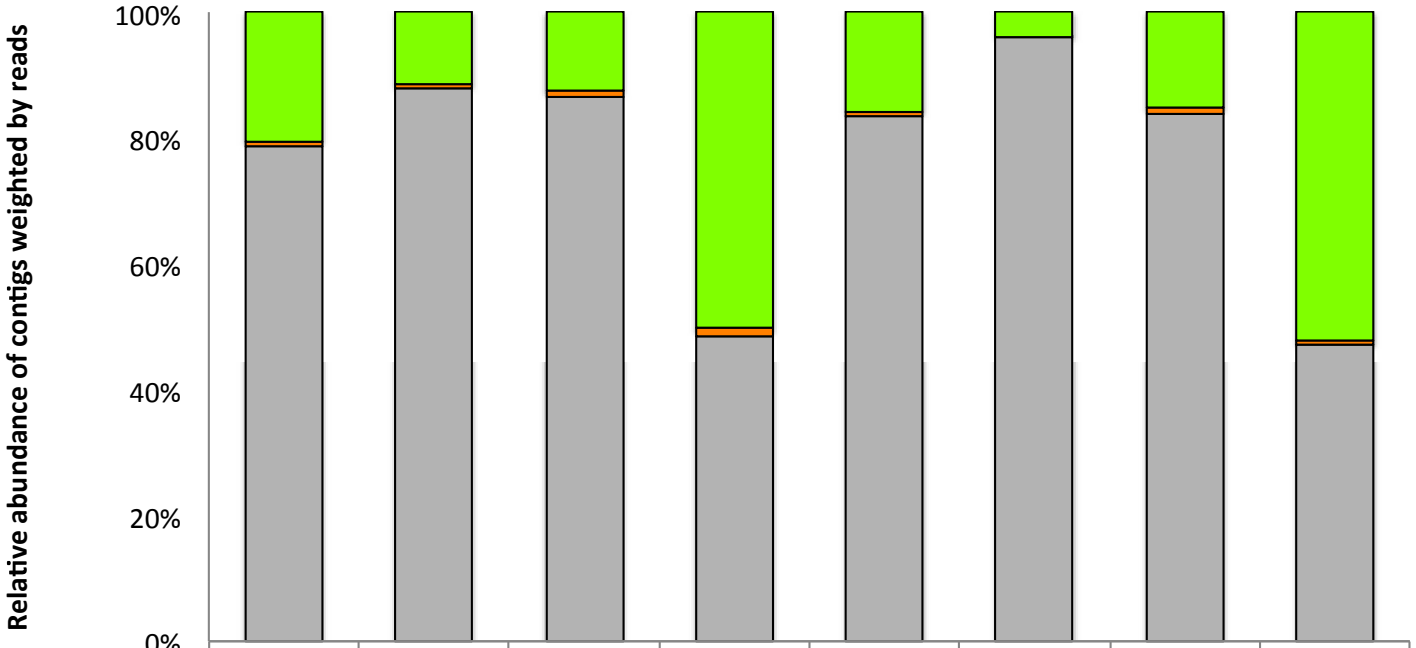
Abbreviation: PE, paired–end.

Read mapping to contigs was performed using default settings in Bowtie 2. Numbers in parentheses indicate specific alignment rate.

5.3.3. Taxonomic profile of the Great Lakes ballast and harbor water viromes

A wide array of viruses was discovered in all ballast and harbor waters from the Great Lakes. Among all assembled contigs, about 22.8% were assigned to viral taxa, but 77.2% had no or low levels of amino acid similarity to known viral sequences in the NCBI RefSeq database (Figure 5.2A). Of the contigs with similarity to known viruses, 34 different viral families were identified, consisting of 15 double–stranded (ds) DNA (69.7%), six single–stranded (ss) DNA (19.4%), one dsRNA (0.1%), and 12 ssRNA (6.8%) viruses (Figure 5.2B). These represented viruses infecting a wide range of hosts, including bacteria (62.1%), vertebrates/invertebrates (12.6%), algae (2.9%), plants (1.6%), amoebae (1.0%), and fungi/protozoa (0.01%; Figure 5.2C).

A



	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Assigned	20.4	11.6	12.6	50.3	15.7	3.9	15.3	52.3
Not assigned	0.9	0.4	1.0	1.1	0.8	0.1	0.8	0.7
No hits	78.7	88.0	86.4	48.6	83.5	96.0	83.9	47.0

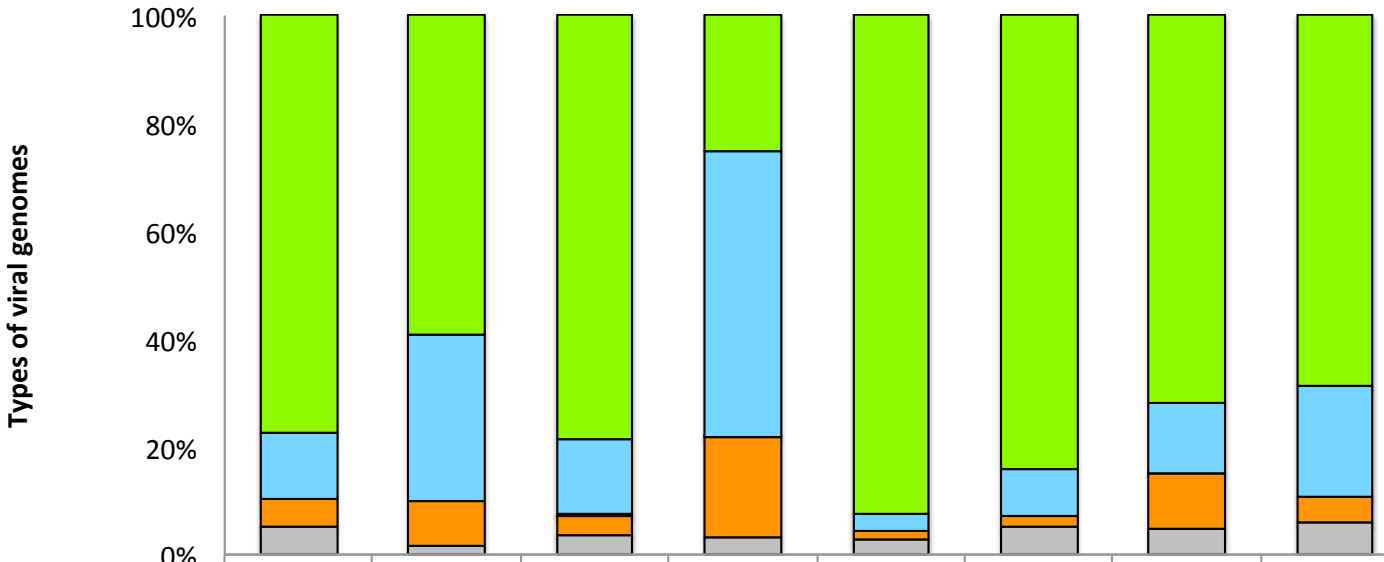
Figure 5.2. Taxonomic profile of the ballast and harbor water viromes. Relative abundance of contigs weighted by sequence reads based on taxonomic assignment of contigs (A). Contigs that were assigned to viral taxa but did not meet the selected MEGAN

Figure 5.2 (cont'd)

parameters were placed under "Not assigned," and contigs that did not have any hits to known sequences in the databases were placed under "No hits." Types of viral genomes (B), types of virus hosts (C), and contigs assigned to viral families (D) in the ballast and harbor water viromes. Viral families whose maximum relative abundances across eight viromes less than 3% were represented as "Others[a]". Unassigned contigs at the family level were represented as "Others[b]."

Figure 5.2 (cont'd)

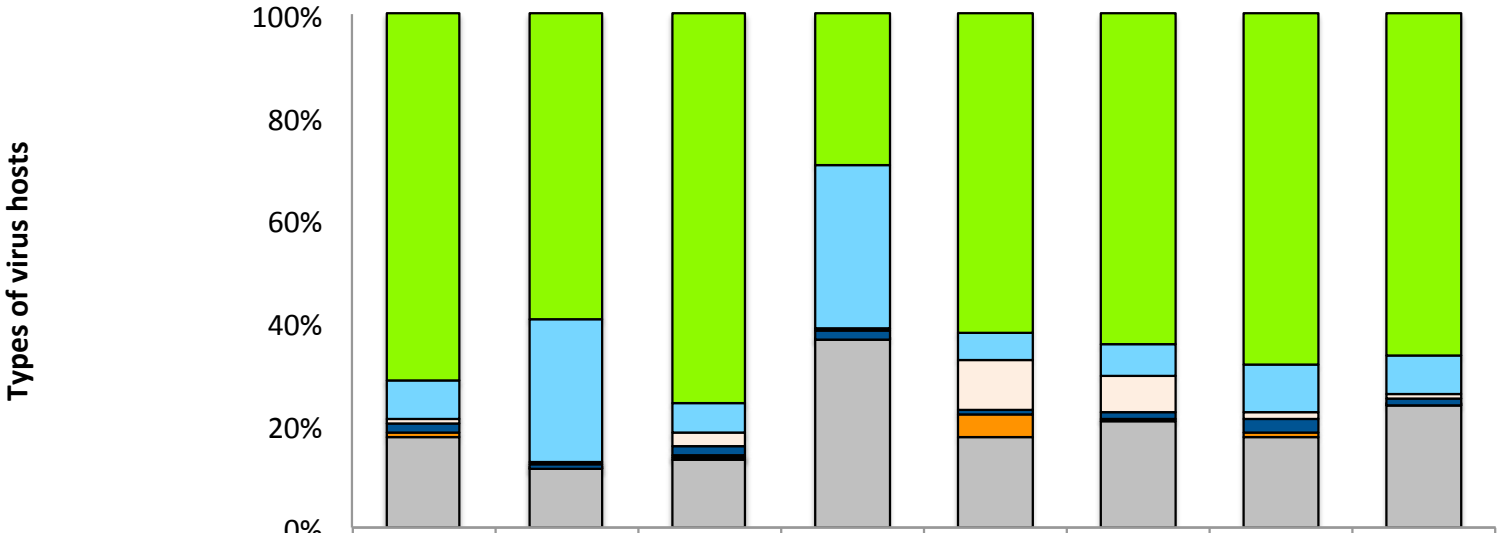
B



	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
dsDNA viruses	77.6	59.3	78.5	25.3	92.4	84.1	71.7	68.9
ssDNA viruses	12.1	30.6	13.9	52.8	3.1	8.8	13.3	20.4
dsRNA viruses	0.0	0.0	0.3	0.1	0.0	0.0	0.2	0.0
ssRNA viruses	5.0	8.4	3.6	18.7	1.6	2.1	10.2	4.7
Unassigned	5.2	1.7	3.8	3.1	2.9	5.0	4.7	6.0

Figure 5.2 (cont'd)

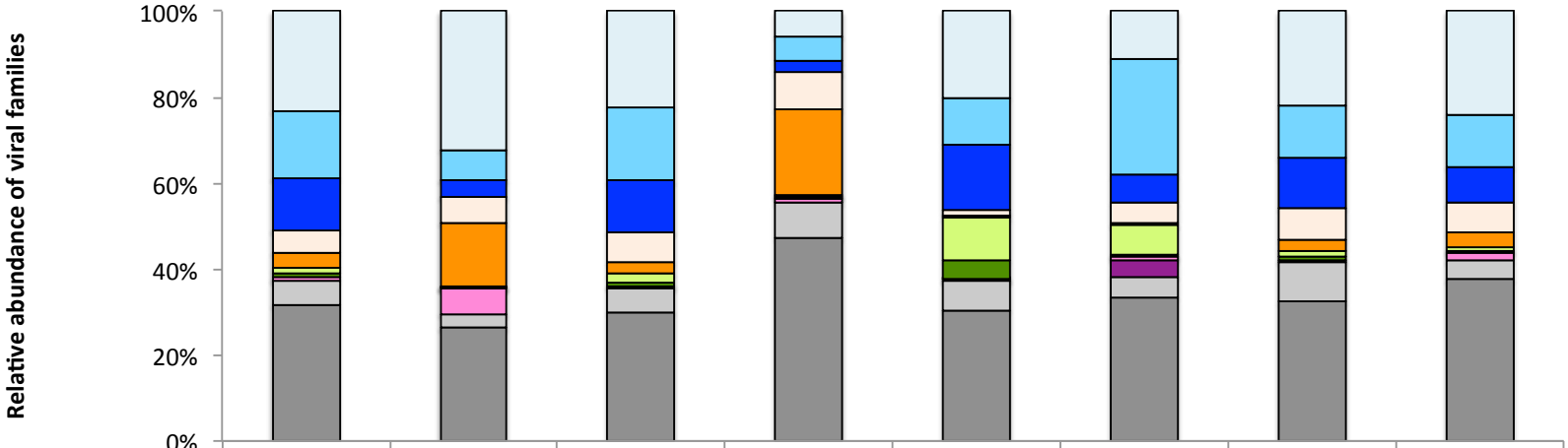
C



	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Bacteria	71.5	59.3	75.7	29.4	62.1	64.3	68.2	66.5
Vertebrates/Invertebrates	7.2	28.0	6.0	31.8	5.2	6.1	9.5	7.5
Algae	1.2	0.3	2.3	0.5	9.7	7.0	1.3	0.8
Plants	1.5	1.0	1.9	1.8	1.0	1.5	2.4	1.3
Amoebae	1.0	0.1	0.6	0.1	4.5	0.2	0.9	0.3
Fungi/Protozoa	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Unknown host	17.5	11.3	13.4	36.4	17.6	20.9	17.7	23.7

Figure 5.2 (cont'd)

D



	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Siphoviridae	23.2	32.2	22.2	5.7	20.1	11.1	21.8	24.0
Podoviridae	15.6	6.9	17.3	5.9	10.8	27.0	12.3	12.2
Myoviridae	12.0	4.1	11.8	2.6	15.6	6.5	11.6	8.5
Microviridae	5.2	6.0	7.0	8.6	0.9	4.9	7.4	6.8
Parvoviridae	3.8	14.8	2.7	19.8	0.7	0.4	2.6	3.6
Phycodnaviridae	1.2	0.3	2.3	0.5	9.7	7.0	1.3	0.8
Mimiviridae	1.0	0.1	0.6	0.1	4.4	0.2	0.9	0.3
Circoviridae	0.7	6.0	0.6	1.1	0.3	0.9	0.6	1.9
Ascoviridae	0.0	0.0	0.0	0.0	0.3	3.7	0.0	0.0
Others[a]	5.7	3.3	5.5	8.3	6.6	5.0	9.2	4.5
Others[b]	31.6	26.3	29.9	47.1	30.5	33.4	32.4	37.5

Relative abundance of viral families revealed that more than half (average 52.5%) of the assigned contigs in each virome were homologous to dsDNA phages, belonging to the order of *Caudovirales* (*Myoviridae*, *Podoviridae*, *Siphoviridae*, and unclassified *Caudovirales*) with the exception of the IB virome (Lake Huron) with a low relative abundance of 19.3% (Figure 5.2D). These dsDNA phages were associated with 62 different bacterial hosts, with the majority being *Cellulophaga* (average 14.1%) followed by *Synechococcus* (9.1%) and *Pelagibacter* (6.9%; Table 5.5). Along with *Synechococcus*, a number of contigs was found to be associated with phages whose hosts belong to cyanobacteria such as *Prochlorococcus* (3.3%). Moreover, contigs most similar to those infecting bacteria in genera containing human pathogens, including *Burkholderia* (1.3%), *Klebsiella* (0.3%), *Pseudomonas* (3.6%), *Salmonella* (2.1%), and *Vibrio* (6.0%), were detected.

Table 5.5. Distribution of dsDNA phage hosts identified in the ballast and harbor water viromes

Host	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Acinetobacter	0.62	0.46	0.62	0.00	0.63	0.54	0.65	0.00
Actinoplanes	0.96	0.82	0.93	0.00	0.48	0.00	0.60	0.56
Aeromonas	1.33	0.97	1.59	0.77	1.68	0.43	1.64	0.42
Aggregatibacter	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00
Agrobacterium	0.00	0.00	0.54	0.00	0.48	0.00	0.44	0.00
Alteromonas	0.47	0.00	0.45	0.00	0.54	0.00	0.45	0.00
Arthrobacter	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00
Azospirillum	0.90	0.00	0.87	0.00	0.93	0.00	1.17	0.68
Bacillus	2.52	0.56	3.07	0.00	3.41	2.90	3.66	2.00
Bdellovibrio	3.71	2.61	3.34	5.87	2.13	0.64	3.55	1.99
Burkholderia	1.80	0.46	2.07	0.90	1.44	0.43	2.72	0.64
Campylobacter	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00
Caulobacter	1.66	1.28	2.16	0.00	2.57	1.39	2.35	0.54
Celeribacter	0.67	0.66	0.84	0.77	0.99	0.00	0.58	0.72
Cellulophaga	12.92	11.91	8.60	22.46	9.28	14.59	9.91	22.44
Clavibacter	0.74	0.00	1.23	1.51	0.54	0.00	0.91	0.00
Clostridium	0.00	0.56	0.00	0.00	1.44	0.97	0.00	0.00
Colwellia	0.00	0.66	0.00	1.96	0.00	0.43	0.00	0.00
Croceibacter	0.61	0.00	0.75	0.00	0.48	0.00	0.64	0.57

Table 5.5 (cont'd)

Host	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Cronobacter	1.96	0.61	1.09	1.30	3.47	1.82	1.98	0.88
Delftia	0.00	0.00	0.00	0.00	0.75	0.00	0.44	0.00
Enterobacter	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00
Enterococcus	0.00	0.46	0.00	0.00	0.00	0.32	0.00	0.00
Escherichia	0.86	1.07	0.70	1.67	1.89	0.97	0.72	1.36
Flavobacterium	2.42	1.74	2.18	5.50	1.05	0.64	1.85	4.62
Hamiltonella	0.00	0.00	0.00	0.00	0.51	0.00	0.00	0.50
Iodobacter	0.00	0.46	0.00	0.00	0.00	0.00	0.00	0.00
Klebsiella	0.00	0.00	0.00	0.00	0.99	0.64	0.47	0.00
Lactobacillus	0.00	0.00	0.00	0.00	0.00	0.32	0.49	0.46
Lactococcus	0.53	0.00	1.61	1.02	2.87	0.86	1.36	1.02
Mycobacterium	2.08	1.84	3.02	0.00	2.13	1.50	2.53	0.95
Myxococcus	5.43	5.62	5.94	3.95	3.41	1.50	4.82	2.98
Pantoea	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00
Pelagibacter	5.90	9.86	6.16	9.13	8.08	4.61	4.78	6.49
Planktothrix	0.71	0.00	0.66	1.43	4.37	0.86	0.97	0.00
Prochlorococcus	2.59	4.85	2.85	1.47	2.78	5.58	2.26	3.97
Pseudoalteromonas	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00
Pseudomonas	3.89	1.07	7.10	1.96	3.50	1.72	6.68	2.81

Table 5.5 (cont'd)

Host	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Psychrobacter	1.15	1.43	0.79	1.22	0.00	2.15	0.82	1.38
Puniceispirillum	3.08	4.34	3.25	3.14	3.17	3.97	2.59	4.41
Ralstonia	3.49	2.61	3.78	3.47	2.78	11.27	2.97	4.58
Rhizobium	1.50	0.51	1.69	1.26	1.47	0.54	1.91	1.10
Rhodococcus	4.08	7.05	3.89	1.26	2.07	2.25	4.13	3.91
Rhodothermus	0.99	0.51	0.43	0.00	0.99	1.07	0.74	1.28
Riemerella	1.04	0.97	0.62	1.88	0.72	0.43	0.60	1.53
Roseobacter	0.71	0.72	0.91	1.35	0.60	0.86	1.02	0.86
Escherichia	0.86	1.07	0.70	1.67	1.89	0.97	0.72	1.36
Salinivibrio	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00
Salmonella	1.95	3.17	2.00	3.42	1.68	0.97	1.45	2.49
Sinorhizobium	0.00	0.00	0.43	0.00	0.00	0.32	0.59	0.00
Sphingomonas	0.83	0.51	0.00	0.00	0.69	0.75	0.53	0.85
Staphylococcus	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00
Stenotrophomonas	0.00	0.00	0.00	0.00	0.48	0.00	0.00	0.00
Streptococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
Streptomyces	0.59	0.46	0.62	0.00	0.63	0.32	0.63	0.61
Synechococcus	8.84	14.61	6.60	8.23	9.67	8.69	6.84	9.17
Tetrasphaera	0.72	0.51	0.57	0.00	0.00	0.00	0.76	0.88

Table 5.5 (cont'd)

Host	AB (Erie)	BB (Ontario)	BH (Superior)	IB (Huron)	IH (Superior)	MB (Michigan)	MH (Superior)	PB (Ontario)
Thalassomonas	2.20	1.89	2.46	1.55	1.20	1.39	1.79	1.61
Thermoanaerobacterium	1.48	3.17	1.46	2.53	1.17	5.26	1.15	2.34
Vibrio	8.84	4.04	7.92	4.16	6.11	3.97	9.65	3.34
Xanthomonas	2.55	4.96	3.00	2.36	3.20	9.44	2.36	2.07
Xylella	0.68	0.00	0.80	1.02	0.00	0.00	0.88	0.46
Yersinia	0.00	0.00	0.00	0.00	0.60	0.32	0.00	0.00

Values are represented as percentage of number of contigs in each virome.

Vertebrate (including those that could infect humans) and invertebrate viruses were present in all samples (average 12.7%). Two ballast water viromes, BB (Lake Ontario; 28.0%) and IB (Lake Huron; 31.8%), had higher relative abundances of vertebrate and invertebrate viruses due to significantly higher abundances of *Parvoviridae*, which is capable of infecting either vertebrates or invertebrates (14.8% and 19.8% for the BB and IB viromes, respectively). The contigs related to *Alphatetraviridae* (insect viruses), *Iflaviridae* (insect viruses), and unassigned *Picornavirales* (vertebrate/invertebrate viruses) were present in at least one of the ballast water viromes but not in any of the Lake Superior harbor water viromes (Table 5.6). In contrast, contigs related to *Reoviridae* (vertebrate/invertebrate viruses) were detected only in one of the harbor water viromes, BH (Lake Superior).

