January 2014

Investigating the Diversity of Single-Stranded DNA Bacteriophages in Marine Environments

Max Stephen Hopkins
University of South Florida, hopkins.max@gmail.com

Investigating the Diversity of Single-Stranded

DNA Bacteriophages in Marine Environments

by

Max Hopkins

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Biological Oceanography
with a concentration in Marine Microbiology
College of Marine Science
University of South Florida

Major Professor: Mya Breitbart, Ph.D.
John Paul, Ph.D.
Lauren McDaniel, Ph.D.

Date of Approval:
June 19, 2014

Keywords: marine phage, marine virus, Gokushovirus, *Microviridae*, cyanophage

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

There are estimated to be 10^{30} virus-like particles in the world's oceans. Most are viruses that infect bacteria, called 'bacteriophages' or simply 'phages'. Phages exert tremendous influence on marine biogeochemical cycling because they are responsible for about half of all bacterial death in the oceans, causing nutrient release into the dissolved and particulate organic matter pools. Traditional paradigms of phage biology held that most of these ocean phages belonged to the *Caudovirales* group: phages that contain a double-stranded DNA genome within a geometric capsid 'head' to which a 'tail' is joined, in one of several morphological variants, that is the main structure allowing the phage to interact and infect a host bacterium. Compared to tailed phages, small, non-tailed, single-stranded DNA-containing phages have been an historical afterthought; believed to exist only in specialized, niche environments. However, recent studies harnessing advances in technology have revealed that single-stranded DNA phages are ubiquitous to nearly every marine environment yet tested.

Small, icosahedral, single-stranded DNA bacteriophages of the subfamily *Gokushovirinae* (family *Microviridae*) exemplify the difficulty that viruses can present as study subjects. They are difficult to visualize by epifluorescence microscopy and contain a paucity of genetic and protein material. As a result, recognition of their importance in marine environments has lagged behind that of tailed, double-stranded DNA bacteriophages. This thesis seeks to redress this knowledge gap.

The first chapter expands knowledge of gokushovirus diversity in the environment by developing a degenerate PCR assay to amplify a portion of the major capsid protein (MCP) gene of gokushoviruses. Over 500 amplicons were sequenced from ten diverse environmental samples (sediments, sewage, seawater and freshwater), revealing the ubiquity and high diversity of this understudied phage group. The data was aggregated in several informative ways. Multiple alignments were combined with a predicted 3D-structure to reveal regions of both high and low conservation. Viewed in a phylogenetic framework, many gokushovirus MCP clades contained samples from multiple environments, although distinct clades dominated the different sample types. Some environments, particularly pelagic sediments, appear as hotbeds of gokushovirus diversity, while freshwater springs were the least diverse.

The second chapter used the same primer set to detect gokushovirus communities at 0 m and 100 m depth in two seasons from three years at the Bermuda Atlantic Time-series Study (BATS) site. As a result of twenty-six years of constant sampling, the annual hydrodynamic cycling of BATS is very well understood. This wealth of knowledge allows us to hypothesize that the winter deep mixing layer will act to connect the viral communities between 0 m and 100 m. Conversely, in summer when stratification occurs, viral communities at the two depths will become divergent. We find compelling evidence to support this hypothesis.

The final chapter of this thesis details continuing efforts to characterize the first non-tailed, single-stranded DNA, temperate phage to infect a member of the globally important genus of marine autotroph, *Synechococcus.* Efforts undertaken have spanned genomic, metagenomic and proteomic methodologies. The lack of culturable, phage-host

model systems for small, single-stranded DNA phages is today one of the most glaring impediments to increased understanding of these viruses. In combination with the data presented on environmental diversity, steps taken towards establishing this *Synechococcus* phage as a culturable model system makes this thesis a major contribution to the understanding of environmental ssDNA phages.

INTRODUCTION

This thesis is focused on elucidating the diversity of small, icosahedral single-stranded DNA (ssDNA) bacteriophages (viruses that infect bacteria), often simply called phages, in aquatic environments. The first DNA genome to be completely sequenced belonged to the diminutive ssDNA phage ϕX174, initiating the genomic sequencing era in 1977 (Sanger et al 1977). Knowledge of phage biology and ecology has expanded rapidly since that time, and phages are currently recognized as the most abundant biological entities on the planet, exerting significant driving forces on bacterial diversity and global biogeochemistry (Breitbart 2012). Despite the early characterization of ssDNA phages, their double-stranded DNA (dsDNA) counterparts have received a disproportionate amount of attention over the past three decades. As of 2011, more than 80% of the completely sequenced phage genomes in Genbank belonged to tailed dsDNA phages of the *Caudovirales* (Krupovic et al 2011). The *Caudovirales* also account for the vast majority (96%) of phages characterized by electron microscopy (Ackermann 2007). Culture-based studies, combined with pulsed-field gel electrophoresis studies (Steward et al 2000) and early metagenomic methods that excluded ssDNA phages (Breitbart et al 2002), created the general paradigm that dsDNA tailed phages belonging to the *Caudovirales* dominate in environmental communities. However, recent studies have challenged this dogma by demonstrating the abundance of nontailed viral particles and DNase-insensitive viral genomes in the oceans (Brum et al 2013; Steward et al 2012).

Single-stranded DNA viruses, both bacteriophages and eukaryotic viruses, have garnered attention for having inherently high mutation rates which are estimated to be on par with rates for RNA viruses (Duffy and Holmes 2009; Duffy et al 2008; Raney et al 2004). This is thought to be a result of oxidative deamination, which single-stranded DNA is several hundred-fold more likely to experience than double-stranded DNA (Frederico et al 1990).

Currently only two ssDNA phage families have been adopted by the International Committee for the Taxonomy of Viruses (ICTV); family *Inoviridae* and family *Microviridae* (Fane 2005). The *Inoviridae* are long, filamentous phages such as M13 of *E. coli* which enter into stable, non-lytic relationships with host enteric bacteria and steadily produce progeny through budding*.* The *Microviridae* are small, T=1 icosahedra that are primarily lytic, although they have been discovered in integrated prophage form (Cherwa and Fane 2011; Krupovic and Forterre 2011). ICTV further divides the family *Microviridae* into subfamilies *Gokushovirinae,* which infect obligate intracellular parasites (e.g. Chlamydia) and the 'true' *Microvirinae* that infect enterobacteria and have a surface-spike protein. Recent reports have suggested the existence of additional ssDNA morphotypes and genotypes which, when fully characterized, may be divergent enough to warrant formation of new families of ssDNA phage (Holmfeldt et al 2013; McDaniel et al 2006). The icosahedral ssDNA subfamily *Gokushovirinae* (which means "very small' in Japanese) and related, as yet uncharacterized phage are the focus of this thesis.

Icosahedral, ssDNA phages belonging to the family *Microviridae* have been present in culture collections since the 1920s, yet until 2006, this phage family had not been described in the oceans, one of the most extensively studied ecosystems in terms of microbial ecology. In 2006, next-generation 454 pyrosequencing was applied to viral metagenomics, requiring the introduction of a non-specific amplification technique (rolling circle amplification; RCA) to obtain sufficient starting quantities of DNA. The first study to utilize this approach found that the recognizable sequences from an 80 meter deep viral metagenome from the Sargasso Sea were dominated by sequences similar to the *Gokushovirinae* subfamily (Angly et al 2006). This finding was unexpected since gokushoviruses had previously only been reported to infect parasitic bacteria (*Chlamydia*, *Bdellovibrio*, *Spiroplasma*) and were believed to be successful in a fairly narrow niche (Brentlinger et al 2002; Cherwa and Fane 2011). The Angly et al (2006) study relied on the use of RCA, which is known to preferentially enrich for circular ssDNA elements (Kim and Bae 2011), so conclusions cannot be drawn regarding the abundance of gokushoviruses. Despite this caveat it was surprising to find environmental settings with significant community-composition of heretofore human- and agriculturally-associated phages.

Building upon the Angly et al (2006) study, viral metagenomic studies employing RCA have uncovered novel ssDNA phages in a variety of environments (reviewed in Rosario and Breitbart 2011), including freshwater aquifers (Smith et al 2013), freshwater lakes (López-Bueno et al 2009; Roux et al 2012a), stromatolites (Desnues et al 2008), soils (Kim et al 2008), coastal estuaries (Labonté and Suttle 2013a; McDaniel et al 2008; McDaniel et al 2013), seawater (Labonté and Suttle 2013b) and reclaimed water (Rosario et al 2009). A 2012 data-mining study assembled 81 additional complete *Microviridae* genome sequences from various environments and human gut/stool samples (Roux et al 2012b). A key finding of this comprehensive study was an intriguing emergent phylogenetic topology for the *Gokushovirinae* subfamily with dichotomous clading of environmental vs. 'human-associated' gokushoviruses (Hopkins et al 2014; Roux et al 2012b).

Although there is abundant genomic evidence for the presence of ssDNA phages in the environment, to date there has been only one fully-characterized ssDNA phage cultured from an environmental sample; a Baltic Sea-derived *Bacteroidetes* isolate (Holmfeldt et al 2013). The authors suggest that this may be a highly divergent variant of *Microviridae* based on the syntenous gene layout, however this will await subsequent annotation of the unknown gene features. There is a pressing need for culturing representatives of environmental ssDNA phage in order to determine the host range and effects of these seemingly ubiquitous viruses.

This thesis seeks to expand on earlier studies of environmental ssDNA bacteriophage. Chapters one and two describe the development and application of degenerate PCR primers for amplifying environmental *Gokushovirinae*, thus providing two datasets that enhance previous understanding of the diversity of gokushoviruses in aquatic environments. The third chapter details ongoing efforts to characterize an elusive, putatively icosahedral ssDNA temperate phage from a *Synechococcus* culture.

References:

Ackermann HW (2007). 5500 Phages examined in the electron microscope. *Archives of Virology* **152:** 227-243.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006). The marine viromes of four oceanic regions. *PLoS Biology* **4:** e368.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* **99:** 14250-14255.

Breitbart M (2012). Marine viruses: truth or dare. *Annual Review of Marine Science* **4:** 425-448.

Brentlinger KL, Hafenstein S, Novak CR, Fane BA, Borgon R, McKenna R, Agbandje-McKenna M (2002). *Microviridae*, a family divided: isolation, characterization, and genome sequence of ϕMH2K, a bacteriophage of the obligate intracellular parasitic bacterium *Bdellovibrio bacteriovorus*. *Journal of Bacteriology* **184:** 1089-1094.

Brum JR, Schenck RO, Sullivan MB (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME Journal* **7:** 1738-1751.

Cherwa JE, Fane BA (2011). *Microviridae*: microviruses and gokushoviruses. *eLS*.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452:** 340-343.

Duffy S, Holmes EC (2009). Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology* **90:** 1539-1547.

Duffy, S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9:** 267-276.

Fane B. (2005). Family *Microviridae*. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) Virus Taxonomy, Classification and Nomenclature of Viruses, 8th ICTV Report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press: San Diego, USA.

Frederico LA, Kunkel TA, Shaw BR (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29:** 2532-2537.

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

Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M, Breitbart M (2014). Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *The ISME Journal* **doi:** 10.1039/ismej.2014.43.

Kim K-H, Chang H-W, Nam Y-D, Roh SW, Kim M-S, Sung Y, Jeon CO, Oh H-M, Bae J-W (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and Environmental Microbiology* **74:** 5975-5985.

Kim K-H, Bae J-W (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Applied and Environmental Microbiology* **77:** 7663-7668.

Krupovic M, Forterre P (2011). *Microviridae* goes temperate: Microvirus-related proviruses reside in the genomes of *Bacteroidetes*. *PLoS ONE* **6:** e19893.

Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiology and Molecular Biology Reviews* **75:** 610-635.

Labonté JM, Suttle CA (2013**a**). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME Journal* **7:** 2169-2177.

Labonté JM, Suttle CA (2013**b**). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Frontiers in Microbiology* **4**: 404.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009). High diversity of the viral community from an Antarctic lake. *Science* **326:** 858-861.

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

McDaniel LD, delarosa M, Paul JH (2006). Temperate and lytic cyanophages from the Gulf of Mexico. *Journal of the Marine Biological Association of the United Kingdom* **86:** 517-527.

McDaniel LD, Rosario K, Breitbart M, Paul JH (2013). Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental Microbiology* **16:** 570-585.

Raney JL, Delongchamp RR, Valentine CR (2004). Spontaneous mutant frequency and mutation spectrum for gene A of ΦX174 grown in *E. coli*. *Environmental and Molecular Mutagenesis* **44:** 119-127.

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

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

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D (2012a). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* **7:** e33641.

Roux S, Krupovic M, Poulet A, Debroas D, Enault F (2012b). Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE* **7:** e40418.

Sanger F, Air G, Barrell B, Brown N, Coulson A, Fiddes J, Slocombe P, Smith M (1977). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* **265:** 687-695.

Smith RJ, Jeffries TC, Roudnew B, Seymour JR, Fitch AJ, Simons KL, Speck PG, Newton K, Brown MH, Mitchell JG (2013). Confined aquifers as viral reservoirs. *Environmental Microbiology Reports* **5:** 725-730.

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

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G (2012). Are we missing half of the viruses in the ocean? *The ISME Journal* **7:** 672-679.

CHAPTER ONE:

Diversity of single-stranded DNA phages (Family *Microviridae*) in the environment revealed through PCR amplification of the partial major capsid protein

Max Hopkins[1], Shweta Kailasan[2], Allison Cohen[1], Simon Roux[4,5], Kimberly Pause Tucker[3], Amelia Shevenell[1], Mavis Agbandje-McKenna[2], Mya Breitbart[1*]

[1]College of Marine Science, University of South Florida

[2]Department of Biochemistry and Molecular Biology, University of Florida

[3]Department of Biology, Stevenson University

[4]Laboratoire "Microorganismes: Génome et Environnement", Clermont Université, Université Blaise Pascal

[5]CNRS, UMR 6023, LMGE

Publication citation: Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M, Breitbart M (2014). Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *The ISME Journal* **doi:** 10.1039/ismej.2014.43.

Summary:

The small single-stranded DNA (ssDNA) bacteriophages of the subfamily *Gokushovirinae* were traditionally perceived as narrowly-targeted, niche-specific viruses infecting obligate parasitic bacteria, such as *Chlamydia*. The advent of metagenomics revealed gokushoviruses to be widespread in global environmental samples. This chapter expands knowledge of gokushovirus diversity in the environment by developing a degenerate PCR assay to amplify a portion of the major capsid protein (MCP) gene of gokushoviruses. Over 500 amplicons were sequenced from ten diverse environmental samples (sediments, sewage, seawater and freshwater), revealing the ubiquity and high diversity of this understudied phage group. Residue-level conservation data generated from multiple alignments was combined with a predicted 3D-structure, revealing a tendency for structurally-internal residues to be more highly conserved than surface-presenting protein-protein or viral-host interaction domains. Aggregating this dataset into a phylogenetic framework, many gokushovirus MCP clades contained samples from multiple environments, although distinct clades dominated the different sample types. Antarctic sediment samples contained the most diverse gokushovirus communities, while freshwater springs from Florida were the least diverse. Whether the observed diversity is being driven by environmental factors or host-binding interactions remains an open question. The high environmental diversity of this previously overlooked ssDNA viral group necessitates further research elucidating their natural hosts and exploring their ecological roles.

Introduction:

The landmark viral metagenomic comparison of four ocean provinces by Angly et al (2006) study revealed significant presence of *Microviridae* in all environments except the Arctic Ocean. Subsequent viral metagenomic studies employing rolling circle amplification (RCA) have uncovered novel ssDNA phages in a variety of environments (reviewed in

Rosario and Breitbart 2011), including freshwater aquifers (Smith et al 2013), freshwater lakes (López-Bueno et al 2009; Roux et al 2012a), stromatolites (Desnues et al 2008), soils (Kim et al 2008), coastal estuaries (Labonté and Suttle 2013a; McDaniel et al 2008; McDaniel et al 2013), seawater (Labonté and Suttle 2013b) and reclaimed water (Rosario et al 2009). Although metagenomic studies generate sequence fragments, two complete gokushovirus genomes (SARssϕ1 and SARssϕ2) were assembled and PCR-verified from the Sargasso Sea (Tucker et al 2011) and a data-mining study assembled 81 additional complete *Microviridae* genome sequences from various environments and human gut/stool samples (Roux et al 2012b). Roux et al (2012b) revealed the tendency for dichotomous cladding of the *Gokusho-* subfamily between environmental (e.g., SARssϕ1 & -2) vs. 'human-associated' gokushoviruses.

The International Committee on the Taxonomy of Viruses, divides the *Microviridae* into two groups; the *Gokushovirinae* subfamily, and the enterobacteria-infecting ϕX174-type 'true' *Microvirus* genus, for which a subfamily has not been officially adopted (Fane 2005). Since the majority of ssDNA sequences that have been identified in environmental metagenomes are similar to gokushoviruses, this chapter further explores their diversity and environmental distribution by amplifying and sequencing a portion of the gokushovirus major capsid protein (MCP) gene. The selected MCP fragment is flanked by highly conserved motifs to enable efficient amplification and alignment and includes the hypervariable threefold loop believed to dictate host specificity. Results reveal diverse gokushoviruses in all environments examined, demonstrating that ssDNA phages are a pervasive but understudied component of the global environmental virome.

Methods:

Sample collection, processing, and DNA extraction. Samples from ten different sites were examined; six in Florida, USA and four from the Antarctic shelf. Several methods were used to purify viruses and concentrate DNA from these environmental samples, which were mostly samples of opportunity prepared for other projects. Surface water samples were collected in August 2012 from Wall Springs (freshwater (FW); 100 liters), Wall Estuary (saline (SW); 100 liters), and Bayboro Harbor (SW; 200 liters). GPS coordinates, salinity and temperature data are recorded in Table 1.1. Water was strained through 100 μm Nitex mesh, then concentrated down to ~100 ml using a 100 kD tangential flow filter (GE Healthcare, Pittsburg USA) as described previously (Thurber et al 2009). The retentate was filtered through a 0.22 μm Sterivex filter (Millipore, Billerica, MA) to remove bacteria and larger cells. Viral DNA was extracted from the concentrate using the MinElute Virus Spin Kit (Qiagen, Valencia, CA) following the standard kit protocol and eluted into 50 μl of water.

Freshwater samples were collected by snorkelers from Three Sisters Springs in Florida in May 2009. A sterile 60 ml syringe was used to collect 50 ml of water directly from the spring boil (~3 m below the surface). Water samples were immediately filtered through a 0.22 μm Sterivex filter and then on to a 0.02 μm Anotop filter (Whatman, Maidstone, UK), which was frozen at -80^{o}C until extraction. DNA was extracted from the Anotop filter using the Masterpure complete DNA and RNA purification kit (Epicenter, Madison, WI) as described previously (Culley and Steward 2007;Tucker et al 2011).

Surface sediment samples from Wall Spring, Wall Estuary and Hillsborough River were collected with conical tubes directly below their corresponding water samples. The Antarctic margin marine sediments (n=4; sites #4, 11, 14, 15) were collected in February 2012 during the British Services Antarctic Expedition (http://www.bsae2012.co.uk/science.html). Surface grab samples were taken in Marguerite Bay (~68°S, 68°W) from water depths between 200 and 425 m and the upper 0 to 2 cm of sediment was subsampled immediately into conical tubes and frozen at -20°C until processing. DNA extractions were performed from a starting mass of ~250 mg sediment. Sediment samples with high water content were first spun at 7000 xg for 4 min in order to obtain a cohesive sediment plug that could be adjusted for mass. The ~250 mg solid was combined with reagents from the PowerSoil® DNA extraction kit (MoBio, Carlsbad, CA, USA) and vigorously homogenized and disrupted with 1 min of bead beating followed by 10 min of vortexing. The extraction was performed following the manufacturer's protocol with a final elution volume of 100 µl.

The sewage sample was collected in February 2009 from a wastewater treatment plant in Manatee County, Florida. Virus particles were purified from 1.2 liters of sample by filtering through 0.45 µm and 0.2 µm Sterivex filters (Millipore, Billerica, MA, USA). Virus particles were further concentrated and purified using PEG precipitation followed by cesium chloride (CsCl) gradient centrifugation with composite collection in a density range from 1.2-1.5 g/ml (Thurber et al 2009). Viral DNA was extracted using the MinElute Virus Spin Kit (Qiagen, Valencia, CA, USA).

Degenerate PCR for amplification of Microviridae. Degenerate PCR primers were designed using the standalone version of the PhiSiGns utility (Dwivedi et al 2012). Initially, PhiSiGns failed to generate acceptable primers for all of the extant *Gokushovirinae* due to the highly divergent nature of SpV4; therefore, SpV4 was excluded from the design. Degenerate PCR primers MCPf (5'- CCYKGKYYNCARAAAGG – 3') and MCPr (5' – AHCKYTCYTGRTADCC – 3') are designed to amplify an 895 nt fragment of the major capsid protein (MCP) from the remaining extant *Gokushovirinae* (Chp1, NC_001741; Chp2, NC_002194; Chp3, NC_008355; Chp4, NC_007461; CPAR39, NC_002180; ϕCPG1, NC_001998; BdϕMH2K, NC_002643; SARssϕ1, HQ157199; SARssϕ2, HQ157198). These extant genomes from which the primers were derived are henceforth referred to as the nine 'reference genomes'.

To enrich for circular, single-stranded DNA templates, 1 μl of the extracted DNA from each sample was subjected to rolling circle amplification (TempliPhi; GE Healthcare, Piscataway, NJ) according to the manufacturer's instructions. This TempliPhi product was diluted ten-fold and used as the target for degenerate PCR. The 50 μl PCR mixture contained 1 U Apex *Taq* DNA polymerase (Genesee Scientific, San Diego, CA), 1X Apex *Taq* reaction buffer, 0.5 μM of each primer, 0.2 mM dNTPs, and 1 μl of the diluted TempliPhi product. The touchdown PCR conditions were (i) 3 min of initial denaturation at 94^{o}C; (ii) 32 cycles of 60 sec of denaturation (95^{o}C), 45 sec of annealing (47^{o}C with a 0.1^{o} decrease/cycle), 90 sec of extension (72^{o}C); and (iii) 10 min of final extension at 72^{o}C.

The resulting PCR products were visualized using gel electrophoresis. One sample, the Bayboro Harbor estuary concentrate, yielded multiple PCR products of different sizes, so the band most similar in size to the positive control was excised and gel-purified (Zymo, Irvine, CA). The verified PCR products were given a poly-adenine tail using Sigma Taq polymerase, ligated into TOPO TA vector (Invitrogen, Grand Island, NY) and subsequently transformed into OneShot© competent *E. coli* (Invitrogen) and plated with X-gal (20 mg/ml). White colonies were picked and inserts were size verified by PCR with M13 primers. 48 clones from each sample were Sanger sequenced with the M13F primer by Beckman Genomics (Danvers, MA).

Sequence analysis. Sequences were trimmed for quality and vector removal using Sequencher (Gene Codes, Ann Arbor, MI). Trimmed sequences were compared against the Genbank non-redundant (nr) database using a batch BLASTX search (cutoff e=.05) to confirm that the amplicons were similar to the MCP of known *Microviridae*. Sequences that did not have BLASTX similarity to *Microviridae* were considered to be nonspecific amplification and therefore removed from further analyses. *Microviridae* sequences were recovered with high efficiency from most environments, with the exception of the sewage sample, in which ~60% of the sequenced clones were not similar to *Microviridae*.

Sequences with BLASTX similarity to *Microviridae* were dereplicated at the 97% nucleotide level using FastGroup II (Yu et al 2006), which was also used to compute the Shannon-Weiner Diversity Index (Shannon and Weaver 1949). The sequences were provisionally translated into amino acid format and aligned using the ClustalW algorithm in MEGA5 with subsequent manual adjustment (Tamura et al 2011). After obtaining optimal amino acid alignment, the alignment was back-toggled and exported for phylogenetic construction. The phylogeny in Figure 1.2 was generated using the PhyML package (Guindon et al 2010); a maximum-likelihood method employing the GTR model with support values determined by approximate Likelihood-Ratio Test (Anisimova and Gascuel 2006). The phylogeny in Figure 1.3 is a maximum-likelihood tree generated in the FastTree package (Price et al 2009) using the Whelan-And-Goldman residue model from a maximum-likelihood training ‘intree’ generated in MEGA5 from a ClustalW alignment (Tamura et al 2011). The clade-based hidden Markov models combined in Supplementary Figure 1.1 were calculated using ‘hmmbuild’ within the HMMER3.0 package (Eddy 2008) and visualized using LogoMat (Schuster-Böckler et al 2004).