Table 5.6. Relative abundance of the viral families in the ballast and harbor water viromes

Viral family	Host	AB	BB	BH	IB	IH	MB	MH	PB
dsDNA viruses									
Alloherpesviridae	Vertebrates	0.00	0.00	0.00	0.00	0.23	0.85	0.00	0.00
Ascoviridae	Invertebrates	0.00	0.00	0.00	0.00	0.31	3.74	0.00	0.00
Baculoviridae	Invertebrates	0.00	0.00	0.00	0.00	0.36	0.01	0.00	0.00
Herpesviridae	Vertebrates	0.17	0.00	0.00	0.04	0.94	0.01	0.09	0.13
Iridoviridae	**Vertebrates/Invertebrates**	**0.18**	**0.06**	**0.13**	**0.09**	**0.95**	**0.06**	**0.57**	**0.18**
Marseilleviridae	Amoebae	0.00	0.00	0.00	0.00	0.14	0.02	0.00	0.00
Mimiviridae	**Amoebae**	**1.03**	**0.13**	**0.58**	**0.15**	**4.39**	**0.17**	**0.89**	**0.27**
Myoviridae	**Bacteria**	**12.05**	**4.06**	**11.76**	**2.65**	**15.59**	**6.49**	**11.60**	**8.55**
Nudiviridae	Invertebrates	0.00	0.03	0.00	0.00	0.13	0.01	0.00	0.00
Phycodnaviridae	**Algae**	**1.20**	**0.28**	**2.32**	**0.52**	**9.69**	**6.98**	**1.30**	**0.80**
Podoviridae	**Bacteria**	**15.65**	**6.91**	**17.27**	**5.93**	**10.80**	**27.00**	**12.31**	**12.16**
Polydnaviridae	Invertebrates	0.00	0.00	0.00	0.15	0.27	0.02	0.00	0.12
Poxviridae	Vertebrates/Invertebrates	0.00	0.00	0.00	0.10	0.33	0.03	0.14	0.41
Reoviridae	Vertebrates/Invertebrates	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00
Siphoviridae	**Bacteria**	**23.15**	**32.24**	**22.24**	**5.73**	**20.05**	**11.08**	**21.79**	**23.96**
Caudovirales	Bacteria	10.64	7.90	12.56	4.76	8.99	7.43	10.91	5.72
Unclassified *Caudovirales*	Bacteria	1.09	0.53	1.52	0.24	0.80	3.77	1.20	0.60
Herpesvirales	Vertebrates/Invertebrates	0.10	0.04	0.33	0.01	0.16	0.01	0.16	0.03
dsDNA viruses	Unassigned	8.13	5.35	6.53	3.55	12.70	11.14	8.04	7.74
Unclassified dsDNA viruses	Unassigned	1.05	0.27	0.70	0.12	1.20	1.74	0.68	0.54
Unclassified dsDNA phages	Bacteria	3.14	1.52	2.35	1.29	4.41	3.57	2.03	7.68

Table 5.6 (cont'd)

Viral family	Host	AB	BB	BH	IB	IH	MB	MH	PB
ssDNA viruses									
Circoviridae	**Vertebrates**	**0.74**	**6.00**	**0.64**	**1.15**	**0.35**	**0.92**	**0.57**	**1.86**
Geminiviridae	**Plants**	**0.24**	**0.30**	**1.33**	**0.77**	**0.16**	**0.01**	**0.38**	**0.39**
Inoviridae	Bacteria	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00
Microviridae	**Bacteria**	**5.18**	**6.01**	**7.04**	**8.64**	**0.92**	**4.89**	**7.37**	**6.81**
Nanoviridae	Plants	0.00	0.12	0.00	0.13	0.00	0.11	0.00	0.29
Parvoviridae	**Vertebrates/Invertebrates**	**3.78**	**14.83**	**2.72**	**19.84**	**0.73**	**0.41**	**2.60**	**3.61**
ssDNA viruses	Unassigned	0.78	1.48	0.39	1.45	0.12	1.29	0.76	3.23
Unclassified ssDNA viruses	Unassigned	1.44	1.86	1.79	20.84	0.73	1.19	1.58	4.24
dsRNA viruses									
Totiviridae	Fungi/Protozoa	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00
dsRNA viruses	Unassigned	0.00	0.00	0.14	0.10	0.00	0.00	0.17	0.00
ssRNA viruses									
Alphatetraviridae	Invertebrates	0.00	0.03	0.00	0.14	0.00	0.00	0.00	0.00
Benyviridae	Plants	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Dicistroviridae	**Invertebrates**	**0.81**	**0.38**	**0.29**	**2.45**	**0.09**	**0.02**	**1.65**	**0.31**
Iflaviridae	Invertebrates	0.00	0.02	0.00	0.17	0.00	0.00	0.00	0.00
Leviviridae	Bacteria	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.00
Nodaviridae	Vertebrates/Invertebrates	0.10	0.06	0.11	0.11	0.08	0.00	0.52	0.00
Ourmiavirus	Plants	0.10	0.00	0.00	0.00	0.00	0.02	0.36	0.00
Picornaviridae	Vertebrates	0.00	0.04	0.00	0.08	0.00	0.00	0.13	0.00

Table 5.6 (cont'd)

Viral family	Host	AB	BB	BH	IB	IH	MB	MH	PB
Secoviridae	Plants	0.00	0.08	0.00	0.16	0.00	0.00	0.24	0.00
Sobemovirus	Plants	0.00	0.03	0.00	0.07	0.00	0.03	0.00	0.00
Tombusviridae	**Plants**	**1.21**	**0.16**	**0.57**	**0.56**	**0.83**	**1.35**	**1.40**	**0.28**
Virgaviridae	Plants	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.30
Picornavirales	Vertebrates/Invertebrates	0.58	6.17	0.56	1.75	0.15	0.00	1.57	0.24
Unassigned Picornavirales	Vertebrates/Invertebrates	0.00	0.05	0.00	0.36	0.00	0.00	0.00	0.05
Environmental samples <*Picornavirales*>	Vertebrates/Invertebrates	0.76	0.26	1.00	5.32	0.09	0.02	1.50	0.53
Tymovirales	Plants	0.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00
ssRNA positive-strand viruses	Unassigned	1.00	0.68	0.62	2.02	0.25	0.54	1.51	0.24
Unclassified +ssRNA viruses	Unassigned	0.46	0.09	0.40	5.42	0.06	0.09	0.88	2.76
ssRNA viruses	Unassigned	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.01
Unassigned									
Viruses	Unassigned	4.52	1.10	2.81	2.48	2.49	4.86	3.90	4.77
Unclassified phages	Bacteria	0.56	0.12	0.94	0.16	0.38	0.10	0.59	1.00
Satellites	Unassigned	0.16	0.48	0.00	0.42	0.00	0.03	0.19	0.21

Abbreviations: ds, double–stranded; ss, single–stranded.

Bold letters indicated viral families shared by all ballast and harbor water samples.

Contigs belonging to algal viruses, *Phycodnaviridae*, were present in all viromes (average 2.9%), with higher relative abundances in the IH (Lake Superior; 9.7%) and MB (Lake Michigan; 7.0%) viromes. Viruses infecting plants were also present in all viromes (average 0.17%). The contigs related to *Benyviridae*, *Nanoviridae*, *Sobemovirus*, *Tymovirales*, and *Virgaviridae* were present in at least one of the ballast water viromes but not in any of the Lake Superior harbor water viromes (Table 5.6). Viruses infecting fungi/protozoa, *Totiviridae*, were only present in one of the Lake Superior harbor water viromes, BH (0.12%). Contigs belonging to amoeba viruses, *Mimiviridae*, were present in all viromes, while *Marseilleviridae* were present only in one of the harbor water viromes, IH (Lake Superior), and in the MB (Lake Michigan) virome.

5.3.4. Viral pathogens of fish and shrimp in the Great Lakes ballast and harbor waters

Our metagenomic data allowed for an in-depth examination of the types of viruses that might be considered to be key pathogens of fish and shrimp in the Great Lakes. Of the assigned contigs in the ballast and harbor water viromes, 75 contigs were identified as viral pathogens of fish or shrimp (Table 5.7). The identified contigs had lower amino acid similarity (23−44%) to known viruses in the RefSeq database except one contig in the MH virome (Lake Superior) identified as the infectious spleen and kidney necrosis virus (ISKNV) with 72% amino acid similarity. Mapping of reads to complete reference genomes of the identified viral pathogens of fish and shrimp showed low mapping rates (data not shown) except for the koi herpesvirus (KHV). Read mapping to five different KHV genomes exhibited slightly higher mapping rates with an average coverage of 7.6%

(Table 5.8). Overall, the low mapping rate of reads together with low amino acid similarity of contigs indicated that these viral pathogens are potentially novel or genetically diverse.

Table 5.7. Contigs identified as viral pathogens of fish and shrimp by BLASTX search against the NCBI RefSeq database

Virus	Taxonomic classification	AB	BB	BH	IB	IH	MB	MH	PB
Fish									
Infectious spleen and kidney necrosis virus	*Iridoviridae, Megalocytivirus*	3	0	0	0	0	0	1	0
Koi herpesvirus	*Alloherpesviridae, Cyprinivirus*	2	0	1	0	5	0	3	0
Striped Jack nervous necrosis virus	*Nodaviridae, Betanodavirus*	1	0	0	0	0	0	0	0
Shrimp									
Infectious myonecrosis virus	Unclassified *Totiviridae*	1	0	5	0	0	0	1	0
Macrobrachium rosenbergii nodavirus	*Nodaviridae*, *Alphanodavirus*	2	0	0	0	0	0	0	0
Taura syndrome virus	*Dicistroviridae, Aparavirus*	0	0	0	2	0	0	1	1
White spot syndrome virus	*Nimaviridae, Whispovirus*	3	1	0	13	2	0	4	23

Table 5.8. Summary statistics of read mapping to five reference genomes of koi herpesvirus

	# Reads[a]	# Bases[b]	Coverage		# Reads	# Bases	Coverage		# Reads	# Bases	Coverage
	AP008984.1				**DQ177346.1**				**DQ657948.1**		
AB	23,012	898	7.2	AB	26,689	938	8.2	AB	25,914	864	8.4
BB	235	515	0.1	BB	335	621	0.1	BB	269	609	0.1
BH	32,133	596	10	BH	33,819	799	10.5	BH	33,348	674	10.6
IB	7,943	631	2.5	IB	8,933	769	2.7	IB	8,505	570	2.8
IH	25,479	992	8	IH	32,293	944	9.7	IH	30,563	877	10.3
MB	9,667	743	3	MB	11,636	875	3.4	MB	10,841	827	3.7
MH	5,374	725	1.7	MH	6,359	893	1.9	MH	6,033	631	2
PB	69,106	705	21.8	PB	82,956	1,002	24.9	PB	78,189	1,008	26.6

	KJ627438.1				**NC009127.1**		
AB	26,068	1,006	8.2	AB	25,914	938	8.2
BB	267	636	0.1	BB	269	621	0.1
BH	33,431	703	10.5	BH	33,348	799	10.5
IB	8,559	753	2.7	IB	8,505	769	2.7
IH	30,900	1,179	9.8	IH	30,563	944	9.7
MB	10,902	802	3.5	MB	10,841	875	3.4
MH	5,893	948	1.8	MH	6,033	893	1.9
PB	79,030	917	25.2	PB	78,189	1,002	24.9

[a] Number of aligned reads to a reference genome sequence.

Table 5.8 (cont'd)

[b] Number of bases of a reference genome sequence aligned by reads.

Bold letters indicated complete genome sequences of Koi herpesvirus retrieved from the NCBI RefSeq database.

5.3.5. Comparison among the Great Lakes ballast and harbor water viromes

Comparisons of viromes of ballast waters from Lakes Erie, Huron, Michigan, and Ontario and harbor waters from Lake Superior were conducted to examine the similarities between lakes and to inspect whether the harbor waters near vessels were reflective of the respective ballast waters. For virome comparison, annotation-independent approaches were used as they use all contigs present in the virome data sets while annotation-dependent approaches use a low proportion of the contigs with similarity to known sequences in the existing database (average 22.8% of the contigs from Figure 5.2A). It should be noted that two annotation-independent approaches used in this research have different ways to analyze shared homologues of contigs in each virome. TBLASTX–based comparison uses a shared number of contigs between viromes while the QUAST tool uses a shared number of aligned bases in contigs.

The difference in contig profiles analyzed by two methods were consistent with only slight variations as observed in Figure 5.3A and 5.3B. PCoA analyses between viromes suggested that the three Lake Superior harbor water viromes (BH, IH, and MH) grouped together and were distinct from the respective ballast water viromes (BB, IB, and MB from Lakes Ontario, Huron, and Michigan, respectively). The ballast water virome, AB, from Lake Erie showed a similar contig profile with the Lake Superior harbor waters. In addition, the two ballast water viromes from Lake Ontario, BB (open lake) and PB (Port of Hamilton), showed different contig profiles to each other (Table 5.9 and 5.10). The similarity in the contig profiles between IB (Lake Huron) and MB (Lake Michigan) was somewhat expected because Lakes Huron and Michigan are two halves of one lake and considered to be hydrologically the most connected.

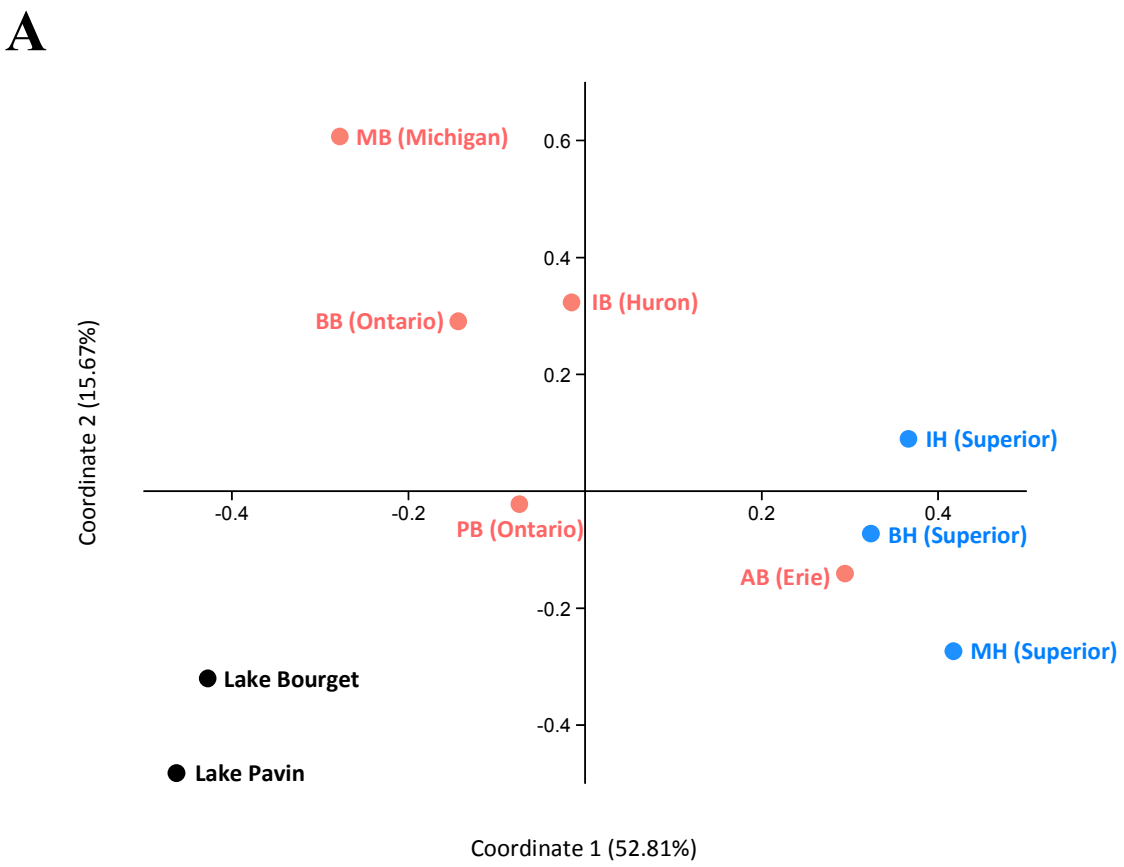


Figure 5.3. PCoA of virome distances based on contig profiles using TBLASTX (A) and QUAST (B). Ballast and harbor water viromes were represented in red and blue circles, respectively. An outlier group (viromes of two temperate freshwater lakes in France) was represented in black circles.

Figure 5.3 (cont'd)

B

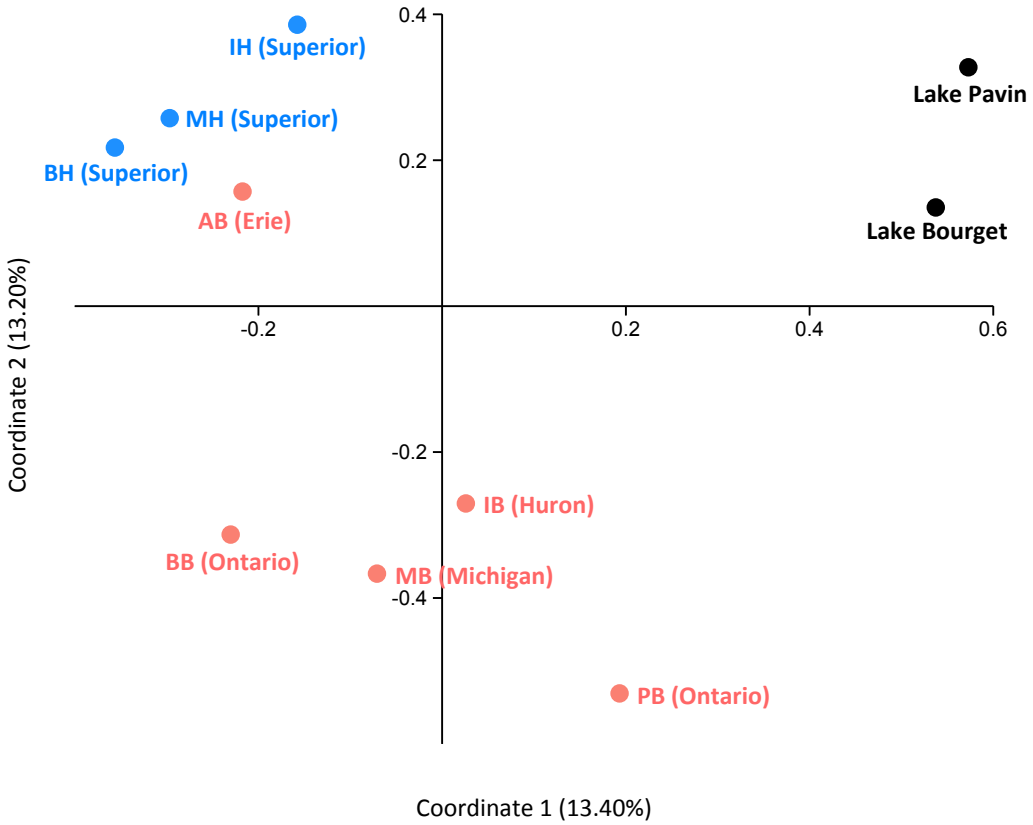


Table 5.9. Similarity matrix of contigs between ballast and harbor water viromes using TBLASTX

	BB (Ontario)	**BH** (Superior)	**IB** (Huron)	**IH** (Superior)	**MB** (Michigan)	**MH** (Superior)	**PB** (Ontario)	**Lake Bourget**	**Lake Pavin**
AB (Erie)	51.231	66.292	56.861	53.405	35.337	66.004	61.762	35.000	30.810
BB (Ontario)		50.582	69.964	38.089	63.207	41.040	70.684	57.067	50.189
BH (Superior)			52.867	50.478	33.949	63.363	57.151	31.139	27.238
IB (Huron)				41.875	53.904	45.529	76.330	47.711	41.054
IH (Superior)					32.705	53.005	48.103	24.896	21.102
MB (Michigan)						27.113	58.301	51.399	45.638
MH (Superior)							50.976	26.194	22.927
PB (Ontario)								63.379	58.066
Lake Bourget									77.670
Lake Pavin									

Lake Bourget and Lake Pavin viromes from two temperate freshwater lakes in France were included as an outlier group.

Table 5.10. Similarity matrix of contigs between ballast and harbor water viromes using QUAST

	BB (Ontario)	**BH** (Superior)	**IB** (Huron)	**IH** (Superior)	**MB** (Michigan)	**MH** (Superior)	**PB** (Ontario)	**Lake Bourget**	**Lake Pavin**
AB (Erie)	6.013	4.926	3.547	2.920	3.895	3.754	1.687	0.342	0.267
BB (Ontario)		9.393	6.468	3.826	10.861	7.216	11.264	2.198	0.435
BH (Superior)			2.654	4.142	3.274	5.751	1.227	0.252	0.240
IB (Huron)				1.586	3.319	2.095	6.955	0.585	0.215
IH (Superior)					2.052	3.321	0.711	0.134	0.130
MB (Michigan)						3.175	7.752	1.388	0.321
MH (Superior)							0.938	0.252	0.184
PB (Ontario)								2.919	1.891
Lake Bourget									2.161
Lake Pavin									

Lake Bourget and Lake Pavin viromes from two temperate freshwater lakes in France were included as an outlier group.

5.3.6. Comparison of the Great Lakes ballast and harbor water viromes with other aquatic viromes

The Great Lakes ballast and harbor water viromes were also compared to previously published viromes in freshwater and marine water. The hierarchical clustering tree analysis revealed that aquatic viromes could be classified into two representative groups, freshwater virome and marine water virome (Figure 5.4), highlighting that these different environments contain unique virome signatures. Among the freshwater virome, the Great Lakes ballast and harbor water viromes were closely related to each other and clustered with two viromes from the oligomesotrophic Lake Pavin and the mesotrophic Lake Bourget in France. Viromes from desert ponds and an aquaculture pond generated a subgroup with the temperate lake viromes. Viromes from an Antarctic lake and reclaimed and potable water were aggregated individually and distant from the other freshwater viromes.

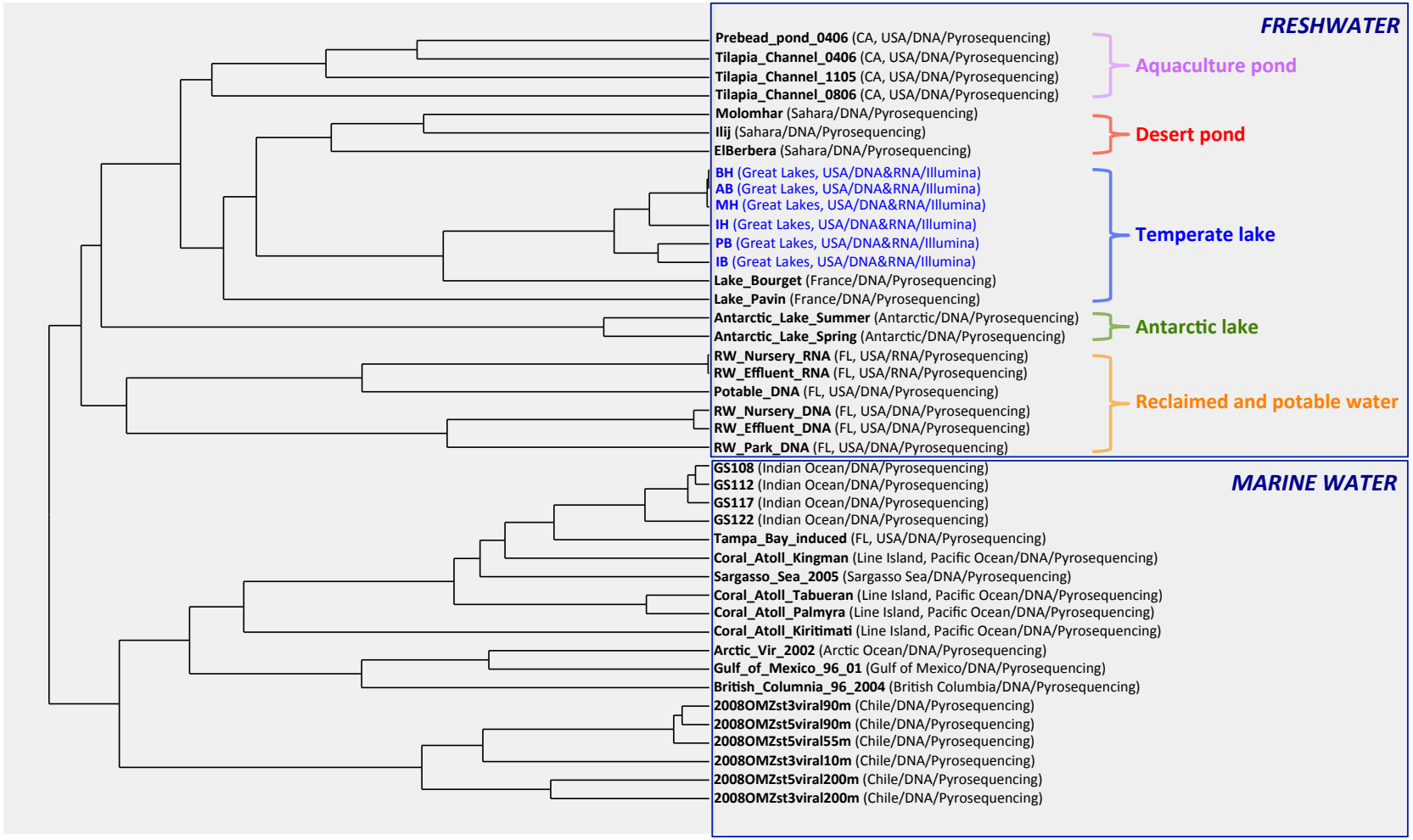


Figure 5.4. Comparison of the ballast and harbor water viromes with published freshwater viromes based on sequence similarity with a cross-TBLASTX search. The six viromes (AB, BH, IB, IH, MH, and PB) from this research are highlighted in blue. TBLASTX–

Figure 5.4 (cont'd)

based comparison was not available for the other two viromes (BB and MB), due to insufficient numbers of input contigs. Virome names were retrieved from the MetaVir. Sampling site, virome fraction, and sequencing platform of each virome were shown in the parentheses.

5.4. Discussion

This research investigated viromes in ballast and harbor waters originating from the Great Lakes including Lakes Erie, Huron, Michigan, Ontario, and Superior (Figure 5.1). The Great Lakes, located in northeastern North America, form the largest group of freshwater lakes on Earth (U.S. EPA). The sampling location at the Port of Duluth–Superior, located at the western part of Lake Superior (Figure 5.1), is the busiest and largest port on the Great Lakes and receives the largest volume (more than 18 million gallons a day) of ballast water within the Great Lakes (Minnesota Pollution Control Agency, 2008; U.S. EPA).

5.4.1. Diversity of viruses in ballast water

In contrast to what is known about the diversity of metazoans transported by ships' ballast water, little is known about the diversity of viruses in ballast water. Moreover, most studies have only focused on the level of VLPs in ballast water using microscopic approaches (Ruiz et al., 2000; Drake et al., 2002; Drake et al., 2007; Leichsenring and Lawrence, 2011). The present research gives us the first insight into the diversity of viruses and their associated host populations in ballast waters across the Great Lakes. Understanding this diversity including types of viruses that are being discharged will assist in defining ballast water treatment and potential ecological risk.

Viruses in the ballast waters from Lakes Erie, Huron, Michigan, and Ontario were characterized and compared with those of Lake Superior harbor waters where they were being discharged. Among 34 viral families identified, 12 viral families were shared by all ballast and harbor waters (highlighted in bold in Table 5.6), representing 62.2% of the

total abundance of phylogenetic groups in the viromes. This suggested that the majority of viromes among the different lakes were similar in the phylogenetic types. However, the presence of six viral families (*Alphatetraviridae*, *Beniviridae*, *Iflaviridae*, *Nanoviridae*, *Sobemovirus*, and *Virgaviridae*) detected in at least one of the ballast waters but not in any of the Lake Superior harbor waters suggested potential opportunities for non-native viruses to be discharged with ballast water into the Port of Duluth–Superior.

In this research, annotation-independent approaches enabled a more comprehensive comparison of viromes from the various lakes (Rodriguez–Brito et al., 2010; Roux et al., 2012). The Great Lakes is a single interconnected hydrologic system. However, each lake is unique due to size, topography, and land use impacts. In addition, nearshore and harbor waters are different from open waters in these lakes (as shown by different contig profiles of the two Lake Ontario viromes, BB and PB). It was somewhat unexpected that ballast water, AB, taken from the Port of Toledo located on Lake Erie showed an indistinguishable virome signature from the Lake Superior harbor waters (Figure 5.3A and 5.5B). Port of Toledo has been characterized as "the port of the greatest concern" for ballast water mediated invasions throughout the Great Lakes by U.S. EPA (U.S. EPA, 2008). The Port of Toledo received the second-largest amount of ballast water following the Port of Duluth-Superior from vessels whose sources were outside the Great Lakes (U.S. EPA, 2008). We hypothesize that this may have contributed to the indistinguishing profile of the AB virome from the Lake Superior harbor water viromes. Further water sampling and investigation of hydrographic characteristics are needed to get better insight into similarities between these geographically distinct samples.

This research added new virome data sets associated with freshwater

environments to the currently limited virome database. To date, many virome studies have focused on marine environments, and others have investigated viromes in human designed/managed freshwater environments (Rodriguez–Brito et al., 2010; Breitbart et al., 2002; Angly et al., 2006; Bench et al., 2007; Rosario et al., 2009; Williamson et al., 2012; Dinsdale et al., 2008; Rooks et al., 2010; Abbai et al., 2012). Virome studies of natural freshwater environments can be found in two extreme environments, an Antarctic lake and desert ponds, as well as in temperate freshwater lakes in France (Roux et al., 2012; Lopez–Bueno et al., 2009; Fancello et al., 2013). These previous studies examined either DNA or RNA viruses (but not both) and were limited to the lower sequencing yield of earlier technologies such as Roche 454 pyrosequencing.

This new work was the first comprehensive research of both DNA and RNA viruses originated from temperate freshwater lakes in the Great Lakes system. Comparison of the Great Lakes ballast and harbor water viromes with previously characterized aquatic viromes is important to determine how inclusion of an RNA virus fraction affects virome clustering. The comparison demonstrated that freshwater viromes are distinct from marine water viromes, providing evidence of hierarchical clustering according to salinity levels despite vast geographic distances. The difference between freshwater and marine water viromes presented in this research is consistent with previous studies (Roux et al., 2012; Logares et al., 2009). This is due to the dominance of dsDNA phages in our viromes, and thus inclusion of an RNA virus fraction did not significantly affect virome clustering. It is noteworthy that different approaches used to prepare and analyze viromes (e.g., sample preparation, sequencing platform, bioinformatics workflow) have the potential to undermine this comparison. Currently,

mid-ocean exchange of ballast water is widely used to comply with the IMO ballast water discharge guidelines (IMO). The underlying principle of this practice is to replace coastal water in ballast tanks with oceanic water. This can, however, introduce viruses associated with the marine water environment to freshwater environments such as the Great Lake basin. The impact of the transport of viruses between biomes on host populations should be further investigated.

5.4.2. Implications and control of viruses in ballast water

With the elevated public attention on the introduction and spread of invasive species, for example VHSV, which caused extensive losses of wild fish in the Great Lakes, this research examined major emerging viral pathogens of fish and shrimp using metagenomics approaches (Bain et al., 2010). Homology searches against existing databases tentatively identified three and four groups of viruses causing diseases in fish and shrimp, respectively. These viruses are listed by the World Organization for Animal Health (OIE) as causing notifiable diseases of fish and shrimp (Walker and Winton, 2010). Notably, KHV (formally classified as the Cyprinid herpesvirus 3) has appeared within the Great Lakes basin with multiple mortality events since 2004 (Grimmett et al., 2006; Garver et al., 2010; Cornwell et al., 2015). It causes diseases and mass mortality in common and koi carp (Cyprinus carpio and C. carpo koi, respectively) and has become the most dramatic example of an emerging disease of fish (Hedrick et al., 2002). The identification of these potential viral pathogens in ballast water is important in improving our understanding of what ballast water treatment would be needed to inactivate viruses in the future. In addition, the presence of viral pathogens with lower amino acid similarity

to known viruses means that these viruses are potentially novel or genetically diverse. Further investigations (e.g., phylogenetic approach, gene–specific PCR) are required to confirm the identification of these viral pathogens.

The diverse viral populations in ballast water and the movement of these viruses around the world have potential impacts on phytoplankton, animal health, and even human health (Ruiz et al., 2000; Drake et al., 2007; Altug et al., 2012). This reinforces the need for ballast water treatment for controlling potential viral invasion. Mid-ocean exchange of ballast water is an interim solution to control the introduction of aquatic invasive species. A few studies have reported that this practice is not effective in reducing the total number of bacteria and VLPs (Leichsenring and Lawrence, 2011; Ducklow, 2000). Over the past few years, special efforts have been made among the scientific and industrial communities to develop technologies for ballast water treatment because of the need for vessels to establish ballast water treatment systems onboard by 2016 according to the U.S.C.G (David and Gollasch, 2015). Several ballast water treatment systems such as filtration, deoxygenation, biocides, and ultraviolet treatment have been developed (Tsolaki and Diamadopoulos, 2010; Lloyd's Register Website). Unfortunately, these techniques have only been tested with marine water focusing on reducing the level of phytoplankton and bacteria. Thus, the effective- ness of ballast water treatment technologies in removing viruses in freshwater environments is currently unknown.

Metagenomics approaches have the potential to overcome the limitations of traditional methods for the detection and characterization of viruses such as cell culture and gene-specific PCR and, thus, provide the opportunity to explore composition and

taxonomic diversity of uncultured viruses. However, the metagenomic workflow from sample collection to bioinformatics analysis is experimentally and computationally challenging at each step, and potential biases may be introduced in estimating viral diversity (Beerenwinkel et al., 2012). On sampling, approximately 60 L of ballast water collected from a ballast tank from each vessel in this research may not be representative of the 5 million gallons of ballast water typically carried by a vessel (Carlton et al., 1995). A more intensive sampling design may be needed to aid in resolving the viral diversity associated with ballast waters. Additionally, the bioinformatics analysis is limited by a lack of viral reference genomes and the need to assemble the short reads for appropriate virome comparison and annotation. The paucity of viral reference databases affects the ability to identify viral pathogens and to do comparisons of the functional capacity of the viromes, when traits cannot be identified (Delwart, 2007; Rosario and Breitbart, 2011; Bibby, 2013). The assembly process also skews the analysis by including primarily the most abundant organisms. With large data sets, sequences that are rare do not assemble into contigs and are therefore not included in the contig analysis (Thomas et al., 2012). However, efforts to sequence more deeply than typical, as is done in this research, will help to address these issues.

The findings of the present research have several important implications. First, ballast and harbor waters originating from the Great Lakes harbored diverse viruses including viral pathogens associated with fish and shrimp (with low amino acid similarity), emphasizing the need for implementing ballast water discharge limits for viruses and treatment. Second, viromes were distinct among the Great Lakes and formed a specific group of temperate freshwater viromes but separated from viromes associated

with marine environments and engineered freshwater systems, suggesting the potential transfer and introduction of viruses between biomes and to the Great Lakes through ballast water discharge. Looking forward, the results of this research will assist in identifying potential viral invasions via ships' ballast water and evaluating ballast water quality and standards to protect public and ecosystem health from invasive species.

APPENDIX

Table 5.11. Accession number of the Illumina HiSeq sequencing data

Sample	Accession number	Release data
Great Lakes ballast and harbor waters under SRP048255		
AB	SRX717585	10−09−2014
BB	SRX717357	10−09−2014
BH	SRX717563	10−09−2014
IB	SRX717564	10−09−2014
IH	SRX717576	10−09−2014
MB	SRX717579	10−09−2014
MH	SRX717582	10−09−2014
PB	SRX717585	10−09−2014

REFERENCES

REFERENCES

1. Abbai, N. S.; Govender, A.; Shaik, R.; Pillay, B., Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. *Molecular Biotechnology* **2012,** 50 (1), 39−48.

2. ABS Website. Available at http://ww2.eagle.org/content/dam/eagle/publications/2014/BWTAdvisory14312rev3.pdf.

3. Altug, G.; Gurun, S.; Cardak, M.; Ciftci, P. S.; Kalkan, S., The occurrence of pathogenic bacteria in some ships' ballast water incoming from various marine regions to the Sea of Marmara, Turkey. *Marine Environmental Research* **2012,** 81, 35−42.

4. Angly, F.; Felts, B.; Breitbart, M.; Salamon, P.; Edwards, R.; Carlson, C.; Chan, A.; Haynes, M.; Kelley, S.; Liu, H.; Mahaffy, J.; Mueller, J.; Nulton, J.; Olson, R.; Parsons, R.; Rayhawk, S.; Suttle, C.; Rohwer, F., The marine viromes of four oceanic regions. *PLoS Biology* **2006,** 4 (11), 2121−2131.

5. Bain, M. B.; Cornwell, E. R.; Hope, K. M.; Eckerlin, G. E.; Casey, R. N.; Groocock, G. H.; Getchell, R. G.; Bowser, P. R.; Winton, J. R.; Batts, W. N.; Cangelosi, A.; Casey, J. W., Distribution of an Invasive Aquatic Pathogen (Viral Hemorrhagic Septicemia Virus) in the Great Lakes and Its Relationship to Shipping. *PLoS One* **2010,** 5 (4), e10156.

6. Beerenwinkel, N.; Günthard, H. F.; Roth, V.; Metzner, K. J., Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology* **2012,** 3, 329.

7. Bench, S. R.; Hanson, T. E.; Williamson, K. E.; Ghosh, D.; Radosovich, M.; Wang, K.; Wommack, K. E., Metagenomic character- ization of Chesapeake bay virioplankton. *Applied and Environmental Microbiology* **2007,** 73 (23), 7629−7641.

8. Bergh, O.; Borsheim, K.; Bratbak, G.; Heldal, M., High abundance of viruses found in aquatic environments. *Nature* **1989,** 340 (6233), 467−468.

9. Bibby, K., Metagenomic identification of viral pathogens. *Trends in Biotechnology* **2013,** 31 (5), 275−9.

10. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.; Segall, A.; Mead, D.; Azam, F.; Rohwer, F., Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* ***2002,*** 99 (22), 14250−14255.

11. Carlton, J. T.; Reid, D. M.; Van Leeuwen, H., Shipping study the role of shipping in

the introduction of nonindigenous aquatic organisms to the coastal waters of the United States (other than the Great Lakes) and an analysis of control options, National Technical Information Service Distributor, Washington, DC, **1995**.

12. Clark, B., Will Viral Hemorrhagic Septicemia (VHS) Be the Straw That Breaks the Camel's Back? The Balkanization of Great Lakes Ballast Water Law. *Minnesota Journal of International Law* **2009,** 18 (1), 227−264.

13. Cornwell, E. R.; Anderson, G. B.; Coleman, D.; Getchell, R. G.; Groocock, G. H.; Warg, J. V.; Cruz, A. M.; Casey, J. W.; Bain, M. B.; Bowser, P. R., Applying multi-scale occupancy models to infer host and site occupancy of an emerging viral fish pathogen in the Great Lakes. *Journal of Great Lakes Research* **2015,** 41 (2), 520−9.

14. David, M.; Gollasch, S., Global Maritime Transport and Ballast Water Management Issues and Solutions. In Invading Nature - Springer Series in Invasion Ecology 8, Springer, Dordrecht, The Netherlands, **2015**.

15. Delwart, E., Viral metagenomics. *Reviews in Medical Virology* **2007,** 17 (2), 115−131.

16. Dinsdale, E. A.; Edwards, R. A.; Hall, D.; Angly, F.; Breitbart, M.; Brulc, J. M.; Furlan, M.; Desnues, C.; Haynes, M.; Li, L.; McDaniel, L.; Moran, M. A.; Nelson, K. E.; Nilsson, C.; Olson, R.; Paul, J.; Brito, B. R.; Ruan, Y.; Swan, B. K.; Stevens, R.; Valentine, D. L.; Thurber, R. V.; Wegley, L.; White, B. A.; Rohwer, F., Functional metagenomic profiling of nine biomes. *Nature* **2008,** 452 (7187), 629−32.

17. Drake, L. A.; Choi, K. H.; Ruiz, G. M.; Dobbs, F. C., Global redistribution of bacterioplankton and virioplankton communities. *Biological Invasions* **2001,** 3 (2), 193−199.

18. Drake, L.; Doblin, M.; Dobbs, F., Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. *Marine Pollution Bulletin* **2007,** 55 (7−9), 333−341.

19. Drake, L. A.; Ruiz, G. M.; Galil, B. S.; Mullady, T. L.; Friedmann, D. O.; Dobbs, F. C., Microbial ecology of ballast water during a transoceanic voyage and the effects of open ocean exchange. *Marine Ecology Progress Series* **2002,** 233, 13−20.

20. Ducklow, H. W., Bacterial production and biomass in the oceans. In Microbial Ecology of the Oceans, Kirchman, D. L., Ed., Wiley-Liss, New York, **2000**.

21. Elsayed, E.; Faisal, M.; Thomas, M.; Whelan, G.; Batts, W.; Winton, J., Isolation of viral haemorrhagic septicaemia virus from muskellunge, Esox masquinongy (Mitchill), in Lake St Clair, Michigan, USA reveals a new sublineage of the North American genotype. *Journal of Fish Diseases* **2006,** 29 (10), 611−619.

22. European Food Safety Authority, Scientific Opinion of the Panel on Animal Health and Welfare on a request from the European Commission on possible vector species and live stages of susceptible species not transmitting disease as regards certain fish

diseases. *European Food Safety Authority Journal* **2007,** 584, 1−163.

23. Fancello, L.; Trape, S.; Robert, C.; Boyer, M.; Popgeorgiev, N.; Raoult, D.; Desnues, C., Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *Isme Journal* **2013,** 7 (2), 359−369.

24. Ferrate Treatment Technologies Website, Available at http://www.ferratetreatment.com/ballastwater.htm.

25. Fischer, M. G.; Allen, M. J.; Wilson, W. H.; Suttle, C. A., Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** 107 (45), 19508− 19513.

26. Fredricks, R., Aquatic nuisance species, manditory ballast water management, and alternative ballast water treatment standards to protect the marine environment. Submitted to U.S. commission on ocean policy, Boston, Massachusetts, **2002,** Available at http://govinfo.library.unt. edu/oceancommission/publicomment/northeastcomments/fredricks_ comment.pdf.

27. Garver, K. A.; Al-Hussinee, L.; Hawley, L. M.; Schroeder, T.; Edes, S.; LePage, V.; Contador, E.; Russell, S.; Lord, S.; Stevenson, R. M.; Souter, B.; Wright, E.; Lumsden, J. S., Mass mortality associated with koi herpesvirus in wild common carp in Canada. *Journal of Wildlife Diseases* **2010,** 46 (4), 1242−51.

28. Grimmett, S. G.; Warg, J. V.; Getchell, R. G.; Johnson, D. J.; Bowser, P. R., An unusual koi herpesvirus associated with a mortality event of common carp Cyprinus carpio in New York State, USA. *Journal of Wildlife Diseases* **2006,** 42 (3), 658−62.

29. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G., QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **2013,** 29 (8), 1072−1075.

30. Hammer, O.; Harper, D. A. T.; Ryan, P. D., PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **2001,** 4 (1), 1−9.

31. Hedrick, R. P.; Gilad, O.; Yun, S.; Spangenberg, J. V.; Marty, G. D.; Nordhausen, R. W.; Kebus, M. J.; Bercovier, H.; Eldar, A., A herpesvirus associated with mass mortality of juvenile and adult koi, a strain of common carp. *Journal of Aquatic Animal Health* **2000,** 12 (1), 44−57.