Homology Modeling of Gokushoviruses. Structural models of gokushovirus MCPs were built using the homology model-building package *MODELER* (Yang et al 2012). The full-length MCP sequences of the reference extant gokushoviruses (six Chlamydia phages, SARssϕ1&2, BdϕMH2k) were aligned against the MCP of *Microviridae* with available structures (ϕX174, Bacteriophage alpha-3, G4, SpV4) using *CLUSTAL-W2* (Larkin et al 2007). Presence of large insertions (>80 amino acids) at the 3-fold loop region as seen in SpV4 prompted use of the pseudo-atomic model of SpV4 instead of the higher-resolution coliphages (ϕX174, α3 and G4) models as the primary template for model building. Superposition of the homology models was carried out in the *COOT* package (Emsley et al 2010). The online server, *VIPERdb2* (Carrillo-Tripp et al 2009), was used to generate a capsid composed of 60 identical copies of the MCP by icosahedral matrix multiplication (Fig 1.1). *UCSF-CHIMERA* (Pettersen et al 2004) was used to calculate percent conservation values based on the presence of the most prevalent residue at a particular position in the alignment of all the selected gokushoviruses. These values were projected onto a ribbon representation of Chp1 using *PyMOL* (Schrödinger, LLC). Surface representation of Chp1 was generated in *UCSF-CHIMERA.*

Results and Discussion:

Building upon the initial discovery of gokushoviruses in a wide range of natural environments, this study designed a degenerate PCR assay to amplify a portion of the gokushovirus major capsid protein (MCP). Although the amplification of genes conserved within specific viral families (i.e., signature genes) is commonly used to explore the diversity and environmental distribution of dsDNA phages (e.g., Filée et al 2005), this is the first study to examine ssDNA phage diversity with such an approach. The gokushovirus MCP amplicon contains regions of both low and high conservation, presenting an ideal target for studying the diversity of these phages in the environment. The 5' portion of the amplicon is dominated by the hypervariable, threefold interaction loop, while the 3' portion includes three of the eight β-sheets (βF-βG) that comprise the 'β-barrel' motif common to all *Microviridae* (Bull et al 2000) as well as many other viral families with T=1 capsids (Agbandje-McKenna and Kleinschmidt 2011).

Gokushovirus MCPs were recovered from all environments tested (freshwater, estuarine, sediments, sewage). A total of 537 sequences were retained following BLASTX parsing, which were then dereplicated within each sample at 97% identity with gaps, yielding 315 sequences for downstream analyses (Genbank Accession No. KF689226 - KF689540), which are represented in the Figure 1.2 phylogeny. Notably, the average size of the aligned amplicons from the environmental samples was 636 ± 22 nt, compared to 705 ± 32 nt in the Chlamydia phages. The difference in length between environmental and cultured gokushoviruses was largely driven by differences in the length of the threefold loop; in the Chlamydia phages the loop-coding region had an average length of 230 ± 23 nt while in the environmental samples the same region averaged 166 ± 19 nt. The smaller size of environmental phage amplicons as compared to cultured isolates has also been reported for dsDNA phages (Breitbart et al 2004), although the reason for this discrepancy is unknown.

A capsid homology model was generated to examine the amplicon from a structural perspective. High-resolution crystal structures with ~3-3.5 Å resolution are available for bacteriophages ϕX174 (PDB ID: 1ALO), A3 (PDB ID: 1MO6), and G4 (PDB ID: 1GFF), which are all members of the enterobacteria-infecting 'true' Microvirus genus (family *Microviridae*). Structures of 'true' Microvirus capsids by X-ray crystallography revealed internal and external scaffolding proteins, a spike-protein and DNA-binding proteins (named B, D, G and J respectively), in addition to the major capsid protein F (Dokland et al 1997). However, the nucleotide-level sequence identity of the MCPs of the nine reference gokushoviruses to those structurally resolved 'true' microviruses only ranged between 18-20%. A notable difference between the gokushovirus targeted by this study and the 'true' microvirus MCPs is that the 'true' microviruses do not carry large insertions loops (>80 residues) between strands βE and βF found at the 3-fold axis of symmetry (see Fig 1.1).

Although there are no high-resolution models for gokushoviruses, a pseudo-atomic 27Å resolution model for the gokushovirus SpV4 built into cryo-reconstructed density is available (Chipman et al 1998). The major capsid protein encoded by the gokushovirus SpV4 is homologous to the F proteins in enterobacterial 'true' microviruses (like ϕX174, a3, ϕK and G4) and shares a canonical, eight-stranded β-motif (Chipman et al 1998). The gokushovirus genomes do not encode for the pentameric G proteins, which create star-shaped spikes at each of the twelve five-fold vertices of the ϕX174 capsid. Instead, pseudo-atomic modeling of the SpV4 MCP, the only gokushovirus for which a structure has been solved, albeit in low resolution cryo-reconstructed density, suggested the presence of “mushroom-like” protrusions on the surface formed by prominent loops found at each 3-fold axis of symmetry of the MCP (Chipman et al 1998). The gokushoviruses also lack external scaffolding proteins. However, gokushoviruses do encode a VP3 capsid protein which is lost during the maturation of procapsids to infectious virions in Chlamydia phage 2 (Clarke et al 2004). In spite of sharing a low sequence homology, its role is considered analogous to internal scaffold protein B.

Due to the closer sequence homology and presence of the threefold loop, the SpV4 MCP pseudo-atomic model was used as a template to build a homology model for Chp1. A conservation percentage for our total environmental data set (n=315 sequences) was calculated using *UCSF-CHIMERA* based on the presence of the most prevalent residue at a particular position in the full alignment. The conservation percentage was projected onto a threaded model of Chp1 in a red-to-blue spectrum of the least-to-most conserved regions for the amplified region specifically (Fig 1.1).

Because of its prominent, surface-protruding location the ‘threefold loop’ has been predicted to be important for host specificity in the gokushovirus SpV4 (Chipman et al 1998). As further evidence for this hypothesis, an experimental study demonstrated that three Chlamydia phages with the same sequence in their threefold loop motif had the same host-infectivity range (Everson et al. 2003). The reported hypervariability within this region is therefore a proposed mechanism for accessing new host types. Our aggregated conservation analysis found similar hypervariability in the threefold loops of environmental gokushovirus MCP sequences, with conservation ≤10% for much of the length (Fig 1.1).

Only at the downstream base of the threefold protrusion where there is a prominent α-helix did the residue conservation rise to >50%. This helix is one of the most highly conserved motifs of these amplicons and implies that it is inherent to the environmental gokushoviruses as it is to SpV4 and ϕX174-type phages for which the structure is known (Chipman et al 1998;McKenna et al 1992;McKenna et al 1996).

The residues participating in the formation of the first β-strand downstream of the threefold insertion loop, βF (Chipman et al 1998), which is part of the eight sheet β-barrel core, had conservation values of approximately 50%. This conservation percentage in the aligned residues rapidly degenerated in the succeeding loop connecting strands βF and βG. The degenerate connecting loop is modeled to interact with four other protein monomers at the fivefold axis of symmetry, contravening the paradigm of multimeric interaction forcing genetic purity (Bahadur and Janin 2008). The residues in strands βG and βH had successively higher levels of conservation, rising to levels greater than 80%. The high level of conservation maintained at the β-core and α-helical regions suggests the importance of these residues in viral capsid assembly. Strand βH runs internally through the core of the structure, emerging to participate in a twofold interaction with an adjacent monomer at the twofold axis of symmetry, indicated by the oval in Figure 1.1. Residues in the loop emerging from this strand show a rapid shift from a high to low value of conservation percentage.

Overall, the environmental gokushovirus MCP amplicons show poor conservation on the surface-exposed regions, while maintaining high conservation at the interior of the capsid. This model suggests that the high sequence and possibly structural variance at these surface-exposed regions may facilitate rapid co-evolution of phages with their hosts and allow for exploration of new host space (Breitbart 2012;Paterson et al 2010).

Phylogenies built with the full-dataset alignment yielded chaotic, irreproducible trees due to an inability to align the highly divergent threefold loop region (<10% conservation). Upon removal of the threefold loop region, more robust alignments were achieved, revealing a phylogeny with many long-branch singletons and clustered clades of varying cohesion. Despite the fact that all the Chlamydia phages were included as reference genomes when designing the degenerate primers used in this study, the environmental amplicons are only distantly related to these cultured isolates (Fig 1.2). The tight clading of the cultured gokushovirus sequences adjacent to those recovered in this study, combined with the aforementioned finding of aberrantly-long threefold loops in the cultured isolates, reinforces the notion that the 'type strains' for the gokushoviruses are not close representatives of the ssDNA phages that dominate in the environment. It is notable that some of the recovered sequences did cluster with SARssϕ1 and SARssϕ2, uncultured gokushoviruses that were assembled from a metagenomic survey of the Sargasso Sea (Tucker et al 2011).

Overall, an extremely broad diversity of novel gokushovirus MCP sequences was recovered with these primers, reflecting results seen in signature gene amplification studies of dsDNA phages (e.g., (Filée et al 2005;Goldsmith et al 2011)). The recovered level of gokushovirus MCP diversity varied across the different environments. This is quantified in Table 1.1 using the Shannon-Wiener Diversity Index (Shannon and Weaver 1949) as computed in FastgroupII (Yu et al 2006). This diversity metric has been criticized as providing a biologically meaningless numerical output in $\log_e$ 'nats', as well as being highly sensitive to inexhaustive 'species' sampling (Magurran 2004), a shortcoming from our perspective since it is highly unlikely that we have exhaustively sampled all of the gokushovirus 'species' from even the most homogenous of our sample sites. Furthermore it is possible that certain methods used to process our samples (e.g., CsCl centrifugation) may have biased the recovery of gokushoviruses. However, since approximately equal sequencing efforts were applied to each site these diversity estimates are useful for comparison between sites (Soetaert and Heip 1990). The sewage site was excluded from the diversity calculation because of the smaller number of sequences from this site. The Antarctic sediment samples were characterized by extremely high diversity with Shannon scores exceeding 3 nats. Almost all of the combined 196 amplicons from Antarctic sediment samples were unique (i.e., <97% identical), demonstrating that far more sequencing is needed to comprehensively document the gokushovirus diversity in Antarctic margin marine sediments. The next highest diversity was recovered from the Bayboro Harbor estuary, likely due to its combination of marine and terrestrial runoff inputs. The riverine systems (Hillsborough River, Wall Springs) had an intermediate level of diversity. The lowest diversity was found in the pristine Florida spring site (Three Sisters Springs), where the 46 sequences obtained dereplicated into only 6 'unique' sequences.

The phylogenetic analysis reveals some clustering by sample type and location (Fig 1.2). The primary partition in the tree (indicated by a dashed black line) is between the Antarctic sediment sequences (blue squares in Clade 1°A, shaded by site) and the other samples, which all originated from Florida, USA (Clade 1°B). Due to the limited number of samples analyzed in this study and the number of variables differentiating each site, it is not possible to determine which variable (e.g., geography, temperature, depth, salinity) or combination of variables is responsible for this dichotomy in the data. Interestingly, the only other sample type to recruit into the right-half of the tree in any significant abundance also originated from brackish sediments (pink squares, taken from the shallow water interface of Wall Estuary), suggesting that Clade 1^{o}A sequences may be more prevalent in saline sediments than those belonging to Clade 1^{o}B (Fig 1.2).

The point-source radiation of several sequence clusters should be considered as a false attraction of several long-branches; there should be underlying sequence similarity among the grouped branches but the point source rooting of the cluster is an artifact of the radiating tree diagram. However, several secondary clades with high support were identified in the data, especially from the 1°B portion of the tree. Noteworthy secondary clades include: clade 2°A, populated exclusively with sewage sequences; clade 2°B, populated almost exclusively by spring water sequences; the sprawling clade 2°C of saltwater and Antarctic sediment sequences; clade 2°D, containing sequences from two saltwater sites; and 2°E which contains a mixture of springwater and saltwater sequences including SARssϕ1, somewhat removed on a long branch (Fig 1.2). None of the sample sites (icon colors) or sample types (icon shapes) shows exclusive recruitment into a single monophyletic clade that would be the hallmark of pure environmental forcing. To highlight the biochemical differences driving the topology of the phylogeny in Figure 1.2, the sequences belonging to each of these 2° clades were aligned and HMM logos were generated using HMMbuild (Eddy 2008) and visualized using LogoMat (Schuster-Böckler et al 2004). The HMM profiles of each clade were manually aligned relative to each other in an effort to juxtapose homologous regions. The HMMs are shown in Supplementary Figure 1.1.

To place the current dataset in the larger context of *Microviridae* diversity, representative sequences from throughout the tree in Figure 1.2 were integrated into the comprehensive *Microviridae* MCP alignment created by Roux et al. (2012b). This alignment was supplemented with sequences with strong gokushoviral BLASTX hits (e-value≤0.01) from two recent marine metagenomic studies; a single gokushovirus capsid sequence from the dataset of Labonté and Suttle (2013) and 41 sequences from the recent viral community analysis of hadopelagic sediments off of Japan (Yoshida et al 2013). The resulting combined phylogenetic tree is visualized in Figure 1.3, which contains as an outgroup the *Microviridae* subfamily most closely related to *Gokushovirinae,* tentatively named the *Pichovirinae* (Roux et al 2012b).

Within the *Gokushovirinae* subfamily, Roux et al. (2012b) depicted a phylogenetic topology consisting of three apparently coherent groups; two of “Eukaryote-associated” and one of “Environmental” strains from freshwater metagenomes as well as SARssϕ2 and BdϕMH2K. Of those Eukaryote-associated clusters from Roux et al. (2012b), the clade containing human gut associates (as well as a turkey gut associate, ‘Microvirus CA82’) was highly supported and distinct, whereas the clustering of the Chlamydia phages and other human gut associates occupied an unstable position adjacent to the Environmental clade, which suggested a more recent divergence.

This topology is broadly recapitulated in our expanded phylogeny (Fig 1.3), where a well-supported clade of eukaryote-associates, including turkey gut-derived CA82 and human gut derivatives, continues to occur. The fact that a similar bifurcated topology has now been reproduced by both primer-based and primer-independent metagenomic methods (Roux et al 2012b) suggests that there may be a true split between environmental and eukaryotic-associated *Gokushovirinae*. Applying PCR primers designed based on these eukaryotic-associated sequences to environmental samples would enable further assessment of this phylogenetic split.

The weaker eukaryotic-associated clade containing the Chlamydia phages from Roux et al. (2012b) now nests within the "Environmentally-dominated clade", which has been significantly expanded through this study. It is not surprising that the primers used in this study amplified sequences most closely related to the Chlamydia phages, since the primers were designed based largely on the Chlamydia phages (6 of 9 reference sequences). However, all but one of the amplicons that was generated using the primers (colored icons) belonged to the "Environmentally-dominated clade", regardless of the sample type (sewage, aquatic, sediment). We look forward to future work testing both environmental and eukaryote-derived with the same primer sets to determine whether this persistent split is a real feature of the gokushoviral topology.

Conclusion:

The discovery of diverse ssDNA phages in all environments tested is highly significant and prompts many questions for future studies. At present, the hosts for these environmental gokushoviruses remain unknown, as do the ecological effects of these phages on their hosts and ecosystems. To date, all cultured gokushoviruses infect intracellular parasites, a possibility that must be considered when attempting to culture environmental gokushoviruses. Phage BdϕMH2K, infecting the obligate intracellular parasite *Bdellovibrio bacteriovorus*, is currently the only cultured phage belonging to the Environmentally-dominated clade. As opposed to *Chlamydia*, which are obligate parasites of eukaryotic organisms, *Bdellovibrio* parasitizes gram-negative bacteria that are far more abundant in the environment than their eukaryotic counterparts. If the targeting of obligate intracellular parasitic bacteria continues to hold true as a hallmark of the *Gokushovirinae*, parasites of abundant bacteria and single-celled eukaryotes may prove fruitful as an avenue for exploring the hosts of the environmental gokushoviruses.

This study, taken together with the data mining work of Roux et al. (2012b), demonstrates the diverse and cosmopolitan nature of the *Gokushovirinae* subfamily, changing the perception of this group of ssDNA phages from one with a fairly narrow, primarily eukaryote-associated niche to a group of importance for microbial ecology. Another significant implication of these data is that studies utilizing nucleic acid staining and epifluorescence microscopy to enumerate environmental viruses (Patel et al 2007) may be underestimating total viral abundance. The small genome sizes of gokushoviruses and other ssDNA phages produce a weak fluorescence signal that is below the detection limit of most microscopes and flow cytometers (Tomaru and Nagasaki 2007). Along with other recent work (Brum et al 2013;Labonté and Suttle 2013;Steward et al 2012), this study emphasizes the need for a shift in the paradigm that dsDNA *Caudovirales* dominate environmental viral communities.

Acknowledgments:

Funding for this study was provided by grants to MB from the National Science Foundation (Microbial Interactions and Processes MCB-0701984 and Division of Biological Infrastructure DBI-0850206) and the University of South Florida Internal Awards Program. MAM and SK were funded by University of Florida COM Research Funds; MH was funded by a Presidential Fellowship from the University of South Florida; AC was funded by an REU Supplement from the National Science Foundation; SR was supported by a PhD grant from the French Defense Procurement Agency (DGA); AS was funded by NSF ANT-1246378.

Thanks to the members of the British Services Antarctic Expedition 2012 for collecting the Marguerite Bay marine sediment samples and to Dawn Goldsmith and Bhakti Dwivedi for assistance with bioinformatics and primer design, respectively.

References:

Agbandje-McKenna M, Kleinschmidt J (2011). AAV Capsid Structure and Cell Interactions. In: Snyder R, Moullier P (eds). *Methods in Molecular Biology: Adeno-Associated Virus; Methods and Protocols*. Springer. pp 47-92.

Anisimova M, Gascuel O (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology* **55:** 539-552.

Bahadur RP, Janin J (2008). Residue conservation in viral capsid assembly. *Proteins: Structure, Function, and Bioinformatics* **71:** 407-414.

Breitbart M, Miyake JH, Rohwer F (2004). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiology Letters* **236:** 249-256.

Breitbart M (2012). Marine viruses: truth or dare. *Annual Review of Marine Science* **4:** 425-448.

Brum JR, Schenck RO, Sullivan MB (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME Journal* **7:** 1738-1751.

Bull J, Badgett M, Wichman H (2000). Big-benefit mutations in a bacteriophage inhibited with heat. *Molecular Biology and Evolution* **17:** 942-950.

Carrillo-Tripp M, Shepherd CM, Borelli IA, Venkataraman S, Lander G, Natarajan P, Johnson JE, Brooks CL, Reddy VS (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Research* **37:** D436-D442.

Chipman PR, Agbandje-McKenna M, Renaudin J, Baker TS, McKenna R (1998). Structural analysis of the spiroplasma virus, SpV4: implications for evolutionary variation to obtain host diversity among the *Microviridae*. *Structure* **6:** 135-145.

Clarke IN, Cutcliffe LT, Everson JS, Garner SA, Lambden PR, Pead PJ, Pickett MA, Brentlinger KL, Fane BA (2004). Chlamydiaphage Chp2, a skeleton in the ϕX174 closet: scaffolding protein and procapsid identification. *Journal of Bacteriology* **186:** 7571-7574.

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

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452:** 340-343.

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

Eddy SR (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Computational Biology* **4:** e1000069.

Emsley P, Lohkamp B, Scott W, Cowtan K (2010). Features and development of Coot. *Acta Crystallographica* **D66**: 486-501.

Fane B. (2005). Family *Microviridae*. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) Virus Taxonomy, Classification and Nomenclature of Viruses, 8th ICTV Report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press: San Diego, USA.

Filée J, Tétart F, Suttle CA, Krisch H (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 12471-12476.

Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa R-A, Breitbart M (2011). Development of *phoH* as a novel signature gene for assessing marine phage diversity. *Applied and Environmental Microbiology* **77:** 7730-7739.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59:** 307-321.

Kim K-H, Chang H-W, Nam Y-D, Roh SW, Kim M-S, Sung Y, Jeon CO, Oh H-M, Bae J-W (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and Environmental Microbiology* **74:** 5975-5985.

Labonté JM, Suttle CA (2013**a**). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME Journal* **7:** 2169-2177.

Labonté JM, Suttle CA (2013**b**). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Frontiers in Microbiology* **4**: 404.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson J, Gibson T (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009). High diversity of the viral community from an Antarctic lake. *Science* **326:** 858-861.

Magurran AE (2004). *Measuring biological diversity*. Blackwell Science Ltd: Malden MA.

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

McDaniel LD, Rosario K, Breitbart M, Paul JH (2013). Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental Microbiology* **16:** 570-585.

McKenna R, Xia D, Willingmann P, Hag LL, Krishnaswamy S, Rossmann MG, Olson NH, Baker TS, Incardona NL (1992). Atomic structure of single-stranded DNA bacteriophage ΦX174 and its functional implications. *Nature* **355:** 137.

McKenna R, Bowman BR, Ilag LL, Rossmann MG, Fane BA (1996). Atomic structure of the degraded procapsid particle of the bacteriophage G4: induced structural changes in the presence of calcium ions and functional implications. *Journal of Molecular Biology* **256:** 736-750.

Patel A, Noble RT, Steele JA, Schwalbach MS, Hewson I, Fuhrman JA (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols***:** 269-276.

Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, Quail M, Smith F, Walker D, Libberton B, Fenton A, Hall N, Brockhurst M (2010). Antagonistic coevolution accelerates molecular evolution. *Nature* **464:** 275-278.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**: 1605-1612.

Price MN, Dehal PS, Arkin AP (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26:** 1641-1650.

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

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

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D (2012a). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* **7:** e33641.

Roux S, Krupovic M, Poulet A, Debroas D, Enault F (2012b). Evolution and Diversity of the *Microviridae* Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE* **7:** e40418.

Schuster-Böckler B, Schultz J, Rahmann S (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics* **5**.

Shannon CE, Weaver W (1949). *The mathematical theory of communication* University of Illinois Press: Urbana, IL.

Smith RJ, Jeffries TC, Roudnew B, Seymour JR, Fitch AJ, Simons KL, Speck PG, Newton K, Brown MH, Mitchell JG (2013). Confined aquifers as viral reservoirs. *Environmental Microbiology Reports* **5:** 725-730.

Soetaert K, Heip C (1990). Sample-size dependence of diversity indices and the determination of sufficient sample size in a high-diversity deep-sea environment. *Marine Ecology Progress Series* **59:** 305-307.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G (2012). Are we missing half of the viruses in the ocean? *The ISME Journal* **7:** 672-679.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28:** 2731-2739.

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

Tomaru Y, Nagasaki K (2007). Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *Journal of Oceanography* **63:** 215-221.

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

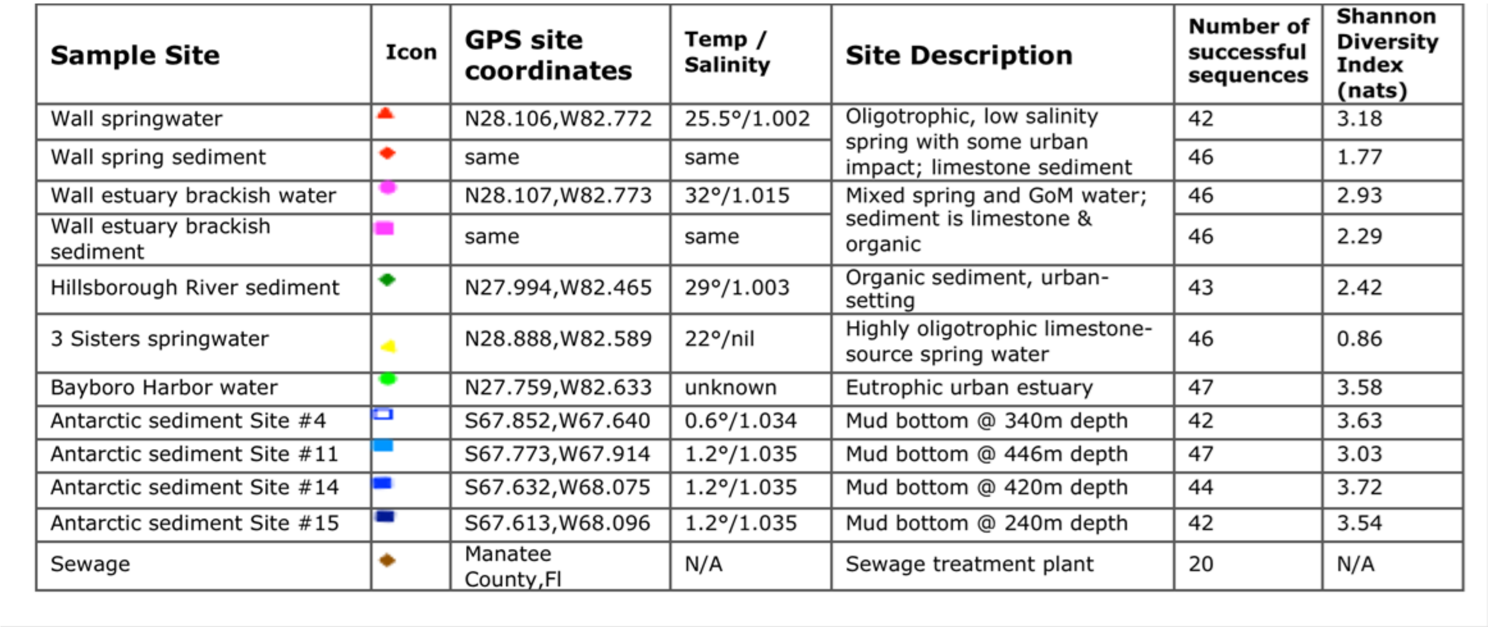
Yang Z, Lasker K, Schneidman-Duhovny D, Webb B, Huang CC, Pettersen EF, Goddard TD, Meng EC, Sali A, Ferrin TE (2012). UCSF Chimera, MODELLER, and IMP: An integrated modeling system. *Journal of Structural Biology* **179**: 269-278.

Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K (2013). Metagenomic analysis of viral communities in (hado) pelagic sediments. *PLoS ONE* **8:** e57271.

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

Tables and Figures:

Table 1.1**:** Description of samples processed in chapter one, including available metadata and results from diversity analysis.



Sample Site	Icon	GPS site coordinates	Temp / Salinity	Site Description	Number of successful sequences	Shannon Diversity Index (nats)
Wall springwater	▲	N28.106,W82.772	25.5°/1.002	Oligotrophic, low salinity spring with some urban impact; limestone sediment	42	3.18
Wall spring sediment	◆	same	same		46	1.77
Wall estuary brackish water	●	N28.107,W82.773	32°/1.015	Mixed spring and GoM water; sediment is limestone & organic	46	2.93
Wall estuary brackish sediment	■	same	same		46	2.29
Hillsborough River sediment	◆	N27.994,W82.465	29°/1.003	Organic sediment, urban-setting	43	2.42
3 Sisters springwater	◀	N28.888,W82.589	22°/nil	Highly oligotrophic limestone-source spring water	46	0.86
Bayboro Harbor water	●	N27.759,W82.633	unknown	Eutrophic urban estuary	47	3.58
Antarctic sediment Site #4	□	S67.852,W67.640	0.6°/1.034	Mud bottom @ 340m depth	42	3.63
Antarctic sediment Site #11	■	S67.773,W67.914	1.2°/1.035	Mud bottom @ 446m depth	47	3.03
Antarctic sediment Site #14	■	S67.632,W68.075	1.2°/1.035	Mud bottom @ 420m depth	44	3.72
Antarctic sediment Site #15	■	S67.613,W68.096	1.2°/1.035	Mud bottom @ 240m depth	42	3.54
Sewage	◆	Manatee County,Fl	N/A	Sewage treatment plant	20	N/A

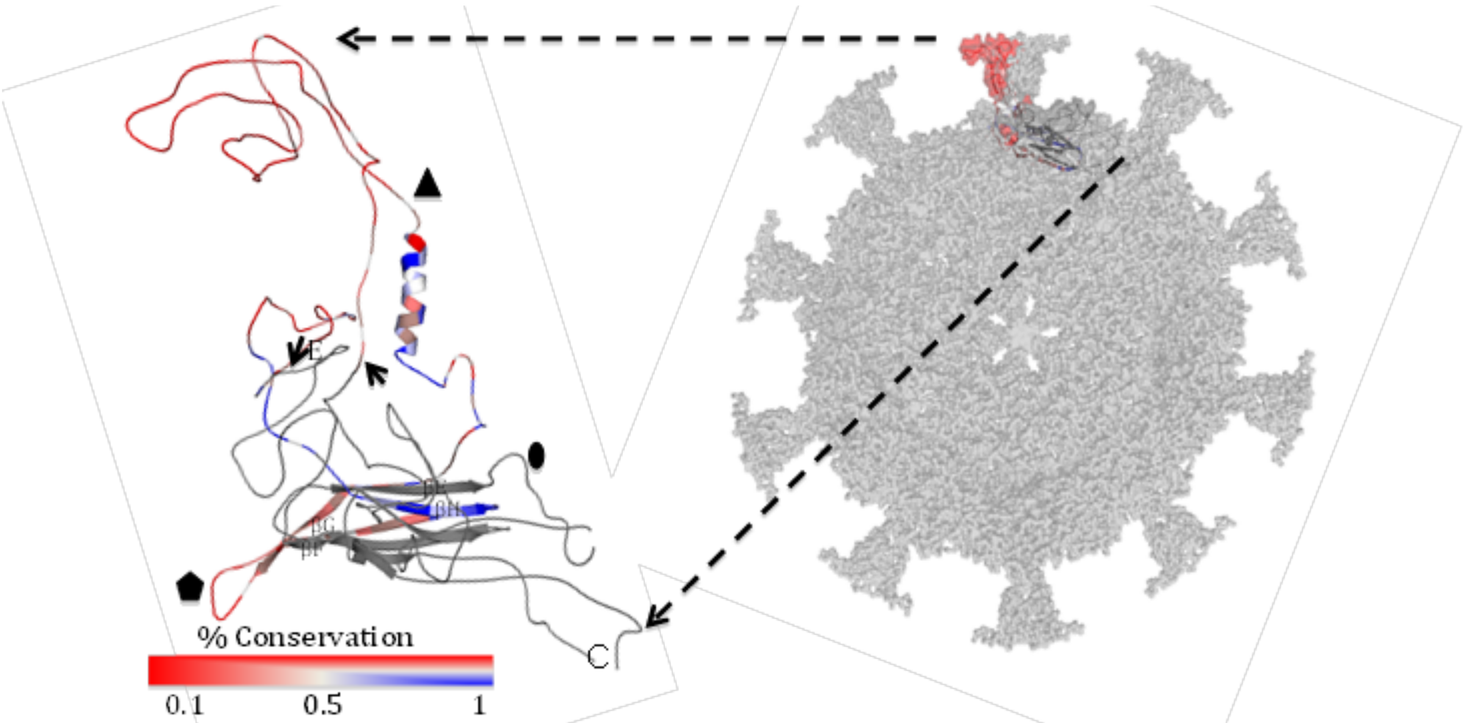
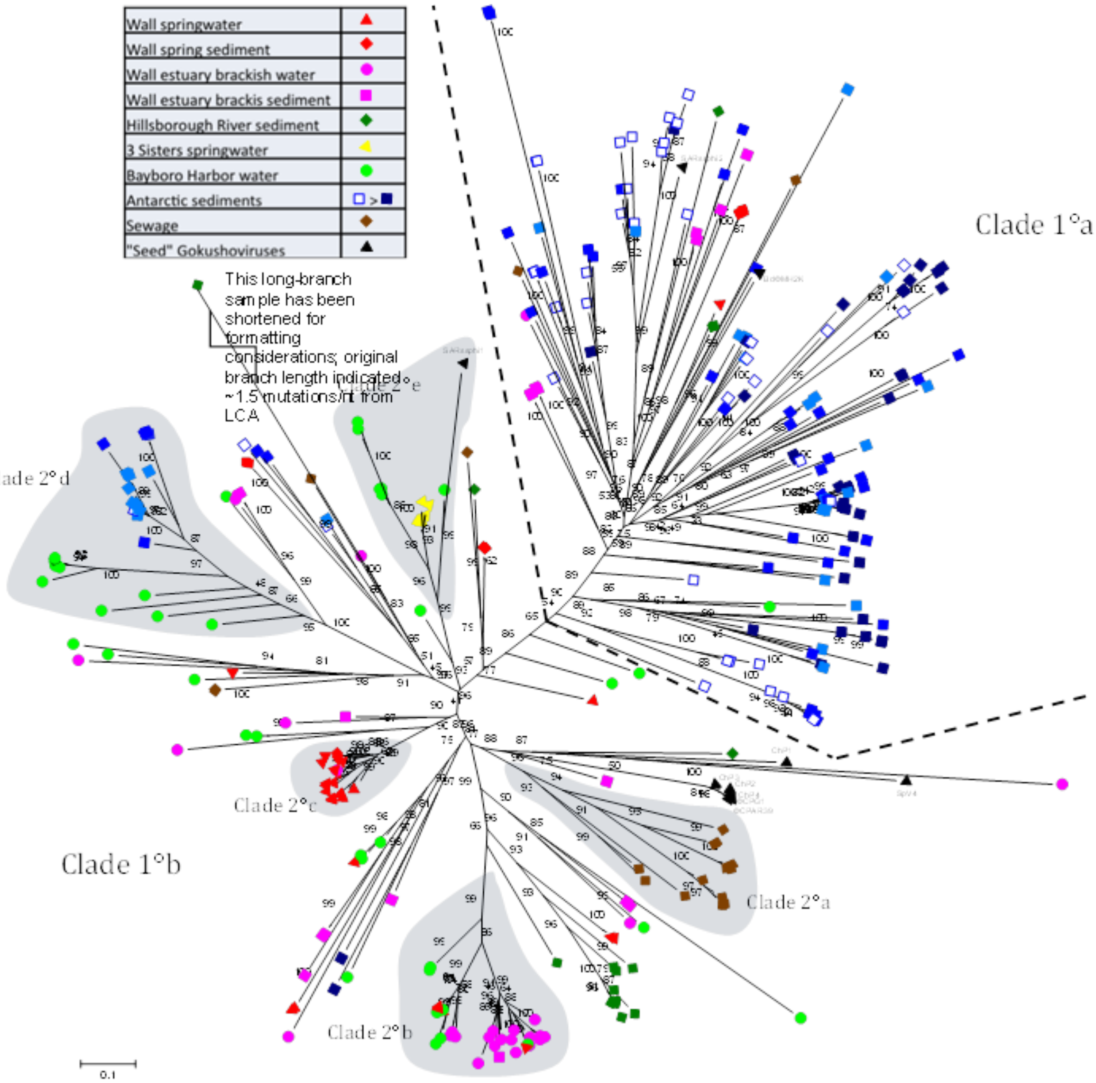


Figure 1.1: Surface representation of a Chlamydiaphage-1 (ChP1) capsid homology model is shown in grey looking down the 5-fold axis of symmetry of an icosahedron (right). The inset to the left is a cartoon representation of predicted ChP1 MCP homology model, in which the amplicon region from this study is highlighted with arrows (S: Start; E: End) and has been colored according to residue conservation of the full aligned dataset from blue (most conserved; 1) to red (least conserved; 0.1). The N-terminal (N) and C-terminal (C) end of the MCP have been labeled along with the 2-, 3- and 5-fold axes of symmetry for an icosahedron shown as an oval, triangle and pentagon, respectively. The β-strands (βE-G), of the eight stranded β-barrel, that are contained within the amplicon and referenced in the discussion are also labeled.



This long-branch sample has been shortened for formatting considerations; original branch length indicated ~1.5 mutations/nt from LCA

Figure 1.2: Unrooted maximum likelihood phylogeny of 316 novel Gokushovirus MCP sequences, unique at 97%. Statistical support values are percentages calculated by the aLRT method. Sequences are coded by color (sampling site) and shape (sample type), as shown in the legend. A primary division in the dataset is shown with a dashed line and annotation, and small secondary clades of interest are indicated by shading and annotation. The icons are dereplicated, unique sequences, which in one instance represents as many as 19 recovered sequences.

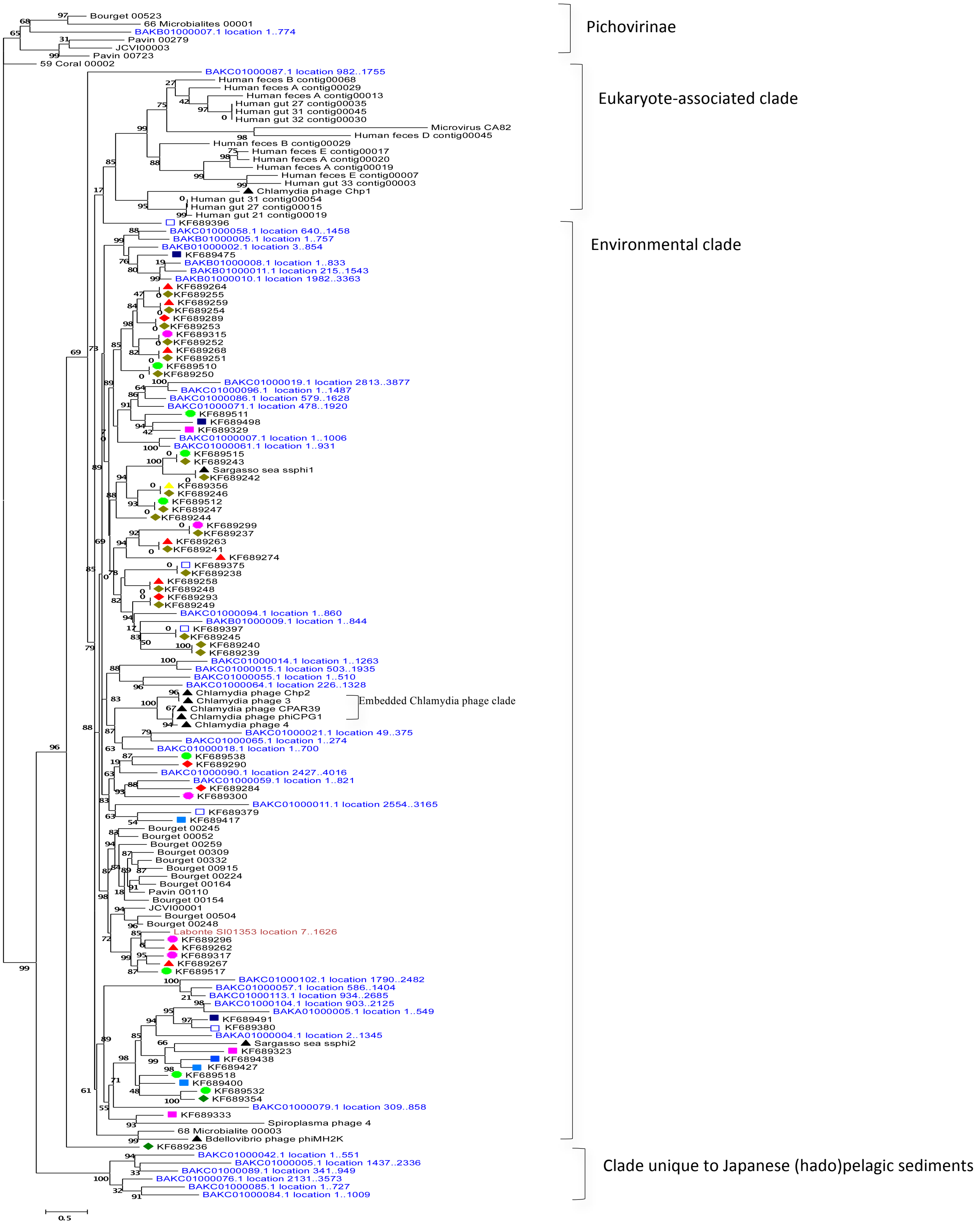

Bourget 00523
66 Microbialites 00001
BAKB01000007.1 location 1..774
Pavin 00279
JCVI00003
Pavin 00723
59 Coral 00002
Pichovirinae
BAKC01000087.1 location 982..1755
Human feces B contig00068
Human feces A contig00029
Human feces A contig00013
Human gut 27 contig00035
Human gut 31 contig00045
Human gut 32 contig00030
Microvirus CA82
Human feces D contig00045
Human feces B contig00029
Human feces E contig00017
Human feces A contig00020
Human feces A contig00019
Human feces E contig00007
Human gut 33 contig00003
Chlamydia phage Chp1
Human gut 31 contig00054
Human gut 27 contig00015
Human gut 21 contig00019
Eukaryote-associated clade
KF689396
BAKC01000058.1 location 640..1458
BAKB01000005.1 location 1..757
BAKB01000002.1 location 3..854
KF689475
BAKB01000008.1 location 1..833
BAKB01000011.1 location 215..1543
BAKB01000010.1 location 1982..3363
Environmental clade
KF689264
KF689255
KF689259
KF689254
KF689289
KF689253
KF689315
KF689252
KF689268
KF689251
KF689510
KF689250
BAKC01000019.1 location 2813..3877
BAKC01000096.1 location 1..1487
BAKC01000086.1 location 579..1628
BAKC01000071.1 location 478..1920
KF689511
KF689498
KF689329
BAKC01000007.1 location 1..1006
BAKC01000061.1 location 1..931
KF689515
KF689243
Sargasso sea ssphi1
KF689242
KF689356
KF689246
KF689512
KF689247
KF689244
KF689299
KF689237
KF689263
KF689241
KF689274
KF689375
KF689238
KF689258
KF689248
KF689293
KF689249
BAKC01000094.1 location 1..860
BAKB01000009.1 location 1..844
KF689397
KF689245
KF689240
KF689239
BAKC01000014.1 location 1..1263
BAKC01000015.1 location 503..1935
BAKC01000055.1 location 1..510
BAKC01000064.1 location 226..1328
Chlamydia phage Chp2
Chlamydia phage 3
Chlamydia phage CPAR39
Chlamydia phage phiCPG1
Chlamydia phage 4
Embedded Chlamydia phage clade
BAKC01000021.1 location 49..375
BAKC01000065.1 location 1..274
BAKC01000018.1 location 1..700
KF689538
KF689290
BAKC01000090.1 location 2427..4016
BAKC01000059.1 location 1..821
KF689284
KF689300
BAKC01000011.1 location 2554..3165
KF689379
KF689417
Bourget 00245
Bourget 00052
Bourget 00259
Bourget 00309
Bourget 00332
Bourget 00915
Bourget 00224
Bourget 00164
Pavin 00110
Bourget 00154
JCVI00001
Bourget 00504
Bourget 00248
Labonte SI01353 location 7..1626
KF689296
KF689262
KF689317
KF689267
KF689517
BAKC01000102.1 location 1790..2482
BAKC01000057.1 location 586..1404
BAKC01000113.1 location 934..2685
BAKC01000104.1 location 903..2125
BAKA01000005.1 location 1..549
KF689491
KF689380
BAKA01000004.1 location 2..1345
Sargasso sea ssphi2
KF689323
KF689438
KF689427
KF689518
KF689400
KF689532
KF689354
BAKC01000079.1 location 309..858
KF689333
Spiroplasma phage 4
68 Microbialite 00003
Bdellovibrio phage phiMH2K
KF689236
BAKC01000042.1 location 1..551
BAKC01000005.1 location 1437..2336
BAKC01000089.1 location 341..949
BAKC01000076.1 location 2131..3573
BAKC01000085.1 location 1..727
BAKC01000084.1 location 1..1009
Clade unique to Japanese (hado)pelagic sediments
0.5

Figure 1.3: Rooted neighbor-joining tree combining sequences from (Roux, Krupovic et al. 2012) (n=60, black typeface with no icons), sequences from (Yoshida et al. 2013) (n=40, blue typeface), a sequence from (Labonté and Suttle 2013) (n=1, red typeface) and sequences from this study (43 environmental, 20 sewage; black typeface with icons from Figure 2) to give a comprehensive view of the *Gokushovirinae,* with the *Pichovirinae* as an outgroup at the top. Bootstrap values were calculated out of 100 replicates. Clades referenced in the discussion are annotated with brackets to the right.

CHAPTER TWO:

Interannual survey of the diversity of single-stranded DNA phages (Family *Microviridae*) at the Bermuda Atlantic Time-series Study by depth and season

Max Hopkins[1], Dawn Goldsmith[1], Mya Breitbart[1]

[1]College of Marine Science, University of South Florida

Summary:

Since 1988, the Bermuda Atlantic Time-series Study (BATS) has collected monthly data and samples in the Sargasso Sea. As a consequence the water column dynamics are among the best understood of any marine study site in the world. The mixing regime is known to undergo a marked shift between winter, when there is extensive mixing, and summer when stratification occurs. Here we present signature gene data on the diversity of the small single-stranded DNA *Gokushovirinae* bacteriophages across depths, seasons and years. This data supplements a recent landmark study by Parsons et al (2011) showing an annually-reoccurring summertime subsurface peak in viral abundance at BATS. We hypothesized that during winter months, when mixing occurs throughout the upper 300 m, the gokushovirus communities at 0 m and 100 m would be similar. Conversely, in summer when the water column is stratified, heterogeneity ought to emerge between the gokushovirus communities at 0 m and 100 m. The analyses presented here lend strong

support to this hypothesis and also suggest interesting connectivity between seasonal and interannual depth cohorts.

Introduction:

The Bermuda Atlantic Time-series Study (BATS) was established in 1988 as one of two time-series studies funded under the auspice of the Joint Global Ocean Flux Study (Michaels and Knap 1996). The BATS site at (31°40'N,64°10'W), typically approximated to within 20 km at time of sampling, lies at 4500 m depth at the site of a preexisting deep sea sediment flux study. BATS was positioned south of the venerable Hydrostation S (Schroeder and Stommel 1969) to be out of the lee of the Bermuda seamount from the prevailing southwesterly current and thus better reflect pelagic conditions throughout a deeper water column. Net flow is 5 cm/s to the southwest as a result of a Gulfstream countercurrent, although local mesoscale eddies create local flow conditions as high as 50 cm/s (Siegel and Deuser 1997).

Sampling cruises have occurred approximately monthly since 1988 with no major gaps (Steinberg et al 2001). The use of a 24 x 12 liter rosette allows the full water column to be sampled to <100 m resolution in two casts (Michaels and Knap 1996). Core measurements collected in real time by an attached CTD and those performed later on the samples are physical, chemical and basic biological parameters such as Chlorophyll-a concentration. The combined continuity and depth resolution of the time series make the annual and interannual dynamics of this water column among the best-understood of any offshore study site. An example of the phenomena that such abundant data has brought to light is the annually recurring shift in the depth of the wind-driven mixing layer. In winter months (December-March) when the strongest wind shear occurs, the mixed layer depth (MLD) extends through the upper 150 m and as deep as 300 m in some years (Steinberg et al 2001). Conversely, in the summer (June-September) when winds are light, the water column stratifies such that only the upper 50 m is being actively mixed.

'Ancillary studies', many requiring extra casts from the cruise vessel, have provided some of the most interesting additional short-term time series data to enrich the core time series (Michaels and Knap 1996). A thymidine-incorporation study by Carlson et al (1996), was the first to report on an annually reoccurring 40-80 m subsurface maximum in bacterial biomass and production which begins in late spring and persists through summer every year. Ancillary viral data began to be added at BATS comparatively recently. In 2006, Angly et al compared a viral metagenome taken at 80 m at BATS against viral metagenomes from three other ocean provinces (Angly et al 2006). Somewhat surprisingly, the BATS viral metagenome revealed high occurrence of heretofore human- and agriculturally-associated *Gokushovirinae*. Viral time-series data was incorporated for the first time between 2000-2010 in a landmark decadal study by Parsons et al (2011). The study showed that viral abundance peaked in early fall lagging the summertime bacterial peak by two or three months. This general observation contained nuances based on host taxa, such as the fact that although viral abundance covaries positively with *Prochlorococcus*, it correlates negatively with SAR11 and *Synechococcus* abundance. Such paired time series data allows for the possibility of creating 'guilt-by-association' inferences about phage-host relationships.

In 2011, informed by the Angly et al findings, the Breitbart Lab sequenced two complete gokushovirus genomes from BATS named as SARssϕ1 (Genbank accession number: HQ157199) and SARssϕ2 (Genbank accession number: HQ157198) (Tucker et al 2011). At the time, these were the first gokushoviruses to be fully sequenced from an environmental sample. The primers used in this study are based in part on these genomes (see Chapter 1; Hopkins et al. 2014). The primers are used to characterize the 0 m and 100 m gokushovirus communities at BATS in March and September during three years. We hypothesize that during winter months, when mixing occurs throughout the upper 300 m, the gokushovirus communities at 0 m and 100 m will be similar. Conversely, in summer when the water column is stratified, the 0 m and 100 m communities ought to differentiate from one another.

Methods:

Sample collection and preparation. The data reported here are from samples collected from BATS in March and September of 2008, 2010 and 2011 at both 0 m and 100 m. The 2010 and 2011 samples were collected through additional Niskin bottle casts to the standard 24 bottle rosette casts used to generate the BATS time series. The 2008 samples were taken during a cruise planned and funded by Dr. Mya Breitbart (NSF Funding MCB-0701984).

An average volume of 245 liters (ranging from 90-383 liters; see Table 2.1) was tangential-flow filtered (TFF) with a molecular weight cutoff of 100 kDa, resulting in a final concentrate volume of approximately 50 ml for a several thousand-fold concentration factor. The concentrates were 0.22 μm filtered to remove bacteria and stored at 4°C until Polyethylene Glycol (PEG) treatment and Cesium Chloride (CsCl) density centrifugation following the protocol of (Thurber et al 2009). PEG 8000 was used to pellet the viruses out of the concentrate and the resuspended pellet was loaded onto a CsCl density step gradient (1.2 g/ml; 1.5 g/ml; 1.7 g/ml) and ultracentrifuged (22,000 rpm on a Beckman SW40Ti rotor for 3 hours at 4°C).

A CsCl density fraction of 1.5 g/ml was collected by puncturing through the sidewall of the ultracentrifuge tube and collecting the outflow. Formamide extraction was performed on the concentrated viruses following the protocol of Green and Sambrook (2012). The CsCl viral fractions from September 2008 were additionally centrifugally concentrated using a Microcon centrifugal filter (Millipore, Billerica, MA) according to the manufacturer's instructions. Following formamide extraction DNA was resuspended in sterile water. The extracted DNA was then amplified using the RCA-based GenomiPhi V2 DNA kit (GE Healthcare, Piscataway, NJ).