32. Hulo, C.; de Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P., ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research* **2011,** 39, D576-582.

33. International Marine Organization Website. Available at http://globallast. imo.org.

34. Leichsenring, J.; Lawrence, J., Effect of mid-oceanic ballast water exchange on virus-like particle abundance during two trans-Pacific voyages. *Marine Pollution Bulletin*

2011, 62 (5), 1103−1108.

35. Lloyd's Register Website. Available at http://www.lr.org/Images/BWTT_ June%202011_tcm155-222616.pdf.

36. Logares,R.;Bråte,J.;Bertilsson,S.;Clasen,J.L.;Shalchian- Tabrizi, K.; Rengefors, K., Infrequent marine-freshwater transitions in the microbial world. *Trends in Microbiology* **2009,** 17 (9), 414−22.

37. Lopez-Bueno, A.; Tamames, J.; Velazquez, D.; Moya, A.; Quesada, A.; Alcami, A., High Diversity of the Viral Community from an Antarctic Lake. *Science* **2009,** 326 (5954), 858−861.

38. Lumsden, J. S.; Morrison, B.; Yason, C.; Russell, S.; Young, K.; Yazdanpanah, A.; Huber, P.; Al-Hussinee, L.; Stone, D.; Way, K., Mortality event in freshwater drum Aplodinotus grunniens from Lake Ontario, Canada, associated with viral haemorrhagic septicemia virus, Type IV. *Diseases of Aquatic Organisms* **2007,** 76 (2), 99−111.

39. Minnesota Pollution Control Agency, Findings of fact-in the matter of a request for issuance of the SDS general permit MNG300000 for ballast water discharges from vessels transiting Minnesota state waters of Lake Superior, **2008,** Available at www.pca.state.mn.us/ index.php/view-document.html?gid=10739.

40. Mokili, J.; Rohwer, F.; Dutilh, B., Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2012,** 2 (1), 63−77.

41. National Research Council (U.S.), Committee on Ships' Ballast Operations, Stemming the tide: controlling introductions of nonindigenous species by ships' ballast water, National Academy Press, Washington, DC, **1996**.

42. Nissimov, J. I.; Worthy, C. A.; Rooks, P.; Napier, J. A.; Kimmance, S. A.; Henn, M. R.; Ogata, H.; Allen, M. J., Draft genome sequence of the coccolithovirus EhV-84. *Standards in Genomic Sciences* **2011,** 5 (1), 1−11.

43. Pagnucco, K. S.; Maynard, G. A.; Fera, S. A.; Yan, N. D.; Nalepa, T. F.; Ricciardi, A., The future of species invasions in the Great Lakes- St. Lawrence River basin. *Journal of Great Lakes Research* **2015,** 41, 139−149.

44. Rodriguez-Brito, B.; Li, L.; Wegley, L.; Furlan, M.; Angly, F.; Breitbart, M.; Buchanan, J.; Desnues, C.; Dinsdale, E.; Edwards, R.; Felts, B.; Haynes, M.; Liu, H.; Lipson, D.; Mahaffy, J.; Martin- Cuadrado, A.; Mira, A.; Nulton, J.; Pasic, L.; Rayhawk, S.; Rodriguez- Mueller, J.; Rodriguez-Valera, F.; Salamon, P.; Srinagesh, S.; Thingstad, T.; Tran, T.; Thurber, R.; Willner, D.; Youle, M.; Rohwer, F., Viral and microbial community dynamics in four aquatic environments. *Isme Journal* **2010,** 4 (6), 739−751.

45. Rohwer, F.; Edwards, R., The Phage Proteomic Tree: a genome- based taxonomy for

phage. *Journal of Bacteriology* **2002,** 184 (16), 4529−4535.

46. Rooks, D. J.; Smith, D. L.; McDonald, J. E.; Woodward, M. J.; McCarthy, A. J.; Allison, H. E., 454-pyrosequencing: a molecular battiscope for freshwater viral ecology. *Genes* **2010,** 1 (2), 210−26.

47. Rosario, K.; Breitbart, M., Exploring the viral world through metagenomics. *Current Opinion in Virology* **2011,** 1 (4), 289−97.

48. Rosario, K.; Nilsson, C.; Lim, Y.; Ruan, Y.; Breitbart, M., Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **2009,** 11 (11), 2806−2820.

49. Roux, S.; Enault, F.; Robin, A.; Ravet, V.; Personnic, S.; Theil, S.; Colombet, J.; Sime-Ngando, T.; Debroas, D., Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS One* **2012,** 7 (3), e33641.

50. Roux, S.; Tournayre, J.; Mahul, A.; Debroas, D.; Enault, F., Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **2014,** 15 (1), 76.

51. Ruiz, G.; Rawlings, T.; Dobbs, F.; Drake, L.; Mullady, T.; Huq, A.; Colwell, R., Global spread of microorganisms by ships - Ballast water discharged from vessels harbours a cocktail of potential pathogens. *Nature* **2000,** 408 (6808), 49−50.

52. Seiden, J. M.; Way, C. J.; Rivkin, R. B., Bacterial dynamics in ballast water during trans-oceanic voyages of bulk carriers: environ- mental controls. *Marine Ecology Progress Series* **2011,** 436, 145−159.

53. Sun, B.; Mouland, R.; Way, C.; Rivkin, R. B., Redistribution of heterotrophic prokaryotes through ballast water: A case study from the west coast of Canada. *Aquatic Invasions* **2010,** 5 (1), 5−11.

54. Thomas, T.; Gilbert, J.; Meyer, F., Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2012,** 2 (1), 3.

55. Tsolaki, E.; Diamadopoulos, E., Technologies for ballast water treatment: a review. *Journal of Chemical Technology and Biotechnology* **2010,** 85 (1), 19−32.

56. U.S. Environmental Protection Agency Website. Available at http://www. epa.gov/glnpo/basicinfo.html.

57. U.S. Environmental Protection Agency. Predicting future introductions of nonindigenous species to the Great Lakes; National Center for Environmental Assessment: Washington, DC, **2008,** Available at http:// cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=190305.

58. Walker, P. J.; Winton, J. R., Emerging viral diseases of fish and shrimp. *Veterinary*

Research **2010,** 41 (6), 51−75.

59. Wilhelm, S. W.; Carberry, M. J.; Eldridge, M. L.; Poorvin, L.; Saxton, M. A.; Doblin, M. A., Marine and freshwater cyanophages in a Laurentian Great Lake: Evidence from infectivity assays and molecular analyses of g20 genes. *Applied and Environmental Microbiology* **2006,** 72 (7), 4957− 4963.

60. Williamson, S.; Allen, L.; Lorenzi, H.; Fadrosh, D.; Brami, D.; Thiagarajan, M.; McCrow, J.; Tovchigrechko, A.; Yooseph, S.; Venter, J., Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS One* **2012,** 7 (10), e42047.

61. Willner, D.; Hugenholtz, P., From deep sequencing to viral tagging: Recent advances in viral metagenomics. *BioEssays* **2013,** 35 (5), 436−442.

62. Wommack, K. E.; Colwell, R. R., Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* **2000,** 64, 69−114.

CHAPTER 6

TRANSPORTING OCEAN VIROMES: INVASION OF THE AQUATIC BIOSPHERE

Abstract

Studies of marine viromes (viral metagenomes) have revealed that viruses are highly diverse and exhibit biogeographic patterns. A growth in global commerce and maritime traffic may accelerate spread of these diverse and non−cosmopolitan viruses from one part of the world to another. Here, metagenomic analyses demonstrated that ballast water moves around viromes (including viral pathogens) unique to geographic and environmental niches. Furthermore, the results from this research show that virus richness is governed by local environmental conditions and different viral groups have different responses to environmental variation. These results identify the ballast water as a contributing factor to virome transport and increased exposure of the aquatic bioshpere to viral invasion.

6.1. Introduction

Viruses are the most undiscovered and mysterious part of the biosphere. Their role as pathogenic entities is well recognized but the array of viral infections throughout the tree of life including archaea, bacteria, and eukaryotes is immense and we have only scratched the surface to reveal the global genetic diversity of viruses. This has limited our understanding of the ecological role of phages and other viral groups in biogeochemical cycling, as well as gene exchange (Wommack et al., 2015). Our knowledge of the viral predator−prey interactions is poor and viral life histories have not been well described. Viral−host specificity that was once considered a well−known biological principal is now being challenged, as the concept of plant viral infections of humans and other animals is being proposed (Balique et al., 2015). Recent global surveys of the ocean viral metagenome (virome), which focused mainly on DNA viruses infecting bacteria, have suggested that marine viruses, particularly phages are highly diverse and can exhibit distinctive biogeographic patterns (Angly et al., 2006; Brum et al., 2015). While these studies have revealed a diverse array of DNA phages in marine environments and that local environmental conditions play an important role in structuring their diversity, little is known about the diversity of RNA viruses and eukaryotic viruses in the oceans and their global transport and disease potential.

Oceanic and coastal anthropogenic pollution is growing in part as a function of global commerce and increasing maritime traffic. It is estimated that ocean−going cargo vessels transport as high as 12 billion tons of ballast water each year, transferring the aquatic life from one part of the world to another (GloBallast Partnerships, 2009). Global movement of non−indigenous species within ballast tanks across natural barriers has

threatened coastal ecosystem and biodiversity. The metazoan ballast invaders have been well studied and described since about the 1980s (Drake et al., 2007). However, the mechanisms of microbial invasions are still unclear despite the potential of microorganisms to influence the ecological functioning of biological communities and ecosystems at a global scale (Amalfitano et al., 2015). Ruiz et al. (2000) provided a hypothesis that the likelihood of invasions goes up with increasing inoculation concentration and that genetic diversity of the microbial component in ballast water including viruses must be examined to further understand the global transport of pathogens. More than a decade later, this call to improve our scientific knowledge has remained unanswered despite the advancement of metagenomics using high−throughput sequencing.

6.2. Materials and methods

6.2.1. Sample collection

A total of 14 samples were collected from the Port of Los Angeles/Long Beach (LA/LB), including 11 ballast waters and three surface harbor waters over a one−week period on March 2014 (Table 6.1). Access to the port was gained by California State Lands Commission, and the ballast water sampling was approved by the captains of vessels. The sampling was conducted under the supervision of ballast water inspectors as well as chief officers of vessels. Samples were transported to a lab in the Cabrillo Marine Aquarium in San Pedro, CA and processed within 12 hours of sample collection. An additional 10 samples were collected from the Port of Singapore, including five ballast

waters and five surface harbor waters over a two−week period on May 2014. Access to the port was gained by Port of Singapore Authority, and the ballast water sampling was approved by an anonymous shipping company and by the captains of vessels. The sampling was conducted under the supervision of the chief officers of vessels. Samples were transported to a lab in National University of Singapore (NUS), Singapore and processed within 12 hours of sample collection. Type of vessels whose ballast waters were sampled included container ship (8), bulk carrier (3), tanker ship (1), car carrier (1), cruise ship (1), and refrigerated cargo carrier (1). For sample collection, ballast waters were sampled mainly through ballast tank manholes (14 samples). When an access to ballast tank manholes was not available, samples were collected via ballast water pipelines (two samples). Prefix ‘C’ and ‘S’ were used to differentiate samples collected from the Port of LA/LB and the Port of Singapore, respectively (e.g. ‘CADO’ or ‘SCB’). Table 6.2 summarized engineered, management, and environmental variables of the ballast and harbor water samples.

Table 6.1. Summary of sampling information

Sample	Collection date	Sampling location	Sample type	Vessel type	Sampling method
CADO	03/01/14	Port of LA/LB	Ballast water	Bulk carrier	Manhole
CASC	03/04/14	Port of LA/LB	Ballast water	Bulk carrier	Manhole
CATL	03/02/14	Port of LA/LB	Ballast water	Container ship	Manhole
CBAL	03/06/14	Port of LA/LB	Ballast water	Refrigerated cargo carrier	Manhole
CCAR	03/03/14	Port of LA/LB	Ballast water	Cruise ship	Ballast water pipeline
CCEB	03/07/14	Port of LA/LB	Ballast water	Bulk carrier	Ballast water pipeline
CCOS	03/07/14	Port of LA/LB	Ballast water	Container ship	Manhole
CLIB	03/02/14	Port of LA/LB	Ballast water	Container ship	Manhole
CNAD	03/05/14	Port of LA/LB	Ballast water	Car carrier	Manhole
CSAG	03/05/14	Port of LA/LB	Ballast water	Cargo ship	Manhole
CTUL	03/04/14	Port of LA/LB	Ballast water	Tanker ship	Manhole
CILB	03/05/14	Port of LA/LB	Harbor water	NA	Bucket
COLB	03/03/14	Port of LA/LB	Harbor water	NA	Bucket
CWLA	03/04/14	Port of LA/LB	Harbor water	NA	Bucket
SCB	05/14/14	Port of Singapore	Ballast water	Container ship	Manhole
SGB	05/21/14	Port of Singapore	Ballast water	Container ship	Manhole
SMB	05/25/14	Port of Singapore	Ballast water	Container ship	Manhole
SQB	05/29/14	Port of Singapore	Ballast water	Container ship	Manhole
SRB	05/18/14	Port of Singapore	Ballast water	Container ship	Manhole
SCH	05/14/14	Port of Singapore	Harbor water	NA	Bucket

Table 6.1 (cont'd)

Sample	Collection date	Sampling location	Sample type	Vessel type	Sampling method
SGH	05/21/14	Port of Singapore	Harbor water	NA	Bucket
SMH	05/25/14	Port of Singapore	Harbor water	NA	Bucket
SQH	05/29/14	Port of Singapore	Harbor water	NA	Bucket
SRH	05/18/14	Port of Singapore	Harbor water	NA	Bucket

Abbreviations: LA/LB, Los Angeles/Long Beach; NA, information not applicable.

Table 6.2. Summary of engineered, management, and environmental variables

	Engineered and management variables						Environmental variables			
Sample	BWE date	BWE latitude (degrees North)	BWE longitude (degrees East)	Distance from shoreline (kilometer)	Distance from shoreline (nautical mile)	Time in ballast tank (day)	pH	Salinity (ppt)	Temperature (ºC)	Turbidity (NTU)
CADO	02/17/14	46.00	-168.00	752.24	406.17	12	7.8	32.2	20.0	0.2
CASC	02/19/14	30.00	163.00	1096.54	592.08	13	8.0	34.7	21.3	0.2
CATL	01/14/14	30.31	-140.93	1837.30	992.06	47	7.0	32.9	15.4	0.2
CBAL	03/01/14	13.30	-104.80	569.87	307.70	5	8.0	32.9	17.7	0.1
CCAR	03/01/14	31.00	-117.00	62.82	33.92	2	7.2	13.7	20.9	5.3
CCEB	02/23/14	36.75	175.83	1076.92	581.49	12	7.9	14.0	21.8	11.0
CCOS	02/24/14	37.14	151.09	748.93	404.39	11	7.9	32.5	17.5	0.5
CLIB	01/22/14	NA	NA	0.00	0.00	39	7.7	33.0	15.2	6.8
CNAD	02/06/14	49.48	-147.41	892.23	481.76	27	7.7	32.4	16.2	0.6
CSAG	03/03/14	38.00	-124.00	85.21	46.01	2	7.8	29.8	16.1	0.3
CTUL	02/04/14	26.65	-121.30	384.89	207.83	28	7.8	14.7	18.4	0.1
CILB	NA	NA	NA	0.00	0.00	0	7.8	31.9	17.3	0.4
COLB	NA	NA	NA	0.00	0.00	0	8.0	12.6	17.4	0.5
CWLA	NA	NA	NA	0.00	0.00	0	7.9	29.7	18.3	0.6
SCB	05/10/14	20.29	113.89	171.09	92.38	4	6.9	14.0	29.9	0.0
SGB	02/03/14	12.50	47.08	109.86	59.32	107	7.9	20.7	30.0	0.0
SMB	05/14/14	34.54	122.14	171.71	92.72	11	7.7	18.9	29.8	4.0

Table 6.2 (cont'd)

	Engineered and management variables						Environmental variables			
Sample	BWE date	BWE latitude (degrees North)	BWE longitude (degrees East)	Distance from shoreline (kilometer)	Distance from shoreline (nautical mile)	Time in ballast tank (day)	pH	Salinity (ppt)	Temperature (ºC)	Turbidity (NTU)
SQB	05/05/14	22.54	116.31	37.78	20.40	24	7.9	20.0	29.9	3.0
SRB	05/17/14	3.72	105.41	68.42	36.94	1	7.5	32.9	30.1	0.0
SCH	NA	NA	NA	0.00	0.00	0	7.6	11.8	30.3	10.0
SGH	NA	NA	NA	0.00	0.00	0	6.0	27.6	30.5	10.0
SMH	NA	NA	NA	0.00	0.00	0	7.9	30.9	30.2	0.0
SQH	NA	NA	NA	0.00	0.00	0	7.9	31.2	30.9	5.0
SRH	NA	NA	NA	0.00	0.00	0	8.0	30.6	30.2	3.0

Ballast water exchange was not performed on the CLIB sample. Thus, most recent ballast water uptake date was used to calculate duration of water in the ballast tank.

Abbreviations: BWE, ballast water exchange; NA, information not applicable.

6.2.2. Preparation and sequencing of viromes

Ballast and harbor water samples were stored at 4 °C and processed for tangential flow filtration (TFF) within 12 hours of sample collection. For samples collected from the Port of LA/LB, the filtration concentrates (approximately 300−500 mL) were transported overnight to Michigan State University (MSU) at 4 °C. For samples collected from the Port of Singapore, the PEG concentrates (20 mL) were transported to MSU at 4 °C with the import permit approved by United States Centers for Disease Control and Prevention (U.S. CDC). Samples were stored immediately at –80 °C upon arrival for further processing. Details on virome preparation, including concentration and purification of viral particles and extraction and amplification of viral nucleic acids, and metagenomic sequencing of ballast and harbor waters were described in the Chapter 4.

6.2.3. Analysis of viromes

Preprocessing of raw sequence reads, *de novo* assembly of sequence reads, and taxonomic classification of contiguous reads (contigs) were performed as previously described in the Chapter 4.

To investigate emerging viral pathogens, contigs most similar to viruses infecting human, fish, and shrimp were extracted from the data sets. These contigs were again BLASTX−searched ($E < 10^{-3}$) against the National Center for Biotechnology Information (NCBI) non−redundant (nr) database (downloaded in April 2014) and any contigs more similar to other proteins were excluded.

The summary of viral taxonomic classification was tabulated as a 72 × 83 matrix (72 viromes in rows × 83 taxonomic groups in columns). This matrix was used for

statistical analyses performed by both the vegan package (Oksanen et al., 2013) in R Statistics Environment (R Development Core Team, 2010) and PAST statistical package (Hammer et al., 2001). Variation in virome composition of 72 samples was determined using Bray−Curtis similairty matrix and a Principal Coordinates Analysis (PCoA) plot. To test for statistically significant differences between prior groupings of the samples made according to geographic origins, one−way Analysis of Similarities (ANOSIM, with 9999 permutations) was carried out. Similarity Percentages (SIMPER) analysis was further performed to identify discriminating taxonomic groups by comparing relative abundances of viral families in the prior groupings. Virus richness was calculated by counting a total number of viral families identified in each data set. To compare differences in virus richness between ballast and harbor waters, Bonferroni−corrected pairwise t−tests were conducted, when statistical differences were found. Spearman's correlation coefficient was computed to examine relationships between viral families and geographical locations and virus richness and variables.

6.2.4. Data access

All Illumina sequencing reads from viromes of the ballast and harbor waters collected from the Port of LA/LB and the Port of Singapore are available in the NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under accession number SRP061842.

6.3. Results and discussion

6.3.1. Influence of global shipping on transport of the ocean virome

By examining variation in virome composition of ballast and harbor waters between geographic locations, a hypothesis that the movement of ballast water across the global shipping network transports the ocean virome was tested. In this regard, viral communities in 24 ocean−captured ballast and harbor waters were explored at two distinct geographic locations, the Port of LA/LB and the Port of Singapore, among the world's busiest container ports (Figure 6.1). A potential bias in virome preparation was minimized by generating three technical replicates for each sample, which contained concentrated and purified viral particles. The resulting 72 ballast and harbor water virome data sets comprised 3.8 billion 100−base pair (bp) paired−end Illumina reads with an average of 52.2 ± 30.9 (mean ± s.d.) million reads (Table 6.3). The virome data sets captured genomes of both DNA and RNA viruses present in ballast and harbor waters.

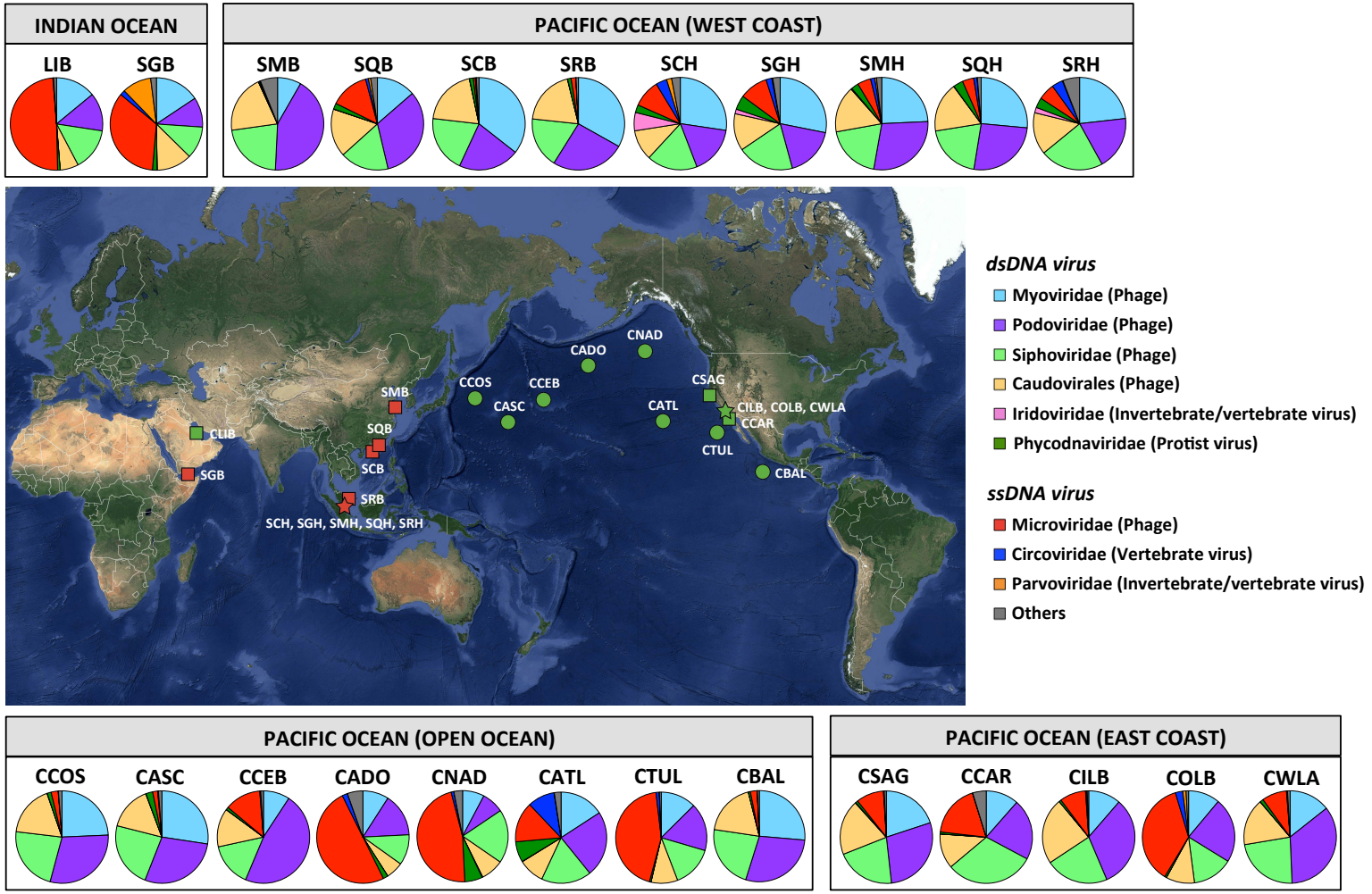


Figure 6.1. Relative distribution of viromes from ballast and harbor waters. Pie charts represent a mean relative abundance of assigned viral families (three replicates from 24 samples). 'Others' are viral families whose maximum relative abundances across

Figure 6.1 (cont'd)

viromes are less than 3% (including RNA viruses). Vessels with ballast waters arriving in the Port of Los Angeles/Long Beach are shown as a green star and the Port of Singapore as a red star. Circles and squares in the map indicate ballast waters exchanged beyond and within 200 nautical miles from nearest shoreline, respectively.

Abbreviations: ds, double−stranded; ss, single−stranded.

Table 6.3. Overview of the sequence reads and the assembled contigs of the virome libraries

Sample	# Raw reads	# Trimmed reads	# Contigs	Mean contig length (bp)	Max contig length (bp)	Min contig length (bp)	Mapping rate (%)	# Reads assigned	% Reads assigned
CADO1	14,607,432	13,308,260	19,513	678	28,244	200	80.38	5,316	27.20
CADO2	47,497,818	45,803,187	63,987	700	33,280	200	89.06	17,854	27.90
CADO3	37,570,710	36,333,456	63,975	723	21,784	200	91.48	18,480	28.90
CASC1	37,551,690	36,160,661	61,390	795	24,693	200	95.08	19,361	31.50
CASC2	38,253,232	36,985,728	58,188	759	21,232	200	94.38	18,082	31.10
CASC3	37,533,220	36,297,181	63,907	769	18,835	200	93.91	19,988	31.30
CATL1	34,359,206	33,335,991	12,672	823	10,731	200	91.44	4,270	33.70
CATL2	31,784,446	30,702,894	21,214	814	19,080	200	90.85	6,978	32.90
CATL3	35,312,320	33,989,357	12,263	817	11,481	200	83.44	4,039	32.90
CBAL1	37,209,228	35,726,700	146,395	705	50,141	200	78.80	44,535	30.40
CBAL2	35,756,178	34,286,197	156,225	692	55,044	200	78.05	45,968	29.40
CBAL3	33,272,332	31,960,090	157,594	703	71,291	200	78.71	47,174	29.90
CCAR1	29,568,966	28,146,134	78,108	742	91,506	200	80.18	22,378	28.70
CCAR2	29,479,562	28,261,709	80,919	745	66,715	200	81.65	23,554	29.10
CCAR3	31,327,906	30,282,791	83,828	744	75,417	201	81.77	24,373	29.10
CCEB1	23,248,862	22,587,076	65,591	638	49,484	200	85.32	19,718	30.10
CCEB2	24,140,902	23,369,141	54,958	634	70,129	200	85.29	15,627	28.40
CCEB3	24,952,210	24,212,835	62,155	634	48,055	200	85.52	17,982	28.90
CCOS1	28,052,426	27,260,459	115,461	620	35,015	200	70.31	30,171	26.10

Table 6.3 (cont'd)

Sample	# Raw reads	# Trimmed reads	# Contigs	Mean contig length (bp)	Max contig length (bp)	Min contig length (bp)	Mapping rate (%)	# Reads assigned	% Reads assigned
CCOS2	30,236,070	29,432,911	128,422	641	42,054	200	79.81	35,251	27.40
CCOS3	33,362,054	32,442,080	160,803	642	42,054	200	77.40	43,934	27.30
CLIB1	36,164,890	35,225,792	53,227	682	11,235	200	91.22	15,088	28.30
CLIB2	35,261,898	34,296,086	52,785	671	18,623	200	90.13	14,549	27.60
CLIB3	32,347,902	31,437,934	53,247	682	24,960	200	91.82	14,898	28.00
CNAD1	39,500,638	38,447,724	15,193	766	39,830	200	92.76	4,587	30.20
CNAD2	37,560,740	36,158,873	29,814	767	39,830	200	92.13	9,296	31.20
CNAD3	36,205,340	34,849,088	28,260	750	39,830	200	92.85	8,667	30.70
CSAG1	38,187,976	36,775,949	149,942	662	30,279	200	79.48	43,958	29.30
CSAG2	31,009,228	29,787,151	116,100	671	41,664	200	76.14	34,240	29.50
CSAG3	31,064,530	29,717,774	114,531	658	35,820	200	72.39	32,244	28.20
CTUL1	31,274,876	29,920,546	76,282	675	40,358	200	80.98	23,201	30.40
CTUL2	31,226,186	29,939,481	71,454	665	37,725	200	85.17	21,602	30.20
CTUL3	44,474,860	42,640,202	105,978	700	44,058	200	81.70	32,343	30.50
CILB1	30,570,756	29,606,336	83,196	699	44,653	200	73.03	29,250	35.20
CILB2	32,573,746	31,458,234	83,107	701	48,857	200	74.38	29,252	35.20
CILB3	28,251,066	27,365,531	83,317	712	47,827	201	73.19	29,418	35.30
COLB1	39,244,856	37,812,202	87,275	682	32,441	200	82.81	28,448	32.60
COLB2	34,613,158	33,229,277	79,395	696	39,135	200	79.20	25,846	32.60

Table 6.3 (cont'd)

Sample	# Raw reads	# Trimmed reads	# Contigs	Mean contig length (bp)	Max contig length (bp)	Min contig length (bp)	Mapping rate (%)	# Reads assigned	% Reads assigned
COLB3	34,245,908	32,779,113	82,806	711	41,163	200	79.02	26,518	32.00
CWLA1	34,460,992	32,955,955	97,182	689	35,449	201	68.98	33,460	34.40
CWLA2	26,397,776	25,308,773	84,053	690	33,121	201	68.90	29,284	34.80
CWLA3	30,014,418	28,826,319	84,891	677	38,550	202	68.74	29,448	34.70
SCB1	126,128,094	117,409,118	357,370	670	53,404	200	78.75	137,806	38.60
SCB2	74,146,924	66,122,820	230,004	647	40,022	200	69.38	85,258	37.10
SCB3	84,646,886	76,295,549	254,294	645	36,716	200	70.41	93,643	36.80
SGB1	97,607,434	84,370,686	137,212	676	87,302	200	71.83	39,229	28.60
SGB2	96,637,490	82,341,630	118,364	662	71,340	200	71.80	33,412	28.20
SGB3	71,187,728	62,579,553	109,122	660	71,807	200	70.16	30,797	28.20
SMB1	126,044,402	114,700,024	35,482	822	73,029	200	86.62	11,795	33.20
SMB2	60,925,126	50,499,522	21,333	871	29,874	201	89.80	7,522	35.30
SMB3	67,434,564	54,949,625	19,601	857	38,813	200	90.41	6,823	34.80
SQB1	81,591,554	69,220,915	88,691	680	48,079	200	78.79	25,224	28.40
SQB2	76,326,568	63,574,755	94,551	686	64,257	200	80.40	27,154	28.70
SQB3	84,313,476	73,836,039	108,913	685	69,918	200	82.88	31,313	28.80
SRB1	93,066,422	80,395,102	163,997	656	77,305	200	75.72	59,914	36.50
SRB2	70,083,208	60,349,646	147,261	638	108,071	200	76.75	51,418	34.90
SRB3	92,558,750	77,608,795	143,923	644	80,621	200	80.00	50,471	35.10

Table 6.3 (cont'd)

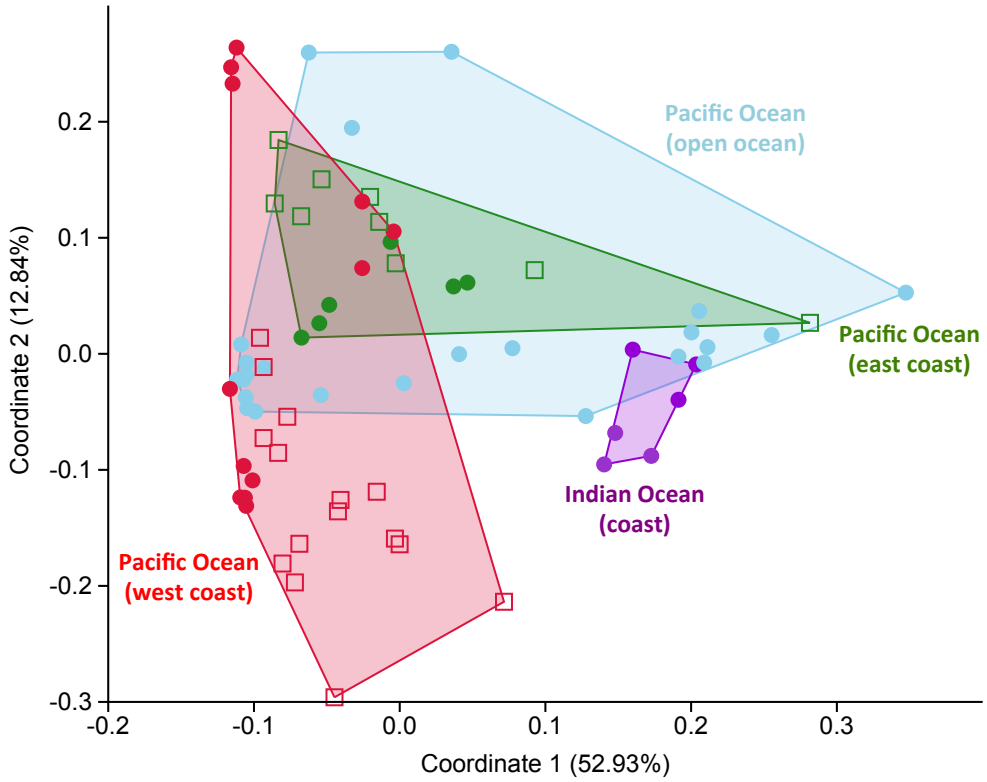
Sample	# Raw reads	# Trimmed reads	# Contigs	Mean contig length (bp)	Max contig length (bp)	Min contig length (bp)	Mapping rate (%)	# Reads assigned	% Reads assigned
SCH1	93,570,936	84,327,180	144,980	658	47,495	200	75.20	41,042	28.30
SCH2	80,664,836	71,032,395	118,896	709	43,653	200	79.50	32,931	27.70
SCH3	65,941,410	54,646,377	83,693	711	24,541	201	72.75	22,272	26.60
SGH1	66,233,144	57,815,171	153,994	665	34,070	200	67.46	49,218	32.00
SGH2	37,860,268	32,434,851	83,473	634	31,217	203	70.25	23,305	27.90
SGH3	45,656,852	38,072,127	90,330	631	22,488	201	71.79	25,386	28.10
SMH1	48,017,348	38,808,697	87,102	665	34,097	200	70.30	22,753	26.10
SMH2	56,983,010	48,424,911	80,164	662	33,235	200	72.85	20,749	25.90
SMH3	38,090,032	31,883,992	75,976	659	39,915	203	71.21	20,002	26.30
SQH1	50,596,026	39,013,324	95,734	641	29,520	200	65.47	30,474	31.80
SQH2	45,563,044	35,354,084	94,034	650	52,577	200	63.68	29,884	31.80
SQH3	99,916,222	83,470,829	190,810	655	40,333	200	73.44	61,299	32.10
SRH1	54,374,024	46,113,360	99,609	687	44,903	200	75.64	26,183	26.30
SRH2	195,662,920	174,705,545	165,569	713	44,904	200	84.33	46,082	27.80
SRH3	89,281,486	74,004,799	115,618	699	31,717	200	82.78	30,869	26.70

Here, the focus of research was narrowed to taxonomically describable viruses in ballast and harbor waters. To increase the probability of obtaining a significant similarity with reference sequences in the NCBI viral database, 3.4 billion high quality reads of the 72 samples were assembled, generating a total of 7.0 million contigs with an average of 97,357 ± 57,922 contigs with a mean length of 696.7 bp. As reported in other virome studies of marine environment (Angly et al., 2006; Hurwitz and Sullivan, 2013; Martinez et al., 2014; Winter et al., 2014), but not limited to, BLASTX searches ($E < 10^{-5}$) against the reference sequences revealed the enormous genetic diversity of viruses in the oceans, which cannot be uncovered using publicly available sequence database. Among the contigs homologous to known viruses (30.6 ± 0.03%), the majority was associated with double−stranded (ds) DNA phages (*Myoviridae*, 18.8 ± 8.4%; *Podoviridae*, 24.6 ± 9.5%; *Siphoviridae*, 19.1 ± 4.4%; and unclassified *Caudovirales*, 14.4 ± 4.7%) followed by single−stranded (ss) DNA phage, *Microviridae* (16.3% ± 17.0%). Along with phages, viruses infecting a broad range of hosts, including archaea, fungi, invertebrate, plant, protist, and vertebrate were present at different abundances in the viromes.

To explain variation in virome composition between geographic locations, all 72 samples were visualized with a PCoA plot (Figure 6.2A). Most of variance (65.8%) between different geographic origin was explained in this analysis, where the presence of distinct viromes in specific ocean realms was found. The significance of this difference was demonstrated by ANOSIM ($R = 0.233$, $p < 0.001$) and low ANOSIM R−value was associated with indistinct separation of ballast water samples originating from open Pacific Ocean from the other clusters (Figure 6.2B). This further suggested that marine viromes are not structured only by geographic patterns but also by local environmental

conditions as reported by a recent study (Brum et al., 2015). Pairwise comparisons showed that viromes of eastern Pacific Ocean along the west coast of America were separated from those of either Indian Ocean ($R = 0.691$, $p < 0.001$) or western Pacific Ocean bordering Eastern Asia ($R = 0.278$, $p < 0.001$), while this separation was not observed with those of open Pacific Ocean.

A



B

Overall R = 0.233 (p < 0.001)			
Sample type	**Pacific Ocean (open ocean, n = 24)**	**Pacific Ocean (east coast, n = 15)**	**Indian Ocean (coast, n = 6)**
Pacific Ocean (east coast, n = 15)	R = - 0.004 (p = 0.427)		
Indian Ocean (coast, n = 6)	R = 0.092 (p = 0.118)	**R = 0.691 (p < 0.001)**	
Pacific Ocean (west coast, n = 27)	**R = 0.143 (p < 0.01)**	**R = 0.278 (p < 0.001)**	**R = 0.914 (p < 0.001)**

Figure 6.2. Influence of geography on virome composition. 72 virome data sets were compared with each other using the Bray−Curtis similarity matrix based on relative abundance of viral families. **A,** Principal Coordinates Analysis (PCoA) plot presenting the difference in the virome composition. Convex hulls were used to group observations by ocean. Closed and open symbols represent ballast and harbor waters, respectively. **B,** Analysis of similarity (ANOSIM) result to identify the difference in the virome composition. Bold text indicates a significant difference between ocean viromes.

Next, it was identified that ssDNA phage, *Microviridae* (32.3%) and dsDNA phages, *Podoviridae* (18.1%) and *Myoviridae* (16.0%) contributed most to the virome dissimilarity between geographic origins (Table 6.4). Correlation analyses between these phage groups and geographical variation revealed that *Myoviridae* had the strongest relationship with geographic location followed by *Microviridae* (Figure 6.3). Relative abundance of *Myoviridae* had a highly significant negative correlation with latitude ($R = -0.671$, $p < 0.0001$) and a positive correlation with longitude ($R = 0.484$, $p < 0.0001$). In contrast to the *Myoviridae*, response of *Microviridae* to geographical variation demonstrated a positive correlation with latitude ($R = 0.387$, $p < 0.001$) and a negative correlation with longitude ($R = -0.476$, $p < 0.0001$). Unlike these two phage families, *Podoviridae* had a weak correlation only with longitude ($R = 0.281$, $p < 0.05$), suggesting that each viral family has different specificity to geographic location. Thus, specific viral families may have unique geographic and environmental niches and these relationships may be masked if better resolution of the genomic diversity is not ascertained.

Table 6.4. Summary of Similarity Percentage (SIMPER) analysis

Viral family	Average dissimilarity	Contribution %	Cumulative %	Mean abundance			
				Pacific Ocean (open ocean)	Pacific Ocean (East coast)	Indian Ocean	Pacific Ocean (West coast)
Microviridae	10.37000	32.30000	32.30000	21.80000	16.50000	41.90000	5.58000
Podoviridae	5.79400	18.05000	50.36000	24.70000	27.90000	11.90000	25.60000
Myoviridae	5.14700	16.04000	66.39000	16.60000	13.60000	14.80000	24.50000
Caudovirales	3.19500	9.95400	76.35000	12.30000	16.20000	9.40000	16.50000
Siphoviridae	3.09400	9.64100	85.99000	18.20000	22.20000	13.10000	19.50000
Phycodnaviridae	0.96160	2.99600	88.98000	2.65000	0.81600	1.25000	2.47000
Circoviridae	0.93820	2.92300	91.91000	1.78000	0.69300	0.99800	1.57000
Parvoviridae	0.72060	2.24500	94.15000	0.03700	0.31300	5.13000	0.42300
Iridoviridae	0.39860	1.24200	95.39000	0.09490	0.07780	0.02530	1.19000
Mimiviridae	0.26280	0.81870	96.21000	0.59100	0.23400	0.26000	0.81000
Poxviridae	0.24470	0.76250	96.98000	0.61500	0.18300	0.12100	0.22000
Geminiviridae	0.22880	0.71300	97.69000	0.33800	0.17500	0.63400	0.35300
Bacilladnavirus	0.18450	0.57500	98.26000	0.00138	0.77400	0.00050	0.00647
Nanoviridae	0.13950	0.43470	98.70000	0.08490	0.03620	0.03660	0.36400
Reoviridae	0.05035	0.15690	98.86000	0.01990	0.00043	0.00252	0.13400
Inoviridae	0.04674	0.14560	99.00000	0.01190	0.00366	0.00668	0.13200
Tombusviridae	0.04413	0.13750	99.14000	0.00000	0.00309	0.00005	0.13000
Tectiviridae	0.04138	0.12890	99.27000	0.02160	0.00454	0.29900	0.01740

Table 6.4 (cont'd)

Viral family	Average dissimilarity	Contribution %	Cumulative %	Mean abundance			
				Pacific Ocean (open ocean)	Pacific Ocean (East coast)	Indian Ocean	Pacific Ocean (West coast)
Virgaviridae	0.02625	0.08179	99.35000	0.00002	0.00094	0.00004	0.07780
Marseilleviridae	0.02376	0.07404	99.42000	0.04410	0.01910	0.01910	0.07280
Herpesviridae	0.02352	0.07329	99.50000	0.08980	0.05780	0.05670	0.05930
Ourmiavirus	0.01858	0.05790	99.55000	0.00013	0.00004	0.00025	0.05520
Nudiviridae	0.01739	0.05418	99.61000	0.01020	0.00127	0.00053	0.04630
Picobirnaviridae	0.01739	0.05418	99.66000	0.02430	0.00016	0.00511	0.03320
Totiviridae	0.01569	0.04887	99.71000	0.00921	0.00004	0.00521	0.03880
Picornavirales	0.01157	0.03606	99.75000	0.00025	0.00000	0.00014	0.03440
Baculoviridae	0.00944	0.02941	99.78000	0.01590	0.01520	0.01000	0.02630
Alloherpesviridae	0.00760	0.02369	99.80000	0.01590	0.01200	0.02410	0.01310
Polydnaviridae	0.00720	0.02242	99.82000	0.00488	0.00618	0.00035	0.01640
Cystoviridae	0.00648	0.02018	99.84000	0.00162	0.00010	0.00000	0.01790
Nodaviridae	0.00455	0.01419	99.86000	0.00000	0.00011	0.00000	0.01350
Hytrosaviridae	0.00394	0.01228	99.87000	0.00161	0.00506	0.00091	0.00962
Herpesvirales	0.00383	0.01193	99.88000	0.00697	0.00591	0.00945	0.00856
Ascoviridae	0.00333	0.01036	99.89000	0.00776	0.00012	0.00227	0.00301
Dicistroviridae	0.00330	0.01029	99.90000	0.00033	0.00001	0.00072	0.00955
Sobemovirus	0.00328	0.01023	99.91000	0.00000	0.00016	0.00010	0.00966