Degenerate PCR for amplification of Microviridae. Degenerate PCR primers MCPf (5'- CCYKGKYYNCARAAAGG – 3') and MCPr (5' – AHCKYTCYTGRTADCC – 3') were designed using the standalone version of the PhiSiGns utility (Dwivedi et al 2012) based upon extant *Gokushovirinae* at the time (Chp1, NC_001741; Chp2, NC_002194; Chp3, NC_008355; Chp4, NC_007461; CPAR39, NC_002180; ϕCPG1, NC_001998; BdϕMH2K, NC_002643; SARssϕ1, HQ157199; SARssϕ2, HQ157198). The primers amplify a 900 nt fragment of the major capsid protein (MCP). These primers were applied as documented in Chapter 1 and Hopkins et al (2014). The 50 µL PCR mix consisted 1 U Apex Taq DNA polymerase, 1X Apex Taq reaction buffer, 0.5 µM of each primer, 0.2 mM dNTPs and 1 µL of template DNA (GenomiPhi product). The touchdown PCR conditions were (i) 3 min of initial denaturation at 94ºC; (ii) 32 cycles of 60 s of denaturation (95ºC), 45 s of annealing (47ºC with a 0.11ºC decrease/cycle), and 90 s of extension (72ºC); and (iii) 10 min of final extension at 72ºC.

Cloning and sequencing of MCP amplicon. The PCR product was visualized through gel electrophoresis. In the case of March 2010, 0 m and 100 m, multiple band sizes were present, so in this situation the band of interest was excised and purified using the Zymoclean DNA Gel Recovery kit (Zymo, Irvine, CA). All other single band samples were purified using the DNA Clean & Concentrator -25 kit (Zymo, Irvine, CA). Terminal 3' adenylation was performed using Sigma-Aldrich REDTaq (Sigma-Aldrich, St. Louis, MO) at 72°C for 10 min. The adenylated products were ligated into the TOPO TA cloning vector for sequencing (Invitrogen, Carlsbad, CA) and transformed through heatshock into competent DH5α *E. coli* cells. Transformed cells were plated overnight on plates containing ampicillin (50 μg/ml of media) and X-gal (100 μl per large plate). White colonies were screened, and positive transformants with correctly-sized inserts were sequenced with the M13F primer by Beckman Genomics (Danvers, MA). There were marked differences in efficiency of transformation between depth-time cohorts. A full 96 well plate of clones was sent for sequencing for 11 depth-time cohorts, however the number of sequence returns that were Blast-identified as *Gokushovirinae* is recorded in Table 2.1. For the 0 m March 2010 sample, multiple attempts were made but not enough transformed clones could be harvested to warrant the sequencing.

Data analysis for signature genes. Data processing was performed entirely by Dr. Dawn Goldsmith. Initially, vector and low-quality sequences were trimmed with Sequencher 4.7 (Gene Codes, Ann Arbor, MI). The sequences were then dereplicated at the 98% nucleotide sequence identity level with gaps using FastGroupII (Yu et al 2006). Dereplicated BATS sequences were aligned with reference sequences at the amino acid level using Muscle (Edgar 2004) with the default parameters as implemented by

TranslatorX (Abascal et al 2010). Regions of low conservation were trimmed using Gblocks (implemented by TranslatorX) using the options for a less stringent selection (Talavera and Castresana 2007). Back-translated nucleotide alignments were used to build maximum-likelihood phylogenetic trees with FastTree version 2.1 (Price et al 2010). Branch supports in FastTree were calculated using the Shimodaira-Hasegawa-like approximate likelihood ratio test on 1000 resamplings. TreeCollapseCL 4 (Hodcroft 2013) was used to collapse branches with support below 50. Hierarchical clustering was performed in R (Pinheiro et al 2013) from a Bray-Curtis dissimilarity matrix based on OTU abundance data using the picante package (Kembel et al 2010). Jaccard stability means were computed using the fpc package (Hennig 2013) to bootstrap the dendrograms. The Jaccard similarity value, which represents the stability of the cluster, is averaged for every bootstrapping of the clustering (1000 times), resulting in a Jaccard stability mean for each cluster. Clusters with Jaccard stability means of 75 and greater are considered valid, stable clusters, while clusters with Jaccard stability means between 60 and 75 indicate patterns in the data (Hennig 2013). Both the Figure 2.1 dendrogram and the Figure 2.2 phylogeny are the sole work of Dr. Dawn Goldsmith.

Results and Discussion:

The dendrogram in Figure 2.1 was generated by assigning all 843 sequences, irrespective of depth-time cohort, into 163 OTUs at 98% nucleotide similarity. Each depth-time cohort was then defined as being composed of a unique subset of the full-dataset OTUs; a 'fingerprint' of sorts. A Bray-Curtis algorithm determined the pairwise distance between each cohort based on OTU composition, so cohorts that are more alike recruit together into clades. Figures 2.1 and 2.2 share the same legend coding, such that shapes denote the sample year, which are then subdivided between March (winter) in purple and September (summer) in green. The monthly subsets are further distinguished by shading, between the 0 m surface samples (light shade) and the 100 m depth samples (dark shade). Figure 2.1 shows that in two years (2008 and 2011) the purple March samples tend to cluster by year, irrespective of depth shading. Alternatively, for the green colored September cohorts, when clustering is apparent it is only by shading (depth), irrespective of year (shape).

Figure 2.1 shows that the mixing layer depth (MLD) is the dominant driver of similarity or differentiation between the *Gokushovirinae* cohorts at the MCP gene sequence level. When the water column is well mixed throughout the upper 100 m, as it typically is in March (Steinberg et al 2001), the viral populations at 0 m and 100 m are blended together to create a relatively homogeneous viral population (different shades of purple). This result is also reflected in the Figure 2.2 phylogeny wherein two clades (arbitrarily designated Clade “A” and “B”) contain the majority of March sequences. Clade A contains the light and dark purple triangles denoting March 2008 sequences, whereas Clade B contains a more dispersed clade of light and dark purple circles from the March 2011 cohort. The fact that these March surface and 100 m depth sequences clade together on both the dendrogram and the phylogeny shows that the *Gokushovirinae* at these depths in winter months represent a unified genetic community as judged by MCP coding sequence. While the March samples are closely related between the surface and 100 m, the phylogenetic clades in which the purple icons dominate tend to be populated by a single year (shape) suggesting that there is little continuity from year to year in the gokushovirus community.

As shown in both the dendrogram (Figure 2.1) and the phylogeny (Figure 2.2), in September when the mixed layer is shallow and the water column is stratified the sequences segregate according to depth (shade of green). This is in keeping with the overall hypothesis of this study and with the aforementioned findings for March. Interestingly, however there appears to be more continuity across years in the September 100 m (dark green shading) gokushovirus community as shown by the dendrogram clustering of different years (green shapes). This is especially true for the 2010 and 2011 samples (dark green square and circle, respectively), which are temporally consecutive samples, though not for the further temporally removed September 2008 100 m sequences (dark green triangles). This result is apparent from the phylogeny in the clades designated "C" and "D". There is also evidence of connectivity between the 100 m gokushovirus populations in stratified September water column and well-mixed March water column, which is evident in Clade B. This may imply that sub-surface gokushoviruses act as a source of continuity and are periodically raised to the surface through mixing dynamics where they encounter novel environmental and evolutionary pressures.

Although there is a degree of interannual continuity in the September 100 m cohorts, the September 0 m sequences (light green shading) account for many of the long-branch singletons in the phylogeny in Figure 2.2. A potential reason for this is that viruses recovered from surface waters in summer are subject to a high degree of UV-light stress (Wommack et al 1996). UV-B wavelengths (280-320 nm) have been shown to be especially deleterious to viruses, creating pyrimidine dimers (Weinbauer et al 1997). This light stress has the net effect of raising mutation rates among exposed viruses. This may be all the more true of ssDNA-based viruses which have been shown to have spontaneous mutation rates as high as 10^{-3} substitutions/site/year, on par with rates for RNA viruses (Duffy et al 2008).

Conclusion:

Sequencing of the *Gokushovirinae* major capsid protein signature gene from 11 depth-time samples from BATS supported the hypothesis that the depth of the mixed layer either unifies or differentiates the viral communities between the surface and at depth. In March, when the near-surface water column is well mixed, the viral communities at 0 m and 100 m were highly similar, clading together by both dendrogram and phylogenetic analyses. In contrast, in September, when the water column is stratified, the viral populations at 0 m and 100 m formed distinct clades. Further research should focus on unifying phage signature gene data with bacterial taxonomical abundance data to examine correlative inferences about which hosts these phages are infecting. Knowledge of hosts would allow deconvolution of the forcing due to physical factors (e.g. mixing depth) versus host factors.

Acknowledgments:

Funding for this research was granted to Mya Breitbart by the National Science Foundation (NSF Funding MCB-0701984). MH and DG were both funded under the USF Presidential Fellowship. Special thanks to Rachel Parsons of BIOS for providing the samples and for her foundational work and to the crew of the R/V Atlantic Explorer, from which the samples were collected.

References:

Abascal F, Zardoya R, Telford MJ (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38:** W7-W13.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006). The marine viromes of four oceanic regions. *PLoS Biology* **4:** e368.

Carlson CA, Ducklow HW, Sleeter TD (1996). Stocks and dynamics of bacterioplankton in the northwestern Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **43:** 491-515.

Duffy S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9:** 267-276.

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

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

Goldsmith, D (2014). PhD Dissertation, "Chapter 3: Depth and seasonal variation in viral diversity in the northwestern Sargasso Sea." Submission to the USF Electronic Thesis and Dissertation database pending.

Green MR, Sambrook J (2012). *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, New York.

Hennig C (2013). fpc: Flexible procedures for clustering. R package version 2.1-5.

Hodcroft E (2013). TreeCollapseCL 4.

Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M, Breitbart M (2014). Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *The ISME Journal* **doi:** 10.1039/ismej.2014.43.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26:** 1463-1464.

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

Parsons RJ, Breitbart M, Lomas MW, Carlson CA (2011). Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *The ISME Journal* **6:** 273-284.

Pinheiro J, Bates D, DebRoy SS, Sarkar D (2013). D., and the R Development Core Team 2013. nlme: Linear and Nonlinear Mixed Effects Models. *R package version*: 3.1-103.

Price MN, Dehal PS, Arkin AP (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5:** e9490.

Schroeder E, Stommel H (1969). How representative is the series of Panulirus stations of monthly mean conditions off Bermuda? *Progress in Oceanography* **5:** 31-40.

Siegel D, Deuser W (1997). Trajectories of sinking particles in the Sargasso Sea: modeling of statistical funnels above deep-ocean sediment traps. *Deep Sea Research Part I: Oceanographic Research Papers* **44:** 1519-1541.

Steinberg DK, Carlson CA, Bates NR, Johnson RJ, Michaels AF, Knap AH (2001). Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 1405-1447.

Talavera G, Castresana J (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56:** 564-577.

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

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

Weinbauer MG, Wilhelm SW, Suttle CA, Garza DR (1997). Photoreactivation compensates for UV damage and restores infectivity to natural marine virus communities. *Applied and Environmental Microbiology* **63:** 2200-2205.

Wommack KE, Hill RT, Muller TA, Colwell RR (1996). Effects of sunlight on bacteriophage viability and structure. *Applied and Environmental Microbiology* **62:** 1336-1341.

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

Tables and Figures:

Table 2.1: Sample collection data including viral counts and number of sequences obtained from each sample in chapter two. This table was adapted from the PhD dissertation of Goldsmith 2014.

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

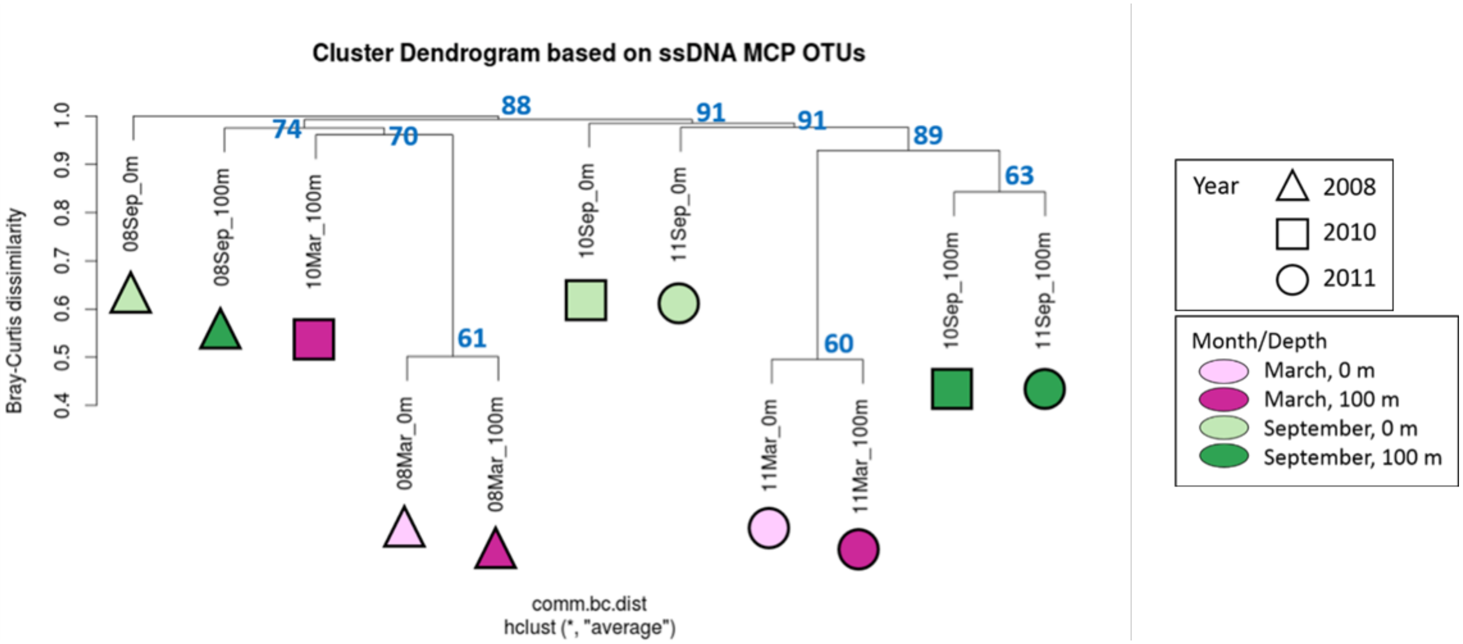


Figure 2.1: Dendrogram illustrating hierarchical clustering of Sargasso Sea samples based on ssDNA MCP OTUs (98% sequence identity). Clustering is calculated from Bray-Curtis dissimilarity of the samples. Branch supports are shown where support is greater than 50 and represent Jaccard stability means. Jaccard stability means > 75 represent valid, stable clusters. Jaccard stability means from 60 to 75 indicate the presence of patterns in the data. Figure from Goldsmith (2014).

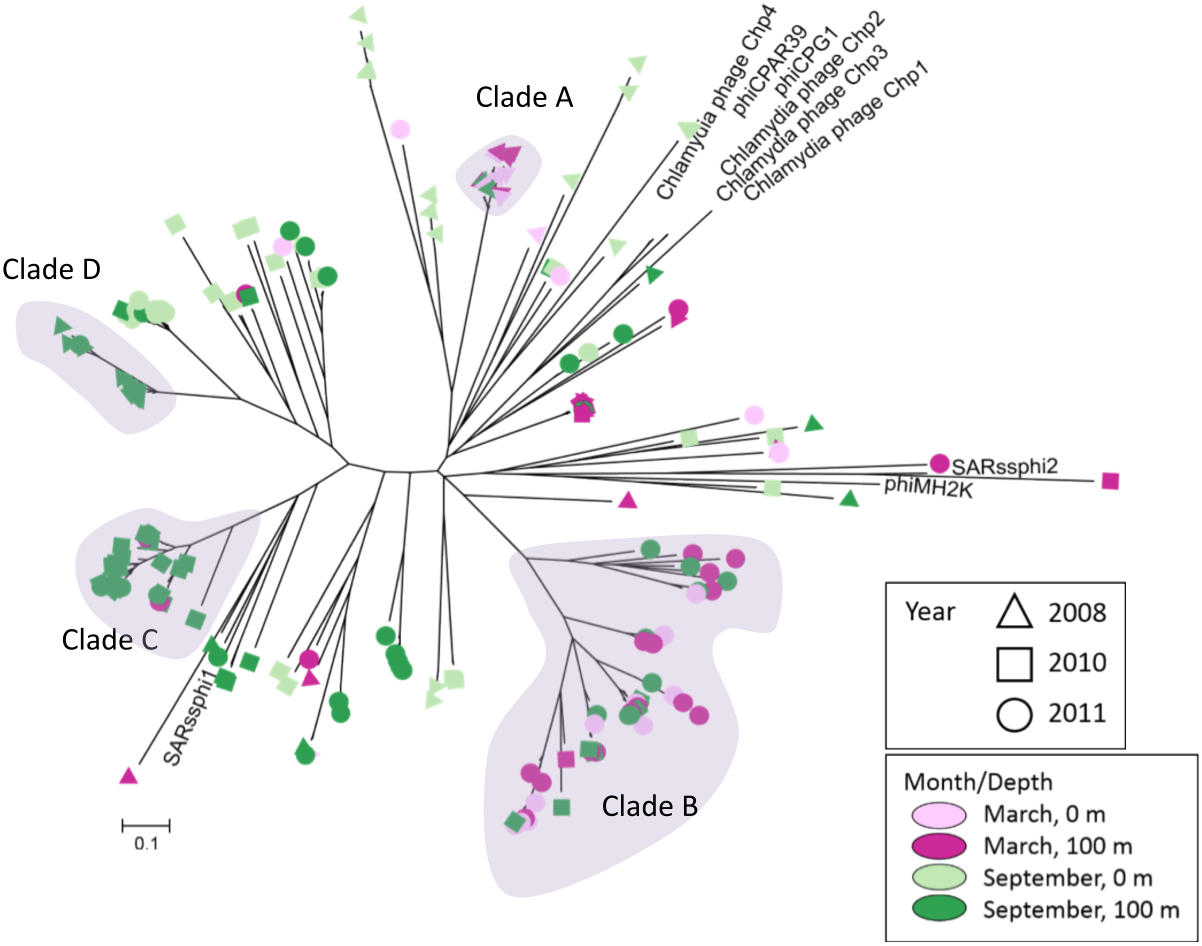


Figure 2.2: Phylogenetic tree showing the relationship among gokushovirus MCP sequences from environmental viruses sampled in the Sargasso Sea (indicated by colored shapes) and MCP sequences from fully sequenced reference gokushoviruses (indicated by names). The scale bar represents substitutions per site. Figure from Goldsmith (2014).

CHAPTER THREE: *Synechococcus* N phage

Summary:

There is currently overrepresentation of double-stranded DNA (dsDNA) bacteriophages in culture collections worldwide. The few single-stranded DNA (ssDNA) phages that are in culture tend to be clinical isolates such as those infecting *Chlamydia* or *E. coli*. While recent studies (including the first two chapters of this thesis) have identified a large diversity of ssDNA viruses in the marine environment, there are currently no marine ssDNA phage isolates in culture. This is a critical deficiency if knowledge of this under-studied portion of the global virome is to be advanced, since a great many assays require a cultured model system. This chapter details efforts to characterize one such potential model system.

Introduction:

In 2006, Dr. Lauren McDaniel (committee member) reported a small icosahedral phage particle that was inducible by Mitomycin C or continuous high light from strain 'N' (GM9914) of a non-axenic, clonal *Synechococcus* culture collection of 26 isolates (designated A-Z) from the eastern Gulf of Mexico (McDaniel et al 2006). The fuchsia-colored isolate contains phycoerythrin, placing it into cluster 5.1 of Bergey's taxonomical system (Garrity et al 2004). Dr. McDaniel sequenced the large subunit of the ribulosebisphosphate carboxylase gene (RbcL) and on the basis of this signature gene sequence determined that the N strain was closely related to Woods Hole strain 7803 (WH7803) (McDaniel et al 2006).

Previous experimentation demonstrated that exposure of the culture to Mitomycin C, or shifting the light regime from a 12 hr-light, 12 hr-dark diel cycle to intense 24 hr high light when the culture is at logarithmic growth phase causes a rapid degradation in *Synechococcus* cells and a concomitant rise in small virus-like particles. The particles are distinct but near the limit of detection for epifluorescence microscopy using SYBR Gold (Life Technologies, Grand Island, NY). Particle concentrates spotted onto formvar grids and counter-stained with uranyl acetate were viewed by electron microscopy for this study (Figure 3.1) and by McDaniel et al (2006). Micrographs from both studies identify spherical, somewhat amorphous, particles approximately 60 nm in diameter with no visible tail appendage.

Using Oligreen DNA stain (Life Technologies, Grand Island, NY) to quantify DNA degradation, Dr. McDaniel applied the ssDNA-preferential restriction enzyme S-1 nuclease to DNA extractions from N phage. Phage λ (dsDNA) was used as a negative control and Phi X174 (ssDNA) was used as a positive control. The N phage experienced a 3x reduction in quantifiable DNA versus Phi X174 and a 5x reduction versus the mostly undigested Lambda DNA (McDaniel et al 2006). The result suggested that the small icosahedral particles most likely contained ssDNA, and/or that its nucleic acid was subject to unusually rapid degradation or at least poor retention of fluorescent signal.

N phage is a highly desirable subject for genomic sequencing since it represents the second report of a temperate, small ssDNA icosahedral phage following the report of temperate *Microviridae* (Krupovic and Forterre 2011) and the only one yet to be experimentally induced. In addition, the putative host for N phage is *Synechococcus*, which would make N phage the first ssDNA phage of cyanobacteria, and indeed the first temperate phage of cyanobacteria if conclusively proved, following more tenuous reports (e.g., Sullivan et al 2009). However, caution is needed in drawing these conclusions since the *Synechococcus* culture is not axenic and the nucleic acid type and genome of N phage has not been definitively characterized.

Methods and Protocols:

Growth and culture maintenance: The cultures were maintained and grown on SN ("Seawater Natural") media adapted from Waterbury et al (1988). Oligotrophic seawater, collected on a cruise from offshore in the Gulf of Mexico and subsequently stored in the darkness for several months before use, makes up the base of the SN media. The seawater is 0.2μm filtered, autoclaved and then left overnight to off-gas, before the addition of six supplemental nutrient solutions:

- Na_2EDTA (1 g/L, pH 8.0): 5.6 ml/liter SN
- Na_2CO_3 (4 g/L): 2.6 ml/liter SN
- K_2HPO_4 (6.1 g/L): 2.6 ml/liter SN
- $NaNO_3$ (300 g/L): 2.5 ml/liter SN
- Vitamin B-12 (1.0 mg/L): 1 ml/liter SN
- Cyano Trace Metals: 1 ml/liter SN
    - $ZnSO_4$ x $7H_2O$ (0.222 g/L)
    - $MnCl_2$ x $4H_2O$ (1.4 g/L)
    - $Co(NO_3)_2$ x $6H_2O$ (0.025 g/L)
    - Na_2MoO_4 x $2H_2O$ (0.390 g/L)
    - Citric acid hydrate (6.250 g/L)
    - Ferric ammonium citrate (6.0 g/L)

Once the nutrients have been added the media must be used within a few days' time, although this can be extended with refrigeration.

In an acid-washed sterile 500 ml Erlenmeyer flask, 150-225 ml of fresh SN was inoculated with 10-25% by volume of seed culture in stationary phase. Previous work by Dr. McDaniel indicated that a total volume of <250 ml is most conducive to a high titer viral induction. The newly inoculated flasks were placed into a white light incubator at 26°C on a 12/12 light-dark cycle simulating environmental conditions. Optical density was measured at least once per day at a 750 nm wavelength with the logarithmic growth phase indicated by A_{750nm}=0.090. When multiple flasks were grown, they were shifted closer or further from the light relative to each other in order to stimulate or suppress the growth and synchronize the simultaneous arrival of all flasks at logarithmic phase.

When the synchronized growth flasks had reached log phase growth as indicated by Optical Density, they were moved to a high light table, for viral induction by round-the-clock light. Dr. McDaniel recommends at least 90 µmol photons $m^{-2}s^{-2}$ of light energy. Fluorescent tube lights (48-inch) suspended <2 ft above the flasks have been particularly effective for large-scale inductions.

After an initial 12-18 hr high light period, when photo-bleaching of the pink *Synechococcus* has become apparent, the flasks are continuously monitored by epifluorescence microscopy counts of virus-like particles (VLPs). One milliliter of bleached culture is centrifuged at 17,000 x g to pellet out cellular detritus. For a strong induction a 1:20-dilution is appropriate for enumeration, so 50 µl of supernatant is diluted into 950 µl of diluent containing 2% w/v formalin. The full milliliter of diluted phage is vacuum filtered onto a 0.02 µm filter (Whatman, Pittsburgh, PA), stained in the dark for 11 minutes with 25x SYBR Gold stain (Life Technologies, Grand Island, NY) and mounted on a slide with antifade solution (Glyerol/PBS buffer with 0.1% final concentration p-Phenylenediaminedi-HCl). Viruses are enumerated under a 100x, oil immersion objective with a 10x ocular. When the flasks have achieved a high viral titer of 10^8-10^9 particles per milliliter, they are removed from the high light table for downstream processing.