Table 6.4 (cont'd)

Viral family	Average dissimilarity	Contribution %	Cumulative %	Mean abundance			
				Pacific Ocean (open ocean)	Pacific Ocean (East coast)	Indian Ocean	Pacific Ocean (West coast)
Asfarviridae	0.00326	0.01015	99.92000	0.00892	0.00000	0.00029	0.00149
Nimaviridae	0.00296	0.00922	99.93000	0.00265	0.00782	0.00249	0.00263
Birnaviridae	0.00280	0.00873	99.94000	0.00005	0.00000	0.00062	0.00816
Bicaudaviridae	0.00260	0.00810	99.95000	0.00365	0.00420	0.00642	0.00154
Bidnaviridae	0.00240	0.00748	99.96000	0.00619	0.00111	0.00021	0.00140
Salterprovirus	0.00216	0.00673	99.96000	0.00632	0.00002	0.00145	0.00044
Adenoviridae	0.00187	0.00583	99.97000	0.00245	0.00261	0.00094	0.00190
Lipothrixviridae	0.00187	0.00582	99.98000	0.00269	0.00191	0.00048	0.00374
Ampullaviridae	0.00096	0.00299	99.98000	0.00002	0.00068	0.00022	0.00244
Potyviridae	0.00085	0.00265	99.98000	0.00015	0.00000	0.00000	0.00243
Alvernaviridae	0.00080	0.00250	99.98000	0.00008	0.00000	0.00050	0.00222
Picornaviridae	0.00075	0.00235	99.99000	0.00000	0.00002	0.00000	0.00224
Astroviridae	0.00068	0.00212	99.99000	0.00000	0.00000	0.00000	0.00203
Caulimoviridae	0.00062	0.00192	99.99000	0.00160	0.00002	0.00020	0.00031
Closteroviridae	0.00048	0.00149	99.99000	0.00139	0.00012	0.00000	0.00003
Tymoviridae	0.00038	0.00118	99.99000	0.00002	0.00000	0.00000	0.00111
Plasmaviridae	0.00025	0.00077	99.99000	0.00026	0.00004	0.00083	0.00028
Polyomaviridae	0.00024	0.00076	99.99000	0.00000	0.00000	0.00002	0.00072

Table 6.4 (cont'd)

Viral family	Average dissimilarity	Contribution %	Cumulative %	Mean abundance			
				Pacific Ocean (open ocean)	Pacific Ocean (East coast)	Indian Ocean	Pacific Ocean (West coast)
Rudiviridae	0.00023	0.00072	99.99000	0.00003	0.00006	0.00013	0.00061
Alphatetraviridae	0.00020	0.00063	100.00000	0.00000	0.00000	0.00000	0.00060
Permutotetraviridae	0.00020	0.00062	100.00000	0.00000	0.00000	0.00000	0.00059
Corticoviridae	0.00020	0.00061	100.00000	0.00011	0.00016	0.00000	0.00041
Marnaviridae	0.00017	0.00054	100.00000	0.00000	0.00000	0.00000	0.00052
Partitiviridae	0.00016	0.00049	100.00000	0.00030	0.00000	0.00002	0.00021
Retroviridae	0.00013	0.00040	100.00000	0.00028	0.00000	0.00011	0.00012
Hepeviridae	0.00012	0.00038	100.00000	0.00000	0.00000	0.00000	0.00036
Secoviridae	0.00012	0.00037	100.00000	0.00000	0.00002	0.00000	0.00035
Umbravirus	0.00008	0.00025	100.00000	0.00000	0.00000	0.00000	0.00024
Alphaflexiviridae	0.00008	0.00025	100.00000	0.00004	0.00000	0.00000	0.00020
Chrysoviridae	0.00007	0.00022	100.00000	0.00003	0.00000	0.00000	0.00019
Caliciviridae	0.00003	0.00010	100.00000	0.00000	0.00000	0.00000	0.00010
Papillomaviridae	0.00003	0.00008	100.00000	0.00002	0.00000	0.00000	0.00006
Bromoviridae	0.00003	0.00008	100.00000	0.00000	0.00000	0.00000	0.00008
Iflaviridae	0.00002	0.00008	100.00000	0.00000	0.00000	0.00000	0.00007
Luteoviridae	0.00002	0.00008	100.00000	0.00000	0.00000	0.00000	0.00007
Carmotetraviridae	0.00002	0.00006	100.00000	0.00000	0.00000	0.00000	0.00006

Table 6.4 (cont'd)

Viral family	Average dissimilarity	Contribution %	Cumulative %	Mean abundance			
				Pacific Ocean (open ocean)	Pacific Ocean (East coast)	Indian Ocean	Pacific Ocean (West coast)
Coronaviridae	0.00002	0.00006	100.00000	0.00000	0.00000	0.00000	0.00005
Hypoviridae	0.00001	0.00004	100.00000	0.00000	0.00000	0.00000	0.00004
Fuselloviridae	0.00001	0.00004	100.00000	0.00000	0.00000	0.00000	0.00004
Betaflexiviridae	0.00001	0.00003	100.00000	0.00000	0.00000	0.00000	0.00003
Cilevirus	0.00001	0.00003	100.00000	0.00000	0.00000	0.00000	0.00003
Bunyaviridae	0.00001	0.00003	100.00000	0.00000	0.00000	0.00000	0.00003
Malacoherpesviridae	0.00001	0.00002	100.00000	0.00000	0.00000	0.00000	0.00002
Endornaviridae	0.00001	0.00002	100.00000	0.00000	0.00000	0.00000	0.00002
Rhabdoviridae	0.00001	0.00002	100.00000	0.00000	0.00000	0.00000	0.00002
Turriviridae	0.00000	0.00001	100.00000	0.00000	0.00000	0.00000	0.00001
Leviviridae	0.00000	0.00001	100.00000	0.00000	0.00000	0.00000	0.00001

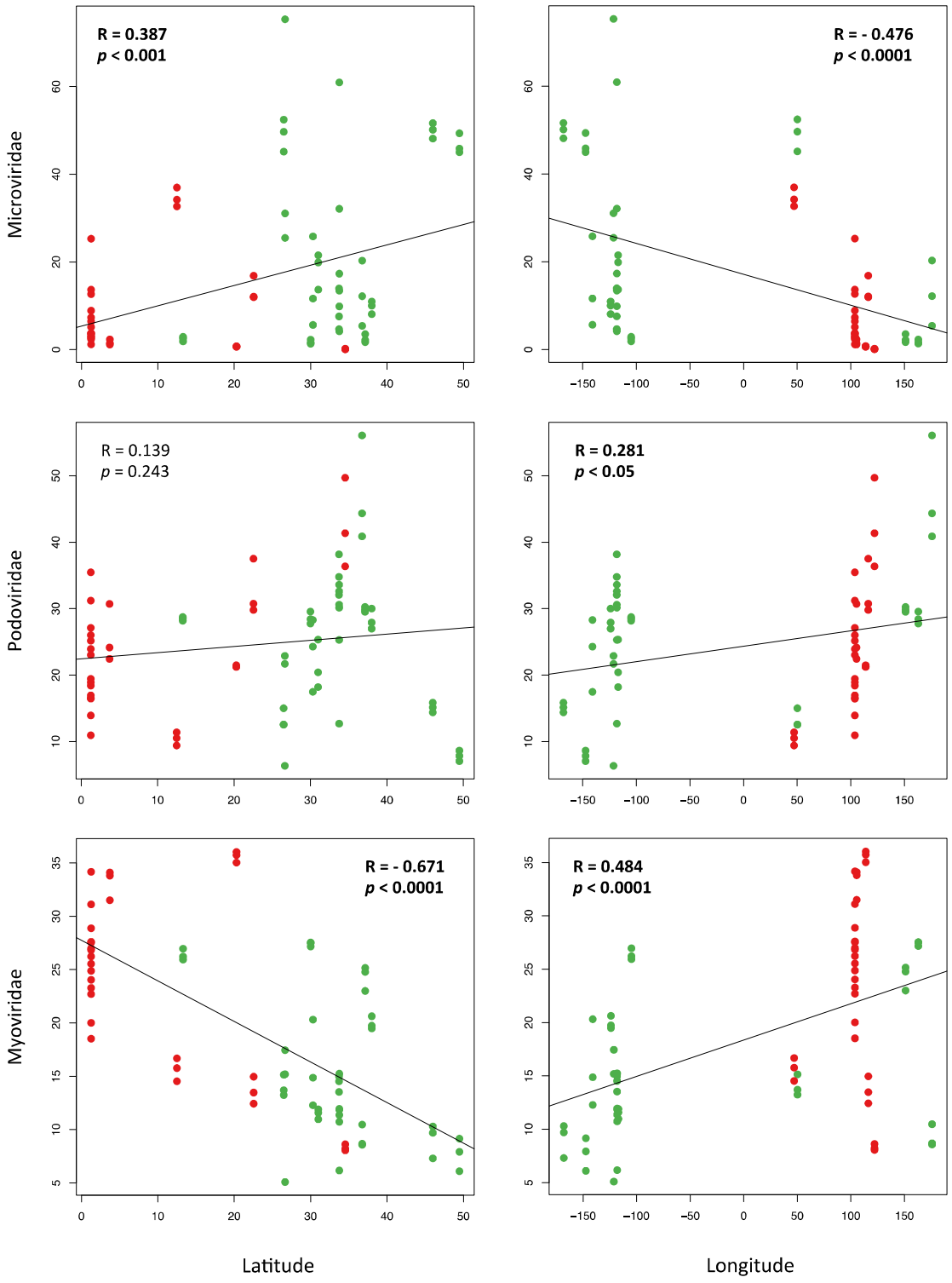


Figure 6.3. Response of the top three viral families contributing most to the differences between oceans to geographical variation. Relationship between relative abundances of *Microviridae*, *Podoviridae*, and *Myoviridae* and samples' geographic origin was

Figure 6.3 (cont'd)

examined. Latitude and longitude are expressed in decimal scale. R was the Pearson correlation coefficient for the relative abundance of viral families against the either latitude or longitude in 72 data sets. Bold text indicates a statistical significance. Green and red dots represent vessels with ballast waters arriving in the Port of Los Angeles/Long Beach and the Port of Singapore, respectively.

6.3.2. Effect of engineered, management, and environmental variables on the ocean virome

Ballast water exchange operation has been considered to be efficient to prevent the introduction of non−indigenous species based on previous findings where lower viral abundances (low number of viral particles) were found in the mid−ocean relative to coastal environments (Cochlan et al., 1993; Boehme et al., 1993; Culley and Welschmeyer, 2002). Due to limited ecological protection afforded by ballast water exchange operation, a more stringent ballast water discharge standard has been issued and awaiting additional research and technological advances (David and Gollasch, 2015). This so called 'Phase 2 standard' is based on regulating the number of organisms that are discharged with ballast water below the specific limits (Department of Homeland Security, 2012). Considering the environmental impact of viruses on host population even at a low concentration, however, potential use of viral abundance, which focuses on viral−like particles, as a regulatory parameter might not meet the goal of preventing viral invasions through ballast water. A better understanding of the types of viruses (virus richness) in ballast water would improve our ability to assess the risk of exposure of marine fauna and flora to viruses and potentially the risk to humans. Efficacy of ballast water exchange in reducing the number of different viruses was evaluated by comparing virus richness between ballast and harbor waters. Here, virus richness was defined as a total number of identified viral families in the data sets. Overall, virus richness varied considerably across samples (ranged from 20 to 56; Figure 6.4A). The ballast and harbor waters collected from the Port of Singapore (41.4 ± 6.6) had significantly higher virus richness than those from the Port of LA/LB (28.5 ± 4.1, $p < 0.0001$; Figure 6.4B). When

comparing virus richness between ballast and harbor waters at each port, harbor waters (44.1 ± 7.7) had significantly higher virus richness than ballast waters (38.7 ± 3.9) in the Port of Singapore ($p < 0.05$), while no significant difference was observed in the Port of LA/LB.

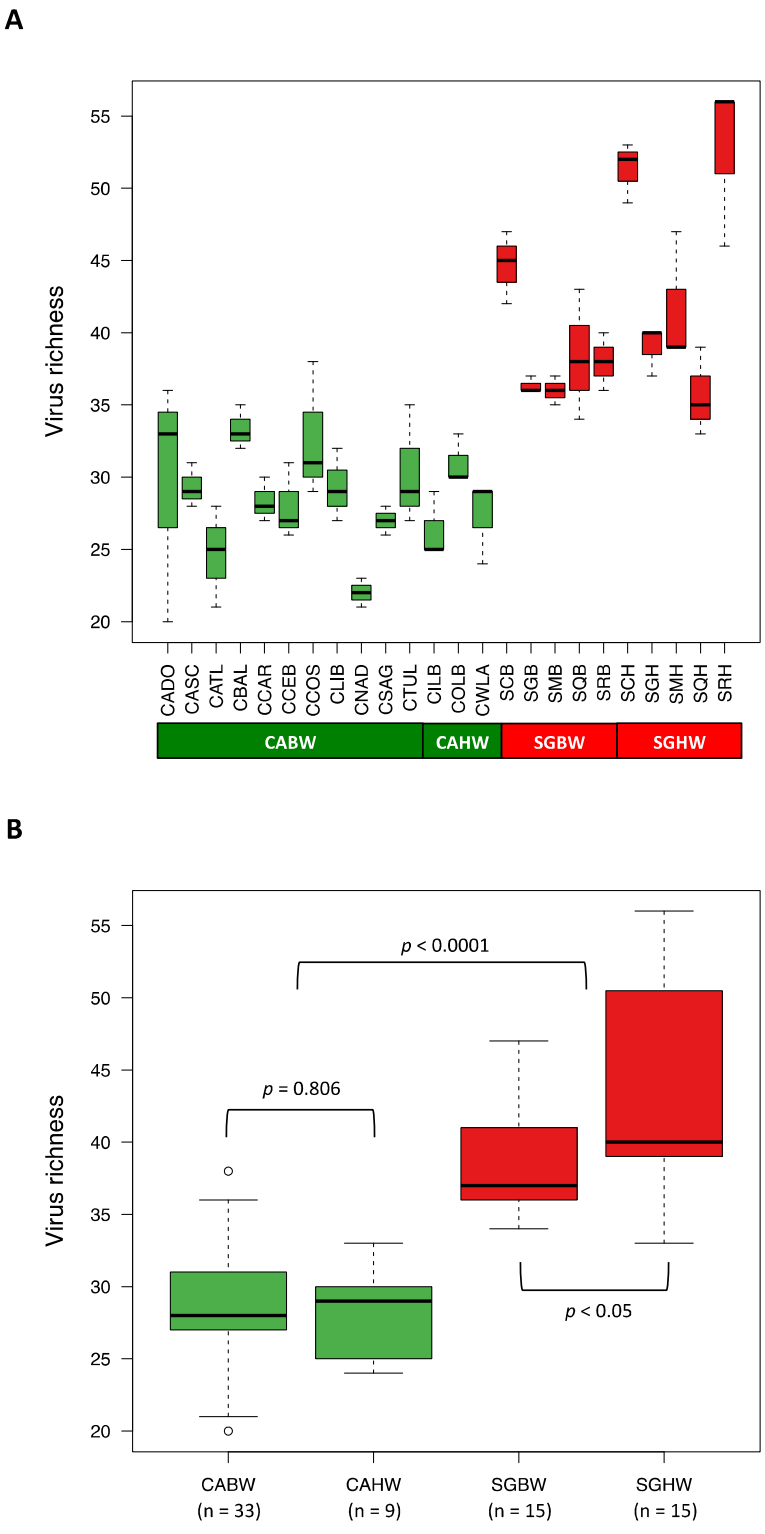


Figure 6.4. Comparison of virus richness between ballast and harbor waters. Virus richness was defined as a total number of identified viral families. **A,** Boxplot presenting

Figure 6.4 (cont'd)

virus richness of individual sample. **B,** Boxplot presenting virus richness of ballast and harbor water groups. Bonferroni−corrected pairwise t−tests were conducted to test for significant differences in virus richness between groups. Black lines within boxplots represent median values and whiskers indicate minimum and maximum values. CABW, ballast water from the Port of Los Angeles/Long Beach (LA/LB); CAHW, harbor water from the Port of LA/LB; SGBW, ballast water from the Port of Singapore; SGHW, harbor water from the Port of Singapore.

Due to an inconsistent pattern observed between the two ports, it was further hypothesized that other variables rather than type of water (either coastal or mid−ocean water) play a more important role in determining virus richness. As the current ballast water management requires a minimum of 200 nautical miles (1 nautical mile = 1.852 kilometers) from any shoreline to conduct ballast water exchange (IMO, 2004), a significance of distance from shoreline on virus richness was investigated. A correlation analysis using 72 data sets indicated that lower virus richness was shown in ballast water replaced farther from any shoreline (Figure 6.5A). As all vessels arriving in the Port of Singapore did not meet the distance requirement (> 200 nautical miles) of ballast water exchange, significance of distance on virus richness of the samples from the Port of LA/LB and the Port of Singapore was also analyzed separately to avoid any bias. A statistically significant decrease in virus richness was not observed with increased distance from shoreline in samples either from the Port of LA/LB or the Port of Singapore. This indicated that 200 nautical miles limit was not efficient in reducing virus richness of ballast water discharged into two studied ports. Effect of an important engineered variable, water storage duration in ballast tanks, on virus richness was also investigated. Again, no significant relationship was observed between virus richness and duration of water in ballast tanks, suggesting that viruses are less susceptible to harsh environmental conditions in ballast tanks, e.g., lack of light, low oxygen, and temperature fluctuations. Together with a previous finding where no significant variation in viral abundance was found over time and before and after ballast water exchange (Leichsenring and Lawrence, 2011), management or engineered variables was not considered to play a major role in determining abundance or richness of viruses present in

ballast water.

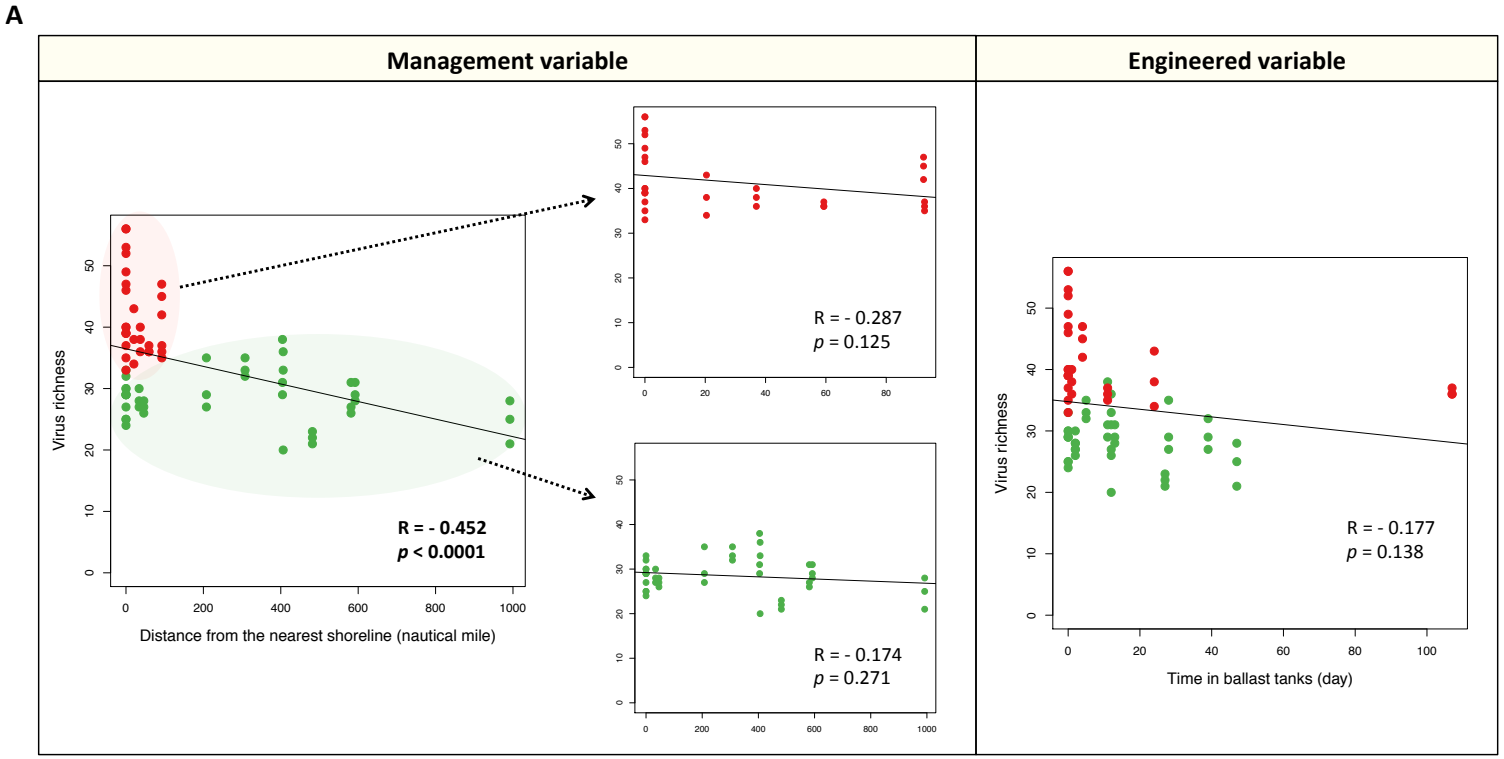
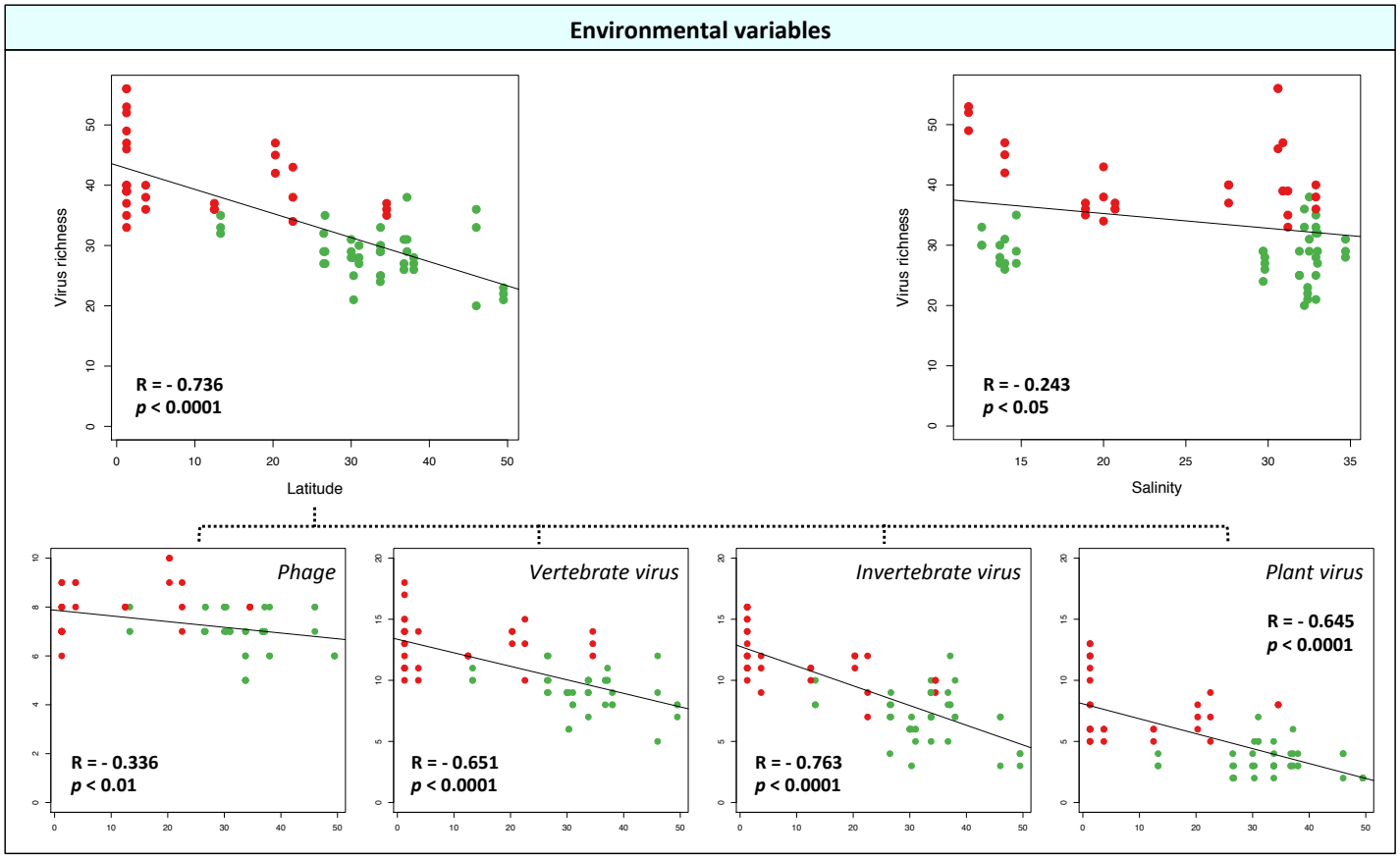


Figure 6.5. Effect of engineered, management, and environmental variables on the differences in the ocean virome. Response of virus richness to engineered, management, and environmental variables was examined. **A,** Relationship between virus richness and engineered and management variables. **B,** Relationship between virus richness and environmental variables. R was the Pearson correlation coefficient for the virus richness against the variables. Bold text indicates a statistical significance. Green and red dots represent ballast and harbor waters collected from the Port of Los Angeles/Long Beach and the Port of Singapore, respectively.

Figure 6.5 (cont'd)

B



As virus richness varied across samples and neither engineered nor management variables affected virus richness, effect of environmental variables on virus richness was next investigated in ballast water. To this end, water temperature and salinity were selected as they have been reported to be important for virus survival and infectivity (Danovaro et al., 2011). Increased water salinity had a slight inverse relationship to virus richness, but its impact on virus richness was less significant than water temperature (Figure 6.5B). As a vessel approaches a destination port, water temperature in ballast tanks becomes similar to that of surrounding environment. Therefore, latitude of samples' geographic origin was used as representative of original water temperature based on their significant relationship ($R = -0.744$, $p < 0.0001$, data not shown). A correlation analysis revealed that viruses were present in higher richness near the equator and lower richness at higher latitudes. Furthermore, each host group (e.g., phage, vertebrate virus) showed different degrees of relationship with temperature, and the weakest relationship was found in phage group. Importantly, this result suggested restricted geographical distribution of other eukaryotic (including animal and plant) viral groups with strong implications regarding invasion of local biological systems (unlike the homogeneous distribution of phages across the oceans).

6.3.3. Potential invasion by rare viral pathogens

Given a significant increase in global ship traffic and its continuous movement of ballast water, the potential for viral pathogens present in ballast and harbor waters was examined to address the host populations at risk of infection. In this research, a number of contigs were found to be associated with viruses causing diseases in a wide

range of hosts (data not shown). Viral contigs most similar to those infecting human, fish, and shrimp with a notable similarity with reference sequences were found, which were related to significant public health problems or direct economic impact due to reductions in fisheries and aquaculture production (Figure 6.6, Table 6.5).

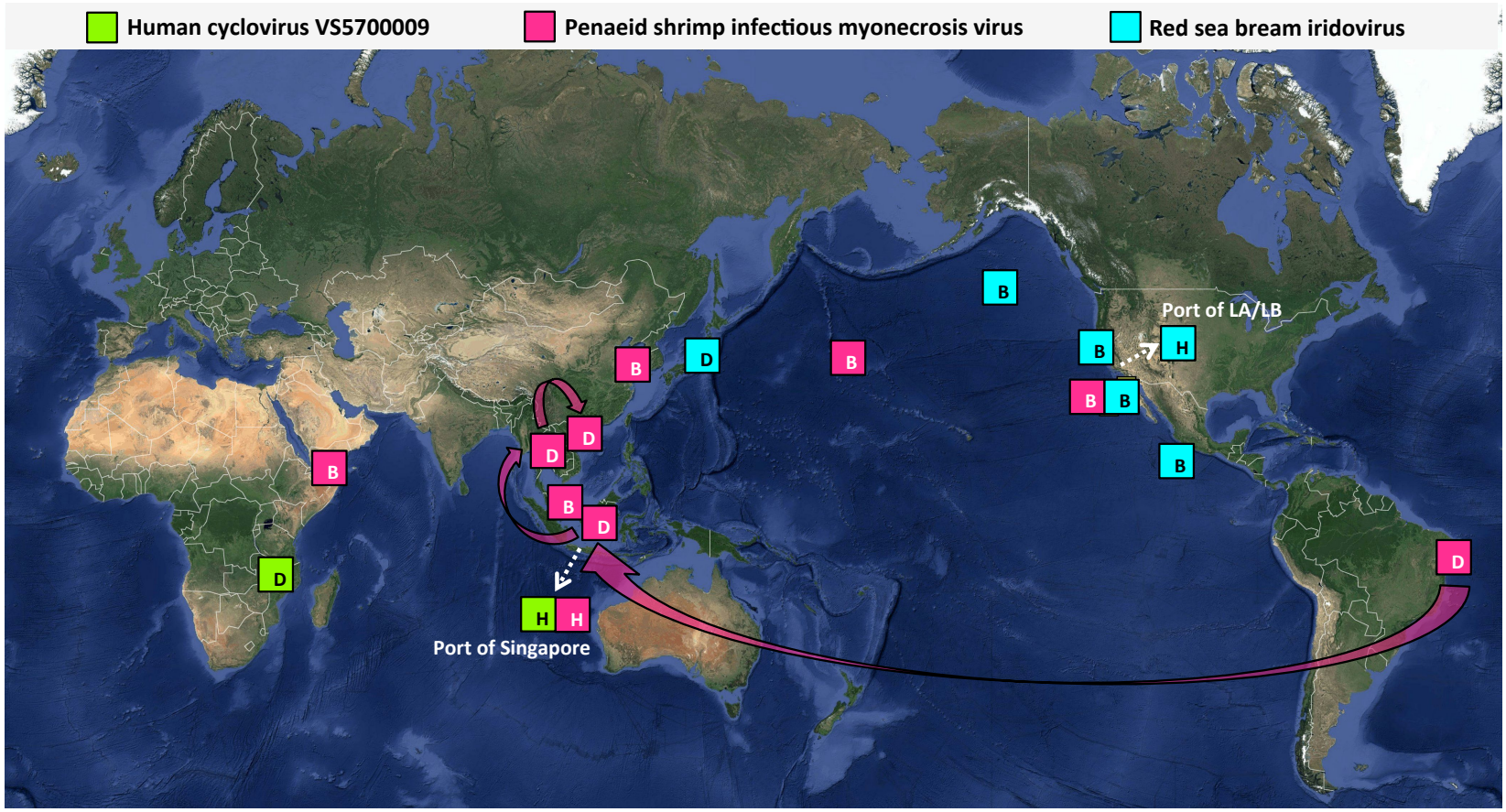


Figure 6.6. Global distribution of eukaryotic viral pathogens. Samples containing viral pathogen−associated contigs were represented in the map. B, ballast water; H, harbor water; D, where viral pathogen−induced disease was found.

Table 6.5. Identified viral pathogens by performing BLASTX searches against non−redundant (nr) database

Human cyclovirus VS5700009

Sample	Contig	GenBank accession number	Putative gene	% Identify	Alignment length (bp)	Bit score	E-value
SCH1	contig-100_4444	YP008130363.1	Rep	92.34	209	414	5.00E-139
SCH1	contig-100_11801	YP008130363.1	Rep	89.08	238	450	9.00E-157
SCH1	contig-100_88126	YP008130363.1	Rep	100.00	26	56.2	4.00E-07
SCH2	contig-100_3519	YP008130363.1	Rep	87.50	240	442	1.00E-149
SCH2	contig-100_11958	YP008130363.1	Rep	89.17	240	452	1.00E-157
SCH3	contig-100_2617	YP008130363.1	Rep	90.48	210	403	1.00E-134
SCH3	contig-100_7094	YP008130364.1	Cap	47.37	190	185	3.00E-53
SCH3	contig-100_69104	YP008130363.1	Rep	94.21	121	251	6.00E-83
SGH2	contig-100_65734	YP008130363.1	Rep	96.51	86	189	1.00E-58
SRH2	contig-100_144318	YP008130363.1	Rep	94.83	116	242	2.00E-79
SRH3	contig-100_16759	YP008130363.1	Rep	92.72	206	416	3.00E-144

Table 6.5 (cont'd)

Penaeid shrimp infectious myonecrosis virus

Sample	Contig	GenBank accession number	Putative gene	% Identify	Alignment length (bp)	Bit score	E-value
CCEB3	contig-100_35587	YP529549.2	RdRp	71.85	135	209	2.00E-61
CTUL3	contig-100_46064	YP529549.2	RdRp	27.27	143	50.8	1.00E-04
SGB1	contig-100_8554	YP529549.2	RdRp	22.58	310	61.2	8.00E-07
SGB1	contig-100_99145	YP529549.2	RdRp	32.17	115	62.8	2.00E-09
SGB3	contig-100_5276	YP529549.2	RdRp	22.73	374	62.8	3.00E-07
SMB1	contig-100_11800	YP529549.2	RdRp	67.44	258	368	4.00E-120
SMB1	contig-100_7705	YP529549.2	RdRp	24.09	303	71.2	2.00E-10
SMB1	contig-100_12153	YP529549.2	RdRp	24.02	229	58.2	2.00E-06
SMB3	contig-100_1958	YP529549.2	RdRp	24.18	306	74.7	5.00E-11
SMB3	contig-100_3046	YP529549.2	RdRp	73.03	393	598	0.00E+00
SMB3	contig-100_463	YP529549.2	RdRp	23.58	424	66.6	5.00E-08
SMB3	contig-100_6031	YP529549.2	RdRp	30.38	260	129	2.00E-30
SRB1	contig-100_74399	YP529549.2	RdRp	34.36	163	95.5	3.00E-20
SCH1	contig-100_32416	YP529549.2	RdRp	63.03	238	322	5.00E-103
SCH1	contig-100_40709	ABN05324.1	Structural protein	80.38	209	332	2.00E-105
SCH1	contig-100_52797	YP529549.2	RdRp	66.32	193	275	8.00E-86
SCH1	contig-100_71529	YP529549.2	RdRp	74.47	47	85.1	1.00E-16
SCH1	contig-100_82030	YP529549.2	RdRp	85.23	149	282	3.00E-89
SCH1	contig-100_92953	YP529549.2	RdRp	33.01	103	66.6	1.00E-10

Table 6.5 (cont'd)

Penaeid shrimp infectious myonecrosis virus

Sample	Contig	GenBank accession number	Putative gene	% Identify	Alignment length (bp)	Bit score	E-value
SCH2	contig-100_103781	ABN05324.1	Structural protein	50.00	94	101	4.00E-23
SCH2	contig-100_23465	YP529549.2	RdRp	84.66	176	330	2.00E-105
SCH2	contig-100_59905	YP529549.2	RdRp	76.69	163	256	2.00E-78
SCH2	contig-100_6098	YP529549.2	RdRp	69.88	83	135	2.00E-30
SMH1	contig-100_16497	YP529549.2	RdRp	63.64	264	357	6.00E-116
SMH1	contig-100_29857	YP529549.2	RdRp	76.17	193	315	4.00E-101
SMH1	contig-100_68796	ABN05324.1	Structural protein	62.22	45	55.8	4.00E-07
SMH2	contig-100_26022	YP529549.2	RdRp	66.46	161	231	1.00E-68
SMH2	contig-100_65524	YP529549.2	RdRp	56.64	113	145	7.00E-39
SMH3	contig-100_11987	YP529549.2	RdRp	64.52	279	391	1.00E-128
SRH2	contig-100_137724	YP529549.2	RdRp	76.19	105	167	2.00E-46

Table 6.5 (cont'd)

Red sea bream iridovirus

Sample	Contig	GenBank accession number	Putative gene	% Identify	Alignment length (bp)	Bit score	E-value
CBAL1	contig-100_140465	BAK14240.1	Cytosine DNMTs	50.48	105	104	4.00E-26
CBAL3	contig-100_91584	BAK14240.1	Cytosine DNMTs	45.52	134	118	8.00E-31
CNAD2	contig-100_18980	BAK14240.1	Cytosine DNMTs	53.85	143	148	4.00E-42
CNAD2	contig-100_22365	BAK14240.1	Cytosine DNMTs	49.12	114	108	7.00E-27
CNAD3	contig-100_1525	BAK14240.1	Cytosine DNMTs	44.23	104	94.4	1.00E-29
CSAG2	contig-100_112823	BAK14240.1	Cytosine DNMTs	59.46	74	88.6	5.00E-20
CTUL3	contig-100_27687	BAK14240.1	Cytosine DNMTs	47.47	158	149	2.00E-41
CILB3	contig-100_11773	BAK14240.1	Cytosine DNMTs	50.00	114	113	2.00E-26
COLB2	contig-100_68594	BAK14240.1	Cytosine DNMTs	45.61	114	97.8	2.00E-23

Abbreviations: Rep, replication−associated protein; Cap, capsid protein; RdRp, RNA−dependent RNA polymerase; DNMTs, DNA methyltransferase.

In three harbor waters collected from the Port of Singapore, a small ssDNA virus was detected that was closely related to human cyclovirus VS5700009 (CyCV−VS5700009) within the family *Circoviridae*. The translated amino acid (aa) seqences of 10 contigs showed best BLASTX matches to replication−associated protein (Rep, GenBank accession number YP008130363.1) and one contig to capsid protein (Cap, GenBank accession number YP008130364.1) of viral genome with 88.6% overall amino acid (aa) similarity (ranged from 47.4% to 100%). Human CyCV−VS5700009 was recently identified in patients with unexplained paraplegia from Malawi by using a metagenomics approach in an attempt to identify unknown human viruses (Smits et al., 2013). Together with two subsequent findings of a novel cycloviruses from human samples in Vietnam and Madagascar (Van Tan et al., 2014; Garigliany et al., 2014), these viruses are considered to be associated with central nervous system infection in humans. Cycloviruses have been found in different sample types from different hosts, including mammals and insects (Garigliany et al., 2014) but they have not yet been reported in environmental water samples. Considering strategic location of the Port of Singapore in the heart of Southeast Asia and its connection to numerous ports worldwide, the finding of human CyCV−VS5700009 in the Singapore harbor waters should be noted and the further risk to host populations from this viral pathogen needs to be investigated.

A small icosahedral dsRNA virus that is most closely related to penaied shrimp infectious myonecrosis virus (PsIMNV) was found in the Singapore harbor waters as well as five ballast waters (one from western Asia, two from southeastern Asia, and two from the open Pacific Ocean). PsIMNV is a member of the genus *Giardiavirus* in the family *Totiviridae*. 27 contigs showed best matches to RNA−dependent RNA polymerase

(RdRp, GenBank accession number YP529549.2) with 53.1% overall aa identity (ranged from 22.6% to 85.2%) and three contigs to structural protein (GenBank accession number ABN05324.1) of PsIMNV genome with 64.2% overall aa similarity (ranged from 50.0% to 80.4%). PsIMNV has created long−distance distribution in global aquaculture, beginning from Brazil and subsequently spreading to Indonesia, Thailand, and Hainan Province in China (Walker and Winton, 2010). The finding of PsIMNV in ballast and harbor waters from southeastern Asia was not surprising given the previously reported geographic distribution of PsIMNV. However, the presence of PsIMNV especially in two ballast waters originating from open Pacific Ocean and being discharged in the Port of LA/LB is worthy of close attention as PsIMNV has not been reported in North America.

In four ballast waters whose geographic origins were close to North America as well as harbor waters of the Port of LA/LB, a large dsDNA virus, red sea bream iridovirus (RSIV) was detected, which belongs to the newest genus *Megalocytivirus* within the family *Iridoviridae*. Nine contigs had homologies with cytosine DNA methyltransferase region of RSIV genome (GenBank accession number BAK14240.1) with 49.5% overall aa similarity (ranged from 44.2% to 59.5%). While RSIV was found in samples whose geographic origins were close to North America in this research, outbreaks of RSIV−induced disease have been occurred mainly in Asia (Ito et al., 2013). The result from this research could not reveal epidemiology or transmission pattern of these viral pathogens. Nevertheless, the finding of the viral pathogens in ballast waters suggested that long−distance distribution of these pathogens could be initiated by continuous movement of ballast water.

6.4. Conclusions

Ballast water is one of the most important vectors for transferring and spreading marine species throughout the world. Although our understanding of marine viruses (mostly phages) has improved vastly due to technological advancement, factors influencing viral diversity and their fate and transport in marine environments are largely unknown. The use of metagenomic tools provided direct evidence that ballast water harbors a high diversity of viruses and transports them across global marine environments. Driven by international regulations, demand for on−board ballast water treatment approaches has emerged. However, the efficacy of current and novel ballast water treatment methods in reducing or eliminating the potential for virus introduction is largely unexplored. Moreover, significant questions remain in addressing ballast water management challenges, such as which viral pathogens or groups should be targeted or are all viruses equal in their capacity to initiate disease and invasion processes?

There is still much to learn about the geographic distribution of viral species and the role of ballast water as a medium for the spread of invasive viruses. The potential global impact of invasive viruses on marine biogeochemical cycles and ecosystem health warrants further research.

APPENDIX

Table 6.6. Accession number of the Illumina HiSeq sequencing data

Ocean−captured ballast and harbor waters under SRP061842		
CADO1	SRX1125275	10−09−2014
CADO2	SRX1162758	10−09−2014
CADO3	SRX1162807	10−09−2014
CASC1	SRX1162759	10−09−2014
CASC2	SRX1162760	10−09−2014
CASC3	SRX1162761	10−09−2014
CATL1	SRX1162762	10−09−2014
CATL2	SRX1162808	10−09−2014
CATL3	SRX1162809	10−09−2014
CBAL1	SRX1162763	10−09−2014
CBAL2	SRX1162810	10−09−2014
CBAL3	SRX1162764	10−09−2014
CCAR1	SRX1162765	10−09−2014
CCAR2	SRX1162766	10−09−2014
CCAR3	SRX1162811	08−25−2015
CCEB1	SRX1162812	08−25−2015
CCEB2	SRX1162813	08−25−2015
CCEB3	SRX1162767	08−25−2015
CCOS1	SRX1162814	08−25−2015
CCOS2	SRX1162768	08−25−2015
CCOS3	SRX1162815	08−25−2015
CLIB1	SRX1162771	08−25−2015
CLIB2	SRX1162772	08−25−2015
CLIB3	SRX1162773	08−25−2015
CNAD1	SRX1162774	08−25−2015
CNAD2	SRX1162775	08−25−2015
CNAD3	SRX1162776	08−25−2015
CSAG1	SRX1162779	08−25−2015
CSAG2	SRX1162780	08−25−2015
CSAG3	SRX1162781	08−25−2015
CTUL1	SRX1162802	08−25−2015

Table 6.6 (cont'd)

Sample	Accession number	Release data
CTUL2	SRX1162803	08−25−2015
CTUL3	SRX1162828	08−25−2015
CILB1	SRX1162769	08−25−2015
CILB2	SRX1162770	08−25−2015
CILB3	SRX1162816	08−25−2015
COLB1	SRX1162777	08−25−2015
COLB2	SRX1162778	08−25−2015
COLB3	SRX1162817	08−25−2015
CWLA1	SRX1162804	08−25−2015
CWLA2	SRX1162805	08−25−2015
CWLA3	SRX1162806	08−25−2015
SCB1	SRX1162818	08−25−2015
SCB2	SRX1162819	08−25−2015
SCB3	SRX1162782	08−25−2015
SGB1	SRX1162786	08−25−2015
SGB2	SRX1162820	08−25−2015
SGB3	SRX1162787	08−25−2015
SMB1	SRX1162791	08−25−2015
SMB2	SRX1162792	08−25−2015
SMB3	SRX1162821	08−25−2015
SQB1	SRX1162795	08−25−2015
SQB2	SRX1162823	08−25−2015
SQB3	SRX1162796	08−25−2015
SRB1	SRX1162797	08−25−2015
SRB2	SRX1162798	08−25−2015
SRB3	SRX1162799	08−25−2015
SCH1	SRX1162783	08−25−2015
SCH2	SRX1162784	08−25−2015
SCH3	SRX1162785	08−25−2015
SGH1	SRX1162788	08−25−2015
SGH2	SRX1162789	08−25−2015

Table 6.6 (cont'd)

Sample	Accession number	Release data
SGH3	SRX1162790	08−25−2015
SMH1	SRX1162793	08−25−2015
SMH2	SRX1162822	08−25−2015
SMH3	SRX1162794	08−25−2015
SQH1	SRX1162824	08−25−2015
SQH2	SRX1162825	08−25−2015
SQH3	SRX1162826	08−25−2015
SRH1	SRX1162800	08−25−2015
SRH2	SRX1162827	08−25−2015
SRH3	SRX1162801	08−25−2015

REFERENCES

REFERENCES

1. Amalfitano, S.; Coci, M.; Corno, G.; Luna, G., A microbial perspective on biological invasions in aquatic ecosystems. *Hydrobiologia* **2015,** *746*, (1), 13-22.