Viral purification and concentration: Irrespective of whether the targeted viral fraction is nucleic acids or protein, the initial concentration steps have been the same, adapted from Molecular Cloning (Green and Sambrook 2012). The step-by-step protocol is included as Appendix #1. Initially, the photobleached culture is transferred to a 460 ml centrifuge bottle into which 1M NaCl by weight is dissolved, in order to keep the viral particles in suspension by preventing their ionic association to the cellular detritus pellet. The large-scale culture is spun at 9500 x g for 10 min, pelleting cellular debris. The clear supernatant is decanted onto a 0.2 µm Nalgene filter tower, without displacing the large, colorful pellet. The attached sterile bottle containing the filtrate is disengaged from the tower and the next step is nuclease digestion to degrade any non-encapsidated nucleic acids. Nuclease digestions were typically run 1-2 hours at room temperature with RNase ONE (Promega, Madison, WI) and either RQ-1 DNase (Promega, Madison, WI) for genomic isolations or the less expensive bovine pancreatic DNase I (Sigma, St. Louis, MO) for proteomic work. The nuclease-treated filtrate was then poured into 460ml centrifuge bottles, saturated with PEG 6000 at 10% w/v and stored overnight at 4°C. Following low-temperature incubation, the filtrates were centrifuged for 10min at 9500 x g at 4°C. The PEG supernatant was vacuum aspirated away with a sterile Pasteur pipette and the viral pellet resuspended in TE buffer. The proteomics protocol proceeds directly to CsCl density gradient ultracentrifugation whereas the genomics protocol subjects the resuspended viral pellet to a chloroform separation step using a Phase-lock™ gel tube to fractionate away remaining contaminant PEG.

For the proteomics work only acid washed centrifuge bottles were used, in order to prevent contaminants from being introduced at any step. The genomic isolation protocol only calls for acid washed bottles in the PEG precipitation phase.

Standard genomic isolation protocol: Initially a standard genomics-based approach was used to attempt characterization of the N phage genome, following a protocol adapted from Green and Sambrook's Molecular Cloning (2012). The step-by-step protocols are included in Appendix #2. In brief, formamide is used to disrupt the viral capsids, and thereafter a series of proteinase and chloroform treatments remove impurities from the sample as the DNA is precipitated and concentrated with ethanol.

Due to the small size of the viral particles and low recovery of nucleic acids, the first five N phage inductions sought to generate TempliPhi product from the extracted DNA. TempliPhi (GE Lifesciences, Pittsburgh, PA) uses a strand-displacing DNA polymerase from the *Bacillus*-infecting Φ29 Podophage. Because of the strand displacing ability of the polymerase, the resulting exponential amplification of DNA occurs as large, branching, full-genome concatemers; a phenomenon known as rolling circle amplification (RCA). After initial failures, a TempliPhi product was generated that produced different gel electrophoresis banding patterns from the positive and negative controls.

The random hexamer driven Whole Genome Amplification™ kit (WGA; Qiagen) was used generate small fragments from the TempliPhi product that could be readily cloned and sequenced. The WGA products were viewed by gel electrophoresis and fragments in the 400-1000 nucleotide size range were excised. Following Zymo Gel Cleanup (Zymo, Irvine, CA), ligation into a TOPO TA vector (Life Technologies, Grand Island, NY), and transformation into competent DH5-α *E. coli* (Life Technologies, Grand Island, NY), 6 clones with inserts around 500 bp were sent for bidirectional sequencing with M13. One subsequent WGA was performed, this time on TempliPhi product generated from DNA extracted using the commercial Zymo Viral Extraction Kit following the PEG concentration step. Ten sequenced clones were generated from this variation on the protocol.

Plasmid prep approach: On prior evidence that a strong plasmid band could be harvested from the culture, and on suspicion that the N prophage may exist as an extrachromosomal circular element, the Qiagen Plasmid Midi Kit (Qiagen, Valencia, CA) was used to target the plasmid (or possible viral replicative form) from log-phase GM9914 culture as well as from lightly induced flasks. Bands in the ~6 kb range were repeatedly obtained when the plasmid prep eluate was visualized on an agarose gel. When this band was excised using both Zymo (Irvine, CA) and MoBio (Carlsbad, CA) gel extraction kits in order to amplify the plasmid using GenomiPhi (GE Lifesciences, Pittsburgh, PA) followed by cloning. However, the GenomiPhi reaction failed. After ruling out failure in the GenomiPhi kit it was suspected that the plasmid was failing to extract from the gel.

Another extracted plasmid fraction was split and tested for restriction digestion with four standard enzymes (BamHI, HindIII, PSTI, XmnI). It was hypothesized that a linearized product might be more readily extractable from the gel and subsequently easier to clone. Enzymes BamHI and PSTI displayed evidence of digestion (Figure 3.2). The digested bands were excised and gel extracted, blunted, tailed and ligated into the TOPO XL high-capacity vector (Invitrogen, Grand Island, NY). Clones were picked and screened using the Herculase long-range Taq polymerase (Agilent, Santa Clara, CA). TOPO XL-transformed clones having ~6 kb inserts in the same size neighborhood of the undigested plasmid were sent to the Operon DNA Sequencing facility (Louisville, KY) for bidirectional sequencing with M13.

Subsequently, we used aforementioned evidence that the plasmid band was digested by BamHI and PSTI to facilitate preferential uptake of the linearized plasmid into pGem vectors that had been digested with the same restriction enzymes to create perfectly complementary overhangs. After transfection, cloning and picking ~50 clones, seven clones with inserts in the 6 kb size range were selected for bidirectional sequencing by Operon.

Preparing a metagenome of the whole N strain Synechococcus culture: In December 2012 we prepared a metagenome for a demonstration of the 454 Jr. Sequencing Platform by Roche Technologies (Branford, CT). Owing to prior difficulty with directly sequencing the viral particles, the decision was made to produce the metagenome from 500 ml of *uninduced* culture growing at logarithmic phase. Although the copy numbers of the viral genome would be low compared to an induced culture, it was believed that this would ensure that the virus sequence was captured and present in the final metagenome in some, perhaps rare, form.

Initial attempts to isolate total DNA in the one microgram quantities demanded by Roche resulted in a swamping of the Qiagen Blood and Tissue Kit. There was too much protein in even small culture pellets and Nanodrop of the final eluate appeared to be heavily contaminated by non-DNA constituents. Following the initial failures to obtain pure DNA in sufficient quantities a new protocol was adapted with help from Short Protocols in Molecular Biology in order to avoid swamping the DNeasy spin column. The new protocol (Appendix #3) involved pelleting, liquid nitrogen freeze-thaw, and chloroform centrifugation to eliminate the abundant protein. After eluting from the DNeasy column with 100 µl of TE, final Nanodrop concentration was 747 ng/µl with a 260 nm/280 nm ratio of 2.03.

Primer design: To follow up on genomics results, such as gap closing of the N strain contigs and evaluating bioinformatics hits (see Results and Data Analysis), primers were frequently required. The data DVD of the metagenome prepared by Roche contains a key Excel file called '454ReadStatus.xls' that contains information on the Newbler assembly, principally data about which reads went into a given contig. When this spreadsheet is sorted by column 'C' which lists the 5' contig, it is possible to extract the list of reads that went into any given contig of interest, all of which are prefixed "HXXO4HP01xxxxx". The sum of the reads obtained are contained in a large fasta file named '1.454Reads.fasta'. Using the web-based Fasta Sequence Extractor from the Fabox applications suite (Villesen 2007), the list of underlying reads taken from the sorted Excel spreadsheet were extracted from the '1.454Reads.fasta' file. The extracted reads were then realigned in Geneious (Biomatters, Auckland, NZ) and regions free of polymorphism with greater than 5x read-coverage were chosen for primer design.

Proteomics methodology: As detailed below in 'Results and Data Analysis' the metagenome did not produce a 'smoking gun' identity for the N phage. It could, however, be used as a search database if the viral proteins were *de novo* sequenced through mass spectrometry. The fundamental challenge of viral proteomics, like viral genomics, is a paucity of starting material to work with. Thus a protocol was adapted that allowed for the input of up to 10 flasks of induced culture.

Briefly, the culture flasks were pooled, cellular detritus was pelleted, supernatant was 0.02 µm -filtered and nuclease-treated and then PEG6000 was added for overnight incubation, as per 'Methods: *Viral Purification and Concentration'* and the Appendix #1 protocol. The resuspended viral pellets were the loaded onto a Cesium Chloride (CsCl) step gradient (2 ml x 1.3 g/ml; 2 ml x 1.5 g/ml; 2 ml x 1.7 g/ml) and ultracentrifuged (29,000 rpm using Sw40Ti rotor for 3 hours at 20°C). The bottom of the ultracentrifuge tube was punctured and purified viral fractions (free from host contaminant protein) were collected sequentially collected in half-milliliter fractions. After determining which fractions had the highest viral titer, fractions were pooled and dialyzed and then further concentrated through the use of an Amicon-15 centrifugal filter unit (Millipore, Billerica, MA). The particles were then denatured, reduced with DTT to eliminate disulfide bridges and run on a NuPAGE Bis-Tris protein gel (Life Technologies, Grand Island, NY). The step-by-step proteomics protocol is attached as Appendix #4 and is meant to commence using material harvested at the conclusion of the Appendix #1 protocol.

Results and Data Analysis:

Sequencing returns from standard genomic isolation > TempliPhi > WGA: Sequencing of the ~500 bp inserts revealed contaminants including *E. coli*, vector sequence and one Flavobacterial symbiont sequence. When WGA was applied to TempliPhi generated from Zymo kit extracted-DNA, eight of ten clones sequenced were *E. coli* related contaminant. The remaining two of the sequences had strong hits to *Synechococcus*

WH7803 hypothetical proteins, which warrant further investigation[1]. A similar protocol was used again in September however this returned only *E. coli* contaminant.

Sequencing returns from restriction digestion of plasmid fraction: When the plasmid fraction harvested from the N strain culture was restriction digested and ligated into TOPO XL, transformed clones having ~6 kb inserts in the same size neighborhood of the undigested plasmid were sent to the Operon DNA Sequencing facility (Louisville, KY) for bidirectional sequencing with M13. These large inserts turned out to be *Rhodobacter* sequence, presumed at the time to be contaminant inserts from the *Synechococcus* commensals present in the non-axenic GM9914 culture.

When restriction enzyme-specific pGem vectors were used to capture the digested plasmid fraction, seven inserts were sent to Operon and bi-directionally-sequenced. This revealed that three of the seven were from the N strain, hitting to WH7803 chromosomal DNA in both the forward and the reverse sequencing direction, and thus considered contaminant in this plasmid search. The remaining four inserts were *Rhodobacter*. Primers designed to test an HMM hit to large_ctg_00174 produced a correctly sized product when applied to several plasmid-uptake clones. After two rounds of bidirectional primer walking the effort to sequence the plasmid from this foothold was abandoned as the plasmid (perhaps not the only one) continued to be related to *Roseobacter*.

Testing the Gokushovirus primers on the N phage: Since the N phage particles are small and icosahedral and implicated as containing ssDNA, it was considered that it might be a gokushovirus. The *Gokushovirinae* PCR primers used in Chapters 1 and 2 of this thesis

[1] A BlastN database comparison of these two WGA contigs and those metagenomic large_contigs with promising Pfam hits yields a perfect match to the hit region of large_contig_00009 to HMM model PF07352: Phage_Mu_Gam

(Hopkins et al 2014) and several other gokushovirus primer pairs conceived by Bhakti Dwivedi, were applied to both the N phage viral concentrate (ruptured at 95°C) as well as to total DNA extracted from the N strain culture. No bands of the expected size were produced by any permutation of template-preparation and primer pair.

Metadata on the Metagenome: The 454 Jr. pyrosequencer produced a metagenome of excellent read quality and length. The gross metrics for the metagenome are displayed in Table 3.1. The reads were assembled into contigs by the proprietary Roche Newbler assembly program. Of the 3,331 total contigs, 3,188 of these were 500 bases or greater in length, which were defined as 'large contigs' (see Table 3.2). These large contigs (n=3,188; all_large_contigs.fasta) became the main target of our search efforts.

Scaffolding and gap closing of all_large_contigs.fasta vs. WH7803: Dr. McDaniel had previously reported a high degree of homology between the RbcL signature genes of the N strain of *Synechococcus* ('GM9914') and WH7803, the type strain of pink *Synechococcus*. It was possible that this similarity could be exploited to identify the N phage lysogenic insertion site, since WH7803 had shown no evidence of harboring a temperate phage under Mitomycin C induction.

Scaffolding of the 'all_large_contigs.fasta' onto WH7803, resulted in the recruitment of 27 contigs leaving 15 gaps unfilled in the WH7803 genome ranging in size from 262 bp – 6353 bp. It was hypothesized that variable prophage insertion could cause a breakdown in the ability of the Newbler assembler to stitch reads together through that area. The decision was made to target these areas with a gap-closing protocol. Since WH7803 is not believed to carry prophage, a discrepancy in the length of the gap closure PCR product versus the predicted length of WH7803 through the gap might have indicated the presence of a phage contained within the N strain but not WH7803. The gaps in the N strain *Synechococcus* were named 1-15 based on their scaffolded position within the WH7803 genome. For every gap-flanking contig, the underlying reads that went into the contig were aligned and visualized to pick regions with unanimous consensus sequence to bind the gap-closing primers.

The DNeasy kit was used to extract concentrated total DNA from both the N strain culture (GM9914) and an in-house WH7803 culture and the gap closing primers were identically applied to these paired total DNA samples. This allowed the product size to be conveniently compared side-by-side. As shown below in Figure 3.3, there were no meaningful (multiple kilobase) discrepancies detected between size of the gap-closing product for WH7803 and the N strain, which could have been attributable to the integrated N prophage.

Initial Blast parsing of all_large_contigs.fasta: Web-based automated annotation pipelines such as VIROME and MG-RAST, predicated on Blast results, were of little assistance in detecting the identity of this prophage within the metagenome. Although a significant number of mobile element genes were detected, a BlastX query of the Genbank non-redundant database ('nr') returned only one explicitly phage-related hit to a phage integrase observed in Large Contig 00068.

The Blast scan did provide a taxonomic look at the makeup of the non-axenic culture. The metagenome was dominated by *Synechococcus* but many Rhodobacter symbionts including *Marinobacter* spp*, Phaeobacter gallaciensis, Oceanicolis* spp*,* were observed as well as *Flavobacter* spp*.*

DataMining with the PhageFinder: deep evolutionary searching of all_large_contigs.fasta: In order to streamline the analysis of the N strain metagenome, we identified the Phage_Finder v2.1 automated prophage scanner [(Fouts 2006): accessible at http://www.mybiosoftware.com/sequence-analysis/818/] as an efficient means of parsing the large contigs. Phage_Finder runs as a 'master' shell program, compiling data returns from 'slave' programs including BlastP, hmmer3 [(Eddy 2009): accessible at http://hmmer.janelia.org/], and a tRNA detector. Repeat hits by slave programs in a single genomic neighborhood build statistical support for the presence of a prophage in that region.

One of the innovative features of Phage_Finder, is the presence of 'built-in' libraries of phage information. The author has compiled a comprehensive database of viral protein sequence (.faa files for BlastP) and viral HMM profiles (.hmm files for hmmsearch) which are updated for each new release. Since phage-related hits are the only objective, queries can simply be compared against these local databases rather than needing to be remotely queried against the totality of the NCBI and Pfam databases at prohibitive computational expense. The reduction in computational expense makes it possible to batch query >3000 individual contigs versus sending each one through a Blast query of the 'nr' database and then an HMMsearch of the Pfam database. The value of being able to do a batch HMMsearch of phage-only profiles against a 6-frame translation of all_large_contigs.fasta for a deep evolutionary search of viral homology within the metagenome was quickly apparent.

The first HMM profile database that was searched was the one built-in to Phage_Finder v2.1; the profiles are mostly derived from the Pfam system but also from the very comparable JCVI TIGR system. Version2.1 is a large and varied database including a full range of structural phage proteins, both surface and interior presenting, as well as enzymes such as integrases and lysozymes. It was assembled through natural language queries containing “phage” and phage-related keywords and thus the *Caudovirales* are overrepresented. Perhaps also as a result, some of the included Pfams are not exclusive to phage, or if they are, the phage proteins are close enough to bacterial proteins to trigger hits to obviously bacterial elements. This was particularly true of enzymatic Pfams, such as integrases, transposases and recombinases that appear to be somewhat interchangeable between phage and bacteria and are found throughout the genomes of the various Rhodobacter symbionts, generating many false positives.

In an effort to further co-opt Dr. Fouts’ search strategy, further specialized databases were generated that contained HMMs of capsid proteins of small icosahedral viruses (infecting all kingdoms) with low triangulation numbers (‘T=’) and both ssDNA and dsDNA. A rich source of icosahedral capsid HMMs was the VIPERdb (Carrillo-Tripp et al 2009); which provided access to the models for obscure families such as *Birnaviridae, Reoviridae, Carmoviridae* in addition to the more obvious candidates like *Microviridae, Parvoviridae, Circoviridae* which had already been tested in earlier search rounds.

There is one major downside to the “PhageFinder method” of metagenomic datamining. Since all Pfam hits are being generated from an ‘onboard’ HMM database then all of the hits will be necessarily phage-related and it allows the searcher to “see only what they want to see”. Phage_finder hits must be subsequently crosschecked against larger non-specific databases such as the GenBank ‘nr’. This was done for every PhageFinder hit region, with results shown in columns D and E of Table 3.3.

Discussion:

Discussion of what makes for an enticing hit: Cross-checking HMM hits with BlastX created four different cases of phage HMM hits. In an approximate rank order of the prevalence these can be summed up as:

- Case 1: A false positive in which a phage Pfam hit has a much stronger Blast hit to a well-characterized non-phage gene from a bacteria in the culture (most common)
- Case 2: An apparent agreement between the Pfam and Blast hits; often this raised the question of whether the Pfam is phage-specific or in fact hitting to a non-phage bacterial mobile element.
- Case 3: An enticing Pfam phage hit with an inconclusive Blast hit to an unfamiliar bacteria, esp. non-Rhodobacter
- Case 4: An enticing Pfam phage hit with a Blast hit to a hypothetical protein of a known member of the culture, esp. *Synechococcus* WH7803.

Cross-checking HMM hits with BlastX frequently revealed that potentially promising HMM-based phage hits most frequently produced Case#1 non-phage Blast hits to functionally characterized ORFs that were much stronger in terms of e-value. For instance, the finding that large_contig_01600 had an hmmsearch hit to the HMM profile of the Carmovirus coat protein (PF08462) at an e-value of 0.087 (see Table 3.3) is promising, since this genus of viruses has capsids are that ~40 nm diameter T=3 icosahedra, perhaps sharing some degree of structural homology with the N phage. However, this HMM hit is invalidated by the finding that the same region of large_contig_01600 hits to a Roseobacter arginyl tRNA-synthetase gene with an overwhelmingly strong e-value of zero.

If the N phage has been detected, then the hit likely falls in the Case #3 or #4 situation. Since available evidence points to the N phage as being highly novel, this allows for the possibility that the prophage has been sequenced and accessioned into the Genbank databases as a hypothetical. This possibility makes ‘contig00004: 29210-29935’, ‘contig00005: 59448-60896’, ‘contig00008: 109622-110827’, ‘contig00009: 23687-24460 and 29285-29851', 'contig00010: 67684-68592' as well as 'contig00019: 17210-17947' very enticing regions since these have overwhelmingly strong Blast hits to hypothetical proteins of *Synechococcus* WH7803, but also hit to phage-related Pfam HMMs. The full list of hits between every individual Pfam HMM query and all_large_contigs (Table 3.3) shows that in several instances there were two or more hits adjacent (or approximately so) within the given metagenomic contig, providing perhaps the strongest evidence of the presence of a prophage. It is important to be mindful that there are likely other prophages present in the culture which are not the N phage and that this may confound efforts to identify the N phage through bioinformatic means.

Discussion of the inadequacy of bioinformatics hits testing methodology: Generating the type of data shown in Table 3.3 raised the vital question of how best to test potential phage regions revealed by bioinformatics. Barring any better methods, a protocol was created to design primers specific to the phage-hit region and then differentially test the primers on an induced, purified viral fraction versus the total, uninduced culture and a negative control: three PCR reactions per primer set. The paired sample PCR strategy had worked well for the gap closing. The hope was that an overwhelmingly strong PCR product would result when the purified N phage viral concentrate was used as template and that this would confirm that the Pfam hit belonged to the N phage. Primers were designed in Geneious (Biomatters, Auckland, NZ) based on the underlying reads, obtained by cross-referencing '454ReadStatus.xls' with '1.454Reads.fasta' as for the gap closing (see Methods).

Upon refinement, a fourth PCR reaction was added to every primer pairs test. To obtain template for the total culture fraction (treatment #1), an aliquot was pulled from the particle prep prior to the DNase treatment to remove host and symbiont DNA. The yield of the completed (DNase digested) particle prep was split and in both fractions, the capsids were popped with 95°C heat treatment (that they would rupture at this temperature is a safe assumption, but an assumption nonetheless). One heated fraction was harvested as template (treatment #2), but the other was treated with Proteinase K (treatment #3) to ensure that any nucleoproteins that could be rendering the viral DNA inaccessible to PCR (a possible reason that the virus has been so recalcitrant to standard genomic isolations) would be removed. The fourth reaction was a negative. A typical result for this protocol is shown in Figure 3.4 for two primer pairs designed against the Pfam hit regions in ctg00909 and ctg02624.

If it is not known whether 'normal' PCR-available DNA is present; if there is no (+) control in order to know the efficacy of the PCR primers in targeting the mystery DNA; then a negative result cannot be properly used to rule out the sequence as belonging to the Nphage. Due to the underlying weakness of this testing methodology, there has been no systematic and rigorous winnowing of promising hits; this is an open line of inquiry that could benefit from a more clever and innovative screening method.

Potential impact of proteomics approach: As of the submission deadline for this thesis, the proteomics work is still ongoing. Using the methodology outlined above and in Indices #1 and #4, several distinct protein populations have been visualized on the NuPAGE Bis-Tris gel stained with Sypro Ruby (both products of Life Technologies, Grand Island, NY). An annotated image of this gel is shown in Figure 3.5.

The distinct bands from this gel will be excised, digested with trypsin and processed by High Performance Liquid Chromatography to further distinguish the molecular weight populations by isoelectric point, as they are fed into the mass spectrometer ion trap. Tryptic-digest length peptides will be sequenced and these will be reconstructed into partial or full-length proteins. These can then be blasted against a 6-frame translation of the all_large_contigs.fasta metagenomic file in hopes of finding the coding region. It is possible that two or more of the *de novo* sequenced protein bands will hit to the same contig of the metagenome. This would provide the 'smoking gun' evidence allowing identification of the N phage genome.

Concluding Remarks:

In concluding what has been an enriching but frequently frustrating and ultimately inconclusive investigation, it may be helpful to consider what is known versus what has been assumed about the N phage. What is known is that when the N strain culture of non-axenic *Synechococcus,* growing at logarithmic phase, is stressed, some component of the culture releases vast quantities of small nucleic acid encapsulations.

Comparatively more has been assumed. We assume that the encapsulations are phages, however they may be vesicles or GTAs. We assume that *Synechococcus* is by far the numerically abundant member of the culture and thus the only component capable of creating such a prolific phage pulse. Future work should test this assumption, perhaps through flow cytometry.

To date, I believe that the approach of trying to sequence the DNA contained in the N phage has been exhausted, by Dr. McDaniel earlier and in the efforts detailed in this thesis. One of four possibilities has occurred. Firstly, it is possible that the N phage genome is truly resistant to all available methods of extraction and sequencing. We know that the DNA is unstable, but it may be inextricably bound by nucleoprotein or rendered inaccessible by some other, as yet unrecognized, mechanism. Secondly, it is possible that the N phage genome has been sequenced but is so divergent that we are unable to detect it through homology searching of any method. Table 3.3 was an effort to address this and it is possible that somewhere on this lengthy list hits to the N phage are hiding in plain view. Thirdly, it is possible that the N 'phage' has no defined genome and is randomly packaging host or total culture DNA, as a vesicle or GTA would. This would explain the many sequencing returns hitting to Rhodobacter symbionts. Lastly, it is possible that the N phage genome is RNA-based.

Future work should focus on non-genomic methods for identifying the N phage. In this regard, the ongoing proteomics work is exciting. However, since a metagenome has never been performed on a particle preparation, this could be of value, especially for identifying vesicle-type packaging (Biller et al 2014). Finally, there is great need for a better method for testing the N phage particles for bioinformatic hits.

Acknowledgments:

MH is funded under the USF Presidential Fellowship. Thanks to Roche Diagnostics for performing a no-cost, high quality metagenome of the N strain culture using the 454 GS Junior sequencing platform. Much gratitude goes to Dr. Lauren McDaniel for her constant generosity, first in sharing the project and subsequently in sharing protocols and wisdom and always being willing to troubleshoot. Thanks also to Dr. John Paul for making his lab available. Thanks to Dr. Kent Seeley of the USF Center for Drug Discovery and Innovation (CDDI) Proteomics Core for proteomics assistance and to Dr. Tony Greco of the USFCMS

Electron Microscopy Lab for providing micrographs. Lastly, thanks to major professor Dr. Mya Breitbart for giving her students the intellectual and laboratory freedom to make the far ranging inquiries that are ultimately the most rewarding, but always providing a backstop of good-natured support.