2. Angly, F.; Felts, B.; Breitbart, M.; Salamon, P.; Edwards, R.; Carlson, C.; Chan, A.; Haynes, M.; Kelley, S.; Liu, H.; Mahaffy, J.; Mueller, J.; Nulton, J.; Olson, R.; Parsons, R.; Rayhawk, S.; Suttle, C.; Rohwer, F., The marine viromes of four oceanic regions. *Plos Biology* **2006,** *4*, (11), 2121-2131.

3. Balique, F.; Lecoq, H.; Raoult, D.; Colson, P., Can Plant Viruses Cross the Kingdom Border and Be Pathogenic to Humans? *Viruses* **2015,** *7*, (4), 2074-2098.

4. Boehme, J.; Frischer, M.; Jiang, S.; Kellogg, C.; Pichard, S.; Rose, J.; Steinway, C.; Paul, J., Viruses, bacterioplankton, and phytoplankton in the southeastern Gulf of Mexico: distribution and contribution to oceanic DNA pools. *Marine Ecology Progress Series* **1993,** *97*, (1), 1-10.

5. Brum, J.; Ignacio-Espinoza, J.; Roux, S.; Doulcier, G.; Acinas, S.; Alberti, A.; Chaffron, S.; Cruaud, C.; de Vargas, C.; Gasol, J.; Gorsky, G.; Gregory, A.; Guidi, L.; Hingamp, P.; Iudicone, D.; Not, F.; Ogata, H.; Pesant, S.; Poulos, B.; Schwenck, S.; Speich, S.; Dimier, C.; Kandels-Lewis, S.; Picheral, M.; Searson, S.; Bork, P.; Bowler, C.; Sunagawa, S.; Wincker, P.; Karsenti, E.; Sullivan, M.; Coordinators, T. O.; Coordinators, T. O., Patterns and ecological drivers of ocean viral communities. *Science* **2015,** *348*, (6237).

6. Cochlan, W.; Wikner, J.; Steward, G.; Smith, D.; Azam, F., Spatial distribution of viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments. *Marine Ecology Progress Series* **1993,** *92*, (1-2), 77-87.

7. Culley, A.; Welschmeyer, N., The abundance, distribution, and correlation of viruses, phytoplankton, and prokaryotes along a Pacific Ocean transect. *Limnology and Oceanography* **2002,** *47*, (5), 1508-1513.

8. Danovaro, R.; Corinaldesi, C.; Dell'Anno, A.; Fuhrman, J.; Middelburg, J.; Noble, R.; Suttle, C., Marine viruses and global climate change. *Fems Microbiology Reviews* **2011,** *35*, (6), 993-1034.

9. David, M.; Gollasch, S., Global Maritime Transport and Ballast Water Management Issues and Solutions. In Invading Nature - Springer Series in Invasion Ecology 8, Springer, Dordrecht, The Netherlands, **2015**.

10. Department of Homeland Security. **2012,** Available at http://www.gpo.gov/fdsys/pkg/FR-2012-03-23/pdf/2012-6579.pdf.

11. Drake, L.; Doblin, M.; Dobbs, F., Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. *Marine Pollution Bulletin* **2007,** *55*, (7-9), 333-341.

12. Garigliany, M.; Hagen, R.; Frickmann, H.; May, J.; Schwarz, N.; Perse, A.; Jost, H.; Borstler, J.; Shahhosseini, N.; Desmecht, D.; Mbunkah, H.; Daniel, A.; Kingsley, M.; Campos, R.; de Paula, V.; Randriamampionona, N.; Poppert, S.; Tannich, E.; Rakotozandrindrainy, R.; Cadar, D.; Schmidt-Chanasit, J., Cyclovirus CyCV-VN species distribution is not limited to Vietnam and extends to Africa. *Scientific Reports* **2014,** *4*.

13. GEF-UNDP-IMO GloBallast Partnerships and IOI: Guidelines for National Ballast Water Status Assessments. GloBallast Monographs No. 17, **2009,** Available at http://globallast.imo.org/wp-content/uploads/2014/11/Mono17_English.pdf.

14. Hammer, O.; Harper, D. A. T.; Ryan, P. D., PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **2001,** *4*, (1).

15. Hurwitz, B.; Sullivan, M., The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *Plos One* **2013,** *8*, (2).

16. International Maritime Organization. International convention for the control and management of ships' ballast water and sediments, **2004,** Available at http://www.uscg.mil/hq/cg5/cg522/cg5224/docs/BWM-Treaty.pdf.

17. Ito, T.; Yoshiura, Y.; Kamaishi, T.; Yoshida, K.; Nakajima, K., Prevalence of red sea bream iridovirus among organs of Japanese amberjack (Seriola quinqueradiata) exposed to cultured red sea bream iridovirus. *Journal of General Virology* **2013,** *94*, 2094-2101.

18. Leichsenring, J.; Lawrence, J., Effect of mid-oceanic ballast water exchange on virus-like particle abundance during two trans-Pacific voyages. *Marine Pollution Bulletin* **2011,** *62*, (5), 1103-1108.

19. Martinez, J.; Swan, B.; Wilson, W., Marine viruses, a genetic reservoir revealed by targeted viromics. *Isme Journal* **2014,** *8*, (5), 1079-1088.

20. Oksanen, J.; Blanchet, F.G.; Kindt, R.; Legendre, P.; O'Hara, R.G.; Simpson, G.L.; Solymos, P.; Stevens, M.H.H.; Wagner, H., *vegan: Community Ecology Package,* R package version 2.0-10, **2013,** Available at http://CRAN.R-project.org/package=vegan.

21. R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing Vienna, Austria, **2010**.

22. Ruiz, G.; Rawlings, T.; Dobbs, F.; Drake, L.; Mullady, T.; Huq, A.; Colwell, R., Global spread of microorganisms by ships - Ballast water discharged from vessels harbours a cocktail of potential pathogens. *Nature* **2000,** *408*, (6808), 49-50.

23. Smits, S.; Zijlstra, E.; van Hellemond, J.; Schapendonk, C.; Bodewes, R.; Schurch, A.; Haagmans, B.; Osterhaus, A., Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010-2011. *Emerging Infectious Diseases* **2013,** *19*, (9), 1511-1513.

24. Van Tan, L.; De Jong, M.; Kinh, N.; Trung, N.; Taylor, W.; Wertheim, H.; Van der Ende, A.; Van der Hoek, L.; Canuti, M.; Crusat, M.; Sona, S.; Uyen, N.; Giri, A.; BKrong, N.; Nghia, H.; Farrar, J.; Bryant, J.; Hien, T.; Chau, N.; Van Doorn, H., Limited geographic distribution of the novel cyclovirus CyCV-VN. *Scientific Reports* **2014,** *4*.

25. Walker, P. J.; Winton, J. R., Emerging viral diseases of fish and shrimp. *Veterinary Research* **2010,** *41*, (6), 51.

26. Winter, C.; Garcia, J.; Weinbauer, M.; DuBow, M.; Herndl, G., Comparison of Deep-Water Viromes from the Atlantic Ocean and the Mediterranean Sea. *Plos One* **2014,** *9*, (6).

27. Wommack, K.; Nasko, D.; Chopyk, J.; Sakowski, E., Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology* **2015,** *53*, (3), 181-192.

CHAPTER 7

CONCLUSIONS

Increasing demand for global trade and the growth in ship traffic have created concerns about the transport and discharge of non−native species including viruses in aquatic environments and the impact on native ecosystems. Currently, our knowledge on the transport of viruses via ships' ballast water has remained limited to viral abundance estimates. This dissertation described for the first time taxonomic characterization of the community structure and diversity of viruses present in ballast water. In this chapter, key results of the present research are summarized. The implications for policy and technological development as well as recommendations for future research are also discussed.

7.1. Summary

The present research integrated environmental virology, metagenomics, and bioinformatics in order to fill critical knowledge gaps of the role of ballast water in the transport of viral assemblages and their compositions. This allowed detailed characterization of ballast water viral communities, including both DNA and RNA viruses and thus, improved the understanding of ballast water−mediated viral transport.

The metagenomic analysis of viruses is a multi−state process, including concentration and purification of virus−like particles (VLPs) from complex environmental water samples and nucleic acid extraction and amplification of viruses with different genome types. In the present research, the use of tangential flow filtration (TFF) with disposable hollow fiber ultrafilters was found to be efficient in simultaneously concentrating different types of viruses in environmental water samples. The TFF procedure described in the Chapter 4 was therefore used to concentrate VLPs in

freshwater− and ocean−captured ballast and harbor waters. Furthermore, the nucleic acid preparation method for high−throughput sequencing described in the Chapter 4 allowed the characterization of DNA and RNA viruses as well as viral pathogens that may be present at a very low level.

As presented in the Chapter 5, the Great Lakes was used as a model system to gain a basic knowledge on taxonomic composition and diversity of freshwater viruses in ships' ballast water. The Great Lakes basin linking North America with ports throughout the world has been invaded by more non−native species than any other freshwater ecosystem in the world (Ricciardi, 2006; Pagnucco et al., 2015) and ballast water has been considered as a contributing factor to the introduction of these non−native species (Mills et al., 1993; Drake and Lodge, 2007). Metagenomic investigation revealed the enormous genetic diversity of viruses in ballast and harbor waters, which could not be uncovered using publicly available sequence database. This demonstrated our limited knowledge of viruses in ballast water. Ballast water was found to harbor diverse viruses, which were largely dominated by double–stranded DNA phages (*Myoviridae*, *Podoviridae*, *Siphoviridae*, and unclassified *Caudovirales*) as well as viral pathogens associated with fish and shrimp at a low level. Comparative metagenomic analyses showed that viral metagenomes (viromes) were distinct among the Great Lakes and formed a specific group of temperate freshwater viromes but separated from viromes associated with marine environments and engineered freshwater systems.

Based on findings where ballast water harbors diverse viruses with characteristic signatures depending on environments, the scope of the rsearch was expanded to marine enviroments (Chapter 6). The Port of Los Angeles/Long Beach (LA/LB) and the Port of

Singapore were chosen as they are the world's busiest container ports and have vessels carrying ballast water of marine origin from worldwide. This research demonstrated through metagenomic analyses that viromes showed geographical differences with major variations observing in several viral families, including *Microviridae*, *Myoviridae*, *Podoviridae*, and *Siphoviridae*. More interestingly, these viral families, which contributed most to the virome dissimilarity, showed different responses to geographical variation. Moreover, ballast water was found to be a contributing factor to transport of not only these phage families but also viral pathogens from one part of the world to another. This research also revealed that virus richness correlates with local environmental conditions but not with ballast water associated engineered (e.g., water storage duration in ballast tanks) or management variables (e.g., distance from nearest shoreline).

7.2. Implications for policy and technological development

Overall, the present research unveiled a high diversity of viruses in ballast water, which has been poorly understood to date due to the difficulty in collection of ballast water and suitable analytical tools for virus analysis. Furthermore, the role of ballast water in initiating long–distance distribution of viruses including viral pathogens was identified, suggesting an increased risk of exposure of aquatic flora and fauna to viruses. These findings emphasize the need for federal agencies to consider the planning and implementing of ballast water discharge limits for viruses, which is currently being considered only by the State of California. This also reinforces the need for ballast water treatment for controlling potential viral invasion.

This research approached an efficacy of current ballast water exchange practice by comparing viromes from various aquatic environments (Chapter 5) and by examining variation in virome composition of ballast and harbor waters between geographic locations (Chapter 6). As revealed in the Chapter 5, the characteristic virome signature observed between different aquatic environments, especially between freshwater and marine environments, implicates the potential introduction of viruses associated with the marine environment to freshwater environment, such as the Great Lakes basin. As shown in the Chapter 6, variations in virome composition of exchanged ballast water and harbor water indicated that ballast water exchange practice does not prevent the potential introduction of non–native viruses. Both findings emphasize that current mid–ocean exchange of ballast water should be viewed not only by the reduction of number of VLPs in ballast water but also by differences in composition of viral communities introduced into native ecosystems through ballast water.

In order to improve our understanding of factors controlling viruses in ballast water, effects of various parameters, including engineered, management, and environmental parameters, on virus richness (type of viruses) as opposed to viral level (number of VLPs) were examined in the Chapter 6. Distance from nearest shoreline (management parameter in this research), which has been considered to be an important factor of virus level, was found to be insignificant in reducing virus richness. Engineered parameter, water storage duration in ballast tanks, was also found to be insignificant in reducing number of different viruses introduced into native ecosystems. However, conditions of local environment, particularly water temperature and salinity, showed close relatedness with virus richness. These findings emphasize the need for considering

an alternative regulatory parameter rather than current use of distance from nearest shoreline or potential use of viral level to meet the goal of preventing viral invasions through ballast water.

Currently, technologies for ballast water treatment are still in the research and development phase. These techniques have been tested mainly with marine water, focusing on reducing the level of phytoplankton and bacteria. There is a lack of information on effectiveness of ballast water treatment technologies in removing viruses, particularly in freshwater environments. Furthermore, the identification of potential viral pathogens of human, fish, and shrimp in ballast water provide valuable information on what ballast water treatment would be needed to inactivate viruses in the future. An improved understanding of viral diversity in the present research will assist in defining ballast water treatment and eventually for shipping industries in complying with increasingly stringent regulatory demands.

7.3. Limitations and recommendations for future research

The present research collected approximately 60 L of water from a ballast tank from each vessel. Given an average of five million gallons of ballast water typically carried by a vessel (Carlton et al., 1995), the ballast water samples collected from this research may not be biologically representative of the contents of the ballast tanks. This could misrepresent true diversity of viruses in ballast water and skew our view of viral diversity. Studies have shown the heterogeneous distribution of organisms contained within ballast tanks, which hinder the collection of representative samples (Murphy et al., 2002; Gollasch and David, 2010; Costa et al., 2015). Other challenges associated with

obtaining biologically representative samples from ballast tanks include: (1) multiple ballast tanks containing ballast water of different origins in a vessel, (2) large volume of ballast water in a vessel, (3) presence of sediment and biofilm in ballast tanks, and (4) irregular shapes of ballast tanks (modified from Murphy et al., 2002). Obtaining biologically representative samples from ballast tanks is critical, as a number of viable organisms and indicator microbes (*Escherichia coli*, intestinal enterococci, and toxicogenic *Vibrio cholera*) discharged with ballast water at the destination ports have to meet D–2 standard (IMO, 2008). In addition, the G2 guideline states that ballast water samples used to determine a ship's compliance must be 'representative' of the 'whole' ballast water to be discharged. Thus, future research needs to investigate the impact of ballast water sampling on the representation of species diversity and to provide clear guidance on how to obtain representative samples.

Although the present research provided new insights into the diversity of viruses and their associated host populations in ballast water, this research is limited to examining viral communities only in discharged ballast water at the destination ports, including the Great Lakes, California, and Singapore. A better understanding of viruses that are being discharged with ballast water at the destination ports is valuable, as it is directly related to the risk of viral invasion at a receiving port. However, investigating changes in viral community structure during the voyages and before and after ballast water exchange practice will determine virus survival in the ballast water over time. This can further suggest which viruses should be targeted for ballast water treatment. In addition, understanding a correlation between changes in viral community structure and environmental or engineered parameters will have implications for developing ballast

water treatment technologies. A more intensive sampling design (e.g., on−board sampling during voyages) is warranted to aid in resolving the viral diversity associated with ballast water as well as in developing ballast water treatment technologies.

Isolation and purification of VLPs from environmental water samples are still challenging components of viral metagenomic studies. As studies are now striving not only to describe viral communities but also quantitatively compare viral abundances across samples, the use of efficient, unbiased, and reproducible workflow for preparing viromes is critical. Previous methodological studies (Thurber et al., 2009; Thurber, 2011; Duhaime and Sullivan, 2012; Hurwitz et al., 2013) have evaluated different VLP concentration and purification and amplification methods using artificial samples of known composition and level of viruses, mainly DNA phages. However, effects of concentration, purification, and amplification methods on recovery of different viral genome types, particularly RNA viruses are still unknown. As different methods of choice heavily affect the type of viruses recovered and can give different views on the viral diversity, the biases introduced in the virome preparation should be considered in downstream analyses, particularly for comparative metagenomic analysis. Thus, future research is needed for a critical evaluation of methods for the virome preparation in complex environmental water samples and impact on resulting sequence data sets.

Another challenges in viral metagenomic studies are analyzing a large amount of sequence data using bioinformatics analysis and the lack of a standardized bioinformatics pipeline for viruses. BLAST searches (Altschul et al., 1990), which are based on similarity to existing databases, are most frequently used to describe the taxonomic profile of metagenomes. However, several factors hamper extracting meaningful

information, resulting in a high percentage of unidentifiable sequences, classified as "unknown". The factors contributing to the limited recovery rate through BLAST searches include: (1) the incompleteness of public sequence databases, (2) the short read lengths produced by high–throughput sequencing technologies, and (3) sequencing errors (Fancello et al., 2012). Computational tools facilitating rapid and robust data analysis of high–throughput sequences are surprisingly lacking in the field of viral metagenomics. Currently, two web–servers, including MetaVir (Roux et al., 2014) and VIROME (Wommack et al., 2011) are available for a comprehensive virome analysis. However, these bioinformatics tools are designed for Roche 454 pyrosequencing data sets and not available for the analysis of assembled sequence data set derived from short reads of Illumina technologies. Recently, MetaVir version 2 tackled these limitations and facilitates analysis of assembled virome sequences (Roux et al., 2014). Additional tools are needed to handle growing number of viromes from different kinds of sequencing technologies for a rapid and robust analysis of high–throughput sequences.

REFERENCES

REFERENCES

1. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D., Basic Local Alignment Search Tool. *Journal of Molecular Biology* **1990,** *215* (3), 403-410.

2. Carlton, J. T.; Reid, D. M.; Van Leeuwen, H., Shipping study the role of shipping in the introduction of nonindigenous aquatic organisms to the coastal waters of the United States (other than the Great Lakes) and an analysis of control options, National Technical Information Service Distributor, Washington, DC, **1995**.

3. Costa, E.; Lopes, R.; Singer, J., Implications of heterogeneous distributions of organisms on ballast water sampling. *Marine Pollution Bulletin* **2015,** *91*, (1), 280-287.

4. Drake, J.; Lodge, D., Rate of species introductions in the Great Lakes via ships' ballast water and sediments. *Canadian Journal of Fisheries and Aquatic Sciences* **2007,** *64*, (3), 530-538.

5. Duhaime, M.; Sullivan, M., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **2012,** *434*, (2), 181-186.

6. Fancello, L.; Raoult, D.; Desnues, C., Computational tools for viral metagenomics and their application in clinical research. *Virology* **2012,** *434*, (2), 162-174.

7. Gollasch, S., David, M., Testing sample representativeness of a ballast water discharge and developing methods for indicative analysis, European Maritime Safety Association, **2010**.

8. Hurwitz, B.; Deng, L.; Poulos, B.; Sullivan, M., Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology* **2013,** *15*, (5), 1428-1440.

9. IMO, Report of the Marine Environment Protection Committee, Annex 3: Resolution MEPC. 173(58) – Guidelines for Ballast Water Sampling (G2), International Maritime Organization, London, **2008**.

10. Mills, E.; Leach, J.; Carlton, J.; Secor, C., Exotic species in the Great Lakes - a history of biotic crises and anthropogenic introductions. *Journal of Great Lakes Research* **1993,** *19* (1), 1-54.

11. Murphy, K.; Ritz, D.; Hewitt, C., Heterogeneous zooplankton distribution in a ship's ballast tanks. *Journal of Plankton Research* **2002,** *24*, (7), 729-734.

12. Pagnucco, K.; Maynard, G.; Fera, S.; Yan, N.; Nalepa, T.; Ricciardi, A., The Future of Species Invasions in the Great Lakes-St. Lawrence River Basin (vol 41, pg 314,

2014). *Journal of Great Lakes Research* **2015,** *41*, 197-197.

13. Ricciardi, A., Patterns of invasion in the Laurentian Great Lakes in relation to changes in vector activity. *Diversity and Distributions* **2006,** *12*, (4), 425-433.

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

15. Thurber R., Methods in Viral Metagenomics. Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats. **2011,** 15–24.

16. Thurber, R.; Haynes, M.; Breitbart, M.; Wegley, L.; Rohwer, F., Laboratory procedures to generate viral metagenomes. *Nature Protocols* **2009,** *4*, (4), 470-483.

17. Wommack, K.; Bhavsar, J.; Polson, S.; Chen, J.; Dumas, M.; Srinivasiah, S.; Furman, M.; Jamindar, S.; Nasko, D., VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **2012,** *6*, (3), 421-433.