References:

Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW (2014). Bacterial Vesicles in Marine Ecosystems. *Science* **343:** 183-186.

Carrillo-Tripp M, Shepherd CM, Borelli IA, Venkataraman S, Lander G, Natarajan P, Johnson JE, Brooks CL, Reddy VS (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Research* **37:** D436-D442.

Eddy, SR (2009). A new generation of homology search tools based on probabilistic inference. *Genome Information* **23:** 205-211.

Fouts DE (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research* **34:** 5839-5851.

Garrity GM, Bell JA, Lilburn TG (2004). *Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology.* Springer, New York.

Green MR, Sambrook J (2012). *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, New York.

Hopkins M, Kailasan S, Cohen A, Roux S, Tucker KP, Shevenell A, Agbandje-McKenna M, Breitbart M (2014). Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *The ISME Journal*. **doi:** 10.1039/ismej.2014.43.

Krupovic M, Forterre P (2011). *Microviridae* Goes Temperate: Microvirus-related proviruses reside in the genomes of bacteroidetes. *PLoS ONE* **6:** e19893.

McDaniel LD, delarosa M, Paul JH (2006). Temperate and lytic cyanophages from the Gulf of Mexico. *Journal of the Marine Biological Association of the United Kingdom* **86:** 517-527.

Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, Sarracino D, Chisholm SW (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. *Environmental Microbiology* **11:** 2935-2951.

Villesen P (2007). FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes* **7:** 965-968.

Waterbury JB, Willey JM (1988). Isolation and growth of marine planktonic cyanobacteria. Springer, New York.

Tables and Figures:

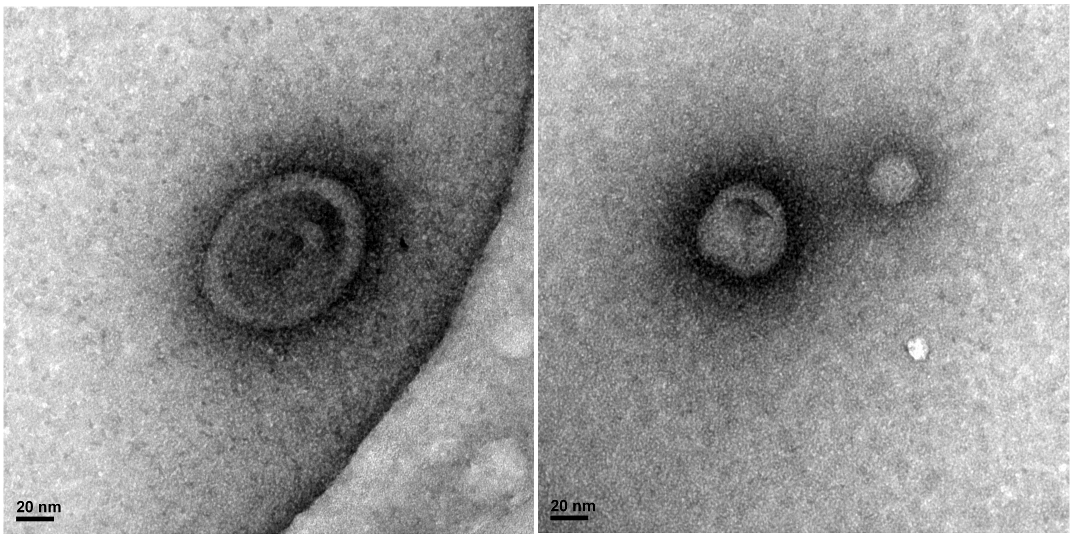


Figure 3.1: Electron micrographs of abundant particles believed to be an icosahedral bacteriophage produced by the N strain *Synechococcus.* Images produced by T. Greco of the USFCMS Electron Microscopy Lab.

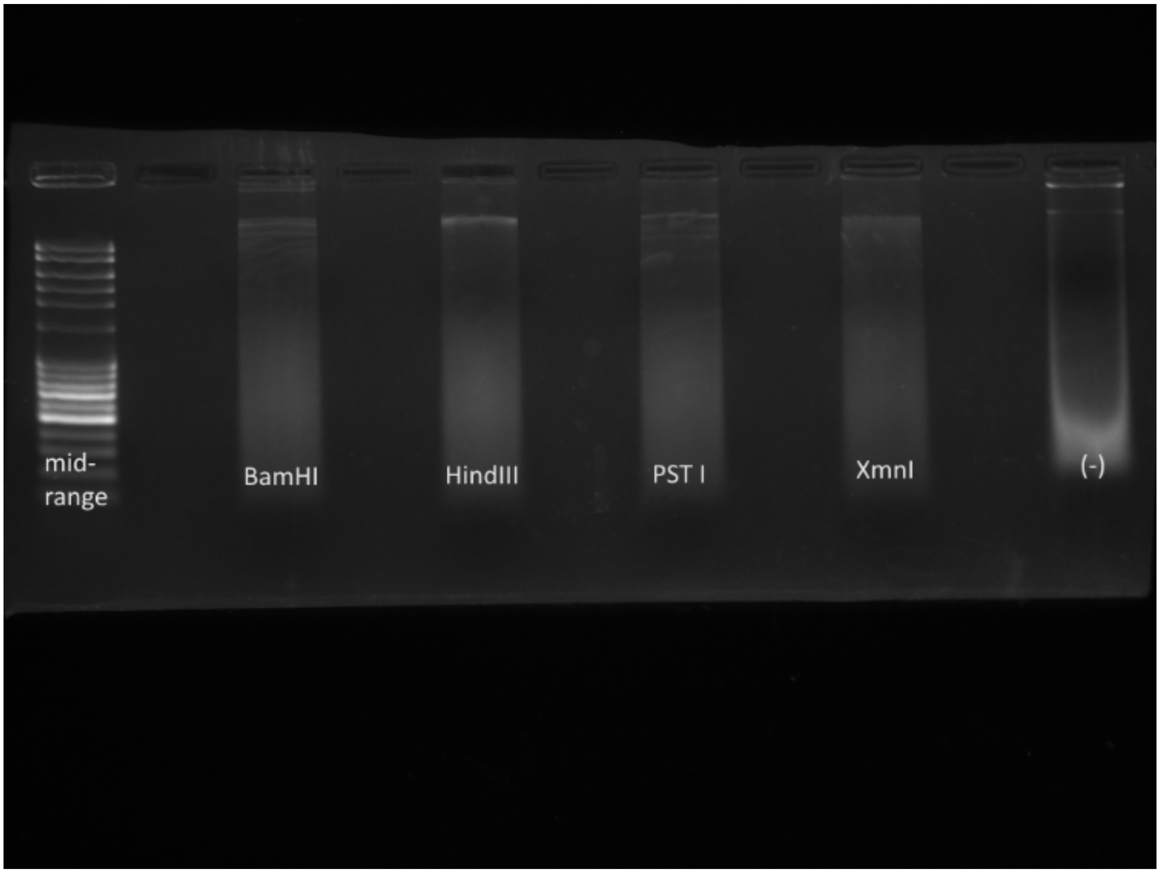


Figure 3.2: The results of plasmid digestion, by four restriction enzymes.

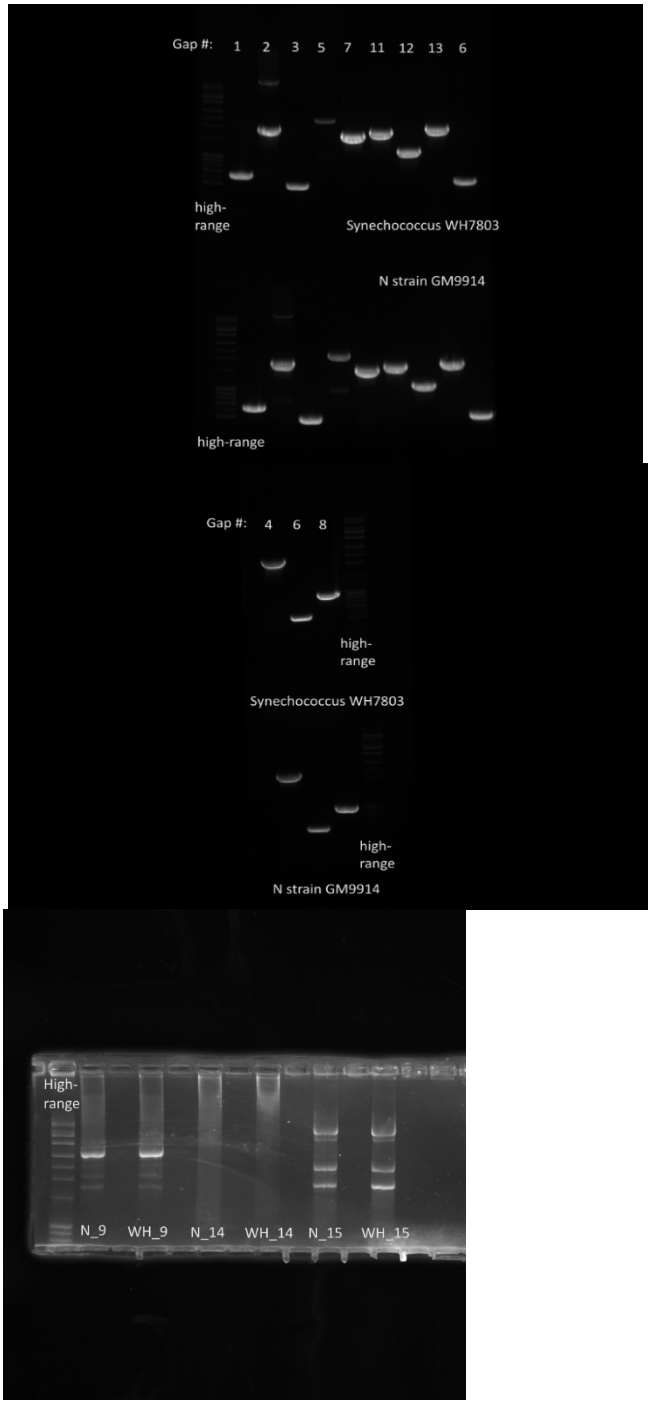


Figure 3.3: The results of the paired WH7803 vs N strain gap-closing experiment. In the upper and middle panels the two different *Synechococcus* strains are juxtaposed top and bottom, while in the bottom panel the strains are compared side-by-side. The three gaps in the bottom panel (gap #'s 9, 14 and 15) were large enough (3-6kb) that Herculase taq was used to generate the product, thus the separate treatment.

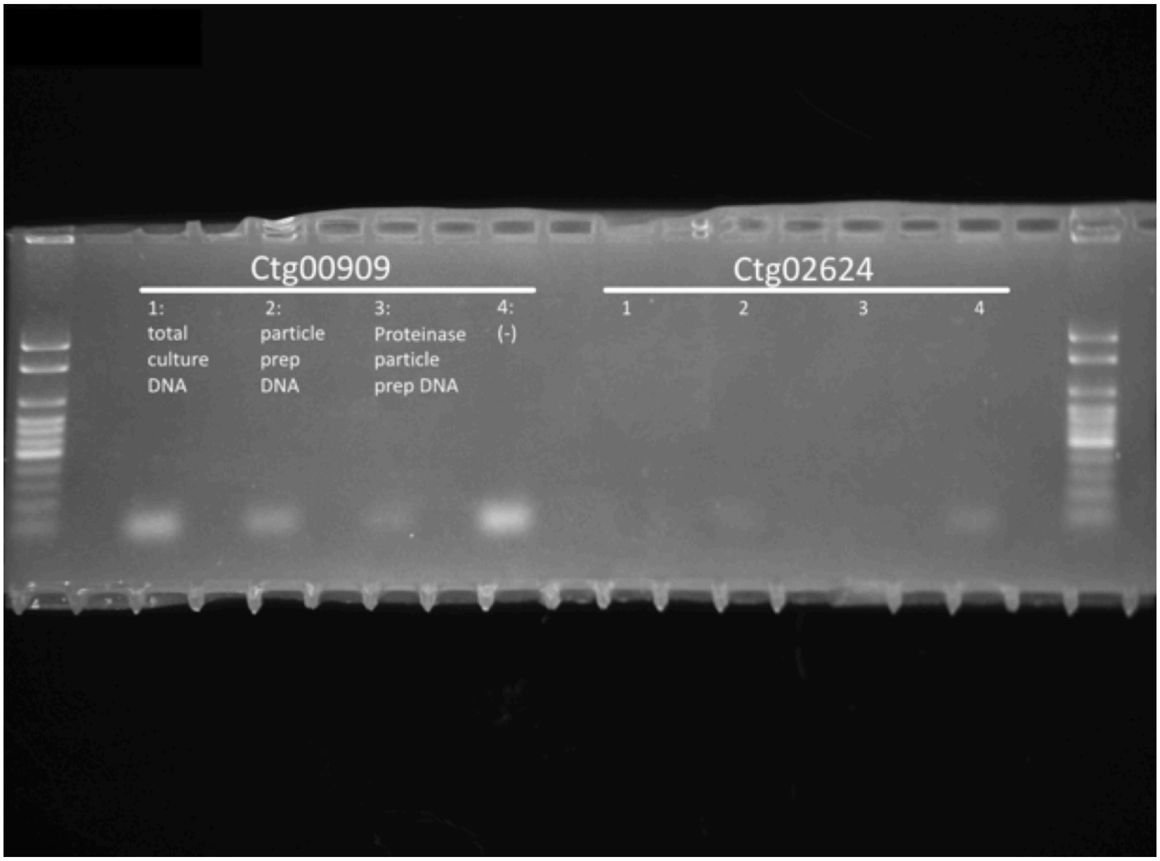


Figure 3.4 A typical and inconclusive result for the bioinformatic hits testing PCR primer-based protocol outlined in the Results and Data Analysis section.

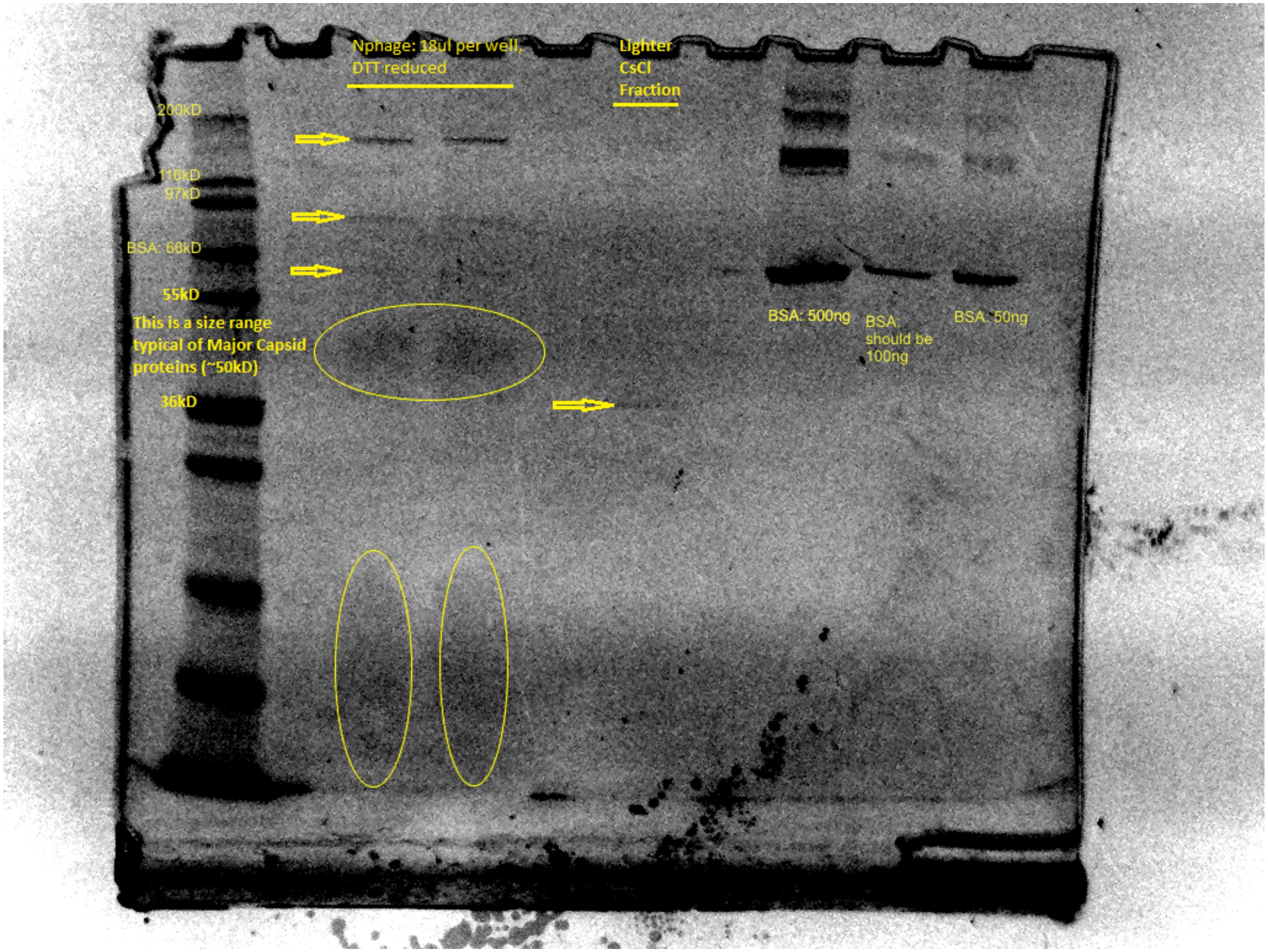


Figure 3.5 Annotated protein gel showing two adjacent, identical lanes of concentrated N phage protein, flanked on the left by a molecular weight ladder and on the left by three, semi-quantitative BSA standards.

Table 3.1: Gross metrics for the N strain metagenome

Sample	HQ Reads	HQ Bases	Avg Read Length	Mode Read Length
Synecho_N	187,218	81,809,059	437	486

Table 3.2: Assembly statistics for the N strain metagenome

All Contig		Large Contig				
Contigs	Bases	Contig	Bases	N50	%Q40+	Cov.
3,331	6.15 Mb	3,188	6.11 Mb	1.95 Kb	97.36%	9.41

Table 3.3: The results of cross-checking phage-related Pfam hits against the Genbank nr database.

hmmsearch Pfam hit region of all_large_ctg	Pfam hit	Pfam evalue	tophit of BlastX search of pFam hit region against Genbank nr	BlastX evalue
VJ4_contig00001_132622_131609_142	PF01464: SLT: transglycosylase SLT domain	5.80E-07	WP_011933308| lytic transglycosylase [Synechococcus WH7803]	0
VJ4_contig00001_273857_272949_291	PF01510: Amidase_2: N-acetylmuramoyl-L-alanine amidase	3.50E-16	WP_011933163| N-actylmuramoyl-L-alanine amidase [Synechococcus WH7803]	0
VJ4_contig00002_77782_76751_11	PF03389: MobA/MobL family	0.0073	WP_011935112| hypothetical protein [Synechococcus WH7803]	0
VJ4_contig00003_103934_103179_107	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	2.90E-08	WP_011932555| Phage integrase family protein [WH7803]	0
VJ4_contig00003_103934_103179_107	TIGR02225: recomb_XerD: tyrosine recombinase XerD	0.00098	WP_011932555| Phage integrase family protein [WH7803]	0
VJ4_contig00003_103934_103179_107	TIGR02249: integrase_gron: integron integrase	0.028	WP_011932555| Phage integrase family protein [WH7803]	0
VJ4_contig00003_169959_169459_187	TIGR01555: phge_rel_HI1409: phage-associated protein, HI1409 family	0.07	WP_011932626| hypothetical protein [WH7803]	6.00E-103
VJ4_contig00003_225810_225124_249	PF00959: Phage_lysozyme: phage lysozyme	1.10E-05	WP_011932688| muramidase endolysin [WH7803]	1.00E-148
VJ4_contig00003_225810_225124_249	PF06737: Transglycosylas: transglycosylase-like domain	0.0049	WP_011932688| muramidase endolysin [WH7803]	1.00E-148
VJ4_contig00003_79445_78981_82	PF00436: Single-strand binding protein family	2.80E-11	WP_011932529| Single-strand binding protein [WH7803]	1.00E-80
VJ4_contig00004_29935_29210_22	TIGR02216: phage conserved hypothetical protein	0.031	WP_011934302| hypothetical protein [WH7803]	4.00E-174
VJ4_contig00005_106524_105787_120	PF00239: Resolvase: resolvase, N terminal domain	0.032	WP_011933783| macrolide ABC transporter [WH7803]	4.00E-168
VJ4_contig00005_23395_23907_30	PF12289: Rotavirus VP1 structural protein	0.048	WP_011933873| cytochrome c-550 [WH7803]	2.00E-110
VJ4_contig00005_59448_60896_67	TIGR02242: phage tail protein domain	0.034	WP_011933835| hypothetical protein [WH7803]	0.00E+00
VJ4_contig00005_93347_92196_104	PF02368: Big_2: bacterial Ig-like domain, group 2	0.0026	WP_011933798| translocase component YidC [WH7803]	0.00E+00
VJ4_contig00006_107371_107925_116	PF03245: Phage_lysis: bacteriophage lysis protein	0.096	WP_011932399| C-phycoerythrin class I beta chain [WH7803]	1.00E-127
VJ4_contig00006_114300_113803_123	TIGR01673: holin_LLH: phage holin, LL-H family	0.076	YP_001224215.1| C-phycoerythrin class II alpha chain [Synech...	3.00E-102
VJ4_contig00007_106864_106481_111	PF05666: Fels1: Fels-1 prophage protein-like	0.044	YP_001226073.1| universal stress protein family protein [Syn...	3.00E-74
VJ4_contig00008_110827_109622_110	TIGR01554: major_cap_HK97: phage major capsid protein, HK97 family	0.77	YP_001224041.1| hypothetical protein SynWH7803_0318 [Synecho...	0
VJ4_contig00008_117749_115692_118	PF01464: SLT: transglycosylase SLT domain	1.20E-24	YP_001224049.1| lytic transglycosylase [Synechococcus sp. WH...	0
VJ4_contig00008_42497_41496_37	PF07902: Gp58: gp58-like protein	0.05	YP_001223973.1| Ycf48-like protein [Synechococcus sp. WH 780...	0
VJ4_contig00008_42966_42538_38	PF05876: Terminase_GpA: phage terminase large subunit GpA	0.065	YP_001223974.1| rubredoxin [Synechococcus sp. WH 7803] >ref|...	2.00E-66
VJ4_contig00009_24460_23687_28	PF07352: Phage_Mu_Gam: bacteriophage Mu Gam like protein	0.014	YP_001225839.1| hypothetical protein SynWH7803_2116 [Synecho...	2.00E-180
VJ4_contig00009_29851_29285_34	PF05929: Phage_GPO: phage capsid scaffolding protein (GPO) serine peptidase	0.0099	YP_001225845.1| hypothetical protein SynWH7803_2122 [Synecho...	1.00E-84
VJ4_contig00009_30244_33849_36	PF04582: Reovirus sigma C capsid protein	0.25	YP_001225847.1| chromosome segregation ATPase [Synechococcus...	0
VJ4_contig00009_8179_7190_10	PF01018: GTP1_OBG: GTP1/OBG	2.20E-59	YP_001225818.1| GTPase ObgE [Synechococcus sp. WH 7803] >ref...	0
VJ4_contig00010_67684_68592_57	PF01510: Amidase_2: N-acetylmuramoyl-L-alanine amidase	1.50E-08	YP_001223822.1| hypothetical protein SynWH7803_0099 [Synecho...	0
VJ4_contig00010_8946_9689_11	PF03245: Phage_lysis: bacteriophage lysis protein	0.0094	YP_001223865.1| Serine acetyltransferase [Synechococcus sp. ...	3.00E-158
VJ4_contig00012_12148_11378_10	PF06805: Lambda_tail_I: bacteriophage lambda tail assembly protein I	0.015	YP_001225795.1| amino acid ABC transporter periplasmic prote...	0

Table 3.3 (Continued)

VJ4_contig00014_48213_47083_51	TIGR02642: phage_xxxx: uncharacterized phage protein	4.7	YP_001223746.1| chaperone protein DnaJ [Synechococcus sp. WH...	0
VJ4_contig00016_42845_43243_45	PF00436: Single-strand binding protein family	2.30E-28	YP_001223894.1| single-stranded DNA-binding protein [Synecho...	1.00E-71
VJ4_contig00019_17947_17210_21	PF05666: Fels1: Fels-1 prophage protein-like	5.40E-18	YP_001225376.1| hypothetical protein SynWH7803_1653 [Synecho...	3.00E-156
VJ4_contig00021_2698_4101_3	PF03374: ANT: phage antirepressor protein KilAC domain	0.0027	WP_009807551.1| putative RepA protein [Roseobacter sp. MED19...	3.00E-126
VJ4_contig00031_192_1_1	PF00239: Resolvase: resolvase, N terminal domain	1.50E-09	YP_001531746.1| recombinase [Dinoroseobacter shibae DFL 12 =...	1.00E-27
VJ4_contig00032_3043_3699_7	PF04404: ERF: Erf superfamily	2.10E-30	WP_002178214.1| Erf family protein [Leptospira interrogans] ...	6.00E-17
VJ4_contig00053_3700_2762_5	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	0.00012	YP_001280023.1| hypothetical protein PsycPRwf_1124 [Psychrob...	6.00E-23
VJ4_contig00064_2752_2522_4	PF05930: Phage_AlpA: transcriptional regulator, AlpA family	0.048	WP_009570797.1| hypothetical protein [Celeribacter baekdonen...	1.00E-24
VJ4_contig00068_168_1439_1	TIGR02224: recomb_XerC: tyrosine recombinase XerC	4.20E-10	YP_005937537.1| phage integrase family site specific recombi...	3.00E-15
VJ4_contig00068_168_1439_1	TIGR02225: recomb_XerD: tyrosine recombinase XerD	4.20E-06	YP_005937537.1| phage integrase family site specific recombi...	3.00E-15
VJ4_contig00068_168_1439_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	2.00E-13	YP_005937537.1| phage integrase family site specific recombi...	3.00E-15
VJ4_contig00068_2431_3048_3	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	6.80E-05	WP_009159429.1| site-specific recombinase, phage integrase f...	6.00E-19
VJ4_contig00078_2365_2992_5	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	1.60E-08	WP_008204046.1| tyrosine recombinase XerC [Roseobacter sp. S...	6.00E-85
VJ4_contig00078_2365_2992_5	TIGR02224: recomb_XerC: tyrosine recombinase XerC	1.90E-49	WP_008204046.1| tyrosine recombinase XerC [Roseobacter sp. S...	6.00E-85
VJ4_contig00078_2365_2992_5	TIGR02225: recomb_XerD: tyrosine recombinase XerD	2.80E-40	WP_008204046.1| tyrosine recombinase XerC [Roseobacter sp. S...	6.00E-85
VJ4_contig00078_2365_2992_5	TIGR02249: integrase_gron: integron integrase	8.70E-10	WP_008204046.1| tyrosine recombinase XerC [Roseobacter sp. S...	6.00E-85
VJ4_contig00078_2365_2992_5			WP_008204046.1| tyrosine recombinase XerC [Roseobacter sp. S...	6.00E-85
VJ4_contig00087_2179_986_3	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	2.80E-15	WP_007811908.1| integrase [Roseobacter sp. AzwK-3b] >gb|EDM7...	0.00E+00
VJ4_contig00087_227_48_1	PF05930: Phage_AlpA: transcriptional regulator, AlpA family	0.0017	WP_022573273.1| putative transcriptional regulator [Rhodobac...	9.00E-13
VJ4_contig00107_1783_2253_4	PF07484: Collar: phage tail collar domain	4.40E-14	WP_008638658.1| phage Tail Collar domain protein [Bizionia a...	5.00E-36
VJ4_contig00107_2421_2713_5	PF07484: Collar: phage tail collar domain	2.30E-22	WP_003047861.1| Tail collar domain-containing protein [Sphin...	8.00E-24
VJ4_contig00118_224_1369_1	PF00665: rve: integrase core domain	2.50E-17	WP_023666458.1| transposase [Rhodobacter sp. CACIA14H1] >gb|...	0
VJ4_contig00121_1751_2560_4	PF00665: rve: integrase core domain	2.20E-33	WP_009159952.1| transposase [Thalassobium sp. R2A62] >gb|EET...	4.00E-178
VJ4_contig00121_485_1183_2	PF00665: rve: integrase core domain	3.60E-23	WP_008553311.1| integrase [Rhodobacterales bacterium Y4I] >g...	1.00E-158
VJ4_contig00135_1_732	PF05518: Totivirus coat protein	4.40E+00	WP_013270462| cytochrome C [Rhodobacter]	2.00E-42
VJ4_contig00140_2395_908_2	PF00665: rve: integrase core domain	4.10E-11	WP_021120852.1| Mobile element protein [Salipiger mucosus] >...	0
VJ4_contig00143_1233_847_3	PF06199: Phage_tail_2: phage major tail protein 2	1.30E-05	WP_008333594.1| hypothetical protein [Maritimibacter alkalip... 129	2.00E-35
			WP_018047695.1| hypothetical protein [Nitrospina sp. AB-629-... 46.6	1.00E-04
			gb|ADD94567.1| hypothetical protein [uncultured phage MedDCM-OCT...	0.002

Table 3.3 (Continued)

VJ4_contig00143_364_1_1	TIGR01760: tape_meas_TP901: phage tail tape measure protein, TP901 family, core region	1.50E-09	WP_017726771.1| tail protein, partial [Bacillus sp. L1(2012)]	3.00E-11
VJ4_contig00165_434_1_1	PF07471:Phage_Nu1: phage DNA packaging protein Nu1	0.02	YP_004303750.1| Isopentenyl-diphosphate delta-isomerase, typ... 159	2.00E-46
VJ4_contig00174_2075_789_2	PF03389: MobA/MobL family	1.30E-42	YP_771879.1| mobilization protein [Roseobacter denitrificans...	0
VJ4_contig00193_1337_2186_3	PF03374: ANT: phage antirepressor protein KilAC domain	0.019	WP_009815136.1| transposase [Roseovarius nubinhibens] >gb|EA...	3.00E-180
VJ4_contig00193_1337_2186_3	PF00665: rve: integrase core domain	7.40E-10	WP_009815136.1| transposase [Roseovarius nubinhibens] >gb|EA...	3.00E-180
VJ4_contig00193_236_742_1	PF06056: Terminase_5: putative terminase, ATPase subunit, gpP-like	0.0011	YP_682883.1| transposase [Roseobacter denitrificans OCh 114]...	4.00E-68
VJ4_contig00193_739_1227_2	PF00665: rve: integrase core domain	1.40E-13	YP_682883.1| transposase [Roseobacter denitrificans OCh 114]...	2.00E-112
VJ4_contig00197_51_1742_1	PF05565: Sipho_Gp157: siphovirus Gp157	7.2	WP_008555366.1| methyl-accepting chemotaxis sensory transduc...	0
VJ4_contig00199_2162_1515_3	PF09299: Mu-transpos_C: Mu transposase, C-terminal	1.00E-13	YP_001170391.1| plasmid replication initiator protein-like p...	8.00E-42
VJ4_contig00268_1388_426_2	TIGR02224: recomb_XerC: tyrosine recombinase XerC	6.70E-12	WP_009809258.1| integrase [Roseobacter sp. MED193] >gb|EAQ45...	2.00E-149
VJ4_contig00268_1388_426_2	TIGR02225: recomb_XerD: tyrosine recombinase XerD	1.10E-10	WP_009809258.1| integrase [Roseobacter sp. MED193] >gb|EAQ45...	2.00E-149
VJ4_contig00268_1388_426_2	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	3.80E-21	WP_009809258.1| integrase [Roseobacter sp. MED193] >gb|EAQ45...	2.00E-149
VJ4_contig00271_1490_1068_3	PF07022: Phage_CI_repr: bacteriophage CI repressor protein	0.0017	NB: hit to wrong contig portion	
VJ4_contig00276_1971_802_2	PF01464: SLT: transglycosylase SLT domain	0.0067	YP_004693239.1| lytic murein transglycosylase [Roseobacter l...	0.00E+00
VJ4_contig00283_1_822_1	PF03432: Relaxase/Mobilisation nuclease domain	0.16	YP_004262422.1| Tex-like protein [Cellulophaga lytica DSM 74...	0.00E+00
VJ4_contig00299_1264_1940_2	PF07455: Psu: phage polarity suppression protein	0.013	YP_002520451.1| Methionine synthase [Rhodobacter sphaeroides...	2.00E-139
VJ4_contig00302_1187_1456_2	PF05973: Gp49: Gp49-like PF05973 family	5.20E-18	WP_018912738.1: hypothetical protein [Thiomonas sp. FB-6]	6.00E-31
VJ4_contig00334_257_1_1	PF06763: Minor_tail_Z: prophage minor tail protein Z	0.0084	WP_005612405.1| hypothetical protein [Ruegeria mobilis] >gb|...	2.00E-20
VJ4_contig00346_1210_305_2	PF00665: rve: integrase core domain	2.60E-22	WP_008281775.1| transposase [Roseovarius sp. TM1035] >gb|EDM...	2.00E-171
VJ4_contig00376_1_1093_1	TIGR02224: recomb_XerC: tyrosine recombinase XerC	0.00013	YP_008975731.1| Integrase [Phaeobacter gallaeciensis DSM 266...	0.00E+00
VJ4_contig00376_1_1093_1	TIGR02225: recomb_XerD: tyrosine recombinase XerD	0.0017	YP_008975731.1| Integrase [Phaeobacter gallaeciensis DSM 266...	0.00E+00
VJ4_contig00376_1_1093_1	TIGR02249: integrase_gron: integron integrase	0.00035	YP_008975731.1| Integrase [Phaeobacter gallaeciensis DSM 266...	0.00E+00
VJ4_contig00376_1_1093_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	3.30E-13	YP_008975731.1| Integrase [Phaeobacter gallaeciensis DSM 266...	0.00E+00
VJ4_contig00381_691_1802_2	PF06950: DUF1293: PF06950 family	0.048	WP_010441423.1| peptide ABC transporter substrate-binding pr...	0.00E+00
VJ4_contig00397_993_1331_4	TIGR01673: holin_LLH: phage holin, LL-H family	0.0096	WP_008207263.1| nitrogen regulatory protein P-II 1 [Roseobac...	4.00E-65
VJ4_contig00418_756_1751_2	PF01018: GTP1_OBG: GTP1/OBG	4.80E-66	WP_022700616.1| GTPase CgtA [Oceanicaulis alexandrii]	2.00E-155
VJ4_contig00419_1753_1271_3	TIGR02419: C4_traR_proteo: phage/conjugal plasmid C-4 type zinc finger protein, TraR family	0.00046	WP_008207216.1| molecular chaperone DnaK [Roseobacter sp. SK...	8.00E-86
VJ4_contig00438_1728_1049_2	TIGR02224: recomb_XerC: tyrosine recombinase XerC	1.30E-61	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	3.00E-104
VJ4_contig00438_1728_1049_2	TIGR02225: recomb_XerD: tyrosine recombinase XerD	3.90E-74	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	3.00E-104

Table 3.3 (Continued)

VJ4_contig00438_1728_1049_2	TIGR02249: integrase_gron: integron integrase	5.40E-32	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	3.00E-104
VJ4_contig00438_1728_1049_2	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	8.60E-39	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	3.00E-104
VJ4_contig00516_568_8_1	TIGR01554: major_cap_HK97: phage major capsid protein, HK97 family	4.8	WP_008555198.1| molecular chaperone GrpE [Rhodobacterales ba...	7.00E-81
VJ4_contig00526_1_1174_1	PF04466: Terminase_3: phage terminase large subunit	2.20E-45	WP_022697912.1| terminase [Maricaulis sp. JL2009]	0.00E+00
VJ4_contig00526_1_1174_1	TIGR01547: phage_term_2: phage terminase, large subunit, PBSX family	1.10E-23	WP_022697912.1| terminase [Maricaulis sp. JL2009]	0.00E+00
VJ4_contig00526_1174_1623_2	TIGR01555: phge_rel_HI1409: phage-associated protein, HI1409 family (further search suggests CRISPR-related)	5.70E-28	ETV63474.1| HI1409 family phage-associated protein [Pseudomon...	2.00E-45
VJ4_contig00547_1596_895_3	PF04582: Reovirus sigma C capsid protein	2.00E+00	YP_008974865.1| Methyl-accepting chemotaxis protein [Phaeoba...	6.00E-70
VJ4_contig00598_462_965_3	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	2.50E-07	WP_023659524.1| Site-specific recombinase XerD [Congregibact...	5.00E-145
VJ4_contig00601_914_303_2	TIGR02219: phage_NlpC_fam: putative phage cell wall peptidase, NlpC/P60 family	0.0047	WP_008205681.1| nucleoside-triphosphate diphosphatase [Roseo...	5.00E-122
VJ4_contig00608_1154_1537_3	PF04582: Reovirus sigma C capsid protein	3.40E-02	WP_008270656.1| hypothetical protein [Flavobacteriales bacte...	2.00E-76
VJ4_contig00625_1_391_1	PF03374: ANT: phage antirepressor protein KilAC domain	0.017	YP_007547109.1| Transcriptional regulator, AraC [Bibersteini...	1.00E-12
VJ4_contig00632_1515_1080_4	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	5.90E-15	WP_009571413.1| integrase [Celeribacter baekdonensis] >gb|EK...	4.00E-60
VJ4_contig00660_273_863_2	PF08765: mor transcription activator family	0.0036	YP_008975082.1| Bacterial mobilization protein (MobC) [Phaeo...	2.00E-128
VJ4_contig00660_850_1491_3	PF03432: Relaxase/Mobilisation nuclease domain	2.70E-14	YP_008975083.1| Type IV secretory pathway, VirD2 component (...	6.00E-149
VJ4_contig00660_850_1491_3	PF03389: MobA/MobL family	0.00057	YP_008975083.1| Type IV secretory pathway, VirD2 component (...	6.00E-149
VJ4_contig00664_218_1015_1	TIGR01669: phage_XkdX: phage uncharacterized protein, XkdX family	0.01	YP_002501423.1| FkbM family methyltransferase [Methylobacter...	4.00E-31
VJ4_contig00710_1449_257_2	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	6.20E-23	WP_009570795.1| integrase [Celeribacter baekdonensis] >gb|EK...	2.00E-169
VJ4_contig00710_1449_257_2	TIGR02224: recomb_XerC: tyrosine recombinase XerC	1.60E-13	WP_009570795.1| integrase [Celeribacter baekdonensis] >gb|EK...	2.00E-169
VJ4_contig00710_1449_257_2	TIGR02225: recomb_XerD: tyrosine recombinase XerD	6.20E-13	WP_009570795.1| integrase [Celeribacter baekdonensis] >gb|EK...	2.00E-169
VJ4_contig00710_1449_257_2	TIGR02249: integrase_gron: integron integrase	2.10E-07	WP_009570795.1| integrase [Celeribacter baekdonensis] >gb|EK...	2.00E-169
VJ4_contig00753_527_1045_2	PF00436: Single-strand binding protein family	7.10E-39	WP_009827963.1| single-stranded DNA-binding protein [Rhodoba...	1.00E-74
VJ4_contig00766_1010_585_2	TIGR02763: chlamy_scaf: scaffolding protein	0.013	WP_009804468.1| C4-dicarboxylate ABC transporter [Oceanicola...	3.00E-78
VJ4_contig00779_263_1255_1	PF00665: rve: integrase core domain	3.70E-11	WP_009803713.1| transposase [Oceanicola batsensis] >gb|EAQ02...	0.00E+00
VJ4_contig00779_263_1255_1	PF06056: Terminase_5: putative terminase, ATPase subunit, gpP-lie	0.017	WP_009803713.1| transposase [Oceanicola batsensis] >gb|EAQ02...	0.00E+00
VJ4_contig00849_296_54_1	PF00665: rve: integrase core domain	0.00075	WP_007153488.1| transposase [Marinobacter algicola] >gb|EDM4...	2.00E-49
VJ4_contig00909_1319_792_2	PF07232: Putative rep protein (DUF1424)	0.0078	WP_009801190.1| hypothetical protein [Oceanicaulis sp. HTCC2...	6.00E-75
VJ4_contig00933_1_1227_1	PF00665: rve: integrase core domain	3.60E-24	WP_008557381.1| transposase [Rhodobacterales bacterium Y4I] ...	3.00E-142
VJ4_contig00946_323_1_1	TIGR01610: phage_O_Nterm: phage replication protein O, N-terminal domain	0.0056	WP_009802230.1| MarR family transcriptional regulator [Ocean...	4.00E-53

Table 3.3 (Continued)

VJ4_contig00987_1_857_1	TIGR02225: recomb_XerD: tyrosine recombinase XerD	0.049	WP_009810478.1| hypothetical protein [Roseobacter sp. MED193...	0.00E+00
VJ4_contig00987_1_857_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	5.20E-10	WP_009810478.1| hypothetical protein [Roseobacter sp. MED193...	0.00E+00
VJ4_contig00995_645_1_1	PF01464: SLT: transglycosylase SLT domain	8.10E-08	YP_006573690.1| TypeIV secretory lytic transglycosylase-like protein [Phaeoba...	1.00E-101
VJ4_contig01007_1_771_1	TIGR02218: phg_TIGR02218: phage conserved hypothetical protein BR0599	3.30E-70	WP_009803363.1| hypothetical protein [Oceanicaulis sp. HTCC2...	2.00E-61
VJ4_contig01007_783_1184_2	TIGR02219: phage_NlpC_fam: putative phage cell wall peptidase, NlpC/P60 family	2.40E-62	WP_022701825.1| peptidase [Oceanicaulis alexandrii]	7.00E-49
VJ4_contig01019_645_1_1	PF05119: Terminase_4: phage terminase, small subunit	0.029	WP_008173889.1| methyl-accepting chemotaxis protein [Marinob...	6.00E-106
VJ4_contig01034_1248_1_1	TIGR02642: phage_xxxx: uncharacterized phage protein	0.01	WP_007815625.1| ribonuclease [Roseobacter sp. AzwK-3b] >gb|E...	0.00E+00
VJ4_contig01052_284_1241_1	PF08774: VRR_NUC: VRR-NUC domain	0.0018	WP_023916667.1| DEAD/DEAH box helicase [Rhodobacter capsulat...	4.00E-142
VJ4_contig01102_1_1217_1	TIGR01644: phage_P2_V: phage baseplate assembly protein V	0.0087	WP_008175737.1| type IV secretion protein Rhs [Marinobacter ...	0.00E+00
VJ4_contig01102_1_1217_1	PF04717: Phage_base_V: phage-related baseplate assembly protein	5.40E-25	WP_008175737.1| type IV secretion protein Rhs [Marinobacter ...	0.00E+00
VJ4_contig01102_1_1217_1	PF05954: Phage_GPD: phage late control gene D protein	3.30E-07	WP_008175737.1| type IV secretion protein Rhs [Marinobacter ...	0.00E+00
VJ4_contig01102_1_1217_1	PF06890: Phage_Mu_Gp45: bacteriophage Mu Gp45 protein	0.023	WP_008175737.1| type IV secretion protein Rhs [Marinobacter ...	0.00E+00
VJ4_contig01109_186_1208_2	PF05876: Terminase_GpA: phage terminase large subunit GpA	9.30E-56	YP_003963669.1| Phage terminase GpA [Ketogulonicigenium vulg...	4.00E-134
VJ4_contig01128_657_1_1	PF08273: Prim_Zn_Ribbon: zinc-binding domain of primase-helicase	0.014	YP_682097.1| DNA primase [Roseobacter denitrificans OCh 114]...	3.00E-117
VJ4_contig01131_139_441_1	TIGR01554: major_cap_HK97: phage major capsid protein, HK97 family	0.11	WP_009803065.1| prolipoprotein diacylglyceryl transferase [O...	1.00E-17
VJ4_contig01143_845_213_1	PF00665: rve: integrase core domain	5.90E-24	YP_004690060.1| integrase [Roseobacter litoralis Och 149] >r...	2.00E-146
VJ4_contig01207_1178_797_3	PF02661: Fic: Fic/DOC family	4.90E-10	WP_020230201.1| death-on-curing protein [Acidovorax sp. MR-S...	1.00E-21
VJ4_contig01286_1_382_1	PF00436: Single-strand binding protein family	0.0088	WP_008270181.1| hypothetical protein [Flavobacteriales bacte...	2.00E-75
VJ4_contig01326_139_819_1	PF07471: Phage_Nu1: phage DNA packaging protein Nu1	7.30E-08	YP_355289.1| putative phage terminase large subunit [Rhodoba...	1.00E-16
VJ4_contig01326_821_1136_2	PF05876: Terminase_GpA: phage terminase large subunit GpA	4.90E-13	YP_003963669.1| Phage terminase GpA [Ketogulonicigenium vulg...	5.00E-28
VJ4_contig01359_1_922_1	PF02661: Fic: Fic/DOC family ("Filamentation Induced by Camp")	2.40E-16	BAH89846.1| cell filamentation protein [uncultured bacterium]	2.00E-148
VJ4_contig01372_1113_186_1	PF01018: GTP1_OBG: GTP1/OBG	1.10E-47	WP_008269232.1| GTPase CgtA [Flavobacteriales bacterium ALC-...	0.00E+00
VJ4_contig01401_1_638_1	PF01018: GTP1_OBG: GTP1/OBG	0.035	WP_008174106.1| diguanylate cyclase [Marinobacter manganoxyd...	2.00E-174
VJ4_contig01435_1087_731_3	TIGR02224: recomb_XerC: tyrosine recombinase XerC	0.00057	WP_004530414.1| integrase [Burkholderia pseudomallei] >gb|ED...	2.00E-09
VJ4_contig01435_1087_731_3	TIGR02225: recomb_XerD: tyrosine recombinase XerD	0.045	WP_004530414.1| integrase [Burkholderia pseudomallei] >gb|ED...	2.00E-09
VJ4_contig01435_1087_731_3	TIGR02249: integrase_gron: integron integrase	0.0063	WP_004530414.1| integrase [Burkholderia pseudomallei] >gb|ED...	2.00E-09
VJ4_contig01435_1087_731_3	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	5.90E-09	WP_004530414.1| integrase [Burkholderia pseudomallei] >gb|ED...	2.00E-09
VJ4_contig01465_94_1076_1	PF00239: Resolvase: resolvase, N terminal domain	0.039	WP_020894619.1| 6-phosphofructokinase [Winogradskyella psych...	0.00E+00

Table 3.3 (Continued)

VJ4_contig01518_117_1_1	PF07022: Phage_CI_repr: bacteriophage CI repressor protein	0.065	WP_021100344.1| Transcriptional regulator AglR, LacI family ...	4.00E-14
VJ4_contig01545_869_1054_2	TIGR02224: recomb_XerC: tyrosine recombinase XerC	0.00019	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	4.00E-13
VJ4_contig01545_869_1054_2	TIGR02225: recomb_XerD: tyrosine recombinase XerD	1.20E-07	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	4.00E-13
VJ4_contig01545_869_1054_2	PF02899: Phage_integr_N: phage integrase, N-terminal SAM-like domain	2.80E-07	WP_009801770.1| tyrosine recombinase XerD [Oceanicaulis sp. ...	4.00E-13
VJ4_contig01550_423_1_1	TIGR02215: phage_chp_gp8: phage conserved hypothetical protein, phiE125 gp8 family	1.00E-29	WP_009803354.1| hypothetical protein [Oceanicaulis sp. HTCC2...	1.00E-36
VJ4_contig01550_423_1_1	PF05135: Phage_connect_1: Phage gp6-like head-tail connector protein	0.00064	WP_009803354.1| hypothetical protein [Oceanicaulis sp. HTCC2...	1.00E-36
VJ4_contig01550_942_610_2	PF05065: Phage_capsid: phage capsid family	3.90E-30	WP_009803353.1| phage capsid protein [Oceanicaulis sp. HTCC2...	2.00E-64
VJ4_contig01550_942_610_2	TIGR01554: major_cap_HK97: phage major capsid protein, HK97 family	9.90E-48	WP_009803353.1| phage capsid protein [Oceanicaulis sp. HTCC2...	2.00E-64
VJ4_contig01562_1_396_1	TIGR02419: C4_traR_proteo: phage/conjugal plasmid C-4 type zinc finger protein, TraR family	7.60E-06	WP_009801381.1| molecular chaperone DnaK [Oceanicaulis sp. H...	8.00E-76
VJ4_contig01580_148_1043_1	PF00665: rve: integrase core domain	9.50E-22	YP_001170391.1| plasmid replication initiator protein-like p...	5.00E-109
VJ4_contig01600_1020_1_1	PF08462: Carmovirus coat protein	0.087	WP_008207196.1| arginyl-tRNA synthetase [Roseobacter sp. SK2...	0.00E+00
VJ4_contig01607_1_398	PF00979: Reovirus_capsid	1.50E-02	WP_011537503| hypothetical protein [Ruegeria spp. TM1040]	8.50E-02
VJ4_contig01608_1_1022_1	PF06322: Phage_NinH: phage protein NinH	0.0011	WP_007350778.1| cytosine deaminase [Marinobacter sp. ELB17] ...	7.00E-136
VJ4_contig01637_697_1026_3	PF00665: rve: integrase core domain	1.90E-05	EGQ64217.1| integrase catalytic subunit [Acidithiobacillus sp...	2.00E-84
VJ4_contig01663_279_860_2	PF00665: rve: integrase core domain	4.80E-12	YP_008972529.1| transposase [Leisingera methylohalidivorans ...	7.00E-99
VJ4_contig01675_374_1_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	9.90E-05	WP_008554125.1| integrase [Rhodobacterales bacterium Y4I] >g...	3.00E-19
VJ4_contig01716_1_1007_1	PF04582: Reovirus sigma C capsid protein	0.42	WP_009803219.1| chromosome segregation protein SMC [Oceanica...	1.00E-30
VJ4_contig01753_992_18_1	TIGR02224: recomb_XerC: tyrosine recombinase XerC	0.024	WP_009802760.1| putative fatty-acid--CoA ligase [Oceanicauli...	1.00E-148
VJ4_contig01775_1_237_1	PF00959: Phage_lysozyme: phage lysozyme	3.60E-10	WP_007799815.1| Phage-related lysozyme [Pelagibaca bermudens...	7.00E-27
VJ4_contig01784_1_989_1	PF02368: Big_2: bacterial Ig-like domain, group 2	0.56	WP_008175271.1| hypothetical protein [Marinobacter manganoxy...	0.00E+00
VJ4_contig01843_1_574_1	PF05565: Sipho_Gp157: siphovirus Gp157	0.0057	WP_023659525.1| Plasmid replication region DNA-binding prote...	1.00E-117
VJ4_contig01851_326_971_2	PF05954: Phage_GPD: phage late control gene D protein	0.00016	WP_008175938.1| type IV secretion protein Rhs [Marinobacter ...	3.00E-132
VJ4_contig01888_960_684_3	TIGR01610: phage_O_Nterm: phage replication protein O, N-terminal domain	0.032	WP_008561894.1| MarR family transcriptional regulator [Ruege...	2.00E-22
VJ4_contig01980_1_614_1	PF00665: rve: integrase core domain	2.20E-19	WP_003167669.1| integrase core domain protein [Brevundimonas...	6.00E-94
VJ4_contig02011_774_1_1	PF09077: Phage-MuB_C: Mu B transposition protein, C terminal	0.0073	WP_022699990.1| flagellin [Oceanicaulis alexandrii]	1.00E-106
VJ4_contig02081_553_1_1	PF02661: Fic: Fic/DOC family ("Filamentation Induced by Camp")	0.016	WP_022700071.1| ComL family lipoprotein [Oceanicaulis alexan...	3.00E-86
VJ4_contig02087_624_1_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	0.064	WP_007812904.1| glmZ(sRNA)-inactivating NTPase [Roseobacter ...	1.00E-60
VJ4_contig02138_1_885_1	PF09158: MotCF: bacteriophage T4 MotA, C-terminal	0.025	WP_008271852.1| polyphosphate kinase [Flavobacteriales bacte...	0.00E+00
VJ4_contig02148_696_881	PF08398: Parvovirus coat protein VP1	5.70E-02	WP_022728538| hypothetical protein [Fondinicurvata sedimis]	5.00E-15

Table 3.3 (Continued)

VJ4_contig02166_771_172_2	PF04582: Reovirus sigma C capsid protein	8.60E-07	WP_020898177.1| Flagellar motor rotation protein MotB [Winog...	3.00E-91
VJ4_contig02264_838_539_3	PF05930: Phage_AlpA: transcriptional regulator, AlpA family	0.00087	YP_004263898.1| DNA-binding domain-containing protein [Cellu...	6.00E-44
VJ4_contig02265_1_852_1	TIGR01725: phge_HK97_gp10: phage protein, HK97 gp10 family	0.013	WP_008271211.1| cytochrome C biogenesis protein [Flavobacter...	5.00E-157
VJ4_contig02301_839_1_1	PF04582: Reovirus sigma C capsid protein	0.012	YP_757808.1| OmpA/MotB domain-containing protein [Maricaulis...	2.00E-75
VJ4_contig02393_588_256_2	PF09114: MotA_activ: transcription factor MotA, activation domain	0.019	WP_007121018.1| ArsR family transcriptional regulator [Ocean...	3.00E-46
VJ4_contig02466_793_1_1	PF07471: Phage_Nu1: phage DNA packaging protein Nu1	0.0074	WP_019387545.1| molecular chaperone DnaK [Flavobacteriaceae ...	8.00E-156
VJ4_contig02522_696_1_1	PF06810: Description: Phage_GP20: phage minor structural protein GP20	0.24	ETS31995.1| Plasmid replication region [Photorhabdus temperat...	2.00E-23
VJ4_contig02526_275_776_2	PF01018: GTP1_OBG: GTP1/OBG	2.00E-58	WP_022693350.1| GTPase CgtA [Ponticaulis koreensis]	2.00E-56
VJ4_contig02561_767_156_2	PF06056: Terminase_5: putative terminase, ATPase subunit, gpP-like	0.052	WP_008225238.1| LuxR family transcriptional regulator [Roseo...	6.00E-111
VJ4_contig02579_7_564_1	PF02661: Fic: Fic/DOC family ("Filamentation Induced by Camp")	2.10E-07	YP_005371724.1| filamentation induced By CAMP protein Fic [C...	7.00E-47
VJ4_contig02590_1_756_1	PF04582: Reovirus sigma C capsid protein	0.057	WP_009807235.1| capsule polysaccharide transporter [Oceanico...	5.00E-100
VJ4_contig02609_272_490_2	TIGR02216: phage conserved hypothetical protein	1.90E-24	YP_008975554.1| phage hypothetical protein [Phaeobacter gall...	2.00E-11
VJ4_contig02624_353_700_2	PF01766: Birnavirus VP2 protein	0.017	WP_022700682.1| hypothetical protein [Oceanicaulis alexandrii]	7.00E-07
VJ4_contig02639_365_743_3	PF05136: Phage_portal_2: phage portal protein, lambda family	2.90E-08	WP_008333601.1| hypothetical protein [Maritimibacter alkalip...	4.00E-23
VJ4_contig02665_1_737_1	PF00665: rve: integrase core domain	3.40E-32	WP_009159952.1| transposase [Thalassobium sp. R2A62] >gb|EET...	3.00E-168
VJ4_contig02746_388_567_2	PF05930: Phage_AlpA: transcriptional regulator, AlpA family	6.30E-20	YP_957938.1| phage transcriptional regulator AlpA [Marinobac...	4.00E-12
VJ4_contig02761_1_364_1	PF05929: Phage_GPO: phage capsid scaffolding protein (GPO) serine peptidase	0.019	WP_023008699.1| chemotaxis protein [Marinobacter]	1.00E-40
VJ4_contig02784_703_232_2	PF06056: Terminase_5: putative terminase, ATPase subunit, gpP-like	0.023	WP_010138654.1| LuxR family transcriptional regulator [Ocean...	5.00E-52
VJ4_contig02826_1_691_1	TIGR02419: C4_traR_proteo: phage/conjugal plasmid C-4 type zinc finger protein, TraR family	0.028	WP_009808543.1| osmotically inducible protein C [Roseobacter...	3.00E-102
VJ4_contig02831_136_495_2	PF05973: Gp49: Gp49-like PF05973 family	3.40E-19	YP_006964838.1| Tad-like protein [Paracoccus marcusii] >ref|...	1.00E-53
VJ4_contig02839_51_686_1	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	0.015	WP_005979190.1| integrase [Ruegeria lacuscaerulensis] >gb|EE...	2.00E-110
VJ4_contig02945_652_144_2	PF01510: Amidase_2: N-acetylmuramoyl-L-alanine amidase	0.069	YP_167252.1| methyltransferase, FkbM family [Ruegeria pomero...	2.00E-32
VJ4_contig02951_407_1_1	PF02336: Densovirus Capsid protein VP4	0.042	WP_008271794.1| methylmalonyl-CoA carboxyltransferase [Flavo...	1.00E-68
VJ4_contig03040_92_604_2	TIGR02224: recomb_XerC: tyrosine recombinase XerC	0.0042	WP_009416122.1| site-specific recombinase, phage integrase f...	1.00E-50
VJ4_contig03040_92_604_2	TIGR02225: recomb_XerD: tyrosine recombinase XerD	0.0032	WP_009416122.1| site-specific recombinase, phage integrase f...	1.00E-50
VJ4_contig03040_92_604_2	TIGR02249: integrase_gron: integron integrase	1.40E-06	WP_009416122.1| site-specific recombinase, phage integrase f...	1.00E-50
VJ4_contig03040_92_604_2	PF00589: Phage_integrase: site-specific recombinase, phage integrase family	9.20E-07	WP_009416122.1| site-specific recombinase, phage integrase f...	1.00E-50
VJ4_contig03105_1_345_1	PF08765: Mor: mor transcription activator family	0.0049	WP_020896441.1| Transcription termination factor Rho [Winogr...	2.00E-59

Table 3.3 (Continued)

VJ4_contig03122_440_150_2	PF04582: Reovirus sigma C capsid protein	0.004	WP_017732158.1| hypothetical protein [Nafulsella turpanensis]	1.00E-12
VJ4_contig03127_1_555_1	PF00665: rve: integrase core domain	3.10E-10	WP_007153488.1| transposase [Marinobacter algicola] >gb|EDM4...	5.00E-105

CONCLUSION

This thesis represents a unified body of work on small, ssDNA bacteriophages in the marine environment, an understudied component of the global virome. A primer set targeting the gokushoviruses has demonstrated the ubiquity of these phages. We found gokushoviruses in every aquatic sample tested, and in some samples such as pelagic, deep-sea sediments they were found in staggering diversity. The same primers coupled with physical oceanographic data from BATS were able to not only detect gokushoviruses at surface and depth, but also to test and support a hypothesis about the annual and interannual community dynamics of these viruses. In addition, advances were made toward the establishment of a highly novel, ssDNA phage-host system in the globally important *Synechococcus* genus. However, much more work remains to be done to redress the knowledge gap between small, ssDNA phages and tailed, dsDNA phages.

For the environmental gokushoviruses the upmost need is to establish a phage-host model system in culture which would allow for a centuries worth of assays to be deployed to shed light on their particular life cycle and replicative dynamics. On this front, Chapter 1 can be instructive for future work. The finding in Figure 1.3, of a dichotomous topology with an emergent 'environmental' clade, in which the *Bdellovibrio* phage BdφMH2k is the only cultured isolate, is particularly instructive. This suggests that the traditionally-regarded niche of gokushoviruses as infecting obligate intracellular parasitic bacteria may indeed remain true in marine environments. Thus efforts to culture marine gokushoviruses should focus on plaque assaying with marine bacteria that intracellularly parasitize other marine bacteria, in the style of *Bdellovibrio bacteriovorus.* Efforts to culture *Gokushovirinae,* not presented in this thesis, employed Cesium Chloride density centrifugation to try to bias the viral inoculum of the plaque assay towards the gokushoviruses. Given the correct host, this would be an efficient direction for future work.

Several new technologies relying on proximity association of phages to their hosts also hold promise for identifying the host bacterial taxa of gokushoviruses. The 'polony' technique (a portmanteau of 'pcr' + 'colony'; Mitra and Church 1999) is one such promising technology in which exponential proliferations of identical PCR amplicons are immobilized in one spot ('colony') within a thin polyacrylamide gel matrix. By probing gokushoviral polonies (generated using the primers from Chapters 1 & 2) and bacterial polonies (generated from 16s rRNA primers) with differently colored hybridization probes, it would be possible to detect proximal associations between the phage and the host. Many such proximal associations would build statistical support for a host relationship. A related but distinct technique of phage-applied fluorescence *in situ* hybridization (phageFISH; Allers et al 2013) uses digoxigenin-labeled probes that hybridize both genes within the phage-of-interest and the 16s rRNA gene of the host with different colors. Through use of horseradish peroxidase-conjugated antibodies the fluorescent signal of the probes is massively amplified allowing real-time observation of phage-host interaction cycles.

One of the most interesting unresolved aspects of the gokushoviruses is the hypervariability of the three-fold insertion loop, believed to play a role in host binding and specificity. Assuming that the whole genome is not subject to hypervariability, then the mechanism by which nucleotide-level hypervariability is generated in only one discrete portion of the viral genome is a mystery. One potential mechanism that has precedence in the literature is the possibility that gokushoviruses contain a diversity-generating retroelement (Doulatov et al 2004). Clever bioinformatics search strategies of a large assemblage of full gokushoviral genomes might facilitate the search for such a mechanism. Here again though, an environmental culture system would be of great utility.

The N phage presented in Chapter 3 remains tantalizingly close to characterization. Currently, the greatest promise lies in the proteomic approach. If the distinct protein populations visualized in Figure 3.5 can be *de novo* sequenced by MALDI-TOF mass spectrometry, then this may provide an immediate indication of where in the metagenome the cryptic phage genome is located. Perhaps we will gain only a 'foothold' in the metagenome that will inform further tests to determine whether the implicated metagenomic sequence is actually belonging to the N phage. In either case, the novelty of the phage and the global biogeochemical importance of the *Synechococcus* host, makes this an important puzzle to solve.

The pursuit of viral ecology has come to be regarded as the notional equivalent of attempting to discover the extent of an indeterminately large, dark space having found oneself in the middle of the dark with only a handheld flashlight. Some 'light beams', such as metagenomics, are powerful and diffuse, whereas others, like PCR, are focused and intense. Alone, we can only light up a small region of the viral darkspace. It is only through the concerted effort of the many, that the extent and shape of the space may become known.

References:

Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, Amann R, Sullivan MB (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology* **15:** 2306-2318.

Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF (2004). Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431:** 476-481.

Mitra RD, Church GM (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research* ***27*****(24),** e34-e39.

Parsons RJ, Breitbart M, Lomas MW, Carlson CA (2011). Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *The ISME Journal* **6:** 273-284.

APPENDIX #1: Viral purification and concentration

Concentration of virus sample with polyethylene glycol (PEG):

- Prepare the viral sample of interest, i.e. by prophage induction or infectious lysis, or environmental sample. It is helpful to keep track of the volumes of lysate (measure when prepping samples) so you don't have to measure volumes later. Make sure to check that the viruses are present and count the abundance by SYBR Gold before proceeding.
- Centrifuge at 9500 x g for 10 min to remove cell debris
- Pool lysates if necessary and 0.2 µm filter (may be necessary to do sequential filtration like 1 µm, 0.4 µm then 0.2 µm for large volumes or cultures with lots of debris).
- Measure volume of lysate/sample (if not already known). Add sodium chloride to 1M final concentration (29.2 g NaCl/500 ml). Swirl to mix.
- DNase/RNase digest to remove any non-encapsidated (i.e. non-viral) nucleic acids: RQ-1 DNase, 2.5 µl/ml and RNase One 0.1 µl/ml of lysate
- Digest at room temp, 1 hour.
- Take the pooled lysates and add Polyethylene Glycol at 10% wt/vol (i.e. 10 g in 100 ml). Use PEG 6000 for small viruses and PEG 8000 for larger viruses.
- Dissolve the PEG completely using a sterile stir bar (or by shaking). Pour the viral sample with PEG into the desired centrifuge bottles and refrigerate the samples, on ice for at least two hours. Overnight is better. Be sure to mark the bottles to know where the viral pellet will be.
- Centrifuge the samples at 9500 x g at 4°C x 20 min.
- Aspirate the supernatant with a sterile Pasteur pipette. Invert and drain for several minutes.
- Re-suspend the viral pellet in 300-500 µl of the appropriate diluent (usually sterile, 0.2 µm filtered ASW for marine viruses). Let the pellet soak for approx 1 hour, be sure to also thoroughly wash the sides of the bottle to retrieve as many viruses as possible.
- Place the resuspended viruses in a clean tube (size depends on total volume). Combine with an equal volume of chloroform, vortex gently for 30 seconds. Centrifuge at 3000 x g for 10 min.
- Aspirate the aqueous phase (top, contains your viruses, the PEG goes with the chloroform)
- Check the precipitated viruses again by SYBR. Also enumerate here, can calculate your percent recovery.
- Purify the viruses using a Cesium Chloride density gradient. if indicated.

APPENDIX #2: Formamide extraction of viral DNA

Extract DNA from Viruses or GTA's after PEG precipitation:

- **Measure the volume of the PEG purified virus/GTA sample. It is usually 500 µl after the PEG step but is sometimes more if samples are pooled.**
- **Add: 0.1 volume 2M Tris, 0.5 volumes 0.5M EDTA, 1 volume of formamide and 2 µl of 10 mg/ml glycogen per milliliter of sample. This step usually needs to be done in a 15 ml centrifuge tube to accommodate the volumes required.**
- Incubate at room temp for 30 mins. (sometimes more DNA is recovered with incubation at 65°C, determine what works better for your samples).
- Add 2 volumes of room temp 100% ethanol
- Centrifuge at 12,000 rpm x 20 min to precipitate the DNA. Aspirate the supernatant and drain.
- Wash pellet with 70% ethanol x 2. Resuspend in 363.5 µl of TE.
- Transfer the DNA to 1.5 ml centrifuge tube. Add 19.2 µl 10% SDS and 1.9 µl of 20 mg/ml proteinase K. Place in a 37°C water bath x 1 hour.
- Add 64.1 µl 5M NaCl and mix, Add 51.3 µl of CTAB/NaCl solution (see recipe at end, this solution is very viscous so be careful pipetting). Place in 65°C water bath x 10 min.
- Place the sample in a phase-lock gel tube (should be a total volume of 500 µl) Add equal volume chloroform (500 µl), stir (I use a needle to stir the gel) and shake to mix well and centrifuge at max speed in microfuge x 4 min.
- Aspirate the supernatant and put in fresh phase-lock tube. Add equal volume of phenol/chloroform (250 µl of each) , mix and centrifuge x 4 min.
- Transfer to another fresh gel-lock tube. Add equal volume of chloroform again, mix and centrifuge 4 min. Aspirate and measure supernatant.
- Add 0.7 volumes of cold isopropanol, mix gently. Centrifuge in the cold (use refrigerated, pre-cooled centrifuge). Decant/aspirate the supernatant.
- Wash with 1 volume 70% ethanol, spin again, 5 min. Aspirate the alcohol carefully and allow the pellet to air dry for 10-15 minutes.
- Resuspend in desired volume of DI water and check quantity/quality with Nanodrop.

CTAB/NaCl solution: Dissolve 4.1 g NaCl in 80 ml DI water (or 2.05 g in 40 ml). Add 10 g (5 g) CTAB while stirring. Heat to 65°C to dissolve, then bring volume to 100 ml (50 ml).

Proteinase K: Is prepared in buffer in the left-hand freezer in the enzyme box. Thaw ahead of time. There is usually some precipitation, mix well (not vortex, it's an enzyme) before use. If you need to make more, the recipe is in the Sambrook manual.

APPENDIX #3: Total DNA extraction from large-scale cyanobacterial culture

Steps 4-5 are taken from Ausubel *et al,* Short Protocols in Molecular Biology, "Preparation of Plant DNA using CTAB" (p. 2-10) and steps 7-10 are taken from the DNeasy Blood & Tissue Handbook. Special thanks to Dr. Andy Millard for inspiration.

1. Starting with ~250ml (or more, or less) of high-density culture, centrifuge 15min at 5000 x g at room temperature in an acid washed centrifuge bottle.
2. Decant off supernatant but retain enough liquid to pipette up the cell pellet and transfer into a 50ml falcon tube. Weigh the pellet.
3. Repeatedly snap freeze and thaw the pellet by holding the bottom of the falcon tube in liquid nitrogen.
4. Add (4ml/gram of pellet) of prewarmed CTAB extraction solution (see next page for recipe); mix thoroughly and incubate ~1hr in a 65°C water bath, mixing occasionally.
5. Add 1 volume of 24:1 chloroform/isoamyl alcohol, mix by inversion and centrifuge 5 min at 7500 x g at 4°C. The result will be a clear aqueous top layer and a green chloroform bottom layer with a thick protein film at the interface.
6. Pipette off the clear aqueous top layer, conservatively avoiding the protein film, and transfer to a 15ml phase-lock tube. Mix with a sterile stir-needle. Centrifuge 5 min at 1500 x g at room temperature. The phase lock gel layer will form over the aqueous layer, so a pipette must be used to carefully punch through and withdraw the aqueous layer. (NB: This phase-lock step may be unnecessary)
7. (Begin Step 2 of "Purification of Total DNA from Animal Tissues" in the DNeasy Blood & Tissue Handbook) Add 10% by volume of Proteinase K (conc?) and incubate >1hr at 56°C, mixing occasionally.
8. Vortex well. Add 1 volume of Buffer AL, vortex. Add 1 volume 100% Ethanol, vortex.
9. Pipette the mixture from step 8, ≤700µl at a time, into a DNeasy Mini spin column placed in a 2ml collection tube. Centrifuge for 1 min at 6000 x g. Discard flow through and reload onto the column. Depending on the starting volume from step 8, this step will repeat 10-20 times.
10. Continue with step 5 of the Animal Tissues protocol from p.30 of the DNeasy Blood & Tissue Handbook. The final elution in step 7 should be performed with 100µl of prewarmed TE buffer after 5 minutes of room temperature incubation on the membrane.

CTAB extraction solution
-2% (w/v) CTAB powder
-100mM Tris-Cl, pH 8
-20mM EDTA, pH 8
-1.4M NaCl
-adjust volume to 10ml with nuclease free water
-0.2µm filter
-Immediately before use add 2% 2-mercaptoethanol

APPENDIX #4: Protocol for protein isolation of large-scale N phage induction

This protocol is for concentrating and extracting viral protein, subsequent to the "Viral Purification and Concentration" protocol, attached as Appendix 1 to this thesis. *** denotes recommended epifluorescence checkpoint.

1. Having incubated the 0.2μm-filtered viral sample with PEG 6000 (10% w/v) overnight at 4°C in 500ml acid-washed centrifuge bottle, centrifuge the samples at 9500 x g at 4°C for 20min.
2. Aspirate off the supernatant with a sterile Pasteur pipette, holding the bottle at an angle, pellet up. Invert and drain the centrifuge bottle for several minutes.
3. Resuspend the viral pellets with enough 0.2μm-filtered SM buffer such that the total volume pooled across all bottles is ~6ml of SM (e.g. 2ml per bottle if two bottles were used).
4. Incubate the pellet at room temperature for ≥1hr (oscillator recommended. ***(1:100 dilution recommended)
5. Add sterile 1.7g/ml CsCl to the ~6ml of SM to bring the density to 1.15g/ml
6. In a clear Beckman Ultra-Fuge™ tube build density gradient using sterile, 0.2μm-filtered density fractions as follows: 2ml x 1.7g/ml; 2ml x 1.5g/ml; 2ml x 1.3g/ml. Density boundaries should be visible.
7. Load the density-adjusted, SM viral suspension onto the CsCl gradient (I recommend using a wide-mouthed 1000ml pipette tip). Use waste CsCl to counterbalance another tube to within 0.01grams.
8. Using the SwTi40 rotor, centrifuge at 29,000rpm for 3hrs at 20°C
9. Gently remove the tube and clean the bottom with EtOh
10. With 10 or more sterile microfuge tubes open and pre-labelled in sequential order, place the Ultra-Fuge™ tube in a stand and puncture the bottom of the tube with an 18G needle.
11. Collect 0.5-1ml of drainage in each microfuge tube. ***(1:200 dilution recommended)
12. Into a pre-conditioned Float-A-Lyzer G2 (Spectrum Labs, Rancho Dominguez, CA: see manual for conditioning instructions) combine the two or three CsCl drainage fractions with the highest viral titer.
13. Dialyze the filled Float-A-Lyzer dialysis buffer (see recipe below) overnight. Do a buffer change after approximately the first hour.

Note: the dialysis steps may be unnecessary at the expense of viral loss. It may be possible to proceed directly to centrifugal concentration of the high viral titer CsCl-collected fractions.

14. Transfer the full contents of the Float-A-Lyzer by pipette into an Amicon Ultra-15 Centrifugal Filter with a 50kDa cutoff. ***
15. Centrifuge the Amicon filter in a fixed angle rotor at 5,000 x g for 15min, with the filter panel oriented tangentially to the circumference of the rotor. Repeat until total concentrate volume is ~100µl.

Note: the viral concentrate can be stored overnight, but the protein gel should be run as soon as possible. The following steps are tailored to the NuPAGE Bis-Tris precast SDS-PAGE gel with 10 wells x 1mm thickness (well capacity ~30µl).

16. For every 20µl of viral concentrate harvested in step 15, add 7µl of 4xLDS buffer and 3µl of 10xReductant for a total volume of 30µl per well. Incubate the mixture at 70°C for 10min. (The Mark12 ladder does not require any added reagents or preparation and may be added to the NuPAGE gel directly).
17. Load each 30µl aliquot into a well. Leaving a blank lane between every Nphage and non-Nphage sample.
18. Fill the NuPAGE gel rig with ~800ml of running buffer (see below). Run the gel for approximately 45 minutes with current and voltage specified by NuPAGE manufacturer (Life Technologies, Grand Island, NY).
19. Remove the gel from the rig. Using a sterile blade crack the plastic casing at the corners and then along the seams. Onto a clean glass pane, remove the gel from the case.
20. Stain and destain the gel with Sypro Ruby Protein Stain™ according to manufacturers specifications (Life Technologies, Grand Island, NY).

Dialysis solution ingredients per 1 liter (prepare 2 liters per dialysis in separate 1 liter cylindrical beaker)

-1 liter MilliQ water
-10 mM NaCl
-50 mM of Tris-HCl
-10 mM of MgCl

Running buffer per 1 liter

-950ml of MilliQ water
-50ml of 20x concentrated MOPS running buffer